We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

5,000
Open access books available

125,000
International authors and editors

140M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Chapter

Detection of Breast Cancer in Mammograms through a New Features Technique

Anwar Yahy Ebrahim

Abstract

This research proposes a new framework for detection of breast cancer in mammograms. It extracts certain dynamic features to distinguish between benign and malignant mammograms. To this aim, this framework uses set of various techniques. First step we have achieved improvement on breast mammogram to improve the image accuracy based on this framework, after new method has been used for features extraction. New methods named Sparse Principal Component Analysis and Weighted Sparse Principal Component Analysis are used to select the distinctive features of the mammograms. The analyzed mammograms are then identified as benign or malignant through codebook technique is more efficient than other on the MIAS data set. The proposed framework tested on MIAS data set achieved an overall classification accuracy of 98% with codebook classifier for sequential selection of benign and malignant mammograms. Suggested method achieves good results when we have verified on various mammograms.

Keywords: chest cancer, mammograms feature extraction, weighted features, codebook design technique

1. Introduction

There are a number of renowned and probable causes for chest cancer. These can be split into seven broad classes: hormonal factors, age, proliferate chest disease, family history of chest cancer, lifestyle factors and [1–5]. Estimates show with the development of technology, radiation scientists have the opportunity to advance their interpretation of image using computer technology capabilities that can develop image resolution from mammography [6–11]. A variety of computer assisted diagnostic systems were proposed such as [12, 13]. In this paper, enhanced Principle Component Analysis (PCA) was used to extract features. Although PCA has been widely applied in the area, but the features considered in this study have not been extracted before [14]. Further, these extracted features are reduced to the best features only. This process is accomplished by two variations of PCA as Sparse Principle Component Analysis (SPCA) [15, 16] and Weighted Sparse Principle Component Analysis (WSPCA). The choice of the (ideally “small”) number of principal components (PCs) to include into the description of the data without losing too much information was somewhat arbitrary [14]. Codebooks represents is final optimized codebook for samples will be generated. It can represent the attributes of the mammograms images more adequately. Proposed technique realized
quite perfect. Project is ordered displayed in stage first. Stage second presents related work. Stage third defines the suggested method. Stage forth contains experimental outcomes and conclusion is presented in stage fifth.

2. Proposed technique

The projected method is split into four major phases as presented in Figure 1. The first phase is representing enhancement by applying histogram equalization, the second phase is representing feature selection, the third phase is representing codebook and the final phase is representing Design Classifier. Every part of these four phases is defined below one after another.

2.1 Enhancement for image

In this phase, the improvement is focused in flat regions avoid over development decreased influence of edge shadowing.

2.2 Features extraction

Features play an important applied DCT projected method.
2.2.1 Discrete cosine transform (DCT)

Features Discrete cosine transform (DCT) is applied for converting the signal into its frequency parts. DCT has the property of separability and symmetry. 2-Dimensional DCT of the input is presented by the following equation:

\[ C(u, v) = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos \left( \frac{\pi(2x + 1)u}{2N} \right) \cos \left( \frac{\pi(2y + 1)v}{2N} \right) \]

where 0 \( \leq u \leq N \), & 0 \( \leq v \leq N \), and \( a(u)a(v) = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } u,v \neq 0 \\ \sqrt{\frac{2}{N}} & \text{for } u=v=0 \end{cases} \)

2.3 Feature selection

In the past, researchers used to reduce the dimensions apply PCA here. Each PC is basically a linear combination of all the original features. This makes the results difficult to interpret [14, 16]. Various approaches have been attempted to overcome this problem. We present a novel technique called WSPCA applying LASSO (elastic net) to generate modified PCs with sparse loadings. Important features are selected based on their weights. The aim behind is to use WSPCA to construct a regression framework in which PCA is reconstructed exactly, and use LASSO to construct modified PCs with sparse loadings. Then important features are selected with adaptive feature's weights to find the best loading vector corresponding the features to achieve high accuracy. The uncorrelated linear combinations are called principal components, which express maximal variations in the data. This provided the researchers with a method of transforming the original high-dimensional dataset into one of the much lower dimension. This method was devised inevitably at the cost of some information loss (variance) and limited ability to interpret new variables and analysis. SPCA can successfully derive sparse loadings.

Despite of its positive aspects, SPCA is not efficient in identifying important features with high accuracy. It also lacks a better step to choose its regulation parameter [14, 16]. WSPCA uses strict criterion and flexible control for selected the important features. To fit our WSPCA models for both features weights expression arrays and regular multivariate features, an efficient algorithm is proposed. In addition, we propose a novel form to calculate the total difference of the modified PCs. In this study, the algorithm for WSPACA in parallel to PCA and SPCA is presented in detail with example: let DCT features (variables) \( F = (F_1, F_2, \ldots F_p) \) represent a p-dimensional random vector with a multivariate normal distribution. It is possible that some features correlate with one another. For instance, if the variables \( F_1 \) and \( F_2 \) are highly correlated, such that the correlation index between \( F_1 \) and \( F_2 \) approaches 0.9, then either \( F_1 \) or \( F_2 \) could be eliminated from the analysis as its role is duplicated by the other. By doing this, the basis of the original features is altered to a more efficient set by using linear combinations. In the general p-dimensional case, this leads to a candidate set of new features. The explained steps are presented in Algorithm 1.
Algorithm 1

Step 1: Suppose A beginning at V [1: k], the loadings of the headmost k (PCs).

Step 2: Assumed a constant A = [\(\alpha_1\ldots\alpha_k\)], fix the next elastic net issue

\[
X_W = \sum_{j=1}^{n} W_j X_j, j = 1, \ldots, n
\]

(2)

Step 3: \(\beta_{\text{WSPCA}} = \left(\alpha_j^T X_W^T X_W - \frac{\lambda}{\pi}ight) + \text{Sign} (\alpha_j^T X_W^T X_W), j = 1, \ldots, k
\]

(3)

Step 4: For a fixed \(\beta_j = [\beta_{\text{SPCA}1}, \ldots, \beta_{\text{SPCA}k}]\), PCA can be found via compute the SVD of the features matrix, calculate the SVD of \(X_W^T X_W = UDV^T\),

(4)

then update A = UVT.

(5)

Step 5: reiterate Steps 4–5, until concourse.

Step 6: Normalization \(\hat{V}_j = \frac{\beta_{\text{WSPCA}j}}{\|\beta_{\text{WSPCA}j}\|_2}\)

(6)

In step 1, the presented PCs are the linear combinations of all original features, V is the response vector (nonzero components) and it is less than or equal to k, given an integer k with \(1 \leq k \leq p\). In Step 2, A is a vector matrix. In Step 3, assumed variables of X are presented in \((n \times p)\) matrix, where n rows represent an independent feature from features (number of observations) and p is the number of variables (dimensions), where is spare coefficients, j be the predictors for nonzero entries, is feature vector, \(XTX\) is represent (covariance matrix) transpose for vector matrix by row vector of features, where represents the norm in the constraint. In the present research, in order to find the optimal number of features, \(\lambda\) is penalty by directly imposing a constraint on PCA and \(\lambda, j = 0\) call SPCA criterion r. B = \([\beta0, \beta1, \ldots, \beta k]^T\), where its regression coefficients represent the optimal minimizing. In Step 4, SVD is a singular value decomposition, UD are PCs, the columns of \(V^T\) are the consistent loading of the PCs eigenvectors, V diagonalizes the covariance matrix \(XTX\), U are called Eigen values of the covariance matrix, D is the diagonal matrix, which has the eigenvvalues of covariance matrix. \(XTX\) and V are the Eigen—genes, which represent the sparse loading of feature matrix. In Step 6, \(W_j\) is weighted features, and, \(\beta_j = [\beta_{\text{SPCA}1}, \ldots, \beta_{\text{SPCA}k}]\). Then \((XW)\) was calculated, which represents weighted feature matrix. Where X is a new feature matrix of SPCA and represents eight types of features.

Coefficients for WSPCA technique were obtained by minimizing both SPCA and weighted feature matrix [17, 18]. In Step 7, represents highly correlated by weighted features among all features, is penalty by directly imposing an constraint on PCA and \((\lambda, j = 0)\), represents to exclude redundant features with very little variation from other features that sufficiently represents it. This is where adaptive weights were used for penalizing different coefficients in the 1 penalty, Here, we can ignore the penalty part in calculating Step 8.

Then, \(AW = UVT\) was updated where PCs were selected for displaying the selected features. Thus, a large dimensionality decrease was realized. Then after \((V_j)\), normalization was calculated for approximated weighted sparse principal components. Step 9 was where \(\beta_{\text{WSPCA}}\) was the WSPCA coefficient.

2.4 Codebook design

After representing each set of features, hierarchical clustering groups the features selection based on similarity to build a hierarchy of clusters. This clustering approach starts with each object as a single class and merges objects into the classes until all objects are in one cluster [18, 19]. The proposed technique needs to define a dimension measure allowing comparison of two classes. The operation of a hierarchical clustering is illustrated in Figure 2.

As an example, seven labeled patterns are shown in Figure 2a, in this research these seven labeled pattern can be consider as seven fragmented windows, which is
then group together in a single cluster. Figure 2b represents the binary tree corresponding to the patterns in Figure 2a. In the binary tree, each patterns are the leaves, each branching points are the similarity between sub-trees. Horizontal cuts using different line patterns in the tree represents classes.

The distance between the two classes can be calculated as the minimum, maximum or the average of the dimensions between attributes of patterns in different clusters. This research employed the average-link method for clustering. In this method, the distance between two categories is defined as the average of the dimensions between all the objects in the two categories. This method is expressed by the next equation.

$$\text{Dist}(c_i, c_j) = \frac{\text{avg}}{x, y \in c_i, y \in c_j} \text{Dist}(x, y)$$

where, $c_i$ and $c_j$ be two categories. Dist defines the dimension between $c_i$ and $c_j$.

In addition, since the number of classes for each mammogram is not known, this study uses the distance criterion to represent the number of cluster. For each mammogram the proposed technique generates the clusters from the important features. In this research, the important features clusters are also termed as codebooks.

3. Outcomes and discussion

We have applied widely presented datasets MIAS [20]. The database image of 69 mammograms were being benign, 54 malignant also 207 normal Improvement has been done by histogram equalization. Outcomes have been display in Figure 3. Once the codebook for important features are generated, the proposed technique
sorts the classes according to the cardinality and keeps only those classes which have sufficient number of features. As a codebook produced from feature selection are illustrated in Figure 3, respectively.

Figure 3. Results by histogram equalization (a, a1); original images (b, b1); and histogram equalized images.

Figure 4. Mammograms: codebook obtained from the important features on a mammograms sample.
In codebook there are different number classes. Each class contains relatively homogeneous groups of similar forms, which are dissimilar to elements in the other classes. These classes are separated by the black window in the codebook as illustrated in Figure 4.

Once the codebook is generated for each mammograms sample, the next step is to determine how to use this information to represent mammograms sample recognition as discussed in the following section.

4. Verification

These codebooks contain different information about a mammograms image and complement each other. It would therefore be a good idea the codebooks to compare two mammogram images. When two mammogram images are compared, the proposed technique computes the distance between them using their codebooks. The final dimension between the two mammogram samples is calculated as a weighted combination of the two distances (Table 1).

<table>
<thead>
<tr>
<th>Techniques</th>
<th>SPCA technique (%)</th>
<th>WSPCA technique (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>88</td>
<td>89</td>
</tr>
<tr>
<td>Bayesian</td>
<td>89</td>
<td>91</td>
</tr>
<tr>
<td>Decision tree</td>
<td>94</td>
<td>95</td>
</tr>
<tr>
<td>Codebook design</td>
<td>96</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 1. Comparison of achievement measurement of various classifiers with SPCA and WSPCA techniques.

5. Conclusion

Suggested method is improved for test the breast cancer from mammograms. This technique achieves this testing in multiple stages. The preprocessing stage on improve image accuracy. Features selection by SPCA and WSPCA has been achieved. Codebook generated for each mammograms sample represent classify as a normal or nonnormal. The tests display projected method provides especially perfect outcomes.

Author details

Anwar Yahy Ebrahim
Department of Computer Science, Babylon University, Babylon, Iraq

*Address all correspondence to: anwaralawady@gmail.com

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
References


Techniques Accuracy (%) Sensitivity (%) Specificity (%) KNN 76.2 77.2 77.5 Neural Network 85.3 84.1 85.3 SVM 86.3 87 87.3 Bayesian 87.3 89.3 89.6


Detection of Breast Cancer in Mammograms through a New Features Technique
DOI: http://dx.doi.org/10.5772/intechopen.89062


