International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 7 (2019) pp. 1638-1643
© Research India Publications. http://www.ripublication.com

Microarray Image Analysis Using Adaptive Data Clustering Algorithms

B.SivaLakshmi[1], N.Naga Malleswara Rao[2]

[1]Phd Research Scholar, Acharya Nagarjuna University, Vijayawada, AP, India.

[2]Professor, Dept of IT, RVR and JC College of Engineering, Guntur, AP, India

Abstract

Microarrays are used in various applications like disease diagnosis, drug discovery and bio-medical research. A Microarray image contains thousands of spots and each of the spot contains multiple copies of single DNA sequence. The analysis of microarray image is done in three stages: gridding, segmentation and information extraction. The microarray image analysis takes the spot intensity data as input and produces the spot metrics as output which are used in classification and identification of differently expressed genes. The intensity of each spot indicates the expression level of the particular gene. In this paper, a noise removal method based on Bi-dimensional Empirical Mode Decomposition (BEMD) and wavelets is presented with different interpolation functions used in decomposition of image. Generally, clustering algorithms are used for segmentation of microarray image. Clustering algorithm such as K-means, Fuzzy c-means etc., has been used in the literature. The main requirements for any clustering algorithm is the number of clusters K. Estimating the value of K is difficult task for given data. This paper presents adaptive data clustering algorithms which generates accurate segmentation results with simple operation and avoids the interactive input K (number of clusters) value for segmentation of microarray image. The qualitative and quantitative results shows that adaptive data clustering algorithms are more efficient than normal data clustering algorithms in segmenting the spot area, thus producing more accurate expression-ratio.

1. INTRODUCTION

The work flow of microarray image analysis was separated into four stages [1].

I. *Image merging*, is the construction of the combined eight-bit image from intensity measurements of both red (Cy5) and green (Cy3) fluorescent dye, that is computationally efficient in doing subsequent gridding and segmentation steps. The combine image *I* is obtained by using some arbitrary function *f* ie.,$I(i,\ j)=f(R(i,\ j),G(i,\ j)$ where *R* is an image corresponding to red channel and *G* is an image corresponding to green channel.

II. *Gridding* [2], is the mechanism of identification of location of the gene spots in the image without any overlapping. The problem of gridding is divided into two stages, *sub-gridding* and *spot-detection*. *Sub-gridding* refers to finding the block index corresponding to a spot on the microarray image, while *spot-detection*, is finding the location *(i, j)* of a specified spot in that indexed block.

III. *Segmentation* [3], is the problem of classifying the pixels of image into a set of non-overlapping regions based on specific criteria. In microarray image, the pixels can be classified into spot, background or noise.

IV. *Information Extraction* [4], includes the calculation of metrics such as Means and Medians, Standard deviation, Diameter, Expression Ratio etc in the region of every gene spot on the microarray image. The expression-ratio measures the transcription abundance between the two sample gene sets. The positive or negative expression ratio indicates the over-expression or under-expression between the control and treatment genes.

Figure 1 show the overall process involved in microarray image analysis.

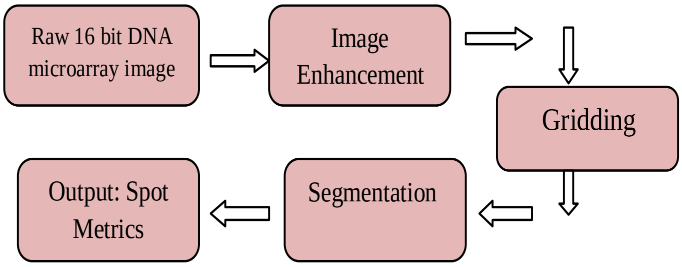


Figure 1: Microarray Image Analysis

2. NOISE REMOVAL USING BEMD PLUS WAVELETS

If the microarray image contains noise, the quality of the edges extracted from the image will be poor. This edge information is primary source for automatic gridding of microarray image. Empirical mode decomposition [5,6] is a signal processing method that nondestructively fragments any non-linear and non-stationary signal into oscillatory functions by means of a mechanism called shifting process. These oscillatory functions are called Intrinsic Mode Functions (IMF), and each IMF satisfies two properties, (a) the number of zero crossings and extrema points should be equal or differ by one. (b) Symmetric envelopes (zero mean) interpret by local maxima and minima [7]. The signal after decomposition using EMD is non-destructive means that the original signal can be obtained by adding the IMFs and residue. The first

IMF is a high frequency component and the subsequent IMFs contain from next high frequency to the low frequency components. The shifting process used to obtain IMFs on a 2-D signal (image) is summarized as follows:

a) Let I(x,y) be a Microarray image used for EMD decomposition. Find all local maxima and local minima points in I(x,y).

b) Upper envelope Up(x,y) is created by interpolating the maxima points and lower envelope Lw(x,y) is created by interpolating minima points. This interpolation is carried out using following redial basis functions shown in table 1.

c) Compute the mean of lower and upper envelopes denoted by Mean(x,y).

$$Mean(x,y) = \frac{(Up(x,y) + Lw(x,y))}{2} \qquad (1)$$

d) This mean signal is subtracted from the input signal.

$$Sub(x,y) = I(x,y) - Mean(x,y) \qquad (2)$$

e) If Sub(x,y) satisfies the IMF properties, then an IMF is obtained .

$$IMF_i(x,y) = Sub(x,y) \qquad (3)$$

f) Subtract the extracted IMF from the input signal. Now the value of I(x,y) is

$$I(x,y) = I(x,y) - IMF_i(x,y) \qquad (4)$$

Repeat the above steps (b) to (f) for the generation of next IMFs.

g) This process is repeated until I(x,y) does not have maxima or minima points to create envelopes.

Original Image can be reconstructed by inverse EMD given by

$$I(x,y) = \sum_{i=1}^{n} IMF_i(x,y) + res(x,y) \qquad (5)$$

The mechanism of de-noising using BEMD-DWT is summarized as follows

a) Apply 2-D EMD for noisy microarray to obtain IMFi (i=1, 2, …k). The kth IMF is called residue.

b) The first intrinsic mode function (IMF1) contains high frequency components and it is suitable for denoising. This IMF1 is denoised with mean filter. This de-noised IMF1 is represented with DNIMF1.

c) The denoised image is reconstructed by the summation of DNIMF1 and remaining IMFs given by

$$RI = DNIMF1 + \sum_{i=2}^{k} IMF_i \qquad (6)$$

Where RI is the reconstructed band. The flow diagram of BEMD-DWT filtering is shown in figure 2.

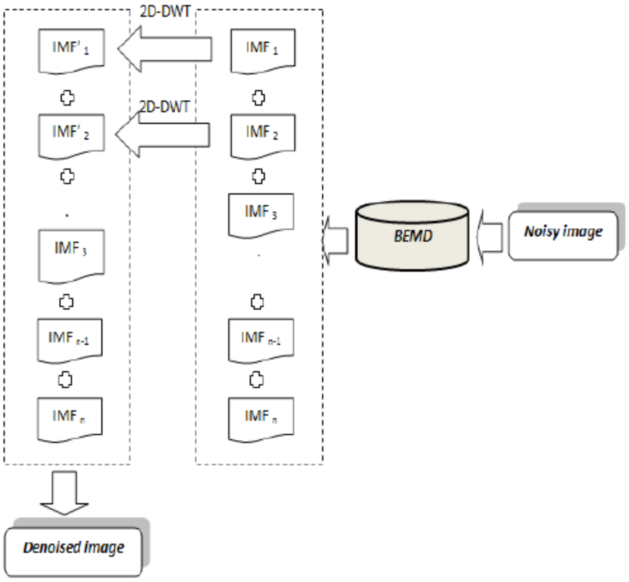


Figure 2: Image Denoising using BEMD + WAVELET

Table 1: Different Interpolation Functions

Function	**Description**
$\varphi(x) = x$	Linear function
$\varphi(x) = \sqrt{x^2 + 1}$	Multi Quadrique Function
$\varphi(x) = e^{-x}$	Exponential Function
$\varphi(x) = x\log(x)$	Logarithmic Function
$\varphi(x) = x^2\log(x)$	Spline Function

Gridding is the process of dividing the microarray image into blocks (sub-gridding) and each block again divided into sub-blocks (spot-detection). The final sub-block contains a single spot and having only two regions spot and background. Existing algorithms for gridding are semi-automatic in nature requiring several parameters such as size of spot, number of rows of spots, number of columns of spot etc. In this paper, a fully automatic gridding algorithm designed in [8, 9] is used for sub-gridding and spot-detection.

3. SEGMENTATION

Many microarray image segmentation approaches have been proposed in literature. Fixed circle segmentation [10], Adaptive circle Segmentation Technique [11], Seeded region growing methods [12] and clustering algorithms [13] are the methods that deal with microarray image segmentation problem. This paper mainly focuses on clustering algorithms.

These algorithms have the advantages that they are not restricted to a particular spot size and shape, does not require an initial state of pixels and no need of post processing. These algorithms have been developed based on the information about the intensities of the pixels only. In this paper, adaptive data clustering algorithms are proposed, in which for the selection of K value, creatively put forward the number of connected domain images meet requirements comparing with iterative variables, and when the two are equal, the value of K is the value of iterative variable. The improvement of the above part can greatly improve the accuracy of image segmentation and also optimize the optimization of algorithm structure to a certain extent.

4. DATA CLUSTERING ALGORITHMS

K-MEANS ALGORITHM:

1. Randomly consider K initial clusters {C_1, C_2,……,C_k} from the m*n image pixels {I_1, I_2, I_3,……,I_{m*n}}.
2. Assign each pixel to the cluster C_j {j=1,2,…..K} if it satisfies the following condition

$$D(I_i, C_j) < D(I_i, C_q), q = 1, 2, ..., K \quad j \neq q \tag{7}$$

Where D(. , .) denotes the dissimilarity measure.

3. Find new cluster centroid as follows

$$C_i^{\wedge} = \frac{1}{n_i} \sum_{I_j \in C_i} I_j, i = 1, 2, ...K \tag{8}$$

Where n_i is the number of pixels belonging to cluster C_i.

4. If

$$C_i^{\wedge} = C_i, i = 1, 2, ..K \tag{9}$$

Then stop.

Else continue from step 2 [14].

K-MEDOIDS ALGORITHM

The k-medoids clustering algorithm [15] for segmentation of microarray image is described as follows:

1. Randomly consider K initial medoids {M_1, M_2,……,M_k} for the clusters {C_1, C_2,……,C_k} from the m*n image pixels {I_1, I_2, I_3,……,I_{m*n}}.

   A cluster medoid is a point that is located centrally with in the cluster. It is the point that has minimum sum of distances to other points of the cluster.

2. Pixel assignment to clusters is done using the condition given by

$$D(I_i, M_j) < D(I_i, M_q), q = 1, 2, ..., K \quad j \neq q \tag{10}$$

Where D(. , .) denotes the dissimilarity measure.

3. Find new medoids $M_i^{\wedge}$ belonging to clusters C_i , $i = 1, 2, ...K$. It is the pixel value with minimum total dissimilarity to all other points.
4. If

$$M_i^{\wedge} = M_i, i = 1, 2, ..K \tag{11}$$

Then stop.

Else continue from step 2.

FUZZY C-MEANS ALGORITHM

1. Randomly consider K initial clusters {C_1, C_2,……,C_k} from the m*n image pixels {I_1, I_2, I_3,……,I_{m*n}}.
2. The membership matrix u_{ij} is initialized with value from 0 to 1 and value of m=2. The summation of pixel memberships representing particular cluster should be equal to 1.
3. Pixel assignment to clusters is done using the condition given by

$$u_{ij}^m D(I_i, C_j) < u_{iq}^m D(I_i, C_q), q = 1, 2, ..., K \quad j \neq q \tag{12}$$

Where D(. , .) denotes the dissimilarity measure.

4. Find new membership and cluster centroid values as follows

$$u_{ik} = \frac{1}{\sum_{j=1}^{K} \left(\frac{D(C_i, I_k)}{D(C_j, I_k)}\right)^{\frac{1}{m-1}}}, for 1 \leq i \leq K$$ u_{ik} denotes the k^th^ object in the i^th^ cluster.

$$C_i^{\wedge} = \frac{\sum_{j=1}^{n} u_{ij}^m I_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{13}$$

Where n is the number of pixels belonging to cluster C_i.

5. Continue 2-3 until each object is assigned to the cluster that has maximum membership [16].

5. ADAPTIVE DATA CLUSTERING ALGORITHM

The basic idea of any clustering algorithm is to cluster the objects closest to them by clustering the K points in the space. Iteratively, the values of centroid of clusters are updated one by one until the best clustering results are obtained. Determining the correct K value is the key to the success of the any clustering algorithm. In this paper we have implemented Adaptive k-means clustering algorithm for estimation of K-value, the same procedure can be used for remaining algorithms presented in this paper.

The K-means algorithm takes Euclidean distance as the similarity measure, which is to find the optimal classification of an initial cluster center vector, so that the evaluation index is minimum. The error square sum criterion function is used as a clustering criterion function. Although the algorithm of K-means is efficient, value of K should be given in advance, and the selection of K value is very difficult to estimate. In the proposed method, we start with the selection of K = 2, that is, image segmentation starts from two clusters, and then the image is segmented. Finally, we determine the number of segmentation results based on the maximum connected domain algorithm [17]. If the image number of the final segmentation result matches the K value, the K value is selected correctly. If the K value does not match, the K value at the beginning will be increased until the above two values match. This procedure for selection K-value can be used for K-medoids, K-modes, Fuzzy c-means , Fuzzy K-medoids clustering algorithms leading to adaptive clustering algorithms.

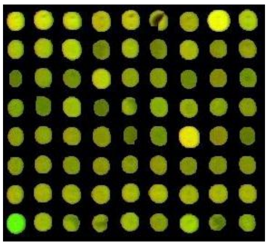
The adaptive k-means clustering algorithm :

```
For K=2 to 10
{
    Randomly consider K initial clusters {C1, C2,……,Ck} from the m*n image pixels {I1, I2, I3,……,Im*n}.
1.  Assign each pixel to the cluster Cj {j=1,2,…..K} if it satisfies the following condition
        $D(I_i,C_j) < D(I_i,C_q), q = 1,2,...,K$
        $j \neq q$                                                   (14)
     Where D(. , .) denotes the dissimilarity measure.
2.  Find new cluster centroid as follows
        $C_i^{\wedge} = \frac{1}{n_i} \sum_{I_j \in C_i} I_j, i = 1,2,...K$     (15)
     Where ni is the number of pixels belonging to cluster Ci.
3.  If
        $C_i^{\wedge} = C_i, i = 1,2,..K$                           (16)
    Then stop.
    Else continue from step 2.
    Compare the maximum connected domain results
    If equal to K print segmented result and break;
    else continue with incremented value of K;
}
```

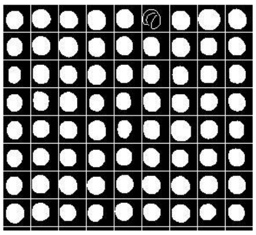
6. EXPERIMENTAL RESULTS

Qualitative Analysis: The proposed clustering algorithm is performed on two microarray images drawn from the standard microarray database corresponds to breast category aCGH tumor tissue. MImage 1 consists of a total of 38808 pixels and MImage 2 consists of 64880 pixels. The noise removal in microarray image is done using BEMD + Wavelets method. In BEMD decomposition process, the envelopes are created using different interpolation methods. Out of these spline interpolation gives better envelopes based on number of maxima and minima values. Table 2 shows PSNR values of different nonlinear filtering algorithms for different values of sigma (Gaussian Noise) added to the image in-order to find the performance of the proposed noise removal algorithm [Done on MImage 1]. Gridding is performed on the input images by the method proposed in [18], to segment the image into compartments, where each compartment is having only one spot region and background. The gridding output is shown in figure 3. After gridding the image into compartments, such that each compartment is having single spot and background, compartment no 1 from MImage 1 and compartment no 8 from MImage 2 are extracted. The image compartments are segmented using Adaptive K-means clustering algorithm. The segmentation result is shown in figure 3. Table 3 shows the quantitative evaluations of clustering algorithms using MSE [19, 20]. The results confirm that Adaptive fuzzy c-means algorithm produces the lowest MSE value for segmenting the microarray image.

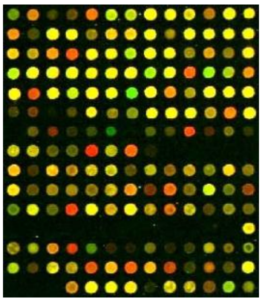
MImage 1



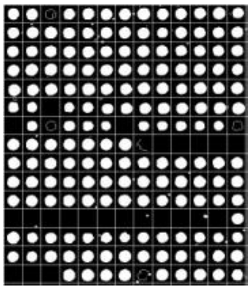
Gridded Image



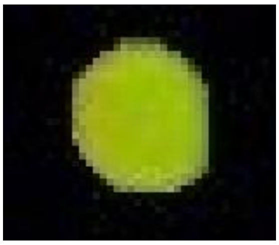
MImage 2



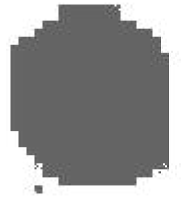
Gridded Image

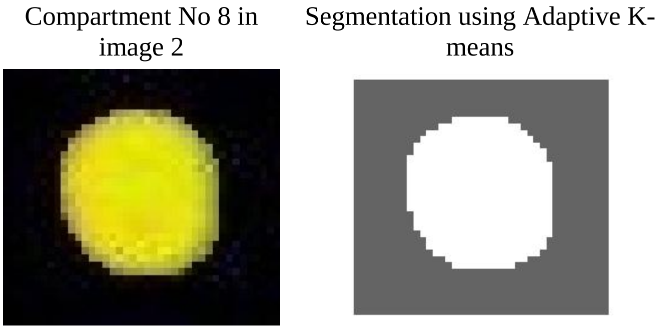


Compartment No 1 in image 1



Segmentation using multiple feature fcm





Compartment No 8 in image 2

Segmentation using Adaptive K-means

Figure 3: Gridding and segmentation results

Table 2: PSNR values of Noise Removal

Noise level (Sigma)	Median	Wavelet	BEMD + Wavelet
5	34.2	35.7	37.9
10	29.5	32.5	34.8
15	25.4	27.3	29.7
20	21.1	23.2	26.2

Table 3: MSE values

	Normal Clustering		Adaptive Clustering	
Method	Compartment No 1	Compartment No 8	Compartment No 1	Compartment No 8
K-means	96.2	95.8	95.9	94.5
K-medians	95.1	94.2	94.7	93.2
K-mode	95.2	94.3	94.8	93.6
Fuzzy c-means	90.7	90.2	89.01	89.6
Fuzzy K-medoids	91.8	91.4	90.2	90.1

7. CONCLUSIONS

Microarray technology provides simultaneous monitoring of thousands of gene expression levels. The main steps in microarray image analysis are gridding, segmentation and information extraction. This paper presents a noise removal methodology using BEMD + Wavelets with different interpolation methods used in BEMD decomposition process. Adaptive clustering algorithms are used for segmentation of microarray image. These algorithms are extensions to the k-means clustering algorithm. These algorithms have been developed based on the information about the intensities of the pixels only. The main requirements for any clustering algorithm is the number of clusters K. Estimating the value of K is difficult task for given data. This paper presents adaptive data clustering algorithms which generates accurate segmentation results with simple operation and avoids the interactive input K (number of clusters) value for segmentation of microarray image. The qualitative and quantitative results shows that adaptive data clustering algorithms are more efficient than normal data clustering algorithms in segmenting the spot area, thus producing more accurate expression-ratio. Out of these algorithms presented in this paper, Adaptive fuzzy c-means clustering algorithm produces better segmentation result. Log ratio of R/G gives the abundance of expression level of the corresponding gene.

REFERENCES

[1] M.Schena, D.Shalon, Ronald W.davis and Patrick O.Brown, “Quantitative Monitoring of gene expression patterns with a complementary DNA microarray”, Science, 270,199, pp:467-470.

[2] Wei-Bang Chen, Chengcui Zhang and WenLin Liu, “An Automated Gridding and Segmentation method for cDNA Microarray Image Analysis”, 19th IEEE Symposium on Computer-Based Medical Systems.

[3] Tsung-Han Tsai Chein-Po Yang, WeiChiTsai, Pin-Hua Chen, “Error Reduction on Automatic Segmentation in Microarray Image”, IEEE 2007.

[4] J.Harikiran, et.al. “Vector Filtering Techniques for Impulse Noise Reduction with Application to Microarray images”, International Journal of Applied Engineering Research”, volume 10, Number 3, pp. 7181-7193, 2015.

[5] J.Harikiran, A.Raghu, Dr.P.V.Lakshmi, Dr.R.Kiran Kumar, “Edge Detection using Mathematical Morphology for Gridding of Microarray Image”, International Journal of Advanced Research in Computer Science, Volume 3, No 2, pp.172-176, April 2012.

[6] B.Sivalakshmi, N.Nagamalleswara rao,”Microarray Image Analysis Using Genetic Algorithm”, IAES Indonesian Journal of Electrical Engineering and Computer Science, Volume 4, No. 3, pp.561-567, 2016.

[7] J.Harikiran et.al. “Fuzzy C-means with Bi-dimensional empirical Mode decomposition for segmentation of Microarray Image”, International Journal of Computer Science Issues, volume 9, Issue 5, Number 3, pp.273-279, 2012

[8] J.Harikiran, Dr.P.V.Lakshmi, Dr.R.Kirankumar, “Automatic Gridding Method for Microarray Images”, Journal of Theoretical and Applied Information Technology, Volume 65, Number 1, pp.235-241, 2014.

[9] J.Harikiran, D.Ramakrishna, B.Avinash, Dr.P.V.Lakshmi, Dr.R.Kiran Kumar, “A New Method of Gridding for Spot Detection in Microarray Images”, Computer Engineering and Intelligent Systems, Volume 5, No 3, pp.25-33, 2014.

[10] M.Eisen, ScanAlyze User’s manual, 1999,

[11] J.Buhler, T.Ideker and D.Haynor, “Dapple:Improved Techniques for Finding spots on DMA Microarray Images”, Tech. Rep. UWTR 2000-08-05, University of Washington, 2000.

[12] R.Adams and L.Bischof, "Seeded Region Growing", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 16,no. 6, pp.641-647, 1994.

[13] D.RamaKrishna, J.Harikiran et.al." "Various Versions of K-means Clustering Algorithm for Segmentation of Microarray Image", International Journal of Electronics Communication and Computer Engineering, Volume 4, Issue 1, pp.1554-1558, 2012.

[14] J.Harikiran, Dr.P.V.Lakshmi, Dr.R.Kiran Kumar, "Fast Clustering Algorithms for Segmentation of Microarray Images", International Journal of Scientific & Engineering Research, Volume 5, Issue 10, pp 569-574, 2014.

[15] J.Harikiran, P.V.Lakshmi," Extensions to the K-means Algorithm for Segmentation of cDNA Microarray Images", CSI Communications, December 2015.

[16] J.Harikiran, Dr.P.V.Lakshmi, Dr.R.Kiran Kumar, "Multiple Feature Fuzzy C-means Clustering Algorithm for Segmentation of Microarray image", IAES International Journal of Electrical and Computer Engineering", Vol. 5, No. 5, pp. 1045-1053, October 2015.

[17] W Zuo, Research on connected region extraction algorithms [J]. Comp. Appl. Softw. 23(1), 97–98 (2006).

[18] Dr.R.Kiran Kumar, B.Saichandana et.al. "Dimensionality Reduction and Classification of Hyperspectral Images using Genetic Algorithm", IAES Indonesian Journal of Electrical Engineering and Computer Science, Volume 3, No 3, pp.503-511, September 2016.

[19] B.Saichandana et.al. "Image Fusion in Hyperspectral Image Classification using Genetic Algorithm", IAES Indonesian Journal of Electrical Engineering and Computer Science, Volume 2, No. 3, June 2016, pp.703-711.

[20] B.Saichandana, et.al, "Clustering Algorithm combined with Hill climbing for Classification of Remote Sensing Image", IAES International Journal of Electrical and Computer Engineering, volume 4, No 6, December 2014, pp.923-930.