Journal of Theoretical and Applied Information Technology
30th November 2011. Vol. 33 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MAMMOGRAM CLASSIFICATION USING MAXIMUM DIFFERENCE FEATURE SELECTION METHOD

[1]R.NITHYA, [2]B.SANTHI

[1]Asstt Prof., School of Computing, SASTRA University, Thanjavur, Tamilnadu, India-613402

[2] Prof., School of computing, SASTRA University, Thanjavur, Tamilnadu, India -613402

E-mail: nithya.r@mca.sastra.edu, shanthi@cse.sastra.edu

ABSTRACT

This paper developed a CAD (Computer Aided Diagnosis) system based on neural network and a proposed feature selection method. The proposed feature selection method is Maximum Difference Feature Selection (MDFS). Digital mammography is reliable method for early detection of breast cancer. The most important step in breast cancer diagnosis is feature selection. Computer automated feature selection is reliable and also it helps to improve the classification accuracy. GLCM (Gray Level Co-occurrence Matrix) features are extracted from the mammogram. The extracted features are selected based on a proposed MDFS method. Experiments have been conducted on datasets from DDSM (Digital database for Screening Mammography) database. Several feature selection methods are available. The accuracy of the model depends on the relevant feature selection. The proposed MDFS method selects only essential features and eliminates the irrelevant features. The experiment results show that neural network based model with proposed feature selection method improved the classification accuracy.

Keywords: *Artificial Neural Network (ANN), Breast Cancer, GLCM, Mammogram, Feature Selection*

1. INTRODUCTION

Currently breast cancer is common disease among women. The computer classification system can reduce the number of unnecessary biopsies. Abnormal breast classified into two types: Mass and Calcifications. Calcifications consist of two types: Microcalcifications and Macrocalcifications [1]. Masses are identified by their shape and margin characteristics. Most of the breast cancer is detected by presence of microcalcifications. Micro calcifications are small calcium deposit and appear as group of bright spots in mammograms. The important factor needed in this disease is early detection and accurate diagnosis. Currently mammography is effective and low cost method to detect breast cancer [2]. Digital mammography is proven as efficient tool to detect breast cancer at early stage. Usually biopsy is unnecessary to detect breast cancer at early stage. The symptoms of breast cancer include mass, changes in shape and dimension of breast. The earlier the cancer is detected, the better treatment can be provided. In the last ten years, several computer aided diagnosis systems are developed to automate detection of breast cancer. Normal pattern typically have smooth surfaces. Conversely, abnormal pattern presents rough and complex surfaces.

Several types of features are extracted from the digital mammograms including region-based features, shape-based features, texture based features and position based features. Texture feature have been widely used to classify normal and abnormal pattern in digital mammogram. In this paper texture based GLCM features are extracted. Feature selection is commonly used in breast cancer classification. The increased dimensionality of data makes both training and testing of classification method difficult [3]. Feature selection helps to enhances classification accuracy. It is necessary to identify and remove irrelevant or redundant features. Feature selection is an important step before any classification scheme. The success of classification scheme depends on features selected. The advantage of feature selection including improvement of the prediction performance, reduces training times and faster performance of the classifier [4]. Several feature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

selection techniques including software packages exist for obtaining minimal feature set. Reducing the dimensionality of the raw input variable space is an important step in pattern recognition

This paper presents the development of CAD system for the detection of normal and abnormal pattern in the breast. The proposed system consists of three major steps: The first step is the feature extraction. The second step is the feature selection using proposed MDFS method. The third step is the classification process using neural network technique, mammogram classified into normal and abnormal pattern.

This paper is organized as follows: Section 2 summarizes the available research on the breast cancer detection. Section 3 describes the proposed methodology. Section 4 demonstrates performance measures. Section 5 discusses experiment results. Finally, in Section 6 conclusion arrives.

2. RELATED WORK

The numbers of research work are conducted in the area of breast cancer detection and classification. A computer system that performs automatic cancer detection can assist the radiologist by providing second opinion and reduce unnecessary biopsy. A.Michael et al, [2] applied a hypothesis test to determine whether the feature can discriminate or not. They conducted the experiments on DDSM database. Brijesh Verma et al, [3] developed a computer aided diagnosis system for digital mammograms based on neural-genetic algorithm feature selection method and obtained accuracy was 85% on mammograms from DDSM. Mohamed A.Alofe et al, [4] used filter model and wrapper model to feature selection. They conducted experiments on MIAS database and obtained accuracy was 100%. Hui-Ling Chen et al, [5] proposed rough set-based feature selection and they obtained accuracy was 96% on mammograms from UCI machine learning repository. M.Vasantha et al, [6] proposed hybrid feature selection method for mammogram classification on DDSM database. The highest classification accuracy obtained by this approach was 96%. Pasi Luukka [7] introduced feature selection method based on fuzzy entropy measures and obtained accuracy was 98.28%. C.Cheng-Lung Huang et al, [8] used support vector machine based feature selection and obtained accuracy was 86%. The Sequential Floating Forward Selection (SFFS) is used by B.Prathiba et al, [9] to reduce the feature dimensionality. The SFFS finds the most discriminate features by sequentially adding and deleting features. Two well known feature selection techniques including forward selection and backward selection is used by Shu Ting-Lio et al, [10] and the experiments tested on UCI machine learning repository. L.wei et al, [11] used sequential backward selection method for the purpose of selecting the most relevant features. In this work 18 features were extracted, out of which 12 features were finally selected for the classification of benign and malign pattern.

3. METHODOLOGY

The overview of proposed methodology is depicted in figure 1.

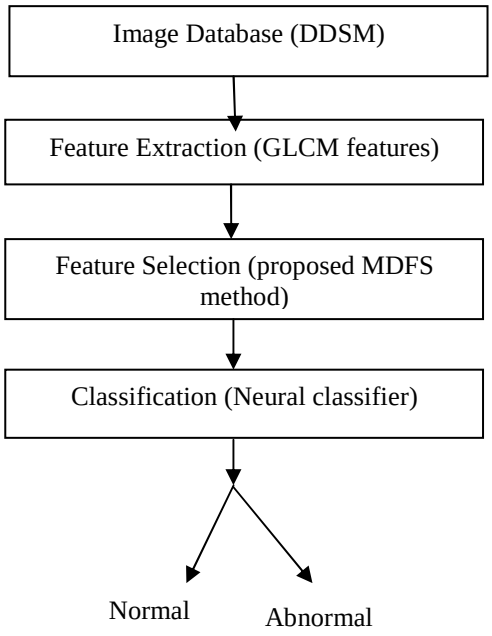

Figure 1. Proposed methodology

3.1. Image Database

In this experiment, real-world breast cancer database from the DDSM was chosen. The mammograms are provided by university of South Florida. The positions of individual mass and micro calcifications are marked. The mammograms in DDSM downloaded from the website located at http://marathon.csee.usf/edu/Mammography/DDSM. The database contains more than 2500 samples.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2. Feature extraction

The features are extracted from the mammograms using GLCM. GLCM calculates the probability of a pixel with the gray level i occurring in a specific spatial relationship to a pixel with the value j [9] .The number of gray levels in the image determines the size of the GLCM. GLCM calculated in 4 angles (0^0,45^0,90^0,135^0) and 4 distances(1,2,3,4) [12-13]. The 18 descriptors extracted from GLCM texture measurement including autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares: variance, sum average, sum variance, sum entropy, difference entropy, information measure of correlation, information measure of correlation 2 and inverse difference moment normalized. The features for normal and abnormal patterns are shown in table 1.

Following notations are used to describe the various GLCM features:

G is the number of different gray levels in an image. P referred to as GLCM. μ is the mean value of P. μ_x and μ_y are the means of P_x and P_y. σ_x and σ_y are standard deviations of P_x and P_y. P_x (i) is the ith entry in the matrix obtained by summing the rows of P (i, j).

$P_x(i) = \sum_{j=0}^{G-1} P(i,j)$ $P_y(j) = \sum_{i=0}^{G-1} P(i,j)$

$\mu = \sum_{i,j=0}^{G-1} i\, P(i,j)$

$\mu_x = \sum_{i=0}^{G-1} i \sum_{j=0}^{G-1} P(i,j) = \sum_{i=0}^{G-1} i\, p_x(i)$

$\mu_y = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} j\, P(i,j) = \sum_{i=0}^{G-1} j\, p_y(j)$

$\sigma_x^2 = \sum_{i=0}^{G-1} (i-\mu_x)^2 \sum_{j=0}^{G-1} P(i,j)$

$= \sum_{i=0}^{G-1} (P_x(i) - \mu_x(i))^2$

$\sigma_y^2 = \sum_{i=0}^{G-1} (j-\mu_y)^2 \sum_{j=0}^{G-1} P(i,j)$

$= \sum_{i=0}^{G-1} (P_y(j) - \mu_y(j))^2$

$P_{x+y}(k) = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)$ for k=0, 1….. , 2(G-1)

$HXY = -\sum_{i,j=0}^{G-1} p(i,j) log_2 p(i,j)$

$HXY1 = -\sum_{i,j=0}^{G-1} p(i,j) log_2 (p_x(i)\, p_y(i))$

$HXY2 = -\sum_{i,j=0}^{G-1} p_x(i) p_y(i) log_2 (p_x(i)\, p_y(i))$

Expressions of GLCM descriptors are:

1) Autocorrelation
   $= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (p_x - \mu_x)(p_y - \mu_y)/\sigma_x \sigma_y$
2) Contrast = $\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)(i-j)^2$
3) Correlation
   $= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)\, X\, (i\, X\, j) - (\mu_x X \mu_y)/ \sigma_x \sigma_y$
4) Cluster prominence
   $= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)(i + j - \mu_x - \mu_y)^4$
5) Cluster shade
   $= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)(i + j - \mu_x - \mu_y)^3$
6) Dissimilarity = $\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} |i-j| P(i,j)$
7) Energy = $\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)^2$
8) Entropy = - $\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j) \log(P(i,j))$
9) Homogeneity = $\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{P(i,j)}{1+|i-j|}$
10) Maximum probability = max(i,j)P(i,j)
11) Sum of squares: variance
    $= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} P(i,j)\, (i-\mu)^2$
12) Sum average(sa) = $\sum_{i=0}^{2G-2} i P_{x+y}(i)$
13) Sum variance = $\sum_{i=0}^{2G-2} (i - sa)^2\, p_{x+y}(i)$
14) Sum entropy
    $= -\sum_{i=0}^{2G-2} P_{x+y}(i) \log(P_{x+y}(i))$
15) Difference entropy
    $= -\sum_{i=0}^{G-1} P_{x+y}(i) \log(P_{x+y}(i))$
16) Information measure of correlation
    $= \frac{HXY - HXY1}{\max\{HX, HY\}}$
17) Information measure of correlation2
    $= \sqrt{(1 - \exp[-2.0(HXY2 - HXY)}$
18) Inverse difference moment normalized
    $= \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \frac{P(i,j)}{1+|i-j|^2}$

3.3. Feature selection

Feature selection is important step in breast cancer detection and classification. After the features are extracted, it is found that not all

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

features used in differentiating between normal and abnormal pattern. The advantage of feature selection is to limit the number of input features to achieve optimum accuracy and also reduce computation complexity. In order to improve the efficiency of feature selection, this paper proposes a maximum difference feature selection method to determine whether the features are relevant or not.

3.3.1. Maximum difference feature selection (MDFS)

The total 18 GLCM features are extracted from the mammograms. It is difficult to select features that discriminate between normal and abnormal tissue. However the use of all the features, results the high dimensioned feature vector that degrade the classification accuracy as well as increase the computational complexity. So feature selection method is necessary to select most relevant features. Classification of normal and abnormal pattern is difficult because both pattern exhibit similar characteristics. The focus of this method is to eliminate similar feature between normal and abnormal pattern. The maximum difference features are selected using sample normal and abnormal mammograms. The basic idea of this algorithm is to identify features that are dissimilar between normal and abnormal pattern. Using this method, top five features are selected. The proposed MDFS method applied for 50 normal and abnormal mammograms are shown in table 2.

Algorithm

Step 1: Extract feature from N normal mammograms. Let it be A.
Step: Extract feature from N abnormal mammograms. Let it be B.
Step 3: Compute sum of feature for N normal mammograms.

$$S1 = \sum_{i=1}^{N} A_i$$

Step 4: Compute sum of feature for N abnormal mammograms.

$$S2 = \sum_{i=1}^{N} B_i$$

Step 5: Compute feature difference (D) between normal and abnormal mammograms

If S1>S2

$$D = (S1-S2) / (S1+S2).$$

else

$$D = (S2-S1) / (S1+S2).$$

Step 6: Repeat step 1 to 5 for all 18 features.
Step 7: Assign rank value to each feature based on D in descending order.
Step 8: Select most relevant top five features.

3.4. Classification

This experiment uses three layer artificial neural network with input layer, hidden and output layer [14]. The sigmoid activation function used for both hidden layer and output layer. The weight values between input and hidden layer, the weight values between hidden and output layer of neural network is updated to achieve optimum classification. The classification process is divided into the training phase and the testing phase. In the training phase, known data are given and the classifier is trained. In testing phase, unknown data are given and the classification is performed using trained classifier. The selected features are normalized and given as input to neural classifier [3]. One hidden layer is used in neural network. One node is used in the output layer which has been trained to represent 1 for normal cases and 0 for abnormal cases.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1. Sample data- GLCM features for normal and abnormal pattern

Feature no	GLCM Features	Normal	Abnormal
1	Autocorrelation	9.1504	7.4782
2	Contrast	0.7890	0.1363
3	Correlation	0.5613	0.9642
4	Cluster Prominence	22.9289	69.4536
5	Cluster Shade	-2.9112	3.0820
6	Dissimilarity	0.4118	0.0998
7	Energy	0.1890	0.3118
8	Entropy	2.0868	1.5162
9	Homogeneity	0.8471	0.9544
10	Maximum probability	0.3368	0.4557
11	Sum of squares: Variance	9.4439	7.5398
12	Sum average	5.8807	4.7554
13	Sum variance	20.8125	18.5144
14	Sum entropy	1.6375	1.4269
15	Difference entropy	0.8260	0.3289
16	Information measure of correlation	-0.3423	-0.7440
17	Information measure of correlation2	0.7660	0.9150
18	Inverse difference moment normalized	0.9887	0.9980

Table 2. Proposed MDFS method applied for GLCM features

Feature no		GLCM Features	S1	S2	D	Rank
1		Autocorrelation	390.52	298.573	0.1334	11
2	*	Contrast	39.061	18.101	0.3667	2
3		Correlation	29.818	40.673	0.1539	8
4		Cluster Prominence	1241.832	1739.708	0.1669	7
5	*	Cluster Shade	-61.105	150.688	2.36421	1
6	*	Dissimilarity	23.503	12.114	0.3197	3
7		Energy	8.99	13.271	0.1923	6
8		Entropy	106.653	87.883	0.0964	13
9		Homogeneity	40.569	44.79	0.0494	16
10		Maximum probability	16.85	20.829	0.1056	12
11		Sum of squares: Variance	405.148	304.877	0.1412	9
12		Sum average	266.034	222.012	0.0902	14
13		Sum variance	858.687	647.29	0.1403	10
14		Sum entropy	82.603	75.907	0.0422	17
15	*	Difference entropy	44.093	27.123	0.238	5
16	*	Information measure of correlation	-15.173	-27.893	0.295	4
17		Information measure of correlation 2	35.888	41.479	0.0722	15
18		Inverse difference moment normalized	49.4265	49.7332	0.0030	18

Journal of Theoretical and Applied Information Technology
30th November 2011. Vol. 33 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. PERFORMANCE MEASURES

Three performance measure terms Accuracy (AC), Sensitivity (SE) and Specificity (SP) are used to evaluate the performance of the classifier [1]. Sensitivity is a proportion of positive cases that are well detected by the test. Specificity is a proportion of negative cases that are well detected by the test. Classification accuracy is depends on the number of samples correctly classified. They are defined as follows

AC= (TP+TN) / (TP+FP+TN+FN)

SE=TP/ (TP+FN)

SP= TN / (TN+FP)

where, TP is the number of true positives; FP, the number of false positives; TN, the number of true negatives; FN, the number of false negatives. Confusion matrix is shown in Table 3.

TP- predicts abnormal as abnormal.
FP- predicts abnormal as normal.
TN- predicts normal as normal.
FN- predicts normal as abnormal.

Table 3. Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

5. EXPERIMENT RESULTS

The experiment described here uses the DDSM database. It includes four steps: 1) feature extraction 2) feature selection 3) feature normalization 4) classification. Features are extracted using GLCM. The extracted features are selected by proposed MDFS method. In MDFS, the sum values are computed for normal and abnormal features. Differences are calculated between normal and abnormal sum values. Rank assignment depends on the significant difference level. Thus, the proposed feature selection algorithm selects most relevant five features. The selected features are cluster shade, contrast, dissimilarity, information measure of correlation and difference entropy. The selected features are normalized between 0 and 1. The normalized features are fed into neural classifier. The weights are adjusted in the experiment to achieve optimum classification. A total of 125 normal cases and 125 abnormal cases are used for the experiments. The experiments were run using 200 cases (100 normal, 100 abnormal) for training and 50 cases (25 normal, 25 abnormal) for testing. The training set used for training the network and test set used for estimating the accuracy of the model. In order to check the efficiency of the proposed method, MDFS method is compared with random feature selection
method. Table 4-5 and figure 2-3 represents confusion matrix for proposed feature selection method and random feature selection method .The results shows that the proposed MDFS is better than
random selection and shown in table 6.

Table 4.Confusion matrix for proposed MDFS method

Actual	Predicted	
	Abnormal (Positive)	Normal (Negative)
Abnormal (Positive)	22(TP)	3(FP)
Normal (Negative)	0(FN)	25(TN)

Table 5.Confusion matrix for random feature selection method

Actual	Predicted	
	Abnormal (Positive)	Normal (Negative)
Abnormal (Positive)	24(TP)	1(FP)
Normal (Negative)	0(FN)	25(TN)

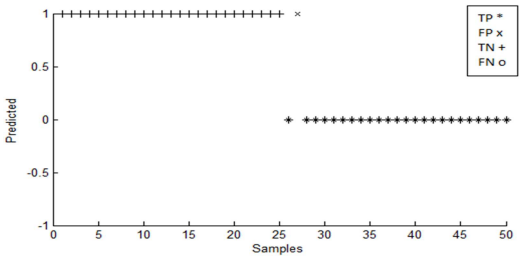

Figure 2. Confusion matrix for proposed MDFS method

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

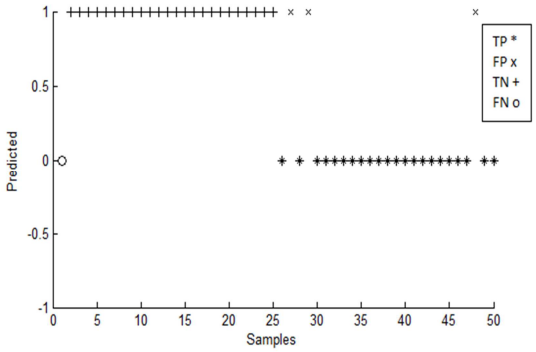

Figure 3. Confusion matrix for random feature selection method

Table 6. Performance measures comparison

Feature selection method	Features	Accuracy (%)	Sensitivity (%)	Specificity (%)
Proposed MDFS	Cluster shade, dissimilarity, contrast, difference entropy and information measure of correlation.	98	100	96
Random selection	Maximum probability, sum variance, sum entropy, information measure of correlation2, inverse difference moment normalized	94	100	89

6. CONCLUSION

In this work, a CAD system for normal and abnormal breast detection of mammograms has been presented. In this paper simple and effective algorithm for feature selection is proposed. The feature selection method that have used in the proposed CAD system had given promising results in classification between normal and abnormal pattern. Five features are considered to be the most significant features of a digital mammogram for classification. Randomly selected features obtained 94% accuracy, whereas proposed maximum difference feature selection method yielded 98% accuracy. Thus the proposed algorithm outperforms random selection method. Future work will examine the performance of the proposed feature selection method with the fuzzy techniques.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] Brijesh Verma, Peter McLeod and Alan Klevansky, "Classification of benign and malign patterns in digital mammograms for the diagnosis of breast cancer", Expert System with Applications, 37, pp.3344-3351, 2010.

[2] Michael A.Yachoub, A.S.Mohamed and Yasser M.Kadah," A CAD system for the detection of malignant patterns in digitized mammogram films", CARIO International Biomedical Engineering Conference, 2006.

[3] Brijesh Verma and Ping Zhang,"A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms", Applied Soft Computing, 7, pp. 513-525, 2007.

[4] Mohamed A.Alolfe,"Feature selection in computer aided diagnostic system for microcalcification detection in digital mammograms",26th National Radio Science Conference, 2009.

[5] Hui-Ling Chen, Bo Yang, Jie Liu and Da-You Liu,"A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis", Expert System with Applications, 38,pp.9014-9022,2011.

[6] M.Vasantha and V.Subbiahbharathi," Classification of mammogram images using hyprid features", European Journal of Scientific Research, Vol.57, No.1, 2011.

[7] Pasi Luukka,"Feature selection using fuzzy entropy measures with similarity classifier", Expert system with Applications, 38, pp.4600-4607, 2011.

[8] C.Cheng-Lung Huang, Hung-Chang Liao and Mu-Chen Chen," prediction model building and feature selection with support vector machine in breast cancer diagnosis", Expert Systems with applications, 38, pp.578-587, 2008.

[9] B.N.Prathibha and V.Sadasivam,"A kernel discriminant analysis in mammogram classification using with texture features in wavelet domain", International journal on computational intelligence, Vol.1, Issue.1, 2010.

[10] Shu-Ting Lio and Bor-Wen Cheng,"Diagnosing breast masses in digital mammography using feature selection and ensemble methods", Medical Systems,2010.

[11] Liyang Wei, Yongyi Yang and Robert M.Nishikawa," Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis", Pattern Recognition, 42, pp.1126-1132, 2009.

[12] A.M.Khuzi, R.Besar and W.M.D. Wan Zaki, "Texture features selection for masses detection in digital mammogram", Biomed, proceedings, pp.629-632, 2008.

[13] R.Nithya and B.Santhi,"Classification of normal and abnormal patterns in digital mammograms for the diagnosis of breast cancer", International Journal of Computer Applications, vol.28, No.6, 2011.

[14] A. Papadopoulos, D.I.Fotiadis and A.Likas," Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines", Artificial Intelligence in Medicine, 34, pp.141-150, 2005.