IMPLEMENTATION AND ANALYSIS OF K-MEANS AND FUZZY C MEANS CLUSTERING TECHNIQUES

SNEHALI D. SABLE, ALPANA DESHMUKH

1Student, M.E. (VLSI and Embedded) GHRIET, Pune India  
2Asst. Prof. Dept of E&TC, GHRIET, Pune India
Email: 1snehalis.22@gmail.com, 2deshmukhalpana81@gmail.com

Abstract- Breast cancer is the second most common cause of cancer death in women. Mammography is the best available technique used for earlier detection. Mammography is a special case of CT scan who adopts X-ray method & uses the high resolution film so that it can detect well the tumors in the breast. Mammographic mass detection is an important task for the early diagnosis of breast cancer. However, it is difficult to distinguish masses from normal regions because of their abundant morphological characteristics and ambiguous margins. In the proposed work breast tumor detection and class of image is obtained by using fuzzy K-means & fuzzy C-means clustering technique is proposed. K Means algorithm is Centroid Based and Fuzzy C Means is Representative Object Based. These two algorithms are to be implemented and the performance is to be analyzed based on their clustering result quality.

Keywords- Breast Cancer, Mammograms, Clustering Technique, Fuzzy C means (FCM), Fuzzy K Means (FKM).

I. INTRODUCTION

Breast image analysis can be performed using X-rays, magnetic resonance, nuclear medicine or ultrasound [9].

1.1 X-Ray Mammography

X-Ray Mammography is commonly used in clinical practice for diagnostic and screening purposes [2]. Mammography provides high sensitivity on fatty breast and excellent demonstration of micro calcifications; it is highly indicative of an early malignancy.

1.2 MRI of the Breast

Magnetic Resonance Imaging is the most attractive alternative to Mammography for detecting some cancers which could be missed by mammography. In addition, MRI can help radiologists and other specialists determine how to treat breast cancer patients by identifying the stage of the disease [1, 2]. To improve the mass detection performance, it is essential to effectively pre-process mammogram to preserve both the intensity distribution and morphological characteristics of regions.

1.3 Breast Ultrasound

Ultrasound, also known as sonography, uses sound waves to look inside a part of the body. A gel is put on the skin of the breast and a handheld instrument called a transducer is rubbed in the gel and pressed against the skin. The transducer transmits the sound waves through the breast. Echoes from the sound waves are picked up and converted by a computer into a black and white picture that is shown on a computer screen. This test is painless and does not expose you to radiation. Ultrasound has become a valuable tool to use along with mammograms because it is widely available, non-invasive, and costs less than other options.

II. METHODOLOGY

2.1 Pre-processing

Mammograms are medical images that are difficult to interpret. Hence pre-processing is essential to improve the quality. It will prepare the mammogram for the next two process segmentation and feature extraction. Digitization noise and high frequency components in the mammography images are removed by using median filter. [1] The selective median filters have proved to be good because they have some very interesting properties:

1. They can smooth the transient changes in signal intensity (e.g., noise);
2. They are very effective for removing the impulsive noises from the signals;
3. They can preserve the edge information in the filtered signal;
4. They can be implemented by using very simple digital nonlinear operations. Edges are the more important factor in the segmentation of mammogram.

BLOCK DIAGRAM
2.2 Segmentation
The goal of segmentation is to find out the suspicious or abnormal mass region from mammogram. A mass is space occupying lesion and usually appears as a bright region on a mammogram. So contrast enhancement is implemented in order to extract the brighter region.

2.3 Discrete Wavelet Transform (DWT)
The wavelet transform (WT) has gained widespread acceptance in signal processing and image compression. Because of their inherent multi-resolution nature, wavelet-coding schemes are especially suitable for applications where scalability and tolerable degradation are important. Wavelet transform decomposes a signal into a set of basis functions. These basis functions are called wavelets. Wavelets are obtained from a single prototype wavelet $\psi(t)$ called mother wavelet by dilations and shifting; where $a$ is the scaling parameter and $b$ is the shifting parameter [2]

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left( \frac{t-b}{a} \right)$$

Image Separation into four frequency bands:
In algorithm here is shown one level discrete wavelet transform. You can also increase the level of DWT by applying this process more than one time. Second and third level DWT gives the better compression ratio.[2] But it will come with loss of some information. First level DWT is quite reasonable for both achieving high compression ratio and also got quality (less MSE).

2.4 Inverse Discrete Wavelet Transform
Inverse Discrete Wavelet Transform (IDWT) is applied to the image which is obtained at the output of DWT. But before applying IDWT to the image, four frequency bands of the image which are obtained after performing DWT operation are De-noised i.e. any kind of synthetic or natural noise that may appear in an image is removed, using image de noising functions which are present in the MATLAB 7.9 After performing IDWT operation on an image, we get the recomposed image which more clearly reveals the mass containing lesion in the original medical image of a breast.[1]

2.5 K Means Clustering Technique
The main idea behind the k-means algorithm is the minimization of an objective function usually taken up as a function of the deviations between all patterns from their respective cluster centres. The K-means algorithm partitions a dataset into $k$ predefined number of clusters that will try to minimize the intra-cluster distance based on Euclidean distance (Jain et al., 1999). K-means algorithm is very fast and simple algorithm [3]

In statistics and machine learning, k-means clustering is a method of cluster analysis which aims to partition ‘$n$’ observations in to ‘$k$’ clusters in which each observation belongs to the cluster with the nearest mean [7]. For a given set of observation $(x_1, x_2,...x_n)$, where each observation is a d-dimensional real vector, then k-means clustering aims to partition the ‘$n$’ observations in to ‘$k$’ sets $(k<n), \{S_1, S_2,...S_n\}$ so as to minimize the within cluster sum of squares (WCSS) in eqn (1).

$$\text{Arg min} \sum_{i=1}^{k} \sum_{x_i \in S_i} \| x_i - \mu_i \|^2 \ldots (1)$$

Where, $\mu_i$ is the mean of $S_i$. The number of cluster $k$ is assumed to be fixed in k-means clustering.

Standard algorithm:
Given an initial set of k-means in which may be specified randomly or by some heuristic, the algorithm produces by alternating between two steps.

I. Assignment Step
Assign each observation to the cluster with the closest mean (i.e. partition the observation according to the voronoi diagram generated by the means) in equation (2).

$$S_i = \{ x_j : \| x_j - m_i \|^2 \leq \| x_j - m_i \|^2 \} \forall i = 1 \ldots k$$

II. Update Step
Calculate the new means to be the centroid of the observations in the cluster in equation (3)

$$m_i^{(k+1)} = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j \ldots (3)$$

The algorithm is usually very fast, it is common to run it multiple times with different starting conditions. Theoretically it has been shown that there exist certain point sets on which k-means takes super-polynomial time, but practically it is not so far.

a. $\text{K}$ initial "means" (in this case $k=3$) are randomly selected from the data set (shown in colour)
b. $\text{K}$ clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.
c. The Centroid becomes the new means
d. Steps (b) and (c) are repeated until convergence has been reached.

2.6 Fuzzy C Means Clustering
Fuzzy C-means (FCM) is a method of clustering technique which allows one piece of data to belong to two or more clusters. This method was developed by Dunn in 1973 and improved by Bezdek in 1981 and it is frequently used in pattern recognition. [3] FCM clustering is one of well known unsupervised clustering Techniques, which can be used for unsupervised image segmentation. The measurement data considered from an unsupervised fuzzy clustering technique is only used to reveal the underlying structure of the data and segment the image in regions with similar spectral properties, so this method has not relationship between pixels in spatial domain; but it just depends on the spectral domain. [4]

The Fuzzy C-means algorithm, also known as fuzzy ISODATA, is one of the most frequently used
methods in pattern recognition. Fuzzy C-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. It is based on the minimization of objective function to achieve a good classification. ‘\( J_m \)’ is a squared error clustering criterion, and solutions of minimization are least squared error stationary point of \( J \) in equation \( (4) \).

\[
j_m = \sum_{i=1}^{k} \sum_{j=1}^{m} u_{ij} \| x_i - c_j \|^2 \quad (4)
\]

Where \( 1 \leq m \leq \infty \) is any real number greater than 1, is the degree of membership of the cluster \( j \), is the \( d \)-dimensional measured data, is the dimension centre of the cluster and is any norm expressing the similarity between any measured data and the centre. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of member ship \( u_{ij} \) in equation \( (5) \) and the cluster centres’ \( c_j \) by equation \( (6) \).

\[
u_{ij} = \frac{1}{\sum_{k=1}^{N} \left( \frac{||x_i - c_k||}{||x_i - c_k||} \right)} \quad ... (5)
\]

\[
c_j = \frac{\sum_{i=1}^{N} u_{ij} x_i}{\sum_{i=1}^{N} u_{ij}} \quad ... (6)
\]

The iteration will stop when

\[
\max_u | u_{ij}^{k+1} - u_{ij}^k | < \epsilon \quad ... (7)
\]

Where \( \epsilon \) is the termination criterion between 0 & 1, whereas \( k \) is the iteration steps. This procedure converges to a local minimum or a saddle point of \( J_m \).

The fuzzy c means algorithm composed of following steps.

1. Initialize \( U = [u_{ij}] \) matrix, \( U^{(0)} \).
2. At k-step calculate the centre vectors \( C^{(k)} = [C_j] \) with \( U^{(k)} \).
3. Update, \( U^{(k)} \), \( U^{(k+1)} \)

\[
u_{ij} = \frac{1}{\sum_{k=1}^{N} \left( \frac{||x_i - c_k||}{||x_i - c_k||} \right)} \quad ... (5)
\]

4. If \( | u_{ij}^{k+1} - u_{ij}^k | < \epsilon \) then STOP, otherwise return to step 2.

### III. FEATURES EXTRACTION

Texture feature is useful in differentiating normal and abnormal pattern. Texture is an alteration and variation of surface of the image. Texture is characterized as the space distribution of gray levels in neighbourhood. There are two types of texture measures first order and second order. In the first order texture measure are statistics calculated from individual pixel. In second order relationship between Neighbour pixels are considered.

\[
\text{Contrast} = \sum_{i,j=0}^{N-1} P_{ij} (i - j)^2
\]

\[
\text{Energy} = \sum_{i,j=0}^{N-1} (P_{ij})^2
\]

\[
\text{Homogeneity} = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+|i-j|}\cdot
\]

Correlation = \( \sum_{i,j=0}^{N-1} P_{ij} \left( \frac{(i-\mu)(j-\mu)}{\sigma^2} \right) \)

\( P_{ij} = \text{Element } i,j \text{ of the normalized symmetrical GLCM} \)

\( N \) is number of gray levels in the image

The GLCM mean, calculated as:

\[
\mu = \sum_{i,j=0}^{N-1} i P_{ij}
\]

\[
\sigma^2 = \sum_{i,j=0}^{N-1} P_{ij} (i - j)
\]

Where,

\[
\text{Contrast is the contrast between a pixel and its neighbour.}
\]

\[
\text{Energy is the sum of squared elements in SGLD or uniformity.}
\]

\[
\text{Homogeneity is closeness of the distribution of elements in SGLD.}
\]

Correlation shows how correlated a pixel is to its neighbour over the whole image.

### IV. APPROXIMATE REASONING

In the approximate reasoning step the tumour area is calculated using the binarization method. That is the image having only two values either black or white (0 or 1). Here 256x256 jpeg image is a maximum image size. The binary image can be represented as a summation of total number of white and black pixels.

\[
I = \sum_{W=0}^{255} \sum_{H=0}^{255} f(0) + f(1)
\]

\[
\text{Pixels} = \text{Width} \times \text{Height} = 256 \times 256
\]

\( f(0) \) = white pixel (digit 0)

\( f(1) \) = black pixel (digit 1)

No of white pixel ; \( P = \sum_{W=0}^{255} \sum_{H=0}^{255} f(0) \)

Where,

\[
P = \text{number of white pixels (width\cdot height)}
\]

1 Pixel = 0.264 mm

The area calculation formula is

\[
\text{Size of tumour, } S = [\sqrt{P} \cdot 0.264] \text{ mm}^2
\]

\( P \) = no of white pixels; \( W \) = width; \( H \) = height.

![Fig 1. GUI for Tumour Detection](image)
V. RESULTS AND DISCUSSION

In this paper I have presented a novel approach to identify the presence of breast cancer mass and calcification in mammograms using K-means and Fuzzy C-Means clustering for clear identification of clusters. Combining these we have successfully detected the breast cancer area in raw mammograms images. The results indicate that this system can facilitate the doctor to detect the breast cancer in the early stage of diagnosis as well as classify the total cancer affected area. This will help doctor to take or analyze in which stage of cancer the patient have and according to which he/she can take necessary and appropriate treatment steps.

VI. PERFORMANCE EVALUATION OF K-MEANS & FUZZY C-MEANS

The main difference is that, in FCM, each point has a weighted associated with a particular cluster. So, a point doesn’t in a cluster as much as has a weak or strong association to cluster, which is determined by the inverse distance to the centre of the cluster. FCM will tend to run slower than K-means, since it is actually doing more work [6]. Each point is evaluated with each cluster, and more operations are involved in each evaluation. K-means just needs to do a distance calculation, whereas FCM needs to do a full inverse-distance weighting. [10] In this proposed work we make these differences or weakness our strong point for full detection of breast cancer. From this we were able to find out the masses as well as the cancerous area i.e. how far the cancer has affected the breast [5]. A real-time system can be implemented using suitable data acquisition software and hardware interface with digital mammography systems.

<table>
<thead>
<tr>
<th>Image</th>
<th>Feature</th>
<th>Method</th>
<th>Centroid Energy</th>
<th>Centroid Distance</th>
<th>Mean Energy</th>
<th>Mean Distance</th>
<th>PSNR</th>
<th>SNR</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM1</td>
<td>Cancer</td>
<td>K-means</td>
<td>0.025</td>
<td>0.002</td>
<td>0.025</td>
<td>0.025</td>
<td>30.94</td>
<td>30.85</td>
<td>55.85</td>
</tr>
<tr>
<td>CAM2</td>
<td>Cancer</td>
<td>FCM</td>
<td>0.024</td>
<td>0.001</td>
<td>0.024</td>
<td>0.001</td>
<td>33.24</td>
<td>33.24</td>
<td>58.84</td>
</tr>
<tr>
<td>CAM3</td>
<td>Cancer</td>
<td>K-means</td>
<td>0.012</td>
<td>0.001</td>
<td>0.012</td>
<td>0.001</td>
<td>32.45</td>
<td>32.45</td>
<td>55.85</td>
</tr>
<tr>
<td>CAM4</td>
<td>Cancer</td>
<td>FCM</td>
<td>0.013</td>
<td>0.001</td>
<td>0.013</td>
<td>0.001</td>
<td>33.24</td>
<td>33.24</td>
<td>58.84</td>
</tr>
<tr>
<td>CAM5</td>
<td>Cancer</td>
<td>K-means</td>
<td>0.021</td>
<td>0.001</td>
<td>0.021</td>
<td>0.001</td>
<td>32.45</td>
<td>32.45</td>
<td>55.85</td>
</tr>
<tr>
<td>CAM6</td>
<td>Cancer</td>
<td>FCM</td>
<td>0.023</td>
<td>0.001</td>
<td>0.023</td>
<td>0.001</td>
<td>33.24</td>
<td>33.24</td>
<td>58.84</td>
</tr>
<tr>
<td>CAM7</td>
<td>Cancer</td>
<td>K-means</td>
<td>0.020</td>
<td>0.001</td>
<td>0.020</td>
<td>0.001</td>
<td>32.45</td>
<td>32.45</td>
<td>55.85</td>
</tr>
<tr>
<td>CAM8</td>
<td>Cancer</td>
<td>FCM</td>
<td>0.019</td>
<td>0.001</td>
<td>0.019</td>
<td>0.001</td>
<td>33.24</td>
<td>33.24</td>
<td>58.84</td>
</tr>
</tbody>
</table>

Table I Result of feature extraction process

REFERENCES
