A Novel Approach for Cancer Detection in MRI Mammogram Using Decision Tree Induction and BPN

Dr.V.Saravanan tvsaran@hotmail.com
H.O.D,Department of M.C.A,
Karunya Deemed University,
Coimbatore, Tamil Nadu, India

S.Pitchumani Angayarkanni pitchu_mca@yahoo.com
Assistant Professor, Department of Computer Science,
Lady Doak College, Madurai,Tamil Nadu , India

Abstract

An intelligent computer-aided diagnosis system can be very helpful for radiologist in detecting and diagnosing micro calcifications patterns earlier and faster than typical screening programs. In this paper, we present a system based on fuzzy-C Means clustering and feature extraction techniques using texture based segmentation and genetic algorithm for detecting and diagnosing micro calcifications' patterns in digital mammograms. We have investigated and analyzed a number of feature extraction techniques and found that a combination of three features, such as entropy, standard deviation, and number of pixels, is the best combination to distinguish a benign micro calcification pattern from one that is malignant. A fuzzy C Means technique in conjunction with three features was used to detect a micro calcification pattern and a neural network to classify it into benign/malignant. The system was developed on a Windows platform. It is an easy to use intelligent system that gives the user options to diagnose, detect, enlarge, zoom, and measure distances of areas in digital mammograms. The present study focused on the investigation of the application of artificial intelligence and data mining techniques to the prediction models of breast cancer. The artificial neural network, decision tree, Fuzzy C Means, and genetic algorithm were used for the comparative studies and the accuracy and positive predictive value of each algorithm were used as the evaluation indicators. 699 records acquired from the breast cancer patients at the MIAS database, 9 predictor variables, and 1 outcome variable were incorporated for the data analysis followed by the 10-fold cross-validation. The results revealed that the accuracies of Fuzzy C Means were 0.9534 (sensitivity 0.98716 and specificity 0.9582), the decision tree model 0.9634 (sensitivity 0.98615, specificity 0.9305), the neural network model 0.96502 (sensitivity 0.98628, specificity 0.9473), the genetic algorithm model 0.9878 (sensitivity 1, specificity 0.9802). The accuracy of the genetic algorithm was significantly higher than the average predicted accuracy of 0.9612. The predicted outcome of the Fuzzy C Means model was higher than that of the neural network model but no significant difference was observed. The average predicted accuracy of the decision tree model was 0.9635 which was the lowest of all 4 predictive models. The standard deviation of the 10-fold cross-validation was rather unreliable. The results showed that the genetic algorithm described in the present study was able to produce accurate results in the classification of breast cancer data and the classification rule identified was more acceptable and comprehensible.

Keywords: Fuzzy C Means, Decision Tree Induction, Genetic Algorithm, Data Mining, Breast Cancer, Rule Discovery.

1. INTRODUCTION

Breast cancer is one of the major causes for the increased moralitcause many among women especially in developed countries. It is second most common cancer in women. The World Health Organization's International estimated that more than 1,50,000 women worldwide die of breast cancer in year. In India , breast cancer accounts for 23% of all the female cancer death followed by cervical cancer which accounts to 17.5% in India. Early detection of cancer leads to significant improvements in conservation treatment[1]. However, recent studies have shown that the sensitivity of these systems is significantly decreased as the density of the breast increased while the specificity of the systems remained relatively constant. In this work **w**e have developed automatic neuron genetic algorithmic approach to automatically detect the suspicious regions on digital mammograms based on asymmetries between left and right breast image.

Diagnosing cancer tissues using digital mammograms is a time consuming task even highly skilled radiologists because mammograms contain low signal to noise ratio and a complicated structural background. Therefore in digital mammogram, there is still a need to enhance imaging, where enhancement in medical imaging is the use of computers to make image clearer. This may aid interpretation by humans or computers. Mammography is one of the most promising cancer control strategies since the cause of cancer is still unknown[2]. Radiologist turn to digital mammography as an alternative diagnostic method due to the problems created by conventional screening programs. A digital mammogram is created when conventional mammogram is digitized; through the use of a specific mammogram is digitizer or a camera, so it can be processed by the computer. Image enhancement methods are applied to improve the visual appearance of the mammograms. Initially the mammogram image is read from the dataset and partial filter (Combination of Low and high Pass filter) is applied to remove the noise from the image.

Fuzzy C Means clustering with texture based segmentation ,decision tree induction and genetic algorithm techniques were used in efficient detection of cancerous masses in mammogram MRI images . The selected features are fed to a three-layer Backpropagation Network hybrid with FCM,GA (BPN-FCM-GA) for classification and the Receiver Operating Characteristic (ROC) analysis is performed to evaluate the performance of the feature selection methods with their classification results.

In this paper various steps in detection of microcalcification such as i) Preprocessing and enhancement using Histogram Equalization and Parital filter ii) Fuzzy C Means clustering iii) Texture based Segmentation iv) Decision Tree Induction V) Genetic Algorithm V) System is trained using Back Propagation Network VI) FROC analysis is made to find how accurate the detection of cancerous masses in automatic detection system.

2. PREPROCESSING & ENHANCEMENT TECHNIQUES:

One of the most important problems in image processing is denoising. Usually the procedure used for denoising, is dependent on the features of the image, aim of processing and also post-processing algorithms [5]. Denoising by low-pass filtering not only reduces the noise but also blurs the edges. Spatial and frequency domain filters are widely used as tools for image enhancement. Low pass filters smooth the image by blocking detail information. Mass detection aims to extract the edge of the tumor from surrounding normal tissues and background, high pass filters (sharpening filters) could be used to enhance the details of images. PSNR, RMS, MSE, NSD, ENL value calculated for each of 161 pairs of mammogram images clearly shows that Partial low and high pass filter when applied to mammogram image leads to best Image Quality[3].

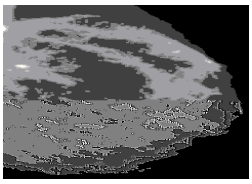

FIGURE 1: Image enhancement using histogram equalization and Partial Filters

3. FUZZY C MEANS CLUSTERING

This algorithm aims at detecting microcalcifications and suspicious areas. In the process of detecting, it may detect other areas that look like a microcalcification. It is up to the user to decide whether the resulting detection is a microcalcification or some other area[4]. The algorithm is simple and based on a fuzzy technique where the size of the microcalcification can be "about the size of a microcalcification." It uses a 8*8 window to scan over the entire digital mammogram and locate microcalcifications or other abnormalities:

WHILE entire 8*8 image has not been examined by 4 window
MOVE 8*8 window to next position
RECORD x,y position and grey level value of pixel with largest grey level in window
IF pixels ***surrounding*** the largest pixel are ***as bright as*** the largest pixel grey level value AND ***outer*** pixels are ***darker*** than the largest pixel grey level value
THEN largest pixel position is the center pixel of a microcalcification
area
END IF
END WHILE

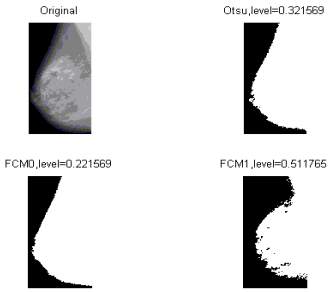

FIGURE 2: segmentation using Fuzzy C means

No. of iteration:36, Time:0.79Seconds

4. FEATURE SELECTION

The main aim of the research method proposed was to identify the effectiveness of a feature or a combination of features when applied to a neural network[5]. Thus, the choice of features to be extracted was important. The following 14 features were used for the proposed method:

average histogram, average grey level, energy, odified energy, entropy, modified entropy, number of pixels, standard deviation,modified standard deviation, skew, modified skew, average boundary grey level, difference, and contrast. The formulas for entropy, energy, skew, and standard deviation were modified so that the iterations started with the first pixel of the pattern and ended at the final pixel. Traditionally, the formulas for these features have iterations starting with the lowest grey level possible and ranging to the highest grey level possible. This modification was done in an attempt to achieve a better classification rate than its traditional version.

Initially, the method determined the ranking of single features from best to worst by using each feature as a single input to the neural network. After this was completed, a combination of features was tested and a best feature or a combination of
features was determined.

1) First Feature Vector (Ten Features): Average histogram, average grey level, number of pixels, average boundary grey, difference, contrast , energy, entropy, standard deviation, and

skew.

2) Second Feature Vector (14 Features): Average histogram, average grey level, number of pixels, average boundary grey, difference, contrast, modified energy, modified entropy, modified standard deviation, and modified skew.The most significant feature or combination of features were selected based on neural-network classification. It was done as follows. We started with a single feature by feeding it to the genetic algorithm and neural network and analyzing the classification rate. If it was increased or unchanged by adding a particular feature, then we included this feature to the input vector. Otherwise, we removed this feature and added another feature to the existing input vector and repeated the whole process again. The total 8190 combinations were investigated and the combinations with the best classification rate were selected for the development of the our CAD system.

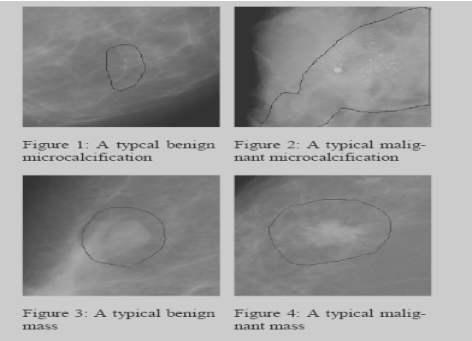

Figure 1: A typcal benign microcalcification

Figure 2: A typical malignant microcalcification

Figure 3: A typical benign mass

Figure 4: A typical malignant mass

5. DECISION TREE INDUCTION METHOD

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems[6]. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5, and Breiman et al.'s CART .A decision tree is a non-linear discrimination method, which uses a set of independent variables to split a sample into progressively smaller subgroups. The procedure is iterative at each branch in the tree, it selects the independent variable that has the strongest association with the dependent variable according to a specific criterion (e.g., information gain, Gini index, and Chi-squared test. In our study, we chose to use J4.8 algorithm as our decision tree method and all subjects according to whether or not they were likely to have breast cancer.

Decision rules	**Fitness value(train data, 453 cases)**	**Accuracy(train data, 453 cases)**	**Accuracy (test data, 246 cases)**
If 5.6 < clump Thickness < 7.2 AND 1.8 < Marginal adhesion < 4.0 AND 3.2 < single Epithelia , 8.6 AND 2.1<normal nucleoid <3.1 THEN class=benign else Malignnant	1	0.9993	0.9878

TABLE 1: Decision Rule with corresponding Accuracy

6. GENETIC ALGORITHM

A genetic algorithm is an iterative procedure until a pre-determined stopping condition (usually the number of generation). Genetic algorithm involves a population of individuals, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. The simple genetic algorithm as a pseudo code is:

Step 1. Generate an initial population of strings randomly.

Step 2. Convert each individual of the current population into If–Then rule.

Step 3. Evaluate each of If–Then rules from training dataset.

Step 4. Select parents for the new population.
Step 5. Create the new population by applying selection, crossover and mutation to the parents.
Step 6. Stop generation if a stopping condition is satisfied, otherwise go to step 3.

a. FITNESS EVALUATION OF RULES

The role of the fitness function is to encode the performance of the rule
numerically. In our study, the objective of the GA method is to find the accurate and general rules among all the rules in the population. Thus, the GA method uses the composite fitness function consisting of accuracy and coverage. To measure the accuracy and coverage of the rule, we use the following definitions: when a rule is used to classify a given training instance, one of the four possible concepts can be observed: true positive (tp), false positive (fp), true negative (tn) and false negative (fn). The true positive and true negative are correct classifications, while false positive and false negative are incorrect classifications. For a two-class case, with class 'yes' and 'no', the four concepts can be easily understood with the following descriptions:
true positive(TP): the rule predicts that the class is 'yes' (positive) and the class
of the given instance is in fact 'yes' (true);
false positive(FP): the rule predicts that the class is 'yes' (positive) but the
class of the given instance is in fact 'no' (false);
true negative: the rule predicts that the class is 'no' (negative) and the class of
the given instance is in fact 'no' (false);
false negative: the rule predicts that the class is 'no' (negative) but the class of
the given instance is in fact 'yes' (true). Using these concepts, we present a very simple fitness function defined as:
Maximize Fitness Function =TP/(TP+FP)

7. NEURAL NETWORK MODEL

We use the popular BPN architecture in building the neural network diagnostic model. Cybenko indicated that one-hidden-layer network is sufficient to model any complex system with any desired accuracy, the designed network model will have only one hidden layer. We use 9 input nodes in the input layer, the initial number of hidden nodes to be tested was chosen to be 6, 7, 8, 9, 10 and the network has only one output node status of the breast cancer or without breast cancer. As Rumelhart [**33**] concluded that lower learning rates tend to give better network result, we decided to incorporate the lower learning rates in our test models. The prediction results of the BPN networks with combinations of different hidden nodes learning rate are summarized in Table 2. As illustrated in Table 2, the {9,6,1} topology with a learning rate of 0.1 gives the best result (i.e. the lowest testing RMSE). The diagnostic result using the designed BPN model stratified by 10-fold cross-validation can be summarized in Table 4. From the result in table 9, we can observe the average correct classification rate is 95.02%. It can be observed that BPN has the highest average correction classification rate in comparison with the decision tree.

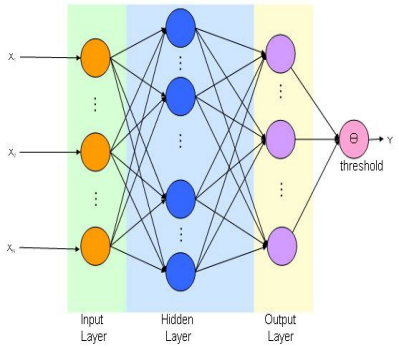

FIGURE 3: Graphical Representation of ANN

Number of Hidden Nodes	Learning Rate	Training RMSE	Testing RMSE
6	0.01 0.03 0.05 0.1 0.5	0.1146 0.0917 0.0749 0.1082 0.5082	0.0600 0.0433 0.0210 0.0089 0.6878
7	0.01 0.03 0.05 0.1 0.5	0.1164 0.0089 0.1015 0.0872 0.6976	0.0768 0.0434 0.0201 0.0111 0.5213
8	0.01 0.03 0.05 0.1 0.5	0.1342 0.0726 0.0947 0.0653 0.6976	0.0692 0.0455 0.0209 0.0128 0.5213
9	0.01 0.03 0.05 0.1 0.5	0.1133 0.0704 0.0472 0.1082 0.5082	0.0762 0.0430 0.0210 0.0092 0.5213
10	0.01 0.03 0.05 0.1 0.5	0.1169 0.0761 0.1069 0.1082 0.5082	0.0633 0.0416 0.0228 0.0146 0.6878

TABLE 2: ANN MODEL PREDICTION RESULT

Prediction Variables used in cancer detection:
1. Clump thickness
2. Uniformity of cell size
3. Uniformity of cell shape
4. Marginal adhesion
5. Single epithelial cell size
6. Bare nuclei
7. Bland chromatin
8. Normal nucleoli
9. Mitoses
10. Class (benign/malignant)

Stability Test Results of GA Model:

Group	1-5					6-10				
Evolution Generation	1	2	3	4	5	6	7	8	9	10
Mean Fitness Value	0.93	0.83	0.89	0.81	0.86	0.85	0.77	0.90	0.96	0.90
Optimal Fitness Value	1	0.976	0.983	0.988	1	0.988	0.984	1	1	1

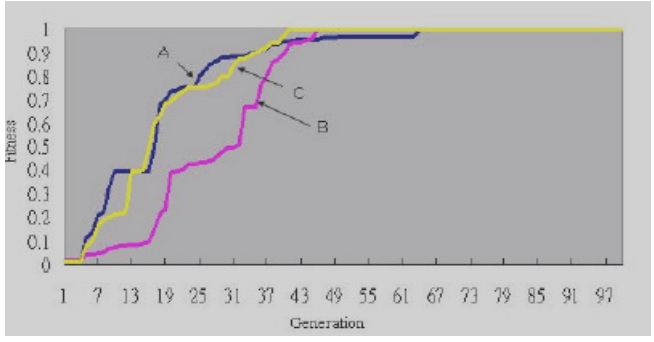

FIGURE 2: Convergence Graph of GA

8. CONCLUSION

In the present study, various data mining models including Fuzzy C Means , decision tree, and artificial neural network were used to compare with the genetic algorithm model by evaluating the prediction accuracy, sensitivity, and specificity. Ineach model, the 10-fold crossover validation was used to compare the results of these 3 models. After testing 10 times, accuracies of Fuzzy C Means was 0.9534 (sensitivity 0.98716 and specificity 0.9582), the decision tree model 0.9634 (sensitivity 0.98615, specificity 0.9305), the neural network model 0.96502 (sensitivity 0.98628, specificity 0.9473), the genetic algorithm model 0.9878 (sensitivity 1, specificity 0.9802). The accuracy of the genetic algorithm was significantly higher than the average predicted accuracy of 0.9612.
In the future, genetic algorithm can be used in combination with many different modifications such as: (1) the use of a different selection method like the steady state selection; (2) the development of adaptive genetic algorithm (AGA) for the parameters varying with populations; (3) the modification of chromosomal encoding; and (4) the development of exclusive fitness functions for different diseases.

Fold	Logistic Regression					Decision Tree					ANNs				
No	Confusion Matrix		Accuracy	SEN	SPE	Confusion Matrix		Accuracy	Sen	Spe	Confusion Matrix		Accuracy	Sen	Spe
1	447	12	0.9628	0.9696	0.9495	441	18	0.9399	0.9483	0.9230	444	15	0.9506	0.9568	0.9387
	14	226				24	216				20	230			
2	446	13	0.9628	0.9716	0.9458	435	24	0.9443	0.9645	0.9104	439	20	0.9442	0.9585	0.9170
	13	227				16	244				19	221			
3	447	12	0.9628	0.9696	0.9497	437	22	0.9384	0.9541	0.9087	443	16	0.9542	0.9651	0.9333
	14	226				21	219				16	224			
4	447	12	0.9670	0.9759	0.9502	437	22	0.9413	0.9583	0.9094	443	16	0.9585	0.9714	0.9341
	11	229				19	221				13	227			
5	447	12	0.9642	0.9717	0.9497	429	30	0.9284	0.9554	0.88	440	19	0.9427	0.9544	0.9201
	13	227				20	220				21	219			
6	446	13	0.9642	0.9737	0.9460	438	21	0.9470	0.9647	0.9142	439	20	0.9442	0.9585	0.9170
	12	228				16	224				19	221			
7	447	12	0.9628	0.9696	0.9495	437	22	0.9442	0.9625	0.9102	443	16	0.9556	0.9672	0.9336
	14	226				17	223				15	225			
8	447	12	0.9628	0.9696	0.9459	437	22	0.9427	0.9604	0.9098	442	17	0.9570	0.9714	0.9303
	14	226				18	222				13	227			
9	447	12	0.9642	0.9717	0.9497	437	27	0.9446	0.9732	0.8941	440	19	0.9484	0.9628	0.9214
	13	227				12	228				17	223			
10	446	13	0.9642	0.9737	0.9460	440	19	0.9642	0.9737	0.9460	442	17	0.9470	0.9567	0.9282
	12	228				20	220				20	220			
Mean			0.9637	0.9716	0.9482			0.9435	0.9615	0.9105			0.9502	0.9628	0.9273
S.D.			0.0013	0.0021	0.0019			0.0089	0.0080	0.0171			0.0057	0.0062	0.0078

TABLE 3: Relationship between Logistic Regression, Decision Tree and ANN

REFERENCES

[1] Barr E, *The handbook of artificial intelligence*, vol. 1-3, William Kaufmann,Los Altos 1982.

[2] Laurikkala J, Juhola M, *A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence*, Comput Methods
Programs Biomed **55** (1998), no. 3, 217-228.

[3]. N. Karssemeijer, *Computer-Assisted Reading Mammograms, European Radiol.*, vol. 7, pp. 743–748, 1997.

[4] L. Mascio, M. Hernandez, and L. Clinton, "Automated analysis for microcalcifications in high resolution mammograms," *Proc. SPIE—Int.Soc. Opt. Eng.*, vol. 1898, pp. 472–479, 1993.

[5.] L. Shen, R. Rangayyan, and J. Desaultels, *Detection and Classification Mammographic Calcifications, International Journal of Pattern Recognition and Artificial Intelligence*. Singapore: World Scientific, 1994, pp. 1403–1416.

[6.] F. Aghdasi, R.Ward, and B. Palcic, "Restoration of mammographic images in the presence of signal-dependent noise," in *State of the Art in Digital Mammographic Image Analysis*. Singapore: World Scientific, 1994, vol. 7, pp. 42–63.

[7.] Y. Chitre, A. Dhawan, and M. Moskowtz, "Artificial neural network based classification of mammographic microcalcifications using image structure features," in *State of the Art of Digital Mammographic Image, Analysis*. Singapore: World Scientific, 1994, vol. 7, pp. 167–197.

[8.] Bosch. A.; Munoz, X.; Oliver.A.; Marti. J., *Modeling and Classifying Breast Tissue Density in Mammograms*, Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on Volume 2, Issue , 2006 Page(s): 1552 – 15582.

[9] Dar-Ren Chena, Ruey-Feng Changb, Chii-Jen Chenb, Ming-Feng Hob, Shou-Jen Kuoa, Shou-Tung Chena, Shin-Jer Hungc, Woo Kyung Moond,*Classification of breast ultrasound images using fractal feature, ClinicalImage*, Volume 29, Issue4, Pages 234-245.

[10] Suri, J.S., Rangayyan, R.M.: *Recent Advances in Breast Imaging, Mammography,and Computer-Aided Diagnosis of Breast Cancer*. 1st edn. SPIE (2006)

[11] Hoos, A., Cordon-Cardo, C.: *Tissue microarray pro.ling of cancer specimens and cell lines: Opportunities and limitations*. Mod. Pathol. 81(10), 1331–1338 (2001)

[12] Lekadir, K., Elson, D.S., Requejo-Isidro, J., Dunsby, C., McGinty, J., Galletly, N.,Stamp, G., French, P.M., Yang, G.Z.: *Tissue characterization using dimensionality reduction and .uorescence imaging.* In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 586–593. Springer, Heidelberg (2006).