International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 1, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Various Versions of K-means Clustering Algorithm for Segmentation of Microarray Image

D. Rama Krishna
Assistant Prof., Deptt. of CSE, GIT, GITAM University

J. Harikiran
Member IEEE & Assist. Prof., GIT, GITAM University

Dr. P. V. Lakshmi
Professor, Department of IT, GIT, GITAM University

Dr. K. V. Ramesh
Professor, Department of Civil, GIT, GITAM University

***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 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. We used four versions of clustering algorithms called K-means, Moving K-means, Fuzzy K-means and Fuzzy Moving K-means for microarray image segmentation that separate the spots from the background. The experimental results show that Fuzzy Moving K-means have segmented the spots of the microarray image more accurately than other three algorithms.**

***Keywords* – K-means Algorithm, Clustering Segmentation, Microarray Image.**

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].

Microarray expression technology helps in the monitoring of gene expression for tens and thousands of genes in parallel. During the biological experiment, the mRNA of two biological tissues of interest is extracted and purified. Each of the mRNA samples are reverse transcribed into complementary DNA (cDNA) copy and labeled with two different fluorescent dyes resulting in two fluorescence-tagged cDNA (red Cy5 and green Cy3). The tagged cDNA copies, called the sample probe, are hybridized with the slide's DNA spots. The hybridized glass slides are fluorescently scanned at different wavelengths (corresponding to the different dyes used), and two digital images are produced, one for each population of mRNA. Each digital image contains a number of spots of various fluorescence intensities. The intensity of each spot is proportional to the hybridization level of the cDNAs and the DNA dots, the gene expression information is obtained by analyzing the digital images [3].

The processing of the microarray images [4] 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. 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.

In this paper, we used three versions of conventional k-means clustering algorithm for segmentation of microarray image. The qualitative and quantitative results show that Fuzzy Moving K-means clustering algorithm has segmented the spots from the background than other three 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 Fuzzy Moving K-means Clustering algorithm, Section 6 presents Experimental results and finally Section 7 report conclusions.

II. 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 segmenting the microarray image is summarized as follows:

Algorithm K-means(x,n,c)

Input:

N: number of pixels to be clustered;

$x=\{x_1,x_2,x_3,\ldots\ldots,x_N\}$: pixels of microarray image

$c=\{c_1,c_2,c_3,\ldots,c_j\}$: clusters respectively. Here we group the pixels into two clusters, foreground and background, j=2.

Output:

cl: cluster of pixels

Begin

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:

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

International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 1, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

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

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

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

End.

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

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

Input:

N: number of pixels to be clustered;

$x=\{x_1,x_2,x_3,\ldots\ldots,x_N\}$: pixels of microarray image

$c=\{c_1,c_2,c_3,\ldots,c_j\}$: clusters respectively. Here we group the pixels into two clusters, foreground and background, j=2.

Output:

cl: cluster of pixels

Begin

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:

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

Step 3: The fitness for each cluster is calculated using

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

All centers must satisfy the following condition:

$$f(c_s) \geq \alpha_a f(c_l) \quad (5)$$

where α_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 and c_l are recalculated according to:

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

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

The value of α_a is then updated according to:

$$\alpha_a = \alpha_a - \alpha_d / n_c \quad (8)$$

The above process are repeated until (5) is fulfilled. Next all 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.

$$f(c_s) \geq \alpha_a f(c_l) \quad (9)$$

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 K-Means Clustering Algorithm

The Fuzzy K-means [8] is an unsupervised clustering algorithm. The main idea of introducing fuzzy concept in the Fuzzy K-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 segmented. It can be used with any number of features and number of classes. The fuzzy K-means is an iterative method which tries to separate the set of data into a number of compact clusters. The Fuzzy K-means algorithm is summarized as follows:

Algorithm Fuzzy K-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=2: foreground and background clusters;

m=2: the fuzziness parameter;

Output:

u: membership values of pixels and segmented 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.

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

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:

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

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.

V. Fuzzy Moving K-Means Clustering Algorithm

In the Fuzzy Moving K-means clustering algorithm [9], the membership function is used in addition to the Euclidian distance to control the assignment of the members to the proper center. The algorithm minimizes the sensitivity to the noisy data by updating the moving member function. It is not obligatory for the members of the center with the largest fitness value to follow the center with the smallest fitness value. The Fuzzy Moving K-means clustering algorithm is summarized as follows:

Input:

N: number of pixels to be clustered;

$x=\{x_1,x_2,x_3,\ldots\ldots,x_N\}$: pixels of microarray image

$c=\{c_1,c_2,c_3,\ldots,c_j\}$: clusters respectively. Here we group the pixels into two clusters, foreground and background, j=2.

m=2: the fuzziness parameter;

Output:

u: membership values of pixels and segmented 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.

$$\sum_{j=1}^{c} u_{ij} = 1 \qquad (13)$$

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:

$$F = \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij}^{m} \, || x_j - c_i ||^2 \qquad (14)$$

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

Step 3: The fitness for each cluster is calculated using

$$f(c_k) = \sum_{t \in c_k} (|| x_t - c_k ||)^2 \qquad (15)$$

All centers must satisfy the following condition:

$$f(c_s) \geq \alpha_a \, f(c_l) \text{ and } m(c_{sk}) > m(c_{lk}) \qquad (16)$$

where α_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, $m(c_{sk})$ is the membership value of point k according to the smallest centre and $m(c_{lk})$ is the membership value of point k according to the largest centre. 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:

$$C_s = 1/n_{cs} \left(\sum_{t \in c_s} x_t \right) \qquad (17)$$

$$C_l = 1/n_{cl} \left(\sum_{t \in c_l} x_t \right) \qquad (18)$$

The value of α_a is then updated according to:

$$\alpha_a = \alpha_a - \alpha_a / n_c \qquad (19)$$

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

Compute the updated membership values u_{ij} belonging to clusters for each pixel according to 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

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}}. \qquad (20)$$

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

$$f(c_s) \geq \alpha_a \, f(c_l) \quad \text{and} \quad m(c_{sk}) > m(c_{lk}) \qquad (21)$$

International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 1, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

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 four clustering methods, K-means, Moving K-Means, Fuzzy K-means and Fuzzy Moving K-means respectively. 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 after each clustering methods have been performed are shown in figure 1.

The number of pixels clustered for each method has been presented in Table 1. The tabulated values show that: K-means method clustered 12986 pixels as foreground and 17822 pixels as background. Moving K-means method clustered 13774 pixels as foreground and 17034 pixels as background. Fuzzy K-means method clustered 14163 pixels as foreground and 16645 pixels as background. Fuzzy Moving K-means method clustered 15019 pixels as foreground and 15789 pixels as background.

Table 1: The number of pixels clustered as spots and background

Method	Spots	Background
K-means	12986	17822
Moving K-means	13774	17034
Fuzzy k-means	14163	16645
Fuzzy Moving K-means	15019	15789

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

$$MSE = \frac{1}{N}\sum_{j=1}^{k}\sum_{i \in c_j} ||v_i - c_j||^2 \qquad (22)$$

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 Fuzzy Moving K-means algorithm produces the lowest MSE value for segmented microarray image.

Table 2: MSE values of four segmented Images

Method	MSE Values
K-means	322.781
Moving K-means	286.392
Fuzzy K-means	198.327
Fuzzy Moving K-means	146.322

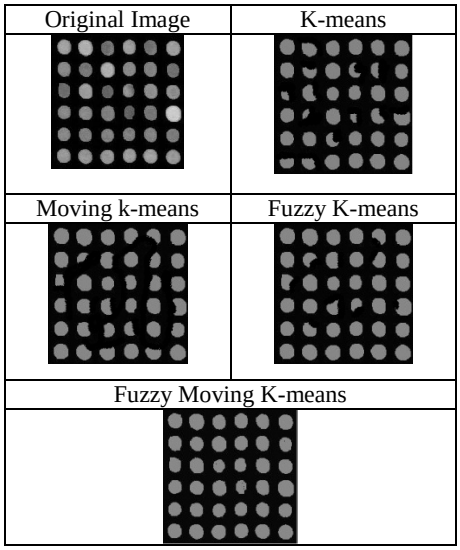

Fig.1. Segmented results using four clustering algorithms

VI. CONCLUSION

This paper has presented four clustering algorithms namely K-means, Moving K-means, Fuzzy K-means and Fuzzy Moving K-means for the segmentation of microarray image. The qualitative and quantitative analysis done proved that Fuzzy Moving K-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] Eleni Zacharia and Dimitirs Maroulis, "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] 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," Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation", IEEE 2009.