

# Multiobjective Clustering with SVM Based Ensembling for Analysis of Gene Expression Data

**Ujjwal Maulik**

Department of Computer Science and Engineering  
Jadavpur University  
Kolkata 700032, INDIA  
Email: [umaulik@cse.jdvu.ac.in](mailto:umaulik@cse.jdvu.ac.in)

# Outline

## ① Background and Motivation

Clustering

Motivation

Genetic Algorithm

Multiobjective Optimization

MOO Definitions

NSGA-II

## ② Multiobjective Fuzzy Clustering

Initial Population

Fitness Computation

Genetic Operators

Combining Pareto-optimal Clustering Solutions

## ③ Application to Microarray Gene Expression Data

Microarray Gene Expression Data

Data Sets

Experimental Results

## ④ Summary

# Data Clustering

- Clustering is a popular **unsupervised** pattern classification technique which partitions the input space containing  $n$  objects into  $K$  regions based on some **similarity/dissimilarity** measure.
  - The value of  $K$  may or may not be known *a priori*.
- Output of a clustering technique is a  $K \times n$  matrix  $U = [u_{ki}]$ .
  - $u_{ki}$  denotes the membership degree of  $i$ th object to the  $k$ th cluster.
  - For **crisp clustering**,  $u_{ki} \in \{0, 1\}$ .
  - For **fuzzy clustering**,  $0 < u_{ki} < 1$ . (better suited for noisy data and overlapping clusters).

# Fuzzy C-means Clustering I

- Given  $K$  (number of clusters), the Fuzzy  $C$ -means (FCM) algorithm is implemented in 4 steps:

Step 1: Choose  $K$  random points as initial cluster centers.

Step 2: Compute the fuzzy membership values  $u_{ik}$  as follows:

$$u_{ik} = \frac{1}{\sum_{j=1}^K \left( \frac{D(v_i, x_k)}{D(v_j, x_k)} \right)^{\frac{2}{m-1}}}, \text{ for } 1 \leq i \leq K; \ 1 \leq k \leq n,$$

Step 3: Recompute the cluster centers  $v_i$  as follows:

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m}, \ 1 \leq i \leq K.$$

Step 4: Go back to Step 2, stop when no more change in the cluster centers.

# Fuzzy C-means Clustering II

- FCM algorithm minimizes the following criterion:

Global fuzzy cluster variance

$$J_m = \sum_{k=1}^n \sum_{i=1}^K u_{ik}^m D^2(v_i, x_k), \quad 1 \leq m \leq \infty.$$

## Fuzzy C-means – Limitations

- Gets stuck at local optima depending on the choice of the initial cluster centers.
  - Solution - Clustering based on global optimization technique such as Genetic Algorithm (GA).
- Optimizes single objective function  $J_m$  – May not be capable of capturing different characteristics of data sets.
  - Solution – Multiobjective Clustering.

### Why multiobjective clustering?

Simultaneous optimization of multiple objectives may lead to higher quality solutions and an improved robustness towards different data properties.

# Genetic Algorithm

- 1 Encode a possible solution of the problem in a form of **chromosome** (string).
- 2 Randomly generate a population of chromosomes.
- 3 Decode each chromosome to get an individual.
- 4 Evaluate the fitness of each individual.
- 5 Perform *selection, crossover and mutation*.
- 6 Repeat steps 3, 4 and 5 until a stop condition is true.
- 7 **Elitism** may be incorporated.
- 8 The best-fit chromosome of the last generation population is considered as the final solution.

# Multiobjective Optimization (MOO)

In many real world problems we have to simultaneously optimize two or more different objectives which are often competitive in nature.

Finding a single solution in these cases is very difficult.

Optimizing each criterion separately may lead to good value of one objective while some unacceptably low value of the other objective(s).

# MOO Problem Statement

- Find the vector of the decision variables:

$$\overline{x}^* = [x_1^*, x_2^*, \dots, x_n^*]^T$$

- which will satisfy the  $m$  inequality constraints:

$$g_i(\overline{x}) \geq 0, \quad i = 1, 2, \dots, m,$$

- the  $p$  equality constraints

$$h_i(\overline{x}) = 0, \quad i = 1, 2, \dots, p,$$

- and optimizes the vector function (consisting of  $k$  objective functions):

$$\overline{f}(\overline{x}) = [f_1(\overline{x}), f_2(\overline{x}), \dots, f_k(\overline{x})]^T.$$

# Domination Relation and Pareto-Optimality I

## Domination Relationship

Let  $a$  and  $b$  be two solutions. Then  $a$  is said to dominate  $b$  iff

$$\forall i \in \{1, \dots, k\}, f_i(b) \leq f_i(a)$$

and

$$\exists j \in \{1, \dots, k\}, f_j(b) < f_j(a).$$

i.e., for all functions  $f_i$ ,  $a$  has a higher or equal value than that of  $b$  and also there exists at least one function  $f_j$  for which  $a$ 's value is strictly greater than that of  $b$ .

## Domination Relation and Pareto-Optimality II

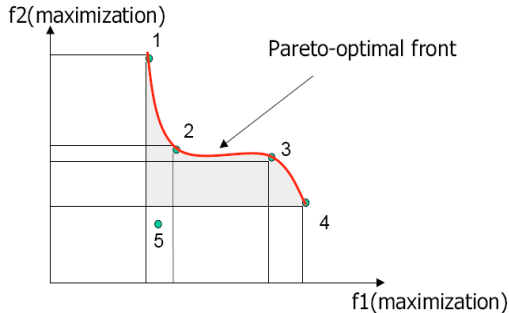
### Non-dominated Set

- Among a set of solutions  $P$ , the non-dominated set of solutions  $P'$  are those that are not dominated by any solution in the set  $P$ .
- A solution  $a$  is called non-dominating with respect to all the solutions if there exists no solution  $b$  that dominates  $a$ .

### Pareto-optimal Set

The non-dominated set of entire search space  $S$  is globally Pareto optimal set.

# Non-domination: Example



Solutions 1, 2, 3 and 4 are non-dominating to each other.

Solution 5 is dominated by 2, 3 and 4, not by 1.

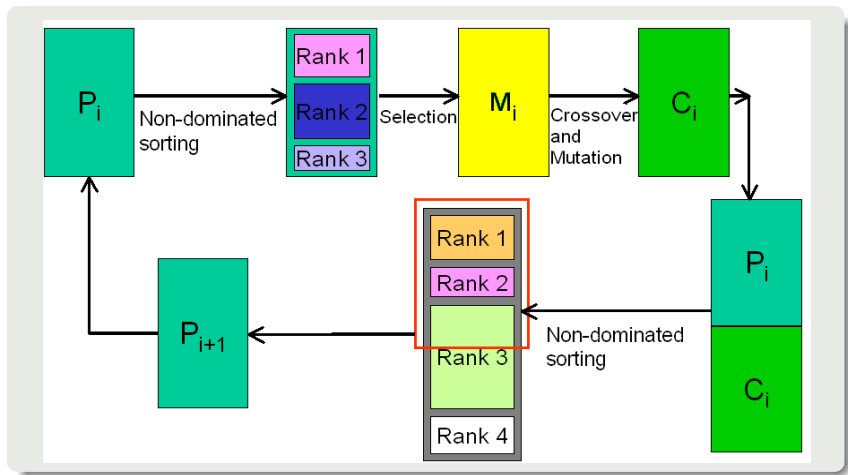
# Multiobjective Optimization Algorithms

- Multiobjective GAs are more popular primarily because of their population based nature.
- Available Algorithms
  - **Non-Pareto approach**
    - Vector Evaluated GA (VEGA): non-Pareto
  - **Pareto-based approach**
    - Non-dominated Sorting GA (NSGA and **NSGA-II**)
    - Niche Pareto GA (NPGA)
    - Strength Pareto Evolutionary Algorithm (SPEA and SPEA2)
    - Pareto Archived Evolutionary Strategy (PAES)
    - Pareto Envelope-based Selection Algorithm (PESA and PESA-II)
    - Archived Multiobjective Simulated Annealing (AMOSA)

# Non-dominated Sorting GA-II

- Proposed by K. Deb et. al. (2002).
- **Non-dominated Sorting**
  - It is based on several layers of classifications of the individuals.
  - Non-dominated individuals get a certain dummy fitness value (**Rank**) and then are removed from the population.
  - The process is repeated until the entire population is classified.
- **Diversity Maintenance**
  - Concept of **Crowding Distance** of individuals in a non-dominated front.
  - Selection based on Crowding distance.
- **Elitism**
  - Non-dominated individuals of parent and child populations are carried to the next generations.
- Time complexity:  $O(MN^2)$  (where  $M$  is the number of objectives and  $N$  is the population size).

# NSGA-II Flowchart



# NSGA-II based Multiobjective Fuzzy Clustering

- Chromosome Representation

- Cluster centers are encoded in the chromosomes.
- For a  $d$  dimensional space  $\text{length of chromosome} = d \times K$

$$\{(v_{11}, v_{12}, \dots, v_{1d}), (v_{21}, v_{22}, \dots, v_{2d}), \dots, (v_{K1}, v_{K2}, \dots, v_{Kd})\}$$

- Example

- Let  $d = 2$ ,  $K = 3$ .
- i.e., two-dimensional space, number of clusters = 3.
- Chromosome: 51.6 72.3 18.3 15.7 29.1 32.2 represents 3 cluster centers (51.6, 72.3), (18.3, 15.7) and (29.1, 32.2).

# Initial Population

- Each chromosome in the initial population encodes  $K$  random data points as  $K$  cluster centers.

```
For each chromosome  $i$  in the population
  For each cluster  $j$ 
     $p$  = randomly chosen point from the data set;
    Population[ $i$ ][ $j$ ] =  $p$ ;
  End
End
```

# Fitness Computation

This consists of three phases.

- **Phase 1:** Extract the cluster centers encoded in the chromosome and compute the fuzzy membership matrix.
- **Phase 2:** Recompute the cluster centers and update the chromosome with the new cluster centers. Recompute the fuzzy membership matrix.
- **Phase 3:** Fitness computation
  - **First objective:** Xie-Beni ( $XB$ ) cluster validity index

$$XB(U, V; X) = \frac{\sum_{i=1}^K (\sum_{k=1}^n u_{ik}^2 D^2(v_i, x_k))}{n(\min_{i \neq j} \{D^2(v_i, v_j)\})}$$

- **Second objective:** Fuzzy cluster variance

$$J_m = \sum_{j=1}^n \sum_{k=1}^K u_{kj}^m D^2(v_k, x_j)$$

- Both  $XB$  and  $J_m$  are to be minimized in order to obtain highly **compact** and **well-separated** clusters.

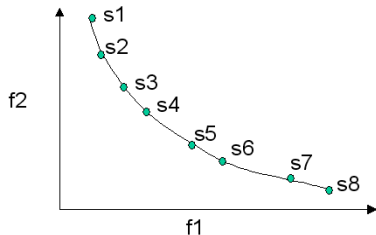
# Genetic Operators

- **Selection** – Crowded binary tournament selection.
- **Crossover** – Single point crossover with a fixed crossover probability.
  - For chromosomes of length  $K$ , a random integer  $p$  is generated in the range  $[1, K]$ . The portions of the chromosomes lying to the right of  $p$  are exchanged to produce two offspring.
  - Centers are considered indivisible.
- **Mutation** – Floating point mutation with fixed mutation probability.
  - A number  $\delta$  in the range  $[0, 1]$  is generated with uniform distribution.
  - If the value at a gene position is  $v$ , after mutation it becomes
$$v = v \pm 2\delta v, \quad \text{if } v \neq 0,$$
$$v = v \pm 2 * \delta, \quad \text{if } v = 0.$$
- Executed with fixed population size and for fixed number of generations

## Obtaining Final Solution from Non-dominated Front

- Multiobjective method produces a set of non-dominated solutions in the final generations. It is needed to obtain a solution from this set.
- For each non-dominated solution, first the clustering label vector is computed from the solution by assigning each point to the cluster to which it has the highest membership.
- Thereafter the label vectors are reordered so that they correspond to each other.
- Next, the points which are assigned to the same cluster by at least 50% of the clustering solutions are obtained.
- These points are taken as the training set. The remaining points are assigned a class label using Support Vector Machine (SVM) classifier.

## Selecting Final Solution



$s1 = \{1 \ 1 \ 1 \ 2 \ 2 \ 3 \ 3 \ 3 \ 3 \ 4 \ 4\}$   
 $s2 = \{1 \ 2 \ 1 \ 2 \ 3 \ 3 \ 4 \ 3 \ 2 \ 4 \ 4\}$   
 $s3 = \{1 \ 1 \ 2 \ 1 \ 2 \ 3 \ 3 \ 4 \ 2 \ 4 \ 4\}$   
 $s4 = \{1 \ 1 \ 2 \ 2 \ 2 \ 3 \ 3 \ 3 \ 4 \ 4 \ 4\}$   
 $s5 = \{1 \ 1 \ 1 \ 1 \ 2 \ 2 \ 3 \ 3 \ 1 \ 4 \ 4\}$   
 $s6 = \{1 \ 1 \ 4 \ 3 \ 2 \ 2 \ 3 \ 4 \ 4 \ 4 \ 4\}$   
 $s7 = \{1 \ 1 \ 3 \ 3 \ 2 \ 3 \ 3 \ 3 \ 3 \ 4 \ 4\}$   
 $s8 = \{1 \ 2 \ 3 \ 1 \ 2 \ 3 \ 4 \ 3 \ 3 \ 3 \ 4\}$

- Applying 50% voting rule, the consensus clustering label vector becomes  $s = \{1 \ 1 \ ? \ ? \ 2 \ 3 \ 3 \ 3 \ ? \ 4 \ 4\}$ .
- Points 1, 2, 5, 6, 7, 8, 10 and 11 are taken as training points for a Support Vector Machine (SVM) classifier.
- Points 3, 4 and 9 are classified using the trained SVM classifier.

# Application to Microarray Gene Expression Data

**Microarray data can be viewed as an  $n \times m$  matrix:**

- Each of the  $n$  rows represents a **gene** (or a clone, ORF, etc.).
- Each of the  $m$  columns represents an **experimental condition** (a sample, a time point, etc.).
- Each element  $e_{ij}$  represents the expression level of the  $i$ th gene under the  $j$ th condition. It can either be an absolute value (e.g. Affymetrix GeneChip) or a relative expression ratio (e.g. cDNA microarrays).
- A row/column is sometimes referred to as the **expression profile** of the gene/condition.

# Microarray Matrix

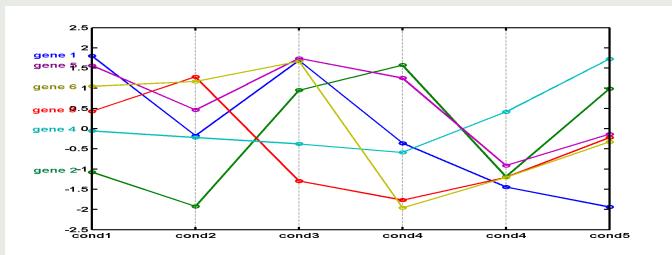
A microarray matrix with 6 genes and 6 conditions.

GENE	cond 1	cond 2	cond 3	cond 4	cond 5	cond 6
gene 1	1.801	-0.174	1.687	-0.359	-1.444	-1.939
gene 2	-1.075	-1.926	0.953	1.575	-1.189	0.987
gene 3	0.427	1.286	-1.295	-1.768	-1.205	-0.22
gene 4	-0.056	-0.221	-0.377	-0.589	0.415	1.727
gene 5	1.565	0.462	1.742	1.253	-0.911	-0.136
gene 6	1.048	1.168	1.668	-1.961	-1.205	-0.325

The values are proportional to expression levels  
**green** = low, **red** = high, **black** = no expression

# Microarray Matrix

A microarray matrix with 6 genes and 6 conditions.



Profile plots are graphical representation of the microarray matrix

## Data Sets for Experiments

Data Sets	Original Number of Genes	Number of Genes after Preprocessing	Number of Time points
Yeast Sporulation	6118	474	7
Yeast Cell Cycle	6000	384	17
Arabidopsis Thaliana	138	138	8
Human Fibroblasts Serum	8613	517	13
Rat Central Nervous System	112	112	9

# Comparison of Different Kernel Functions in MOGA-SVM

**Performance Metric: Silhouette Index** - ranges between -1 and 1, larger value indicates better clustering.

Algorithm	Spor	Cell	Thaliana	Serum	Rat
	$K = 6$	$K = 5$	$K = 4$	$K = 6$	$K = 6$
MOGA-SVM (linear)	0.5852	0.4398	0.4092	0.4017	0.4966
MOGA-SVM (polynomial)	0.5877	0.4127	0.4202	0.4112	0.5082
MOGA-SVM (sigmoidal)	0.5982	0.4402	0.4122	0.4112	0.5106
MOGA-SVM (RBF)	0.6283	0.4426	0.4312	0.4154	0.5127
MOGA (without SVM)	0.5794	0.4392	0.4011	0.3947	0.4872

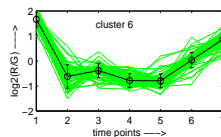
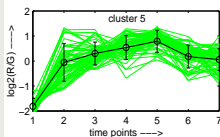
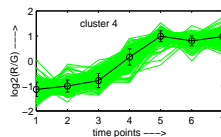
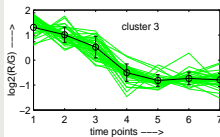
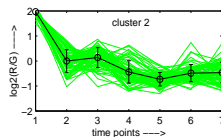
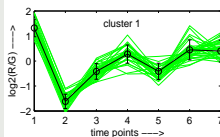
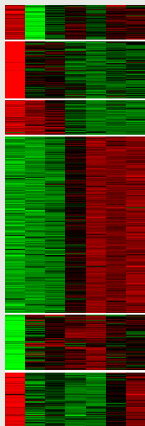
## Comparison among Different Algorithms

**Performance Metric: Silhouette Index** - ranges between -1 and 1, larger value indicates better clustering.

Algorithm	Sporulation		Cell cycle		Thaliana		Serum		Rat CNS	
	$K$	$s(C)$	$K$	$s(C)$	$K$	$s(C)$	$K$	$s(C)$	$K$	$s(C)$
MOGA-SVM	6	<b>0.6283</b>	5	<b>0.4426</b>	4	<b>0.4312</b>	6	<b>0.4154</b>	6	<b>0.5127</b>
MOGA	6	0.5794	5	0.4392	4	0.4011	6	0.3947	6	0.4872
MOGA <sub>crisp</sub> -SVM	6	0.5971	5	0.4271	4	0.4187	6	0.3908	6	0.4917
FCM	7	0.4755	6	0.3872	4	0.3642	8	0.2995	5	0.4050
SGA	6	0.5703	5	0.4221	4	0.3831	6	0.3443	6	0.4486
Average linkage	6	0.5007	4	0.4388	5	0.3151	4	0.3562	6	0.4122
SOM	6	0.5845	6	0.3682	5	0.2133	6	0.3235	5	0.4430
CRC	8	0.5622	5	0.4288	4	0.4109	10	0.3174	4	0.4423

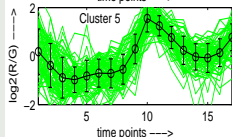
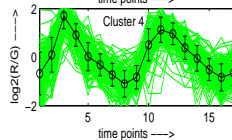
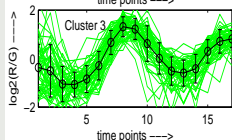
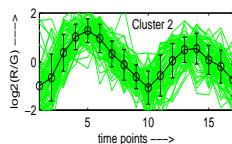
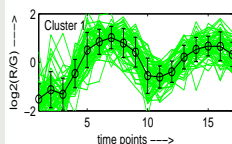
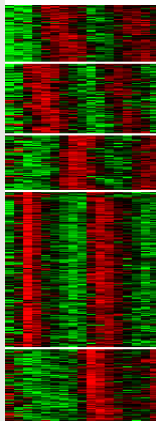
# Heatmap and Cluster Profile Plots

## Yeast Sporulation Data



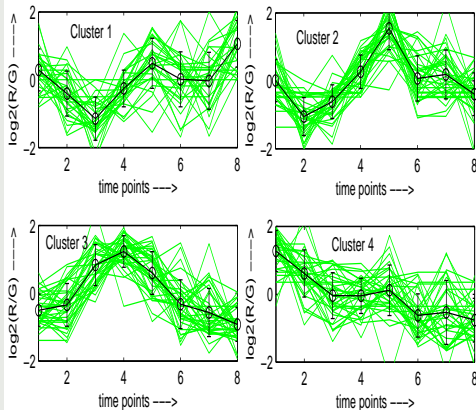
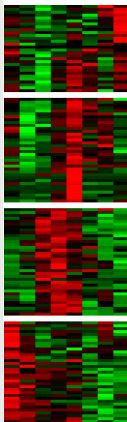
# Heatmap and Cluster Profile Plots

## Yeast Cell Cycle Data



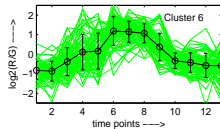
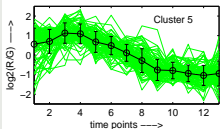
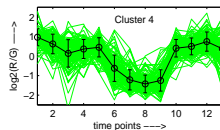
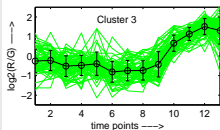
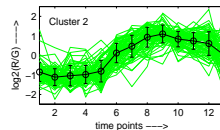
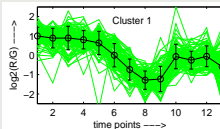
# Heatmap and Cluster Profile Plots

## Arabidopsis Thaliana Data



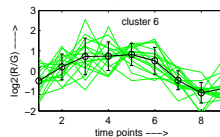
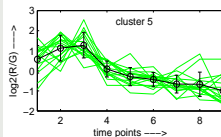
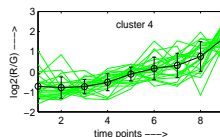
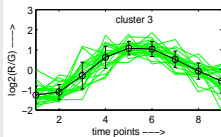
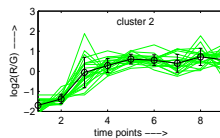
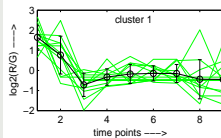
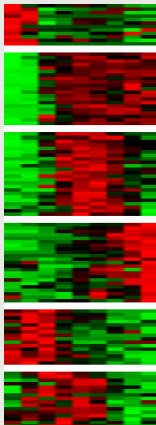
# Heatmap and Cluster Profile Plots

## Human Fibroblasts Serum Data



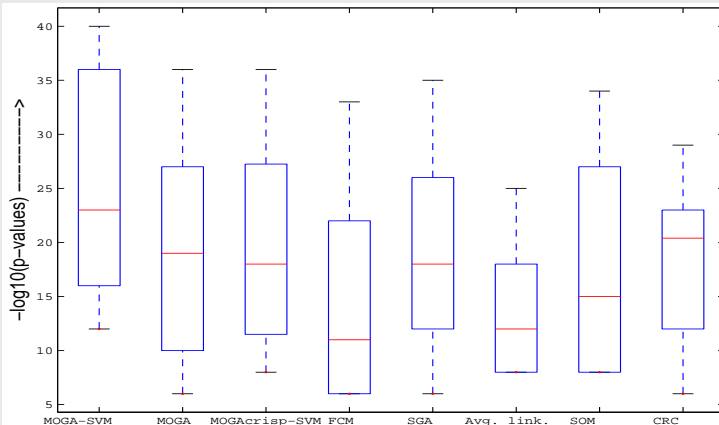
# Heatmap and Cluster Profile Plots

## Rat Central Nervous System Data



# Biological Significance Test

Boxplots for  $p$ -values of Significant GO terms



## Conclusion and Future Scope

- Fuzzy C-means clustering often gets stuck at local optimum – Solution is GA-based clustering.
- Fuzzy C-means clustering optimizes single cluster validity index which may not be equally applicable to different variety of data sets – Solution is multiobjective GA-based clustering.
- NSGA-II based multiobjective fuzzy clustering algorithm is proposed and it is integrated with SVM for improved results.
- Proposed method is applied for clustering genes in microarray gene expression data sets.

## Other Areas of Work in Computational Biology

- Prediction of miRNA targets and TSSs.
- Regulatory network analysis - incorporating miRNAs in the regulatory network.
- miRNA differential expression analysis in Alzheimer's.
- Integrating miRNAs with PPIN.
- Metaheuristic optimization techniques in rational drug design.

# References I

- U. Maulik, S. Bandyopadhyay and A. Mukhopadhyay, Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics, Springer, Heidelberg, Germany, 2011.
- U. Maulik, A. Mukhopadhyay and S. Bandyopadhyay, “Combining Pareto-Optimal Clusters using Supervised Learning for Identifying Co-expressed Genes”, BMC Bioinformatics, Vol. 10, No. 27, 2009.
- A. Mukhopadhyay, S. Bandyopadhyay and U. Maulik, “Multi-class Clustering of Cancer Subtypes through SVM based Ensemble of Pareto-optimal Solutions for Gene Marker Identification”, PLoS One, vol. 5, no.11, art. id. e13803, 2010.
- A. Mukhopadhyay and U. Maulik and S. Bandyopadhyay, “Multiobjective Evolutionary Approach to Fuzzy Clustering of Microarray Data”, Analysis of Biological Data: A Soft Computing Approach, Vol. 3, Chapter 13, pp. 303-326, World Scientific, 2007.

## References II

- S. Bandyopadhyay, A. Mukhopadhyay and U. Maulik, "*An Improved Algorithm for Clustering Gene Expression Data*", Bioinformatics, Vol. 23, No. 21, pp. 2859-2865, 2007.
- U. Maulik and A. Mukhopadhyay, "*Simulated Annealing based Automatic Fuzzy Clustering combined with ANN Classification for Analyzing Microarray Data*", Computers and Operations Research, 2009 (in press).
- U. Maulik, A. Mukhopadhyay and S. Bandyopadhyay, "*Finding Multiple Coherent Biclusters in Microarray Data using Variable String Length Multiobjective Genetic Algorithm*", IEEE Transactions on Information Technology in Biomedicine, 2009.
- A. Mukhopadhyay and U. Maulik, "*Towards Improving Fuzzy Clustering using Support Vector Machine: Application to Gene Expression Data*", Pattern Recognition, Vol. 42, No. 11, pp. 2744-2763, 2009.

## References III

- U. Maulik, A. Mukhopadhyay, S. Bandyopadhyay, M. Q. Zhang and X. Zhang, "*Multiobjective Fuzzy Biclustering in Microarray Data: Method and a New Performance Measure*", In Proc. Int. Conf. WCCI 2008 (CEC 2008), Hong Kong, pp. 1536-1543, June 2008.
- A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "*Multi-objective Genetic Clustering with Ensemble Among Pareto Front Solutions: Application to MRI Brain Image Segmentation*", In Proc. Int. Conf. ICAPR 2009, Kolkata, India, pp. 236-239, February 2009.
- A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "*Refining Genetic Algorithm based Fuzzy Clustering through Supervised Learning for Unsupervised Cancer Classification*", In Proc. EvoBIO 2009, Tubingen, Germany, Lecture Notes in Computer Science, Vol. 5483, pp. 191-202, April 2009.

## References IV

- A. Mukhopadhyay, S. Bandyopadhyay and U. Maulik, "*Analysis of Microarray Data using Multiobjective Variable String Length Genetic Fuzzy Clustering*", IEEE Congress on Evolutionary Computation (CEC 2009), Norway, pp. 1313-1319, May 2009.
- A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "*Unsupervised Cancer Classification through SVM-boosted Multiobjective Fuzzy Clustering with Majority Voting Ensemble*", IEEE Congress on Evolutionary Computation 2009 (CEC 2009), Norway, pp. 255-261, May 2009.

# Thank You

Email - [umaulik@cse.jdvu.ac.in](mailto:umaulik@cse.jdvu.ac.in)