International Journal of Electronics Signals and Systems

Volume 1 | Issue 2 Article 2

September 2011

Hybrid Image Mining Methods to Classify the Abnormality in Complete Field Image Mammograms Based on Normal Regions

Aswini Kumar Mohanty
SOA University, Khandagiri Bhubaneswar, asw_moh@yahoo.com

P. K. Champati
Department of Computer Science, ABIT, Cuttack, pkchampati@gmail.com

Manas Rajan Senapati
Department of computer Science, Gandhi Engineering College, Bhubaneswar, manas_senapati@sify.com

Saroj Kumar Lena
Department of Computer Science, Modi University, Rajstan, sarojln@yahoo.com

Available online at www.interscience.in

Hybrid Image Mining Methods to Classify the Abnormality in Complete Field Image Mammograms Based on Normal Regions

[1]Aswini Kumar Mohanty, [2]P.K. Champati, [3]Manas Rajan Senapati, [4]Saroj Kumar Lena
[1]SOA University, Khandagiri Bhubaneswar, [2]Department of Computer Science, ABIT, Cuttack
[3]Department of computer Science, Gandhi Engineering College, Bhubaneswar,
[4]Department of Computer Science, Modi University, Rajstan
E-Mail:-asw_moh@yahoo.com, pkchampati@gmail.com, manas_senapati@sify.com, sarojln@yahoo.com

***Abstract*—** Breast Cancer now becomes a common disease among woman in developing as well as developed countries. Many non-invasive methodologies have been used to detect breast cancer. Computer Aided diagnosis through, Mammography is a widely used as a screening tool and is the gold standard for the early detection of breast cancer. The classification of breast masses into the benign and malignant categories is an important problem in the area of computer-aided diagnosis of breast cancer. We present a new method for complete total image of mammogram analysis. A mammogram is analyzed region by region and is classified as normal or abnormal. We present a hybrid technique for extracting features that can be used to distinguish normal and abnormal regions of a mammogram. We describe our classifier technique that uses a unique re-classification method to boost the classification performance. Our proposed hybrid technique comprises decision tree followed by association rule miner shows most proficient and promising performance with high classification rate compared to many other classifiers. We have tested this technique on a set of ground-truth complete total image of mammograms and the result was quite effective.

Keywords*: Mammogram, feature extraction, data mining classifier, decision tree, association rule mining*

I. INTRODUCTION

Breast cancer is the leading cause of cancer-related death among women aged 15-54. The earlier breast cancer is detected, the higher is the chance of survival. Screening mammography is the only method currently available for the reliable detection of early and potentially curable breast cancer. Several studies have shown retrospectively that 20% to 40% of breast cancers fail to be detected at screening [1]. A computer-aided detection (CAD) system has been developed as a second reader. The performance of the radiologists can be increased 5-15% by providing the radiologists with results from a CAD system as a "second opinion" [2]. However, the majority of mammograms are normal. Among the false positive readings of normal mammograms, only 15%-34% actually show malignancy at histological examination [3]. An accurate computer-aided system to identify normal mammograms would reduce radiologists' work load, allow them to focus more on suspicious cases and to improve screening performance. In this paper, we propose a new method of full-field mammogram analysis based on the identification of normal regions. First, a classifier for identifying normal regions is trained from a set of features extracted from normal and ground-truth cancerous regions extracted from the DDSM (Digital Database for Screening Mammography) database [4]. Using an overlapped block technique, this classifier is used to analyze full-field mammograms. This approach is independent of the type of abnormality, and may complement computer-aided detection.

The rest of the paper is organized as follows: Section II introduces the Mammogram database to normalize with breast back ground separation method including enhancement followed by feature extraction. A cascading classifier using decision tree and association rule is introduced for classification taking different features like texture, gabor, curvilinear and multi-resolution features. Association rule classifier is now being more used in classification due to its faster execution and accuracy. Association rules are derived between various features component extracted from mammogram images and employed for classification based on their inter class dependencies. These rules are then employed for the classification of a commonly DDSM dataset and rigorous experimentation is performed to evaluate the rule as a result of which it incurs accuracies as high as 95% which surpasses the accuracy rate of other classifier It is specially suiting to

binary decision tree rule classifier due to both are used in data binarization A decision tree recursively subdivides regions in the feature space into different subspaces, using different thresholds in each dimension to maximize class discrimination. Ideally, for a given subspace, the process stops when it only contains patterns of one class. However, in practice, sometimes it is not possible or is computationally prohibitive to use such a stopping criterion, and the algorithm stops when most of the patterns belong to the same region. Section III describes the result of hybrid classifier in comp are to other classifier in terms of Area under ROC. The area as much as covered, lead to better accuracy level and classification performance. The misclassification in decision tree classifier is well captured by ARM and the performance is well the level of 97%.The success rate for classification enhances due to similar employment of data binarization. The final section IV describes about conclusion and future work

II. FULL-FIELD MAMMOGRAM ANALYSIS

The following sections discuss each step of our complete full-field mammogram analysis technique and are outlined in Figure 1.

A. Mammogram Database

All of the mammograms used in this study are obtained from the Digital Database for Screening Mammography (DDSM) [4].Each mammogram has been "normalized" to optical density and linearly mapped to an 8-bit gray level image. We use the breast-background separation method described in [5] to segment out the breast area. The segmented image is ready for full-field analysis.

B.. Enhancement Based on H Representation

A standardized mammogram representation can be based on modeling of the X-ray physics of the image formation process. We used the techniques described in [6,7] that models the complete imaging process and compensates the degrading factors, such as scattering. The resulting image, known as the H representation, records the height of non-fatty tissue in the breast for each pixel in the image. This measurement is intrinsic to the breast. In our experiment, we used a simplified transform based on a mono-energetic hint and an enhancement step to remove the background. We call the processed image I and all of features will be extracted from I.

C. *Regional Feature Extraction*

Each full-field mammogram is analyzed by overlapped moving blocks. The region covered by each block is 512 by 512 pixels. There are four types of features extracted from each region: curvilinear features, texture features, Gabor features, and multi-resolution features. Curvilinear features: Though normal breast tissue may have very different appearance, unequivocally normal breast areas are characterized by curvilinear markings. These curvilinear structures are the ductal structures of the breast tissue.

We used a line detection algorithm we previously developed [8,9] to extract the curvilinear structures in each region. The algorithm is robust to noise and is capable of extracting quasi-linear curves of different widths and angles. A set of features was extracted from the detected curvilinear structures to characterize the region. There were total 18 curvilinear features extracted for each region, capturing the statistical nature of the line pixels.

Texture features: Texture information is characterized by the spatial arrangement of the pixel intensities. This can be specified by a 2D spatial dependence matrix known as the Gray Level Co-occurrence Matrix (GLCM) [10, 11]. GLCM is one of the best known texture analysis methods. We extracted 16 features from the GLCM, as defined in [8] and additional cluster features as defined in [12].

Gabor features: Gabor filters has been used for texture analysis for many years [9,13]. The advantage of Gabor filters is that they provide simultaneous localization in both the spatial and frequency domains. In the study, the highest and lowest frequencies of the Gabor filter-bank were chosen to suit our analysis. We chose 4 orientations and 4 scales for the Gabor filter-bank. We obtained the mean and standard deviation of the energy of each Gabor filtered image. Hence, there were 32 Gabor features extracted from each region. Multi-resolution features: The last type of features was obtained from nonlinear wavelet decomposition. A special nonlinear wavelet transform' the Quincunx Wavelet transforms [14], was used in our study. Only the first four even level wavelet decomposition images were retained for feature extraction. There are five features were extracted from each decomposition for a total of 20 features. The above four types of features combined to form a 86- feature vector associated with each

1024×1024 region. These will be used to train a cascading classifier.

D.. A Cascading Classifier for Identifying Normal Regions

A cascading classifier, shown in Figure 2, was trained using the 1024 × 1024 regions. These regions were manually extracted from screening mammograms different than the mammograms used for testing. All of normal regions were extracted from normal mammograms and cancerous regions were extracted from cancer cases with the cancer in the center of the region. A total of 460 training regions were used, which consisted of 296 normal and 164 cancer regions. The training procedure was performed only once. After the training, the classifier is used to analyze each full-field mammogram region by region.

The two-stage cascading classification system (in Figure 2) is a special case of the stacked generalization [15, 16, 17] due to its layered structure. The first stage may correctly classify most of the abnormal regions while separating out as many of the normal regions. A decision tree classifier described in [11, 18, 19] was used as the first stage classifier because it is one of the most powerful classification tools. Misclassification costs could be specified to retain almost all training cancerous regions. The decision tree classifier was based on a hierarchy of multiple decision variables (features), which made it difficult to evaluate the performance using a Receiver Operation Curve (ROC). Therefore to improve the classification performance, a second-stage classifier was used. Only those regions classified as "abnormal" by the decision tree classifier were classified by the second-stage Association rule mining classifier [20] due to the advantage of extracting set of rules, using a fuzzy approach to select the items and does not require the user to provide the thresholds. ARMC proceeds by combing on one hand, the weighted voting and the decision list algorithms. On the other hand, a fuzzy method is used in order to distinguish the important rules from the less important ones for each class. In this study, the second-stage classifier was Association rule miner classifier [21, 22, 23] with CFS feature selection [24, 25.] This two-stage cascading classifier system has the classification power of a decision tree and the simplicity of the ROC analysis of a rule classifier. Our experiments showed that it performed better than a decision tree or an association rule mining classifier [26, 27, 28].

E.. Full-field Analysis Using Overlapped Regions

The cascading classifier was used to analyze a full-field mammogram using an overlapped, moving block technique. The moving block size is 512 × 512. First, each mammogram was expanded by mirroring 128 pixels along the boundary to reduce the edge effects. The breast area is analyzed by 5 overlapped blocks. The block is centered on a pixel and then is moved by 128 pixels up, down, right, and left. Using the two-stage cascading classifier on each block, the classification result (normal or abnormal) of each block is obtained; therefore each sub-region is classified 5 times. A majority voting scheme is used to determine the final classification (Figure 1).

Finally, a full-field mammogram is classified as a cancer image if one or more sub-regions are abnormal; otherwise, the mammogram is classified as a normal.

III. RESULTS

Our two-stage cascading classifier was trained from an independent training set of 164 ground-truth cancerous regions and 296 normal regions. Among the 164 ground-truth cancerous regions, 53 were masses, 56 were spiculations and 55 were calcifications. The first-stage decision tree classifier was constrained to retain nearly every cancerous region.

This resulted in a True Positive Fraction (TPF) of 0.981 at a False Positive Fraction (FPF) of 0.294. The regions (including 161 true positives and 87 false positives) classified as "abnormal" were then refined by the second-stage linear classifier. Our two-stage classifier system had an overall performance, Az = 0.9756, where Az is the area under the ROC. Figure 3 shows the comparison with an Association ruler classifier, with Az of 0.9576.

Table 1. Normal Classification on Cancer Mammograms

Number of Correct Classifications of Different Cancers		
Calcifications	*Mammograms Tested*	25
	Classified As Abnormal	24
Masses	*Mammograms Tested*	22
	Classified As Abnormal	21
Speculations	*Mammograms Tested*	24
	Classified As Abnormal	22

The classifier was then used to analyze full-field mammograms. We tested 71 cancer mammograms and 76 normal mammograms. Among the 71 cancer mammograms, 25 were calcification images, 22 were mass images and 24 were speculation images. Table 1 shows the performance on full-field cancer mammograms. The true positive rate is 0.944. Most of misclassified cancer images are calcifications and rate is 0.917. The region of analysis might be too large for small clusters of calcifications. Excluding calcifications, we obtained 95.5 percentage correct classifications on mass and spiculation images. We believe the reason for misclassification is due to the subtlety of the breast cancers. 73 normal mammograms are classified correctly, i.e. the true negative rate is 0.96. Most of misclassification is due to high breast density of these normal mammograms.

Table 2. Performance of the classifiers in terms of Area under the roc curves on oroginal feature spaces

Performance of classifiers	
Classifiers	***Area under ROC curves Az***
KNN	0.752
MLP	0.821
PNN	0.803
RBF	0.825
ROUGH SET &FUZZY	0.840
K-MEAN &SVM	0.850

IV. CONCLUSION

A new full-field mammogram analysis method was presented. Our initial results are encouraging. In compare to other classifiers considering the area under ROC but still lacks of good optimal result on mammograms which are dense and fatty and difficult to extract features due to high intensity. This study can be extended using a larger data base of both scanned images and images obtained from a digital mammography system. Better and well optimized methods can be employed for feature selection and reduction to achieve faster computation along with other data mining methods on large size database. Due to high density in breast masses of normal mammograms, the misclassification rate is 0.04. Suitable feature selection and reduction will improve classification performance and reduce data inconsistency. Parameter values as well as block size play a vital role in the system's performance and an investigation of this relation and perhaps automation of their selection is needed to further improve system's robustness The obtained calcification and the no. of misclassification can be more optimal by increasing the no. of training set and testing test samples

REFERENCE

[1] C. Beam, P. Layde, and D. Sullivan, "Variability in the interpretation of screening mammograms by us radiologists, findings from a national sample," Archives of Internal Medicine, vol. 156, pp. 209–213, 1996.

[2] E.L. Thurfjell, K.A. Lernevall, and A.S. Taube, "Benefit of independent double reading in a population based mammography screening program," Radiology, vol. 191, pp. 241–244, 1994.

[3] J.Y. Lo, J.A. Baker, P.J. Kornguth, J.D. Iglehart, and C.E. Floyd, "Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features," Radiology, vol. 203, pp. 159–163, 1997.

[4] M. Heath, K.W. Bowyer, D. Kopans, R. Moore, and Jr. P. Kegelmeyer, "The digital database for screening mammography," Proceedings of the 5th International Workshop on Digital Mammography, pp. 212–218, June 11-14 2000.

[5] T. Ojala, J. N¨appi, and O. Nevalainen, "Accurate segmentation of the breast region from digitized mammograms" Computerized Medical Imaging and Graphics, vol. 25, pp. 47–59, 2001.

[6] Ralph Highnam and Michael Brady, "Mammographic Image Analysis", Kluwer Academic Publishers, Dordrecht, 1999.

[7] R.J. Ferrari, A.C.P.L.F. de Carvalho, P.M.A. Marques, A.F. Frere, "Computerized classification of breast lesions: shape and texture analysis using an artificial neural network", Image Process. Appl., pp. 517–521, 1999.

[8] S. Liu, The Analysis of Digital Mammograms: "Spiculated Tumor Detection and Normal Mammogram Characterization", Ph.D. Thesis, School of Electrical and Computer Engineering, Purdue University, May 1999.

[9] R. J. Ferrari, R.M. Rangayyan, J.E.L. Desautels, and A.F. Frere, "Analysis of Asymmetry in Mammograms via Directional Filtering With Gabor Wavelets", IEEE Transactions on Medical Imaging, vol.20, pp. 953–964, 2001.

[10] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification" IEEE Transaction On Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610–621, November 1973.

[11] Diaz, L.K., Sneige, N. (2005) "Estrogen receptor analysis for breast cancer: Current issues and keys to

increasing testing accuracy", Advance Anat Pathol, Vol. 12, pp. 10-19

[12] R.W. Conners, M.M. Trivedi, and C.A. Harlow, "Segmentation of a high-resolution urban scene using texture operators," Computer Vision, Graphics and Image Processing, vol. 25, no. 3, pp. 273–310, 1984.

[13] B.S. Manjunath and W. Y. Ma. "Texture features for browsing and retrieval of image data," IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 837–842, August 1996.

[14] J. Kova¢cevi´c and M. Vetterli, "Non separable multidimensional perfect reconstruction filter banks and wavelet bases for Rn," IEEE Transactions on Information Theory, vol. 38, no. 2, pp. 535–555, March 1992.

[15] D. Wolpert, "Stacked generalization," Neural Networks, vol. 5, 1992.

[16] Ioanni Pratikakis, Anna Karahaliou, Katerina Vassiou, Vassilis Virvilis, Dimitrios Kosmopoulos, abd Starvrous Peratonis(2007), eMedl"Web-Based E-Training For Multimodal Breast Imaging" Proceedings of world Academy of Science, engineering and technology vol 25 November 2007.

[17] Holden, N., Freitas, A. A Hybrid PSO/ACO Algorithm for Discovering Classification Rules in Data Mining, In Journal of Artificial Evolution and Applications (JAEA), 2008.

[18] S.B. Gelfand, C.S. Ravishankar, and E.J. Delp, "An iterative growing and pruning algorithm for classification tree design," IEEE Transaction on Pattern Analysis Machine Intelligence, vol. 13, pp. 163–174, 1991.

[19] A. Mosavi, "Multiple Criteria Decision-Making Preprocessing Using Data Mining Tools", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 1, March 2010

[20] Roselin, R.; Thangavel, K.;"International conference on computing communication and networking technologies"(ICCCNT),IEEE Explore,pp.1-6, 2010, DOI:10.1109/ICCCNT.2010.5592607

[21] Irina Tudor, "Association Rule Mining as a Data Mining Technique"; BULETIN Vol. LXNo. 1/2008, pp.49 – 56, Seria Matematică - Informatică – Fizică

[22 Maria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coman, "Application of Data Mining Techniques for Medical Image Classification", Proceeding of the Second International Workshop on Multi Media Data Mining in conjunction with ACM SIGKDD Conference-2001

[23] Harleen Kaur and Siri Krishan Wasan "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science (2): pp. 194-200, 2006.

[24] P. Somol, P. Pudil, J. Novovicova, and P. Paclik, "Adaptive floating search methods in feature selection," Pattern Recognition Letters, vol. 20, pp. 1157–1163, 1999.

[25] Michal Haindl1, Petr Somol1, Dimitrios Ververidis, Constantine Kotropoulos "Feature Selection Based on Mutual Correlation" progress in pattern recognition, image analysis and applications, lecture notes in computer science vol. 4225/2600,569-577,DOI 10.1007/11892755_59

[26 Masala, G. L.;Tangaro, S.;Golosio, B.;Oliva, P.;Stumbo, S.;Bellotti, R.;de Carlo, F.;Gargano, G.; Cascio, D.;Fauci, F.;Magro, R.;Raso, G.;Bottigli, U. ;Chincarini, A.;de Mitri, I.;de Nunzio, G.; Gori, I.; Retico, A.; Cerello, P.; Cheran, S. C.; Fulcheri, C.; Lopez Torres, E."Comparative study of feature classification methods for mass lesion recognition in digitized mammograms" Il Nuovo Cimento C, vol. 30, Issue 3, p.305-316 DOI: 10.1393/ncc/i2007-10241-y

[27] Leonardo de Oliveira Martins, Geraldo Braz Junior, Aristófanes Correa Silva, Anselmo Cardoso de Paiva, Marcelo Gattass"Detection of Masses in Digital Mammograms using K-Means and Support Vector Machine

❑❑❑

.