International Conference on Innovative Applications in Engineering and Information Technology(ICIAEIT-2017)

International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) *Volume.3,Special Issue.1,March.2017*

Multiple Feature Clustering Algorithm for Segmentation of cDNAMicroarray Image

B. Sridevi[1] ***K.Srinivasa Rao*** [2] ***V. Gayathri***[3] ***Dr. G. Lavanya Devi***[4]

[1]***M.Tech Student, Dept. of Computer science and system engineering,A.UCollege of Engineering (A), Visakhapatnam.*** bsridevi205@gmail.com

[2]***Research Scholar, Dept. of Computer science and system engineering, A.UCollege of Engineering (A),Visakhapatnam.*** sri.kurapati@gmail.com

[3]***Research Scholar, Dept. of Computer science and system engineering, A.UCollege of Engineering (A), Visakhapatnam.*** vallepu.gayathr@gmail.com

[4]***Assistant Professor Dept. of Computer science and system engineering, A.UCollege of Engineering (A), Visakhapatnam.***lavanyadevig@yahoo.co.in

Abstract: A Deoxyribonucleic Acid (DNA) microarray is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip forming an array. The analysis of DNA microarray images allows the identification of gene expressions to draw biological conclusions for applications ranging from genetic profiling to diagnosis of cancer. The output of image analysis is a matrix consisting of a measure of intensity of each spot in the image. This measure denotes gene expression ratio (transcription abundance) between the test and control samples for the corresponding gene. The positive expression indicates the over-expression, while negative expression indicates under-expression between the control and treatment genes. The DNA microarray image analysis includes three tasks: gridding, segmentation and intensity extraction. The segmentation step of microarray image analysis has been implemented in this paper. In this paper, multiple feature FCMclustering algorithm for microarray image segmentation that separate the spots from the background is proposed. The experimental results show that multiple feature Fuzzy C-means have segmented the spots of the microarray image more accurately than K-means, Moving K-means and Fuzzy C-means.

I.INTRODUCTION

Microarrays, widely recognized as the next revolution in molecular biology, enable scientists to analyze genes, proteins and other biological molecules on a genomic scale [1]. A microarray is a collection of spots containing DNA deposited on the solid surface of glass slide. Each of the spot contains multiple copies of single DNA sequence [2].

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

1. 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 fie., 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.
2. Gridding 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.
3. Segmentation 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.

1. 4. Information Extraction 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 treatmentgenes.Fig 1 shows the overall process involved in microarray image analysis.

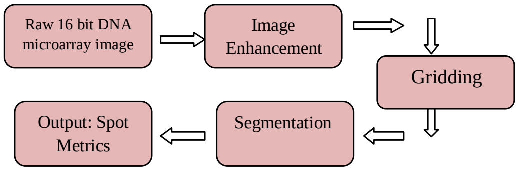


Fig 1: Microarray Image Analysis

In digital image segmentation applications, clustering technique is used to segment regions of interest and to detect borders of objects in an image. Clustering algorithms are based on the similarity or dissimilarity index between pairs of pixels. It is an iterative process which is terminated when all clusters contain similar data. In order to segment the image, the location of each spot must be identified through gridding process. Hirata [5] presented an automatic gridding method by using the horizontal and vertical profile signal of the image to perform the image gridding. The algorithm can satisfy the requirements of microarray image segmentation.

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 (one feature). But in the microarray image segmentation problem, not only the pixel intensity, but also the distance of pixel from the center of the spot and median of intensity of a certain number of surrounding pixels influences the result of clustering. In this paper, multiple feature fuzzy c-means clustering algorithm is proposed, which utilizes more than one feature. The qualitative and quantitative results show that multiple feature fuzzy C-means clustering algorithm has segmented the image better than other clustering algorithms. The paper is organized as follows: Section 2 presents the K-means clustering algorithm, Section 3 presents Moving K-means clustering algorithm, Section 4 presents Fuzzy K-means clustering algorithm, Section 5 presents multiple feature clustering, Section 6 presents Experimental results and finally Section 7 report conclusions.

II.K-MEANS CLUSTERING ALGORITHM

The K-means clustering algorithm for segmenting the microarray image [6] is summarized as follows:

Step 1: cluster centroids are initialized.

Step 2: compute the closest cluster for each pixel and classify it to that cluster, ie: the objective is to minimize the sum of squares of the distances given by the following:

$$\Delta_{ij} = \| x_i - c_j \|. \quad \arg\min \sum_{i=1}^{N} \sum_{j=1}^{C} \Delta_{ij}^{2} \qquad (1)$$

Step 3: Compute new centroids after all the pixels are clustered. The new centroids of a cluster is calculated by the following

$$c_j = \frac{1}{Nj} \sum x_i \text{where } x_i \text{ belongs to } c_j. \qquad (2)$$

Step 4: Repeat steps 2-3 till the sum of squares given in equation is minimized.

The K-means clustering algorithm has many weaknesses which are as follows:

1. The number of clusters K, must be determined before the algorithm is executed.
2. The algorithm is sensitive to initial conditions. It produces different results for different initial conditions.
3. The K-means algorithm may be trapped in the local optimum. As a result, the trapped clusters would represent wrong group of data.
4. Data which are far away from the centers may pull the centers away from the optimum location, leading to poor representation of data.

To avoid this problem, the maximal and minimal observed values in the target area for the intensities are used instead of random starting points for the two clusters in the segmentation of cDNA microarray images. This provides a meaningful representation of foreground and background and assures the convergence to an adequate optimum.

III.MOVING K-MEANS CLUSTERING ALGORITHM

The Moving K-means clustering algorithm is the modified version of K-means proposed in [7]. It introduces the concept of fitness to ensure that each cluster should have a significant number of members and final fitness values before the newposition of cluster is calculated. The Moving K-means clustering algorithm for segmenting the microarray image is summarized as follows:

Step 1: cluster centroids are initialized (minimal and maximal values of pixels in the target area)

Step 2: compute the closest cluster for each pixel and classify it to that cluster, ie: the objective is to minimize the sum of squares of the distances given by the following:

$$\Delta_{ij} = \| x_i\text{-}c_j\|. \quad \arg\min \sum_{i=1}^{N} \sum_{j=1}^{C} \Delta_{ij}^{2} \qquad (3)$$

Step 3: The fitness for each cluster is calculated using

$$f\,(c_k) = \sum_{t \in c_k} (\,\| x_t\text{-}c_k\|\,)^2 \qquad (4)$$

All centers must satisfy the following condition:

$$f\,(c_s) \geq \alpha_a\, f(c_l) \qquad (5)$$

where$\alpha_a$is small constant value initially with value in range $0<\alpha_a<1/3$,c_s and c_l are the centers that have the smallest and the largest fitness values. If (5) is not fulfilled, the members of c_l are assigned as members of c_s, while the rest are maintained as the members of c_l. The positions of $c_s$andc_l are recalculated according to:

$$C_s = 1/n_{cs}\Big(\sum_{t \in c_s} x_t\Big) \qquad (6)$$

$$C_l = 1/n_{cl}\,\Big(\sum_{t \in c_l} x_t\Big) \qquad (7)$$

The value of α_a is then updated according to:

$$\alpha_a = \alpha_a\text{-}\ \alpha_a/n_c \qquad (8)$$

The above process are repeated until (5) is fulfilled. Nextall data are reassigned to their nearst center and the new center positions are recalculated using (2).

Step 4: The iteration process is repeated until the following condition is satisfied.

$$f\,(c_s) \geq \alpha_a\, f(c_l) \qquad (9)$$

The Moving K-means algorithm has the following drawbacks:

1. The Moving K-means algorithm is sensitive to noise.
2. For some cases of Moving k-means, the clusters or centers are not located in the middle or centroid of a group of data, leading to imprecise results.
3. The fitness concept in the Moving k-means algorithm lead to a problem where some members of centers with the largest fitness are enforced to be assigned as a members of a center with the smallest fitness.

IV.FUZZY C-MEANS CLUSTERING ALGORITHM

The Fuzzy C-means algorithm [8] is summarized as follows:

Step_1: Initialize the membership matrix $u_{ij}$is a value in (0,1) and the fuzziness parameter m (m=2). The sum of all membership values of a pixel belonging to clusters should satisfy the constraint expressed in the following.

$$\sum_{j=1}^{c} u_{ij}=1 \qquad (10)$$

for all i= 1,2,…….N, where c (=2) is the number of clusters and N is the number of pixels in microarray image.

Step_2: Compute the centroid values for each cluster c_j. Each pixel should have a degree of membership to those designated clusters. So the goal is to find the membership values of pixels belonging to each cluster. The algorithm is an iterative optimization that minimizes the cost function defined as follows:

$$F= \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij}^{\,m}\ \| x_j\text{-}c_i\|^2 \qquad (11)$$

Whereu_{ij}represents the membership of pixel $x_j$in the ith cluster and m is the fuzziness parameter.

Step_3: Compute the updated membership values u_{ij} belonging to clusters for each pixel and cluster centroids according to the given formula.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{2/(m-1)}},$$

and

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}}.$$

Step_4: Repeat steps 2-3 until the cost function is minimized.

V.MULTIPLE FEATURE CLUSTERING

The clustering algorithms used for microarray image segmentation are based on the information about the intensities of the pixels only. But in microarray image segmentation, the position of the pixel and median value of surrounding pixels also influences the result of clustering and subsequently that leads to segmentation. Based on this observation, multiple feature clustering algorithm is developed for segmentation of microarray images. To apply clustering algorithms on a single spot, we take all the pixels that are contained in the spot are, which is obtained after gridding process, and create a dataset D = {x_1, x_2, x_3, x_4, x_5,……,x_n}, where x_i = [$x_i^{(1)}$, $x_i^{(2)}$, $x_i^{(3)}$] is a three dimensional vector that represents the ith pixel in the spot region. We use three features, defined as follows

$x_i^{(1)}$: Represents the pixel intensity value.

$x_i^{(2)}$:Represents the distance from pixel to the center of the spot region.

The spot center is calculated as follows:

1. Apply edge detection to the spot region image using canny method.
2. Perform flood-fill operation on the edge image using imfill method.
3. Obtain label matrix that contain labels for the 8-conneted objects using bwlabel function.
4. Calculate the centroid of each labeled region (connected component) using regionprops method.

$x_i^{(3)}$:Represents the median of the intensity of surrounding pixels.

For each pixel in the spot region, once the features are obtained forming the dataset D, then the FCM clustering algorithm is applied.

VI.EXPERIMENTAL RESULTS

The proposed four different clustering algorithms are performed on a sample microarray slide that has 48 blocks, each block consisting of 110 spots. A sample block has been chosen and 36 spots of the block have been cropped for simplicity. The sample image is a 198*196 pixel (gray scale) image that consists of a total of 38808 pixels.

The segmentation step implemented separately by three clustering methods, K-means, Moving K-Means, Fuzzy C-means with multiple features for each pixel. These methods are implemented in such a way that the grayscale intensity value of all the pixels in the image are grouped into two clusters. The segmented microarray images using multiple feature FCM is shown in figure 2.

Quantitative analysis is a numerically oriented procedure to figure out the performance of algorithms without any human error. The Mean Square Error (MSE) is significant metric to validate the quality of image. It measures the square error between pixels of the original and the resultant images. The MSE is mathematically defined as

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i\in c_j} ||v_i-c_j||^2 (13)$$

Where N is the total number of pixels in an image and x_i is the pixel which belongs to the j^{th} cluster. The lower difference between the resultant and the original image reflects that all the data in the region are located near to its centre. Table 2 shows the quantitative evaluations of four clustering algorithms after segmenting the microarray image. The results confirm that Multiple Feature FCM algorithm produces the lowest MSE value for segmented microarray image.

Table 3. Segmentation result for Image 1

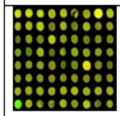
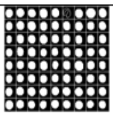
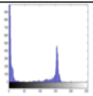
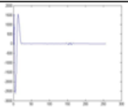
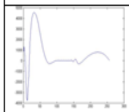
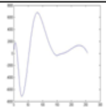
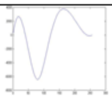
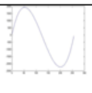
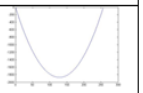
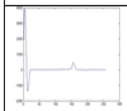
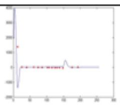
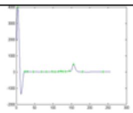
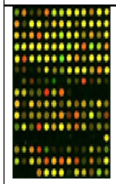
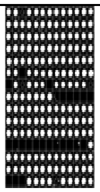
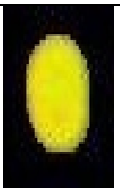
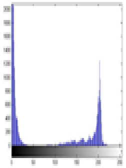
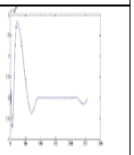
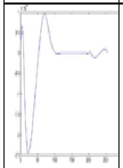
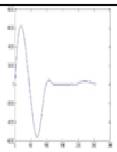
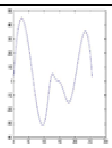
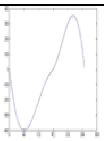
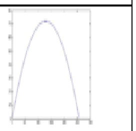
Image 1	Gridded Image	Compartment No 1	Histogram	IMF1
				
IMF2	IMF3	IMF4	IMF5	IMF6
				
Combined IMF	Local Minima	Local Maxima	Centroids	Segmented Image using Multiple features
			No of clusters :2 Centroids are 4, 145	

Table 4. Segmentation result for Image 2

Image 2	Gridded Image	Compartment No 8	Histogram	IMF1
				
IMF2	IMF3	IMF4	IMF5	IMF6
				
Combined IMF	Local Minima	Local Maxima	Centroids	Segmented Image usingMultiple features

VII.CONCLUSIONS

This paper has presented clustering algorithms namely K-means, Moving K-means, Fuzzy K-means and multiple feature FCM for the segmentation of microarray image with multiple features for each pixel. The qualitative and quantitative analysis done proved that Multiple Feature

Fuzzy C-means has higher classification quality of spots than other clustering algorithms. The occurrence of dead centers, center redundancy and trapped center at local minima problems can be avoided. The proposed clustering algorithms are also less sensitive to initialization process of clustering value.

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 Wen-Lin 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, Wei-ChiTsai, Pin-Hua Chen, “Error Reduction on Automatic Segmentation in Microarray Image”, IEEE 2007.

[4] EleniZacharia and DimitirsMaroulis, “Microarray Image Analysis based on an Evolutionary Approach” 2008 IEEE.

[5] R.Hirata, J.Barrera, R.F.Hashinoto and D.o.Dantas, “ Microarray gridding by mathematical morphology”, in Proceedings of 14th Brazilian Symposium on Computer Graphics and Image Processing, 2001, pp: 112-119

[6] VolkanUslan, OmurBucak, “clustering based spot segmentation of microarray cDNA Microarray Images “, International Conference of the IEE EMBS , 2010.

[7] SitiNarainiSulaiman, Nor Ashidi Mat Isa, “ Denoising based Clutering Algorithms for Segmentation of Low level of Salt and Pepper Noise Corrupted Images”, IEEE Transactions on Consumer Electronics, Vol. 56, No.4, November 2010.

[8] LJun-Hao Zhang, Ming Hu HA , Jing Wu,” Implementation of Rough Fuzzy K-means Clustering Algorithm in Matlab”, Proceedings of Ninth International Conference on Machine Learning and Cybernetics”, July 2010.

[9] Nor Ashidi Mat Isa,” Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation”, IEEE 2009.