Classification of Normal, Benign and Malignant Tissues Using Co-occurrence Matrix and Bayesian Neural Network in Mammographic Images

Leonardo de O. Martins, Alcione M. dos Santos, Aristófanes C. Silva and Anselmo C. Paiva
Universidade Federal do Maranhão - UFMA
Av. dos Portugueses, SN, Campus do Bacanga, Bacanga
65085-580, São Luís, MA, Brasil
leonardo@gia.deinf.ufma.br, amirandas@terra.com.br, ari@dee.ufma.br, paiva@deinf.ufma.br

Abstract

This work analyzes the application of the co-occurrence matrix to the characterization of breast tissue as normal, benign or malignant in mammographic images. The method characterization is based on a process that selects, using forward selection technique, from all computed measures which best discriminate among normal, benign and malignant tissues. Then, a Bayesian neural network is used to evaluate the ability of these features to predict the classification for each tissue sample. To verify this application we also describe tests that were carried out using a set of 218 tissues samples, 68 benign and 51 malignant and 99 normals. The result analysis has given an accuracy of 86.84%, which means encouraging results. The preliminary results of this approach are very promising in characterizing breast tissue.

1 Introduction

The breast cancer is the major cause of death by cancer in the female population. It is know that the best prevention method is the precocious diagnosis, what lessens the mortality and enhance the treatment [1]. Every three minutes, a woman is diagnosed with breast cancer, and every 13 minutes, a woman dies from the disease. It is also estimated that one in eight women will be diagnosed with breast cancer in her lifetime, and 1 in 30 will die from it [20]. Mammography is currently the best technique for reliable detection of early, non-palpable, potentially curable breast cancer [1]. The mortality rate from this disease decreased for the first time in 1995, due in part to the increasing utilization of screening mammography [1]. However, radiologists vary in their interpretation of mammograms. In addition, the interpretation is a repetitive task that requires much attention that make it easy to misinterpretation. Therefore, in the past decade there has been tremendous interest in the use of image processing and analysis techniques for Computer Aided Diagnosis (CAD) from digital mammograms, which will replace conventional mammograms in the near future. The goal has been to increase diagnostic accuracy as well as the reproducibility of mammographic interpretation.

CAD systems can aid radiologists by providing a second opinion and may be used in the first stage of examination in the near future, allowing the variability reduction among radiologists interpretation of mammograms [14]. For that, it is important to develop many techniques to detect and recognize suspicious lesions and also to analyze and discriminate them. One of the methods reported for lesions diagnosis from mammograms are the neural networks. Neural networks have been used as a final step for classification, after feature selection. In [25], a neural-genetic algorithm for feature selection in conjunction with neural and statistical classifiers has obtained a classification rate of 85.0% for a test set. Recently, metaheuristics have been applied in identification of suspicious region in mammograms [15], using the border points, enhanced by a genetic algorithm, and nipple position, identified by an ant colony system. The results obtained with a set of mammograms indicate that their method can improve the sensitivity and reliability of the systems for automated detection of breast tumors. A computer aided neural network classification of regions of suspicion on digitized mammograms is also presented in [5], in which a Radial Basis Function Neural Network (RBFNN) is used to complete the classification, fed by features selected out by a new technique based on independent component analysis. Experiments in the MIAS Database have shown a recognition accuracy of 88.23% in the detection of all kinds of abnormalities and 79.31% in the task of distinguishing between benign and malignant regions, outperforming in both cases standard textural features, widely used for cancer
In this paper we intend to investigate the effectiveness of a classification method done through the use of a Bayesian neural network with a set of tissues textural features, computed from the co-occurrence matrix applied to mammographic images.

This work is organized as follows. In section 2, we present the techniques for feature extraction and tissue classification. Next, in section 3, the results are shown and we discuss about the application of the techniques under study. Finally, section 4 presents some concluding remarks.

2 Material and Methods

2.1 Image Acquisition

For the development and evaluation of the proposed system, we used the mini-MIAS [24] database. This database contains left and right breast images for a total of 161 patients with ages ranging from 50 to 65. The spatial resolution of the image is $50 \mu m \times 50 \mu m$. All images are digitized at a resolution of $1024 \times 1024$ pixels and at 8-bit grey scale level. All images also include the locations of any abnormalities that may be present. The existing data in the collection consists of the location of the abnormality (like the center of a circle surrounding the tumor), its radius, breast position (left or right), type of breast tissues (fatty, fatty-glandular and dense) and tumor type if it exists (benign or malignant).

2.2 Bayesian Neural Networks

In this paper, we focus exclusively on feedforward networks with a single hidden layer with $m$ nodes and $k$ outputs, besides do not allow direct connections from the inputs and outputs. The particular form of the neural model we will work is

$$f_k(x, w) = \phi_k\{w_{k0} + \sum_{j=1}^{m} w_{kj}\phi_j\{w_{j0} + \sum_{i=1}^{p} w_{ij}x_i\}\}$$

where $x$ is the input vector with $p$ explanatory variables, $x_i$ and $\phi$ represents the activation function and the set of all weights (parameters), represented by the vector $w$, including input-hidden weights, biases and hidden-output weights. If the neural model is used for classification problem, the output $f_k(x, w)$ is the final value used for classification process.

In the Bayesian approach to learning neural network [4], [16], [17], the objective is to find the weights posterior distribution mode. To obtain the posterior distribution of the weights, we need to specify the prior distribution, which is a probability distribution that represents the prior information associated with the weights of the network, and the data likelihood. Firstly, we will discuss how we choose the prior distribution of the weights.

Many implementations of Bayesian Neural Networks (BNN) use Gaussian distribution, with zero mean and some specified width, as the priors for all weights and biases in the network. To specify the prior distributions, the weights were divided into three separate groups: bias terms, input to hidden weights and hidden to output weights.

We consider a Gaussian prior with zero mean and unknown variance $1/\lambda$, for the input to hidden weights, where $\lambda$ is a precision parameter. Instead of fixing the $\lambda$ value, we regard it as another parameter. We would then call it a hyperparameter to separate it from weights and biases. Now, we need to specify a hyperprior distribution for $\lambda$.

Although there are several ways to implement the required hyperprior, we choose a Gamma hyperprior, with mean and specified shape parameter [2]. This process can be extended, where each input weight have different priors and hyperpriors. This process is called Automatic Relevance Determination (ARD) [21]. Using this prior distribution, it is possible determine the relative importance of the different inputs. The relevance of each input is considered to be inversely proportional to the variance of this distribution.

In order, the prior distribution for hidden to output weights was also considered Gaussian with zero mean and unknown variance $\lambda$. We use Gamma hyperprior with mean and specified shape to hiperparameter $\lambda$.

Finally, all biases terms are then assumed to be distributed according to a gaussian prior with mean zero and variance $\lambda$, where the Gamma hyperprior, with mean and specified shape, was again used to the hyperparameter $\eta$. For a multi-class classification problem, the targets are represented by a binary label vector in which a single component is set to one to denote the correct class and the others component are set to zero.

Consider a network with one output $f_k(x, w)$ for each class. In this case, the outputs of the trained network are interpreted as class probabilities, where it is common to use a generalization of the logistic sigmoid activation function. Therefore, the probability that a target $y_i$ has value $j$ is

$$p(y_i = j \mid x, w) = \frac{\exp(f_j(x, w))}{\sum_k \exp(f_j(x, w))}$$ (2)
Assuming that $(x_1, y_1), \ldots, (x_n, y_n)$ are independents and identically distributed according to multinomial distribution, we have the following likelihood function for the training data

$$P(D \mid w) \propto \prod_{i=1}^{n} \prod_{k=1}^{c} f_k(x_i, w)^{y_i} \quad (3)$$

After observing the data, using Bayes theorem and likelihood, prior distribution is updated to the posterior distribution

$$P(w, \varphi \mid D) \propto \frac{P(D \mid w, \varphi)P(w, \varphi)}{P(D)}$$

$$= \frac{P(D \mid w, \varphi)P(w \mid \varphi)P(\varphi)}{\int \int P(D \mid w, \varphi)P(w, \varphi)\text{d}w\text{d}\phi} \quad (4)$$

The denominator in the Equation 4 is sometimes called normalizing constant that ensures that the posterior distribution integral is equal to one. This constant can be ignored since it is irrelevant to the first level of inference. Therefore, the theorem may also be written as

$$P(w, \varphi \mid D) \propto P(D \mid w, \varphi)P(w \mid \varphi)P(\varphi) \quad (5)$$

Given a training data, to find the weight vector $w^*$, corresponding to the maximum of the posterior distribution, is equivalent to minimize the error-function $E(w)$, which is given by

$$E(w) = -\ln P(D \mid w) + \ln P(w, \varphi) \quad (6)$$

where $P(D \mid w)$ is the likelihood function presented in Equation 3.

In the Bayesian learning, the posterior distribution is used to find the predictive distributions for the target values in the new test case given the inputs for that case as well as inputs and target for training cases [9]. To predict the new output $y_{n+1}$ for new input $x_{n+1}$, predictive distribution is obtained by integrating the predictions of the model with respect to the posterior distribution of the model parameters

$$P(y_{n+1} \mid x_{n+1}, D) = \int \int P(y_{n+1} \mid x_{n+1}, w, \varphi)P(w, \varphi \mid D)\text{d}w\text{d}\varphi \quad (7)$$

The posterior distribution presented in Equation 5 is very complex and its evaluation requires high-dimensional numerical integration, then it is impossible to compute it exactly. In order to make this integral analytically tractable, we need to introduce some simplifying approximations.

There are different approaches to calculate the posterior distribution. In [17] is used a Gaussian approximation for the posterior distribution, while in [21] is introduced a hybrid Monte Carlo method. Another approach to approximate the posterior distribution uses Markov Chain Monte Carlo method (MCMC) [22]. For a review of these methods see, for instance [3].

The idea of MCMC is to draw a sample of values $w(t), t = 1, \ldots, M$ from the posterior distribution of network parameters. In this work, we used Gibbs sampling [8] to generate samples to the posterior distribution.

Gibbs sampling is perhaps the simplest MCMC method and it is applicable when the joint distribution is not known explicitly, but the full conditional distribution of each parameter is known. In a single iteration, Gibbs sampling involves sampling one parameter from full conditional distribution given all other parameters.

Gibbs sampling requires that all conditional distributions of the target distribution can be sampled exactly. When the full conditional distribution was unknown, it was used the Metropolis-Hasting algorithm [12] or adaptive sampling procedure [10]. For more details of this method, see [7].

We can observe that the integral of Equation 7 is the expectation of function $f_k(x(n+1), w)$ with respect to the posterior distribution of the parameters. This expectation can be approximated by MCMC, using a sample of values $w(t)$ drawn from the posterior distribution of parameters. These values are then used to calculate

$$y_{n+1} \approx \frac{1}{M} \sum_{t=1}^{M} f_k(x(n+1), w(t)) \quad (8)$$

### 2.3 Co-occurrence Matrix

The co-occurrence matrix or Spatial Gray Level Dependence Method – SGLDM is a texture analysis technique that has been frequently used in 2D image segmentation and identification [13], [18] and [11]. Specific applications to medical images can be found in [19] and [18].

Co-occurrence matrix displays the grayscale spatial-dependency along different angular relationships, horizontal, vertical and two diagonal directions in image. A co-occurrence matrix is specified by relative frequencies $P(i, j, r, \Theta)$ with which two pixels, separated by distance $r$, occur in a texture along the direction of angle $\Theta$, one with grayscale $i$ and the other with grayscale $j$. A co-occurrence matrix is therefore a function of distance $r$, angle $\Theta$ and grayscale. It is assumed that a textured image $Y$ is defined over a $M = N_x \times N_y$ finite lattice $\Omega$:

$$\Omega = \{(k, l), (m, n) : 1 \leq (k, l), (m, n) \leq M\} \quad (9)$$

The co-occurrence matrices used in this work show texture characteristics along $0^\circ, 45^\circ, 90^\circ$ e $135^\circ$ as follow:

$$P(i, j, 0^\circ) = \#\{(k, l), (m, n) \in \Omega), |k - m| = 0, |l - n| = r, Y(k, l) = i, Y(m, n) = j\} \quad (10)$$
et al. just 5 of the 13 measures proposed by Haralick texture are produced. Therefore, in this paper we used co-occurrence matrix.

R matrices. The frequency normalising constant necessary to normalise the frequencies of the occurrence due to the angular relationships. To overcome this, it is

where # denotes the number of elements. It is observed that the co-occurrence matrix is symmetrical because of

\( P(i, j, r, \Theta) = P(j, i, r, \Theta) \).

The total number of pairs of compared pixels is different due to the angular relationships. To overcome this, it is necessary to normalise the frequencies of the occurrence matrices. The frequency normalising constant \( R \) is explicitly defined as the frequency of pairs of compared pixels in each co-occurrence matrix.

Based on above co-occurrence matrix, many different texture are produced. Therefore, in this paper we used just 5 of the 13 measures proposed by Haralick et al. [11] to perform pattern recognition based on co-occurrence matrix \( P \). The measures used in this work are: Contrast, Homogeneity, Inverse Difference Moment, Entropy and Energy.

### 2.4 Selection of Most Significant Features

Our main objective is to identify the effectiveness of a feature or a combination of features when applied to a bayesian neural network. Thus, the choice of features to be extracted is important.

Forward selection is a method to find the “best” combination of features (variables) by starting with a single feature, and increasing the number of used features, step by step [6]. In this approach, one adds features to the model one at a time. At each step, each feature that is not already in the model is tested for inclusion in the model. The most significant of these feature is added to the model, so long as it’s p-value is below some pre-set level.

### 2.5 Evaluation of the classification method

Sensitivity and specificity are the most widely used statistics to describe a diagnostic test. Sensitivity is the proportion of true positives that are correctly identified by the test and is defined by \( Se = TP/(TP+FN) \). Specificity is the proportion of true negatives that are correctly identified by the test and is defined by \( Sp = TN/(TN+FP) \). Where FN is false-negative, FP is false-positive, TN is true negative and TP is true positive diagnosis.

### 3 Experimental Results

The regions of interest (ROI) were manually extracted from each image based on the information provided by the MIAS database. From this database, we selected 119 abnormal (68 benign and 51 malignant mammograms) and 99 normal tissues samples from each group, summing up 218 mammograms. To each mammogram, a ROI was manually selected, containing the lesion, in the case of the benign and malignant mammograms, as shown in Figure 1. For the normal mammograms, the ROI was randomly selected. Only the pectoral muscle was not considered as a possible ROI, although tissue and fatty tissue were. If the tissues had different sizes, it was rescaled each ROI. Therefore, they were resized to 50x50 pixels.

Figure 1. Illustration of a ROI example of malignant tissue.

To compute the co-occurrence matrix we used four directions \( 0^\circ, 45^\circ, 90^\circ \) e \( 135^\circ \). For each direction we compute the measures for the 1 pixel distance. These parameters were empirically defined, after exhaustive tests. We used the five measures defined in the Section 2.3. This method had the number of gray levels equal to 8, 16, 32, 64, 128 and 256. Using these parameters, the co-occurrence matrix generated 120 measures for each ROI (4 directions \( \times \) 1 distance \( \times \) 5 measures \( \times \) 6 number of quantization levels).

To make feasible the computation we need to select from all the obtained measures which were the minimum set that has the power to discriminate benign from malignant tissues. To do it, we used the forward technique that reduced the number of variables to 17. To reduce the multiple
linear correlation among the variables, we examine pair-wise correlations between variables, and remove one from each pair with high correlation (absolute correlation value above 0.9). Thus, the numbers of variables decreased to 8. This shows that for the used database these were the more adequate characteristics to describe the tissues texture.

We generated a BNN with 8 inputs, a single-layer feedforward with a fixed number $m$ of hidden nodes and three output nodes. Several values for $m = 5, 6, \ldots, 18$ were used in the experiments, and we report just the best one, that was obtained with $m = 12$.

The nonlinear activation function used for the hidden units was the logistic sigmoid, which produces an output between 0 and 1. For the output units, we used the logistic sigmoid function, but Equation 2 is also used because these outputs are probabilities and must sum to one.

We used the Gaussian prior distribution as described in Section 2.2, with three separate weight-groups. The prior over network parameters are

$$
\begin{align*}
&w_i \mid \lambda_i \sim N(0, \lambda_i^{-1}), \quad i = 1, \ldots, 96. \\
&v_j \mid \lambda_j \sim N(0, \lambda_j^{-1}), \quad j = 1, \ldots, 36. \\
b_k \mid \lambda_k \sim N(0, \lambda_k^{-1}), \quad k = 1, \ldots, 15. 
\end{align*}
$$

where $w_i$ represents the input-hidden weights, $v_j$ the hidden-output weights and $b_k$ the biases terms.

A convenient form for the hyperprior distributions is vague Gamma distribution. Here, we considered all hiperparameters distributed according Gamma distribution with the scale and shape parameter equal to 0.001. The priors for different parameters and hiperparameters are all independent.

The software WinBUGS [23] was used to implement the Bayesian neural network. Through WinBUGS, we specified the model described in Section 2.2. Next, the software simulated the posterior distribution values for each parameter of interest, using the Metropolis-within-Gibbs procedure. We computed a single chain of a MCMC sampler in WinBUGS for each parameter of interest (weights and bias). We simulate 20000 iterations, and discarded the 10000 first in each sequence.

The posterior distribution samples for the model parameters were used to estimate the predictive distribution for the new test inputs. For each iteration $t$, the BNN has parameters $w(t)$ and produces an output $y(t)$, for an input vector $x$. Thus, for each test sample, we calculate the arithmetic mean of the M network outputs, according to Equation 8. As we can see in Section 2.2, this value is a good approximation of Equation 7.

The experiment was configured with 180 training samples and 38 samples for tests. For the training, we obtained an accuracy of 100.0%. Table 1 shows the results obtained with the proposed method for the tests. Based on the Table 1, we can see that the method obtained a mean success rate of 86.84% on discriminating malignant from benign and normal tissues.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Benign</th>
<th>Normal</th>
<th>Malignant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>8</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Correct</td>
<td>5</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>Correct (%)</td>
<td>62.5</td>
<td>95.0</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Table 1. Detailed accuracy analysis in the breast tissue characterization.

To compute sensitivity and specificity more precisely, we divided results obtained in this multiclass classification problem in three binary classification problems. Thus, we examined the efficiency of the proposed method in classify between benign vs. normal, malignant vs. normal and malignant vs. benign tissues. Tables 2, 3 and 4 show the grouped results. Examining Table 2, we can see that the method has a sensibility of 83.33% and a specificity of 95%, considering benign as “positive” and normal as “negative”. For Table 3, which considers only malignant and normal tissues, the method obtained an accuracy of 100%. Finally, considering malignant as “positive” and benign as “negative” the method obtained a sensibility of 90% and a specificity of 71.42%.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Benign</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Normal</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>Correct (%)</td>
<td>83.33</td>
<td>95.0</td>
</tr>
</tbody>
</table>

Table 2. Detailed accuracy analysis in the benign vs. normal t tissue characterization.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Malignant</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malignant</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Correct (%)</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 3. Detailed accuracy analysis in the malignant vs. normal t tissue characterization.

The number of samples studied in our dataset is not sufficiently large enough to allow us to reach definitive conclusions, but preliminary results from this work are very encouraging, demonstrating the potential of BNN use, together with the co-occurrence matrix features, for multiple variables classification and its effectively to
discriminate benign from malignant and normal tissue in mammography.

4 Conclusion

Based on these results, we have observed that such measures provide significant support to a more detailed clinical investigation, and the results were very encouraging when tissue were classified with Co-occurrence matrix and Bayesian Neural Networks. Nevertheless, there is the need to perform tests with a larger database and more complex cases in order to obtain a more precise behavior pattern.

Despite the good results obtained only by analyzing the geometry, further information can be obtained by analyzing the geometry. As a future work, we propose a combination of texture and geometry measures for a more precise and reliable diagnosis. Also we need to compare the BNN results with other neural networks to assert its effectively.

References


<table>
<thead>
<tr>
<th>Tissue</th>
<th>Malignant</th>
<th>Benign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malignant</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Benign</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Correct (%)</td>
<td>90.0</td>
<td>71.42</td>
</tr>
</tbody>
</table>

Table 4. Detailed accuracy analysis in the malignant vs. benign tissue characterization.