Classification of Mammographic Breast Density Using a Combined Classifier Paradigm

Keir Bovis and Sameer Singh
PANN Research, Department of Computer Science, University of Exeter, Exeter, UK

Abstract. In this paper we investigate a new approach to the classification of mammographic images according to breast type. The classification of breast density in this study is motivated by its use as prior knowledge in the image processing pipeline. By utilising this knowledge at different stages including enhancement, segmentation and feature extraction, its application aims to increase the sensitivity of detecting breast cancer. Our implemented discrimination of breast density is based on the underlying texture contained within the breast tissue apparent on a digital mammogram and realised by utilising four approaches to quantifying the texture. Following feature extraction, we adopt a variation on bootstrap aggregation ('bagging') to meet the assumptions of independence in data representation of the input data set, necessary for classifier combination. Multiple classifiers comprising feed-forward Artificial Neural Network (ANN) are subsequently trained with the different perturbed input data spaces using 10-fold cross-validation. The set of classifier outputs, expressed in a probabilistic framework, are subsequently combined using six different classifier combination rules and the results compared. In this study we examine two different classification tasks; a four-class classification problem differentiating between fatty, partly fatty, dense and extremely dense breast types and a two-class problems, differentiating between dense and fatty breast types. The data set used in this study is the Digital Database of Screening Mammograms (DDSM) containing Medio-Lateral Oblique (MLO) views for each breast for 377 patients. For both tasks the best combination strategy was found using the product rule giving an average recognition rate on test of 71.4% for the four-class problem and 96.7% for the two-class problem.

1 Introduction

A report from the National Cancer Institute (NCI) estimates that about 1 in 8 women in the United States (approximately 12.6 percent) will develop breast cancer during their lifetime [17]. Government sponsored mass-screening mammography programs have been proposed as an effective method of increasing survival time for women with breast cancer [1], but the application of Computer Aided Detection (CAD) within screening programs is still to be addressed. To this end, much research has taken place for the development of CAD techniques and systems. For a radiologist interpreting a benign mammogram, there exists an extremely wide variation in the mammographic appearance of the breast. Within the mammogram, radiographically visible density includes ducts, lobular elements, and fibrous connective tissue. The fibrous connective tissue can be of two types, intralobular or extralobular tissue, and this latter tissue type is seen as the major component of gross density variation in mammograms. Breast density is an important factor in the interpretation of a mammogram. In a breast that is considerably dense, the sensitivity of mammography for the early detection of malignancy and large cancers is reduced because of the difficulty in locating ill-defined cancers within an opaque uniform background. The American College of Radiology (ACR) Breast Imaging Reporting and Data System (BIRADS), identifies four major groups for classifying breast density (Kopans [7]): (1) predominantly fat; (2) fat with some fibroglandular tissue; (3) heterogeneously dense.; (4) extremely dense. Examples of these breast types are shown in Figure 1.

Previous work has focused on using the underlying texture to discriminate between breast types. In their study, Miller and Astley [13] investigated texture-based discrimination between fatty and glandular breast types, exper-

Figure 1. Mammograms with differing mammographic breast densities (a) predominantly fat; (b) fat with some fibroglandular tissue; (c) heterogeneously dense.; (d) extremely dense.

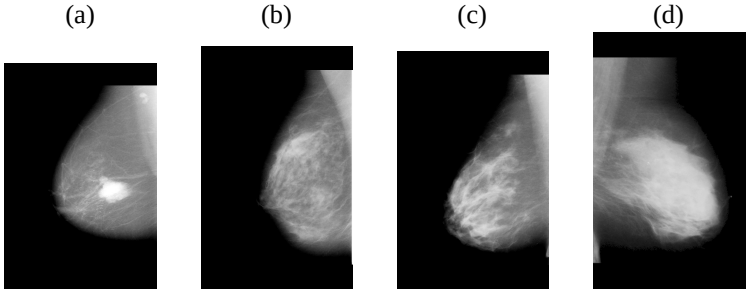
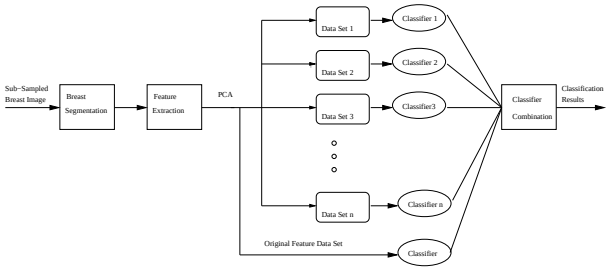


Figure 2. Experimental method overview.



imenting with granulometric techniques and Laws texture masks. Taylor et. al. [16] similarly investigated the classification of fatty and dense breast types using an automated method of extracting the Region Of Interest (ROI) based on texture.

In this study we examine two different classification tasks; a four-class classification problem differentiating between breast densities following the BIRADS classification and a two-class problem, differentiating between dense and fatty breast types. In fulfilling these classification objectives we extract four groups of texture features from segmented digital mammograms. To improve the performance and robustness of the classifiers we use classification combination rules proposed by Kittler et al. [6]. The remainder of this paper is organised as follows; Section 2 describes the experimental method used including feature extraction, data dimensionality reduction, classifier training, testing and combination; Section 3 details the experimental results and finally conclusions from the study are discussed in Section 4.

2 Experimental Method

We evaluated 377 mammograms from the Digital Database of Screening Mammograms (DDSM) [10]. Accompanying each mammogram is a rating of its density according to the BIRADS system determined by an expert radiologist. Of the 377 mammograms our data set is split as follows; predominantly fatty (n=74); fat with some fibroglandular tissue; (n=81) heterogeneously dense; (n=96) extremely dense (n=126). The images vary in size and are converted to an 8-bit grey scale level. A block diagram identifying the major components of the proposed systems is given in figure 2.

2.1 Image Pre-processing

The original grey scale image is initially aligned on the y-axis such that the image is orientated with the nipple pointing right. It is then sub-sampled by a factor of four to reduce the computational complexity for the subsequent breast extraction component.

2.2 Breast Segmentation

As our approach aims to utilise the whole breast for texture feature extraction, a mechanism for segmenting the breast from its background is required. Previous studies in breast/background segmentation [8,12] have difficulties due to the inherent noise within digitised mammograms. We adopt the technique proposed by Chandrasekhar and Attikiouzel [4] with an additional step to trim the profiles, removing the top 20% and bottom 10% of the images to facilitate the removal of poor segmentation that might still include noise.

2.3 Feature Extraction

Previous approaches to classifying breast type have examined the underlying texture within the breast. In this study we employ four approaches for determining texture; 1) By constructing Spatial Grey Level Dependency (SGLD) matrices [5] using the directions $\{0^o, 45^o, 90^o, 135^o\}$ and pixel distances {2, 4, 6}, we extract 15 features. These features include angular second moment, contrast, correlation, inverse different moment, sum average, sum variance, sum entropy, entropy, difference average, difference variance, difference entropy, information measure of correlation I, information measure of correlation II, inertia, variance ; 2) Following the application of the Fourier transform, we extract the total spectral energy from 10 equidistant analysing rings from the power spectrum [11]; 3) By convolving each mammographic image with each combination of Laws' texture masks [9], we extract the

total texture energy for this mask combination for use as a feature. 4) Following application of the Discrete Wavelet Transform (DWT), four features (standard deviation, mean, skewness and kurtosis) characterising the distribution of wavelet coefficients are extracted from 3 different sub-bands and at 3 different scales of the transformed image. In addition to these four groups of texture features, we extract a series of statistical features; entropy, standard deviation, mean, skewness and kurtosis were extracted from the grey scales comprising the segmented breast image; a circularity shape feature [14] and the fractal dimension using the Hurst coefficient described by Russ [15].

Following feature extraction we use Principal Component Analysis (PCA) to reduce the dimensionality of the data set of 316-dimension feature vector. On applying PCA we select the first *n* components by analysing the eigenvalue spectrum and determine a cut-off point after which the eigenvalues level off at small values. In this study we select the components corresponding to the first thirty eigenvalues.

2.4 Classifier Training /Testing and Combination

To improve the performance and robustness of our developed system, we choose to implement a classifier combination paradigm such that we combine the decision of *n* component classifiers on test using combination rules proposed by Kittler et al. [6]. The assumption of conditional independence between each component classifier is an important aspect of the combination strategy. The combination framework is based on the constraint that each classifier uses its own representation of the input space. This is achieved by training the set of component classifiers with unique perturbations to their training and validation sets thus each classifier forms an independent representations of the input space. Our method is motivated by the frequently cited bootstrap aggregation or 'bagging' [3] method of perturbing the input space. The resultant 'Bagging Predictor' combines the classification decision on test, from *n* component classifiers, each trained individually on a training set created using a bootstrap approximation with replacement. A critical factor in whether bagging will improve the accuracy is the stability of the procedure used for constructing the component classifier. Improvement will occur for unstable procedures where a small change in the data set can result in large changes in the classifier accuracy. Breiman [3] pointed out that Artificial Neural Networks (ANN's) utilised unstable procedures in forming the classifier and hence their use in our study.

To perturb the training set for each classifier, we use a simple randomisation process of the original data set together with a 10-fold cross-validation method [2] to reduce the bias on classifier evaluation. The 10-fold cross-validation method trains an individual classifier for each fold using 90% of data and 10% for testing. To prevent over-fitting of the ANN to the training set, 10% of the training data is additionally used to create a validation set. The process continues 10 times each time training an individual classifier with a different training and testing with a disjoint test partitions. No sample appears simultaneously in training and test. By randomising the data set before creation of the training/validation and testing folds, we introduce different combinations of samples into training/validation thereby providing the required perturbation of the input space. In this way, individual classifiers will be trained on different training sets, leading to different representations of the input space. Testing on these different input space representations leads to diversity in the resultant classifications for individual samples. Our trained classifiers comprise a feed-forward ANN using a back-propagation with momentum learning function (learning rate $\eta = 0.01$, momentum $\mu = 0.5$) together with a softmax activation function [2] to give an estimate of the posterior probabilities. Using the output from each classifier, a soft classification combination may be achieved. We investigate the following six combination rules detailed by Kittler et al. [6] majority vote, sum rule, max rule, min rule, product rule and median rule.

3 Results

We present two sets of results, one for each of the classification problems we are trying to solve. For each problem, we present the results of combining the posterior probabilities resulting from each of the eleven trained classifiers on test, combined using the six combination rules and compare with the single best classification result on test. For the four-class problem we label each sample according to the DDSM density ground truth. For the two-class problem, we label all samples of classes 1 and 2 as belonging to class 1 (fatty) and all samples of classes 3 and 4 as belonging to class 2 (dense).

Table 1 shows the results of the four-class problem. The best single recognition rate obtained from each of the eleven classifiers is 58.3%. By combining the posterior probabilities from each of the eleven classifiers using the product rule we obtain an increased recognition rate of 71.4%. All of the classifier combination rules produced an increase on the single best recognition rate with the exception of the min rule. Similarly, Table 1 shows the

Table 1. Recognition rates for combination rules on four-class problem (a) and two-class problem (b).

(a)

Combination Method	Recognition Rate %
Best single result	58.3
Min rule	40.3
Max rule	58.5
Product rule	71.4
Majority vote	62.7
Median rule	61.2
Sum rule	69.5

(b)

Combination Method	Recognition Rate %
Best single result	77.3
Min rule	90.4
Max rule	90.4
Product rule	96.9
Majority vote	89.1
Median rule	93.1
Sum rule	96.7

results obtained for the two-class problem. In this experiment each of the classifier combination rules produced a recognition rate better than that of the best single classifier at 77.3%. By combining the outputs from each of the classifiers using the product and sum rules, give equally improved recognition rates of 96.7%.

4 Conclusions

Within this study we have investigated the use of a variety of texture features for the classification of breast tissue type and demonstrated its application by classifying segmented breast regions on mammograms. By using classifier combination rules we have been able to maximise the recognition rate on test for the two classification tasks.

The results obtained for the four-class problem indicate that the classification of breast density according to the BIRADS system is a challenging task. The subtlety of breast tissue differentiating the four classes increases the likelihood of confusion in the resultant classification. Equally, our ground-truth relies on an expert radiologists assessment of the breast that may be subject to inter-observer differences. Our results reflect the problem of differentiating between subtle glandular and dense tissue types. Addressing the two-class problem of classifying dense and fatty we have demonstrated that the technique of combining classifier outputs not only improves the underlying performance but also ensure classifier robustness. The results of the two-class problem justify the use of this technique in a proposed CAD system.

References

1. NHS breast screening program review 1999. *NHS Breast Screening Program*, 1999.
2. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
3. Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
4. R. Chandrasekhar and Y. Attikiouzel. A simple method for automatically locating the nipple on mammograms. *IEEE Transactions on Medical Imaging*, 16(5):483–494, 1997.
5. R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems Man and Cybernetics*, SMC-3(6):610–621, 1973.
6. J Kittler, M. Hatef, R. P. W Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.
7. D. B. Kopans. *Breast Imaging*. Lippincott-Raven Publishers, 1998.
8. T. K. Lau and W. F. Bischof. Automated detection of breast tumors using the assymetry approach. *Computers and Biomedical research*, 24:273–295, 1991.
9. K. Laws. Texture image segmentation, ph.d. thesis, dept. of engineering, university of southern california, 1980.
10. R. Moore M. Heath K. Bowyer, D. Kopans and P. Kegelmeyer Jr. The digital database for screening mammography, 2000.
11. V. Hlavac M. Sonka and R. Boyle. *Image Processing, Analysis, and Machine Vision*. PWS Publishing, 1999.
12. A. Mendez, P. Tahoces, M. Lado, M. Aouto, and J. Vidal. Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms. *Medical Physics*, 25(6):957–964, 1998.
13. P. Miller and S. Astley. Classification of breast tissue by texture analysis. *Image and Vision Computing*, 10(5):277–282, 1992.
14. N. Petrick, H. P., Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler. Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classificaation. *Medical Physics*, 23(10):1685–1696, 1996.
15. J. C. Russ. Fractal imension, hurst coefficients, and frequency. *Journal of Computer Assisted Microscopy*, 2:249–257, 1990.
16. P. Taylor, S. Hajnal, M-H Dilhuydy, and B. Barreau. Measuring image texture to separate difficult from easy mammograms. *The British Journal of Radiology*, 67:456–463, 1994.
17. E. J. Feurer L. M. Wun. Devcan: Proability of developing or dying of cancer. Version 4.0, 1999.