Breast Cancer Detection in Mammograms based on Clustering Techniques- A Survey

R.Ramani
Assistant Professor / ECE
V.M.K.V Engineering College,
Salem, Tamilnadu, India

S. Valarmathy
Associate Professor / ECE
V.M.K.V Engineering College,
Salem, Tamilnadu, India

N.Suthanthira Vanitha,
PhD.
Professor / Head / EEE,
Knowledge institute of technology
Salem, Tamilnadu, India

ABSTRACT
Cancer is one of the most leading causes of deaths among the women in the world. Among the cancer diseases, breast cancer is especially a concern in women. Mammography is one of the methods to find tumor in the breast, which is helpful for the doctor or radiologists to detect the cancer. Doctor or radiologists can miss the abnormality due to inexperience's in the field of cancer detection. Segmentation is very valuable for doctor and radiologists to analysis the data in the mammogram. Accuracy rate of breast cancer in mammogram depends on the image segmentation. This paper is a survey of recent clustering techniques for detection of breast cancer**.** These fuzzy clustering algorithms have been widely studied and applied in a variety of application areas. In order to improve the efficiency of the searching process clustering techniques recommended. In this paper, we have presented a survey of clustering techniques.

Keywords: Clustering, Mammogram, Image segmentation, k-means, fuzzy c-means, modified fuzzy c-means, Kernelized Fuzzy C-Means, modified Kernelized Fuzzy C-Means, Hierarchical Clustering.

I.INTRODUCTION
Segmentation is the process of partitioning a digital image into several segments based on pixels. It is a vital and essential part of image examination system. The main process is to represent the image in a clear way. The result of image segmentation is a collection of segments which combine to form the entire image [1]. Real world image segmentation problems actually have multiple goals such as minimize overall deviation, maximize connectivity, minimize the features or minimize the error rate of the classifier etc [2]. Segmentation algorithms can be classified into different categories based on segmentation techniques used such as the features thresholding [3], template matching [4], region based technique and clustering. Each technique has their own limitations and advantages in terms of suitability, performance and computational rate. The low-level segmentation techniques are fast and simple, but these methods simply analyze an image by reducing the quantity of the data to be processed. This problem can result in loss of important information. Clustering is the process of identify group of similar image primitive [5]. Clustering techniques can be classified into supervised clustering demands human interaction to decide the clustering criteria and the unsupervised clustering decides the clustering criteria by itself. Supervised clustering includes hierarchical approaches such as relevance feedback techniques [6], [7] and unsupervised clustering includes density based clustering methods. These clustering techniques are made to perform image segmentation. A variety of clustering techniques have been introduced to make the segmentation more effective. The clustering techniques which are included in this paper such as relevance feedback [8], log based clustering [9], hierarchical clustering [10], graph based, retrieval-dictionary based, filter based clustering etc.

2. CLUSTERING ALGORITHMS
The main purpose of clustering is to divide a set of objects into significant Groups. The clustering of objects is based on measuring of correspondence between the pair of objects using distance function. Thus, result of clustering is a set of clusters, where object within one cluster is further similar to each other, than to object in another cluster. The Cluster analysis has been broadly used in numerous applications, including segmentation of medical images, pattern recognition, data analysis, and image processing. Clustering is also called data segmentation in some applications because clustering partitions huge data sets into groups according to their resemblance.

2.1 K –means clustering segmentation
The K-means algorithms are under the group of squared error based clustering. The k means algorithms are an iterative technique which is used to split an image into k clusters. In statistics and machine learning, k means clustering is a method of cluster analysis which can to portions n observations into k cluster, in which each observation be in the right place to the cluster with the adjacent mean [11],[12]. The basic k means clustering algorithms as follows, i) pick k cluster centre either randomly or based on some heuristic, ii) assign each pixel in image to the come together that minimum the distance between the pixels cluster centre.iii) re-compute the cluster centre's by averaging all of the pixels in the cluster. Repeat last two steps until convergences are attained. K-means clustering key endeavor to partitions the n observation into k sets (k<n) s={s1,s2,s3,…..sk} so as to minimize the within cluster sum of squares.

$$arg\ min \ \sum_{i=1}^{k} \cdot \ \sum_{xj\ \epsilon\ si} \left\| x_j - u_i \right\|^2 \ - - - 1$$

Where μ_i is the mean of points in Si. The majority of the common algorithms use an iterative refinement method. Due to its ubiquity it is often called the k-means algorithms. The k-means algorithm is exceptionally simple and can be

implemented in solving many practical problems. But the performance of the k-means algorithm depends on the initial positions of the cluster centers.

2.2 Adaptive K-means Clustering Algorithm

In [13], an adaptive k-means clustering algorithm implemented for the breast image segmentation for detection of micro calcifications and also a computer based decision system for early detection of breast cancer. This method is implemented to improve the performance of existing K-means approach by varying a variety of values of certain parameters discussed in the algorithm [14], [15], [16]. The modified form of k-means algorithm is called an Adaptive K-means Clustering Algorithm. The modified form of k-means algorithm as follows, the histogram is a count of data points falling in a variety of ranges. The effect is uneven approximation of the frequency distribution of data. The collection of data is called classes, and in framework of histogram they are known as bins, because one can think of them as containers that accumulate data and fill up at a rate equal to the frequency of that data class. The shape of the histogram sometimes is predominantly sensitive to the number of bins. If the bins are too wide, essential information might get absent. By reducing the number of bins and increasing the number of classes in the K-means algorithm, the detection accurateness is found to be increasing. Quantization in terms of color histograms refers to the process of dropping the number of bins by taking colors that are very similar to each other and putting them in the same bin. Perceptibly quantization reduces the information concerning the content of images but as was mentioned this is the swapping when one wants to reduce processing time. Lastly, the detection accuracy was predictable and compared the performance with previous similar research works emphasizing the detection accuracy values. The accuracy of detection has increased [13].

2.3 Fuzzy C-means Algorithm

The fuzzy c means algorithm also referred to as fuzzy ISO data is one of the most habitually used methods in pattern recognition, fuzzy c means is a method of clustering which follows one piece of data to belong to two or more clusters [17]. This method was developed by Dunn in 1973 and enhanced by Bezdek in 1981 and it is recurrently used in pattern recognition, classification, medical image segmentation, etc. FCM is an iterative algorithm, which is used to find cluster centre that minimize a dissimilarity function. FCM uses fuzzy partition such that a given data point can belong to several groups. To achieve a high-quality classification is a squared error clustering criterion and solutions of minimization are at least squared error immobile point of J in the following equation.

$$J_m = \sum_{i=1}^{k} \cdot \sum_{j-i}^{c} \mu_y^x \|xi - cj\|^2 ----2$$

Fuzzy portioning is carried out through an iterative optimization of the objective. The clustering method for both k means and FCM is same but in k means algorithm when it cluster , it takes the mean of the weighted cluster so as to easy to identify masses or the origin point of cancer or tumor. In FCM, it considers that each point has weighted value associated with cluster. To find out how much breast cancer has spread out, this technique helped to doctors or radiologistic [17]. The performance is based on initial cluster centers. FCM also suffers from the presence of outliers and noises so that it is not easy to identify the initial partitions. FCM gives better results than hard k-means algorithm.

2.4 Kernelized Fuzzy C-means Algorithm

In [18], Kernelized FCM algorithms (KFCM) are implemented with spatial constraints on the objective function that could improve the image segmentation. The kernel methods [19] are one of the a large amount researched subjects within machine learning community in the last few years and widely have been applied to pattern recognition and function approximation. The kernel methods consist of: (i) a class of robust non Euclidean distance measures for the original data space to derive novel objective functions and thus clustering the non-Euclidean structures in data. (ii) Enhancing robustness of the original clustering algorithms to noise and outliers, and (iii) still retaining computational straight forwardness. The algorithm is realized by modifying the objective function in the conventional fuzzy c-means (FCM) algorithm using a kernel-induced distance instead of Euclidean distance in the FCM, and thus the consequent algorithm is derived, called as the kernelized fuzzy c-means (KFCM) algorithm, which is better than FCM. In FCM, the membership matrix U is allowed to have not only 0 and 1 but also the elements with any values between 0 and 1, this matrix satisfies the constraints:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots., N ---3$$

In this work, the kernel function K(x, *C*) is taken as the Gaussian radial basic function (GRBF):

$$K(x,\ c) = \exp\left(\frac{-\|x-c\|2}{\sigma^2}\right) ---4$$

The objective function is given by:

$$J_m = 2\sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^m (1 - K(x_j, c_i)) ---5$$

The fuzzy membership matrix u can be obtained from:

$$u_{ij} = \frac{(1-K(x_j,c_i))^{-1/(m-1)}}{\sum_{k=1}^{c}(1-K(x_j,c_k))^{-1/(m-1)}} ----6$$

$$1 \le m < \propto$$

Since the K-means method aims to decrease the sum of squared distances from all points to their cluster centers, this should result in squashed clusters. We use the intra-cluster distance measure, which is simply the median distance between a point and its cluster centre. The equation is given as:

$$intra = midean\left(\sum_{i=1}^{c}\sum_{x\epsilon c_i}\|x - v_i\|^2\right) --7$$

Therefore, the clustering which gives a lowest value for the validity measure will tells us what the ideal value of the clusters. Then the number of clusters is known before estimating the membership matrix. The algorithms incorporate spatial information into the membership function and the validity process for clustering. They have estimated accurate clusters routinely even without prior knowledge of the true tissue types and the number of clusters of given images. KFCM method will be obtained adequate result, which is more compatible with human visual perception.

2.5 Multiple kernel fuzzy c-means (MKFC)

The fuzzy c-means is a popular soft clustering method; its effectiveness is mainly limited to spherical clusters. Applying kernel behavior, the kernel fuzzy c-means algorithm attempts to address this problem by mapping data with nonlinear relationships to correct feature spaces. Kernel combination, or choice, is critical for efficient kernel clustering. Regrettably, for most applications, and is not easy to find the right combination. In [20], multiple kernel fuzzy c-means (MKFC) algorithm implemented which extends the fuzzy c-means algorithm with a multiple kernel learning location. By incorporating multiple kernels and automatically adjusting the kernel weights, MKFC is more vulnerable to hopeless kernels and unrelated features. This makes the choice of kernels less crucial. In addition, show multiple kernel k-means (MKKM) to be a special case of MKFC. Experiments on both artificial and real-world data demonstrated the efficiency of multiple kernel fuzzy c-means algorithm [20]. The kernel based segmentation technique, particularly for images with small resolution and reduced contrast. The application of multiple or multiple kernels in the FKCM (fuzzy kernel c-means) has its advantages. In addition to the flexibility in selecting kernel functions, it also offers a new approach to join different information from multiple heterogeneous or homogeneous sources in the kernel space. Specifically, in image-segmentation problems, the input data engage properties of image pixels sometimes derived from very dissimilar sources. Therefore, it can define different kernel functions purposely for the intensity information and the texture information separately, and then combine these kernel functions and apply the composite kernel in MKFCM to obtain better image-segmentation results.

The general framework of MKFCM aims to minimize the objective function

$$Q = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|\emptyset_{com}(x_j) - \emptyset_{com}(o_i)\right\|^2 ---8$$

To enhance the Gaussian kernel based KFCM - F by adding a local information term in the objective function

$$Q = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}(1 - K(x_j, o_i)) + \alpha\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left(1 - k(x_j, o_i)\right) ---9$$

This implemented algorithm is easy to implement and provides soft clustering results that are immune to irrelevant, redundant, ineffective, and unreliable features or kernels. The merits of MKFCM based image segmentation is flexibility in selection and combination of the kernel functions in different shapes and for different types of information, after combining the different kernels in the kernel space, there is no need to change the computation procedures of MKFCM, this is another advantages to reflect and fuse the image information from multiple heterogeneous or homogeneous sources [21].

2.6 Hierarchical Clustering

The well- known technology in information retrieval is hierarchical clustering [22]. The process of integrating different images and building them as a cluster in the form of a tree and then developing step by step in order to form a little cluster. The steps involved in this process are as follows: the images from a variety of databases are divided into X-sorts. The classification will be calculated by modifying the cluster centers, sorts of the images and stored in the form of matrix N*N continuously which also includes dissimilarity values. At first it calculates the similarities between the queried image and the retrieved image in the image database. Secondly, it identify the similaritie between two closest images (In N*N matrix) and integrate them to form a cluster. Finally all the similarities are grouped to shape a single cluster.

2.7 Wavelet based K-means Algorithm

In [23], wavelet transformation and K – means clustering algorithm are developed for cancer detection in mammograms. The first step is to carry out image segmentation. It allows distinctive masses and micro calcifications from surroundings tissue. The wavelet transformation and K- means clustering algorithm have been used for intensity based segmentation. This method is strong against noise. The processed image is added to the original image to get the sharpen image. Then K-means algorithm is applied to the sharpened image in which the cancer region can be placed using the thresholding method. The combination of noise-robust nature of applied processes and the simple K-means algorithm gives greater results.

The Wavelet based K-means Algorithm various has steps as follows

1. **Step1:** Apply mammogram as an input.

2. **Step2:** Apply wavelet transform on mammogram to attain wavelet decomposed image resulting in four sub bands. These are the LL (Lower resolution version of image), LH (Horizontal edge data), HL (Vertical edge data), & HH (Diagonal edge data) sub bands representing approximation, horizontal, vertical and diagonal components in the form of coefficients, respectively. LL sub band contains low level and the other three (LH, HL, and HH) contain high level details.

3. **Step3:** locate estimate coefficients in LL equal to zero and apply inverse wavelet transform to get a high pass image from the enduring (horizontal, vertical and diagonal) sub bands. We call the resultant image level-1 (L1) detail image.

4. **Step4:** Add L1 to the original image to get a sharpened image.

5. **Step5:** Apply K-means algorithm for segmentation of sharpened image.

6. **Step6: T**hresholding method is applied to detect breast cancer in mammogram.

We apply Discrete Wavelet Transform (DWT) to mammogram images, because wavelets provide frequency information as well as time-space localization. In addition, their multi-resolution character enables us to envisage image at various scales and orientations. The multi-resolution property provides information about various high frequency components at dissimilar levels of decomposition. Over-decomposition should however be avoided, because as the decomposition levels increase, there is a huge risk that lower frequencies become a part of detail components. This may restrict us to use only fewer level of decomposition because lower frequencies will become part of high pass image and decrease effective detail in an image. The Wavelet transform equipped the algorithm noise free because wavelets provide frequency information as well as time-space localization. Then k-means was applied to segment the mammogram. K-means provides a simple and efficient technique of segmentation. In [23], Proved the result is better by comparing with other methods.

2.8 Fuzzy k-c-means Clustering Algorithm

There are several methods available for medical image segmentation such as Clustering methods, Thresholding method, Classifier, Region Growing, Deformable Model, Markov Random Model etc. The specifically k-means and fuzzy c-means clustering algorithms are used for breast cancer image segmentation. These algorithms were combined to come up with another technique called fuzzy k-c-means clustering algorithm[24], which is improved result in terms of time utilization. In Fuzzy k-c-Means the concentration is on making the number of iterations equal to that of the fuzzy c means, and still get a most favorable result. The Fuzzy K-C-Means algorithm has the following steps:

- Read the image into the Matlab environment
- Try to make out the amount of iteration it might possibly do within a given period of time
- Decrease quantity of iteration with distance check
- Obtain the size of the image
- Calculate the distance promising size using repeating structure
- Concatenate the given dimension for the image size
- Repeat the matrix to generate huge data items in carrying out probably distance calculation
- Decrease repeating when possible distance has been attained
- Iterations begin by identifying large component of data the value of the pixel
- Iteration stops when feasible identification elapses
- Time is generated.

The Fuzzy K-C-Means is a method obtained from both fuzzy c-means and k-means but it carries more of fuzzy c-means properties than that of k-means. Fuzzy k-c-means works on gray scale images like fuzzy c-means generates the same number of iterations as in fuzzy c-means. Both fuzzy c-means and k-means are competing in terms of time; fuzzy k-c-means has been programmed to produce the same number of iteration with fuzzy c-means with a quicker operation time. That is fuzzy k-c-means is faster than both fuzzy c-means and k-means. In terms of accuracy, the number iteration is put into consideration. There more iterations and accuracy. Results have been analyzed and recorded [24].

3. CONCLUSION

In this paper, we have accomplished a partial survey of various clustering techniques used for image segmentation. The image segmentation based clustering algorithms can be done in a valuable way. Clustering techniques may help to enhance the efficiency of the image recovery process. All the clustering techniques may obtain satisfaction results but not able to produce 100 % of accuracy. The image segmentation remains a challenging problem in medical image processing, computer vision and still an imminent problem in the world. Further works, we planned to develop a novel efficient clustering technique to produce more accuracy than existing methods for the detection of breast cancer.

REFERENCES

[1] C.Harris and M.Stephens, "A Combined Corner and Edge Detection," Proc.Fourth Alvey Vision Conf., pp.147-151, 1988.

[2]Shirakawa, S., and Nagao, T., "Evolutionary Image Segmentation Based on Multiobjective Clustering". Congress on Evolutionary Computation (CEC '09), Trondheim, Norway, 2466-2473, 2009.

[3] K. S. Chuang., H. L. Tzeng., S. Chen., J. Wu., and T. J. Chen., "Fuzzy C-Means Clustering with Spatial Information for image Segmentation," *Comput. Med. Imaging Graph*, vol. 30, no. 1, pp. 9–15, Jan. 2006.

[4] R. Szeliski., D. Tonnesen., and D. Terzopoulos., "Modeling Surfaces of Arbitrary Topology with Dynamic Particles", *In: Proceedings of CVPR*, pp. 82–87, 1999.

[5] Puzicha, J., Hofmann, T. and Buhmann, J. M.,"Histogram Clustering for Unsupervised Image Segmentation"., Computer Vision and Pattern Recognition , Vol.2.IEEEpress,602-608, 2000.

[6] Zhou XS, Huang TS. Relevance feedback in image retrieval: A comprehensive review. Multimedia Syst;8:536-544, 2003.

[7] Chundi, P., Dayal, U., Sayal, M., Hsu, M: US20077181678, 2007.

[8] Wang JZ, Li J, Wiederhold G. Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Trans pattern Analysis Machine Intell;23:947-963, 2001.

[9] Jin J, Kurniawati R, Xu G, Bai X. Using browsing to improve content-based image retrieval. J Visual Common Image Represent; 12:123-135, 2001.

[10] Huang Min,Sun bo,Xi Jianqing"An Optimized image retrieval method based on Hierarchical clustering and genetic algorithm"I'ntl forum on Information technology and applications,978-0-7695-3600-2/09-IEEE,2009.

[11] T.Kanungo, D.M Mount, N.Netanyahu, C.Piatko,R.Silverman and A.Y.wa(2002),an efficient k-means clustering algorithms,analysis and implementation proc.IEEE conference computer vision and pattern recognition pp.881-892.

[12] Bhagwati charanpatel, Dr.G.R.sinha, an adaptive k-means cluster algorithms for breast image segmentation, international journal of computer applications(0975-8887),vol 10-n 4 ,nov-2010 .

[13] Bhagwati Charan Patel, Dr. G.R.Sinha, "An Adaptive K-means Clustering Algorithm for Breast Image Segmentation", International Journal of Computer *Applications (0975 – 8887), Volume 10– N.4, November 2010.*

[14] Jianbo Shi & Jitendra Malik (1997) Normalized Cuts and Image Segmentation, Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 731-737.

[15] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, & A. Y.Wu (2002) An efficien k-means clustering algorithm: Analysis and implementation Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.881-892.

[16] Lloyd, S. P. (1957). Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, vol. **28** (2), p. 129–137.

[17] An edge detection method for microcalcification cluster in mammogram by yu guang zhang wen Lu Fu yun cheng LI song taishan med.univ.taian,china, in second international conference on bio-medical engineering and informatics 2009.

[18] E.A. Zanaty Sultan Aljahdali Narayan Debnath, "A Kernelized Fuzzy C-means Algorithm for Automatic Magnetic Resonance Image Segmentation.

[19] D.Q. Zhang, and S.C. Chen, "A novel kernelized fuzzy c-means algorithm with application in medical image segmentation," Artif. Intell. Med, vol.32, pp.37–50, 2004.

[20] Hsin-Chien Huang, Yung-Yu Chuang and Chu-Song Chen, "Multiple Kernel Fuzzy Clustering'', IEEE transactions on fuzzy systems. June 16, 2011.

[21] long chen, C.L.Philip chen, " a multiple kernel fuzzy c means algorithms for image segmentation', IEEE Transation on system.man.and cyberrenetics-part b,cybernetics.feb'9 ,2011.

[22] Huang Min,Sun bo,Xi Jianqing"An Optimized image retrieval method based on Hierarchical clustering and genetic algorithm"I'ntl forum on Information technology and applications,978-0-7695-3600-2/09-IEEE,2009.

[23] Shruti Dalmiya, Avijit Dasgupta, Soumya Kanti Datta, "Application of Wavelet based K-means Algorithm in Mammogram Segmentation", International Journal of Computer Applications (0975 – 8887) Volume 52– No.15, August 2012.

[24] Ajala Funmilola A*, Oke O.A, Adedeji T.O, Alade O.M, Adewusi E.A, "Fuzzy k-c-means Clustering Algorithm for Medical Image Segmentation", Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online), Vol 2, No.6, 2012

.