A Hybrid Nelder-Mead Method For Biclustering Of Gene Expression Data

Kavitha M, Dr.N.Arulanand

PG Student, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India
Associate Professor, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, India
Email: kavithacse2@gmail.com, arulnat@gmail.com

ABSTRACT: Biclustering algorithms are used to identify local patterns from gene expression data sets and used to extract biologically relevant information. The fundamental goal of this work is to derive the heuristic approaches to identify the coherent biclusters from gene expression data with minimum MSR (Mean Square Residue) and maximum row variance. Nelder Mead (NM) simplex method is a local search method and very sensitive to the choice of initial points and does not guaranteed to attain the global optimum. The simplex obtained from each iteration continues to shrink and fall into local minima solution. To deal with this problem hybrid optimization approaches namely, Nelder Mead with Levy Flight and Tabu Search with Nelder Mead are proposed and compared. From the analysis, the result shows that NM with levy Flight method performs better to obtain global optima solution when compared and analyzed with NM method and Tabu search with NM.

Keywords: gene expression data, biclustering, heuristic optimization, Nelder Mead, Mean Square Residue, fitness function

1. INTRODUCTION

The cell is the basic functional unit of all living organisms. A central core in the cell called nucleus and inside the nucleus there is an important molecule known as DNA. The DNA strand of living organisms is an enormous sequence of four nucleotides: adenine (A), cytosine(C), guanine (G) and thymine(T). The whole genome can be divided into genes subsequences which encode proteins or regulate cell functioning. The structure of the human genome consists of 40000 genes. A gene is a segment of DNA, which contains the formula for the chemical composition of one particular protein. Gene expression is the process of transcribing a gene's DNA sequence into messenger Ribonucleic Acid (mRNA) sequences, which in turn are later translated into proteins. The fundamental goal of microarray gene expression data analysis is to identify the behavioral patterns of genes. A typical DNA microarray analysis follows a multistep procedure: fabrication of microarrays by fixing properly designed oligonucleotides representing specific genes; hybridization of cDNA populations onto the microarray; scanning hybridization signals and image analysis; transformation and normalization of data; and analyzing data to identify differentially expressed genes as well as sets of genes that are co regulated.The gene expression matrix is the processed data obtained after normalization. Each row in the matrix corresponds to a particular gene and each column corresponds to a particular condition. The expression level for a gene across different experimental condition is cumulatively called the gene expression profile, and the expression level of each gene under an experimental condition is cumulatively called the sample expression profile. An expression profile of a gene or an experimental condition is thought of as a vector and can be expressed in vector space. For example, an expression profile of a gene known as a vector in n dimensional space where n is the number of conditions whereas an expression profile of a condition with m genes can be denoted as a vector in m dimensional space where m is the number of genes. Figure 1.1 shows the gene expression matrix A with m genes across n conditions is considered to be an $m \times n$ matrix. Each element a_{ij} of this matrix represents the expression level of a gene i under a specific condition j, and is represented by a real number.

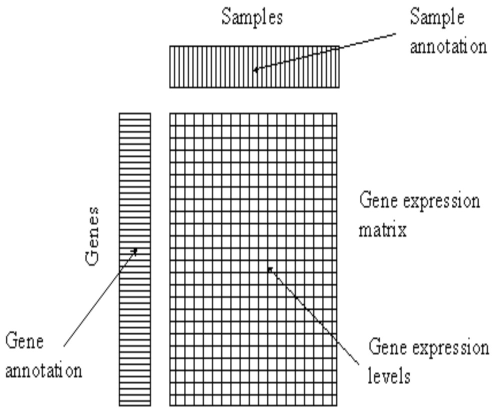

Fig. 1 *Gene expression matrix*

Data mining techniques are essential in order to identify various patterns from gene expression data. Clustering is an important approach for mining significant biological models in gene expression data analysis. Clustering approach used to identify the existing relationships among a set of variables such as biological conditions. The goal is to extract information on how gene expression levels vary among the different conditions, including groups of co-expressed genes. If two different genes exhibit similar expression profile across the conditions, this indicates a common pattern of regulation and possibly reflecting some kind of interaction or relationship between their functions. Clustering of gene expression data can be done in three ways, namely Gene-based clustering, Condition-based clustering and Biclustering.Several clustering methods have been proposed for expression analysis. However, conventional clustering approaches such as hierarchical, self organizing and k-means clustering group the genes over all the conditions, whereas cellular processes are active only under a subset of conditions. Also, a single gene may belong to more than one group, as a gene may be involved in more than one biological process. Therefore, biclustering algorithms have been preferred to standard clustering techniques to identify local patterns from gene expression data sets. Biclustering is a data mining technique which clustersthe rows and columns of a matrix simultaneously.Finding the biclusters in a large microarray gene expression data is a much more complex problem than clustering. The search space for the biclustering problem is 2^{m+n} where m represents the number of genes and n represents the number of conditions. Usually m+n is more than 2000. In fact, it has been proven to be a NP-hard problem and hence heuristic search algorithms are used. It solves the problem quickly and finds the approximate solution when classic methods fail to found. It finds the best solution in the current set of conditions.The rest of the paper is organized as follows: Section II describes the literature survey on various biclustering methods. Section III gives the overview about Nelder Mead method and presents the proposed Nelder Mead with Levy Flight method. The experimental results are analyzed in section IV. Section V presents conclusion.

2. RELATED WORKS

Nelder and Mead proposed a simplex search method and it is a commonly used nonlinear optimization technique, which is a well-defined numerical method. This technique is applied for the problems for which derivatives may not be known. Nelder-Mead method maintain at each step a simplex which is a geometric figure in n dimensions of nonzero volume that is having n+1 vertices. Each iteration of a simplex-based search begins with the n+ 1 vertex and the associated function values. The procedure takes an initial simplex as argument and returns the best solution in another final simplex. The following procedures are used to rescale the simplex based on the local behavior of the function: reflection, expansion, contraction and shrinkage. By following these procedures, the simplex improves itself at each iteration and get closer to the optimum. The Nelder–Mead simplex algorithm is a classical local descent algorithm and it depends on the choice of initial points and does not guaranteed to attain the global optimum.*Hossein Rahami et al.* [8]explained that the genetic algorithm (GA) is a search heuristic that mimic the genetic processes of biological organisms. This heuristic approach is used to explore useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), and uses stochastic direct optimization methods. GA uses random operators, and therefore may result in different solutions each time they are run. GA is used to create a powerful tool that can be used to solve optimization problems like constrained and unconstrained problems, single- and multi-variable, linear and nonlinear problems.In order to make GA a more reliable algorithm and in relation with optimization efficiency many approaches have been suggested. One improvement can be doneis modification of genetic operators and the other is hybridizing of GA with other optimization methods.*Kennedy and Eberhart et al.* [5]proposed Particle Swarm Optimization (PSO).In PSO a number of simple particles are flown in the search space and each evaluates the objective function for its current location. Each particle then decides its next movement through the search space by comparing the history of its own current and best (best fitness value) locations with those of one or more particles of the swarm. The next iteration continuous after all particles has been moved. Finally, the swarm as a whole, like a flock of birds conjointly searching for food, is likely to move close to an optimum solution of the fitness function.If a particle’s current location coincides with the global best position particle, the particle can move away from this pointonly if its previous velocity and inertia (w) are non-zero. If their previous velocities are close to zero, then all particles will stop moving once they reach the global best particle, which may lead to premature convergence of the algorithm. PSO categorized under the global search procedures but requires much computational effort.To improve the Particle swarm optimizers *Shu-Kai S. Fan, Erwie Zahara et al* [4] proposed hybrid simplex search and particle swarm optimization for unconstrained optimization. To deal with the slow convergence of PSO, a hybrid approach called PSO with a local simplex search technique is addressed.Shuffled Frog Leaping (SFL) is swarm intelligence based heuristic optimization algorithm proposed by *Eusuff and Lansey* [7] to solve discrete combinatorial optimization problem. The search begins with a randomly selected population of frogs. Within each memeplex, the frogs are affected by other frog’s ideas. So, they experience a memetic evolution. Memetic evolution assesses the quality of the meme of an individual and improves the individual frog’s performance towards a goal.If the global best particle is trapped into a local minima solution, then all the other particles will also fly toward the local optimum. In this situation, the velocity update for the global best particle is done by merely a small jump for further improvement such that those particles can avoid trapping into the local optimum. To overcome this issue, a mutation heuristic is added to the global best particle.As mentioned in [10] and [11], Tabu Search (TS) allows covering widely the solution space and itstimulates the search towards the solutions far from the current solution. Itavoids the risk of trapping into a local minimum. Globally optimize a function f in a given search domain identifies global minima without being trapped into the local minima solution. To localize a promising area, likely to contain a global minimum, it is essential to explore the whole domain. When a promising area is detected, the next step is applying exploitation procedure and obtains the local optimum solution. This is performed by means of only one method. *Erwie Zahara and Yi-Tung Kaob* [12]proposed hybrid Nelder–Mead and particle swarm optimization for constrained engineering design problems. Constraint handling methods includes gradient repair method and constraint fitness priority-based ranking method, whichare used in NM–PSO as a special operator to deal with satisfying constraints.*Chellamuthu Gunavathi and Kandasamy Premalatha et al* [14] proposed shuffled frog method with levy flight in [2]. In Shuffled Frog Leaping, Levy flight is applied to avoid early and quick convergence of shuffled frog leaping (SFL) algorithm. Animals need to develop their search strategies to avoid spending more time in such unproductive areas. Levy flight approach having this property. To improve the searching strategy and classification of frogs in Shuffled Frog Leaping, an additional parameter LF is added.

3. HEURISTIC OPTIMIZATION METHODS

3.1 Nelder Mead Method

The Nelder Mead (NM) simplex search method is a local search method designed for unconstrained optimization problems. Each iteration begins with the n+ 1 vertex and the corresponding function values. It maintains a non degenerate simplex at each iteration. The method takes an initial simplex as argument and returns the best solution in another final simplex. NM method uses four basic procedures to rescale the simplex: reflection, expansion, contraction and shrinkage. The NM method returns the solution from the final simplex which has the smallest objective value. Each vertex is represented as candidate solution for the problem. Solutions are encoded by means of binary strings of length N+M, where N and M are the number of genes and number of conditions respectively. The mapping function of solution into a binary string representation of a bicluster is set using,

$$yij = \{ \begin{array}{ll} xij \geq 0.5 & 1 \\ otherwise & 0 \end{array}$$

Where, x_{ij} – Random value generated for j^th gene/condition of i^th point

Y_{ij}- Binary string representation of bicluster x_{ij}

Let gene expression data matrix A has R rows and C columns where a cell a_{ij} is a real value that represent the expression level of gene *i* under under condition *j.*

The residue of an element a_{ij} in a submatrix A_{IJ} (where I represent R and J represents C) equals

$$res_{i,j} = a_{i,j} + a_{I,J} - a_{I,j} - a_{i,J}$$

The quality of a biclusters is evaluated by computing the Mean Square Residue (MSR), i.e. the sum of all the squared residues of its elements is given as

$$H(I,J) = \frac{1}{|I||J|} \sum_{i\in I, j\in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

Where

$$a_{iJ} = \frac{1}{|J|} \sum_{j\in J} a_{ij}, \quad a_{Ij} = \frac{1}{|I|} \sum_{i\in I} a_{ij},$$

$$a_{IJ} = \frac{1}{|I||J|} \sum_{i\in I, j\in J} a_{ij} = \frac{1}{|I|} \sum_{i\in I} a_{iJ} = \frac{1}{|J|} \sum_{j\in J} a_{Ij}$$

Low MSR value denotes strong coherence in the biclusters. The row variance is an accompanying score to include the non-trivial biclusters. The row variance can be represented as follows:

$$V(I,J) = \frac{1}{|J|} \sum_{j\in J} (a_{ij} - a_{Ij})^2$$

The fitness function for obtaining coherent biclusters is defined as follows:

$$f(I,J) = H(I,J) + \frac{1}{V(I,J)}$$

First the vertices x_1, x_2… x_n, x_{n+1} of the simplex are sorted in ascending order depending on the objective function value.

$$f(x_1) \leq f(x_2) \leq .. \leq f(x_n)$$

f (x) denotes fitness value of the vertex x.After every iteration the worst vertex is replaced by new vertex using the following steps:

Reflection (R)

The reflection point x_r is calculated using the following formula

$$x_r = x_c + \alpha(x_c - x_{n+1})$$

x_c represents the centroid of the n best points calculated using,

$$x_c = \frac{1}{n+1} \sum_{i=1}^{n+1} x_i$$

If the value of $f_r = f(x_r)$ satisfies the condition $f_1 \leq f_r \leq f_n$then, we replace x_{n+1} with x_r and start next iteration.

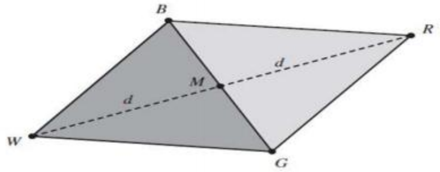

Fig.2 *Reflection*

Expansion (E)

If $f_r < f_i$, calculate expansion vertex x_e

$$x_e = x_c + \beta(x_r - x_c)$$

If $f_e = f(x_e)$ satisfies $f_e < f_r$ replace x_{n+1} with x_e otherwise with $x_{r.}$

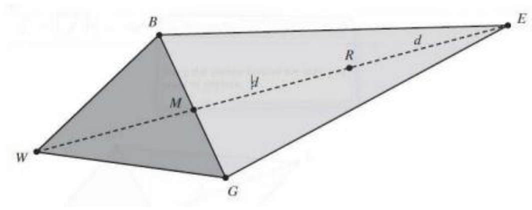

Fig.3*Expansion*

Outside Contraction (OC)

In case $f_n \leq f_r < f_{n+1}$, compute the outside contraction point.

$$x_{oc} = x_c + \gamma(x_r - x_c)$$

If $f_{oc} = f\ (x_e)$ satisfies $f_{oc} \leq f_r$ replace x_{n+1} with x_{oc} otherwise do shrink operation.

Inside Contraction (IC)
In case $f_r \geq f_{n+1,}$ compute the inside contraction point.

$$x_{ic} = x_c - \gamma(x_r - x_c)$$

If $f_{ic} = f\ (x_{ic})$ satisfies $f_{ic} \leq f_{n+1}$ replace x_{n+1} with x_{ic} otherwise do a shrink operation.

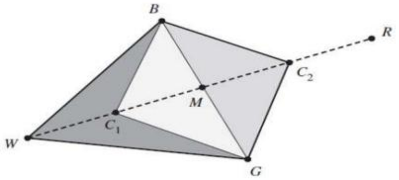

Fig.4*Contraction*

Shrink (S)
Shrink vertex is calculated using,

$$x_i = x_1 + \delta(x_i - x_1)$$

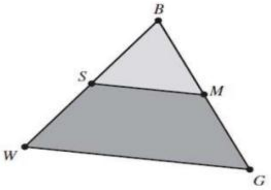

Fig.5*Shrink*

The Nelder–Mead simplex algorithm is a powerful local search algorithm method and it depends on the choice of initial points and does not guaranteed to obtain the global optimum. After certain iterations the simplex continues to shrink. It does not explore the whole search space.

3.2 Nelder Mead with Tabu Search

Tabu search is used to solve various optimization problems. Tabu search [10] allows covering widely the solution space, to stimulate the search towards the solutions far from the current solution.It avoids the risk of trapping into a local minimum. To localize a promising area, thatshould contain a global minimum, it is essential to well explore the whole domain. When a promising area is identified, the appropriate tools are used to exploit this area and obtain the optimum as quickly as possible. Exploration is done by Tabu Search method and exploitation is done by using Nelder-Mead method.Toexplore the whole search area, the algorithm starts from the initial solution which is randomly generated. Then the algorithm generates specified number of neighbors. The objective function is calculated for each accepted neighbor and the neighbor which has the best fitness value becomes the new current solution. This current solution is added to the tabu list of length L. If a new promising area is accepted, stop the exploration process and start the exploitation process. Exploitation is done by NM method. It is used to move towards a minimum.

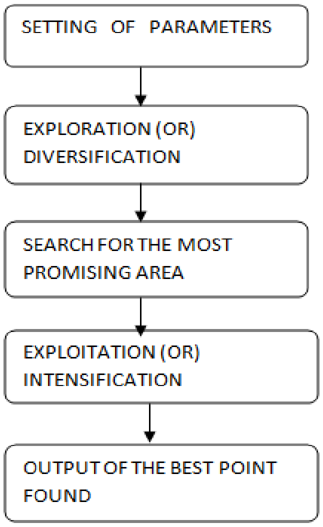
SETTING OF PARAMETERS
EXPLORATION (OR) DIVERSIFICATION
SEARCH FOR THE MOST PROMISING AREA
EXPLOITATION (OR) INTENSIFICATION
OUTPUT OF THE BEST POINT FOUND

Fig.6*NM with Tabu Search*

3.3 Nelder Mead with Levy Flight

Levy Flight (LF) explained by *Barthemy P., Bertolotti J., Wiersma D. S.* [6] is a random walk where the step size has a Levy tailed probability distribution. Levy Flight term is used to refer discrete grid. Levy Flight follows Markov Process. Exponential property of Levy Flight provides it a scale invariant property and used to model data for exhibiting/ showing clusters. NM approach is very sensitive to the choice of initial points. NM method does not explore the whole search space and hence not guaranteed to attain global optimum solution. When the simplex continues to shrink, levy flight method is applied. It adds random step size to the vertex values. So, it helps to explore the search space. Levy flight is carried out by,

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Le'vy\ (\lambda),$$

Where α > 0 is the step size set based on problem of interest. Lambda value taken as 1.5 and EX-OR symbolizes exclusive OR i.e., entry wise multiplication.

ALGORITHM

```
            IF f(R) <f (G), THEN perform
                        Case (i) {either reflection or expansion}
            ELSE perform
                        Case (ii) {either contraction or shrink}
  BEGIN {Case (i)}
IF f (B) <f(R) THEN
Replace W with R
ELSE
                                    Compute E and f(E)
IF f(E) < f(B) THEN
   Replace W with E
            ELSE
   Replace W with R
            ENDIF
  ENDIF
  END {Case(i)}
            BEGIN {Case (ii)}
IF f(R) < f(W) THEN
Replace W with R
```

```
Compute C=(W+M)/2
Or C= (M+R)/2 and f(c)
IF f(C) < f(W) THEN
Replace W with C
ELSE
 IF (Shrinks
Continuously)
                    Apply Levy Flight
                    ELSE
 Compute S and f(S)
Replace W with S
 Replace W with M
      ENDIF
    END {Case (ii)}
```

where f (R) , f (G), f (B), f (E), f (C), f (S), f (W)are the fitness functions of R,G,B,E,C,S and W respectively. The vertices of the simplex at the next iteration consist of x_1, $v_2 \ldots v_{n+1}$.

4. EXPERIMENTAL RESULTS

In this paper experiments are conducted on the dataset Aradabsis Thaliana. It contains 734 genes and 69 conditions. The minimum fitness value obtained for the biclusters with the stopping criterion up to the maximum iteration value as 500. The quality of the biclusters is evaluated by Mean Square Residue (MSR). Low MSR value denotes strong coherence among the genes in the biclusters. The results obtained for the proposed approaches namely, NM with Levy Flight and Tabu Search with NM are compared and analyzed. NM method uses four basic procedures to rescale the simplex: reflection, expansion, contraction or shrink. It is a local search method and does not guarantee to obtain the global optima. After certain iterations the simplex continues to shrink at Nelder-Mead method and hence easily fall into local minima solution. NM method does not explore the whole search space. Whereas, when applying levy flight approach the simplex explores the search area because of its random step size.

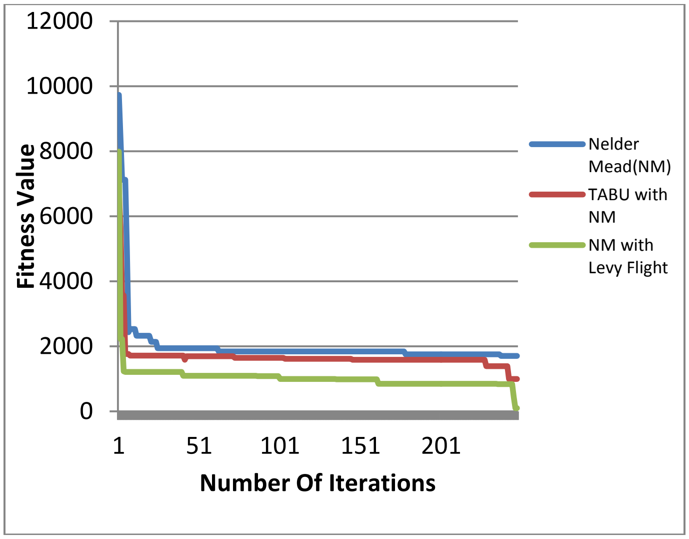

***Fig. 7** Comparison Chart*

***Table 1** Comparison Table*

Optimization Methods	Fitness Value (Iteration 1)	Fitness Value (Iteration 150)	Fitness Value (Iteration 250)	Fitness Value (Iteration 500)
Nelder Mead	9735.39	1840.8	1752.8	1704.6
Tabu Search with Nelder Mead	6769.9	1712.33	1592.33	992.33
Nelder Mead with Levy Flight	7992.56	984.22	838.54	102.04

At Levy flight the next position or function value is evaluated based on two parameters namely, the current position and transition function. The consecutive positions generated by the steps of Levy Flight create a random walk that has a Levy tail distribution. The use of Levy Flight with the NM significantly improves the performance of NM and hence NM with Levy approach explores the search space better than Nelder-Mead method and Tabu Search with NM. Fig.7 shows the comparison analysis chart of the optimization techniques NM, NM with Levy and Tabu search with NM used in this work to discover coherent biclusters in the dataset. X-axis denotes number of iterations and fitness values shown in y-axis. Table 1 shows the fitness values obtained at various iterations of the optimization methods. Fitness value of the NM with Levy Flight method gradually decreases when compared to the other two methods. As shown in the Fig 7, NM with Levy gives the better result when compared to NM and Tabu search with NM because of its random step size. As when compared to the NM method, Tabu with NM performs better but NM with Levy performs better than the later. Tabu search with NM method uses exploration and exploitation steps. The result shows that the fitness value obtained from NM with Levy is minimal.

5. CONCLUSION

The Proposed hybrid Nelder-Mead with Levy Flight method explores the search space better than Nelder-Mead method and it overcomes the poor convergence problem of NM. When the simplex continues to shrink, levy flight method is applied to explore the search space. It helps to find most coherent bicluster. Nelder-Mead with levy flight method returns the solution point from the final simplex which has the smallest objective value. Nelder-Mead and Tabu search with Nelder-Mead methods are implemented for comparative analysis. The result shows that the proposed Nelder-Mead with Levy Flight method outperforms the conventional Nelder-Mead and Tabu Search with Nelder-Mead methods. The proposed method works well for microarray gene expression data to find functionally correlated genes and the behavioral patterns of genes. These patterns have huge significance and application in bioinformatics and clinical search such as drug discovery, treatment planning, accurate diagnosis.

REFERENCES

[1] Y. Cheng, and G.M. Church, "Biclustering of Expression Data," in Proc. of the 8th Conf. Intel. Sys. Mol. Biol., Menlo Park, United States, 2000, pp. 93-103.

[2] R. Balamurugan, A. M. Natarajan, K. Premalatha "Comparative Study on Swarm Intelligence Techniques for Biclustering of Microarray Gene Expression Data" in World Academy of Science, Engineering and Technology International Journal of Computer, Control, Quantum and Information Engineering Vol:8, No:2, 2014

[3] M. Pandi, K. Premalatha "An Advanced Nelder Mead Simplex Method for Clustering of Gene Expression Data" in World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:8, No:4, 2014

[4] Shu-Kai S. Fan, Erwie Zahara "A hybrid modified simplex search and modified particle swarm optimization for unconstrained optimization" in European Journal of Operational Research 181 (2007) 527–548

[5] Riccardo Poli, James Kennedy, Tim Blackwell "Particle swarm optimization-An overview" in Springer Science + Business Media, LLC 2007

[6] Barthemy P., Bertolotti J., Wiersma D. S., "A Levy Flight for light, Nature", 453, 495-498(2008)

[7] M.M. Eusuff and. E. Lansey, "Optimization of water distribution network design using the shuffled frog leaping algorithm, "Journal of Water Resources Planning and Management,vol.129, no.3,pp.210–225,2003

[8] Hossein Rahami, A.Kaveh , Reza NajianAsl "Ahybridmodified Genetic-Nelder MeadSimplexalgorithmforlarge-scale truss optimization" in International Journal Of Optimization In Civil Engineering Int. J. Optim. Civil Eng., 2011; 1:29-46

[9] Sauravjyoti Sarmah and Dhruba K. Bhattacharyya "An Effective Technique for Clustering Incremental Gene Expression data" in IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3, May 2010

[10] Rachid Chelouah, Patrick Siarry "A hybrid method combining continuous Tabu search and Nelder–Mead simplex algorithms for the global optimization of multiminima functions" in European Journal of Operational Research 161 (2005) 636–654

[11] Abdel-Rahman Hedar, Masao Fukushima "Tabu Search directed by direct search methods for nonlinear global optimization" in Elsevier Science 10 June 2004.

[12] Erwie Zahara,Yi-Tung Kaob "Hybrid Nelder–Mead simplex search and particle swarm optimization for constrained engineering design problems" in Expert Systems with Applications 36 (2009) 3880–3886

[13] Shu-Kai S. Fan, Erwie Zahara "A hybrid simplex search and particle swarm optimization for unconstrained optimization" in European Journal of Operational Research 181 (2007) 527–548

[14] Chellamuthu Gunavathi and Kandasamy Premalatha "A Comparative Analysis of Swarm Intelligence Techniques for Feature Selection in Cancer Classification" in Hindawi Publishing Corporation, The Scientific World Journal Volume 2014, Article ID 693831,12 pages

[15] Hüseyin Hakli, Harun Uguz "A novel particle swarm optimization algorithm with Levy flight" in Applied Soft Computing 23 (2014) 333–34