Microarray Image Analysis using k-means Clustering Algorithm

B.SivaLakshmi[1], N.Naga Malleswara Rao[2]
[1]*Phd Research Scholar, Acharya Nagarjuna University, Vijayawada, AP, India.*
[2]*Professor, RVR and JC College of Engineering, Guntur, AP, India*
Email: bolem.sivalakshmi33@gmail.com

Abstract:-By using Microarray Technology, in a single experiment one can study the function of thousands of genes in parallel. 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. Generally, clustering algorithms are used for segmentation of microarray image. 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. Clustering algorithm such as K-means, Moving 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 k-means clustering algorithm is more efficient than normal k-means clustering algorithm in segmenting the spot area, thus producing more accurate expression-ratio.

Keywords: Image processing, Microarray Image Analysis, Clustering algorithms

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.

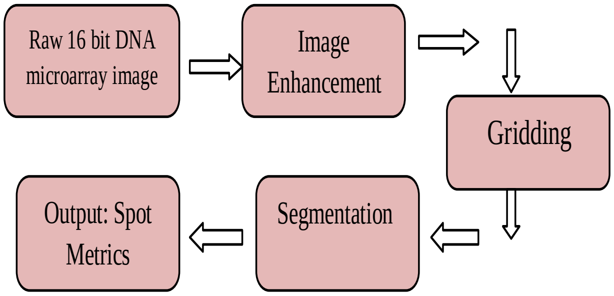


Fig 1: Microarray Image Analysis

2. ENHANCEMENT AND GRIDDING

If the contrast of the microarray image is low, the quality of the edges extracted from the image will be poor. This edge information is primary source for automatic gridding of microarray image [5]. The quality of the spot edges can be improved by applying GA based contrast enhancement algorithm [6] to the original image prior to the computation of Gridding. 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 [7, 8] is used for sub-gridding and spot-detection.

3. SEGMENTATION

Many microarray image segmentation approaches have been proposed in literature. Fixed circle segmentation [9], Adaptive circle Segmentation Technique [10], Seeded region growing methods [11] and clustering algorithms [12] 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 [13, 15]. These algorithms have been developed based on the information about the intensities of the pixels only [14]. 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 values 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.

3.1 K-MEANS CLUSTERING ALGORITHM

The K-means clustering algorithm is a partitioning method which assign pixels to a user-defined mutually exclusive number of clusters (k) in such a way that maximizes the separation of those clusters while minimizing intra-cluster distances relative to the cluster's mean or centroid and returns the index value of cluster, for which the pixel is assigned. The k-means clustering is always better than hierarchical clustering for large amount of data. The algorithm mainly dependent on three parameters 1) numbers of clusters k 2) Selection of initial values of centroids 3) the distance metric (optimization function) used in assignment of pixels to different clusters. This paper presents the implementation of k-means on microarray image with different parameters available in k-means function in MATLAB and shows how best the clusters are obtained using silhouette index.

The k-means clustering is given as follows:

1. Initialization of k centroid values. This initialization in MATALB by using name – ‘Start’ and values ‘plus’, ‘cluster’, ‘sample’, ‘uniform’ and ‘numeric’ in the k-means function. The default value is ‘plus’, which selects k pixel values from image randomly as centroid values.
2. Assignment of pixels to the corresponding cluster based on the minimum value of distance measure from corresponding pixel to centroid.
   The k-means function in MATLAB has following distance measures with name – ‘Distance’ and values – ‘sqeuclidean’, ‘cityblock’, ‘cosine’, ‘correlation’ and ‘hamming’.
3. Up-dation of centroid values by taking mean of pixels belonging to cluster generating new centroids.
4. This process is repeated until there is no change in old and new centroid values or by specifying maximum number of iterations in kmeans function using name –‘MaxIter’ and value – number of iterations (numeric value).

The k-means function in MATLAB returns cluster indices, centroid locations, distances from pixel to centroid and sum of distances from pixel to centroid within cluster. To know the information how given image is segmented into clusters, draw silhouette plot using the cluster indices returned from k-means function. The silhouette plot displays a measure of how close each pixel in one cluster is to points in the neighboring clusters. The silhouette ranges from -1 to +1. +1 indicates pixels are assigned perfectly to one cluster and -1 indicates wrong assignment of pixels to cluster. This measure is calculated using silhouette function in MATLAB with three parameters, input image, cluster indices returned from k-means function and distance measure.

3.2 ADAPTIVE K-MEANS 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 [16, 18]. 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.

The adaptive k-means clustering algorithm is presented below

```
For K=2 to 10
{
    Randomly consider K initial clusters {$C_1$, $C_2$,……,$C_k$} from the m*n image pixels {$I_1$, $I_2$, $I_3$,……,$I_{m*n}$}.
    1. Assign each pixel to the cluster $C_j$ {j=1,2,…..K} if it satisfies the following condition
           $D(I_i, C_j) < D(I_i, C_q), q = 1, 2, ..., K$
           $j \neq q$                                        (1)
       Where D(. , .) denotes the dissimilarity measure.
    2. Find new cluster centroid as follows
           $C_i^{\wedge} = \frac{1}{n_i} \sum_{I_j \in C_i} I_j, i = 1, 2, ...K$      (2)
       Where $n_i$ is the number of pixels belonging to cluster $C_i$.
    3. If
           $C_i^{\wedge} = C_i, i = 1, 2, ..K$              (3)
       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;
}
```

4. EXPERIMENTAL RESULTS

Qualitative Analysis: The proposed clustering algorithm is performed on a microarray images drawn from the standard microarray database corresponds to breast category aCGH tumor tissue. The Image consists of a total of 38808 pixels. Gridding is performed on the input image by the method proposed in [13], to segment the image into compartments, where each compartment is having only one spot region and background. The gridding output is shown in figure 2. After gridding the image into compartments, such that each compartment is having single spot and background, compartment no 8 from image is extracted. First the image is segmented using K-means clustering algorithm in MATLAB. The silhouette plots for the k-means function with different distance measures are shown in figure 2 with different parameters in k-means function. The efficient way to compare the performance of k-means clustering with different functions is done by calculating the mean of silhouette values. The larger mean value means better segmentation. The mean of silhouette values for the above functions executed on single spot image shown in figure 3 is shown in table 1. Then the same image compartments are segmented using Adaptive K-means clustering algorithm. The gridding and segmentation result is shown in figure 3. Table 2 shows the quantitative evaluations of k-means clustering algorithms using MSE [17, 19]. The results confirm that Adaptive k-means clustering algorithm produces the lowest MSE value for segmenting the microarray image.

5. 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. Adaptive clustering algorithms are used for segmentation of microarray image. 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 k-means clustering algorithm is more efficient than normal k-means clustering algorithm in segmenting the spot area, thus producing more accurate expression-ratio. 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, 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.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] M.Anirban, et.al. “Multiobjective Genetic Algorithm based Fuzzy Clustering of Categorical Attributes”, IEEE transactions on Evolutionary Computing, volume 13, number 5, pp.991-1005.

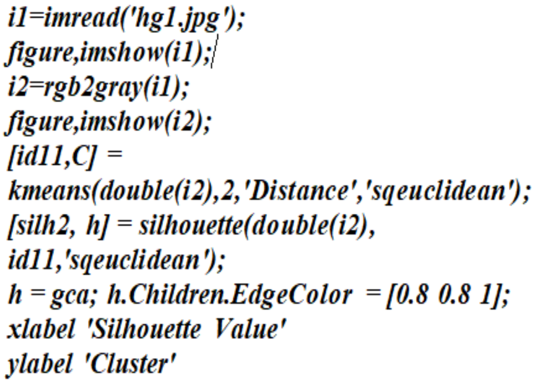
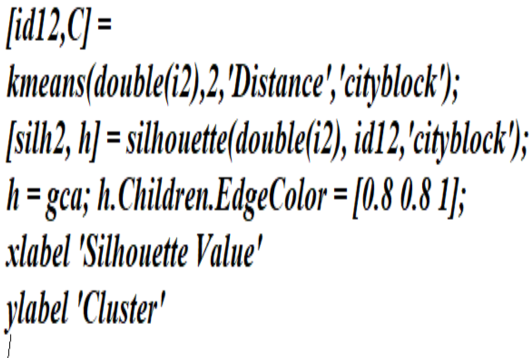
[15] J.Harikiran, P.V.Lakshmi,” Extensions to the K-means Algorithm for Segmentation of cDNA Microarray Images”, CSI Communications, December 2015

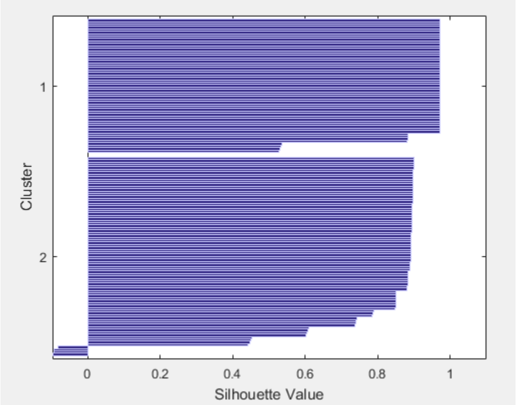
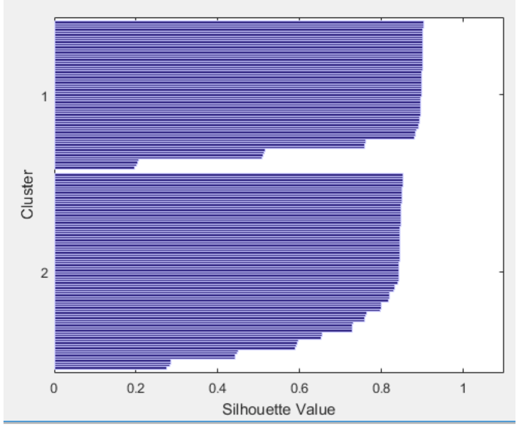
[16] Dr.R.Kiran Kumar, B.Saichandana et.al. “Dimensionality Reduction and Classification of Hyperspectral Images using Genetic Algorithm”, IAES Indonesian Journal of Electrical Engineering and Computer Science, Volume 3, No 3, pp.503-511, September 2016.

[17] 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.

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

[19] 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.


i1=imread('hg1.jpg');
figure,imshow(i1);
i2=rgb2gray(i1);
figure,imshow(i2);
[id11,C] =
kmeans(double(i2),2,'Distance','sqeuclidean');
[silh2, h] = silhouette(double(i2),
id11,'sqeuclidean');
h = gca; h.Children.EdgeColor = [0.8 0.8 1];
xlabel 'Silhouette Value'
ylabel 'Cluster'
	
	[id12,C] =
	kmeans(double(i2),2,'Distance','cityblock');
	[silh2, h] = silhouette(double(i2), id12,'cityblock');
	h = gca; h.Children.EdgeColor = [0.8 0.8 1];
	xlabel 'Silhouette Value'
	ylabel 'Cluster'

[id13,C] = kmeans(double(i2),2,'Start', 'sample', 'Distance','cosine');
[silh2, h] = silhouette(double(i2), id13,'cosine');
h = gca; h.Children.EdgeColor = [0.8 0.8 1];
xlabel 'Silhouette Value'
ylabel 'Cluster'
	[id14,C] = kmeans(im2bw(i2),2,'Distance','hamming', 'MaxIter', 80);
	[silh2, h] = silhouette(double(i2), id14,'hamming');
	h = gca; h.Children.EdgeColor = [0.8 0.8 1];
	xlabel 'Silhouette Value'
	ylabel 'Cluster'
	cluster(i)=mean(silh2);

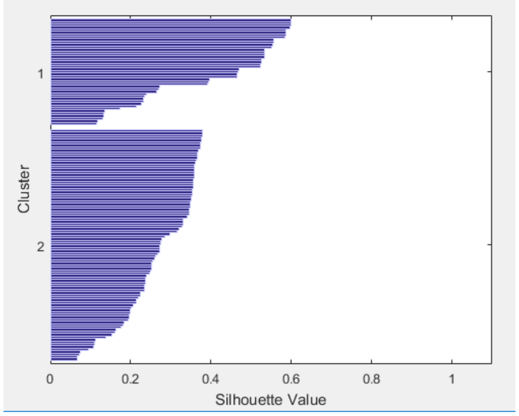
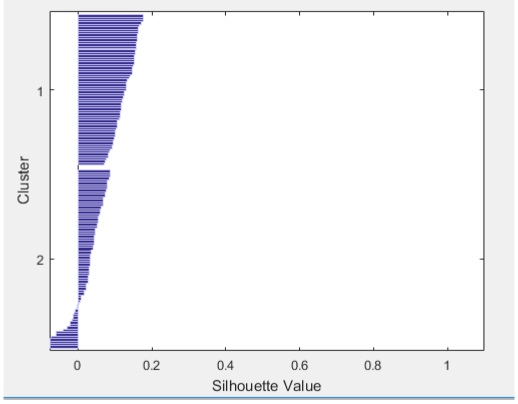
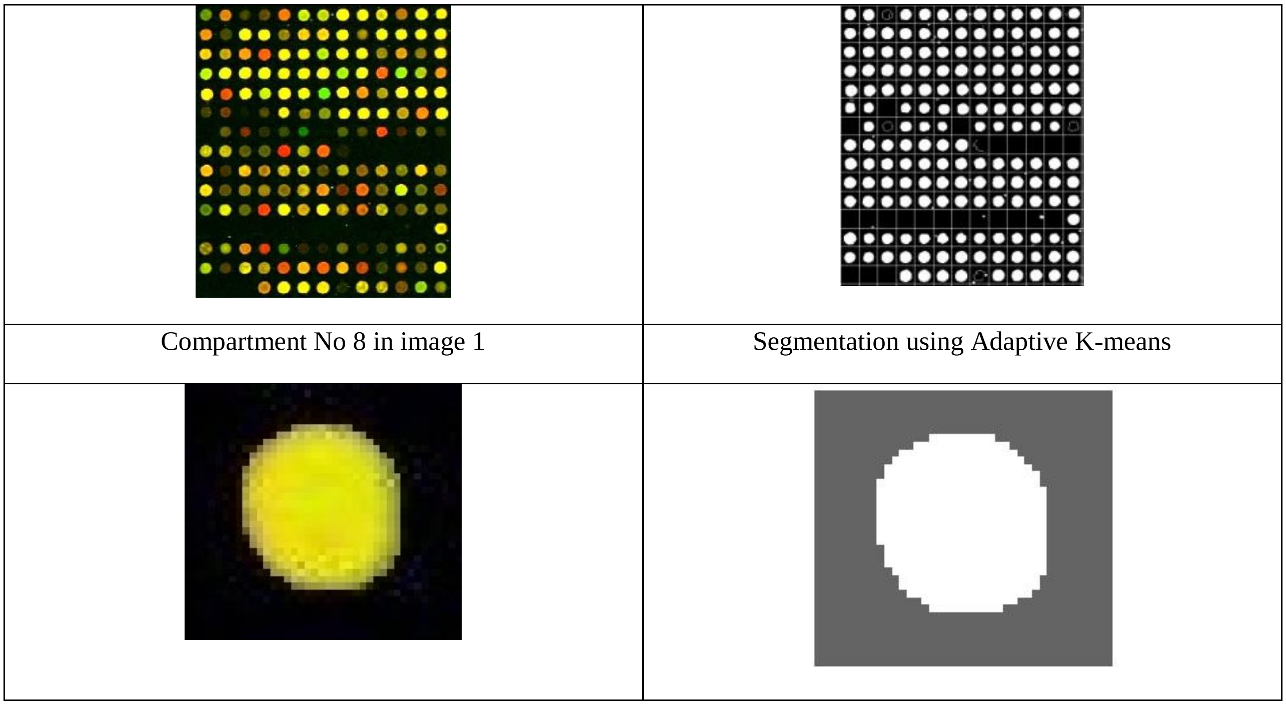
	

Fig 2: Silhouette Index

Table 1: Silhouette values with different distance measures

Distance measure	Silhouette value (mean)
Sqeuclidean	0.8534
City Block	0.7847
Cosine	0.4224
Hamming	0.0735

Image 1	Gridded Image



Compartment No 8 in image 1

Segmentation using Adaptive K-means

Figure 3: Gridding and segmentation results

Table 2: MSE values

	Normal Clustering	Adaptive Clustering
Method	Compartment No 8	Compartment No 8
K-means	95.8	94.5