Mammogram classification method based on GMM and GLCM-PSO-PNN

Xiaojian Zhang[a], Chengjian Wei[b] and Xili Wan[c]

School of computer science and technology, Nanjing Tech University, Nanjing 211816, China.

[a]156162726@qq.com
[b]1277805603@qq.com
[c]1721257367@qq.com

Abstract. Facing the condition that the inefficient training of traditional classifiers in the classification process of mammography, a classification method is proposed combining image processing and supervised learning. Firstly, the improved adaptive median filter enhances the image contrast. Then, according to the result of breast segmentation based on Gauss Mixture Model (GMM), this paper proposed a classification model based on Probabilistic Neural Network optimized (PNN) optimized by Gray Level Co-occurrence Matrix (GLCM) and Particle Swarm Optimization (PSO). The eigenvector extracted from the GLCM can be used as input to simplify the network structure. The smoothing factor optimized by PSO used to train the network can improve accuracy. The results in public mammographic patches database demonstrate that the model can classify the types of mammography effectively and perform better than the previous methods.

Key words: mammogram; gauss mixture model; probabilistic neural network; gray level co-occurrence matrix; particle swarm optimization

INTRODUCTION

According to the latest research, the mortality rate of breast cancer has been increasing year by year, and it is the most easily available cancer in the world, accounting for 25% of the total number of newly diagnosed cases. Early detection and correct diagnosis can greatly improve the survival rate of breast cancer. Masses and a large number of micro calcification clusters are early signs of breast cancer, but also an important basis for distinguishing between benign and malignant tumors. Therefore, using the rapid development of computer-aided diagnosis (CAD) technology to assist experts in analyzing and detecting suspicious lesions and improving the accuracy of early diagnosis of breast tumor has become a research hotspot.

In recent years, domestic and foreign scholars have done a lot of detailed work on CAD of breast tumor recognition. Q. Jin et al. [1] improved the Back Propagation (BP) neural network by adding momentum and adaptation rate to reduce the required time of training the model. Abirami et al. [2] used wavelet transform to extract image features. Multi-layer Perceptron was used as a classifier to obtain better classification results on public datasets. Lin et al. [3] proposed a support vector machine (SVM)-based classification method with better performance than K-nearest neighbors. Literature [4] [5] respectively used Sequential Minimal Optimization (SMO) and Particle Swarm Optimization algorithm, improving the performance of SVM. All the studies above belong to two-class mammogram classification, but mammogram has normal attributes besides benign and malignant. For three-class mammogram classification, Uppal et al. [6] fuse the features of discrete cosine transforms and wavelet transforms, using variety of classifiers for classification. Vani et al. [7] used Extreme Learning Machine (ELM) as classifier and proved that the method is superior to other techniques. Recently, deep learning has received extensive attention in the image processing field. Jadoon et al. [8] used convolutional neural networks (CNN) to classify mammogram and compared the performance of softmax and SVM as output layers. Unlike the method of optimizing classifiers, accurate segmentation of pathological features in images can also improve the accuracy of diagnosis.

Rahimeh et al. [9] proposed a hybrid algorithm based on regions and contours, combined cluster segmentation techniques to identify benign and malignant tumors. Although the methods above show certain effectiveness in the specific classification process, it is difficult to meet the requirements of the auxiliary clinical diagnosis.

A new diagnostic model called GLCM-PSO-PNN combining the advantages of image processing and supervised learning is proposed. Compared with the traditional feed forward neural network, the PNN absorbs the advantages of Radial Basis Function (RBF) neural network to make its performance better. The GLCM-PSO-PNN adopts the GLCMs to extract the main features, and compensates for the disadvantages of the PNN, the increasing of complexity caused by too large training samples, which is unfavorable for hardware implementation [10]. At the same time, we use PSO algorithm to select the best parameter and improve the accuracy without losing the training efficiency of classifier. The experimental results show that the model is better than previous methods. The algorithm and simulation have been implemented in MATLAB.

IMAGE PROCESSING

Contrast enhancement

Due to the uneven contrast of mammograms. In this paper, an improved adaptive median filter is used to enhance the contrast of scaled images. In order to preserve the complete details, the size of sliding window need to change.
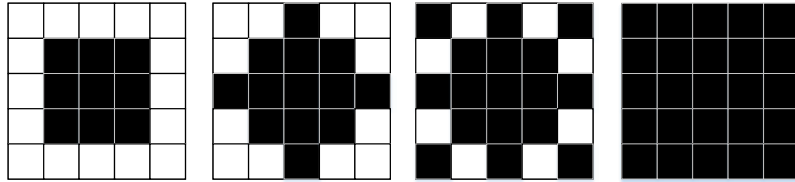

FIGURE 1. Improved sliding window.

The specific steps for processing original images are as follows:

Step 1: The original images belong to RGB mode with size of 1024×1024, which is scaled to 256×256 and converted to Grey.

Step 2: Calculate $A_1 = Z_{med} - Z_{min}$, $A_2 = Z_{med} - Z_{min}$. If $A_1 \leq 0 \parallel A_2 \geq 0$, and the window size is smaller than S_{max}, it will increase as shown in Fig. 1, otherwise, step 2 is entered. （Z_{min}, Z_{max} and Z_{med} are the minimum, maximum, and median values of the image grayscale, respectively)

Step 3: For each pixel Z_{xy} of images, calculate $B_1 = Z_{xy} - Z_{min}$, $B_2 = Z_{xy} - Z_{max}$. If $B_1 > 0$ & $B_2 < 0$, grayscale value does not change; if $B_1 \leq 0 \parallel B_2 \geq 0$, $Z_{xy} = Z_{med}$.

GMM-based Segmentation

After obtaining contrast-enhanced grayscale images, dividing them into N regions based on GMM. Pixels of each region obey the Gaussian distribution. Therefore, the images feature can be represented by a weighted mixed Gaussian distribution:

$$p(x / \lambda) = \sum_{i=1}^{N} w_i g(x \mid u_i, \sigma_i) \qquad (1)$$

Where x is D-dimensional continuous image feature vector. $g(x|\mu_i,\sigma_i)$, $i=1,\cdots N$ is one of Gaussian distributions. w_i, $i=1,\cdots N$ are mixed weights and satisfy:

$$\sum_{i=1}^{N} w_i = 1 \qquad (2)$$

The gray image segmentation can be seen as a two-class problem, so the category of all pixel points can be inferred. The image to be divided is defined as Y=$(y_1,\ldots, y_n)$ where y_n is the gray value of each pixel. Set of pixel points is set to X=$(x_1,\ldots, x_n)$ where $x_n \in$ L. L is 2-Dimension vector (0, 1). According to the maximum a posteriori (MAP) probability criterion, the set of class labels X* is defined as:

$$X^* = \arg \max_{x} \{ P(Y \mid X, \Theta) P(X) \} \qquad (3)$$

The prior probability P(X) obeys the Gibbs distribution, and the joint likelihood probability is defined as:

$$P(Y \mid X, \Theta) = \prod_i P(y_i \mid x_i, \theta_{x_i}) \tag{4}$$

Where $P(y_i \mid x_i, \theta)$ is a Gaussian distribution that obeys parameter $\theta = (\mu, \sigma)$. In the Markov random field problem, parameters set $\Theta = \{\theta_l \mid l \in L\}$ is obtained by prior knowledge. In this paper, the Hidden Markov Random Field-Expiration Maximization (HMRF) framework [11] is used. The parameters set is learned in an unsupervised way. The algorithm is mainly divided into E steps and M steps. First, initialize parameters set Θ^0, then

Step E: Obtain Θ^t after t iterations, and calculate the conditional expectation:

$$Q(\Theta \mid \Theta^t) = E\left[\ln P(X, Y \mid \Theta) \mid Y, \Theta^t\right] \tag{5}$$

Step M: Update $Q(\Theta|\Theta^t)$ to get a more accurate estimate of the segmented image:

$$\Theta^{t+1} = \arg\max_\Theta Q(\Theta \mid \Theta^t) \tag{6}$$

Alternating these two steps, using the current parameters set Θ^t. Label set X_t is estimated according to MAP [12]:

$$X^t = \arg\max_X \left\{ P(Y \mid X, \Theta^t) P(X) \right\} = \arg\min_X \left\{ U(Y \mid X, \Theta^t) + U(X) \right\} \tag{7}$$

In summary, the GMM-based HMRF-EM algorithm segmentation steps are as follows:

Step 1: Initialization parameter Θ^0, where μ is the mean value of gray scale, and σ is 0.

Step 2: Substitute the obtained parameter $\theta = (\mu, \sigma) \in \Theta$ into equation (1). In EM algorithm, the Gaussian mixed distribution is used to calculate the latest parameter Θ^{t+1}.

Step 3: According to equation (3), estimate the set of pixels and classify it into the region with highest probability.

Step 4: Update the parameter set by EM algorithm. As shown in Eq. (7), updating of parameters can maximize the probability of label set while minimizing the total energy.

Step 5: Repeat Step 3 until there is no significant change about energy $U(Y|X, \Theta^t)$ in Eq. (7) or reach 10 iteration.

IMAGE CLASSIFICATION

Feature extraction

After the segmentation, mammary region and adipose tissue can be clearly identified. After that, image features can be extracted by generating Gray Level Co-occurrence matrix. The GLCMs representing images are stored in a $i \times j \times n$ matrix. i and j represent the number of gray levels. n is determined by the fixed positional relationship (d, θ) between two pixels. Where d and θ represent the distance and angle. During the experiment, d = {1}, θ = {0°, 45°, 90°, 135°}. So there are 4 gray level co-occurrence matrices for each region. The following characteristics are calculated: autocorrelation, contrast, correlation, clustering highlighting, clustering shading, energy, entropy, homogeneity, maximum probability, sum of squares, sum of mean, sum of variance, sum of entropy, variance of variance, Differences in entropy, related information measures 1, related information measures 2, normalization of deficits, normalization of inverse moments.

Improved classifier

Traditional PNN is a kind of feedforward neural network. Compared with BP neural network, it has faster training speed and simpler structure, so it is widely used in the field of pattern recognition and fault detection [13]. The PSO-PNN structure is shown in Fig. 2.
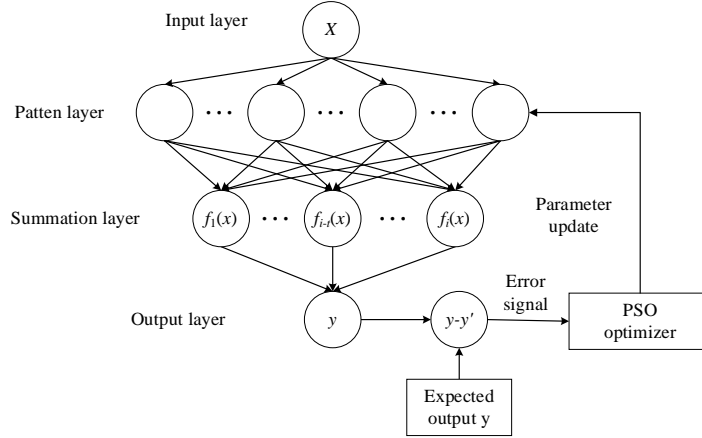
FIGURE 2. Training process

The first layer is input layer. Enter the training samples $X=(x_1,x_2, \ldots ,x_d)^T$ that consists of different features. The second layer is pattern layer which consists of n neurons. The value of n is equal to the sum of categories. Each neuron node has a center. The Gaussian function calculates the distance between the input vector and the center, then returns a scalar value:

$$\phi_{ij}(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma^d}\exp\left[-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}\right] \tag{8}$$

Where i=1, 2...N, N is the number of training samples. d is the dimension of the sample space and x_{ij} is the j-th center of the i-th sample. σ is a smoothing factor that plays an important role in classification. The third layer is summation layer. Sum up the neuron output belonging to the same category in the pattern layer and calculate the average value:

$$f_i(x) = \frac{1}{L}\sum_{j=1}^{L}\phi_{ij}(x) \tag{9}$$

Where L represents the number of training samples of the i_{th} class. The last layer is output layer. Each output are multiplied by the prior probability of the corresponding category. The maximum value of products corresponds to the desired category for the testing sample:

$$y = \arg\max\left(f_i(x)\right) \tag{10}$$

Choosing the smoothing factor in Eq. (8) according to experience does not reflect the probability feature of the sample space well, resulting in lower classification accuracy.

PSO algorithm is easy to implement, besides has good robustness and computational efficiency. Therefore, it is often used to optimize the parameters of the neural network [14]. In the D-dimensional search space, the population Q consists of n particles, $Q = (\sigma_1,..., \sigma_n)$. The position of the i-th particle $\sigma_i=(\sigma_{i1},...,\sigma_{iD})^T$, which is a potential optimal solution of the smoothing factor σ. $\Lambda\sigma_i$ is the velocity of the i-th particle, with the same dimension and position. During the iteration, each particle of the population updates its own velocity $\Lambda\sigma_g$ and position σ_g according to the following eqution:

$$\Lambda\sigma_g^{t+1} = \lambda\Lambda\sigma_g^t + c_1r_1\left(P_i - \sigma_g^t\right) + c_2r_2\left(P_g - \sigma_g^t\right) \tag{11}$$

$$\sigma_g^{t+1} = \sigma_g^t + \Lambda\sigma_g^{t+1} \tag{12}$$

Where P_i and P_g represent the best position of the individual and population respectively. Where λ is the inertia weight, which controls the current speed and its effect to the next speed. c_1 is the cognitive learning rate, c_2 is the social learning rate, which are all non-negative constants, prompting each particle to reach the best position. r_1 and r_2 are two random numbers in the range [0,1], and t is the number of iterations.

In summary, the GLCM-PSO-PNN-based Mammogram classification procedure is as follows:

Step 1: The input data consists of feature vectors which contains 78 elements. The first and second elements are the number and category of mammogram (common, benign, and malignant) respectively. The rest are image features extracted by GLCMs.

Step 2: Initialization parameters, the maximum number of iterations is 30, the population size is 20, and the acceleration factors c1 and c1 are set to 1.4944. The initial value of smoothing factor is 1.5, and limit it into [0.1, 10].

Step 3: The fitness value of the particle is calculated by the mean-square error of the actual output y and the expected output y′ in Fig. 3.

Step 4: The speed and position of the particles are updated according to Eq. (11) and Eq. (12). If the fitness value of a certain particle is higher than any previous moment, it is regarded as the individual extreme value. If higher than all particles, it is regarded as the population extreme value.

Step 5: GLCM-PSO-PNN is trained with optimized parameters. Enter data for testing and analyze the results.

SIMULATION

Data sources and processing

The simulation data was from mini-MIAS Database provided by the Mammographic Image Analysis Society [15]. The data set contains 322 breast tumor cases, among which 66 are benign, 54 are malignant, and the rest are normal. In this paper, the data samples are randomly divided into two sets, where the training set includes 222 samples, and the testing set includes 100 samples. As shown in Fig. 3. The first column are original image (1024 ×1024). Too much useless background and higher pixels are not conducive to subsequent feature extraction. So, the useless region was cropped; then, the RGB image were converted to Grey and enhanced the contrast; finally, features were extracted from segmentation image.
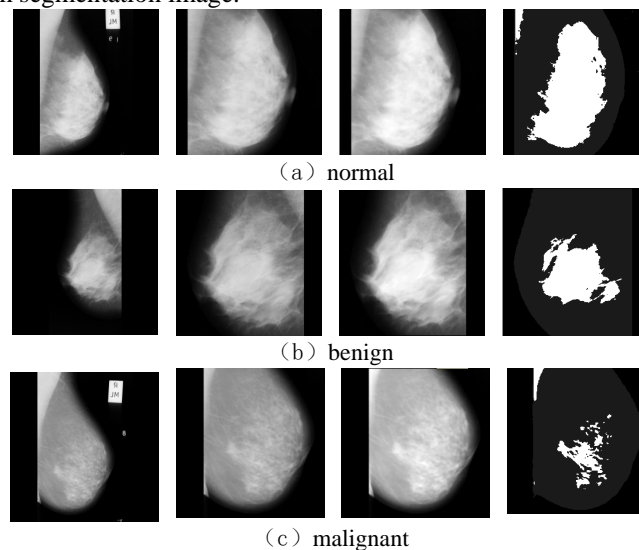

（a）normal

（b）benign

（c）malignant

FIGURE 3. Mammogram processing

Results analysis

The variation of σ and the corresponding accuracy changes with iteration in PSO is shown in Fig. 4. The algorithm converges so quickly that the best smoothing factor is found by less than 20 iterations. After 30 iterations of algorithm, the optimal smoothing factor σ_{best} is 1.96, and the classification accuracy rate of training set is 98.6%. The σ_{best} is applied for training PNN to achieve the discrimination of the testing set. Fig. 5 and Fig. 6 shows the category recognition effect and testing errors of the classifier. Finally, only one case of malignant samples was misjudged as benign, and the classification accuracy reached 99%.
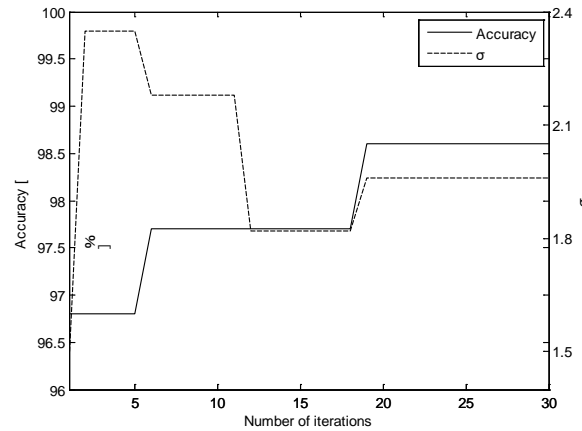
FIGURE 4. The variation of σ and corresponding accuracy with iteration
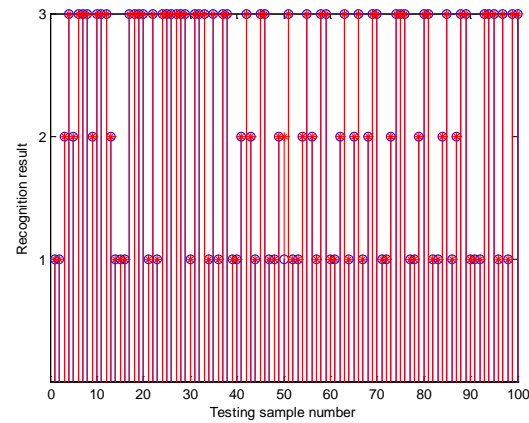
FIGURE 5. Category recognition result
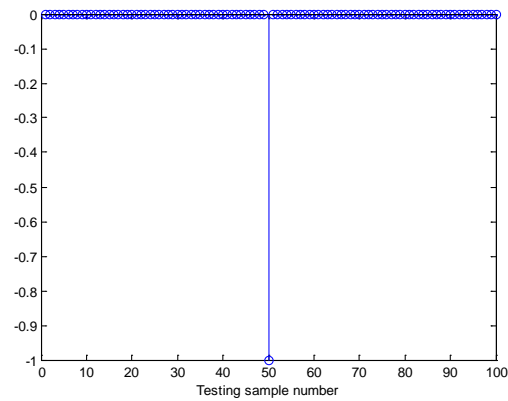
FIGURE 6. Error of testing result

Method comparison

In order to further analyze and verify the performance of this model, we compared the methods as follows: the SVM optimized by PSO [5]; DCT or DWT extracts features, SVM as classifier [6]; the extreme learning machine [7] and Daubechies wavelet extracts features, the radial basis function neural network as classifier [2].Besides, in order

to verify the help of image segmentation, the pre-segmentation images are also used. The comparison results are shown in Table 1.

TABLE 1. Comparison of different Methods.

Method	specificity	sensitivity	Average accuracy
Ours	98%	98%	98%
PSO-SVM	88%	80%	84%
DCTODWT	98%	94%	96%
ELM	90%	98%	94%
DB4+RBF	94%	92%	93%
PSO-PNN	92%	84%	88%

Due to the lack of benign and malignant samples in the mini-MIAS database, the original 100 samples were expanded to 400 by rotating; 300 of them were randomly selected as the training set and the rest were used as test set. As demonstrated by the results in Table I, the specificity, the sensitivity and the average accuracy of our method is higher than others, which proves it can find the malignant tumor more accurately.

CONCLUSION

This paper proposes a novel mammography classification method combined with image segmentation technology. In a visual point view, contrast enhance and segmentation of images allows experts to make more accurate diagnoses with their knowledge and experience. GMM replaces a single Gaussian distribution and can model more complex distributions. At the same time, low-dimensional features are extracted from GLCMs, which simplifies the structure of network and facilitates hardware implementation. The proposed approach incorporates PSO technique to find appropriate smoothing factor for PNN with respect to resulting recognition accuracy. The simulation results show the performance, especially the recognition accuracy of GLCM-PSO-PNN is superior to other methods. Hence the optimized model can assist experts in early diagnosis.

ACKNOWLEDGMENTS

REFERENCES

1. Q. Jin and P.Z. Gao, Study on Breast Cancer Diagnosis Based on Artificial Neural Network. Computer Simulation, no. 06, 2011, pp. 235-238.
2. C Abirami, R Harikumar, S.R.S. Chakravarthy. Performance analysis and detection of micro calcification in digital mammograms using wavelet features International Conference on Wireless Communications, Signal Processing and NETWORKING. IEEE, 2016, pp. 2327-2331.
3. J Lin, D. L. Heinz, A. J. Campbell, et al. Computer aided detection of breast masses in mammography using support vector machine classification SPIE Medical Imaging. International Society for Optics and Photonics, 2011, pp. 843-845.
4. Y. L. Zhang, H. B. Shi, W. L. Shang, et al. Improved method for computer-aided diagnosis of breast cancer based on support vector machines Application Research of Computers, no. 30(8), 2013, pp. 2373-2376.
5. Gu Y, X.Q. Lu, Y Zhao et al. Research on Computer-Aided Diagnosis of Breast Tumors Based on PSO-SVM. Computer Simulation, no. 32(5), 2015, pp. 344-349.
6. M.T.N. Uppal. Classification of mammograms for breast cancer detection using fusion of discrete cosine transform and discrete wavelet transform features 2016.
7. G Vani, R Savitha, N Sundararajan. Classification of abnormalities in digitized mammograms using Extreme Learning Machine International Conference on Control Automation Robotics & Vision. IEEE, 2011:2114-2117.
8. M.M. Jadoon, Q Zhang, I.U. Haq, et al. Three-Class Mammogram Classification Based on Descriptive CNN Features BioMed Research International, 2017:3640901.

9. R Rouhi, M Jafari. Classification of benign and malignant breast tumors based on hybrid level set segmentation Expert Systems with Applications, no. 46(c), 2016, pp. 45-59.

10. P.A. Kowalski, M. Kusy Sensitivity Analysis for Probabilistic Neural Network Structure Reduction. IEEE Trans Neural Netw Learn Syst, 2017, pp. 1-14.

11. Y Zhang, M Brady, S Smith Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging, 2001, 20(1), pp. 45-57.

12. S.A. Shah, N.C. Chauhan. An Automated Approach for Segmentation of Brain MR Images using Gaussian Mixture Model based Hidden Markov Random Field with Expectation Maximization. Journal of the Royal Society of Medicine, no. 2(78), 2015, pp. 380-7.

13. R Ouhibi, S Bouslama, K Laabidi. Faults classification of asynchronous machine based on the probabilistic neural network (PNN) International Conference on Control Engineering & Information Technology. IEEE, 2017, pp. 1-7.

14. Q Ge, C. J. Wei. Approach for optimizing echo state network training based on PSO Computer Engineering and Design, no. 30(8), 2009, pp. 1947-1949.

15. J Suckling. The Mammographic Image Analysis Society Digital Mammogram Database"Exerpta Medica. 1994, 1069.