Fast Clustering Algorithms for Segmentation of Microarray Image

J.Harikiran, Dr.P.V.Lakshmi, Dr.R.Kirankumar

Abstract: Microarray technology allows the simultaneous monitoring of thousands of genes. Based on the gene expression measurements, microarray technology have proven powerful in gene expression profiling for discovering new types of diseases and for predicting the type of a disease. Gridding, segmentation and intensity extraction are the three important steps in microarray image analysis. In this paper, three different clustering algorithms such as K-means, Moving K-means and Fuzzy C-means are used for segmentation of microarray images. In all the traditional clustering algorithms, number of clusters and initial centroids are randomly selected and often specified by the user. In this paper, a hill climbing algorithm for the histogram of the input image will generate the number of clusters and initial centroids required for clustering. It overcomes the shortage of random initialization and achieves high computational speed by reducing the number of iterations. The experimental results show that Fuzzy C-means has segmented the microarray image more accurately than other three algorithms.

Index terms: Microarray Image, Image Processing, Image segmentation

—————————— ◆ ——————————

1 INTRODUCTION

Microarrays widely recognized as the next revolution in molecular biology that enable scientists to monitor the expression levels of thousands of genes in parallel [1]. A microarray is a collection of blocks, each of which contains a number of rows and columns of spots. Each of the spot contains multiple copies of single DNA sequence [2]. The intensity of each spot indicates the expression level of the particular gene [3]. The processing of the microarray images [4] [5] usually consists of the following three steps: (i) gridding, which is the process of segmenting the microarray image into compartments, each compartment having only one spot and background (ii) Segmentation, which is the process of segmenting each compartment into one spot and its background area (iii) Intensity extraction, which calculates red and green foreground intensity pairs and background intensities.

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. An automatic gridding method by using the horizontal and vertical profile signal of the image presented in [5] is used to perform the image gridding. The algorithm can satisfy the requirements of microarray image segmentation.

In this paper, four different clustering algorithms are used for segmentation of microarray image. In the clustering algorithms, parameters such as cluster number and initial centroid positions are chosen randomly and often specified by the user. Instead of randomly initializing the parameters in the clustering algorithms, the hill climbing algorithm on the histogram of input image will automatically determine the cluster centers and the number of clusters in the image. Using hill climbing algorithm as a preliminary stage with clustering algorithms reduces the number of iterations for segmentation and costs less execution time. The qualitative and quantitative results show that Fuzzy C-means clustering algorithm has classified the image better than other clustering algorithms. The paper is organized as follows: Section 2 presents Hill Climbing Algorithm, Section 3 presents K-means clustering algorithm, Section 4 presents Moving K-means clustering algorithm, Section 5 presents Fuzzy C-means clustering algorithm, Section 6 presents Experimental results and finally Section 7 report conclusions.

2 HILL CLIMBING ALGORITHM

Traditional hill-climbing segmentation [11] [12] is a nonparametric algorithm that clusters the colors of an image. The idea is that each cluster is represented by a hill in the histogram, where the hill consists of adjacent colors. In this paper, an extended version of hill climbing algorithm is presented. This Hill climbing algorithm which is used in the preliminary stage for a clustering algorithm is stated as follows:

Input: Histogram of the microarray image .

Output: The number and values of peaks= Number of clusters and initial centroids respectively

Step 1: Start at a non-zero bin of the histogram and make uphill moves until reaching a peak as follows:

1. Compare the number of pixels of the current histogram bin with the number of pixels of the neighboring bins.
2. If the neighboring bins have different number of pixels, the algorithm make an uphill move towards the neighboring bin with large number of pixels.
3. If the immediate neighboring bins have the same number of pixels, the algorithm checks the next neighboring bins, and so on, until two neighboring bins with different number of pixels are found. Then, an uphill move is made towards the bin with larger number of pixels.
4. The uphill is continued until reaching a bin from where there is no possible hill movement. That is the case when the neighboring bins have smaller number of pixels than the current bin. Hence the current bin is identified as peak representing local maximum.
5. If no uphill move is done, the stopping bin is identified as a peak of a hill, and all bins leading to this peak are associated with it.

Step 2: Select another unclimbed bin as a starting bin and perform step 1 to another peak. This step is continued until all the nonzero bins of the histogram are climbed.

Step 3: Thresholding: Find the peaks whose value is higher than one percent of the maximum peak (which is identified as local maximum in step 1 and 2).

Step 4: Remove the peaks which are very close. This is done by checking the difference between the grey levels of the two individual peaks. If the difference is less than 20, then the peak with lowest value is removed.

Step 5: Neighboring pixels that lead to the same peak are grouped together.

Step 6: The identified peaks represent the initial number of clusters of the input image. Thus the number and values of the peaks are saved.

3 K-MEANS CLUSTERING ALGORITHM

K-means is one of the basic clustering methods introduced by Hartigan [6]. This method is applied to segment the microarray image in recent years. The K-means clustering algorithm for segmentation of microarray image is summarized as follows [14]:

Algorithm K-means(x, N, c)

Input:

N: number of pixels to be clustered;

x={x_1, x_2, x_3,……, x_N}: pixels of microarray image

c = {c_1, c_2, c_3,…., c_j}: clusters respectively.

Output:

cl: cluster of pixels

Begin

Step 1: cluster centroids and number of clusters are determined by Hill climbing algorithm.

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

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

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

End.

4 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 new position of cluster is calculated. The Moving K-means clustering algorithm for segmentation of microarray image is summarized as follows:

Algorithm Moving K-means(x, N, c)

Input:

N: number of pixels to be clustered;

x={x_1,x_2,x_3,……,x_N}: pixels of microarray image

c ={c_1,c_2,c_3,….,c_j} : clusters respectively.
Output:
cl: cluster of pixels
Begin
Step 1: cluster centroids and number of clusters are determined by Hill climbing algorithm.
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:

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

Step 3: The fitness for each cluster is calculated using

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

All centers must satisfy the following condition:

$$f(c_s) \geq \alpha_a f(c_l) \tag{5}$$

where α_a is small constant value initially with value in range 0< α_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 and c_l are recalculated according to:

$$C_s = 1/n_{cs} \left(\sum_{t \in c_s} x_t \right) \tag{6}$$

$$C_l = 1/n_{cl} \left(\sum_{t \in c_l} x_t \right) \tag{7}$$

The value of α_a is then updated according to:

$$\alpha_a = \alpha_a - \alpha_a / n_c \tag{8}$$

The above process are repeated until (12) is fulfilled. Next all data are reassigned to their nearest center and the new center positions are recalculated using (9).
Step 4: The iteration process is repeated until the following condition is satisfied.

$$f(c_s) \geq \alpha_a f(c_l) \tag{9}$$

End

5 FUZZY C-MEANS CLUSTERING ALGORITHM

The Fuzzy C-means [8] [9] is an unsupervised clustering algorithm. The main idea of introducing fuzzy concept in the Fuzzy C-means algorithm is that an object can belong simultaneously to more than one class and does so by varying degrees called memberships. It distributes the membership values in a normalized fashion. It does not require prior knowledge about the data to be classified. It can be used with any number of features and number of classes. The fuzzy C-means is an iterative method which tries to separate the set of data into a number of compact clusters. It improves the partition performance and reveals the classification of objects more reasonable. The predefined parameters such as number of clusters and initial clustering centers are provided by Hill climbing algorithm. The Fuzzy C-means algorithm is summarized as follows:

Algorithm Fuzzy C-Means (x,N,c,m)
Input:
N=number of pixels to be clustered;
x = {x_1, x_2 ,..., x_N}: pixels of Microarray image;
c ={c_1,c_2,c_3,….,c_j} : clusters respectively.
m=2: the fuzziness parameter;
Output:
u: membership values of pixels and clustered Image
Begin
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.

$$\sum_{j=1}^{c} u_{ij} = 1 \tag{10}$$

for all i= 1,2,…….N, where c 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:

$$F = \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij}^{m} \| x_j - c_i \|^2 \tag{11}$$

where u_{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

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}}. \qquad (12)$$

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

6 EXPERIMENTAL RESULTS

Qualitative Analysis:

The proposed three clustering algorithms are performed on a two microarray image drawn from the standard microarray database corresponds to breast category aCGH tumor tissue. Image 1 consists of a total of 38808 pixels and Image 2 consists of 64880 pixels. Clustering algorithms with and without hill climbing are executed on the two microarray images. The segmentation result of Fuzzy C-means clustering algorithm with hill climbing on two images is shown in figure 1. The hill climbing algorithm is executed on the histogram of input images for identification of number of clusters and initial centroids which are required for clustering algorithms. The centroids for the first image are 2 and 127, and for the second image the centroids are 17 and 181.

Quantitative Analysis:

Quantitative analysis is a numerically oriented procedure to figure out the performance of algorithms without any human error. The Mean Square Error (MSE) [10] 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

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

Where N is the total number of pixels in an image and xi is the pixel which belongs to the jth 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 1 shows the quantitative evaluations of three clustering algorithms. The results confirm that Fuzzy C-means algorithm produces the lowest MSE value for segmenting the microarray image. As the initial centroids required for clustering algorithms are determined by Hill climbing algorithm, the number of iterative steps required for classifying the objects is reduced. While the initial centroids obtained by hill climbing are unique, the segmented result is more stable compared with traditional algorithms. Table 2 shows the comparison of iterative steps numbers for clustering algorithms with and without Hill climbing.

7 CONCLUSION

This paper has presented three clustering algorithms namely K-means, Moving K-means and Fuzzy C-means combined with hill climbing for the segmentation of microarray image. The qualitative and quantitative analysis done proved that Fuzzy C-means has higher segmentation quality than other clustering algorithms. Clustering algorithm combined with hill climbing overcomes the problem of random selection of number of clusters and initialization of centroids. The proposed method reduces the number of iterations for segmentation of microarray image and costs less execution time.

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, Oct 20;270(5235):467-70.
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. Eleni Zacharia and Dimitirs Maroulis, "Microarray Image Analysis based on an Evolutionary Approach" IEEE 2008.
5. J.Harikiran, B.Avinash, Dr.P.V.Lakshmi, Dr.R.Kiran Kumar,"Automatic Gridding Method for Microarray Images", Journal of Applied Theoritical and Information Technology",Vol 65, No 1, pp 235-241, 2014.
6. Volkan Uslan, Omur Bucak, " clustering based spot segmentation of microarray cDNA Microarray Images ", International Conference of the IEE EMBS , 2010.
7. Siti Naraini Sulaiman, 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, Samy A.Salamah, Umi Kalthum Ngah.," Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation", IEEE Transaction on Consumer Electronics, 12/2009; DOI: 10.1109/TCE.2009.5373781.

10. B.Saichandana, Dr.K.Srinivas, Dr.R.KiranKumar," De-noising based clustering Algorithm for Classification of Remote Sensing Image", Journal of Computing, Volume 4, Issue 11, November 2012.

11. Zhengjian DING, Juanjuan JIA, DIA LI , "Fast Clustering Segmentation Method Combining Hill Climbing for Color Image", Journal of Information and Computational Sciences, Vol 8, pp. 2949-2957.

12. Takumi OHASHI, Zaher AGHBARI, Akifumi MAKINOUCHI," Hill Climbing algorithm for Efficient Color bases Image Segmentation", Signal Processing, Pattern Recognition and Applications, 01/2003.

13. Zhengjian DING, Juanjuan JIA, DIA LI , "Fast Clustering Segmentation Method Combining Hill Climbing for Color Image", Journal of Information and Computational Sciences, Vol 8, pp. 2949-2957.

Table 2: Comparison of iterative step numbers

	Clustering algorithm	Iterative steps (without hill climbing)	Iterative steps (with hill climbing)
IMAGE 1	K-means	10	4
	Moving K-means	14	6
	Fuzzy C-means	17	9

	Clustering algorithm	Iterative steps (without hill climbing)	Iterative steps (with hill climbing)
IMAGE 2	K-means	20	12
	Moving K-means	27	16
	Fuzzy C-means	31	21

Table 1: MSE values

Method	MSE Values (IMAGE 1)	MSE Values (IMAGE 2)
K-means	282.781	346.47
Moving K-means	216.392	298.69
Fuzzy C-means	138.327	198.76

Image 1	Fuzzy K-means

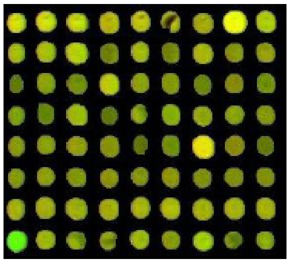
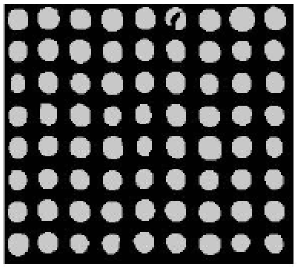

	

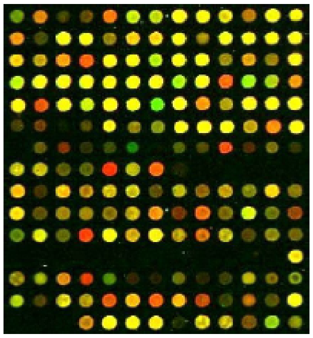
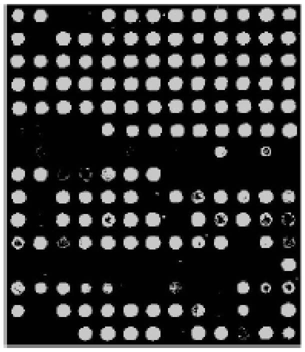
Image 2	Fuzzy K-means
	

Fig 1: Segmentation result