ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

An Enhanced Mammogram Image Classification Using Fuzzy Association Rule Mining

Dr.K.Meenakshi Sundaram [1], P.Aarthi Rani [2] , D.Sasikala [3]

Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode, Tamilnadu, India[1]

Research Scholar, Department of Computer Science, Erode Arts and Science College, Erode, Tamilnadu, India[2]

Assistant Professor and Head, Department of Computer Applications, Sri Vasavi College, Erode,Tamilnadu, India[3]

Abstract: Digital mammogram becomes the most effective technique for early breast cancer detection modality. Processing images require high computational capabilities. Computer image processing techniques will be applied to enhance images. This paper discusses about Data mining is a technique to dig the data from large database for analysis and execution and the image mining technique deals with extracting implicit knowledge with data relationship. This paper, applies image mining technique on mammogram to classify the cancer diseases. It can be classified into normal, benign and malignant. In existing method used association rule mining, decision tree classify a mammogram image and the Fuzzy Association Rule Mining is applied. Experiments have been taken dataset with 300 images taken from MIAS of various types to improve accuracy using minimum number of rules to patterns. The experiments and results of the FARM gives better performance compared with existing method.

Keywords: Mammogram Image**,** FARM**,** Fuzzy association rule mining, gray-level co occurrence matrix.

I. INTRODUCTION

A Mammogram is an x-ray of the breast that can reveal abnormalities (benign or malignant). The procedure involves compressing the breast between two plates and then applying a small dose of radiation to produce an x-ray image. Mammography plays an important role to detect abnormalities in the breast. It gives detailed information about anatomy, morphology and pathologies of breast for screening and diagnosis of breast cancer. There is a difficulty to detect masses in mammograms because sometimes masses seemed to be similar to normal breast tissues on mammograms. It is difficult to distinguish between malignant and benign masses. Irregular shapes have a higher probability of being malignant and regular shapes have a probability of being benign. Difference in regions of the right and left breast is known as bilateral asymmetry of the breast.

II. LITERATURE SURVEY

In the feature extraction and selection step the features that characterize specific region are calculated and the ones that are important are selected for the classification of the mass as benign or malignant. The feature space is very large and complex due to the wide diversity of the normal tissues and the variety of the abnormalities. Dominguez and Nandi [1] performed segmentation of regions via conversion of images to binary images at multiple threshold levels. Li et al. [2] proposed general guidelines for feature extraction and selection of significant features: discrimination, reliability, independence and optimality. They divided features into three categories: intensity features, geometric features and texture features. Pappas [3] used a generalization of K-means clustering algorithm to separate the pixels into clusters based on their intensity and their relative location. Sahiner et al. [4] used K-means clustering algorithm followed by object selection to detect initial mass shape within the ROI. The ROI is extracted based on the location of the biopsied mass identified by a qualified radiologist. Initial mass shape detection is followed by an active contour segmentation method to refine the boundaries of the segmented mass.

Shruti Dalmiya et al. [5] application of Wavelet based K-means Algorithm in Mammogram Segmentation describes on mammography images using wavelet transformation and K – means clustering for cancer tumor mass segmentation. The first step is to perform image segmentation. It allows distinguishing masses and micro calcifications from background tissue and wavelet transformation and K- means clustering algorithm have been used for intensity based

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

segmentation. Szekely et al. [6] used MRF in "fine" segmentation to improve the preliminary results provided by the "coarse" segmentation. In "coarse" segmentation the feature vector is calculated and passed to a set of decision trees that classifies the image segment. After the "fine" segmentation they used a combination of three different segmentation methods: a modification of the radial gradient index method, the Bezier histogram method and dual binarization to segment a mass from the image. Timp et al. [7] designed two kinds of temporal features: difference features and similarity features. Difference features measured changes in feature values between corresponding regions in the prior and the current view. Varela et al. [8] segmented suspicious regions using an adaptive threshold level. The images were previously enhanced with an iris filter. Zheng et al. [9] used an adaptive topographic region growth algorithm to define initial boundary contour of the mass region and then applied an active contour algorithm to modify the final mass boundary contour. Region growing and region clustering are also based on pixel classification. In region growing methods pixels are grouped into regions. A seed pixel is chosen as a starting point from which the region iteratively grows and aggregates with neighbouring pixels that fulfil a certain homogeneity criterion.

III. EXISTING METHODOLOGY

The classification methods of decision tree classifier for mammogram image classification. Mammography is currently most effective image modality for breast cancer screening. Mining informations are from large database to be recognized as key topic in database system. Classification involves two phases namely Training phase and Testing phase. The training phase, the properties of typical image features based class creased. This features space partitions used to classify image. In the existing system the process of mammogram image classification is more complex. Each model proceeds in different ways to accomplish the process. Most of the models noise removal has not been used in the pre-processing stage. Hence we propose a model for better classification.

Association Rule Mining (ARM)

Discovering frequent item sets is the key process in association rule mining. In order to perform data mining association rule algorithm, numerical attributes should be discretized first then continuous attribute values should be divided into multiple segments. Traditional association rule algorithms adopt an iterative method to discovery, which requires very large calculations and a complicated transaction process. Because of this, a new association rule algorithm is proposed in this paper. This paper adopts a Boolean vector method to discovering frequent item sets. In general, the new association rule algorithm consists of four phases as follows:

1) Transforming the transaction database into the Boolean matrix.
2) Generating the set of frequent 1-itemsets L1.
3) Pruning the Boolean matrix.
4) Generating the set of frequent k-item sets Lk (k>1).

1) Transforming the transaction database into the Boolean matrix

The mined transaction database is D, with D having m transactions and n items. Let T={T1,T2,…,Tm} be the set of transactions and I={I1,I2,…,In}be the set of items. We set up a Boolean matrix Am*n, which has m rows and n columns. Scanning the transaction database D, we use a binning procedure to convert each real valued feature into a set of binary features. The 0 to 1 range for each feature is uniformly divided into k bins, and each of k binary features record whether the feature lies within corresponding range.

2) Generating the set of frequent 1-itemset L1

The Boolean matrix Am*n is scanned and support numbers of all items are computed. The support number Ij.supth of item Ij is the number of '1s' in the j^{th} column of the Boolean matrix Am*n. If Ij.supth is smaller than the minimum support number, itemset { Ij } is not a frequent 1-itemset and the j^{th} column of the Boolean matrix Am*n will be deleted from Am*n. Otherwise itemset { Ij } is the frequent 1-itemset and is added to the set of frequent 1-itemset L1. The sum of the element values of each row is recomputed, and the rows whose sum of element values is smaller than 2 are deleted from this matrix.

3) Pruning the Boolean matrix

Pruning the Boolean matrix means deleting some rows and columns from it. First, the column of the Boolean matrix is pruned according to Proposition 2. This is described in detail as: Let I• be the set of all items in the frequent set LK-1,

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

where k>2. Compute all |LK-1(j)| where j belongs to I2, and delete the column of correspondence item j if |LK – 1(j)| is smaller than k – 1. Second, re-compute the sum of the element values in each row in the Boolean matrix. The rows of the Boolean matrix whose sum of element values is smaller than k are deleted from this matrix.

4) Generating the set of frequent k-itemsets Lk

Frequent k-item sets are discovered only by "and" relational calculus, which is carried out for the k-vectors combination. If the Boolean matrix Ap*q has q columns where 2 < q £ n and minsupth £ p £ m, k q c, combinations of k-vectors will be produced. The 'and' relational calculus is for each combination of k-vectors. If the sum of element values in the "and" calculation result is not smaller than the minimum support number minsupth, the k-itemsets corresponding to this combination of kvectors are the frequent k-itemsets and are added to the set of frequent k-itemsets Lk.

Decision Tree Classification (DTC)

A decision tree is typically evaluated by predictive accuracy that considers all errors equally. However, the predictive accuracy might not be appropriate when the data is imbalanced and the costs of different errors vary markedly. The high rate of correct cancerous predictions is required, while allowing for a small to moderate error rate in the majority class. It is more costly to predict a cancerous case as non-cancerous, than otherwise. Moreover, distribution cost sensitive applications can require a ranking or a probabilistic estimate of the instances.

K-means Algorithm

K-means algorithm is a simple but elegant segmentation method. The main advantage of K-means algorithm is its simplicity. Speed of execution is very high. But the problem with K-means algorithm is that if the initial cluster centers are chosen incorrectly this algorithm may not converge. This happens in the case of noisy image mostly. K-Means algorithm is an unsupervised clustering algorithm that classifies the input data points into multiple classes based on their inherent distance from each other. The algorithm assumes that the data features form a vector space and tries to find natural clustering in them. The points are clustered around centroids which are obtained by minimizing the objective. The various steps of K-means algorithm is described as follows,

1) Compute the intensity distribution (also called the histogram) of the intensities.
2) Initialize the centroids with k random intensities.
3) Repeat the following steps until the cluster labels of the image do not change anymore.
4) Cluster the points based on distance of their intensities from the centroid intensities.

IV. PROPOSED METHODOLOGY

In the proposed method for extracting features, a mammogram image is classified shown in Fig.1. In the training phase, the properties of typical image features based class creased. This features space partitions used to classify image. Fuzzy Association Rule uses fuzzy logic to convert numerical attributes to fuzzy attributes thus maintaining the integrity of the information conveyed by such numerical attributes.

Feature extraction

The characteristics of feature in the objects of interest, if selected carefully are represented of the maximum relevant information that the image has to offer for a complete characterization. Feature extraction methodologies analyze objects and images to extract the most prominent features that are represented of the various classes of objects. Features are used as inputs to classifiers that assign them to the class that they represent.

Intensity Histogram Features

Intensity Histogram analysis has been extensively researched in the initial stages of development of this algorithm. The intensity histogram features like mean, variance, entropy etc. are given in Table 1. The values obtained by this work for different types of images are summarized in Table.2. The calculated features are mean, variance, skewness, kurtosis, entropy and energy.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

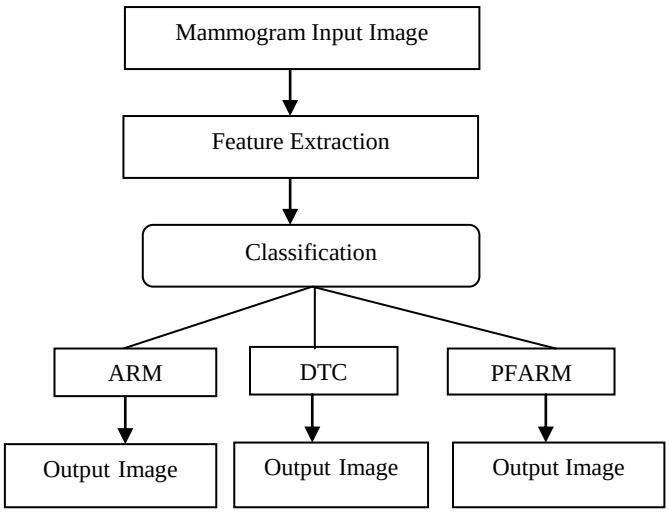

Fig.1 Mammogram image classification system

Table 1. Intensity histogram features

Feature Number assigned	Feature
1	Mean
2	Variance
3	Skewness
4	Kurtosis
5	Entropy
6	Energy

Table 2 Intensity histogram features and their values

Image type	Features					
	Mean	Variance	Skewness	Kurtosis	Entropy	energy
Normal	7.2534	1.6909	-1.4745	7.8097	0.2504	1.5152
Malignant	6.8175	4.0981	-1.3672	4.7321	0.1904	1.5555
Benign	5.6279	3.1830	-1.4769	4.9638	0.2682	1.5690

GLCM Features and GLCM Construction

It is a statistical method that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. By default, the spatial relationship is defined as the pixel of interest and the pixel to its immediate right, but you can specify other spatial relationships between the two pixels. Each element (*I*, *J*) in the resultant GLCM is simply the sum of the number of times that the pixel with value *I* occurred in the specified spatial relationship to a pixel with value *J* in the input image. GLCM is a matrix S that contains the relative frequencies with two pixels: one with gray level value i and the other with gray level j-separated by distance d

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

at a certain angle θ occurring in the image. Given an image window W(x, y, c), for each discrete values of d and θ, the GLCM matrix S(i, j, d, θ) is defined as follows. An entry in the matrix S gives the number of times that gray level i is oriented with respect to gray level j such that $W(x_1, y_1)=i$ and $W(x_2, y_2)=j$, then

$$(x_2, y_2) = (x_1, y_1)+(d\cos\theta, d\sin\theta)$$

Here use two different distances d={1, 2} and three different angles θ={0°, 45°, 90°}. Here, angle representation is taken in clock wise direction.

Feature subset selection

Feature subset selection helps to reduce the space which improves the prediction accuracy and minimizes the computation time. This can be achieved by removing irrelevant, redundant and noisy features, it selects the subset of features that can achieve the best performance in terms of accuracy and computation time. It performs the Dimensionality reduction. Features are generally selected by search procedures. A number of search procedures have been proposed. The selected optimal features are considered for classification. The oscillating search has been fully exploited to select the feature from mammogram which is one of the best techniques to optimize the features among many features.

Fuzzy Association Rule Mining

At the time of fuzzy ARM process, a number of fuzzy partitions are defined on the image domain of each attribute. Fig.2 describes about the fuzzy partition on image domain with different attribute values.

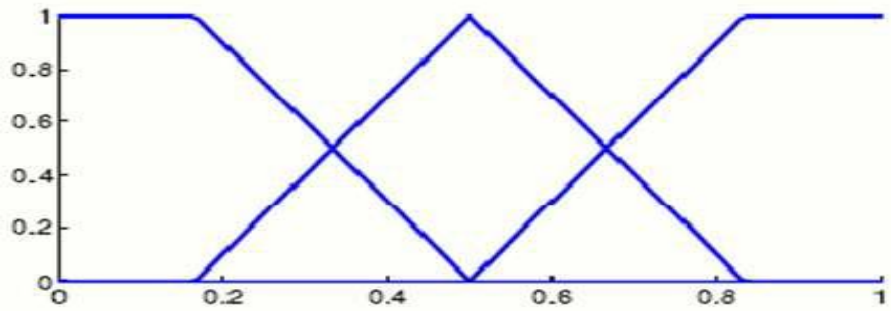

Fig.2 Fuzzy partitions on image domain of different attribute Values

As a result the extended attribute value is in the interval [0,1] is transformed from the original datasets is the transactional database for forming the FARM rules. In order to process this dataset, new measures are used in terms of t-norms. The generation of FARM is directly impacted by the fuzzy measures.

Fuzzy Partitioning Algorithm

Dataset D= { x1 , x2 , ….. xn } where x1 , x2 , ….. xn are different records set of quantitative attributes QA = { q1 , q2 ,….. qr } set of fuzzy partitions FP = { FP1 , FP2 …… FPS }- Set of fuzzy partitions of quantitative attributes qm applied on the image set. Given a dataset D which has both object and their corresponding attributes. Each attribute and the values for it are singled out. A fuzzy partition is obtained by using the ROI containing its corresponding attribute, with each value being uniquely identified by membership function u in these fuzzy partitions.

Proposed Algorithm

Step 1: Read fuzziness parameter m

Step 2: For each $q_p \in QA(p = 1, 2, \ldots \ldots r)$, then FPp = apply_ROI ($q_p$)

Step 3: for each partition t FPi ,Label t appropriately function apply_ROI(q)

Step 4: read k (number of ROI), Find max $\{|\mu_{ij}^{(k+1)} - \mu_{ij}^{(k)}|\} < \delta$

Step 5: for each $\mathbf{x_i} \in \mathbf{D}$ (i = 1, 2,…, N) for each ROI j (j = 1, 2,…, C)

Step 6: calculate μ_{ij}, FP= set of fuzzy partitions after completion of above iteration

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Step 7: return FP.

V.EXPERIMENTATION & RESULTS

The Proposed methodology is experimented with Mammographic Image Analysis Society image databases and the results are presented separately. The images in the database have different sizes and are categorized classes as listed. In particular, a retrieved image is considered a match if and only if it is in the same category as the query.

Table.3. Results obtained by proposed methods.

Actual	ARM	DTC	FARM
Normal	100%	100%	100%
Malignant	92.78%	89.6%	94.71%
Benign	100%	100%	100%

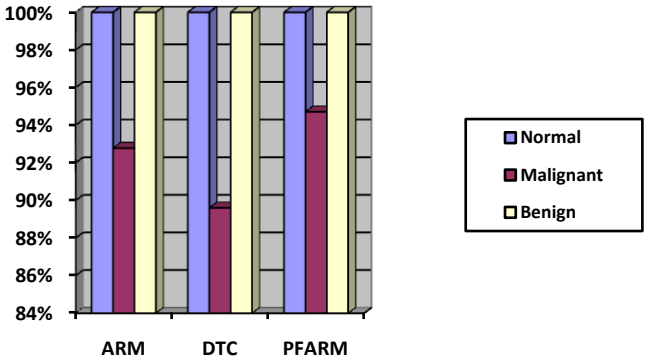

Fig.3 Chart for the results obtained by the proposed method

The fuzzy association rule mining, Decision tree classification and association rule mining using image contents for classification of mammograms. The average accuracy of 95% by using precision and recall measures to evaluation method for mammogram classification. Precision is number of true positive divided by the total number of true positives in dataset. Recall is total number of predictions divided by total number of true positives in dataset. The results using selected features are given in Table.3. and the results are plotted in the Fig.3.

Fig..4. Input image

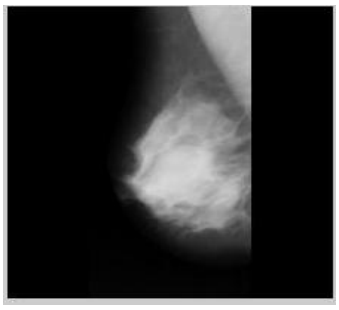

Fig.5. Output image

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

The confusion matrix has been obtained from the testing part. For example, out of 97 actual malignant images 07 images was classified as normal. In case of benign and normal all images are classified correctly shows in Table.4.

Table 4. Confusion matrix of predicted class

Actual	ARM			DTC			FARM		
Normal	99	0	99	99	0	99	99	0	99
Malignant	97	90	7	97	87	10	97	94	3
Benign	104	0	0	104	0	0	104	0	0

VI. CONCLUSION

The mammography is the best method for breast cancer detection. A fuzzy association rule mining is proposed. The main aim of the method used to improve the accuracy of detection and reduce computation cost of mammogram image analysis and can be applied to other image analysis applications. The algorithm uses simple statistical techniques to develop a fuzzy based feature selection for medical images. The proposed method proves that the approach is easier and requires less computation time compared than other existing methods.

REFERENCES

[1] Dominguez, A.R., Nandi and A.F., "Enhanced Multi-Level Thresholding Segmentation and Rank Based Region Selection for Detection of Masses in Mammograms", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 449–452 (April 2007).

[2] Li, H., Wang, Y., Ray Liu, K.J., Lo, S.C.B., Freedman, "Computerized Radiographic Mass Detection— Part II: Decision Support by Featured Database Visualization and Modular Neural Networks", *IEEE Transactions on Medical Imaging* 20(4) (April 2001).

[3] Pappas, T.N., "An Adaptive Clustering Algorithm for Image Segmentation", *IEEE Transactions on Signal Processing* 40(4), 901–914 (1992).

[4] Sahiner, B., Hadjiiski, L.M., Chan, H.P., Paramagul, C., Nees, A., Helvie, M., Shi, J., "Concordance of Computer-Extracted Image Features with BI-RADS Descriptors for Mammographic Mass Margin", *Giger, M.L., Karssemeijer, N. (eds.) Proc. of SPIE Medical Imaging* 2008: *Computer-Aided Diagnosis,* vol. 6915 (2008).

[5] 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.

[6] Szekely, N., Tóth, N., Pataki,B., "A Hybrid System for Detecting Masses in Mammographic Images", *IEEE Transactions on Instrumentation and Measurement* 55(3), 944–951 (2006).

[7] Timp, S., Varela, C., Karssemeijer, N., "Temporal Change Analysis for Characterization of Mass Lesions in Mammography", *IEEE Transactions on Medical Imaging* 26(7), 945–953 (2007).

[8] Varela, C., Tahoces, P.G., Méndez, A.J., Souto, M., Vidal, J.J., "Computerized Detection of Breast Masses in Digitized Mammograms", *Computers in Biology and Medicine* 37, 214–226 (2007).

[9] Zheng, B., Mello-Thoms, C., Wang, X.H., Gur, D., "Improvement of Visual Similarity of Similar Breast Masses Selected by Computer-Aided Diagnosis Schemes", *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, ISBI 2007, April 12-15, pp. 516–519 (2007)