Breast Cancer Detection Using Mammography

GULZAR A. KHUWAJA
Faculty of Computer Studies
Arab Open University
P. O. Box 3322, Safat 13033
KUWAIT

Abstract: - The presence of microcalcification clusters in mammograms contributes evidence for the detection of early stages of cancer. In this paper, a low-cost and high-speed neural network based breast cancer detection algorithm is presented. The microcalcifications are extracted with an adaptive neural network that is trained with cancer/malignant and normal/benign breast mammograms and a best accuracy rate of 99% for the classification of cancer/normal/benign is achieved.

Key-Words: - Breast Cancer Detection, Mammogram, Tumor Pattern Classification, Learning Vector Quantization, Medical Imaging

1 Introduction

Breast cancer is a leading cause of death among women, and its incidence is rising. Cancer is a group of diseases in which cells in the body grow, change, and multiply out of control. Usually, cancer is named after the body part in which it originated. Thus, breast cancer refers to the erratic growth and proliferation of cells that originate in the breast tissue. A group of rapidly dividing cells may form a lump or mass of extra tissue. These masses are called tumors. Tumors can be either cancerous (malignant) or non-cancerous (benign).

To enable early detection, the American Cancer Society (ACS) recommends a baseline mammogram for all women by the age of 40. Mammography has been shown to be effective in screening asymptomatic women to detect occult breast cancers and to reduce mortality by as much as 30% [1, 2]. A microcalcification cluster, an early sign of breast cancer that may warrant biopsy, is commonly defined as three or more microcalcifications present in one cm^2 on a mammogram. The width of a microcalcification is less than 2 mm. These clusters are often difficult to detect due to their small size and their similarity to other tissue structures [3].

In order for mass screening to be cost effective, means need to be developed to achieve it with high accuracy and speed. Even if qualified personnel are available, it is difficult for a radiologist to read screening mammograms in large numbers, since most are free of malignant features, and maintaining the required attention level is extremely difficult [4]. Radiologists misdiagnose 10-30% of the malignant cases; two-thirds of which are retrospectively evident in the mammogram. Of the cases sent for surgical biopsy, only 10-20% are actually malignant [5].

In the detection of breast cancer, false negatives may cause a delay in the diagnosis and treatment of the disease while false positives cause unwarranted biopsy examinations. Breast biopsy is an expensive and disconcerting procedure using tissue pathology techniques, although it is less severe with the advent of large core needle biopsy techniques [6]. Therefore, both sensitivity and specificity need to be maximized, with a relatively higher priority on sensitivity, which has a more vital role. A computer-aided diagnosis (CADx) system assisting the radiologist could have a tremendous impact by helping to correctly diagnose the missed malignant cases and reduce the number of unnecessary surgical biopsies [7] and other oversights that may result from poor mammographic image quality, radiologist fatigue, or alternative sources.

In this paper, a low-cost and high-speed neural network based breast cancer detection algorithm is presented. The microcalcifications are extracted with an adaptive neural network that is trained with cancer/malignant and normal/benign breast mammograms and a best accuracy rate of 99% for the classification of cancer/normal/benign is achieved.

2 LVQ Classifier

LVQ (learning vector quantization) is a supervised classifier that was first studied by Kohonen [8]. To classify an input vector, it must be compared with all prototypes. The Euclidean distance metric is used to select the closest vector to the input vector. And the input vector is classified to the same class as the nearest prototype.

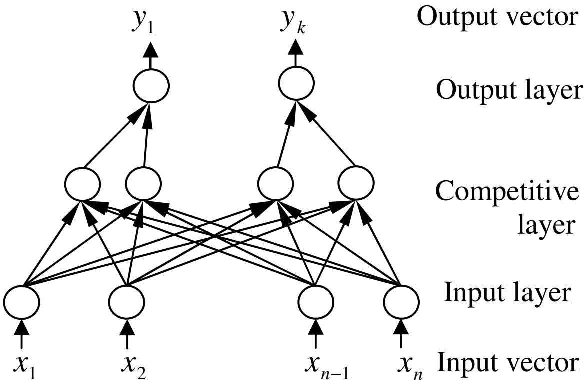


Fig. 1: Architecture of the LVQ Classifier.

The LVQ classifier (Fig. 1) consists of an input layer, a hidden competitive layer, which learns to classify input vectors into subclasses and an output layer, which transforms the competitive layer's classes into target classifications defined by the user. Only the winning neuron of the hidden layer has an output of one and other neurons have outputs of zero. The weight vectors of the hidden layer neurons are the prototypes, the number of which is usually fixed before training begins. The number of hidden neurons depends upon the complexity of the input-output relationship and significantly affects the results of classifier testing. Selection of the number of hidden neurons must be carefully made, as it highly depends on the encompassed variability in the input patterns. Extensive experiments are performed to conduct the suitable number.

For a training set containing *n* input mammograms, each of these images is labeled as being one of *k* classes. The learning phase starts by initiating the weight vectors of neurons in the hidden layer. Then, the input vectors are presented randomly to the network. For each input vector X_j, a winner neuron W_i is chosen to adjust its weight vector:

$$\|X_j - W_i\| \le \|X_j - W_k\|, \quad \text{for all } k \ne i \qquad (1)$$

The weight vector $W_i(t)$ is updated to the next step t+1 as follows:

$$W_i(t+1) = W_i(t) + \alpha(X_j - W_i(t)) \qquad (2)$$

if X_j and W_i belong to the same class

$$W_i(t+1) = W_i(t) - \alpha(X_j - W_i(t)) \qquad (3)$$

if X_j and W_i belong to different classes.

where $0 \le \alpha \le 1$ is the learning rate, which may be kept constant during training or may be decreasing monotonically with time for better convergence [8]. Otherwise, do not change the weights. The training algorithm is stopped after reaching a pre-specified error limit. During the test phase, the distance of an input vector to each processing element of the hidden layer is computed and again the nearest element is declared as the winner. This in turn fires one output neuron, signifying a particular class.

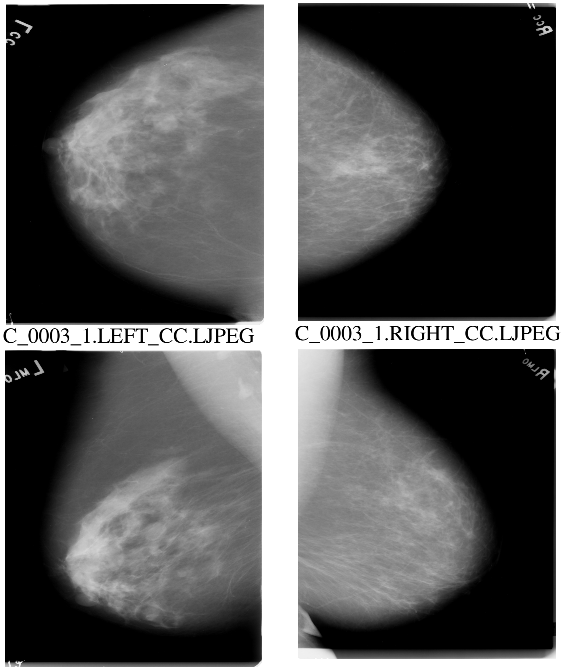
3 Database and Preprocessing

The database selected for this study is Digital Database for Screening Mammography (DDSM) [9], developed at the University of South Florida. The database contains 2620 studies. Each study includes two images of each breast, along with some associated patient information (age at the time of study, ACR (American College of Radiology) breast density rating, subtlety rating for abnormalities, ACR keyword description of abnormalities) and image information (scanner, spatial resolution, etc). Images containing suspicious areas have associated pixel-level "ground truth" information about the locations and types of suspicious regions. Also provided is the software both for accessing the mammogram and truth images and for calculating performance figures for automated image analysis algorithms.

The DDSM is a mixture of normal, benign, benign without callback, and cancer volumes. There are 2620 cases available in 43 volumes. Each volume is a collection of cases of the corresponding type. Normal cases are formed from a previous normal screening exam (pulled from a file) for a patient with a normal exam at least four years later. A normal screening exam is one in which no further "work-up" was required. Cancer cases are formed from screening exams in which at least one pathology proven cancer was found. Benign cases are formed from screening exams in which something suspicious was found, but was determined not malignant (by pathology, ultrasound or some other means). The term benign-without-callback is used to identify benign cases in which no additional films or biopsy was done to make the benign finding. These cases, however, contained something interesting enough for the radiologist to mark.

The DDSM is organized into *cases* and *volumes*. A "case" is a collection of images and information corresponding to one mammography exam of one patient. A "volume" is simply a collection of cases collected together for purposes of ease of distribution. Normally all volumes are available on-line; marathon.csee.usf.edu/Mammography/Database.html. A case consists of between 6 and 10 files. These are an "ics" file, an overview "16-bit PGM" file, four image files that are compressed with loss less JPEG encoding and zero to four overlay files. Normal cases will not have any overlay files.

The four standard views (left/right medio-latral oblique and left/right cranio caudal) from each case of DDSM were digitized on one of four different scanners. Table 1 [9] lists some characteristics for each of these scanners and provides calibration equations to convert pixel values to optical densities. The four images of one case (C-0003) are shown in Fig. 2.


C_0003_1.LEFT_CC.LJPEG C_0003_1.RIGHT_CC.LJPEG

C_0003_1.LEFT_MLO.LJPEG C_0003_1.RIGHT_MLO.LJPEG

Fig. 2: Full breast images of DDSM database.

After digitization, DDSM mammograms were automatically cropped to remove much of the background (non-breast tissue) area. They were then manually processed to darken (digitally zero) pixels in regions that contained patient identifiers and were stored in files using a truly loss-less compression algorithm.

The compressed mammograms (.LJPEG images) are then decompressed and converted to 8-bit resolution Portable Gray Map format (.PGM images) using the provided C program routines [9]. The mammograms (.PGM images) are again cropped manually to leave out any dark space and extract only the breast area using Matlab. Since, for a pattern recognition system, the use of low-resolution images is efficient and practicable [10], the whole set of images is re-sampled as 75×50 pixels. To reduce the image size, a low pass filter is applied to the image before interpolation using the nearest (Euclidean distance) neighbor interpolation method.

4 Algorithm

The most challenging step in the design of a pattern recognition system is the selection of a suitable base model that constitutes its building blocks. The next step is the features selection and extraction method.

4.1 Neural Classifiers

In machine-based detection, a gallery of patterns is first enrolled in the system and coded for subsequent searching. A probe pattern is then obtained and compared with each coded image in the gallery; detection is noted when a suitable match occurs. The challenge of such a system is to perform detection of the pattern despite transformations: location and size of the masses, changes in lighting conditions, common problems of machine vision, and changes due to age. The need is, thus, to find appropriate codings for masses which can be derived from a number of mammograms and to determine in what way, and how well, two such codings shall match before the masses are declared the true positive case.

4.2 LVQ Models

The algorithm based on compact LVQ base model parameters is as follows:

1) Select network parameters:

- Input layer size = Image size (75x50 = 3,750 neurons).
- Training set size (X) = 20 or 60 breast images {X/2 Cancer/Malignant (X/4 LEFT_CC + X/4 RIGHT_CC), X/2 Normal/Benign (X/4 LEFT_CC + X/4 RIGHT_CC)}.
- Number of classes (NC) = 2 (Cancer/ Malignant and Normal/Benign).
- Number of hidden layer neurons (NH) = 2 or 4 or 6 or 10.
- Learning rate (α) = 0.1.
- Set up the target vector which specifies the target class of each pattern in the training set.
- Display update rate =100.
- Arrange the input patterns of the training set as one-dimensional columns in an array (P).
- Number of training epochs (EP) = 1500 or 2500 or 5000.

2) Initialize an LVQ classifier:

- ► Initialization of the weight matrix for competitive layer w1.
- ► Initialization of the weight matrix for linear layer w2.

3) Start training an LVQ classifier based on selected efficient base model parameters with cropped and resized mammograms.
4) Test the trained classifier for cancer/malignant/ normal/benign on both training and test sets of new patients and compute PCCTR and PCCTS (percentage of correct classification for training and test sets respectively).
5) Get the best accuracy rate of cancer/normal/ benign detection and calculate the average and standard deviation of the 50 best networks.
6) Exit.

5 Conclusion

In this paper, we present a low-cost and high-speed neural network based breast cancer detection system that is trained with low-resolution mammograms. The objective is to eliminate bulk storage requirements. The performance is evaluated using randomly chosen 120 digitized mammograms, with 60 cases of pathology proven cancer and 60 normal/benign cases of DDSM (Digital Database for Screening Mammography) database [9]. The tumors are detected based on features extracted with an adaptive learning vector quantization (LVQ) neural network compact architecture. The neural network approach has long been in use for pattern recognition and has recently been adopted for the detection/classification of breast tumors. A best accuracy rate of 99% for the classification of cancer/normal/benign is achieved for 100 DDSM cases. The intent of this work is twofold: a) to compare the breast cancer/normal/benign classification performance of various hidden-layer-size neural networks, and b) to explore the best image resolution for cancer/benign classification. The robustness of the proposed algorithm is demonstrated by calculating standard deviation of 20 best networks.

The main difficulty in the development of the computer-aided detection system is the clinical requirement that the reduction in false-positive biopsies be done without false-negative results. The classification results of the algorithm are very encouraging. The training time was short and the classification rate on both the training data set and testing data set were very high. In breast cancer detection, the cost of false-negative (missed cancer/malignant) is very high, hence, radiologists can use the proposed network system as a computer-aided detection (CADx) tool to obtain a second opinion or to verify their diagnosis. The proposed algorithm could easily be implemented using a commonly available PC.

References:

[1] Tabar, L. Gad, A. Holmberg, L. and Ljungquist, U., 1985, Significant reduction in advanced breast cancer: Results of the first seven years of mammography screening in Kopparberg, Sweden, Diag. Imag. Clin. Med, 54, pp. 158-164.

[2] Fletcher, S. W. Black, W. Harris, R. Rimer, B. K. and Shapiro, S., 1993, Report of the international workshop on screening for breast cancer, J. National Cancer Inst, 85, No. 20, pp. 1644-1656.

[3] Bankman, I. N. Nizialek, T. Simon, I. Gatewood, O. B. Weinberg, I. N. and Brody, W. R., 1997, Segmentation algorithms for detecting microcalcifications in mammograms, IEEE Trans. Inform. Tech. Biomed., 1, No. 2, pp. 141-149.

[4] Rangayyan, R. M. El-Faramawy, N. M. Desautels, J. E. L. and Alim, O. A., 1997, Measures of acutance and shape for classification of breast tumors, IEEE Trans. Med. Imag., 16, No. 6, pp. 799-810.

[5] Giger, M. L., 1993, Computer-aided diagnosis, RSNA Categorical Course Phys., pp. 283-298.

[6] Zheng, Y. Greenleaf, J. F. and Gisvold, J. J., 1997, Reduction of breast biopsies with a modified self-organizing map, IEEE Trans. Neural Networks, 8, No. 6, pp. 1386-1396.

[7] Polakowski, W. E. Cournoyer, D. A. Rogers, S. K. DeSimio, M. P. Ruck, D. W. Hoffmeister, J. W. and Raines, R. A., 1997, Computer-aided breast cancer detection and dagnosis of masses using difference of Gaussians and derivative-based feature saliency, IEEE Trans. Med. Imag., 16, No 6, pp 811-819.

[8] Kohonen T., “Self-Organizing Maps,” *Springer Verlag*, Berlin, Germany, 1995.

[9] Heath, M. Bowyer, K.W. Kopans D. et al, 1998, Current status of the Digital Database for Screening Mammography, *Digital Mammography*, Kluwer Academic Publishers, pp. 457-460.

[10] Khuwaja, G. A., 2002, An adaptive combined classifier system for invariant face recognition, Digital Signal Processing, 12: 21-46.