10.1109/ULTSYM.2012.0423

Combined Naïve Bayes and logistic regression for quantitative breast sonography

Chandra M. Sehgal[1], Theodore W. Cary[1], Alyssa Cwanger[1], Benjamin J Levenback[1], Santosh S. Venkatesh[2],
Departments of Radiology [1], and Department of Electrical Engineering[2],
University of Pennsylvania, Philadelphia, PA 19104, USA.

Abstract - **Sonography is commonly used as an adjunct to mammography for early detection of breast cancer. We are developing methods to classify solid breast masses in sonograms as malignant or benign. The goal of this study was to combine two independent probabilistic classifiers to improve computer-aided diagnosis of breast masses. Naïve Bayes and logistic regression were used for supervised classification of masses from extracted morphological sonographic features, in combination with mammographic BI-RADS (categories 1 to 5) and patient age.**

Solid masses with biopsy-proven diagnoses were analyzed. Training and testing were performed using leave-one-out cross validation. Diagnostic performance was evaluated by the area under the curve (AUC) of the receiver operating characteristic (ROC). Agreement between predictions from the two classifiers was used to differentiate benign and malignant masses. The results show that logistic regression and Naïve Bayes performed with ROC area of 0.902 ± 0.023 and 0.865 ± 0.027, respectively. The combined use of logistic regression and Naïve Bayes demonstrated reduction in biopsies by 48%, with malignancy missed in 2% of cases (false negative rate of 6.4%).

Index Terms – quantitative breast ultrasound; computer-aided diagnosis; breast cancer; machine learning.

I. INTRODUCTION

Breast cancer is a significant cause of morbidity and mortality in women of all races, and early detection of the disease is important for improving patient clinical care. Studies have shown that sonography complements mammography by detecting breast masses that may not be visible on mammograms. Furthermore, compared to mammography ultrasound has good soft tissue discrimination, with no superposition of overlying fibroglandular tissue. Ultrasound can provide independent information on breast masses that could improve the diagnostic outcomes [1-4].

Despite major technical advancements in image quality, only a small proportion (20 to 30%) of masses classified as suspicious on imaging prove to be cancers after biopsy [5]. The financial and emotion cost of the low biopsy yield of imaging is significant. There is considerable ongoing effort to improve diagnostic performance of imaging methods to reduce the unnecessary biopsies. Several research groups have proposed computer-based analysis of breast ultrasound images to improve differentiation between malignant and benign masses [6-19]. In our previous studies [14-17] we identified quantitative sonographic features from computer image analysis that were statistically different for benign and malignant breast masses. These features along with the age of the patients were good predictors of malignancy [18-19]. In this study we expand the feature space to include mammographic BI-RADS category. Furthermore we develop a new approach that combines two independent probabilistic classifiers to improve computer-aided diagnosis of breast masses.

II. MATERIAL AND METHODS

A. Ultrasound data and features

The ultrasound data consisted of the images of 266 masses from patients recommended for biopsy based on mammography. For each mass, 3 to 5 images were obtained in radial and antiradial planes. A trained human observer manually outlined the mass on a computer display using a mouse. Since malignant cancers are known to have more complex shapes and margins than the benign masses, two types of features were extracted that characterized the lesion margins: (1) grayscale features at the margin, and (2) shape (morphometric) features. The extracted features from multiple views were compacted by taking their arithmetic mean. These features are described below [3, 11, 14].

B. Grayscale features

(1) Brightness difference: mean intensity difference between the lesion interior and immediate exterior.

(2) Margin sharpness: The sharpness of the margin. The lesion was divided into 72 sectors centered at the center of mass over 360 degrees, and the fraction that exhibited statistical difference in grayscale at the margin was measured.

(3) Angular variation at margin: Intensity variance in the sectors [see (2) above] around the interior of the lesion.

C. Shape Features

(4) Depth-to-width ratio: The ratio of the ROI's depth to its width on the sonogram, a standard feature used clinically.

(5) Axis ratio: Similar to depth-to-width ratio, the ratio of the major to minor axis of a best-fit ellipse to the ROI.

(6) Tortuosity: The perimeter of the lesion divided by the circumference of its best-fit ellipse. [11], same as elliptically normalized circumference, ENC)

(7) Radius variation: Variance in radius of the lesion from the center of gravity to each pixel in the boundary.

(8) Elliptically normalized skeleton: The number of pixels in a medial-axis skeleton of the shape divided by the circumference of its best-fit ellipse [11].

In addition to image features, patient age and mammographic BI-RADS category of the lesion were also included in the analysis. With the exception of Category 0, which denotes the need for further imaging evaluation, these categories segregate the experts' estimated probabilities of malignancy for the lesion into bins of increasing probability [20]:

D. BI-RADS Categories

Category 0: Needs further evaluation, missing data.

Category 1: Negative, nothing to comment on.

Category 2: Benign lesion.

Category 3: Short-term follow-up suggested, ≤ 2% chance of malignancy.

Category 4: Suspicious abnormality, biopsy should be considered, 3 to 94% chance of malignancy.

Category 5: Highly suspicious, requires biopsy, ≥ 95% chance of malignancy.

Category 6: Known biopsy-proven malignancy.

The actual experts' numeric estimates of the probabilities were not available. The category 1 through 5 indicating increasing probability of malignancy on a nonlinear scale was used as a *numeric* feature.

E. Machine learning algorithms

After the features were extracted from the cases, machine learning algorithms were trained on the features to build prediction models (WEKA, University of Waikato, NZ). Training and testing were performed using leave-one-out cross-validation: training the algorithm on *n*-1 samples in the database and predicting the outcome on the remaining *n*th sample, repeated iteratively until each sample is analyzed. Leave-one-out cross-validation is particularly appropriate for a small dataset, where it is important to train on as much data as possible. Performance for the models built from the learning algorithms was given by the area under the ROC curve (AUC). Algorithms were chosen to test two distinct learning styles: Naïve Bayes and logistic regression. Both methods estimate probabilities of malignancy under different assumptions of conditional independence between the individual attributes or features.

Bayes' Rule of conditional probability states that the posterior probability of malignancy given N conditionally independent features $F_1, \ldots, F_n$ and a priori probability, $p(M)$, is defined as

$$\mathrm{p(M|F_1, \ldots, F_n)} = \prod_{i=1}^{n} \mathrm{p(F_i|M)} \frac{p(M)}{p(\mathrm{F_1, \ldots, F_n})} \quad (1)$$

For categorical attributes, the malignant probability term was determined as the ratio of the number of malignancies in the category to the total number of malignant cases. For numerical attributes, a Gaussian distribution was assumed. The features that were not normally distributed were discretized into nominal attributes using equal-width or equal-frequency binning.

In logistic regression binary classification as malignant or benign is accomplished through thresholding a linear combination of input features, where probability of malignancy is defined a

$$p(M|F_1, \ldots F_n) = \frac{1}{1+exp[-(\alpha+\sum_{i=1}^{n} \beta_i F_i)]} \quad (2)$$

The coefficients α and β_i are determined from the training data.

F. Combining Naïve Bayes and logistic regression

After the Naïve Bayes' and logistic learners were trained, they were combined by evaluating the agreement between the two outputs. The diagnosis was determined by comparing probability of malignancy estimates, p_i, with different decision thresholds (p_{th}):

$$p_i \geq p_{th} \quad i = M,$$
$$p_i < p_{th} \quad i = B, \quad (3)$$

where M and B represent malignant and benign diagnosis. Agreement between decision outcomes for Naïve Bayes and logistic regression was determined at different thresholds (p_{th}) between 0.01 and 0.99. The boundary thresholds ($p_{th} \leq 0.01$ and $p_{th} \geq 0.99$) were not considered as they represent cases where all the masses were identified as malignant or benign. Sensitivity and specificity of diagnosis at the threshold of maximum agreement between the two algorithms was measured.

III. RESULTS

Of the 266 masses, 181 (68%) were benign and 85 (32%) were malignant. The majority of benign lesions were classified as fibroadenoma, while most malignant masses were diagnosed as invasive ductal carcinoma. Figure 1 shows receiver operating characteristic (ROC) curves for the Naïve Bayes and logistic regression models. Logistic regression performs better than Naïve Bayes except for the narrow range of sensitivity from 63 to 82 (Figure 1).

The area under the ROC curve, AUC ± SD, for Naïve Bayes (dotted line, Figure 1) was 0.865 ± 0.027 (95% CI, 0.818 to 0.904). For logistic regression the corresponding area under the curve (solid line, Figure 1) was 0.902 ± 0.023 (95% CI, 0.860 to 0.935). AUC measures the ability of the algorithm to correctly classify malignant and benign masses, where AUC of 1.0 represents perfect diagnostic performance and 0.5 equals random chance. Pairwise comparison of the two ROC curves shows the difference between the areas of 0.037 ± 0.017 (95% CI, 0.004 to 0.070) is significant (P = 0.030).

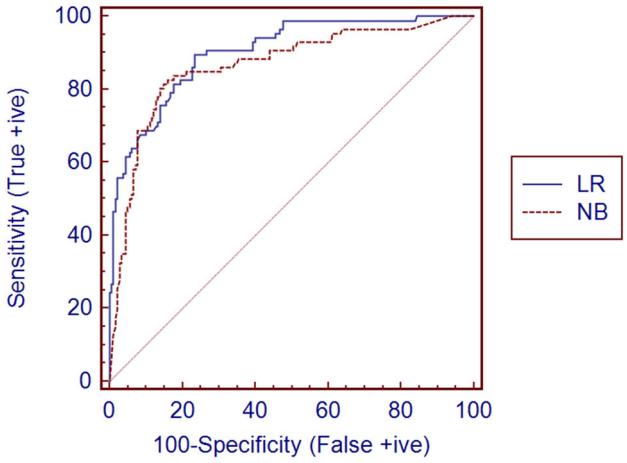

Figure 1: ROC curves for Naïve Bayes (NB) and logistic regression (LR) machine learning models.

Figure 2 shows the agreement between Naïve Bayes and logistic regression based on Equation 3. The agreement is fairly stable at ~88% from probability thresholds (p_{th}) of 0.15 to 0.7. Since more malignant cases will be identified at lower p_{th}, the sensitivity and specificity at p_{th} of 0.15 was measured. At this threshold there was agreement between Naïve Bayes and logistic regression in 233 (87.6%) cases. The remaining 33 (12.5%) were not considered suitable for analysis, rather candidates for further imaging or other diagnostic tests.

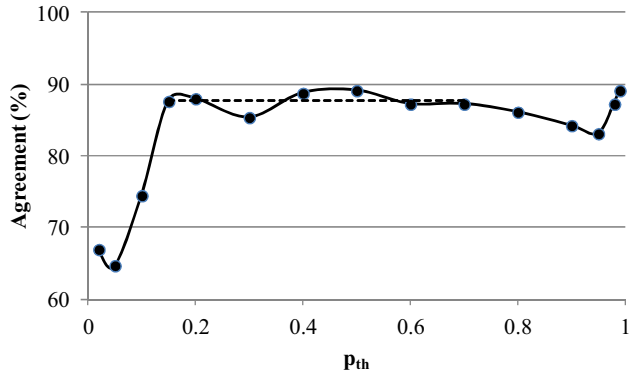

Figure 2: Percent agreement between Naïve Bayes and logistic regression algorithms as a function of threshold probability (p_{th}).

The confusion matrix for the proposed classification scheme is given in Table 1, obtained by cross-classifying the predicted class in the test set with the true class. The sensitivity and specificity, also known as true-positive rate and true-negative rate, for the 233 good cases were 0.94 and 0.68, respectively.

	Malignant	Benign	
Test +ive	73 (31%)	49 (21%)	122 (52%)
Test -ive	5 (2%)	106 (45%)	111 (48%)
	78 (33%)	155 (67%)	233(100%)

Table 1: Confusion matrix for the cases showing agreement between Naïve Bayes and logistic regression

IV. Discussion

There are many ways to construct classification rules, each with its own merits and drawbacks. The results of this study demonstrate that two dissimilar machine learning algorithms, Naïve Bayes and logistic regression, have high diagnostic performance for differentiating malignant and benign lesions. When the two algorithms are compared, logistic regression consistently outperforms Naïve Bayes (Figure 1), except for the sensitivity range ~70 to 80. The area under the ROC curve for logistic regression is 0.902 ± 0.023 compared to 0.865 ± 0.027 for Naïve Bayes. The area under the ROC curve represents the probability that the classifier will rank a randomly chosen positive (malignant) case higher than a randomly chosen negative (benign) case [21]. That is, logistic regression has higher probability than Naïve Bayes to rank a randomly chosen malignant case higher than a randomly chosen benign case. The difference between the two approaches, although relatively small (0.037 ± 0.017), is statistically significant. The lower performance of Naïve Bayes is probably related to its assumption that all the features are conditionally independent. In practice the image features are derived from the same gray level distribution and are thus likely to be correlated and not completely independent.

The attractive aspect of the two proposed algorithms is that they both provide numerical estimates of probability of malignancy. While the probability estimates of the two methods are comparable they are not identical for each case. This difference suggests that the two algorithms might be combined to create an ensemble learner with higher performance. There are many ways of combining algorithms, the simplest approach being a weighted combination of the individual estimates. However, preliminary assessment of weighting showed the approach did not lead to significant improvement in diagnostic performance, although much further work is needed to confirm these results. The lack of improvement could be due to the fact that the difference in the estimates of the two probabilities is already small.

A second approach for combining models is to use consensus between them to decide the final diagnosis. Each machine learning algorithm is treated as an independent observer. The cases where there is no agreement are classified as indeterminate requiring further investigation, possibly passed to an additional machine learning algorithm. The cases for which there is agreement are the only ones suited for diagnostic classification. The results of the study show that the

agreement between the two learning method varies with the threshold p_{th}, but over a broad range of p_{th} there is high level of agreement. At p_{th} of 0.15 there is an agreement for 233 (78 malignant, 155 benign) cases, of which only 122 would will be recommended for biopsy. Since originally all 233 cases would be recommended for biopsy, this represents a significant 48% reduction in the number of biopsies. Of the 122 recommended cases, 49 (32% of the 155 benign cases, 24 % of all the 233 cases) would be benign or unnecessary biopsies, an improvement over the normal rate of 60 to 80%: the biopsy yield increased from 33% (78/233) to 60% (73/122). However, the benefits of the proposed method are at the cost of missing 5 malignant cases, (6.4% of the 78 malignant masses present in the group or 2 % of all the 233 cases).

V. CONCLUSION

The results demonstrate that computer-based quantitative methods improve diagnosis on breast ultrasound and have the potential to reduce the number of biopsies. The results are encouraging and a further reduction in the false negative rates along with feature selection and independent validation would be helpful in making the current method suitable for clinical use in the future.

ACKNOWLEDGMENT

We thank Susan Schultz RDMS for help with patient studies. This work was supported in part by NIH grants CA87526 and CA130946

REFERENCES

[1] Madjar Helmut. The Practice of Breast Ultrasound: Techniques, Findings Differential Diagnosis.. Thieme Medial Publishers, Inc. New York, 2000.

[2] Jackson VP. Management of solid breast nodules: what is the role of sonography? *Radiology. 196(1):14-5, 1995.*

[3] Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology. 196(1):123-34, 1995.*

[4] Arger PH, Sehgal CM, Conant EF, Zuckerman J, Rowling SE, Patton JA. Interreader variability and predictive value of US descriptions of solid breast masses: pilot study. *Academic Radiology. 8(4):335-42, 2001.*

[5] Agency of Healthcare research and quality, Core-Needle Biopsy for Breast Abnormalities: Clinician's Guide*,* effectivehealthcare.ahrq.gov.

[6] Goldberg V, Manduca A, Ewert DL, Gisvold JJ, Greenleaf JF. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. *Medical Physics. 19(6):1475-81, 1992.*

[7] Dumane VA, Shankar PM, Piccoli CW, Reid JM, Forsberg F, Goldberg BB. Computer aided classification of masses in ultrasonic mammography. *Medical Physics. 29(9):1968-73, 2002.*

[8] Shankar PM, Dumane VA, Piccoli CW, Reid JM, Forsberg F, Goldberg BB. Classification of breast masses in ultrasonic B-mode images using a compounding technique in the Nakagami distribution domain. *Ultrasound in Medicine & Biology. 28(10):1295-300, 2002.*

[9] Drukker K, Giger ML, Horsch K, Kupinski MA, Vyborny CJ, Mendelson EB. Computerized lesion detection on breast ultrasound. *Medical Physics. 29(7):1438-46, 2002.*

[10] Chen DR, Chang RF, Kuo WJ, Chen MC, Huang YL. Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks. *Ultrasound in Medicine & Biology. 28(10):1301-10, 2002.*

[11] Chen CM, Chou YH, Han KC, Hung GS, Tiu CM, Chiou HJ, Chiou SY. Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks. *Radiology. 226(2):504-14, 2003.*

[12] Richter K, Heywang-Kobrunner SH, Winzer KJ, Schmitt KJ, Prihoda H, Frohberg HD, Guski H, Gregor P, Blohmer JU, Fobbe F, Doinghaus K, Lohr G, Hamm B. Detection of malignant and benign breast lesions with an automated US system: results in 120 cases. *Radiology. 205(3):823-30, 1997.*

[13] B. Sahiner, H.P. Chan, M.A. Roubidoux, L.M. Hadjiiski, M.A. Helvie, C. Paramagul, J. Bailey, A.V. Nees and C. Blane, Malignant and benign breast masses on 3D US volumetric images: Effect of computer-aided diagnosis on radiologist accuracy, *Radiology* 242, pp. 716–724, (2007).

[14] Sehgal CM, Cary TW, Kangas SA, Weinstein SP, Schultz SM, Arger PH, Conant EF. Computer-based margin analysis of breast ultrasound for differentiating malignant and benign masses. *Journal of Ultrasound in Medicine. 23: 1201-1209, 2004*

[15] Sehgal CM, Arger PH Rowling SE, Conant EF, Reynolds C, Patton JA. Quantitative vascularity of breast masses by Doppler imaging: regional variations and diagnostic implications. *Journal of Ultrasound in Medicine. 19(7):427-40; 441-2, 2000.*

[16] Song, Jae H, Venkatesh, SS, Conant, EA, Arger, PH, Sehgal, CM. Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. *Academic Radiology*, 4:487-495, 2005.

[17] Matsumoto MMS, Sehgal CM, Udupa J Local binary pattern texture-based classification of solid masses in ultrasound breast images Medical Imaging 2012: Ultrasonic Imaging, Tomography, and Therapy, edited by Johan G. Bosch, Marvin M. Doyley, Proc. of SPIE Vol. 8320, pages 83201H-1 to 83201H-8 ·, 2012 SPIE 2012.

[18] Harvey P, Arger PH, Conant EF, Sehgal CM, Differentiation of the solid benign and malignant breast masses by quantitative analysis of the ultrasound images, 2009 IEEE International Ultrasonics Symposium Proceedings, 530 – 533, 2009.

[19] Cary TW, Cwanger A, Venkatesh SS, Conant EF, Sehgal CM, Comparison of naïve Bayes and logistic regression for computer-aided diagnosis of breast masses using ultrasound imaging Medical Imaging 2012: Ultrasonic Imaging, Tomography, and Therapy, edited by Johan G. Bosch, Marvin M. Doyley, Proc. of SPIE Vol. 8320, pages 83200M-1 to 83200M-7, SPIE 2012.

[20] American College of Radiology, http://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/BIRADS/MammoAssessmentCategories.pdf .

[21] Fawcett T, ROC Graphs: Notes and Practical Considerations for Researchers, ROC1.tex, 1-38, 2004