A Comparative Study on K-Means and Fuzzy C-Means Algorithm for Breast Cancer Analysis

A Sathish
Department of Computer Science
PSG College of Arts & Science
Coimbatore.

J Mohana Sundaram
Department of Computer Science
PSG College of Arts & Science
Coimbatore.

***Abstract*- Fuzzy C-Means is a method of clustering which allows one piece of data to belong to two or many clusters. This method is frequently used in pattern recognition. It is based on minimizing functions. Fuzzy Partitioning is carried out through an interactive optimization of the objective function, with the update of membership the cluster centers. Fuzzy C-means is one of them and it is used widely in such applications as a clustering algorithm. In this study, we applied a different clustering algorithm, an artificial immune system (AIS), for data reduction process. We realized the performance evaluation experiments on standard Chainlink and Iris datasets, while the main application was conducted by using Wisconsin Breast Cancer dataset and Pima Indians dataset which were taken from the UCI Machine learning repository. For these datasets, the performance of AIS in data reduction process was compared with fuzzy c-means clustering algorithm in which Multilayer Perceptron (MLP)-Artificial Neural Networks (ANN) was used as a classifier after the data reduction processes.**

Keywords- K-mean, Fuzzy C-Means, C-mean, Breast image, segmentation, detection, CAD

I. INTRODUCTION

With today's improving technology are very high level, Data recording opportunities are also expanding and providing lots of ways for information flow. For large datasets, data mining techniques are affected in three ways: computing time, predictive or descriptive accuracy and representation of the data mining model. Thus some preliminary data pre-processing steps should be conducted before mining in data. Several approaches can be taken into consideration for data reduction for example random sampling of current dataset. Clustering is another alternative to reduce the number of samples by taking only cluster representative sample for all samples in a cluster.

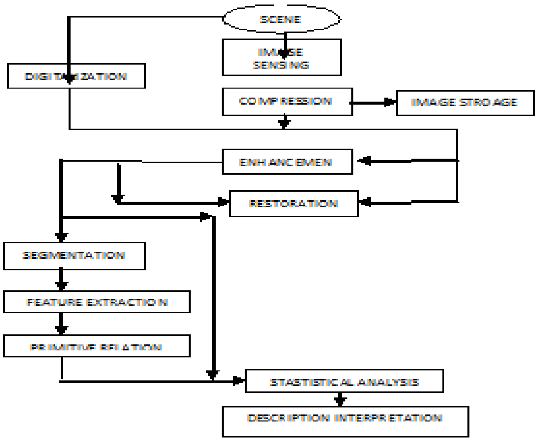


Figure 1: Schematic diagram of different stages of image processing and analysis technique

Breast cancer is most common type of cancer in women, with more than one million cases and nearly half million of deaths occurring worldwide annually [1]. In 2010, there were reported approximately 207090 newly diagnosed cases and 30840 deaths in the United States, and total of 1,638,910 new cancer cases is projected to occur in 2012 [2]. A breast cancer victim's chances for long-term survival are improved by early detection of the disease, and it is very early detection is in turn enhanced by an accurate diagnosis.

II. K-MEANS WITH BREAST CANCER

A breast cancer CAD scheme separates suspicious regions that may contain masses from the background arenchyma – the tissue characteristic of an organ, as distinguished from associated connective or supporting tissues. In other words, such schemes are partition the mammogram into several nonintersecting regions and extract Regions of Interest (ROIs) and suspicious mass candidates from the ultrasound images. While orders of words areas are darker than its surroundings, it has a similar density, a regular shape of variable size. Thus, the image segmentation is essential to maintaining the sensitivity and accuracy of the entire mass detection and classification system.

We have proposed an adaptive K-means segmentation method for detection of micro calcifications in digital mammograms. In the present work, we have made an attempt to improve the performance of existing K-means approach by varying various values of certain parameters discussed in the algorithm.

The K-means algorithm is an iterative technique that is used to partition an image into K clusters. In statistics and machine learning also used, K-means clustering is the one of the most method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The basic algorithm is: Pick K-means cluster centers, either by chance or based on some heuristic.

To assign each pixel in the images to the cluster that minimizes the distance between the pixel and the cluster center; Re-compute the cluster centers by averaging all of the pixels in the cluster repeat last two steps until convergence is attained (e.g. no pixels change clusters. Given a set of observations (x_1, x_2… x_n), where each observation is a d-dimensional real vector, k-means clusters aims to partition the n observations into k sets ($k < n$) $S = \{S_1, S_2 \dots S_k\}$ so as to minimize the Within-Cluster Sum of Squares (WCSS):

Where μ_i is the mean of points in S_i. The most common algorithm uses an iterative alteration technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, particularly it was used in the computer science community. Given an initial set of the k means $m_1 \dots m_{k.}$

1) Which may be specified randomly or by some heuristic, the algorithm takings by alternating between two steps. Assign each observation to the cluster with the closest mean by

2) Calculate the new means to be the centroid of the observations in the cluster.

3) We have modified the algorithm as follows:

The histogram is summary graph showing a count of data points falling in many ranges. The effect is rough approximation of the frequency distribution of data. The group of data is called classes, and in context of histogram they are known as bins, because they 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 particularly sensitive to the number of bins. If the bins are wide, important information might get omitted. By reducing the number of bins and increasing the number of classes in the K-means algorithm, the detection accuracy is found to be increasing. Quantization in terms of color histograms refers to the process of reducing the number of bins by taking colors that are very similar to each other and putting them in the same bin. By default the many number of bins one can obtain using the histogram function is 256. For the purpose of saving time when trying to compare color histograms, it was a one can quantize the number of bins. Observably quantization reduces the information regarding the content of images but as was mentioned this is the tradeoff when one wants to reduce processing time.

III. FUZZY C-MEANS WITH BREAST CANCER

The relative fibroglandular tissue content in the breast, commonly referred to as breast density, has been shown to be the most significant risk factor for breast cancer after age. Currently, the most common approaches to quantify density are based on either semi-automated methods or visual assessment, both of which are highly subjective. This work presents a novel multi-class fuzzy c-means (FCM) algorithm for fully-automated identification and quantification of breast density, optimized for the imaging characteristics of digital mammography. The proposed algorithm involves adaptive FCM clustering based on an optimal number of clusters derived by the tissue properties of the specific mammogram, followed by generation of a final segmentation through cluster agglomeration using linear discriminant analysis. When evaluated on 80 bilateral screening digital mammograms, a strong correlation was observed between algorithm-estimated PD% and radiological ground-truth of $r=0.83$ ($p<0.001$) and an average Jaccard spatial similarity coefficient of 0.62. These results show promise for the clinical application of the algorithm in quantifying breast density in a repeatable manner.

Besides of analysing performance of AIS, we also compared the performance of AIS as a data reduction method on two real world classification problems. They are Diabetes Disease and Breast Cancer classification problems. The related datasets were taken from the UCI data mining repository. In these experimentations however, train & test data partitioning was conducted in a different way. Firstly, the training and testing data were determined and then the training data were reduced with AIS and FCM methods. In this step, data were reduced so that approximate compression ratios of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%and 98% were obtained. Then optimum 13 ANN architectures with optimum parameters were trained with these reduced training data and test data were presented to these trained ANNs to calculate classification ratios. In the same time, these were done two times because 2-fold cross validation method was utilized during the experiments. Let us explain this experimental procedure in more detailed way for the used datasets separately. The Pima Indians Diabetes dataset contains 768 samples taken from healthy and unhealthy persons. 500 of these samples belong to persons with no diabetes problem while the remaining 286 sample are of persons with diabetes. The class information contained in this data set is given by 2 for healthy persons and by 1 for diabetic patients. The number of attributes in samples is 8.

1) FUZZY C-MEANS ALGORITHM

Fuzzy C-means algorithm is also called as ISODATA. It was most frequently used in pattern recognition. Fuzzy C-mean is the method using in clustering. It is using one piece of data to belong to two or more clusters. It always based on minimization of objective functions to achieve a good classification.

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} ||x_i - c_j||^2 \qquad 1 \le m < \infty \tag{1}$$

where m is an real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i_{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function display above, with the updates of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{||x_i - c_j||}{||x_i - c_k||})^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{2}$$

This iteration will stop when $max_{ij}\left\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\right\} < \in,$

Where ε is a termination criterion between 0 and 1, whereas *k* are the iteration steps using in Fuzzy. This procedure converges to a local minimum or a saddle point of *Jm*. The algorithm is generated for the following steps:

1. *Initialize U=[u_{ij}] matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}$=[c_j] with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{3}$$

3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{||x_i - c_j||}{||x_i - c_k||})^{\frac{2}{m-1}}} \tag{4}$$

4. *If || $U^{(k+1)}$ - $U^{(k)}$||< ε then STOP; otherwise return to step 2.*

2) K-MEANS ALGORITHM:

K-Means is a well-known partitioning algorithm used for grouping. Objects are classified as belongings to one of the k groups, the k chosen a priori.

The most common algorithm uses an iterative technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, mainly in the data mining community. These initial set of k means $m_1(1),\dots,m_k(1)$, the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\|^2 \le \left\|x_p - m_j^{(t)}\right\|^2 \forall j, 1 \le j \le k\right\} \tag{5}$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be is assigned to two or more of them.

Update step: This method calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|}\sum_{x_j \in S_i^{(t)}} x_j \tag{6}$$

These was an arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

3) PERFORMANCE EVALUATION OF FUZZY C-MEANS

The most important difference is that in FCM, each has a weighted associated with a specify cluster so, a point doesn’t have cluster as much as a little or more association to cluster. It was very determine by increase distance to the center of the cluster. FCM will tend to run slower than K-means, since it is actually doing more work. Every point is calculated with each cluster, and many operations are involved in each evaluation.

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 the messes and the cancer area.

IV. RESULTS AND DISCUSSIONS

The area details the detections of breast cancer mass and calcification in mammograms using image processing functions. K-means clustering and fuzzy C-means algorithm.

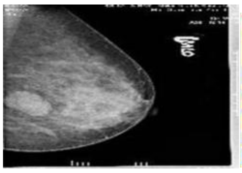


Figure 2: Original breast image

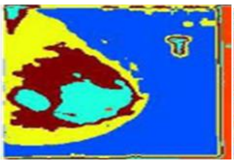


Figure 3: Result from image database

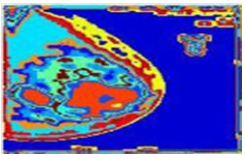


Figure 4: Results for Bins=5

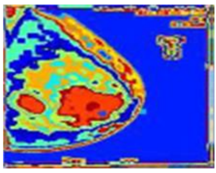


Figure 5: Results for Classes =10.

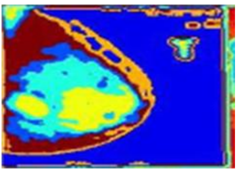


Figure 6: Results for Bins=5, Classes=20.

Figure 6- 9 shows the results for constant value of number of classes and increasing the number of Bins. This can be observed that the affected regions are more accurately located i.e. the identification of affected area with malignant effects gets more prominent.

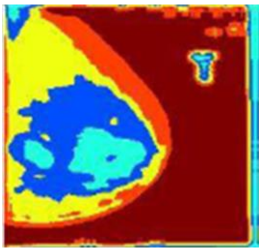


Figure 7: Results for Bins=10 Classes =5

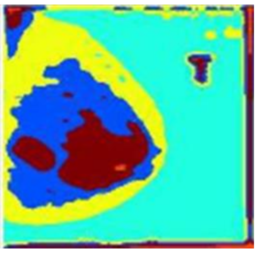


Figure 8: Results for Bins=15 Classes=5

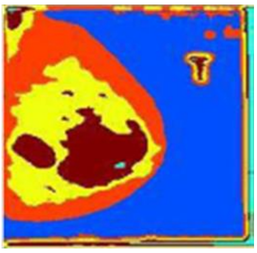


Figure 9: Bins=20,Classes=5.

Finally, the detection accuracy was estimated and compared the performance with previous similar research works emphasizing the detection accuracy values. The results obtained are also in support of anticipation with the findings and diagnosis by a senior radiologist of BSR APPOLO centre of cancer research. The accuracy of detection has increased. Table 1 shows the detection accuracy of future and accessible work.

V. Comparision of Algorithms

The area details the detections of breast cancer mass and calcification in mammograms using image processing functions such as K-means clustering and fuzzy C-means clustering algorithm.

1) Results

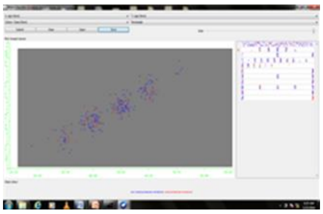
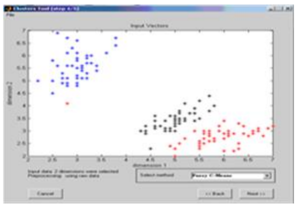


Figure 10: K-Means with Breast Cancer Using Weka



Input data: 2 dimensions were selected
Preprocessing: using raw data

Figure 11: Fuzzy C-Means with Breast Cancer Using Matlab

2) Comparative Analysis of Algorithms

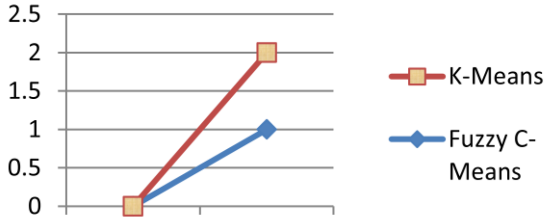
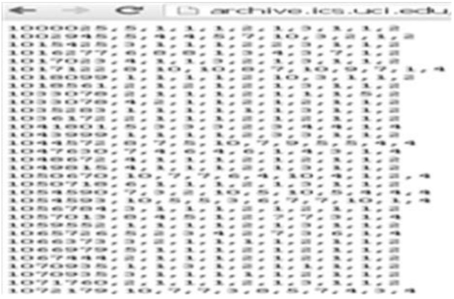


Figure 12: Comparative Analysis



```
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
1015425,3,1,1,1,2,2,3,1,1,2
1016277,6,8,8,1,3,4,3,7,1,2
1017023,4,1,1,3,2,1,3,1,1,2
1017122,8,10,10,8,7,10,9,7,1,4
1018099,1,1,1,1,2,10,3,1,1,2
1018561,2,1,2,1,2,1,3,1,1,2
1033078,2,1,1,1,2,1,1,1,5,2
1033078,4,2,1,1,2,1,2,1,1,2
1035283,1,1,1,1,1,1,3,1,1,2
1036172,2,1,1,1,2,1,2,1,1,2
1041801,5,3,3,3,2,3,4,4,1,4
1043999,1,1,1,1,2,3,3,1,1,2
1044572,8,7,5,10,7,9,5,5,4,4
1047630,7,4,6,4,6,1,4,3,1,4
1048672,4,1,1,1,2,1,2,1,1,2
1049815,4,1,1,1,2,1,3,1,1,2
1050670,10,7,7,6,4,10,4,1,2,4
1050718,6,1,1,1,2,1,3,1,1,2
1054590,7,3,2,10,5,10,5,4,4,4
1054593,10,5,5,3,6,7,7,10,1,4
1056784,3,1,1,1,2,1,2,1,1,2
1057013,8,4,5,1,2,?,7,3,1,4
1059552,1,1,1,1,2,1,3,1,1,2
1065726,5,2,3,4,2,7,3,6,1,4
1066373,3,2,1,1,1,1,2,1,1,2
1066979,5,1,1,1,2,1,2,1,1,2
1067444,2,1,1,1,2,1,2,1,1,2
1070935,1,1,3,1,2,1,1,1,1,2
1070935,3,1,1,1,1,1,2,1,1,2
1071760,2,1,1,1,2,1,3,1,1,2
1072179,10,7,7,3,8,5,7,4,3,4
```

Figure 13: Breast Cancer Dataset from UCI Repository

VI. Conclusion

In the paper there was a comparison between Fuzzy C-Means and K-Means Algorithm. Breast cancer how they differ from in each algorithm we have to display in diagrams. Using Fuzzy C-Mean how the cancer effects and K-Means how they differs that was the main problem for this paper.

References

[1] T.Hieken, J.Harrison, J.Herreros, Velasco, “Correlating sonography, mammography, and pathology in the assessment of breast cancer size,” American Journal of Surgery 182(4) , pp. 351-354, .2001.

[2] I.Saarenmaa, T.Salminen, U.Geiger, P.Heikkinen, S.Hyvrinen, J.Isola, V.Kataja, M.Kokko, R.Kokko, E.Kumpulainen, “The effect of age and density of the breast on the sensitivity of breast cancer diagnostic by mammography and ultasonography,” Breast Cancer Research and Treatment, 67(2) , pp. 117-123, 2001.

[3] W.Pedrycz, J. Waletzky, “Fuzzy clustering with partial supervision,” IEEE Transactions on Systems, Man and Cybernetics,” 27, pp. 787–795, May 1997.

[4] D.T.C. Lai, J.M. Garibaldi, “A comparison of distance-based semi-supervised fuzzy c-means clustering algorithms,” in 2011 IEEE International Conference on Fuzzy Systems, pp. 1580–1586, June 2011.