Medical Image Feature, Extraction, Selection And Classification

[1]M.VASANTHA[*], [2]DR.V.SUBBIAH BHARATHI, [3]R.DHAMODHARAN

1. Research Scholar, Mother Teresa Women's University, KodaiKanal, and

Asst.Professor, Department of Computer applications,St.Peters University , Chennai

Vasantha_spec@yahoo.com, Phone No. 9884759409

2. Dean ,DMI College Of Engineering, Chennai

3. St. Petre' s University, Chennai

Abstract - Breast cancer is the most common type of cancer found in women. It is the most frequent form of cancer and one in 22 women in India is likely to suffer from breast cancer. This paper proposes a image classifier to classify the mammogram images. Mammogram image is classified into normal image, benign image and malignant image. Totally 26 features including histogram intensity features and GLCM features are extracted from mammogram image. A hybrid approach of feature selection is proposed in this paper which reduces 75% of the features. Decision tree algorithms are applied to mammography classification by using these reduced features. Experimental results have been obtained for a data set of 113 images taken from MIAS of different types. This technique of classification has not been attempted before and it reveals the potential of Data mining in medical treatment**.**

Keywords- Breast cancer, Mammogram, Decision tree, Data mining, classification.

1. INTRODUCTION

Breast cancer in India is in rise and rapidly becoming the leading cancer in females (MedIndia 2006) and death toll is increasing at fast rate (Gajalakshmi *et al.*, 2009) and no effective way to treat this disease yet. So early detection becomes a critical factor to cure the disease and improving the surviving rate. Generally the X-ray mammography is a valuable and most reliable method in early detection.

Data mining of medical images is used to collect effective models, relations, rules, abnormalities and patterns from large volume of data. This procedure can accelerate the diagnosis process and decision-making. Different methods of data mining have been used to detect and classify anomalies in mammogram images such as wavelets [2, 6], statistical methods and most of them used feature extracted using image processing techniques [5].Some other methods are based on fuzzy theory [1] and neural networks [3].Most of the Computer Aided Methods proved to be the powerful tool that assists the radiologist to speed up the treatment process.

In this paper we have used classification method called Decision tree classifier for image classification. Classification process typically involves two phases: training phase and testing phase. In training phase the properties of typical image features are isolated and based on this training class is created .In the subsequent testing phase , these feature space partitions are used to classify the image. We have used supervised decision tree method by extracting low level image features for classification. The merits of this method are effective feature extraction, selection and efficient classification. The rest of the paper is organized as follows. Section 2 presents the preprocessing and section 3 presents the feature extraction phase. Section 4 discusses the proposed method of Feature selection and classification. In section5 the results are discussed and conclusion is presented in section 6.

2.PRE-PROCESSING

The mammogram image for this study is taken from Mammography Image Analysis Society (MIAS), which is an UK research group organization related to the Breast cancer investigation. As mammograms are difficult to interpret, preprocessing is necessary to improve the quality of image and make the feature extraction phase as an easier and reliable one. The calcification cluster/tumor is surrounded by breast tissue that masks the calcifications preventing accurate detection and shown in Figures 2.1 .A pre-processing; usually noise-reducing step is applied to improve image and calcification contrast.

In this work an efficient filter referred to as the low pass filter, was applied to the image that maintained calcifications while suppressing unimportant image features.

Figures 2 shows representative output image of the filter for a image cluster in figure 1. By comparing the two images, we observe background mammography structures are removed while calcifications are preserved. This simplifies the further tumor detection step.

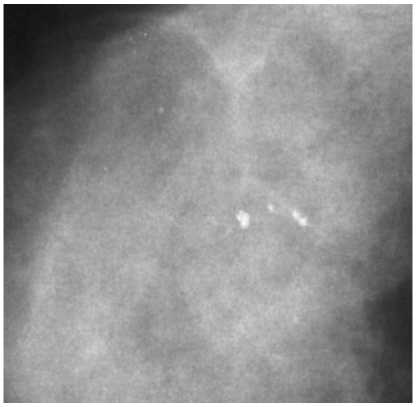

Fig. 1 ROI of a Benign

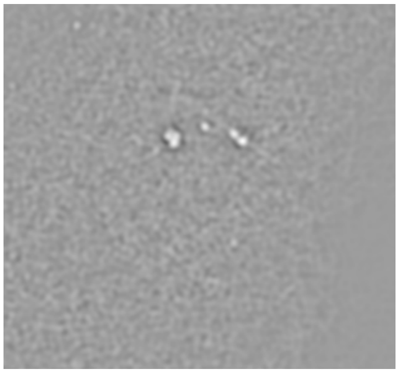

Fig. 2. ROI after Pre-processing Operation

2.1 Histogram Equalization

Histogram equalization is a method in image processing of contrast adjustment using the image's histogram. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to get better contrast. Histogram equalization accomplishes this by efficiently spreading out the most frequent intensity values. The method is useful in images with backgrounds and foregrounds that are both bright or both dark. In particular, the method can lead to better views of bone structure in x-ray images, and to better detail in photographs that are over or under-exposed. In mammogram images Histogram equalization is used to make contrast adjustment so that the image abnormalities will be better visible.

3. FEATURE EXTRACTION

Features, characteristics of the objects of interest, if selected carefully are representative of the maximum relevant information that the image has to offer for a complete characterization a lesion. Feature extraction methodologies analyze objects and images to extract the most prominent features that are representative of the various classes of objects. Features are used as inputs to classifiers that assign them to the class that they represent. In this Work intensity histogram features and Gray Level Co-Occurrence Matrix(GLCM) features are Extracted.

3.1 Intensity Histogram Features

Intensity Histogram analysis has been extensively researched in the initial stages of development of this algorithm. Prior studies have yielded the intensity histogram features like mean, variance, entropy etc. These are summarized in Table 3.1(a) Mean values characterize individual calcifications; Standard Deviations (SD) characterize the cluster. Table 3.1(b) summarizes the values for those features.

Table 3.1(a) Intensity histogram features

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

In this paper , the value obtained from our work for different type of image is given as follows:

Table 3.1.(b) 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
malignan	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

3.2 GLCM Features

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 (horizontally adjacent), 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.

The Following GLCM features were extracted in our research work:

Autocorrelation, Contrast, Correlation, Cluster Prominence, ClusterShade, Dissimilarity Energy, Entropy, Homogeneity, Maximum probability , Sum of squares, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, Information measure of correlation, information measure of correlation, Inverse difference normalized.

The value obtained for the above features from our work for a typical image is given in the following table 3.2

Table 3.2 : GLCM Features and values Extracted from Mammogram Image

Feature No	Feature Name	FeatureValues
1)	autocd	44.1530
2)	contrd	1.8927
3)	corrpd	0.1592
4)	cpromd	37.6933
5)	cshad1	4.2662
6)	dissid	0.8877
7)	energd	0.1033
8)	entrod	2.6098
9)	homopd	0.6645
10)	maxprd	0.6411
11)	sosvhd	0.1973,
12)	savghd	44.9329
13)	svarhd	13.2626
14)	senthd	133.5676
15)	dvarhd	1.8188
16)	denthd	1.8927
17)	inf1hd	1.2145
18)	inf2hd	-0.0322
19)	indncd	0.2863
20)	idmncd	0.9107

4. FEATURE SELECTION

Feature selection helps to reduce the feature space which improves the prediction accuracy and minimizes the computation time. This is achieved by removing irrelevant, redundant and noisy features .i.e., 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 .Popularly used feature selection algorithms are Sequential forward Selection, Sequential Backward selection, Genetic Algorithm and Particle Swarm Optimization.

In this work a combined approach of Greedy stepwise method and Genetic Algorithm is proposed to select the optimal features. The selected optimal features are considered for classification.

4.1 Proposed Hybrid Approach Algorithm:

1. Extract N number of features A1, A2, A3..AN from ROI Of the preprocessed Image
2. Apply Genetic algorithm to select the optimal set containing n1 number of features where n1<N
3. Apply Greedy step wise search to select the best subset containing n2 number of features n2 where n2<N
4. Find the Union of n1 features and n2 features as n features
5. Use the n features where n<N for Classification.

The selected features using GA method are tabulated as follows:

Table 4.1(a): Feature selected By GA method

S.no	Features
1.	Cluster prominence
2.	Energy
3.	Information measure of correlation
4.	Inverse difference Normalized
5.	Skewness
6.	Kurtosis

The selected features using Greedy stepwise method are listed in the following table

Table 4.1(b): Feature selected By Greedy method

S.no	Features
1.	Energy
2.	Mean

By applying the proposed algorithm, it will produce a feature set contain best set of features which is less than the original set and is given in the below table

Table 4.1 (c): Feature selected By proposed Hybrid method

S.no	Features
1.	Cluster prominence
2.	Energy
3.	Information measure of correlation
4.	Inverse difference Normalized
5.	Skewness
6.	Kurtosis
7.	Contrast
8.	Mean

4.2 Classification

The selected features are used for classification. For classification of samples, we have employed the freely available Machine Learning package, WEKA[4] to train our data set using J48 decision tree method. Out of 113 images in the dataset, 80 were used for training and the remaining for testing purposes.

5. EXPERIMENTAL RESULTS

In this paper we used J48 classifier, a decision tree classifier based on C4.5, from WEKA [4] to train and test the features. The average accuracy is 95%. We have used the precision and recall measures as the evaluation metric for mammogram classification. Precision is the fraction of the number of true positive predictions divided by the total number of true positives in the set. Recall is the total number of predictions divided by the total number of true positives in the set .

The testing results using the selected features are given as shown in the table 3

Table 5.1: Results obtained by proposed method

Normal	100%
Malignant	87.5 %
Benign	100%

The confusion matrix has been obtained from the testing part .In this case for example out of 35 actual malignant images 5 images was classified as normal. In case of benign and normal all images are correctly. The confusion matrix is given below (Table 5.2)

Table5.2: Confusion matrix

Actual	Predicted class		
	Benign	Malignant	Normal
Benign	40	0	0
Malignant	0	30	5
Normal	43	0	0

The following graph shows the comparative analysis of our method and various other methods :

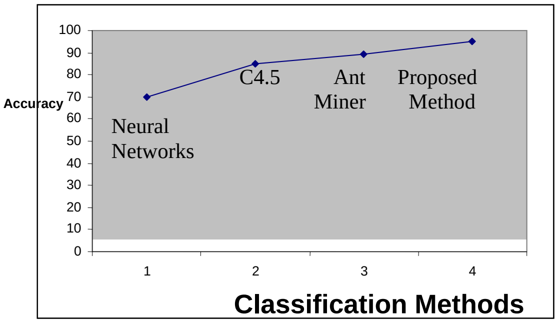

Fig. 3 Performance of the Classifier

6. CONCLUSION

Mammography is one of the best methods in breast cancer detection, but in some cases radiologists face difficulty in directing the tumors. The methods like one presented in this paper could assist the medical staff and improve the accuracy of detection. Our method can reduce the computation cost of mammogram image analysis and can be applied to other image analysis applications. The algorithm uses simple statistical techniques in collaboration to develop a novel feature selection technique for medical image analysis. The value of this technique is that it not only tackles the measurement problem but also provides a visualization of the relation among features. In addition to ease of use, this approach effectively addresses the feature redundancy problem. The method proposed has been proven that it is easier and it requires less computing time than existing methods.

REFERENCES

[1] D.Brazokovic and M.Nescovic ., “Mammogram screening using multisolution based image segmentation”, International journal of pattern recognition and Artificial Intelligence, 7(6):1437-1460,1993. [

[2] C.Chen and G.Lee, “Image segmentation using multitiresolution wavelet analysis and Expectation Maximum(EM) algorithm for mammography” , International Journal of Imaging System and Technology, 8(5):491-504,1997.

[3] I.Christiyanni et al ., “Fast detection of masses in computer aided mammography”, IEEE Signal processing Magazine, Pages:54-64,2000.

[4] Holmes, G., Donkin, A., Witten, I.H.: “WEKA: a machine learning workbench.” In: Proceedings Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, pp. 357-361 (1994).

[5] S.Lai,X.Li and W.Bischof .” On techniques for detecting circumscribed masses in mammograms”, IEEE Trans on Medical Imaging , 8(4):377-386,1989.

[6] T.Wang and N.Karayaiannis, “Detection of microcalcification in digital mammograms using wavelets”, IEEE Trans. Medical Imaging, 17(4):498-509,1998