Multiresolution and Multiscale Geometric Analysis based Breast Cancer Diagnosis using weighted SVM

Yang Wang[1,a] Miaomiao Yin[2,b]

[1]Public computer teaching and research center, Jilin University, Changchun, China

[2]School of Management, Jilin University, Changchun, China

[a]wyangjlu@gmail.com, [b]yinmiaomiao@gmail.com

Key Words: Support Vector Machine, Breast cancer diagnosis, Digital Mammogram

Abstract. This paper presents an approach for breast cancer diagnosis in digital mammogram using multiresolution and multiscale geometric analysis. The proposed method consists of two stages. In the first stage, mammogram images are decomposed into different resolution levels using wavelet transform and curvelet transform, which are sensitive to different frequency bands. A set of the biggest coefficients from each decomposition level is extracted as features vector. In the second stage, classification is performed on a weighted support vector machine (SVM). Due to random selection of samples, it is highly probable that a significantly small portion of the training set is the "mass present" class. To address this problem, we propose to use weighted SVM in a successive enhancement learning scheme to examine all the available "mass present" samples. The proposed approach is applied to the Mammograms Image Analysis Society dataset (MIAS) and classification accuracy of 99.3% is determined over an efficient computation time by successive learning enhancement. Experiment results illustrate that the multiresolution and multiscale geometric analysis-based feature extraction in conjunction with the state-of-art classifier construct a powerful, efficient and practical approach for breast cancer diagnosis.

Introduction

Breast cancer is the second leading cause of death for woman all over the world and more than 8% women will suffer this disease during their lifetime. In 2008, there were reported approximately 182,460 newly diagnosed cases and 40,480 deaths in the United States [1]. Since the causes of breast cancer still remain unknown, early detection is the key to reduce the death rate (40% or more) [2, 3]. The earlier the cancers are detected, the better treatment can be provided. However, early detection requires an accurate and reliable diagnosis which should also be able to distinguish benign and malignant tumors.

Our objective is to develop a breast cancer diagnosis system for mass classification of digital mammograms. Mass classification requires a preprocessing step of segmenting the input image into disjoint areas, such as the breast region, background and content text. A significant body of research has already been devoted to breast segmentation including adaptive thresholding [4, 5], polynomial modeling [6], active contours [7] and classifier-based techniques [8].

The second step in mass classification (as the most effective stage) is feature extraction. Texture is a commonly used feature in the analysis and interpretation of images. Malagelada's approach [6] distinguishes underlying textures in mammography into the three following classes:

(1) Statistical methods: The extracted features of statistical methods include those obtained

from surface variation measurements (smoothness, coarseness) [6] and run-length statistics [3, 8].

(2) Model-based methods: The analysis of texture features in this class is based on prior models such as Markov random fields [9] and fractals [6].

(3) Signal processing methods: In this class, texture features are obtained according to either pixel characteristics or image frequency spectrum including Laws energy filtering [6], Gabor filtering [6], and wavelet [10-12].

Following feature extraction, an appropriate classifier is utilized in breast mass classification. Various methods have been proposed with the most efficient ones known as Bayesian classifier [13], multilayer perceptron [9,14], adaptive neuron fuzzy inference system (ANFIS) classifier [12], radial basis function (RBF) [14], k-nearest neighbors (KNN) [8], decision tree classifier [15], and support vector machines (SVM) [16].

This work is aimed at improving performance of the current mass classification methods using the recently proposed image transforms and classifiers. The novelty of this research is in exploiting the superiority of combination of wavelet and curvelet transform in representing point and line singularities in digital mammogram. Furthermore the structural risk minimization property of successive enhancement learning (SEL) weighted SVM achieves a more efficient mammogram mass classification [17].

Methodology

Feature Extraction. The original mammograms are pixels, and almost 50% of the whole image comprised of the background with a lot of noise. Therefore a cropping operation is applied to the images to cut off the unwanted portions of the images. Regions of Interest (ROI's) are cropped. The cropping process was performed manually, where the given center of the abnormality area is selected to be the center of ROI. Thus, almost all the background information and most of the noise are eliminated. By this method we are sure that no abnormality was suppressed with the background. An example of cropping that eliminates the label on the image and the black background is given in Fig.1. Some examples to the ROI's are presented in Fig.2.

Once the images are cropped as described, both wavelet and curvelet transform methods are applied and the features vectors are extracted. For wavelet, four different decomposition levels based on Daubechies-8 (db8) wavelet function is used. The used levels of decomposition and wavelet functions are selected based on previous work [4, 10 ,11], the number of decomposition levels used for curvelet transform is 4.

In each decomposition level, the obtained coefficients are sorted in descending order. Then, the biggest 50 coefficients are extracted to represent the corresponding mammogram (i.e., feature vector). This means that each mammogram image is represented by 400 coefficients.

Classification Method. After feature extraction, the above features are passed to a successive enhancement learning (SEL) weighted SVM classifier. The basic idea of the successive enhancement learning method is to iteratively select the most representative "mass present" examples from all the training images while keeping total number of training examples small [18]. This scheme can improve the generalization ability of a weighted SVM classifier.

In general, for a mammogram, many ROIs are classified as "mass absent", and very few are labeled as "mass present". Therefore, due to random selection of samples, it is highly probable that a significantly small fraction of the training set is occupied by the "mass present" class. To overcome this problem, we propose to use weighted SVM in a successive enhancement learning scheme to examine all the available "mass present" samples. Misclassified in "mass-present" has severe effects and causes of death. Hence, accuracy in this class is more important than in another and samples of this class should have more contribution in classification. Therefore, the penalty of misclassification for each class must be different, and we need to assign higher weights to "mass present" samples. Hence, we consider an equal penalty for the training samples belonging to same class, and set the ratio of penalties for different classes to the inverse ratio of the training class sizes. The weight of each class is determined as

$$\begin{cases} \dfrac{W_1}{L_2} = \dfrac{W_2}{2L_1} \\ W_1 + W_2 = 1 \end{cases} \tag{2.1}$$

where W_1 , W_2 and L_1 , L_2 are weighted and sample numbers in "mass absent" and "mass present", respectively. The weighted support vector machines compensate for undesirable effects caused by the uneven training class size. This fact improves classification accuracy for the class with small training size.

This method resembles the bootstrap technique, which starts with a reduced training set after the training, retain the classifier by using a new set of images containing some misclassified false positive examples. An interesting property of bootstrap is that the obtained solution converges to the real one as the number of iteration increases [9].

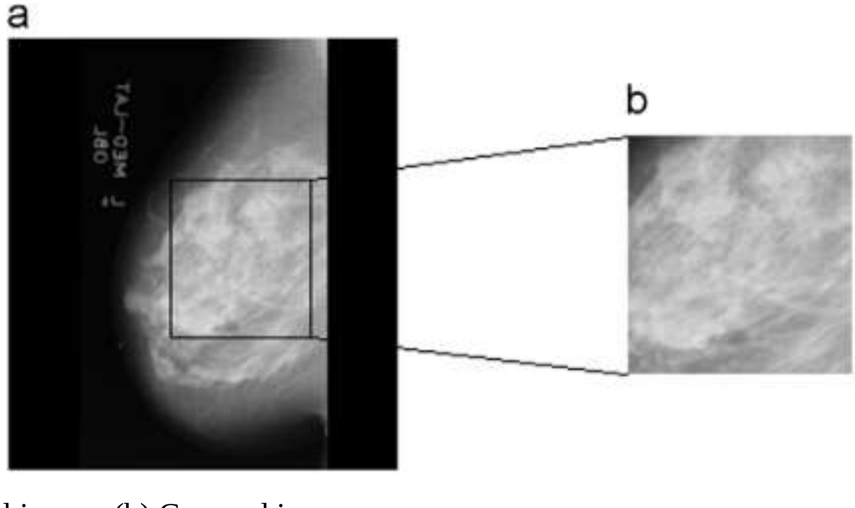

Fig.1 (a) Original image, (b) Cropped image

Fig 2 Sample of the used mammogram images

Results and Discussion

Table 1. Performance comparison of weighted SVM and SEL weighted SVM classifiers using features in detecting normal and abnormal tissues.

Measures	Weighted SVM				SEL weighted SVM			
	Co-occurrence	Wavelets	Contourlets	Wavelets and Curvelets	Co-currence	Wavelets	Contourlets	Wavelets and Curvelets
Mean sensitivity	0.376	0.402	0.458	0.489	0.603	0.601	0.602	0.989
Mean specificity	0.914	0.952	0.977	0.939	0.923	0.954	0.981	0.995
Mean accuracy	0.738	0.780	0.813	0.815	0.851	0.871	0.901	0.993

The original mammograms are 1024×1024 pixels and almost 50% of the whole image comprised of the background with lot of noise. In phase one, we apply a cropping operation to the image to cut off the black parts of the image. Thus, almost all the background information and most of the noise are eliminated. The cropping process was done manually. Regions were extracted with size 128×128 pixels.

The second phase deals with the feature extraction from the ROI's of the set of images. The wavelet and curvelet transform are used to represent the ROI's in multiscale decomposition levels. The ROI's of mammogram images are transformed into four scales. Then the 50 biggest coefficients from each scale decomposition level are used to be the feature vector of the corresponding mammogram.

Phase three is the classification process. Successive enhancement learning weighted SVM classification was accomplished by using the radial basis function (RBF) kernel, and adaptively tuning the parameter C and the kernel parameter to 2^7 and 10^{-6}, respectively.

The dataset used in this work is the Mammographic Image Analysis Society (MIAS). This dataset is selected because of the various cases it includes. It is also widely used in similar research work. It is composed of 322 mammograms of right and left breast, from 161 patients, where 51 were diagnosed as malignant, 64 benign and 207 normal. The abnormalities are classified into micro-calcifications, circumscribed mass, ill-defined mass, speculated mass, architecture distortion, and asymmetry. The method was applied to a set of 90 (60 normal and 30 abnormal cases) mammograms taken from the MIAS dataset. The images were in 8-bit gray resolution format and of size 1024×1024 pixels.

In the medical imaging, the most important performance measures are both specificity and sensitivity. Ideally, one wishes both high specificity and sensitivity measures. Theoretically, however, these two metrics are inversely proportional. Since accuracy is a function of sensitivity and specificity measures together, this descriptor was selected to determine the overall preciseness of the classifier. Experimental results and comparisons between different features and classifier are presented in Tables 1. As Table 1 shows, our approach successfully distinguishes between normal and abnormal tissues while outperforming current methods. The ROC curve is illustrated in Fig. 3. Obviously, the SEL weighted SVM achieves the best accuracy with minimum standard deviation. In order to assess the significance of results, *p*-value is determined as 0.01.

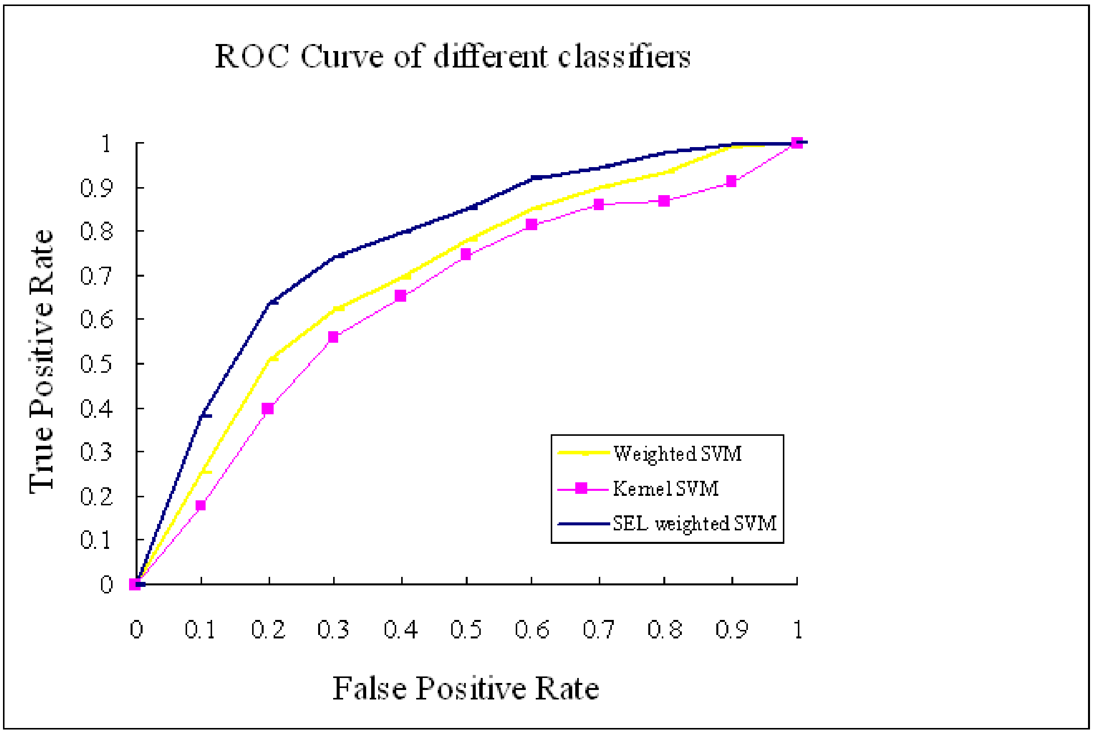

Fig. 3. ROC curve for SEL-weighted SVM, weighted SVM and kernel SVM classifier. It can be seen the area under the ROC curve for SEL-weighted SVM is the best.

Conclusions

Breast cancer diagnosis using digital mammogram is a practical field of investigation. The positive results could affect the mortality ratio of human life. In this paper, we exploited the advantages of multiresolution analysis along with the multiscale geometric analysis and employed SEL weighted SVM classifier to detect and classify the breast masses into normal, benign and malignant cases. Firstly, each mammogram image is decomposed using wavelet and curvelet transforms. The 50 biggest coefficients are extracted from each decomposition level. It was shown that curvelet transform can successfully capture structural information along multiple scales, locations and orientations. This representation offers supplements over the separable 2-D wavelet transform which has limitations in directional analysis of textures. Secondly, the formulation of SVM is based on the principle of structural risk minimization. The weighted SVM compensate for the undesirable effects caused by the uneven training class size. The proposed SEL scheme can further lead to improvement in the performance of the trained weighted SVM classifier. Experimental results demonstrate that SEL weighted SVM approach yielded the best performance when compared to a number of current methods. Mass detection and classification in other tissues based on statistical and intrinsic properties of multiscale geometric analysis certainly seem to be promising areas for future research.

References

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M.J. Thun, Cancer Statisitcs 2008. *CA:A Cancer Journal for Clinicians*, Vol.29(2008) p.71-96.

[2] H. Cheng, X. Shi, R. Min, L. Hu, X. Cai and H. Du, Approach for automated detection and classification of masses in mammograms. *Pattern Recogn*., Vol.39(2006), p.646-668.

[3] H. Cheng, X. Cai, X. Chen and L. Lou, Computer aided detection and classification of microcalcification in mammogram: a survey. *Pattern Recogn. Lett* ., Vol.36(2003), 2967-2991.

[4] G. Kom, A. Tiedeu and M. Kom, Automated detection of masses in mammograms by local adaptive thresholding. *Comput. Biol. Med*., Vol.37(2007), p.37-48.

[5] C. Varelaa, P.G. Tahocesb, A.J. Mndezc, M. Soutoa,and J.J. Vidala, Computerized detection of breast masses in digitized mammograms. *Comput. Biol. Med*., Vol.37(2007), p.214-226.

[6] A. Oliver Malagelada, Automatic mass segmentation in mammographic image. Ph.D. Thesis, Universitat de Girona, *Spain*, 2004.

[7] R. Ferrari and R. Ranayyan, Automatic identification of the pectoral muscle in mammograms. *IEEE Trans Med. Imaging*, Vol.23(2004), p.232-245.

[8] D. Raba, A. Oliver, J. Marti, M. Peracaula and J. Espunya, Breast segementation with pectoral muscle suppression on digital mammograms. *Medical Imaging: Pattern Recognition and Image Analysis*, Vol.3523(2005), p.471-478.

[9] M. Roffilli, Advanced machine learning techniques for digital mammography. Technical Report, Department of Computer Science, University of Bologna, *Italy*, 2006.

[10] L. Semeler, L. Dettori and J. Furst, Wavelet-based texture classification of tissues in computed tomography. in: *IEEE International Symposium on Computer-Based Medical Systems*, 2005.

[11] L. Semeler and L. Dettori, A comparison of wavelet-based and ridgelet-based texture classification of tissues in computed tomography. in: *International Conference on Computer Vision Theory and Appications*, 2006.

[12] R. Mousa, Q. Munib and A. Mousa, Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural network.. *IEEE Trans. Image Process.,* Vol.28(2005), p.713-723.

[13] E.A. Fischer, J.Y. Lo and M.K. Markey, Bayesian networks of BI-RADS discriptors for breast lesion classification. in: *IEEE EMBS*, vol.4, *San Francisco*, 2004, p.3031-3034.

[14] K. Bovis, S. Singh, J. Fieldsend and C. Pinder, Identification of masses in digital mammograms with MLP and RBF nets. *IEEE Trans. Image Process.,* Vol.1(2005), p.342-347.

[15] O.J. Freixener, A. Bosch, D. Raba and R. Zwiggelaar, Automatic classification of breast tissue. in: *Lecture Notes in Computer Science, Pattern Recognition and Image Analysi*s, Springer, Berlin, 2000, p. 431-438.

[16] T. Mu and A.K. Nandi, Detection of breast cancer using v-SVM and RBF networks with self organization selection of centers. in: *Third IEEE International Seminar on Medical Applications of Signal Processing*, 2005.

[17] F. Moayedi, Z. Azimifar, R. Boostani and S. Katebi, A support vector based fuzzy neural network approach for mass classification in mammography. in: *International Conference on Digital Signal Processing*, Britain, 2007.

[18] I. El-Naqa, Y. Yang, M. Wernick, N. Galatsanos and R. Nishikawa, A support vector machine approach for detection of microclassification. IEEE Trans. Medical Imaging, Vol. 21(2002), p.1552-1563.