A Comparison of Cluster Based Segmentation Methods for Medical Image

R. Seetha¹, S. Santhosh baboo²
Research Scholar¹, Associate Professor²
P.G. & Research Department of Computer Science and Applications
Dwarka Doss Goverdhan Doss Vaishnav College
Arumbakkam, Chennai – 600106, TN, India

Abstract: Mammogram is the x-ray of a breast, which is the most effective way for early diagnosis of breast cancer. Identifying the accurate breast region with tumor is a challenging problem and finding the appropriate technique for the same is tough. In this paper, various cluster based segmentation techniques are compared to find the suitable method to segmenting the area of interest, which can be used for further analysis of tumor.

Keywords - cluster; segmentation; region; FCM; K-Means; EM

1. INTRODUCTION

Image segmentation is a process of partitioning an image into coherent objects and grouping together of similar-looking pixels for efficiency of further processing. It can be done by either partitioning the region with coherent internal property or by grouping the coherent tokens in an image. Various methods namely Segment foreground from background, Histogram-based segmentation, Segmentation as clustering, Graph-theoretic segmentation, Interactive segmentation are available for performing segmentation on an image.

This paper deals with the process of segmenting the preprocessed medical mammogram breast image. There are different image segmentation techniques, like pixel based; edged based; neural network based and cluster based. This paper deals with one of the most commonly used efficient pixel based segmentation method based with clusters[1][6]. Image segmentation using clustering [2][3] can be divided into different types like Fuzzy C-means, EM and K-means clustering. In this paper, a best method for segmentation of mammogram image is found, which makes the image easier for further analysis.

2. REVIEW OF LITERATURE

Khaled Hammouda has done a survey on different techniques of an image. It implemented different data clustering technique- k-means, FCMs clustering. It analyze these different technique, it is found that k-means and Fuzzy c-means are preferred when the number of cluster is known, and in such cases Mountain clustering is not usually used because of high number of dimension due to its exponential proportionality to the dimension of the problem. But if the number of cluster is not known, Mountain clustering method is used. Again it is conclude that subtractive clustering is more advantages than Mountain clustering. Lastly it is stated that these different technique can be used in conjunction with other neural or fuzzy system so that it can improve the system performance.

H.D. Cheng, X.H.Jiang, Y.Sun and Jingli Wang did a literature survey on image segmentation technique. They discussed about different approaches of segmentation of monochrome images.

Ibrahim A. Almerhag, Idris S Feghi and Ali A Dulla proposed a new method of k-means clustering algorithm; it proposed a new modified method of conventional k-means algorithm. In conventional k-means algorithm, usually cluster centers are randomly initialized. But here it intruded a new method to initialize the initial cluster centers. It uses minimum and maximum data points in the given data set to initialize the cluster centers.

3. METHODOLOGY

The proposed approach is depicted in Figure-1. The following are the steps of the proposed segmentation technique based on clusters

- Input an image
- Preprocessed Image
- Clustering
- Segmentation
- Performance Evaluation
Figure-1: Proposed Methodology for cluster based Segenting

This research work chooses the digital mammogram image, which is the most effective way for early prediction of tumor, for the clustering and segmenting process in order to identify sustainable and unsustainable region in the selected image. This research work uses breast mammogram images collected from normal and abnormal patient at Aarthi Scans, Thiruvarur. The breast images in DICOM format is taken for analysis. The DICOM is abbreviated as digital imaging and communications in medicine. This improvement of the digitally stored image is prepared with the help of the MATLAB software. The chosen images are pre-processed for the purpose of removing the noise and enhancing the image[5]. The pre-processed image is used in this paper for clustering.

4. CLUSTERING TECHNIQUES

Clustering is a method to divide or assemble a set of data into a specific number of groups[7]. The most commonly used techniques are FCM, k-means and EM clustering method.

4.1. FCM CLUSTERING

Fuzzy C-Mean (FCM) is an unsupervised clustering algorithm that has been applied to wide range of problems involving feature analysis and clustering design. FCM has a wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, and target recognition. With the developing of the fuzzy theory, the fuzzy c-means clustering algorithm based on Ruspini fuzzy clustering theory was proposed in 1980s [16][11].

An image can be represented in various feature spaces, and the FCM algorithm classifies the image by grouping similar data points in the feature space into clusters. This clustering is achieved by iteratively minimizing a cost function that is dependent on the distance of the pixels to the cluster centers in the feature domain. The pixels on an image are highly correlated, i.e. the pixels in the immediate neighborhood possess nearly the same feature data[11]. Therefore, the spatial relationship of neighboring pixels is an important characteristic that can be of great aid in imaging segmentation

FCM Clustering algorithm is frequently used in pattern recognition. FCM assigns membership to each data point on the basis of distance between the cluster and the data point. The algorithm is based on minimization of the objective function

\[
\sum_{i=1}^{n} \sum_{j=1}^{c} U_{ij}^m \|x_i - c_j\|^2
\]

Where \( n \) the number of data is, \( c \) is the number of clusters, \( U_{ij} \) is the degree of membership of \( x_i \) in the cluster \( j \).

There are four main steps in this approach for segmenting the color images

Step 1. Preprocessed Image acquisition.
Step 2. Clustering by using Fuzzy Clustering Method.
Step 3. Segmentation process: Assign label to highest membership value in each column of the partition matrix \( U \) that indicates a data point belongs to which cluster center. Find the segmented images from the labeled matrix.

The FCM clustering process has been carried out with variety of images.

4.2. K-MEANSCLUSTERING

Among the hard clustering algorithms, K-Means algorithm [4][12] is always researcher’s first choice because of its simplicity and high performance ability. It partitions a collection of data into a k number group of data [13]. It classifies a given set of data into k number of disjoint cluster. K-means algorithm consists of two separate phases. In the first phase it calculates the k centroid and in the second phase it takes each point to the cluster which has nearest centroid from the respective data point. There are different methods to define the distance of the nearest centroid and one of the most used methods is Euclidean distance.

Once the grouping is done it recalculate the new centroid of each cluster and based on that centroid, a new Euclidean distance is calculated between each center and each data point and assigns the points in the cluster which have minimum Euclidean distance[14]. Each cluster in the partition is defined by its member objects and by its centroid. The centroid for each cluster is the point to which the sum of distances from all the objects in that cluster is minimized. So K-means is an iterative algorithm in which it minimizes the sum of distances from each object to its cluster centroid, over all clusters.
Here, K is the numbers of clusters to be specified. The formal steps involved in this algorithm are

1. Choose K initial centroids.
2. Determine the distance of each data item to the centroid.
3. Form k clusters by assigning the data items to the closest centroid.
4. Re-compute the centroid of each cluster.
5. Repeat step 3 and step 4 until the centroids do not change their positions.

The aim of the K-Means algorithm is to minimize the squared error function [4]:

\[ J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i' - c_j\|^2 \]

Here, \( \|x_i' - c_j\|^2 \) is a chosen distance measure between a \( x_i' \) data point and the cluster center \( c_j \).

### 4.3. EXPECTATION MAXIMIZATION CLUSTERING

Expectation Maximization (EM) is one of the most common algorithms used for density estimation of data points in an unsupervised setting. The algorithm relies on finding the maximum likelihood estimates of parameters when the data model depends on certain latent variables. In EM, alternating steps of Expectation (E) and Maximization (M) are performed iteratively till the results converge. The E step computes an expectation of the likelihood by including the latent variables as if they were observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the last E step [17]. The parameters found on the M step are then used to begin another E step, and the process is repeated until convergence.

Mathematically for a given training dataset \( \{x_1, x_2, x_3, \ldots, x_m\} \) and model \( p(x, z) \), where \( z \) is the latent variable, we have

\[ l(\theta) = \sum_{i=1}^{m} \log(p(x_i; \theta)) \]

As can be seen from the above equation, the log likelihood is described in terms of \( x, z \), and \( \theta \). But since \( z \), the latent variable is not known, we use approximations in its place. These approximations take the form of E & M steps mentioned above and formulated mathematically below:

**E-Step.** for each \( i: Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta) \)

**M-Step.** for all \( z: \theta := \arg \max \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log P(x^{(i)}, z^{(i)}; \theta) / Q_i(z^{(i)}) \)

where \( Q_i \) is the posterior distribution of \( z^{(i)} \), given the \( x^{(i)} \).

Conceptually, the EM algorithm can be considered as a variant of the K-Means algorithm where the membership of any given point to the clusters is not complete and can be fractional.

### 5. SEGMENTATION

Segmentation is division of an image into meaningful structures, which is an essential step in image analysis, object representation, visualization, and many other image processing tasks. There are many methods for segmenting an image that have been used researchers. Image segmentation is the first step in image analysis and pattern recognition [3]. It is a vital and important component of image analysis technique, plays one of the hardest roles in image processing, and decides the quality of the final result of examination [2]. It is the process of dividing an image into different homogeneous regions. Image segmentation is to cluster pixels into sustainable and unsustainable regions in an image, i.e., regions corresponding to individual pixels based on surfaces, objects, or natural parts of the objects [7]. Several techniques are available that are quite popular, important, and are regularly used for image segmentation. This paper deals with pixel-based segmenting technique.

After performing the clustering process on the image, we can use pixel-based segment process based on Gaussian algorithm. This is a simplest segmentation method based on the Bayes decision theory in pattern recognition. The gray level histogram of the image is computed and then two component densities are extracted (corresponding to the object and the background) from the mixture density associated with the histogram [14]. It is commonly assumed that both the background and the object densities are Gaussian.
Algorithm

Step 1. Compute the mean $\mu$ and standard deviation $\sigma$ of the histogram

$$\mu = \frac{1}{N} \sum_i F(i) \times i$$

$$\sigma = \sqrt{\frac{1}{N} \sum_i F(i) \times (i - \mu)^2}$$

Step 2. Find a least-squares fit of the histogram $f(i)$ by adjusting the parameters $p_1, \pi_1, \sigma_1, p_2, \pi_2, \sigma_2, \ldots$

Step 3. After the parameters of the mixture density have been estimated, a pixel with gray level $x$ is assigned to the Object.

Images can be segmented into regions by the clustering algorithm.

6. RESULT AND DISCUSSION

The mammogram image with noise is taken as the input image and it has been pre-processed in order to remove the noise[5]. The preprocessed image is taken as the input for the clustering process. In this paper, three different clustering methods are considered namely FCM, K-means and EM. These methods provide sustainable and unsustainable pixels, among them, only the sustainable pixels are clustered in this work. Three clusters are formed for each of the methods and their corresponding outputs are shown in Table 2. In Table 3, the clustered pixel values were mentioned along with the average time taken for calculating them.

Table-1: Input Images

<table>
<thead>
<tr>
<th>Input</th>
<th>With Noise</th>
<th>Preprocessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table-2: Segmented Images for various clustering Technique

<table>
<thead>
<tr>
<th>Cluster Value</th>
<th>Cluster Value</th>
<th>Cluster Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table-3: Clustered Pixel and average time runtime for various techniques

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Clustered Pixels</th>
<th>Timing (in secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>101656 8036 3421</td>
<td>3.834</td>
</tr>
<tr>
<td>K-Means</td>
<td>57721 4097 2151</td>
<td>1.620</td>
</tr>
<tr>
<td>EM</td>
<td>14055 3421 1378</td>
<td>5.043</td>
</tr>
</tbody>
</table>

The average time taken by FCM, K-Means and EM are 3.834 Seconds, 1.620 Seconds and 5.043 Seconds respectively. Figure 2 shows the average clustered timing for the three clustering methods and Figure 3 portrays the Pixel Count for FCM, K-Means and EM. Among the three clustering methods used, K-Means gives the better result compared to the other two methods based on the timing and pixel count.
7. CONCLUSION

The noise removed pre-processed image is used as the input data for the segmentation process. In this paper, the segmentation is done by the above discussed methods, which yields sustainable pixels based on different cluster values. For better segmentation of the pixels, three different clustering algorithms are used and the comparison is made to find out the best suited among them for the mammogram image. From the result, it is found that K-Means method is giving the best clusters, based on the average timing and the pixel count value. The resultant segmentation is used in future for the classification process of mammogram image.

REFERENCES


