A Computer-Aided System for Classification of Breast Tumors in Ultrasound Images via Biclustering Learning

Qiangzhi Zhang[1], Huali Chang[1], Longzhong Liu[2], Anhua Li[2], and Qinghua Huang[1(✉)]

[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
qhhuang@scut.edu.cn
[2] The Cancer Center of Sun Yat-sen University, Guangzhou, China

Abstract. The occurance of Breast cancer increases significantly in the modern world. Therefore, the importance of computer-aided recognition of breast tumors also increases in clinical diagnosis. This paper proposes a novel computer-aided diagnosis (CAD) method for the classification of breast lesions as benign or malignant tumors using the biclustering learning technique. The medical data is graded based on the sonographic breast imaging reporting with data system (BI-RADS) lexicon. In the biclustering learning, the training data is used to find significant grading patterns. The grading pattern being learned is then applied to the test data. The k-Nearest Neighbors (k-NN) classifier is used as the classifier of breast tumors. Experimental results demonstrate that the proposed method classifies breast tumors into benign and malignant effectively. This indicates that it could yield good performances in real applications.

Keywords: Computer-aided diagnosis (CAD) · Breast cancer · Biclustering · K-NN · BI-RADS

1 Introduction

Breast cancer has a high death rate and kills millions of women all over the world each year. However, there is still lack of effective treatment yet. So, early detection becomes very important to increase the survival rate of patients [1]. Since ultrasound imaging (USI) has advantages of being non-radiative, non-invasive, real-time and low-cost, it becomes one of the most important methods for breast cancer diagnosis [2]. To reduce the subjective dependency and improve the diagnostic accuracy, computer-aided diagnosis (CAD) system is badly needed to obtain reliable diagnostic results.

Generally, the traditional CAD system for breast cancer based on the USI involves four stages, i.e. image preprocessing, segmentation, feature extraction and selection, and classification. Owing to the complex imaging environment and principle, the USI unavoidably contains a lot of artifacts and noises which leads to speckles, shadows and low contrast. Therefore, a preprocessing procedure for speckle reduction is necessary. Image segmentation is one of key procedures in the CAD system. But it is worth to mention that it is often difficult to segment breast ultrasound (BUS) images due to the speckles and low contrast which are inherent in BUS images. It is also critical to find a feature set of breast tumor which can distinguish between benign and malignant

X. Wang et al. (Eds.): ICMLC 2014, CCIS 481, pp. 24–32, 2014.
DOI: 10.1007/978-3-662-45652-1_3

accurately. However, the traditional CAD system usually performs the diagnosis based on texture, color, Doppler and morphologic features of breast tumors that have some limitations on breast tumor diagnosis. Finally, it is very time consuming to train a good classifier with high classification accuracy. In such a CAD framework, the final diagnostic results will be unreliable if any one of the four stages does not perform well and the whole process is complicated.

A number of CAD systems for breast cancer based on the USI have been presented in recent years. Ding et. al. [3] applies multiple-instance learning method to classify the tumors in BUS images. Chang et. al. [4] proposes an automatic tumor segmentation and a shape analysis technique to improve the distinction between benign and malignant breast tumors. Garra et. al. [5] uses quantitative analysis of ultrasound image texture to improve the ability of ultrasound to distinguish benign from malignant breast tumors. In recent years, Doppler spectral analysis serves as a useful tool in distinguishing between benign and malignant breast tumors. Some studies have been conducted based on the Doppler spectral analysis to achieve improved performance for the classification of breast tumors [6], [7], [8]. A combination of wavelet transform and neural networks techniques is proposed to improve the accuracy of breast tumor classification [9].

The breast imaging reporting and data system (BI-RADS) lexicons are designed to aid radiologists in describing abnormalities for sonographic breast findings. In recent years, many studies have shown that breast tumor analysis can achieve better performance based on the BI-RADS lexicons [10], [11]. On the other hand, biclustering has emerged as one of the most popular tools for data mining. It identifies object with the same attribute or the same function which would be more beneficial for mining important information.

In this paper, the proposed method adopts a new CAD framework for the classification of breast cancers as benign or malignant. Methods for grading the input image data under features recommended by the sonographic BI-RADS lexicons [12] and biclustering grades using the prior label information for mining diagnostic patterns are proposed. Finally, the k-NN classifier is applied to perform classification of tumor types.

This paper is organized as follows. Section 2 provides a brief introduction of the novel CAD system being proposed. Section 3 presents and discusses the experimental results of the proposed method. We conclude this paper in Section 4.

2 Proposed Methods

In our study, each tumor in BUS image is described by 17 features (from the sonographic BI-RADS lexicon) which are graded with 4 grades by experienced clinicians. Then, the biclustering learning algorithm is applied to mine useful information in feature matrix which we get by grading. At last, the k-NN classifier is employed to perform classification.

2.1 Grading of Medical USI Data

The American College of Radiology (ACR) commission has published an ultrasound lexicon for breast cancer called the breast imaging reporting and data system

(BI-RADS) in 2003 to standardize the reporting of sonographic breast cancer finding and as a communication tool for clinicians [13]. This lexicon includes descriptors of features such as mass shape, orientation, lesion boundary, echo pattern margin, posterior acoustic features, and blood supply, etc [12]. Several studies have confirmed the utility of these features described in the sonographic BI-RADS lexicon in distinguishing benign and malignant breast tumors [14], [15], [16].

Many studies have shown that breast tumor in BUS image has a number of sonographic characteristics significantly different for malignant and benign tumors, which allows the classification as either malignant or benign [12], [17], [18]. For example, malignant tumor is commonly hypoechoic lesion with ill-defined border which is "taller than broader", with spiculate and angular margin, duct extension and branch pattern. However, benign tumors usually tend to have a smooth and well circumscribed border, three or fewer gentle lobulations and thin echogenic capsule. Additionally, a BUS image which shows a malignant nodule commonly has internal calcification, posterior shadowing and an abundant blood supply in or around the tumor. However, these characteristics rarely appear or have a comparatively slight degree in BUS image with benign nodules.

Based on above-mentioned, we choose 17 features (according to the sonographic BI-RADS lexicon) to describe each tumor and grade them with 4 grades. Taking the feature of blurred boundary as an example, clinician would give it a grade 1, 2, 3 or 4 (corresponding to not blurred, slightly blurred, relatively blurred and extremely blurred, respectively) based on a given BUS image. Thus, a two-dimensional feature matrix is constructed, and each row represents the features of a BUS image with label (benign or malignant), each column represents the feature grades given by experienced clinicians. The features being used according to the sonographic BI-RADS lexicon are listed in Table 1.

2.2 Biclustering Learning Algorithm

Biclustering algorithm is a data mining method to find statistically significant sub-matrix (also called bicluster) in a data matrix [20], [21], [22], [23]. After constructing a feature matrix by grading, biclustering learning algorithm is performed to get biclusters which will be the training set of k-NN classifier. Details are as follows.

We first apply a traditional agglomerative hierarchical clustering (HC) method [19] on each column of the two-dimension feature matrix A (for convenience, A represents the feature matrix later on) to get bicluster seeds. Given that A has R rows and C columns, bicluster seeds of every column are formulated by:

$$[C_s(i, j), N_{cl}(j)] = HC\_(j, \tau), j = 1, ..., C \quad (1)$$

where *HC* is the agglomerative hierarchical clustering algorithm, τ is the distance threshold for *HC*, $C_s(i, j)$ and $N_{cl}(j)$ are the *i*th bicluster seed and the number of clusters in the *j*th column, respectively.

Table 1. The features being used by proposed system according to the sonographic bi-rads

Features					
1.	Shape	2.	Orientation	3.	Margin
4.	Indistinct	5.	Angular	6.	Microlobulated
7.	Spiculated	8.	The boundary	9.	Echogenicity
10.	Internal echo pattern	11.	Posterior echo pattern	12.	Blood supply
13.	Effect on surrounding tissue ---Edema	14.	Effect on surrounding tissue --- Architectural distortion	15.	Effect on surrounding tissue ---Ducts
16. Calcification in mass		17.	Calcifications out of mass		

As aforementioned, the detected clusters from all columns are treated as bicluster seeds. After all the bicluster seeds are discovered, the next step is to expand each of them along the column direction according to the mean-square-residue (MSR) score [20] of the sub-matrix. MSR score has been mostly used as a natural assessment criterion for the quality of bicluster. Whether or not to delete a column in the refining step is depended on if the newly formed sub-matrix satisfies the criterion of MSR score (i.e. the MSR score is less than a preset threshold δ) [21].

Given an R×C matrix, MSR score is formulated as follows:

$$H(I,J)=\sum_{i\in I,j\in j}\frac{(a_{ij}-a_{iJ}-a_{Ij}+a_{IJ})^2}{|I||J|} \tag{2}$$

$$a_{iJ}=\frac{1}{|J|}\sum_{j\in J}a_{ij},\ a_{Ij}=\frac{1}{|I|}\sum_{i\in I}a_{ij} \tag{3}$$

$$a_{IJ}=\frac{1}{|I||J|}\sum_{i\in I,j\in J}a_{ij} \tag{4}$$

where a_{ij} represents the element value in the matrix corresponding to row i and column j, $H(I, J)$ is the value of MSR score of bicluster. Only if the MSR score of newly formed sub-matrix is less than the preset threshold δ, then accept it as a valid bicluster.

Finally, this study defines a support-based metric as a measure to help make the classification after all valid biclusters have been detected. We make use of the labels of ultrasound images to calculate the support of biclusters discovered in the previous step, which is expressed as:

$$Spt(s_m)=\frac{R_m}{R_i},Spt(s_b)=\frac{R_b}{R_i} \tag{5}$$

where $Spt(s_m)$ and $Spt(s_b)$ denote the supports of malignant and benign, respectively, R_m denotes the number of rows with malignant label in the ith bicluster, R_b denotes the number of rows with benign label, and R_i is the total number of rows of the ith bicluster.

To make a classification decision, we have to know the maximum of the supports, as follows:

$$S(\bullet) = \max(Spt(s_m), Spt(s_b)) \tag{6}$$

where S denotes the maximum value among $Spt(s_m)$ and $Spt(s_b)$, indicating that the type of the bicluster is determined by the signal with the largest support.

In order to obtain more reliable biclusters, we select the bicluster whose S is larger than a pre-defined threshold S_{pre} as a useful classification signal. S_{pre} is set to 0.8 in this paper. The detected biclusters with support information are grouped into three types of signals: benign, malignant and no-action signals. For instance, given that $Spt(s_m)$ = 0.85, $Spt(s_b)$ = 0.15, respectively, the bicluster is regarded as a malignant signal and given a malignant label, because $S = \max(Spt(s_m), Spt(s_b)) = Spt(s_m) = 0.85$ is greater than S_{pre}. For another instance, if the supports of malignant and benign are 0.4, 0.6, respectively, the bicluster is classified as a no-action signal, due to $S = 0.6$ which is below the pre-defined threshold 0.8. Additionally, we remove the useless bicluster whose S is less than S_{pre}. Thus we can achieve several meaningful signals which are more beneficial to classification.

2.3 K-NN Classifier

In pattern recognition, k-NN classifier is one of the most popular classifiers [24]. The idea of k-NN algorithm is quite simple and straightforward. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors measured by a distance function.

After previous steps, we obtain several useful biclusters from the training set with label information. A bicluster is translated into a classification rule by averaging each column. The output classification rule is a vector where the value of an element is the mean of the corresponding indicators values over the BUS images included by the bicluster. Then, calculate the modified Euclidean distances between object in test set and all valid biclusters acquired from training set. If $X=(x_1,x_2,...,x_m)$ and $Y=(y_1,y_2,...,y_n)$ ($m <= n$, $n = 17$) are two samples taken from the training set and testing set, separately. The modified Euclidean distance function used to compute the similarity between two samples is defined as following:

$$D(X,Y) = \frac{1}{m}\sqrt{\sum_{i \in m}(x_i - y_i)^2} \tag{7}$$

where x_i, y_i are grades of features given by clinician.

We set k to be 3 in our study (also try other values, but get best performance when k is set to 3). In order to overcome the problem of small dataset, we adopt the leave-one-out cross validation method to evaluate the performance of our method. The advantage of leave-one-out cross validation is that each instance in one dataset can be used for both training and testing.

The flow diagram of the proposed system is shown in Figure 1.

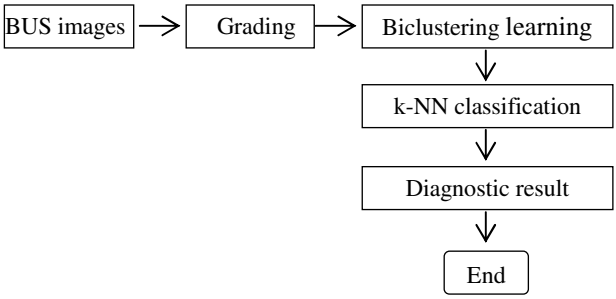

Fig. 1. The flow diagram of the proposed system

3 Experimental Results

The proposed system is developed by VC++6.0 (Microsoft Co. Ltd., USA) and evaluated using a 51 BUS images dataset (29 malignant and 22 benign). The BUS images are provided by the Cancer Center of Sun Yat-sen University and imaged by an IU22 SonoCT System (Philips Medical Systems) with a L12-5 50mm Broadband Linear Array at the imaging frequency of 7.1MHz.

To assess the performance of the proposed method, the specificity, sensitivity and accuracy are used to evaluate the performance of the classification capability of the proposed method. Define the number of correctly classified benign and malignant tumors as true negative (TN) and true positive (TP), respectively. The number of incorrectly classified benign and malignant tumors as false negative (FN) and false positive (FP), respectively.

$$Specificity(TNF) = \frac{TN}{TN + FP} \tag{8}$$

$$Sensitivity(TPF) = \frac{TP}{TP + FN} \tag{9}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{10}$$

In order to validate the accuracy of the proposed scheme, it is compared with the other two systems proposed in [4], [8]. Table 2 illustrates the quantitative analysis of these three methods.

Table 2. The descriptive statistics (syst 1 [ref. 4], syst 2 [ref. 8])

Method	*Accuracy*	*Specificity*	*Sensitivity*
Syst 1	90.95 %	92.50 %	88.89 %
Syst 2	85.60%	79.50%	97.60%
Our syst	94.12%	95.45%	93.10%

From Table 2, we can see that our method achieves the highest overall classification accuracy, specificity, and sensitivity (94.12%, 95.45% and 93.10%). It is obvious that the proposed method has better performance than that of [4] and [8]. Generally speaking, compared with the traditional CAD systems, our method further improves the accuracy of breast tumor classification.

4 Discussions and Conclusions

In this paper, an effective classification system for breast tumor is proposed. Firstly, tumors are characterized according to the BI-RADS lexicon for breast ultrasound and each feature is graded from 1 to 4. Then, a biclustering learning algorithm is applied to find meaningful local-coherent patterns. Detected patterns with support information, namely supervised biclusters, are used as classification basis and grouped into benign, malignant and invalid sets. To the best of our knowledge, this is the first attempt to take the advantage of both biclustering algorithm and supervised learning for CAD system of breast cancer diagnosis. Finally, the k-NN classifier is adopted to classify breast tumor as benign or malignant based on corresponding feature vector and supervised biclusters. Syst. 1 extracts the solidity morphologic features from B-Mode ultrasound images and then a SVM classifier is employed to classify the tumor as benign or malignant. In Syst. 2, textural features, morphologic features and color Doppler features are extracted from the B-Mode and the color Doppler images. Those features are then used to classify benign and malignant tumors. Compared with systems proposed in [4] and [8], a novel CAD method using a new biclustering learning technique for classification of breast tumor as benign or malignant is proposed.

Quantitative experimental results demonstrate that the proposed system significantly improves the accuracy of breast tumor classification. However, the dataset being used in our experiment only has a quite limited number of samples. Accordingly, we will test our method on an expanded dataset with more samples and different tumor types to obtain more reliable results in our future work. With further efforts to improve the proposed scheme, it is expected that this scheme will become more valuable to radiologists and be used as a useful diagnostic aid in real clinical applications.

Acknowledgements. This work is supported by National Natural Science Funds of China (No. 61372007), Natural Science Funds of Guangdong Province (No. S2012010009885), the Fundamental Research Funds for the Central Universities (No. 2014ZG0038), and Projects of innovative science and technology, Department of Education, Guangdong Province (No. 2013KJCX0012).

References

1. Landis, S.H., Murray, T., Bolden, S., Wingo, P.A.: Cancer statistics. CA: A Cancer Journal for Clinicians **48**(1), 6–29 (1998)
2. Chen, D., Chang, R., Wu, W., Moon, W.K., Wu, W.: 3-D breast ultrasound segmentation using active contour model. Ultrasound in Medicine and Biology **29**(7), 1017–1026 (2003)
3. Ding, J., Cheng, H.D., Huang, J., Liu, J., Zhang, Y.: Breast Ultrasound Image Classification Based on Multiple-Instance Learning. Journal of Digital Imaging **25**(5), 620–627 (2012)
4. Chang, R., Wu, W., Moon, W.K., Chen, D.: Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors. Breast Cancer Research and Treatment **89**(2), 179–185 (2005)
5. Garra1, B.S., Krasner, B.H., Horii, S.C., Ascher, S., Mun, S.K., Zeman, R.K.: Improving the Distinction between Benign and Malignant Breast Lesions: The Value of Sonographic Texture Analysis. Ultrasonic Imaging **15**(4), 267–285 (2002)
6. Kuo, W., Chen, D.: Classification of benign and malignant breast tumors using neural networks and three-dimensional power Doppler ultrasound. Ultrasound in Obstetrics & Gynecology **32**(1), 97–102 (2008)
7. Diao, X., Wang, T., Yang, Y., Chen, S.: Computer-aided diagnosis of breast tumor based on B-mode ultrasound and color Doppler flow imaging. In: Proceeding of BMEI 2009 Conference, Tianjin, pp. 1–5 (October 2009)
8. Liu, Y., Cheng, H.D., Huang, J., Zhang, Y., Tang, X., Tian, J., Wang, Y.: Computer Aided Diagnosis System for Breast Cancer Based on Color Doppler Flow Imaging. Journal of Medical Systems **36**(6), 3975–3982 (2012)
9. Chen, D., Chang, R., Kuo, W., Chen, M.: Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks. Ultrasound in Medicine & Biology **28**(10), 1301–1310 (2002)
10. Jales, R.M., Sarian, L.O., Torresan, R., Marussi, E.F., Alvares, B.R., Derchain, S.: Simple rules for ultrasonographic subcategorization of BI-RADS (R)-US 4 breast masses. European Journal of Radiology **82**(8), 1231–1235 (2013)
11. Park, C.S., Lee, J.H., Yim, H.W., Kang, B.J., Kim, H.S., Jung, J.I., Jung, N.Y., Kim, S.H.: Observer agreement using the ACR breast Imaging reporting and data system (BI-RADS)-Ultrasound, first edition (2003). Korean Journal of Radiology **8**(5), 397–402 (2007)
12. Mendelson, E.B., Berg, W.A., Merritt, C.R.B.: Toward a standardized breast ultrasound lexicon, BI-RADS: ultrasound. Seminars in Roentgenology **36**(3), 217–225 (2001)
13. Levy, L., Suissa, M., Chiche, J.F., Teman, G., Martin, B.: BIRADS ultrasonography. European Journal of Radiology **61**(2), 202–211 (2007)
14. Hong, A.S., Rosen, E.L., Soo, M.S., Baker, J.A.: BI-RADS for sonography: positive and negative predictive values of sonographic features. American Journal of Roentgenology **184**(4), 1260–1265 (2005)
15. Thomas, A., Thickman, D., Rapp, C.L., Dennis, M.A., Parker, S.H., Sisney, G.A.: Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology **196**(123) (1995)

16. Rahbar, G., Sie, A.C., Hansen, G.C., Prince, J.S., Melany, M.L., Reynolds, H.E., Jackson, V.P., Sayre, J.W., Bassett, L.W.: Benign versus malignant solid breast masses: US differentiation. Radiology **213**(3), 889–894 (1999)
17. Heinig, J., Witteler, R., Schmitz, R., Kiesel, L., Steinhard, J.: Accuracy of classification of breast ultrasound findings based on criteria used for BI-RADS. Ultrasound in Obstetrics & Gynecology **32**(4), 573–578 (2008)
18. Mainiero, M.B., Goldkamp, A., Lazarus, E., Livingston, L., Koelliker, S.L., Schepps, B., Mayo-Smith, W.W.: Characterization of Breast Masses With Sonography Can Biopsy of Some Solid Masses Be Deferred? Journal of Ultrasound in Medicine **24**(2), 161–167 (2005)
19. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1), 4–37 (2000)
20. Cheng, Y.Z., Church, G.M.: Biclustering of expression data. In: Proceedings of ISMB 2000 Conference, pp. 93–103 (August 2000)
21. Huang, Q.H., Tao, D.C., Li, X.L., Jin, L.W., Wei, G.: Exploiting Local Coherent Patterns For Unsupervised Feature Ranking. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **41**(6), 1471–1482 (2011)
22. Huang, Q.H.: Discovery of time-inconsecutive co-movement patterns of foreign currencies using an evolutionary biclustering method. Applied Mathematics and Computation **218**(8), 4353–4363 (2011)
23. Huang, Q.H., Tao, D.C., Li, X.L., Liew, A.W.C.: Parallelized evolutionary learning for detection of biclusters in gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics **9**(2), 560–570 (2012)
24. Wu, X.D., Kumar, V., Quinlan, J.R., et al.: Top 10 algorithms in data mining. Knowledge and Information Systems **14**(1), 1–37 (2008)