International Journal Of Public Mental Health And Neurosciences
ISSN No: 2394-4668
(An Official publication of Sarvasumana Association)

A Multi-Objective Approach for Revitalizing Gene Expression Data

(Resolving Gene Expression data)

Sushmita Chakraborty
Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, India
29sushmita@gmail.com

Toran Verma
Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, India
toran.verma@rungta.ac.in

***Abstract*—** To locate an array of genes, protein structure and identify the diseases on the basic of gene expression is possible due to the discovery of microarray technology. This kind of analysis is prepared by Genetic Algorithm. Here, proposing a Multi-objective structure combination (FCM and NSGA-II) to analyze the data sets of gene expressions and optimize them by using multi-objective advance. The evaluation of results illustrates that as the quantity of three different data sets of genes expression are clustered, furthermore optimized simultaneously with four objective functions. Fuzzy c-means algorithm is directly applied on the data sets by which the highest membership facts are consider for further processing of clustered datas.

Index Terms: Objective Functions, FCM, NSGA-II, Semi-supervised clustering, Coefficient Entropy & elitism

INTRODUCTION

The objective behind the data mining comes to gather un-identified knowledge moreover new patterns from the data. The methodologies involved in data mining act are analytical study of statistic, structural study and in big data.

With the help of Microaaray Technology for gene expression data, it is getting much easier to solve the huge biological networks that are surrounded with tricky calculations and to find the precise genes very quickly. In 2015, Bandyopadhyay, Saha, Maulik and Deb compared the gene expression data by Semi-FeaClusMOO and AMOSA suggested that absolute Pareto most favorable is taken as heavy aggregation of non-dominant result. Multi-objective optimization [1] used in many areas for optimizing data simultaneously with objective functions, as these methods build up resourceful clustering techniques. The Non-Dominated Sorting GA-II (NSGA-II) is mainly comes under the perception of Multiobjective fuzzy clustering algorithm which used in projected Multiobjective semi-supervised array and the element selection technique is lying on called. In general, the datasets are the collection of records. Each gene expression [3] is noted in different time order. In this report, we suggest a Fuzzy C-Mean clustering technique which gathers gene expression data of three different arrangements along with we introduce four objective functions: ARI-index, XB-index [2], Partioning Clustering index and Coefficient entropy index. To verified the cluster validity indices mentioned above objective functions are adapted. Since, above objective functions are Multi-objective optimization technique is well-known as Non-Dominant Sorting Genetic Algorithm-II is simultaneously optimized the objective functions after the method of FCM clustering. Our proposed mechanism effectively optimizes the action performed by the data sets based on FCM and NSGA-II algorithms.

II. PROBLEM DEFINITION

A. Problem of needing to recover data

A most important difficulty nowadays is that it is problematic to obtain appropriate in sequence regarding genes from experiments that have earlier been well thought-out. Xperrsomics [4] is solving this difficulty by unlocking this unseen in sequence from position to side manual annotation of experiments and complete disparity expression inspection.

B. Traditional model

Compared with the long-established way towards genomic analysis, which has limited inspection and assembly information about a single gene. Microarray technologies presently focus on expressing the levels for huge amount of genes.

C. Difficulty towards data normalization

Appropriate composite actions of microarray experiments, gene expression information repeatedly includes a tremendous sum of interfaces. Hence for this clustering algorithms must.

II. MATHEMATICAL MODELING OF PROBLEM

FCM that generates output data, each of which is grouped via their sorted objective functions and optimize simulteouly by NSGA-II.

$$C = \sum_{C}^{N} \sum_{i}^{I} x_{i,c}^{y} D^2(U_i, V_C), 1 \le y \le \infty \qquad \text{(i)}$$

Where, C= Compactness of fuzzy determined by using D = sum number of genes, depend on the volume of gene expression data measured at 474,384 and 138 respectively, for data, I= Total number of clusters, X= fuzzy membership and why = fuzzy component, VC = represents the cat gene and a is the center of C^{th} cluster, D2 (U_i, Vc) = is the distance involving V_C and U_i., plot the two cluster centers found by the cm function. Fuzzy membership value to calculate the by using eqn (ii):

International Journal Of Public Mental Health And Neurosciences
ISSN No: 2394-4668
(An Official publication of Sarvasumana Association)

$$x_{i,c} = \frac{\left(\frac{1}{D(U_i,V_C)}\right)^{\frac{1}{y-1}}}{\sum_{c-1}^{i}\left(\left(\frac{1}{U_j,V_C}\right)^{\frac{1}{y-1}}\right)} 1 \le c \le C, 1 \le i \le N \qquad \text{(ii)}$$

$$U_c = \frac{\sum_{i=1}^{N} D_{c,i}^2 V_i}{\sum_{i=1}^{N} D_{c,i}^2}, 1 \le c \le C \qquad \text{(iii)}$$

IV. OBJECTIVE FUNCTIONS

i. Xie-Beni Objective Function Index

It is applied to limit the separability in addition to compactness of clusters. From the below eqn (iv) shows the ratio of cluster separation and compactness of the cluster:

$$XB = \frac{\sum_{i=1}^{kmax}\sum_{j=1}^{n} i\, \mu_{j-1}^2 ||x_j - c_i||^2}{n(\min_{i \ne k} ||c_i - c_k||^2)} \qquad \text{(iv)}$$

ii. Adjusted Rand Index

It lies between [-1, +1] and computed as follows:

$$ARI = \frac{\binom{n}{2}\sum_{r=1}^{R}\sum_{c=1}^{c}\binom{t_{rc}}{2} - \left[\sum_{r=1}^{R}\sum_{c=1}^{C}\binom{t_{rc}}{2}\right]}{\binom{n}{2}\left[\sum_{r=1}^{R}\binom{t_{rc}}{2} + \sum_{c=1}^{c}\binom{t_{rc}}{2}\right] - \left[\sum_{r=1}^{R}\binom{t_{rc}}{2} + \sum_{c=1}^{c}\binom{t_{rc}}{2}\right]}$$

(v)

iii. Classification entropy

It is an index of confidence having high membership values in any of the classes. Classification entropy values close to 1 point out the membership between several classes, whereas values nearer to zero indicate more inclusive classification into a single class.

V. PROPOSED METHODOLGY

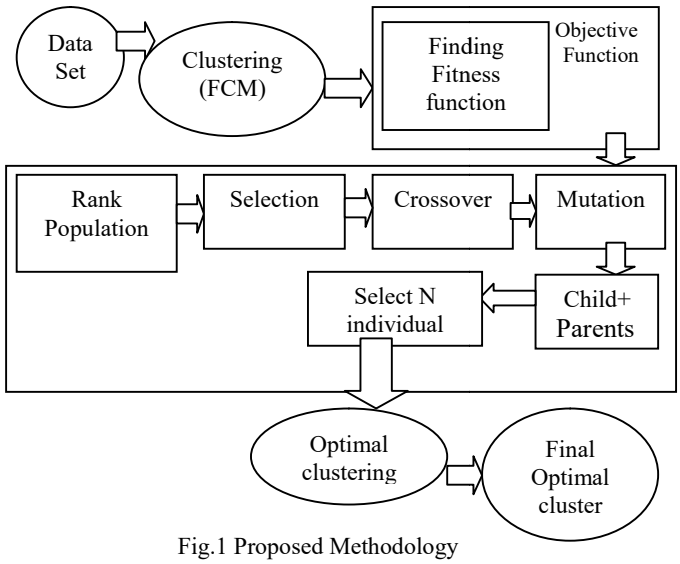

Fig.1 Proposed Methodology

STEP I. DATA SETS (GENE EXPRESSION DATA)

The data set which used in this project is Yeast Sporulation, Yeast Cell Cycle, and Arabidopsis Thaliana. Data logs are transformed and easily downloaded from the above web site. During the formation of genes, some rejected genes [6] are sufficiently ignored.

STEP 2. FUZZY C-MEAN CLUSTERING ALGORITHM

Gene collection and clustering [7] is a salutary model of meaningful mining strategies which is asked for analytic thinking for any variety of information related to excavation. It aids to group the k number of items into C1, C2.... CK on the starting point of like and dissimilar forms. Cluster validity is the operation of squaring off the known cluster.

STEP 3. OBJECTIVE FUNCTION

Refer section (ii) where ARI and XB index objective functions equations are discussed. We took these project four parameters, which is a fusion of minimizing and maximizing various parameters. Classification entropy [8] is a key of confidence having elevated membership values in any of the classes. Classification entropy values close to 1 point out the membership between several divisions, whereas values nearer to zero indicate more inclusive classification interested in a single class.

TABLE 1 PROPERTY OF CLUSTER VALIDITY INDICES

Sl. No	Parameter in CVI	Principle	Definition
1.	Compactness	Intra-Cluster Distance	The summation of the distances of the objects within the same cluster is reduced.
2.	Separability	Intra-Cluster Distance	The distance between any two clusters is Maximized.
3.	Exclusiveness	Probability Density	All data values tend to cluster towards the mean value.

V. WORKING PRINCIPLE BEHIND VALIDATION

XB and Jm (FCM) are considered here precisely to validate the working principle of gene expression data. We have considered three data sets.The Multiobjective optimization [10] method is employed to optimize the prearranged gene expression data, though the number of applications is there like Pattern Recognition, Document Classification, and Information Retrievals and in medical and bio-medical domain has a large application for practicing this technique.

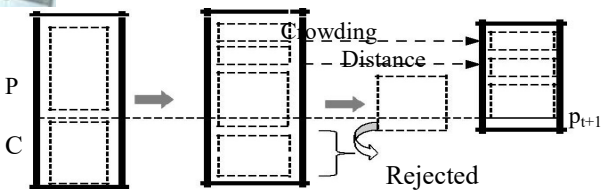

Fig.2 NSGA-II ELITSM PROCESS

To picture out a string like arrangement of centers is used to form a cluster [11]. A position represents an individual cluster, which contains quite a lot of sub-clusters. Now, the four objectives function viewing the diverse value of cluster and compute concurrently optimized by using the NSGA-2 genetic approach. The first two objective functions are referred as unsupervised information and the last as supervised information. To synthesized fuzzy membership values as, j = 1, 2, l, m= 1, 2. . ., n is tallied. Basically XB index is correlated with the entire *variation* σ and least separation *sep* of the clusters. Consider σ and *sep* as:

$$\sigma(U,X;X) = \sum_{i=1}^{K}\sum_{k=1}^{n} u_{i,k}^{2} D^{2}(v_i, x_k) \quad \text{(vi)}$$

And

$$XB(U,V;X) = \frac{\sigma(U,V;X)}{n*sep(v)} \quad \text{(vii)}$$

As mentioned Jm objective functions which generates the overall (global) fuzzy variance shown in below equation. [viii]:

$$jm = \sum_{j=1}^{k}\sum_{k=1}^{n} u_{k,j}^{m} D^{2}(v_k, x_j) \quad \text{(viii)}$$

The main reason for using Jm objective function in this projected methodology is to attain compact clusters. Naturally XB and Jm indices are opposite to each other. To express the compactness of clusters Jm index is used, while XB index considers compactness and separation of the clusters. The two known genetic operators [14], crossover and mutation concluded under crowded selection. While processing NSGA-II algorithm, the parents and the child population are deployed in the direction to achieve the finest outcome of the process known as elitism [12]. In executing NSGA-II algorithm, with support of limited amount of formation.

VI. RESULTS

As discussed above, the three data sets are easility available on website and all datas are in log transformed. Hence there is no need for pre-processing of these data sets (Arabidopsis Thaliana, Yeast Cell Cycle, and Yeast Sporulation [15]). The saturation points are evaluated as well as plotted between objective function and the number of iterations. Results obtained by NSGA-II [12] MOO are also computed in form of clustering. The system program invokes by passing the following parameters.

Input Parameters: number of iteration =100, population amount=8, crossover prospect=0. 8 and mutation prospect=0. 01. Both α and β are set to 0.5. Analysis shows β (majority voting threshold) alter the functions of clustering technique called MOGA. The algorithm has been executed for a range of β values starting from 0.1 to 0.9 with a step size of 0.05 for all the data sets. In Fig 5 shows the graph between Jm and XB index and number of iterations.

The input parameters for achieving this graph of Yeast Sporulation data set are in = 100 (number of iterations), Population = 50, crossover probability=1. 34, α=β=0. 05, fuzzy membership m= 0.01 and the number of clusters = 12. The graph the minimum function of XB and Jm [13] saturate 85 iterations at points 0.5. Jm is to be minimized to acquire compact clusters. XB and Jm indices are to an extent contradictory in nature. XB index is responsible for both compactness and separation for the clusters, whereas Jm only represents the global compactness of the clusters accurately.

TABLE 2 DEPICTS THRESHOLD PARAMETER FUNCTIONALITY

Two Threshold Parameter	Dependability		
α (membership threshold)	Sizes of the training and testing sets	Confidence	It is the highest membership degree higher.
Increased	Reduced	Increased	
Decreased	Increased	Decreased	
β (fuzzy voting)			
Increased	Decreased	Determines the minimum number of non-dominated solutions that agree with each other in the fuzzy voting Context.	
Decreased	Increased		

TABLE 3 SELECTED BEST FITNESS VALUES OF TWO OBJECTIVE FUNCTIONS (JM & XB) FOR YEAST SPORULATION

Population size = 8	Jm (FCM)	XB
1	360.753	0.312
2	360.753	0.312
3	361.992	0.304
4	361.414	0.307
5	361.868	0.304
6	361.180	0.309
7	360.924	0.311
8	360.829	0.312

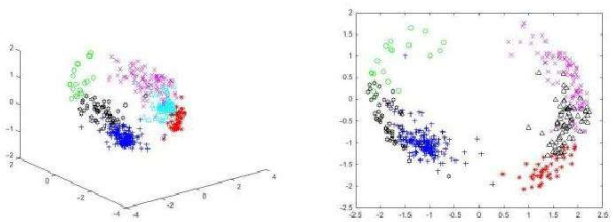

Fig.3 showing clustering (a) after FCM & (b) The sorted clustered of two objective functions for Yeast Sporulation.

TABLE 4 SELECTED BEST FITNESS VALUES OF FOUR OBJECTIVE FUNCTIONS FOR YEAST SPORULATION

POPULATION SIZE 8	JM (FCM)	XB (XIE-BENI)	PC (PARTIONING COEFFICIEN)	CE (CLASSIFICATION ENTROPY)
1	361.18	0.317	0.913	0.253
2	653.91	115.9	0.712	0.869
3	361.8	0.308	0.913	0.25
4	470.93	0.586	0.93	0.195
5	1190.6	31.91	0.616	1.136
6	1066.0	27.50	0.685	0.928
7	854.45	19.36	0.66	0.998
8	821.51	67.89	0.744	0.743

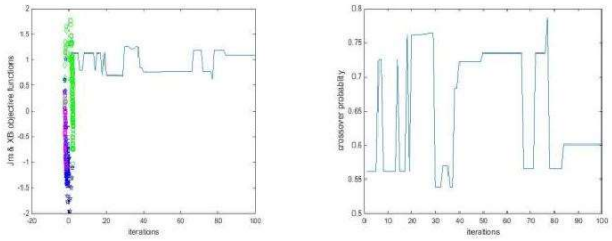

Fig. 4 the (a) Objective functions and (b) crossover probability v/s Number of iteration for Yeast Sporultions data set.

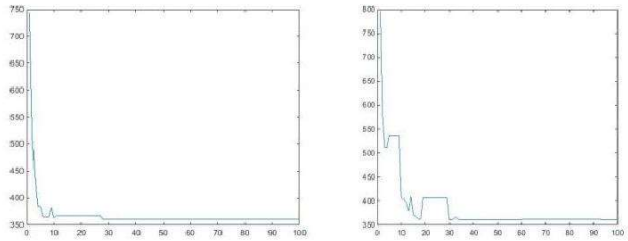

Fig. 6 the graph (a), (b) showing the saturation level after some iteration. The total number of iteration i= 100, in (a) after 29 iteration it gets saturates optimize by only two objective functions whereas (b) it saturates after 90 iteration, optimize by four objective functions. With this observation, in all the experiments hereafter, β value has been kept constant at 0.5.

Outcome for Yeast Cell data set - The values of the different parameter number of generations=100, population size=8, crossover probability=0.8 and mutation probability=0.01. Both α and β are set to 0.5. The fuzzy exponent values of m for the data sets Cell cycle is 1.14. The algorithm has been executed for different number of clusters and the solution giving the best fitness score is considered.

TABLE 6 SELECTED BEST FITNESS VALUES OF TWO OBJECTIVE FUNCTIONS (JM &XB) FOR YEAST SPORULATION

Population size = 6	*Jm*	*XB*
1	*1974.563*	*3.17*
2	*2056.395*	*1.06*
3	*2009.780*	*2.09*
4	*2027.086*	*1.47*
5	*2039.936*	*1.17*
6	*1990.398*	*2.16*

TABLE 5 SELECTED BEST FITNESS VALUES OF FOUR OBJECTIVE FUNCTIONS FOR YEAST CELL

Population	*Jm*	*XB*	*CE*	*PC*
1	*2150.32*	*14785591880972.*	*0.44*	*0.93*
2	*2230.21*	*9366066944580.5*	*0.42*	*0.94*
3	*2206.51*	*4606094586965.6*	*0.42*	*0.94*
3	*2111.37*	*1352359876019.9*	*0.43*	*0.93*
4	*2094.14*	*2815991803246.5*	*0.43*	*0.93*
5	*2211.13*	*3806197357534.3*	*0.42*	*0.94*

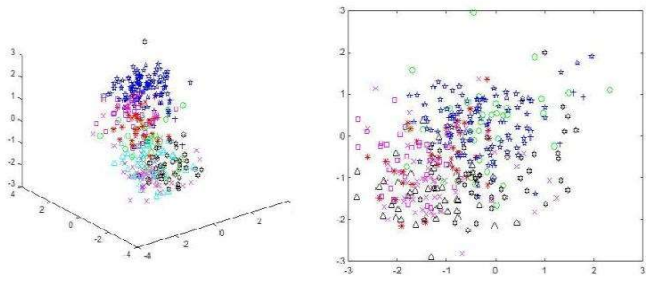

Fig. 5 showing clustering after optimize two objective functions by NSGA-2 (a) after FCM & (b) The final non-dominated Pareto-optimal front obtained by MOGA clustering for Yeast Cell data set.

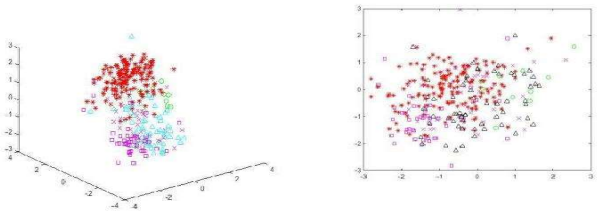

Fig. 6 showing clustering after optimize four objective functions by NSGA-2 (a) after FCM & (b) The final non-dominated Pareto-optimal front obtained by MOGA clustering for Yeast Cell data set.

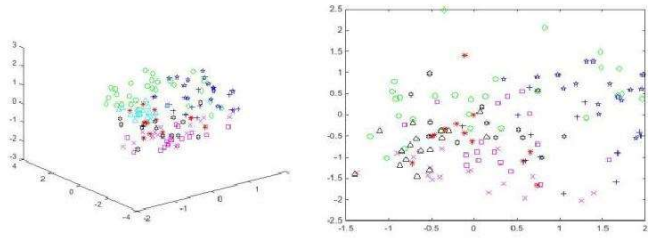

Fig. 7 showing clustering after optimize four objective functions by NSGA-2 (a) after FCM & (b) The final non-dominated Pareto-optimal front obtained by MOGA clustering for Arabidopsis Thaliana data set.

VII. Conclusion and scope of further work

Proposing a new Multiobjective optimization technique for obtaining a optimum outcome via NSGA-II algorithm which forms the sorted clustered genes. The adv of using NSGA-II algorithm is to optimize more than one objective function parallel, here, Xie-Beni (XB) ,Jm, ARI, Partioning Cassification,Coefficint Entropy are the basic objective functions used in this paper to advance towards real world scenario. The results show the saturation points of different objective functions also showing the best fitness value. It main use in satellite image processing and medical diagnosis. In future work we suggest instead of cluster technique some advance classification and fuzzy to be incorporated and enhanced.

REFERENCES

[1] Deb, A. Pratap, S. Agrawal, and T. Meyarivan, A fast and elitist multiobjective genetic algorithm:NSGA-II, IEEE Trans. Evol. Comput., vol. 6, pp. 182197, 2002.

[2] X. Xie and G. Beni., A validity measure for fuzzy clustering. IEEE Trans. on P.A.M.I., vol. 13, no.4, pp. 841846, 1991.

[3] R. Sharan, Click and expander: a system for clustering and visualizing gene expression data. Bioinformatics, vol. 19, p. 17871799, 2003.

[4] A. Ben-Dor and et al, Clustering gene expression patterns. J. Comput. Biol., vol. 6, p. 281297, 1999.

[5] A. Jain, M. N. Murty, and P. Flyn, Data clustering: A review. ACM Computing Surveys, vol. 31, no. 3, pp. 264323, 1999.

[6] J. Handle, J. Knowles, On semi-supervised clustering via Multiobjective optimization, in: Genetic and Evolutionary Computation Conference, pp. 1465-1472, 2006.

[7] C. Zhang, X. Lu, and X. Zhang, Significance of gene ranking for classification of microarray samples, IEEE/ACM TRANS. ON COM. BIO. AND BIOINF. vol. 3, no. 3, pp. 312320, JULYSEPTEMBER 2006.

[8] D. Davies and D. Bouldin., A cluster separation measure. IEEE Trans. On P.A.M.I., vol. 1, no. 2, pp. 224227, 1979.

[9] S. Saha, A. Ekbal, A. K. Alok, Semi-supervised clustering using multiobjective optimization. In Hybrid Intelligent Systems (HIS), IEEE 12th International Conference, pp. 360-365, December 2012 .

[10] S. Bandyopadhyay, S. Saha, U. Maulik and K. Deb, A simulated annealing based multi-objective optimization algorithm: AMOSA, IEEE Trans. Evol. Comput, 12, pp. 269, 2008.

[11] K. Deb, A. Pratap, S. Agrawal, and T. Meyarivan, A fast and elitist multiobjective genetic algorithm:NSGA-II, IEEE Trans. Evol. Comput., vol. 6, pp. 182197, 2002.

[12] X. Xie and G. Beni., A validity measure for fuzzy clustering. IEEE Trans. on P.A.M.I., vol. 13, no.4, pp. 841846, 1991.

[13] R. Sharan, Click and expander: a system for clustering and visualizing gene expression data. Bioinformatics, vol. 19, p. 17871799, 2003.

[14] S. Chu, M. Eisen, J. Mulholland, D. Boston, P. O. Brown, The transcriptional program of Sporulation in budding Yeast, Science, 28 (2), pp. 699-705, 1998.