Asian Journal of Engineering and Applied Technology
ISSN 2249-068X Vol. 7 No. 2, 2018, pp.31-36

Self-Organization Map Based Segmentation of Breast Cancer

A. Arokiya Mary Delphia[1], M. Kamarasan[2] and S. Sathiamoorthy[3]
[1]Research Scholar, [2&3]Assistant Professor, Division of Computer and Information Science
Annamalai University, Annamalai Nagar, Tamil Nadu, India
E-Mail: ks_sathia@yahoo.com, smkrasan@yahoo.com

***Abstract* - Breast cancer is the second major leading cause of cancer fatality in women. Mammography prevails the best method for initial detection of cancers of the breast, capable of identifying small pieces up to two years before they grow large enough to be evident on physical testing. X-ray images of the breast must be accurately evaluated to identify beginning signs of cancerous growth. Segmenting, or partitioning, radiographic images into regions of similar texture is usually performed during the method of image analysis and interpretation. The comparative lack of structure definition in mammographic images and the implied transition from one texture to makes segmentation remarkably hard. The task of analyzing different texture areas can be considered a form of the exploratory report since a priori awareness about the number of different regions in the image is not known. This paper presents a segmentation method by using SOM.**
***Keywords* - Breast Cancer, Mammography, Self-Organizing Map, Euclidean Distance, Validity Measure, Double Bouldin Index**

I. INTRODUCTION

According to the USA Cancer Society, breast cancer is in second place as the most common type of cancer afflicting women but remains the leading cause of cancer mortality in women within the ages of 40 and 55. Recent year in the United States, approximately 180,200 women will be diagnosed with invasive breast cancer. Meanwhile the same year, about 44,190 women will lose the fight against this life-threatening disease [1]. Although the proportion of new breast cancer rose on average 4 percent between the years 1982 and 2017, the percentage rate has tapered off to just over one percent in the years since. Much of this welcome decrease in new breast cancer diagnoses have been attributed to the increased use of mammography to detect early stages of this disease. Although significant prediction technique has been made in the technology of mammography, much work remains to be done to improve overall detection accuracy [2]. Segmentation is the process of partitioning an image into multiple fragments [3]. All the pixels in a region are similar to some characteristic, such as color, intensity, or texture. Artificial neural networks [4] are parallel computational models, comprised of densely interconnected adaptive processing units. An essential feature of these networks is that they learn by example. The adaptive nature of the artificial neural networks makes it more suitable for applications where one has a little or incomplete understanding of the problem to be solved but where training data is readily available.

II. SELF-ORGANIZING MAP (SOM)

A Self-Organizing Map is used to project the high dimensional data on to the two-dimensional map [5]. The dimensional reduction could allow us to visualize the important relationship among the data more easily. The topology structure property which is observed in the brain is also noted in SOM which is not observed in any other artificial neural network. It is said to be topology preserving since it preserves the neighborhood relation of the input pattern. The units that are physically located next to each other will respond to classes of input vectors that are likewise located next to each other.

The basic SOM model consists of an input layer and an output layer. The SOM network consists of neurons which are similar to the neurons in the brain. Each input is fully connected to all the units. The number of neurons in the output layer depends on the number of clusters in the image to be segmented, i.e., the number of clusters is equal to the number of output neurons. Color is one of the essential features used for image segmentation. SOM is used to map patterns in a three-dimensional color space to a two-dimensional space. SOM learns through competition. For each input vector, only one neuron in the network will respond. This mechanism is known as competition. Once a neuron is elected as a winner, the weights of that neuron and the neurons in the neighborhood of the victor are updated. The neighborhood scheme for SOM may be rectangular, hexagonal or circular. The multicomponent values are given as input for training. Initially, the learning rate is set to 0.1, and the neighborhood size is initialized to the maximum of either the height or width of the network divided by two. The weight vectors of the neurons are initialized randomly. For every iteration, the input vectors to be clustered are presented to the network in a random order. The neurons with weight vector that best match the input vector is elected as the winner or the best matching unit (BMU). The winner is elected by using the Euclidean distance method which is as follows

$$\left\|x - W_i^{[k]}\right\| = \min_i \left\|x - W_i^{[k]}\right\| \qquad (1)$$

where x is the input vector, W is the weight of the winning unit i at each iteration k. The winning neuron and the neurons within the neighborhood of the winning unit are updated in such a way that their weights become closer to

the input vector being presented to the network. The weights are updated as follows.

$$W_i^{k+1} = W_i^k + H_{li}^k \text{ (x - } W_i^k) \quad (2)$$

where H is the smoothing kernel defined over the winning neuron. The kernel can be written concerning Gaussian function as

$$H_{li}^k = \alpha^k exp\left(-\frac{d^2(l,i)}{2(\sigma^k)^2}\right) \quad (3)$$

where d is the distance between the winning neuron and the neuron i and σ is the neighborhood distance, and σ k is the learning rate at iteration k. The learning rate and the neighborhood size are updated after each iteration. As the number of iterations increases, the learning rate and the neighborhood decreases. The learning rate is exponentially reduced as follows

$$\alpha^k = \alpha^0 \exp\left(-\frac{k}{T}\right) \quad (4)$$

where σ 0 is the initial learning rate, and T is the total number of iterations which is set to 1000. The decreasing function for the neighborhood is given as follows

$$\sigma^k = \sigma^0 \left(1 - \frac{k}{T}\right) \quad (5)$$

where σ^0 is the initial neighborhood size and σ^k is the neighborhood size at iteration k. The size of the neighborhood is decreased until it encompasses a single unit.

Once the SOM converges, the input is mapped from a high color space to a two-dimensional map. The final result of SOM depends on the initial values of weights, data used for training, and the characteristics of the map such as some nodes in the network, learning rate, and the neighborhood. SOM suffers from a drawback of over-segmentation. So, an optimization method like genetic algorithm is used to identify the optimal number of clusters. The data set identified from SOM is given as an input to an optimization method for identifying the cluster centers.

III. PROPOSED FRAMEWORK FOR THE SEGMENTATION OF BREAST CANCER IMAGES

The following figure 1 represents the proposed framework for the classification of breast cancer cell using Pre-processing step, Segmentation step. In this work, the noise of the image is removed by using median filtering method in the pre-processing step; SOM is used in the segmentation step.

A. Pre-Processing with Median Filter

Noise is an unwanted signal in the image. The noise in the image is of three types Salt and Pepper noise, Impulse noise, Gaussian noise. A Median Filter operates over the window by selecting the median intensity in the window, The advantages of using Median filter is of its robust average ,that is its unrepresentative pixel in the neighborhood does not affective the median value and also it has the quality of preserving sharp edges. The median filter works through the image by pixel by pixel and replaces it with median value of neighboring pixels and the pattern of neighbors is called the window which slides pixel by pixel over the entire image.

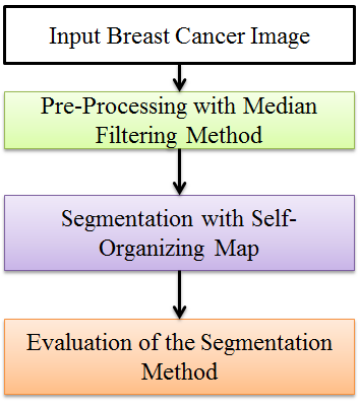

Fig. 1 Proposed Framework for the SOM based Segmentation of Breast Cancer Image

B. Segmentation using SOM

Clustering process in this study uses gray scale values of each pixel as an input to SOM method. Neighborhood topology which is used in SOM method in this study is a linear array or also known as one dimensional (1-D) topology. Calculation of SOM algorithm is split into two stages, the stage of learning, and recognition stage. In this research, to determine the distance does not use Euclidean Distance, but it uses Normalized Euclidean Distance.

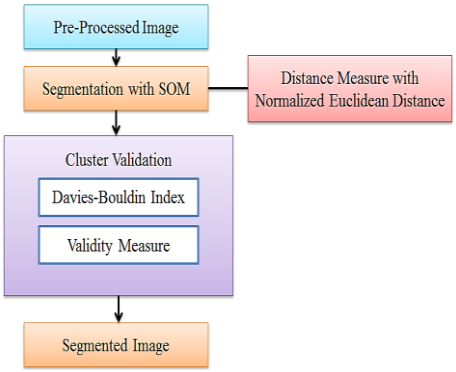

Fig. 2 Segmentation Process of Self-Organizing Map

C. Normalized Euclidean Distance

The computation of the Normalized Euclidean Distance is modified form of the Euclidean Distance [8]. Normalized Euclidean Distance of two vectors, between vector *u* and vector *v* is shown by the following equation

$$d_{(uv)} = \sqrt{\sum_{k=1}^{n} (\bar{u} - \bar{v})^2}$$

Where

$$\overline{u_i} = \frac{u_i}{\|u\|}, \qquad \overline{v_i} = \frac{v_i}{\|v\|}$$

||v|| is normalized value of vector v. The normalized value is expressed in the following equation

$$\|v\| = \left[\sum_{i=1}^{n} v_i^{\,2}\right]^{\frac{1}{2}}$$

IV. CLUSTER VALIDATION

A. Davies-Bouldin Index (DBI)

DBI was preceded in 1979 by David L. Davies and Donald W. DBI is applied to evaluate the clustering results [9]. DBI is a method to measure the ratio of the total within-cluster scatter (a spread of the cluster) and the between-cluster separation (distance between clusters). Below Equation is used to calculate the spread of cluster value.

$$S_i = \frac{1}{T_i}\sum_{x \in C_i} \|x - z_i\|$$

where T_i is a number of member in i[th] cluster (C_i), and z_i is the i[th] cluster center. The distance between clusters is calculated by the Euclidean distance between the center of i[th] cluster and center of j[th] cluster. Following equation is used to calculate its distance.

$$d_{ij} = \|z_i - z_j\|$$

R_{ij} is ratio value between i[th] cluster and j[th] cluster, which is calculated by the following formula.

$$R_{ij} = \left\{\frac{S_i + S_j}{d_{ij}}\right\}$$

Finding the maximum value of the ratio (D_i), it is used to find the value of DBI. Following equation is used to calculate the value of D_i.

$$D_i = \max_{j:j \neq i} R_{ij}$$

Then, DBI value is calculated by using Equation.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} D_i$$

where k is the number of clusters.

DBI with the most minimum value indicates the most optimal clustering results and achieve well-separated cluster.

B. Validity Measure (VM)

VM is one of the indexes to test the validity of clustering results [10]. VM is commonly used in the application of image segmentation based on clustering. VM is calculated using below given equation.

$$VM = y\left(\frac{intra}{inter}\right)$$

where intra is intra-cluster distance, inter is inter-cluster distance, and y is a function of the number of the clusters that is formed. Below equation is used to find the value of intra-cluster distance.

$$intra = \frac{1}{N}\sum_{i=1}^{k}\sum_{x \in C_i} \|x - z_i\|^2$$

where N is a total number of pixels in an image, k is number of clusters, and zi is the center of cluster Ci.

Also, to calculate VM, takes the minimum value of inter-cluster distance [10]. Below Equation is used to find the inter-cluster distance.

$$inter = \min\left(\|z_i - z_j\|\right)$$

where $i = 1,2 \ldots k$, and $j = i+1, \ldots, k$. y is multiplied by the quotient between the distance of intra-cluster and inter-cluster. Below equation is used to calculate y.

$$y = c.N(2,1) + 1$$

where c is a constant value in the range of 15 to 25, N (2,1) is a Gaussian function for the number of clusters (k). The Gaussian function is shown in Equation

$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(k-\mu)^2}{2\sigma^2}\right]}$$

VM should be minimum to obtain an optimal result and achieve well-separated cluster.

IV. RESULT AND DISCUSSION

Following figure 3 presents the breast cancer images considered for the segmentation process whereas

1. Image1.jpg,
2. Image2.jpg,
3. Image3.jpg,
4. Image4.jpg and table I gives the initialization of parameters on SOM method and spatial operations

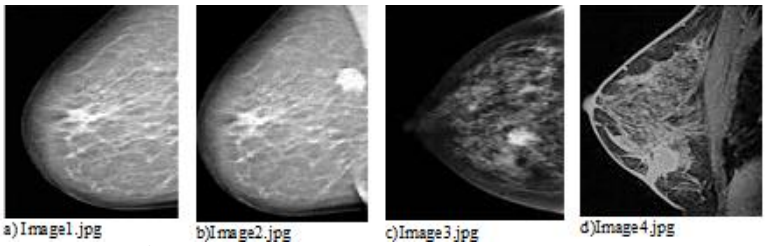

Fig. 3 Breast Cancer Image for Segmentation process a) Image1.jpg b) Image2.jpg c) Image3.jpg d) Image4.jpg

TABLE I INITIALIZATION OF PARAMETERS ON SOM METHOD

S. No.	Parameter	Value
1	A	0.2
2	Epoch (T)	200
3	The radius of the noise filter	3
4	Threshold Region (ThA)	0.003 * total pixels

TABLE II VALIDITY MEASURE AND DAVIS-BOULDIN INDEX FOR THE GIVEN BREAST CANCER IMAGE1.JPG BASED ON THE NUMBER OF CLUSTERS

Number of Clusters	Image1.jpg	
	Validity Measure	Davies-Bouldin Index
2	4.98	2.710
3	4.556	3.632
4	7.851	1.662
5	4.470	2.937
6	4.954	2.380
7	4.819	3.100
8	8.661	3.789
9	8.199	3.982
10	46.11	4.150

TABLE III VALIDITY MEASURE AND DAVIS-BOULDIN INDEX FOR THE GIVEN BREAST CANCER IMAGE2.JPG BASED ON THE NUMBER OF CLUSTERS

Number of Clusters	Image2.jpg	
	Validity Measure	Davies-Bouldin Index
2	3.099	1.396
3	3.752	2.730
4	2.89	1.209
5	1.946	1.897
6	3.321	1.599
7	2.652	2.477
8	8.253	2.474
9	7.577	3.751
10	7.710	3.522

Tests are performed on each image by forming 2 until 10 clusters. From each cluster which is formed, then calculate the value of Validity Measure (VM) and Davies-Bouldin Index (DBI). Each value of VM and DBI with minimum value showed the most optimal number of cluster. Results of validity measurement for each test images are shown in the following table II – table V.

TABLE IV VALIDITY MEASURE AND DAVIS-BOULDIN INDEX FOR THE GIVEN BREAST CANCER IMAGE3.JPG BASED ON THE NUMBER OF CLUSTERS

Number of Clusters	Image3.jpg	
	Validity Measure	Davies-Bouldin Index
2	4.126	2.854
3	6.686	2.323
4	3.833	1.522
5	3.487	2.533
6	1.564	2.987
7	3.122	2.495
8	14.63	2.788
9	12.51	2.779
10	30.28	3.947

TABLE V VALIDITY MEASURE AND DAVIS-BOULDIN INDEX FOR THE GIVEN BREAST CANCER IMAGE4.JPG BASED ON THE NUMBER OF CLUSTERS

Number of Clusters	Image4.jpg	
	Validity Measure	Davies-Bouldin Index
2	2.422	4.947
3	108.2	4.896
4	15.69	2.750
5	12.86	3.221
6	18.77	4.960
7	14.68	3.331
8	13.37	4.944
9	13.79	4.667
10	16.76	4.376

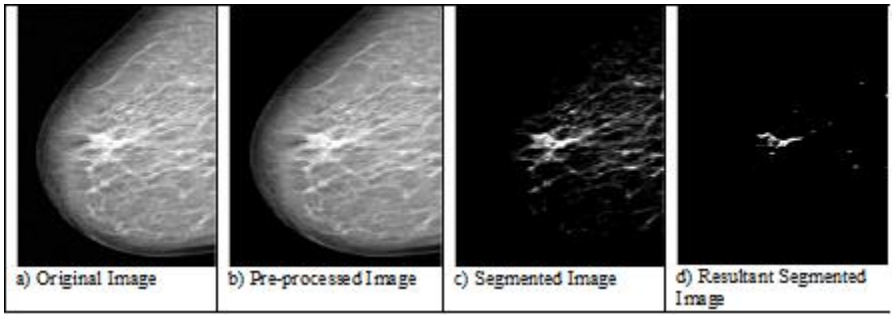

Fig. 4 Results obtained for the given Image1.jpg by Riotous Clustering and SOM Segmentation

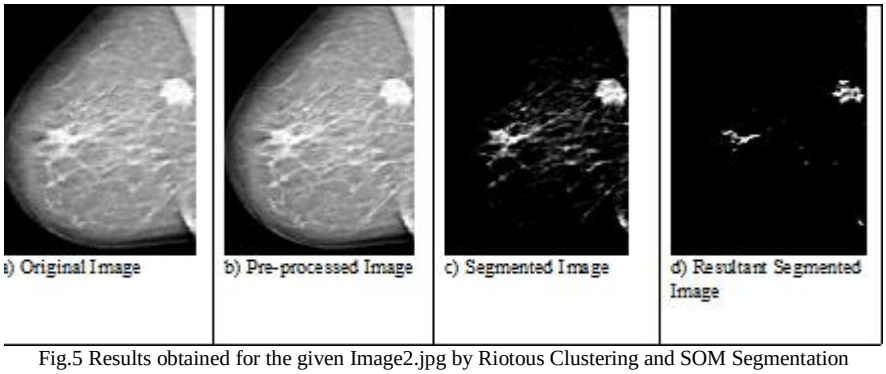

Fig.5 Results obtained for the given Image2.jpg by Riotous Clustering and SOM Segmentation

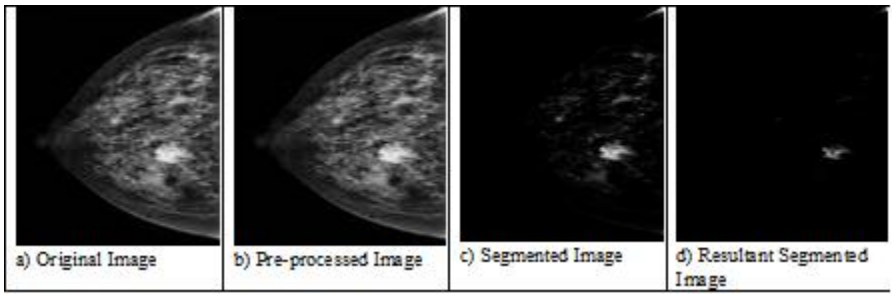

Fig.6 Results obtained for the given Image3.jpg by Riotous Clustering and SOM Segmentation

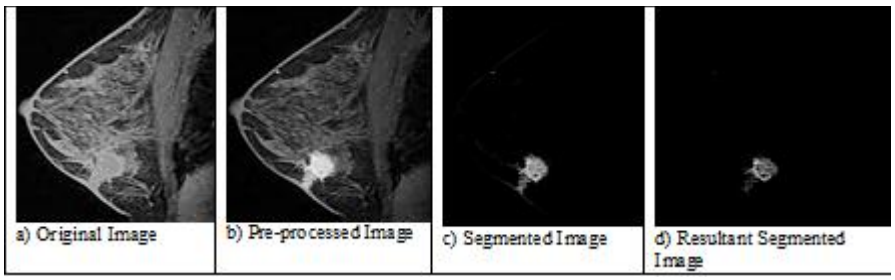

Fig.7 Results obtained for the given Image4.jpg by Riotous Clustering and SOM Segmentation

Above figures 4 to figure 7 presents the result obtained by proposed framework for given image1 to image 4. From the segmented image, the cancerous tissue can be obtained easily.

TABLE VI OPTIMAL CLUSTER FOR THE GIVEN BREAST CANCER IMAGES

S. No	Image Number	Optimal Number of Cluster	
		Validity Measure	Davies-Bouldin Index
1	Image1.jpg	5	2
2	Image2.jpg	5	2
3	Image3.jpg	6	2
4	Image4.jpg	2	2

From the table VI, the cluster number 2 gives the optimal solution by using Davies-Bouldin Index among the other cluster numbers for the given 4 images.

V. CONCLUSION

In this paper, from the proposed framework, Normalized Euclidean Distance performs clustering well and gives segmentation results as in human perception. Using this proposed work, the segmentation of the breast cancer images can be done unsupervised and automatically, by utilizing measurement of cluster validity. Davies-Bouldin Index (DBI) and Validity Measurement (VM) indexes comparatively affords distinct of optimal number of clusters. For each breast cancer images, the optimal numbers of clusters which are developed by DBI, on average are less than the results which are obtained by VM.

REFERENCES

[1] Williams, B. Lovoria, *et al.* "Demographic, psychosocial, and behavioral associations with cancer screening among a homeless population", *Public Health Nursing*, 2018.

[2] Henriksen, Emilie L., *et al.* "The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review", *Acta Radiologica*, 0284185118770917, 2018.

[3] Siddharth Singh Chouhan, Ajay Kaul and Uday Pratap Singh, "Image Segmentation Using Computational Intelligence Techniques: Review", *Archives of Computational Methods in Engineering*, 2018.

[4] Ahmed, O. Isra, Banazier A. Ibraheem and Zeinab A. Mustafa, "Detection of Eye Melanoma Using Artificial Neural Network", *Journal of Clinical Engineering*, Vol. 43, No. 1, pp. 22-28, 2018.

[5] Shukla, Nagesh, *et al.* "Breast cancer data analysis for survivability studies and prediction", *Computer Methods and Programs in Biomedicine*, Vol. 155, pp. 199-208, 2018.

[6] Arora, Shaveta, Madasu Hanmandlu, and Gaurav Gupta. "Filtering impulse noise in medical images using information sets", *Pattern Recognition Letters*, 2018.

[7] Boemer, Fabian, Edward Ratner, and Amaury Lendasse. "Parameter-free image segmentation with SLIC", *Neurocomputing*, Vol. 277, pp. 228-236, 2018.

[8] Park, Young-Seuk, et al. "Multivariate Data Analysis by Means of Self-Organizing Maps", *Ecological Informatics, Springer*, Cham, pp. 251-272, 2018.

[9] Kumar, Krishan, Deepti D. Shrimankar, and Navjot Singh. "Eratosthenes sieve based key-frame extraction technique for event summarization in videos", *Multimedia Tools and Applications*, Vol. 77, No. 6, pp. 7383-7404, 2018.

[10] Ngo, Long Thanh, Trong Hop Dang and Witold Pedrycz, "Towards Interval-Valued Fuzzy Set–based Collaborative Fuzzy Clustering Algorithms", *Pattern Recognition*, 2018.