International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:4, 2014

An Advanced Nelder Mead Simplex Method for Clustering of Gene Expression Data

M. Pandi, K. Premalatha

***Abstract*—**The DNA microarray technology concurrently monitors the expression levels of thousands of genes during significant biological processes and across the related samples. The better understanding of functional genomics is obtained by extracting the patterns hidden in gene expression data. It is handled by clustering which reveals natural structures and identify interesting patterns in the underlying data. In the proposed work clustering gene expression data is done through an Advanced Nelder Mead (ANM) algorithm. Nelder Mead (NM) method is a method designed for optimization process. In Nelder Mead method, the vertices of a triangle are considered as the solutions. Many operations are performed on this triangle to obtain a better result. In the proposed work, the operations like reflection and expansion is eliminated and a new operation called spread-out is introduced. The spread-out operation will increase the global search area and thus provides a better result on optimization. The spread-out operation will give three points and the best among these three points will be used to replace the worst point. The experiment results are analyzed with optimization benchmark test functions and gene expression benchmark datasets. The results show that ANM outperforms NM in both benchmarks.

***Keywords*—**Spread out, simplex, multi-minima, fitness function, optimization, search area, monocyte, solution, genomes.

I. Introduction

GENE expression is the method in which information from gene is used for the generation of gene product. Gene expression data is used to interpret genetic code of a sample. The information regarding building and maintain of cells for an organism is carried by genes. The genes are encoded in long strands of DNA in most of the living organisms. Usually DNA is having a double helix structure. It consists of four types of nucleotide subunits to form a chain. The nucleotide subunits are namely adenine, cytosine, guanine and thymine. Guanine pairs with cytosine and adenine pairs with thymine. Transcription and translation are the two steps in gene expression, in which transcription produces messenger RNA from DNA. Messenger RNA or mRNA is single stranded. In the translation step, defined sequences of amino acids are produced from mRNA. A Micro array experiment evaluates a large number of DNA sequences consisting of genes, cDNA clones or expressed sequence tags under different conditions. These conditions may be a time based or tissue samples based. A gene expression [1] data set from a micro-array experiment can be represented by a real-valued expression matrix. In this matrix, rows represent expression pattern of genes, columns represent expression profile of samples or experimental conditions.

M.Pandi is with the Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu, India (e-mail: mpandi123@gmail.com).

Datasets are represented as set of genes G = $\{g_1, g_2, g_3 \ldots g_n\}$, where g_i represents i^{th} gene in the data set and w_{ij} represents expression profile [2] of i^{th} gene at j^{th} samples/conditions. Fig. 1 represents dataset with n genes and m samples/conditions vector of real numbers represented as follows.

Fig. 1 Gene expression data matrix

The expression levels of various genes can be represented by using Microarray technology. DNA molecules of various genes are placed in discrete spots of a microscope slide. A simple microarray is an N×M array, where N is the number of genes and the number of conditions is given by M. The row in the array represents a gene and columns represent the conditions.

Gene expression profiling provides many ways to study about the gene expression patterns. Co-expressed genes can be identified by the cluster analysis of gene expression data. The main step in analyzing gene expression data is to identify the group of genes that are having the similar expression pattern.

Data mining is an area, where we can extract knowledge from a large database. Knowledge extraction involves many tasks. Clustering is one of the important data mining tasks which is having a number of applications in the area of biology and other disciplines. Here similar objects are grouped in a cluster. Clustering of gene expression data is helpful to understand gene regulation, gene function and cellular processes. While considering the case of gene expression data, the elements are genes. There is no previously defined class label for clustering.

There are mainly two categories of clustering. The first one is hierarchical method and the other one is partitional method. Partitional method will divide the objects to various clusters based on some conditions. Partitional method is faster than

the hierarchical method but this method has a disadvantage that we have to mention the number of clusters in priori.

Hierarchical methods are classified as agglomerative type and divisive type. Hierarchical method will create a tree structure to form clusters. Agglomerative algorithm works with a bottom-up approach while divisive works with a top-down approach.

In the case of partitional and hierarchical, the solutions may be local optimum or may not be necessarily the global solution. This makes worse when the solution space is very large.

Sorting N objects into K groups can be done in many ways which is given by [3]

$$Q(N, K) = \frac{1}{K!}\sum_{i=0}^{K}(-1)^{i}\binom{K}{i}(K-i)^{N} \quad (1)$$

For example, for $Q(25,5)$ there are 2,436,684,974,110,751 ways of sorting 25 objects into 5 groups. If the number of clusters is unknown the objects can be sorted $\sum_{K=1}^{N} Q(N, K)$ ways. For 25 objects this is over 4×10^{18}. Clearly, it is impractical for an algorithm to exhaustively search the solution space to find the optimal solution. Furthermore traditional clustering algorithms search relatively a less subset of the solution space. As a result, the probability of success of these methods is small and it requires for an algorithm with the potential to search large solution spaces effectively. Contrary to the localized searching of the traditional algorithm, the global optimization algorithm [4] performs a globalized search in the entire solution space.

Optimization is the process of selecting the best element from some sets of available alternatives under certain constraints (if any). This process can be solved by minimizing or maximizing the objective or cost function of the problem. In each iteration of the optimization process, choosing the values (e.g. real or integer variables) from within an allowed set is done systematically until the minimum or maximum result is reached or when the stopping criterion is met.

Nelder Mead simplex is one of the best known multi-dimensional, unconstrained optimization methods without derivatives proposed by Nelder and Mead in 1965 [4]. This method is suitable for problems with non-smooth functions as it does not require any derivative information. This method is widely used where the function values are uncertain or subject to noise to solve parameter estimation and similar statistical problems. It can also be used for problems with discontinuous functions which occur frequently in statistics and experimental mathematics. When we consider n dimension there will be n+1 vertices, i.e. 2 dimensional problems will have 3 vertices in which each vertex represents a solution. In each evolution the simplex may move, expand or shrink. When all the three vertices finally converge to a single point, the stopping criterion is met.

The rest of the paper is organized as follows: Section II describes the literature review on Nelder Mead Algorithm and Gene expression data clustering. The overview NM is given in Section III. Section IV presents the ANM algorithm for gene expression data clustering. The experiment results are analyzed and demonstrated in Section V.

II. Related Works

Spendley et al. [5] proposed a simple search method which is modified by Nelder and Mead. The modified method was an unconstrained and non-linear. In the simplex Nelder Mead method, the coordinate with highest function value is replaced with a reflected or extended alternate point. Iteratively changing the coordinate with maximum function value will finally result in an optimum point. Nelder and Mead proposed the Nelder Mead optimization method in the year 1965. Krovi et al., [6] proposed a fitness function to evaluate the partitions formed by only two clusters. This function utilizes the average distance between objects and their respective cluster centroids. A combined Nelder Mead method was proposed by Durand and Alliot [7]. It was tested on two benchmark functions the GA and local optimization technique, which is found ineffective when used separately.

Nazareth et al. [8] proposed a Variant of the Nelder-Mead Algorithm Based on Golden-Section Search. Marco et al. [9] introduced a globalized Nelder Mead method for the optimization purpose. Here globalization is achieved by a method called probabilistic restart and local searches are performed by improved Nelder Mead algorithm. Fazel Famili et al. [10] proposed evaluation and optimization of clustering in gene expression data analysis. This work introduced new cluster quality method called stability. Vito Di Gesú et al. [11] proposed genetic algorithm for clustering of gene expression data called Genclust. The performance was evaluated based on real dataset and have used internal and external validation techniques. Chelouah et al. [12] proposed a hybrid method combining continuous tabu search and Nelder–Mead simplex algorithms for the global optimization of multi-minima functions. Simplex search is used to accelerate the convergence towards a minimum. Tabu Search allows covering widely the solution space, to stimulate the search towards solutions far from the current solution and to avoid the risk of trapping into a local minimum. Kim et al. [13] compared the performance of several clustering methods based on data preprocessing including strategies of normalization or noise clearness.

Ma et al. [14] proposed a novel evolutionary algorithm called evolutionary clustering (EvoCluster). It encodes an entire cluster grouping in a chromosome so that each gene in the chromosome encodes one cluster. Kustra R et al. [15] Introduced clustering expression data that permits integration of various biological data sources through combination of corresponding dissimilarity measures. This work reviews about genomic data fusion and validating results from clustering expression data. Satapathy, S. C. et al. [16] developed an Efficient Hybrid Algorithm for Data Clustering Using Improved Genetic Algorithm and Nelder Mead Simplex Search. In this paper, to improve the accuracy of data clustering IGA is tested with many benchmark test functions.

Kerr G et al. [17] conducted a review on techniques of clustering gene expression data. This work mentions about the limitations and addresses them and provides a framework for the evaluation of clustering in gene expression analyses. Zhihua Du [18] proposed a new clustering algorithm for clustering gene expression data called PK means. This method incorporates Particle Pair Optimizer (PPO), K means and Fuzzy Kmeans for clustering which provide a more accurate result.

Wei Liu et al. [3] proposed a novel methodology for finding the regulation on gene expression data. This work helps to find feature subset to build the classifier for gene expression data analysis. Principal component analysis was employed to construct the classifier. Zahara and Kao [19] proposed a Hybrid Nelder–Mead simplex search and particle swarm optimization for constrained engineering design problems. They introduce embedding constraint handling methods including the gradient repair method and constraint fitness priority-based ranking method in NM–PSO as a special operator to deal with satisfying constraints. Gao et al. [20] have implemented the Nelder-Mead Simplex Algorithm with Adaptive Parameters. Rui Xu et al. [21] conducted a review on clustering algorithm in biomedical research. The work provides an overview of the status quo of clustering algorithms, to illustrate examples of biomedical applications based on cluster analysis, and to help biomedical researchers to select the most suitable clustering algorithms for their own applications.

Nam Pham et al. [22] proposed an improved Nelder Mead method and its application using a quasi-gradient parameter. Wang et al. [23] proposed a parameter identification of chaotic systems by hybrid Nelder–Mead simplex search and differential evolution algorithm. This work is done by suitably fusing the DE-based evolutionary search and NM simplex-based local search, for achieving satisfactory optimization performances. Nagi et al. [24] had done a survey on gene expression data clustering analysis. This work mentions about various approaches to gene expression data analysis using clustering techniques. This work also discuss about the performance of various existing clustering algorithms under each of these approaches and proximity measures. Salome et al. [25] proposed an efficient clustering of gene expression data. This work introduced methods to improve the searching and the clustering performance in genomic data from commonly used clustering techniques. Liu and Yang [26] proposed a new hybrid nelder-mead particle swarm optimization for coordination optimization of directional overcurrent relays. Here PSO is the main optimizer, and the Nelder-Mead simplex search method is used to improve the efficiency of PSO due to its potential for rapid convergence. In the result, the work is compared with actual PSO.

Jaskowiak et al. [27] investigated about the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures in 52 data sets from time course and cancer experiments. This work mentions about commonly employed measures, such as Pearson, Spearman, and Euclidean distance. Recently Balamurugan et al. [28] conducted on a Comparative Study on Swarm Intelligence Techniques for Biclustering of Microarray Gene Expression Data.

III. Nelder Mead Simplex Method

Nelder Mead simplex algorithm, is an algorithm that exploits local information and converges to the nearest optimal point. It is an algorithm searching for local minimum and can be used for multi-dimensional optimizations. It does not have to compute derivatives to move along a function as gradient methods.

Nelder and Mead devised a simple method for finding a local minimum of a function of several variables. A simplex is a triangle for two variables, and the method is a pattern search that compares function values at the three vertices of a triangle. The vertex where f (x, y) is largest is the worst vertex, which rejected and replaced with a new vertex. A new triangle is formed and the search is continued. A sequence of triangle will be generated, which might have different shapes for which the function values at the vertices get smaller and smaller. The coordinates of the minimum point is found by reducing the size of the triangle. The algorithm will find the minimum of a function of N variables which is computationally compact and effective.

A. Initial Triangle BGW

Let f (x, y) be the function that is to be minimized. To Let the vertices of the triangle: **V**k = (xk , yk), k = 1, 2, 3. The function f (x, y) is then evaluated at each of the three points: zk = f (xk , yk) for k = 1, 2, 3. The subscripts are then reordered so that $z1 \leq z2 \leq z3$. We use the notation

$$B = (x1, y1),\ G = (x2, y2),\ \text{and}\ W = (x3, y3) \tag{2}$$

B. Midpoint of the Good Side

The construction process uses the midpoint of the line segment joining B and G. It is found by averaging the coordinates:

$$M = \frac{B + G}{2} = \left(\frac{x1 + x2}{2}, \frac{y1 + y2}{2} \right) \tag{3}$$

C. Reflection Using the Point R

The function decreases as we move along the side of the triangle from W to B, and it decreases as we move along the side from W to G. Hence it is feasible that f (x, y) takes on smaller values at points that lie away from W on the opposite side of the line between B and G. We choose a test point R that is obtained by "reflecting" the triangle through the side BG. First find the midpoint M of the side BG to determine R and then draw the line segment from W to M whose length is d. This last segment is extended a distance d through M to locate the point R in Fig. 2.

The vector formula for R is

$$R = M + (M - W) = 2M - W \quad (4)$$

D. Expansion Using the Point E

If function value R is lesser than function value of W, then the simplex has moved in the correct direction toward the minimum. There exists a possibility that the minimum is just a bit farther than the point R. using this assumption we extend the line segment through M and R to the point E. This forms an expanded new triangle BGE in which the point E is found by moving an additional distance d along the line joining M and R in Fig. 2. If the function value at R is greater than the function value at E, then we have found a better vertex than R.

The vector formula for E is

$$E = R + (R - M) = 2R - M. \quad (5)$$

E. Contraction Using the Point C

If the function values at R and W are the same, then another point must be tested. Perhaps M is having the smaller function, but we cannot replace W with M because we must have a triangle. Consider the two midpoints C1 and C2 of the line segments WM and MR, respectively in Fig. 3.

C is the point with the smaller function value and the new triangle is BGC. The choice between C1 and C2 may be inappropriate for the two-dimensional case, but it is important in higher dimensions.

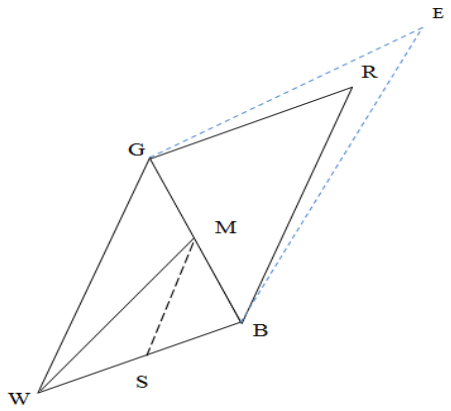

Fig. 2 Operations performed on initial triangle

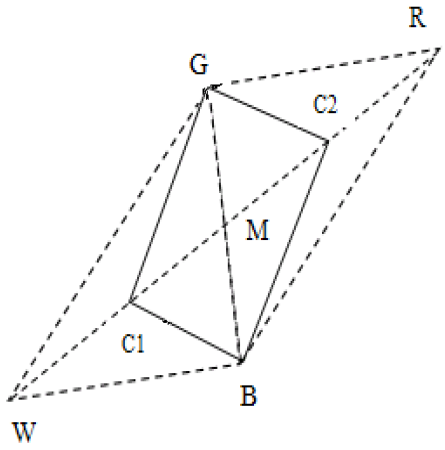

Fig. 3 Finding contraction points

F. Shrink toward B

If the function value at W is not greater than the value at C, then the points G and W must be shrunk toward B in Fig. 2. The point G and W is replaced with M and S respectively, which is the midpoint of the line segment joining B with W.

G. Logical Decisions for Each Step

A computationally efficient algorithm should perform function evaluations only if needed. In each step, a new vertex is found, which replaces W. As soon as a new vertex is found, further investigation is not needed and the iteration step is completed. The stopping criterion is that, finally all the three points of the vertex will become the same point.

H. Logical Decision for Nelder Mead Algorithm

IF $f(R) < f(G)$, THEN Perform Case (i) {either reflect or extend }

ELSE Perform Case (ii) {either contract 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
Compute *S* and *f (S)*
Replace *W* with *S*
Replace *G* 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.

IV. An Advanced Nelder Mead Simplex Method for Clustering Gene Expression Data

A. Problem Statement

The clustering problem is expressed as follows:

The set of M genes $G = \{G_1, G_2, \ldots, G_N\}$ is to be clustered. The genes are to be grouped into non-overlapping clusters $C = \{C_1, C_2, \ldots, C_K\}$ (C is known as a clustering), where K is the number of clusters, $C_1 \cup C_2 \cup \ldots \cup C_K = G$, $C_i \neq \varphi$, and $C_1 \cap C_2 = \varphi$ for $i \neq j$.

Assuming $f : G \times G \rightarrow \Re^+$ is a measure of distance between genes. Clustering is the task of finding a partition $\{C_1, C_2, \ldots, C_K\}$ of G such that

$$\forall\, i, j \in \{1, \ldots, K\},\ j \neq i, \forall\, x \in C_i : f(x, O_i) \geq f(x, O_j)$$

where O_i is one cluster representative of cluster C_i.

The goal of clustering is stated as follows:

Given,

1. A set of genes $G = \{G_1, G_2, \ldots, G_N\}$,
2. A desired number of clusters *K*, and
3. An objective function or fitness function that evaluates the quality of a clustering, the system has to compute an assignment $g : G \rightarrow \{1, 2, \ldots, K\}$ and maximizes the objective function.

The global maximization problem can be defined as follows: Given $f : S \rightarrow \Re$ where $S \subseteq \Re^N$ and N is the dimension of the search space *S*. Find $y \in S$ such that $f(y) \geq f(z), \forall z \in S$. The variable y is called the global maximizer of *f* and *f(y)* is called the global maximum. The process of finding the global optimal solution is known as global optimization [29]. A true global optimization algorithm will find y regardless of the selected starting point $z_0 \in S$. The variable y_L is called the local maximizer of *L* because $f(y_L)$ is the largest value within a local neighborhood, *L*. Mathematically speaking, the variable y_L is a local maximizer of the region *L* if $f(y_L) \geq f(z), \forall z \in L$ where $L \subset S$.

For clustering, two measures of cluster quality are used. One type of measure allows comparing different sets of clusters without reference to external knowledge and is called an internal quality measure. The other type of measures evaluates how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called an external quality measure.

Internal criterion function focuses on producing a clustering solution that optimizes a particular criterion function that is defined over the genes. These genes are part of each cluster and do not take into account the genes assigned to different clusters. The proposed work applies global searching strategies for identifying optimal clusters in the exhaustive search space. Typical objective function in clustering formalizes the goal of achieving high intra-cluster similarity, where genes within a cluster are similar, and low inter-cluster similarity, where genes from different clusters are dissimilar.

This is an internal criterion for the quality of a clustering.

It is formulated by minimizing a formal objective function Mean Squared Error (MSE) distortion.

$$\text{MSE(P)} = \sum_{i=1}^{N} \left\| G_i - C_{p(i)} \right\|^2 \qquad (6)$$

where

N is the number of Genes;

$G = \{G_1, G_2, \ldots, G_N\}$ is a set of *N* gene samples;

$P = \{ p(i) \mid i = 1, \ldots N \}$ is class label of *G*

$C = \{ c_j \mid j = 1, \ldots, K \}$ are *K* cluster centroids.

B. Vertex Representation

The proposed work represents each vertex as a solution. Each vertex will be having d number of values where d is the dimension. An example of cluster representation is given in Fig. 2. The solution represents G_1 is present in cluster *#1*, G_2 is present in cluster *#2*, G_3 is present in cluster *#1* and so on.

$$[\,1 \quad 2 \quad 1 \quad 3 \quad \ldots \quad 2\,]$$

Fig. 4 Cluster representation

At the initial stage, the random number is generated between 0 and 1 and K is the number of clusters. Let v be the generated random number then the cluster value v' is

$$v' = \text{int}\,(vK) + 1 \qquad (7)$$

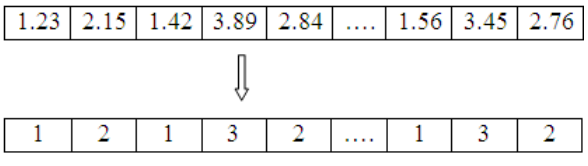

Fig. 5 The representation of vertex for clustering

C. An Advanced Nelder Mead Simplex Algorithm

In the proposed work, the reflection and expansion steps are removed and instead a new concept called spread out is used. This is done based on the assumption that a better point will be available away from the best point and good point.

Expansion was done with the same assumption but in expansion step we will be receiving only one point which is away from the best and good points. Nelder Mead method is suffering from premature convergence and early restart. So the efficiency of algorithm is comparatively less to the other algorithms. To overcome this problem and to increase the global search area, Advanced Nelder Mead is proposed.

1. Initial Triangle BGW

Initially the three points selected are evaluated using the given fitness function and the best, good and worst points are selected accordingly as in existing system.

2. Reflection Using the Point R

The function decreases as we move along the side of the triangle from W to B, and it decreases as we move along the side from W to G. Hence it is feasible that f (x, y) takes on smaller values at points that lie away from W on the opposite side of the line between B and G. We choose a test point R that is obtained by "reflecting" the triangle through the side BG. The vector formula for R is

$$R = M + (M - W) = 2M - W. \quad (8)$$

3. Spread out

For a given initial triangle, the best, good and worst point is found and the midpoint of best and good point is found which is called M1. Next we propose a new operation called spread out. Spread out is done from the worst point towards the good point and best point. Two points E1 and E2 can be found out using the given formula.

$$E1 = 2G - W \quad (9)$$

$$E2=2B-W \quad (10)$$

$$M2 = \frac{E1 + E2}{2} \quad (11)$$

In the given Fig. 6 WG and WB are extended to E1 and E2 respectively. Then the midpoint of E1 and E2 is calculated called M2. Now the corresponding fitness functions for E1, E2 and M2 are found. Among these the best one is compared with W and if found better, then W is replaced with new point. In Fig. 6 if it is found worst, then contraction and shrink operations are performed.

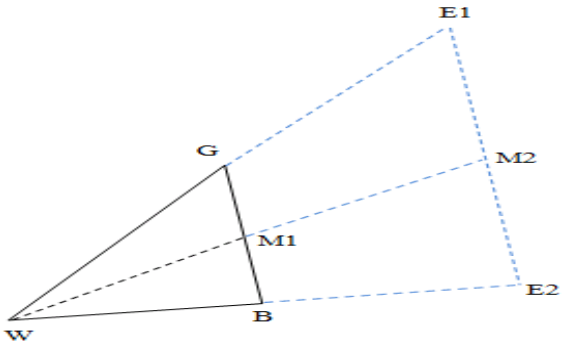

Fig. 6 Finding spread out point and its midpoint

If E2, G, and B are the better points obtained after an iteration, then the new triangle formed is E2 G B.

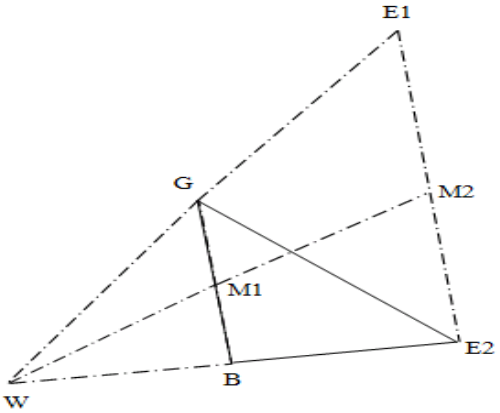

Fig. 7 Finding new triangle with available best points

4. Contraction Using the Point C

If spread out option doesn't work, then another point must be tested. For that we may use the contraction as per in Nelder Mead Simplex method. The point with the smaller function value is called C, and the new triangle is BGC.

5. Shrinking towards the Point B

If contraction also doesn't work then we may go for the shrink option to find a better result. The other two points are shrunk towards the best point. Usually shrinking will be a rare case in practice.

6. Logical Decision for Advanced Nelder Mead Algorithm

BEGIN {Case (i).}
Find the points E1, E2 and M2
Calculate $f(E1)$, $f(E2)$ and $f(M2)$
Select the best point among E1, E2 and M2 and choose as E
IF $E<f(W)$
Replace W with E
ELSE
{Case (ii)}
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
                Compute S and f (S)
                Replace W with S
                Replace G with M
                 ENDIF
        END {Case (ii).}
```

where $f(R)$, $f(G)$, $f(B)$, $f(E1)$, $f(E2)$, $f(M2)$, $f(C)$, $f(S)$, $f(W)$ are the fitness functions of R, G, B, E1, E2, M2, C, S and W respectively.

V. Experimental Results

In this section, the experiments that have been done to evaluate the performance of an NM and ANM for five optimization benchmark test functions are described. Table I shows the optimization function with range, dimension and models. The parameter setting for NM and ANM includes number of vertex (solution) as 30 and number of iteration as 20000 for the benchmark functions.

TABLE I
Benchmark Optimization Functions Used in Experiments

No	Function	Min.	Range	D	C	Formulation
F1	Ackley	0	[-32,32]	30	MN	$f(x) = -a\exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(cx_i)\right) + a + \exp(1)$
F2	Hartman3	-3.86278	[0,1]	3	MN	$f(x) = -\sum_{i=1}^{4} c_i \exp\left[-\sum_{j=1}^{3} a_{ij}(x_j - p_{ij})^2\right]$
F3	Hartman6	-3.32237	[0,1]	6	MN	$f(x) = -\sum_{i=1}^{4} c_i \exp\left[-\sum_{j=1}^{6} a_{ij}(x_j - p_{ij})^2\right]$
F4	Hump camel 6	-1.0316	[-5,5]	2	MN	$f(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x1x2 - 4x_2^2 + 4x_2^4$
F5	Michalewicz	-1.8013	[0,10]	2	MS	$f(x) = -\sum_{i=1}^{d} \sin(x_i)\sin^{2m}\left(\frac{ix_i^2}{\pi}\right)$
F6	Rastrigin	0	[-5.12,5.12]	30	MS	$f(x) = \sum_{i=1}^{n}\left[x_i^2 - 10\cos(2\pi x_i) + 10\right]$
F7	Rosenbrock	0	[-30,30]	30	UN	$f(x) = \sum_{i=1}^{n-1}\left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right]$
F8	Shubert	-186.731	[-10,10]	2	MN	$f(x) = \left(\sum_{i=1}^{5} i\cos((i+1)x_1 + i)\right)\left(\sum_{i=1}^{5} i\cos((i+1)x_2 + i)\right)$

Min: Global Minimum, D: Dimension, C: Characteristics, U: Unimodal, M: Multimodal, S: Separable, N: Non-Separab

TABLE II
Experiment Results for Benchmark Functions Derived from NM and ANM

Function	Best optimum		Mean function Evaluation value	
	NM	ANM	NM	ANM
F1	0	0	9833.65	8509.47
F2	-3.8628	-3.8627	14715.65	9276.85
F3	-3.3224	-3.3223	18313.75	5682.4
F4	-1.0316	-1.0316	9517.85	7051.03
F5	-1.9680	-1.8013	7462.3	652.8
F6	0.5138	0	8357.95	7148.24
F7	1.3078	0	8760	5439.60
F8	-185.75	-186.731	9459	6545.3

Table II shows the result obtained from NM and ANM for the functions listed in Table I. The results show that the mean function evaluation of ANM is lower than NM and the global optimum is obtained for all eight functions by ANM. This paper proposes a new idea of incorporating an operation called spread out to the Nelder mead method. This method finds out new vertices and selects the best among given vertices to form a new triangle. It will increase the global search area and thus provides a better result on optimization. In NM method each and every iteration selects only one vertex from the search area while for ANM, in each iteration six coordinates are evaluated from the search area and the best three coordinates are selected to form a new triangle. Since this process continues for many numbers of iterations, ANM can obtain a better solution than NM.

A. Datasets

An ANM algorithm is tested on four datasets of gene expression data, The Yeast Cell Cycle (YCC) dataset [30] there are more than 6,000 genes during two cell cycles from Yeast measured at 17 time points. A subset of 698 genes is identified based on their peak times of five phases of the cell cycle and annotated. The resulting 698x72 data matrix is standardized (i.e., for each row, the entries are scaled so that the mean is zero and the variance is one) and used for our experiments. Second one is Reduced Yeast Cell Cycle (RYCC). This data set originates in the one by Cho et al. [31]. Ka Yee Yeung extracted 384 genes from the yeast cell cycle data set in Cho et al. to obtain a 384x17 data expression matrix. It is to be pointed out that each gene in the RYCC data set appears also in the YCC data set. However, the

dimensionality of the two data sets is quite different, and this may cause algorithms to behave differently.

Third one is Reduced Peripheral Blood Monocytes (RPBM). We have randomly picked 10% of the cDNAs in each of the 18 original classes. Whenever that percentage is less than one, we have retained the entire class. The result is a 235x139 data matrix, and the true solution is readily obtained from that of PBM. The fourth dataset is Rat Central Nervous System (RCNS) which is a data set obtained by reverse transcription coupled PCR to study the expression levels of 112 genes during rat central nervous system development over 9 time points. This results in a 112x9 data matrix. Wen et al. [32] studied it to obtain a division of the genes into 6 classes, in which four of them are composed of biologically functionally related genes. Such a division is assumed to be the true solution. Previously Wen et al. performed two transformations on the data for each gene, (1) Each row is divided by its maximum value and (2) to capture the temporal nature of the data, and the difference between the values of two consecutive data points is added as an extra data point. So the final data set consists of a data matrix of dimension 112x17, which is the input to our algorithms. The second transformation has the effect to enhance the similarity between genes with closely parallel, but offset, expression patterns. Table III shows the parameter and its value used for clustering gene expression data.

TABLE III
PARAMETER AND ITS VALUE FOR BENCHMARK DATASETS

Parameter	Value
Number of vertex (solutions)	60
Number of iteration	200
Cluster size	3 to 10

Figs. 8-11 correspondingly show the results obtained from NM and ANM for RPBM, RYCC, YCC and RCNS with the varying cluster size from 3 to 10. The results show that all the four datasets and varying cluster size the fitness value obtained from ANM outperforms NM with its new operation spread out and increased search area.

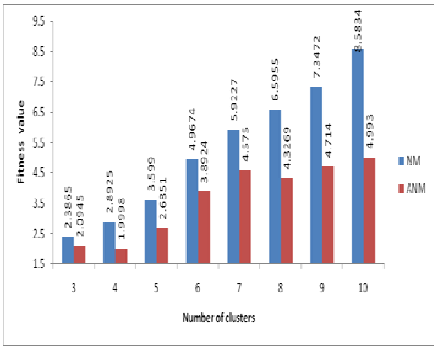

Fig. 8 Experiment results for RPBM data

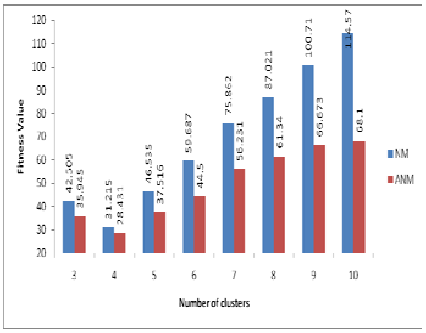

Fig. 9 Experiment results for RYCC data

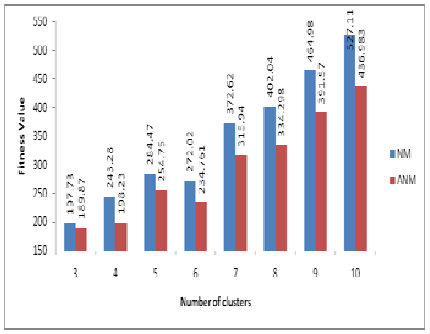

Fig. 10 Experiment results for YCC data

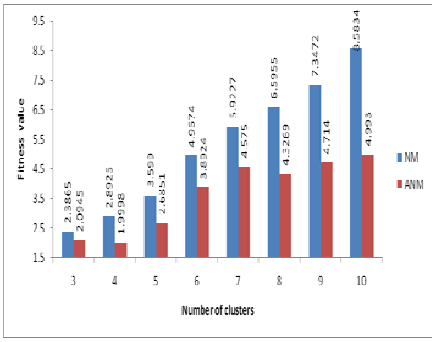

Fig. 11 Experiment results for RCNS data

VI. CONCLUSIONS

Microarrays are useful to simultaneously monitor the expression profiles of thousands of genes under various experimental conditions. Identification of gene cluster is the main goal in gene expression data analysis and is an important task in bioinformatics research. In this work the gene expression data are clustered using NM and ANM. To avoid premature convergence due to stagnation NM is modified as ANM by introducing new spread-out operation. While comparing to Nelder Mead method, the proposed work is having less number of operations performed and comparatively good results are obtained. The steps like reflection and expansion is removed from Nelder Mead method and a single step called spread out is used in the new method. This may increase the global search considerably and will result in a better solution. The spread out operation will give three points including the midpoint which can be

compared with the previous best point and update the new best, good and worst points, but reflection and expansion step may give 2 points. The performance of NM and ANM is analyzed with optimization benchmark test functions and gene expression benchmark data sets. The results show that ANM outperforms NM in both benchmarks.

REFERENCES

[1] Alon, U. Barkai, N. Notterman, D.A. Gish, K. Ybarra, S. Mack, D. and Levine, A.J. "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Array", In Proc. of the Natl. Acad. Sci. U.S.A.,1999, Vol. 96, No. 12, pp. 6745-6750.

[2] C. Ding, "Analysis of Gene Expression Profiles: Class Discovery and Leaf Ordering", In *Proc. of the Int. Conf. Comput. Mol. Biol. (RECOMB)*, Berlin, Germany, 2002, pp. 27-136.

[3] Wei Liu, Bo Wang, Jarka Glassey, Elaine Martin, and Jian Zhao, "A novel methodology for finding the regulation on gene expression data", *Prog. Nat. Sci.*, Vol. 19, pp. 267-272, 2009.

[4] J.A. Nelder, and R. Mead, "A simplex method for function minimization", Comput. J. Vol. 7, pp. 308–313, 1965.

[5] W. Spendley, G.R. Hext, and F.R. Himsworth, "Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation", Technometrics, Vol. 4, pp. 441–461, 1962.

[6] R. Krovi, "Genetic Algorithms for Clustering: A Preliminary Investigation", In Proc. of the 25th Hawaii Int. Conf. Syst. Sci., 1992, Vol. 4, pp. 540-544.

[7] N. Durand, J.M. Alliot, W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, R.E. Smith (Eds.), "A combined Nelder–Mead simplex and genetic algorithm", In Proc. of the Genet. Evol. Comput. Conf. GECCO_99, Morgan Kaufmann, Orlando, FL, USA, 1999, pp. 1–7.

[8] L. Nazareth, P. Tseng, "Gilding the Lily: A Variant of the Nelder-Mead Algorithm Based on Golden-Section Search", *Comput. Optim. Appl*, vol. 22, no. 1, pp. 133–144, 2002.

[9] A. Marco Luersen and Rodolphe Le Riche, "Globalised Nelder-Mead method for Engineering optimization", *J. Comput. Struct*., Vol 3, 10 pages, 2004.

[10] F. Fazel, L. Ganming, L. Ziying, (2004) "Evaluation and optimization of clustering in gene expression data analysis", *BMC Bioinf.* Vol. 20, No. 10, pp. 1535-1545, 2004.

[11] Vito Di Ges , Raffaele Giancarlo, Giosu Lo Bosco, Alessandra Raimondi and Davide Scaturro, "GenClust: A genetic algorithm for clustering gene expression data", *BMC Bioinf.* Vol. 280, No.6, pp. 1-11, 2005.

[12] R. Chelouah and P. Siarry, "A hybrid method combining continuous Tabu search and NeldereMead simplex algorithms for the global optimization of multiminima functions". *Eur. J. Oper. Res.* Vol. 161, No. 3, pp. 636-654, 2005.

[13] S.Y. Kim and H. Toshimitsu, "Evaluation of Clustering based on Preprocessing in Gene Expression Data", *World Acad. Sci. Eng. Technol.,* Int. J. Comput. Inf. Sci. Eng. Vol.1, No. 5, pp. 154-159, 2007.

[14] P.C.H. Ma, K.C.C. Chan, X. Yao, and D.K.Y. Chiu,"An evolutionary clustering algorithm for gene expression microarray data analysis", *IEEE Trans. Evol. Comput.*, Vol. 10, No. 3, pp. 296-314, 2006.

[15] R. Kustra, "A factor analysis model for functional genomics", *BMC Bioinf.*, Vol. 216, No. 7, pp. 1-13, 2006.

[16] Dr. S C Satapathy et.al Article "An Efficient Algorithm for Data Clustering using improved Genetic Algorithm and Nelder Mead Simplex Search" *IEEE Int. Conf. Comput. Intell. Multimedia Appl*. Sivakasi, India Dec 2007.

[17] G. Kerr, H.J. Ruskin, M. Crane, and P. Doolan, "Techniques for clustering gene expression data", *Comput. Biol. Med.*, Vol. 38, pp. 283–293, 2007.

[18] Zhihua Du, YiweiWang, Zhen Ji, "PK-means: A new algorithm for gene clustering", *Comput. Biol. Chem.*, Vol. 32, pp.243-247, 2008.

[19] Erwie Zahara, Yi-Tung Kao, "Hybrid Nelder-Mead simplex search and particle swarm optimization for constrained engineering design problems", *Expert. Syst. Appl. Vol. 36, No.* 2, pp. 3880-3886, 2009.

[20] F. Gao, L. Han, "Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters", *Comput. Optim. Appl*, vol: 51. pp. 259-277, 2010.

[21] Rui Xu and D.C. Wunsch," Clustering Algorithms in Biomedical Research: A *Review", IEEE Rev. Biomed. Eng.,* Vol. 3, pp. 120 – 154, 2010.

[22] Nam Pham and Bogdan M.Wilamowski, "Improved Nelder-Mead simplex method and applications", J. Comput., Vol 3, issue 3, pp:55-63, 2011.

[23] L. Wang, Y. Xu, and L. Li, "Parameter identification of chaotic systems by hybrid Nelder – Mead simplex search and differential evolution algorithm", Expert Syst. Appl., vol. 38, pp. 3238-3245, 2011.

[24] Sajid Nagi, D.K. Bhattacharyya, and J.K. Kalita, "Subspace Clustering in Gene Expression Data Analysis: A Survey, in Machine Intelligence: Recent Advances", *Narosa Publ., Delhi*, pp. 211-219, 2011.

[25] J. Jacinth Salome and R.M. Suresh, "Efficient Clustering for Gene Expression Data", *Int. J. Comput. Appl.,* Vol. 47, pp. 30-35, 2012.

[26] An Liu, and Ming-Ta Yang, "A New Hybrid Nelder-Mead Particle Swarm Optimization for Coordination Optimization of Directional Overcurrent Relays", *Math. Prob. Eng.,* Vol. 2012, pp. 1-18, 2012.

[27] P. A. Jaskowiak and R.J.G.B Campello, "Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data", *Proc. Braz. Symp. Bioinf. Brasilia. Braz*, 2011, pp. 1-8.

[28] R.Balamurugan A.M.Natarajan and K. Premalatha, "Comparative Study on Swarm Intelligence Techniques for Biclustering of Microarray Gene Expression Data.", *World Acad. Sci. Eng. Technol.,* Int. J. Comput. Inf. Sci. Eng. Vol.8, No. 2, pp. 4619-4625, 2014.

[29] P. Gray, W.E. Hart, L. Painton, C. Phillips, M. Trahan, and J. Wagner, "A Survey of Global Optimization Methods", *Tech. Rep.*, *Sandia Nat. Lab*, 2000.

[30] P. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and Futcher, (1998) "Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization", Mol. Biol. Cell, Vol. 9, pp - 3273-3297, 1998.

[31] R. J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L.W.Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart and R. Davis, (1998) "A genome-wide transcriptional analysis of the mitotic cell cycle", *J. Mol. Cell*..Vol:2, pp:65-73, 1998.

[32] X. Wen, S. Fuhrman, G.S. Michaels, G.S. Carr,D.B. Smith, J.L. Barker and R. Somogyi, "Large scale temporal gene expression mapping of central nervous system development". In *Proc. of the Natl. Acad. Sci. U.S.A.2005,* Vol:95.,pp:334-339.