A FEATURE ANALYSIS APPROACH TO MASS DETECTION IN MAMMOGRAPHY BASED ON RF-SVM

Ying Wang, Xinbo Gao, Jie Li

School of Electronic Engineering, Xidian University, Xi'an 710071, China

ABSTRACT

A new approach to mass detection in mammography is presented. The main obstacle of building a mass detection system is the similar appearance between masses and density tissues in breast. Hence, the various features of the extracted regions of interest (ROIs) are analyzed by synthesis. Then the support vector machine (SVM), which is designed later to distinguish masses from normal areas, is employed to classify these ROIs exactly. To further improve the performance of SVM, the relevance feedback (RF) is introduced to filter out the false positives. The experimental results illustrate that SVM classifier can effectively detect the mass areas, and the RF-SVM scheme can be efficiently incorporated into this learning framework to further improve detection performance.

Index Terms—Image analysis, feature extraction, pattern recognition, relevance feedback

1. INTRODUCTION

As well known, breast cancer is one of leading causes of cancer deaths among women. The early diagnosis and treatment can effectively increase the survival chance of patients. Mass is the major signs of early breast cancer on mammograms. However, it is difficult to distinguish masses from normal tissues since the various appearances of the masses and its ambiguous margins.

Most mass detection algorithms consist of two stages: 1) detection of suspicious regions on the mammography; and 2) classification of suspicious regions as mass or normal tissue. The first stage is designed with a very high sensitivity so that a larger number of false positives can be acceptable. The purpose of the second stage is to reduce the false positives as many as possible which is also the decision procedure of detection methods. Some researchers just focused on the second stage of detection algorithms [1-4]. Sahiner *et al.* proposed a texture feature based convolution neural network for this task [1]. Wei investigated the use of global and local multi-resolution texture features to reduce the false positives which detections on a set of manually extracted ROIs [3]. Brake *et al.* defined several features that were designed to capture image characteristics like intensity, location, contrast *etc* to discriminate lesions from normal tissue [4]. However, the above methods either employed not enough features, or used the classic classifier decision tree, neural network *etc*. So the detection performances need further improving.

Support vector machine (SVM) is a statistical learning method based on structural risk minimization. It has good generalization ability, and is able to compress the useful information of high-dimensional spaces into a small number of elements named support vectors. SVM are therefore capable of learning in sparse, high-dimensional spaces, by using very few training examples. It has been already applied to calcifications detection, giving rise to very good results [5]. A featureless approach based on SVM for the detection of masses has been proposed in [6].

To improve the detection result, a new SVM detection method based on some typical features is attempted to develop in this paper. Features of suspicious areas are extracted and classified by the SVM classifier. To further improve the performance, a relevance feedback method is introduced. The proposed feedback learning model can achieve better detection results, which successfully removes more false positives among suspicious areas.

2. RF-SVM CLASSIFIER

False-positive reduction corresponds to a two-class pattern recognition problem, *i.e.*, distinguishing true masses from false signals. SVM is a very effective binary classification algorithm which could solve this kind of problems [7].

In the general case in which the data points are not linearly separable in the input space, a nonlinear transform-ation is used to map the data vector $\boldsymbol{x}$ into a high dimensional space (called feature space) prior to applying the linear maximum-margin classifier. The discriminant function in an SVM classifier has the following form.

$$f(\mathbf{x}) = \sum_{i=1}^{L_s} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{1}$$

where $K(\cdot,\cdot)$ is the kernel function, $\boldsymbol{x}_i$ are so-called support vectors determined from training data, L_s is the number of support vectors, y_i is the class indicator (e.g.,+1 for class 1 and –1 for class 2) associated with each $\boldsymbol{x}_i$, and α_i are constants, also determined from training.

Since the information of original training samples is limited, in order to improve the performance of SVM classifier, the relevance feedback method is introduced to allow the SVM classifier to learn more from feedback procedure. Relevance feedback (RF) has been applied extensively in image retrieval. It is a post-query process to refine the search by using positive and/or negative indications from the user of the relevance of retrieved images. We combine RF with our SVM detection system, hoping that the feedback information can improve the detection performance. According to the relevance feedback algorithm in [8], a feedback learning method based on SVM, named RF-SVM, is provided.

As we know, Eq.(1) is the classification function of SVM. It represents the distance from classification hyper-plane to each sample, which also denotes whether the sample is correctly classified. During the procedure of feedback, history information also plays an important role. The system will achieve good results rapidly if one adds this information into current feedback process. The history information will incorporate to the feedback procedure by associated the weights with every unlabelled images.

$$w(i) = (1-\beta)\, w(i) + \beta f(x_i) \tag{2}$$

where β is an attenuation coefficient. This coefficient not only keeps the effect of history information, but also emphasizes particularly on the requirement of current detection procedure. The method effectively avoids the result run into local solution, and significantly improves the detection performance.

3. REGION FEATURE ANALYSIS

Feature extraction is a key step in most pattern recognition systems. The general guidelines are: 1) features of patterns in different classes should have significantly different values; 2) features should have similar values for the patterns within the same class; 3) these features should not be strongly correlated to each other; 4) some redundant features should be deleted, and a small number of features is preferred for reducing the complexity of the classifier.

Similarity estimation of object areas is based on the typical features which characterized the areas well. Since the feature extraction of mass is based on region, it is important to select those features which could describe region characteristics well. Many useful image features have been proposed, which can be divided into three categories, namely, intensity, geometric, and texture features. To describe the features of mass as well as possible, we summarize 42 typical features in Tab.1.

Intensity and geometric features are described in detail in Table 1, and the texture features will present briefly as follows. Texture features have been successfully applied into medical image analysis, because they can well depict the texture of images such as uniformity, smoothness and difference among adjacent pixels. Mean value and standard derivation of texture energy map are the Laws texture features we employs, which describes the characteristic of image filtered using Laws template [9]. Features based on co-occurrence matrix are some measures related to specific textural characteristics of the image, such as homogeneity, contrast, entropy and energy [10]. Since Daubechies wavelets D_6 and D_{20} could provide a good combination of regular prototype wavelets with varying sizes to extract texture information with varying spatial frequency [11], the energy and entropy of the decomposed wavelet coefficients are computed as wavelet features.

Table 1 Features of mass

Feature Sub-Space	Features
Intensity Features	Contrast; Invariant moment; Mean gray and gradient of ROIs; Standard derivation inside ROIs; Higher order moments of ROIs; Mean gradient of ROIs boundary;
Geometric Features	Circularity; Compactness; Sphericity; Fourier descriptor
Texture Features	Laws texture; Co-occurrence matrix texture; Wavelet transform texture

4. MASS DETECTION SCHEME

The ROIs in mammograms need to be extracted before classification procedure. In order to extract the suspicious areas exactly, a simple but effective method is employed. It begins with morphological enhancement to remove the back-ground noise and the structure noise inside the suspected mass patterns. Then the regions in enhanced images, which take on certain intensity and contrast values, are extracted and selected as seed regions. The ROIs will be obtained later using fuzzy region grow algorithm [12].

Since the limitation of intensity and contrast are not strict, suspicious regions in mammograms are extracted entirely after coarse detection. Thus, there still exist a large number of false positives.

After the feature extraction step, the appropriate SVM model should be chosen. For selecting models with good performance, a widely used statistical method called *m*-fold cross-validation is adopted. In the experiment, the SVM classifier is trained using a 10-fold cross-validation procedure to confirm the best model and parametric setting. Finally, the kernel function employs the RBF kernel.

$$K(x_i, x) = \exp\left[-\frac{(x-x_i)^2}{2\sigma^2}\right] \tag{3}$$

where $\sigma = 3$ and $C = 100$ are used in this paper, as the generalization error is smallest under this setting.

Let T_p , T_n denote the positive and negative training set, F_p , F_n denote the corresponding feedback sample set, the new mass detection method based on RF-SVM can be described in detail as follows.

Step1 Initializing T_p , F_p , T_n , F_n and validation set V , setting $w(i)$=0.

Step2 Reset the training set as follows before training.

$$T_p = T_p \cup F_p \quad (4)$$

$$T_n = T_n \cup F_n \quad (5)$$

Step3 Training the SVM classifier which contains more sample information as the current classifier.

Step4 Classifying set V then saving the misclassified samples to the preliminary feedback set $WF_p^{'}(k)$ and $WF_n^{'}(k)$. The feedback sets $WF_p(k)$ and $WF_n(k)$, which need to com- puter the weights, should be made sure later.

$$WF_p(k) = \left(WF_p^{'}(k) \cup F_p\right) - F_n \quad (6)$$

$$WF_n(k) = \left(WF_n^{'}(k) \cup F_n\right) - F_p \quad (7)$$

Step5 Computing the weights of misclassified samples of $WF_p(k)$ and $WF_n(k)$ respectively based on Eq. (1) and (2), and then absolute values of these weights are arranged in ascending order. The samples that enjoy small scores are selected to experts or feedback directly for the next learning and detection step.

Step6 The procedure will stop when the detection rate is more than 90%, or the amount of samples in F_p and F_n has been less than 1. This means the current classifier has been reached the requirement of system. Otherwise, go to *step 2* for further feedback and learning.

5. EXPERIMENTAL RESULTS

In this study, the mammograms are selected from USF DDSM database [13], all the lesions in images have been marked by experts. These mammograms are of dimension 5000*3000 pixels, with a spatial resolution of 0.05 mm/pixel and 12-bit/16-bit gray scale. In our experiment, training set included 192 images containing 200 mass regions and 200 negative samples extracted from these images. Another data set of 150 mammograms is used to test and evaluate the performance of the proposed algorithm, in which 100 images are validated data in feedback and the others are test data. According to the coarse detection result, the former contains 132 mass areas with 1084 false-positive areas, and the latter contains 64 mass areas with 599 false-positive areas.

Fig.1 gives some detection results of mammograms. Since the testing set has 663 ROIs, the sensitivity of the SVM classifier is 85.9% with a false-positive fraction of 4.8 marks per image. As shown in Fig.2 (a), the mammary has plenty glandular tissues and the ROIs extracted is shown in (b). Since many regions possess similar features with masses, after detection procedure, only a small fraction of false positives are removed. Thus, the relevance feedback method is used to improve the performance.

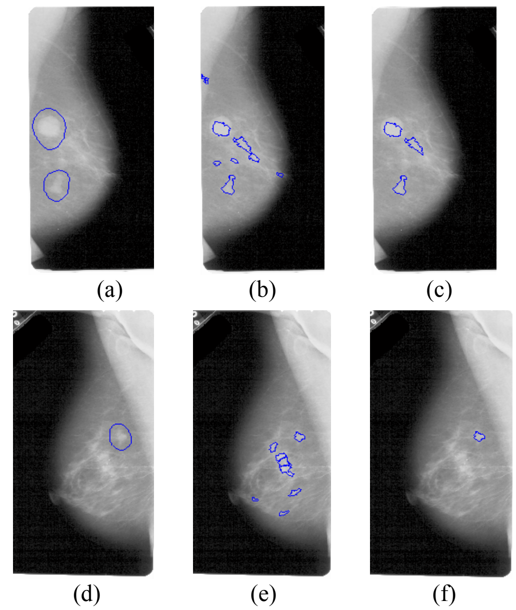

Fig. 1. Detection results of SVM. (a)(d) Original images with expert's mark; (b)(e) ROIs; (c)(f) Detection results

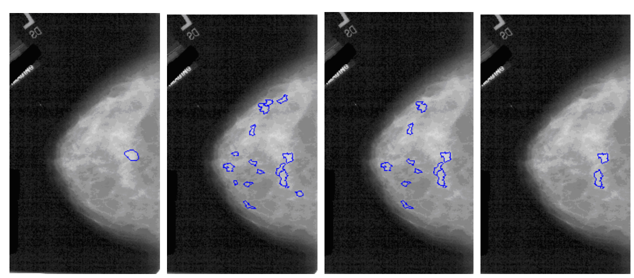

Fig. 2. Detection result of density mammogram of SVM and RF-SVM (a) Original image (b) ROIs (c) Detection result of SVM (d) Detection result of RF-SVM

Fig.3 shows the FROC curves of detection result using SVM and different iterations RF-SVM separately. From the figure, it can be found that the proposed feedback method could improve the detected result. It is also notes that the performance is further improved as feedback more times. In order to verify the efficiency of feedback learning model, Fig.4 shows the detection results on testing set using SVM and RF-SVM which feedback and learns from validation set five times. As we can see from Fig.4, feedback learning model could further improve the detection performance, while remove more false positives in images. SVM classifier trained by feedback learning procedure could make the sensitivity of SVM classifier rise to 90.6% and the false-positive fraction fall to 3.6 marks per image. Compared with the results reported in [6], our method could

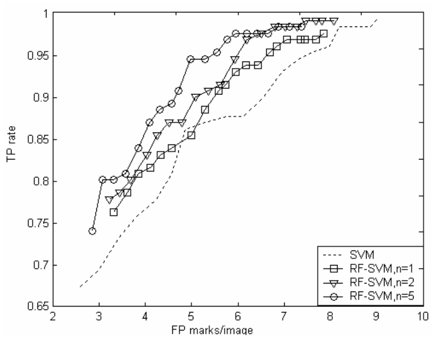

Fig. 3. FROC of RF-SVM detection results

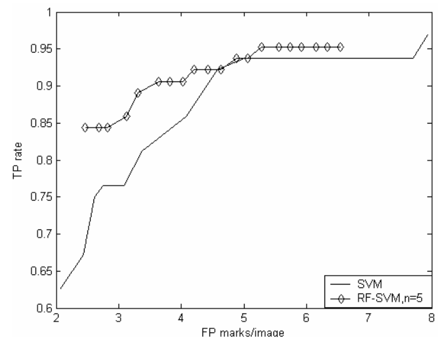

Fig. 4. Detection results of SVM and RF-SVM

obtain higher sensitivity. Then the RF-SVM classifier which feedback run five times is used to classify the ROIs in Fig.2(b), the detection result is shown in Fig.2(d). Obviously, the RF-SVM classifier achieves better detection performance which could remove more false positives.

6. CONCLUSTIONS

In this paper, a mass detection method based on RF-SVM is proposed to distinguish the masses from normal areas correctly. ROIs were extracted firstly, and then the SVM will train and test using features extracted from ROIs. To remove more false positives, relevance feedback method was introduced to improve the performance. Experimental results demonstrate that the SVM classifier can achieve good detection result, and RF-SVM could further improve the detection performance of classifier. We will still study the feedback procedure to improve the generalization ability of the SVM feedback learning model in future.

7. ACKNOWLEDGMENT

This work was supported by the Program for Changjiang Scholars and Innovative Research Team in University, the Program for New Century Excellent Talents in University (NCET-04-0948), and the National Basic Research Program of China (NBRPC, 2006CB705700).

8. REFERENCES

[1] B. Sahiner, H. P. Chan, N. Petrick, *et al.*, “Classification of Mass and Normal Breast Tissue: A Convolution Neural Network Classifier with Spatial Domain and Texture Images,” *IEEE Trans. on Medical Imaging*, vol.15, no.5, pp. 598-610, 1996.

[2] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious, *et al.*, “Computer-Assisted Detection of Mammographic Masses: A Template Matching Scheme Based on Mutual Information,” *Medical Physics*, vol.30, no.8, pp. 2123-2130, 2003.

[3] D. Wei, H. P. Chan, N. Petrick, *et al.*, “False Positive Reduction technique for Detection of Masses on Digital Mammograms: Global and Local Multiresolution Texture Analysis,” *Medical Physics.*, vol. 24, no.6, pp. 903-914, 1997.

[4] G. M. te Brake, N. Karssemeijer, and J. H. Hendriks, “An Automatic Method to Discriminate Malignant Masses from Normal Tissue in Digital Mammograms,” *Physics in Medicine & Biology.*, vol.45, no.10, pp. 2843-2857, 2000.

[5] Issam EI-Naqa, Yongyi Yang, Miles N. Wernick, *et al.*, “A Support Vector Machine Approach for Detection of Microcalcifications,” *IEEE Trans. on Medical Imaging*, vol.21, no.12, pp. 1552-1563, 2002.

[6] R. Campanini, D. Dongiovanni, E. Iampieri, *et al.*, “A Novel Featureless Approach to Mass Detection in Digital Mammograms Based on Support Vector Machines,” *Physics in Medicine & Biology*, vol.49, no.6, pp. 961-975, 2004.

[7] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[8] Issam EI-Naqa, Yongyi Yang, Nikolas P. Galatsanos, *et al.*, “A Similarity Learning Approach to Content-Based Image Retrieval: Application to Digital Mammography,” *IEEE Trans. on Medical Imaging*, vol.23, no.10, pp. 1233-1244, 2004.

[9] P.Miller, K.Bowyer, “Classification of Breast Tissue by Textural Analysis,” *Image and Vision Computing*, vol.10, no.5, pp. 277-283, 1992.

[10] Rafael C. Gonzalez, Richard E. Woods, *Digital Image Processing*, Prentice Hall, New Jersey, 2002.

[11] A.P.Dhawan, Y.Chitre, C.Kaiser-Bonasso. “Analysis of Mammogramphic Microcalcifications Using Gray-Level Image Structure Features,” *IEEE Trans. on Medical Imaging*, vol.15, no.3, pp. 246-259, 1996.

[12] D. Guliato, R. M. Rangayyan, W. A. Carnielli, *et al.*, “Segmentation of Breast Tumors in Mammograms by Fuzzy Region Growing,” *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol.20, no.2, pp. 1002-1005, 1998.

[13] http://marathon.csee.usf.edu/Mammography/Database.html.