PERGAMON

Pattern Recognition 32 (1999) 891–903

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

Classification of microcalcifications in digital mammograms using trend-oriented radial basis function neural network

Osamu Tsujii*, Matthew T. Freedman, Seong K. Mun

Department of Radiology, Georgetown University Medical Center, 2115 Wisconsin Ave. NW, # 603 Washington, DC 20007, USA

Received 4 March 1997; in revised form 10 June 1998

Abstract

We proposed some novel classification features for the microcalcification of mammograms, and selected the effective combined features using Karhunen–Loeve (KL) transformation followed by the restricted Euclidean distance measure, and finally applied the proposed trend-oriented radial basis function neural network (TRBF-NN) to distinguish the benign group from the malignant group and evaluate the performance with the round-robin method. The two-dimensional KL features were more distinguishable than the raw two-dimensional features. The TRBF-NN was able to define the more generalized distribution than those distributions defined by the conventional RBF-NNs. According to the receiver operating characteristic analysis, the proposed system performed better than two trained radiologists.

Keywords: Mammograms; Microcalcification; Classification; Feature selection; Karhunen–Loeve transformation; Euclidean distance measure; Neural network; Radial basis function; Round-robin method; Receiver operating characteristic

1. Introduction

In the United States, breast cancer is the leading cause of death in women between 40 and 55 years of age. At present the mammogram is the only proven method for detecting minimal breast cancer. One important indicator of breast cancer is the presence of clustered microcalcifications. Clustered microcalcifications can be seen on mammograms in 30–50% of cases of breast cancer. However, most mammographic calcifications are benign. Accurate classification of microcalcifications into benign and malignant groups would help improve diagnostic sensitivity as well as reduce the number of unnecessary biopsies.

What features are useful to distinguish benign from malignant calcifications? Various investigators have attempted this distinction [1–6]. Roselli-Del-Turco used three features to distinguish benign from malignant calcifications [3]: (1) size, shape or density of the calcifications, (2) size or shape of the "cluster", and (3) number of microcalcifications. Their analysis of the morphological criteria, which led to a distinction between benign and malignant biopsy results, is one of the most thorough that has been published to date. However, there seems to be no conclusive features which could distinguish benign from malignant calcifications. Wu used a convolution neural network to classify benign and malignant microcalcifications in radiographs of pathologic specimens. [7]. Thiele analyzed 21 texture features (i.e., 16 co-occurrence and five fractal) of the breast tissue surrounding microcalcifications on digitally acquired images during stereotactics biopsy. His Jackknife results misclassified 2 of 18 malignant cases (sensitivity 89%) and 6 of 36 benign cases (specificity 83%) for logistic discriminant

* Corresponding author.

0031-3203/99/$ – See front matter
PII: S0031-3203(98)00099-5

analysis [8]. Jiang extracted eight characteristic features of clustered microcalcifications. Those features described the size, shape irregularity, number and uniformity of individual microcalcifications, and the size and shape of a cluster. When individual microcalcifications were identified by a computer rather than manually, the classification performance of his technique remained comparable to that of the radiologists [9]. However, the above three studies did not quantitatively measure how each feature contributes to classification.

Most classification systems consist of four subsystems: measure, preprocess and/or transformation, feature selection, and classification. In the design of a classification system, a common assumption is that all input features play an important discriminatory role in the classification and are essential for a specified performance. However, this may not always be true in practical applications. If the designer does not have confidence in what the effective features are, some features may be redundant or not as important as others. Our purpose in this paper is to investigate the effectiveness of our feature selection method and the proposed neural network as detailed below:

(1) *Image feature selection method for classification of clustered microcalcifications*: It is not easy to estimate appropriate image features on this classification problem. First, we propose 10 image feature candidates. Second, some redundancy of image features is eliminated through Karhunen–Loeve (KL) transformation. Finally, we determine the most effective KL feature (i.e., eigen vector) plain to classify benign and malignant groups through the restricted Euclidean distance measure (rEDM).
(2) *Trend-oriented radial basis function neural network* (*TRBF-NN*): Since two data distributions of benign and malignant groups were partly overlapped on the selected KL feature plain, a more powerful neural network to regularize the class distribution was desired. We propose the novel cost function which would be minimized through the network training, where the learning equations of the centers and widths of each radial basis function (RBF) are based on the gradient-descent method [10].

2. Materials and method

Our database consists of the Mammographic Image Analysis Society (MIAS) MiniMammographic Database [11], which includes 9 benign and 13 malignant calcification cases, and the Georgetown University Hospital Database, which includes 17 benign and 16 malignant calcification cases. Forty-seven benign and 81 malignant region of interest (ROI) images, a total of 128 ROIs, were selected from $50\,\mu m \times 50\,\mu m$ digitized whole mammograms manually. Each 256×256 pixel ROI image is supposed to contain whole clustered microcalcifications. If calcifications were widely distributed beyond an ROI boundary, additional ROIs were selected from an image while avoiding more than 50% overlap between ROIs. The overview of our proposed method is shown in Fig. 1. First, we extract ten raw image features which are calculated from an original ROI image, the binarized microcalcification image, automatically made in preprocess, and two processed images based on the binarized image. These features are based on three morphological criteria: (1) number, size, and shape of the calcifications, (2) size and shape of the "cluster", and (3) contrast of microcalcifications. Second, we apply KL transformation to ten-dimensional raw feature hyper-space in order to reduce the dimension of the problem. Next, we select the best two-dimensional KL feature plain, separating benign from malignant calcifications, from ten-dimensional KL feature hyper-space using the rEDM. In addition to the KL feature plain, the best plain for raw feature hyper-space is selected for comparison using the rEDM. Finally, we classify them based on the two-dimensional plain using the proposed TRBF-NN and evaluate the performance with the round-robin method, where one sample is tested after the learning based on the remaining 127 samples.

2.1. Preprocessing

The purpose of the preprocessing is to get a binarized image of clustered microcalcifications. Fig. 2a and b is an example of a original gray-scale image and a binarized microcalcification image, respectively. The preprocessing algorithm is summarized as follows:

1. *Subtraction of the averaged image from the original image*: The averaged image is made through applying 23×23 pixels average kernel to the original image. Then the averaged image is subtracted from the original image to eliminate the background trend.
2. *Binarization through the histogram quantization*: The average-subtracted image (i.e., the outcome of step 1) is quantized to 32 levels based on its histogram. The pixels in the maximum level are only used as candidates of clustered microcalcifications.
3. *Opening of the histogram quantized image* (*i.e., the outcome of step* 2): The 3×3 pixels morphological opening filter is applied to the histogram quantized image to remove line artifacts which arise when microcalcifications are in the ductal structures.
4. *Dilation of the opened image* (*i.e., the outcome of step* 3): The 3×3 pixels morphological dilation filter is applied to the opened image to enlarge objects.
5. *AND operation*: The AND operation between the opened image and the dilated image (i.e., the outcome of step 4) is processed to preserve the detail shape of microcalcifications.

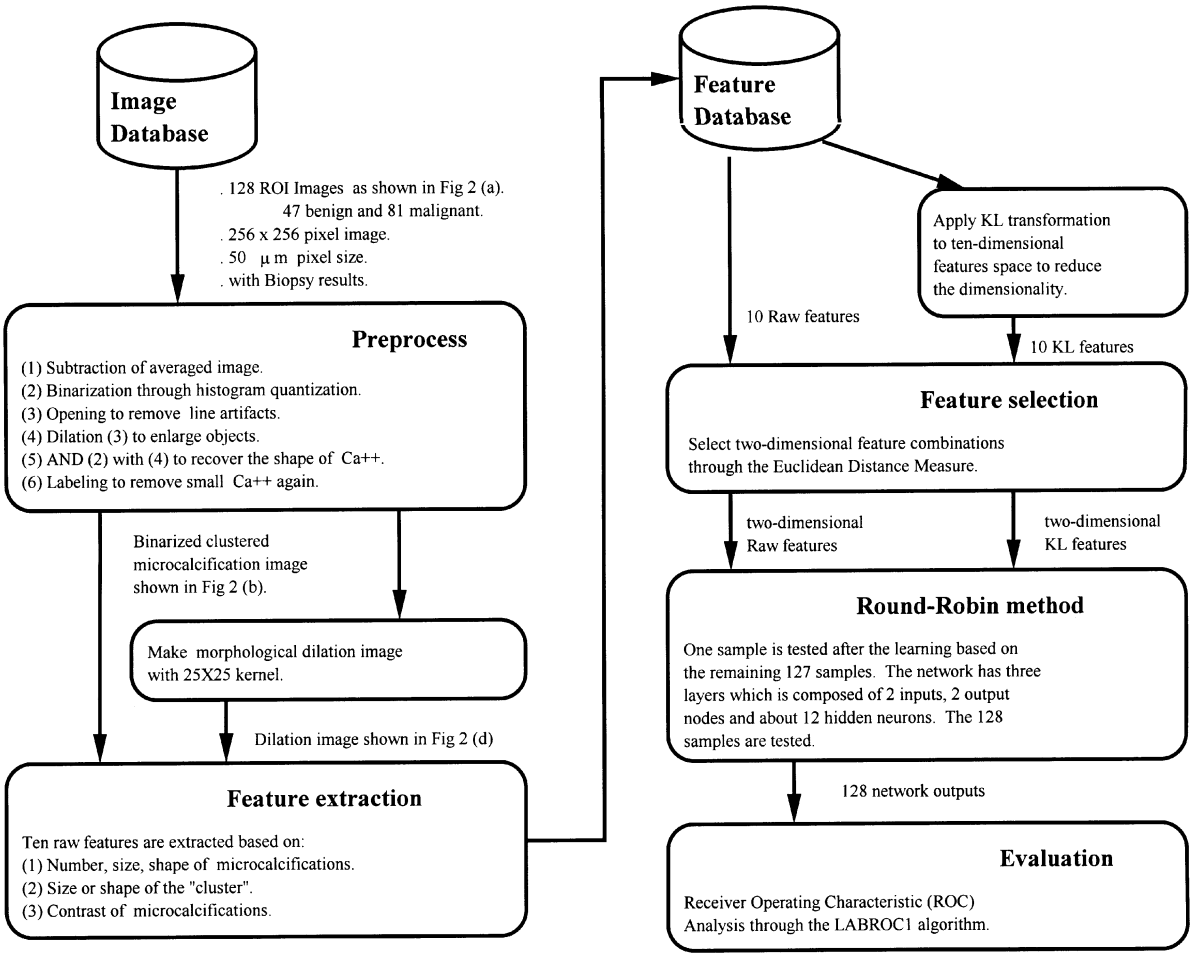

Fig. 1. The overview of the entire method. Ca^{++} represents microcalcifications.

6. *Labeling*: The labeling for the AND image (i.e., the outcome of step 5) is processed to eliminate small objects less than 5 pixels, which are regarded as noise.

All parameters used in the preprocessing are determined on the basis of the experiments.

2.2. Feature extraction

Our image features are based on four images, which are the original image shown in Fig. 2a, the binarized clustered microcalcifications image shown in Fig. 2b, the ellipse image, which fits to the distribution of microcalcifications, shown in Fig. 2c, and the morphological dilation image of the binarized image shown in Fig. 2d. Ten raw features $u_k, k = 0, \ldots, 9$, based on four factors (i.e., number, size, shape, and subtlety) are listed as follows.

u_0: Number N, where N is the number of microcalcifications shown in Fig. 2b.

u_1: Distribution1 N/DA, where DA is the distribution area of microcalcifications shown in Fig. 2c.

u_2: Shape1 AE, where AE is the average circularity of microcalcifications. The higher the circularity, the rounder the microcalcification.

u_3: Shape2 WE, where WE is the weighted average circularity of microcalcifications.

u_4: Distribution2 CA/DA, where CA is the total area of microcalcifications shown in Fig. 2b.

u_5: Size1 BA, where BA is the area of the biggest microcalcification.

u_6: Size2 AA, where AA is the average area of microcalcifications.

u_7: Distribution3 Ax/Bx, where Ax and Bx are the length of the semimajor axis and semiminor axis, respectively, of the ellipse which best fits to the disposition of microcalcifications. The higher the Distribution3, the rounder the cluster.

u_8: Contrast1 C/M, where C and M are the contrast and moment of the original image masked by the 25×25

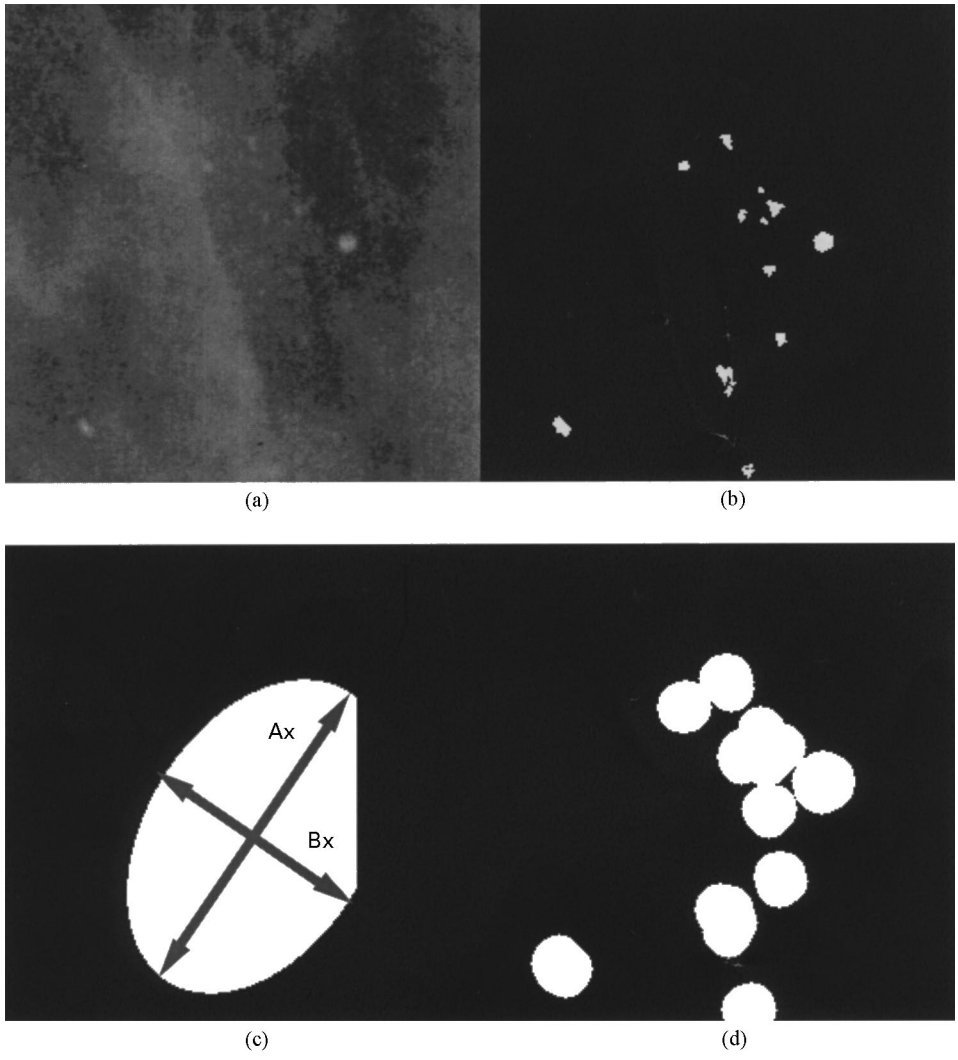

Fig. 2. (a) an original clustered microcalcifications image, (b) a binarized clustered microcalcifications image representing the area and shape of each microcalcification, (c) an ellipse image assumed to represent the area and shape of the microcacification cluster, (d) a morphological dilation image used as the background when calculating the contrast features.

kernel dilated microcalcification image shown in Fig. 2d. The kernel size is determined by some heuristics. The higher the contrast, the more clearly the microcalcifications can be seen.

u_9: Contrast2 $D1/D2$, where $D1$ and $D2$ are the average intensity of microcalcifications and the original image masked by the 25×25 kernel dilated microcalcification image, respectively.

2.3. Feature selection

The reduction of the dimensionality of a problem in order to deliver a system for a real-world application has always been a main concern of researchers. In this study, we reduce the dimensionality from 10 to 2, because a two-dimensional plain is the largest dimension in which a human can recognize the separation of classes without changing the direction of the view. There are two basic approaches to reducing the dimensionality: (1) transformation technique such as Fourier transformation or KL transformation, and (2) selection of a subset of features by evaluating the features based on the available data, e.g., Euclidean distance measure (EDM). A combinational feature evaluation measures the separability of a subset of features which takes into account the combining effect of a set of features. In this study, we apply KL transformation followed by the restricted EDM. Before KL transformation, the normalization of feature values must be processed, because KL features are calculated on the basis of the variance of feature distributions.

The magnitude variation among raw features should be removed. The normalized feature values $x'_k(n)$, $k = 0, \ldots, 9$, for the original feature value $x_k(n)$ are given as

$$x'_k(n) = \frac{x_k(n) - \mu_k}{\sigma_k}, \tag{1}$$

where k is a feature index, and μ_k and σ_k are the mean and the standard deviation, respectively, of the original feature value sequence $x_k(n)$, $n = 0, \ldots, 127$.

KL transformation extracts the lengths of the data distribution in the multi-dimensional feature space using eigensystems. The eigenvalues λ_k, $k = 0, \ldots, 9$, and the eigenvectors $\mathbf{F}_k$ for the covariance matrix $\mathbf{R}$ of the 10 normalized original features are defined as all solutions of

$$\begin{aligned} &|\mathbf{R} - \lambda_k \mathbf{I}| = 0, \\ &\mathbf{R}\mathbf{F}_k = \lambda_k \mathbf{F}_k, \\ &\mathbf{F}_k \neq \mathbf{0}, \end{aligned} \tag{2}$$

where $\mathbf{I}$ is an identity matrix. In this study the eigenvectors $\mathbf{F}_k$, where k is a descending order of the eigenvalues, are called as KL features vk, and KL feature values $v_k(n)$, $k = 0, \ldots, 9$, are defined as inner products between the eigenvector $\mathbf{F}_k$ and the normalized original features $\mathbf{u}'(n) = \{u'_k(n)\}$ [12]. The KL features which have high eigenvalues usually consist of effective raw features, because in general the wider the distribution, the more separability.

Table 1 shows the eigenvalues and the eigenvectors for the covariance matrix of the 10 original features for our database. Using this table we can find the combination ratios of the normalized 10 original features to compose each KL feature. In other words, this table defines the KL features (i.e., vectors) in the ten-dimensional hyperspace composed by the normalized ten original features. According to the empirical theory, the direction having the bigger variances of the data distribution are more effective to discriminate the classes than those of the smaller variances. In that sense the KL features v_0 and v_1, which have the biggest eigenvalues, are regarded as the more effective features for discrimination. However, it is not always true that the combination of the two KL features with the biggest eigenvalues are the best for the classification in the two-dimensional plain. We quantize the separability of the data, projected from the normalized original ten-dimensional space to the KL feature plain, using the restricted EDM. The EDM of two features is defined by

$$\mathrm{EDM}(x_1, x_2) = \frac{p(w_k)p(w_l)}{N(w_k)N(w_l)} \times \sum_{p=1}^{N(w_k)} \sum_{q=1}^{N(w_l)} [\mathbf{x}_p^{(k)} - \mathbf{x}_q^{(l)}]^{q'} [\mathbf{x}_p^{(k)} - \mathbf{x}_q^{(l)}], \tag{3}$$

where x_1, x_2 are two features, $p(w_k)$ and $p(w_l)$ are prior probabilities of occurrence of class w_k and w_l, respectively, $N(w_k)$ is the number of patterns in class w_k, and x_p^k are two element vectors from class w_k [13]. EDM is used to evaluate the separability in the feature space, where the distance between every pair of two-element patterns from a different class is accumulated. However, EDM sometimes presents inappropriate results when some data points are far from the major mass (e.g., inside the standard deviation) of the data. Because such data make EDM values higher than the values expected from the appearance of the data distribution, we introduce the restricted EDM which measures the accumulated distance with the restriction of the data space. The rEDM is defined by

$$\mathrm{rEDM}(x_1, x_2, \alpha) = \frac{p'(w_k)p'(w_l)}{N'(w_k)N'(w_l)} \times \sum_{p=1}^{N'(w_k)} \sum_{q=1}^{N'(w_l)} [\mathbf{x}'^{(k)}_p - \mathbf{x}'^{(l)}_q]^{q'} [\mathbf{x}'^{(k)}_p - \mathbf{x}'^{(l)}_q], \quad |x_1| \leqslant \alpha, |x_2| \leqslant \alpha, \alpha > 0, \tag{4}$$

Table 1
Eigenvalues and eigenvectors of the covariance matrix for the original 10 features

	v_9	v_8	v_7	v_6	v_5	v_4	v_3	v_2	v_1	v_0
Eigenvalues:										
	0.1818	0.2941	0.3855	0.4734	0.0352	1.1460	1.0335	1.6123	2.1326	2.7055
Eigenvectors:										
u'_0	− 0.2305	− 0.2179	0.1436	0.6389	− 0.0766	0.0469	− 0.3457	0.5030	− 0.0651	− 0.2928
u'_1	0.2748	0.0554	0.1728	0.0523	− 0.6093	0.1816	− 0.4310	− 0.0876	0.3503	0.4103
u'_2	− 0.2756	0.5761	0.3346	0.0653	− 0.0444	0.1694	− 0.2559	− 0.4417	− 0.3535	− 0.2445
u'_3	0.4214	− 0.5330	− 0.1638	0.0977	0.0636	0.1545	− 0.2693	− 0.4347	− 0.4439	− 0.1245
u'_4	− 0.2120	− 0.1433	0.2777	0.0615	0.6634	0.1659	− 0.2677	− 0.0846	0.1287	0.5391
u'_5	0.5299	0.2270	0.4139	0.2186	0.0628	− 0.2135	0.2530	0.2607	− 0.4292	0.2879
u'_6	− 0.5358	− 0.3453	0.0968	− 0.0834	− 0.4110	− 0.1099	0.2110	− 0.0489	− 0.4370	0.3981
u'_7	− 0.0638	0.1036	− 0.2670	0.5992	0.0089	− 0.5420	0.0904	− 0.4384	0.1881	0.1587
u'_8	− 0.0040	0.1061	− 0.2264	0.3918	− 0.0339	0.7302	0.4762	− 0.0510	0.0091	0.1418
u'_9	− 0.0310	0.3467	− 0.6560	− 0.0649	0.0522	− 0.0062	− 0.3762	0.2879	− 0.3520	0.3054

where α is a positive value to restrict the data space, and all dashed-variables are modified from Eq. (3) to be within the restricted space. Using the rEDM the contributions from the data, whose absolute values are outside of the coefficient α, are eliminated when measuring the accumulated distance.

We apply the rEDM to both the raw feature space and the KL feature space in order to select the best separable combination (i.e., plain) for each feature space. Before applying the rEDM to KL features, the magnitude variance of KL features is also removed following Eq. (1), otherwise the KL features, which have higher eigenvalues, are always selected as the best features. The normalized KL features are represented as v'_k, $k = 0, \ldots, 9$. The rEDM of raw features are evaluated for comparison with that of KL features. The number of combinations is 45 (i.e., a combination of two from ten). Fig. 3a and b show the rEDM values of 45 two-dimensional combinations for both raw features and KL features, respectively. Those values are the weighted averages of the six rEDMs, where α is varied from 1.0 to 1.5 with 0.1 step. The rEDM values generally have positive correlation with the α values, so the weighted averages are used. The most separable combinations, which indicate the maximum rEDM values, for raw features and KL features are the combination between the raw features u'_0 and u'_3, and the combination between KL features v'_1 and v'_2. The projections of the data from the normalized original ten-dimensional space to both the raw feature plain and the KL feature plain determined in Fig. 3 are shown in Fig. 4a and b, respectively. The data distribution of the KL feature plain is normalized using Eq. (1) after the inner production process. As seen in Fig. 4a, two classes are separated by the line with a 135° slope. It means that there is a positive correlation between two raw features u'_0 and u'_3, i.e., Number and Shape2. The more microcalcifications, the smaller the weighted average circularity, the more likely the microcalcifications are malignant. As seen in Fig. 4b, two classes are separated by the line with a 45° slope, meaning that there is a negative correlation between the KL features v'_1 and v'_2. When there is a negative correlation between two KL features, the bigger the difference between raw component features of two KL features, the more effectively the raw component feature can distinguish two classes. As seen in Table 1, the differences of u'_5, u'_9, and u'_7 between two KL features v'_1 and v'_2 are larger than those of other raw component features, which means that the larger the maximum area, the higher the contrast, and the smaller the ellipsity of the clustered distribution, the more likely the microcalcifications are malignant.

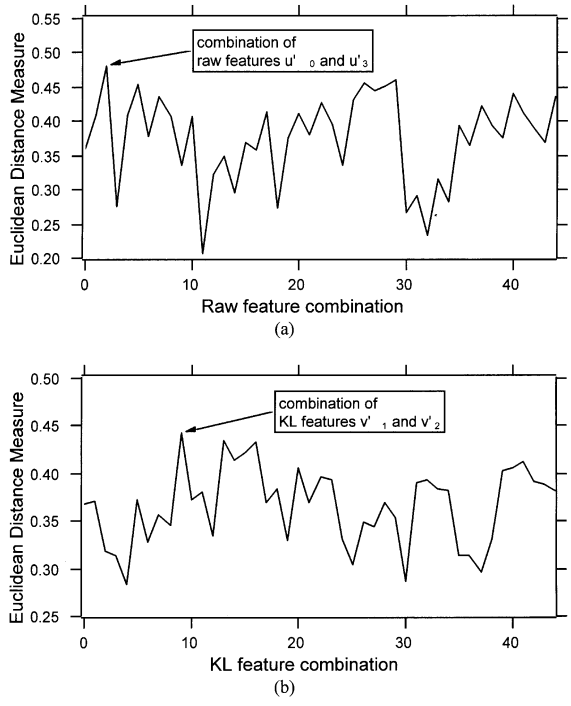

Fig. 3. The averaged rEDMs of 45 two-dimensional combinations for (a) ten raw features and (b) ten KL features.

2.4. Review of RBF neural network

In this section, we review the central features of the RBF-NN [14]. Assume that the neural network has a three layer feedforward architecture as shown in Fig. 5. Input vectors $\mathbf{x}$ are propagated to the hidden units (i.e., RBF neurons), which computes a hyper-spherical function of $\mathbf{x}$, so that the output of the jth hidden unit is given by

$$\phi_j = \phi(\|\mathbf{x} - \mathbf{y}_j\|), \tag{5}$$

where $\mathbf{y}_j$ is the center of the RBF neuron for the jth hidden unit, and $\| \ldots \|$ denotes a distance measure that is generally taken to be the Euclidean norm. The nonlinear function ϕ can be chosen in variety of ways and can, in principle, vary from one hidden unit to the next. For example, we use a Gaussian nonlinearity:

$$\phi(x) = \exp\left(-\frac{x^2}{\sigma^2}\right). \tag{6}$$

The outputs z_i of the neural network are given by the weighted sums of the outputs from the hidden units:

$$z_i = \sum_j w_{ij}\phi_j, \tag{7}$$

where the synoptic weights w_{ij} are adaptive variables that are set during the learning phase. Training data are supplied to the neural network in the form of pairs $(\mathbf{x}_p, t_p)$ of input and target vectors, where $p = 1, \ldots, P$

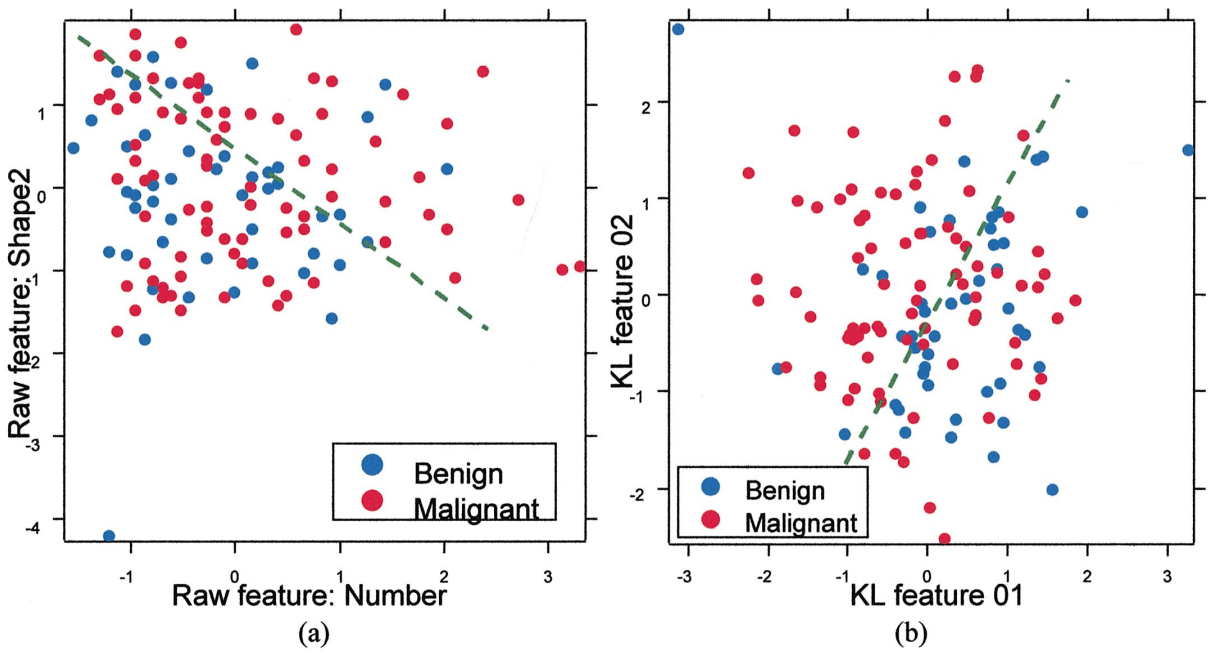

Fig. 4. The data distributions of (a) the raw features and , and (b) the KL features and . Blue dots and red dots represent benign and malignant cases, respectively. As seen in Fig. 4a, when two classes are separated by the line with a 135° slope, there is a positive correlation between two axes, i.e., two raw features, to separate two classes. As seen in Fig. 4b, when two classes are separated by the line with a 45° slope, there is a negative correlation between two axes, i.e., two KL features.

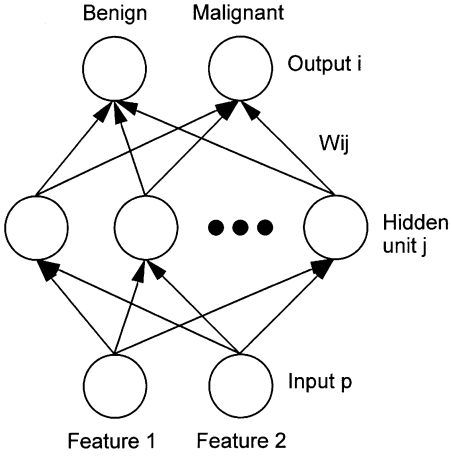

Fig. 5. Architecture of a three layer feedforward neural network used in this study.

labels the individual training pairs. The learning algorithm aims to minimize the sum-of-squares error defined by

$$E = \frac{1}{2}\sum_p \sum_i (z_{ip} - t_{ip})^2, \tag{8}$$

where $z_{ip} = z_i(\mathbf{x}_p)$ denotes the output of the ith output unit when the neural network is presented with the input vector $\mathbf{x}_p$. At a minimum of E we have

$$\frac{\partial E}{\partial w_{ij}} = 0. \tag{9}$$

It is unlikely that the widely used technique of error back-propagation [15], the learning algorithm for the RBF-NN corresponds to the solution of a linear problem. Therefore, the training of the network is a fast procedure as follows:

$$w_{ij} = \sum (\mathbf{M}^{-1})_{kj} \left\{ \sum_p \phi_{kp} t_{jp} \right\}, \tag{10}$$

where the matrix $\mathbf{M}$, which is the covariance matrix of the transformed data, is defined by

$$\mathbf{M}_{kj} = \sum_p \phi_{kp} \phi_{jp}, \tag{11}$$

where $\phi_{kp} = \phi_k(\mathbf{x}_p)$ and $\phi_{jp} = \phi_j(\mathbf{x}_p)$.

2.5. Modifications of RBF-NN

An important consideration in setting up an RBF-NN is the choice of the number, center $\mathbf{y}_j$ and width σ_j of the

RBF neuron. The most natural choice is to let each data sample point in the training set correspond to an RBF center. In this case the number of degrees of freedom in the network equals the number of items of data, and the neural network function fits exactly through each data point. If the data appear regular, but are contaminated by noise, the neural network will learn all the details of the individual data points, rather than representing the underlying trends in the data. This phenomenon is sometimes called overfitting. There are four main ways to avoid overfitting. The first method, regularization, damps out the rapid sensitivity in a full size model (e.g., the full set of RBFs, whose centers correspond to the input vectors from the training data, is retained) by adding a weight penalty term to the minimization criterion [16–18]. For example, minimization of the energy E_s

$$E_S = \frac{1}{2}\sum\sum (Z_{ip} - t_{ip})^2 + \lambda \sum_i \sum_j \omega_{ij}^2 \tag{12}$$

is zero-order regularization [19]. The regularization parameter, λ, has to be chosen a priori or estimated from the data, and controls the degree to which the neural network function is smoothed. The second way to avoid overfitting is to explicitly limit the complexity of the network (i.e., to reduce the number of degrees of freedom) by allowing only a subset of the possible centers to participate. This method has the added advantage of producing small networks. Broomhead and Lowe suggested choosing such a subset randomly from the training inputs [14]. Chen used forward selection to choose the centers of the hidden units to produce small neural networks [20]. Orr has done extensive studies on the examination of regularized forward selection [21]. The third way is to choose the centers of the RBF neurons using a k-means algorithm or a self-organization algorithm such as the topology preserving feature map [22]. The fourth way to avoid overfitting is to find the approximated solution of centers and widths while reducing the number of degrees of freedom for the neural network. At a minimization of the energy function E_s, given by Eq. (12), we have

$$\frac{\partial E_s}{\partial \omega_{ij}} = 0,$$

$$\frac{\partial E_s}{\partial \mathbf{y}_j} = 0,$$

$$\frac{\partial E_s}{\partial \sigma_j} = 0. \tag{13}$$

The weights ω_{ij} are solved, using the same technique as Eq. (10), and are given by

$$w_{ij} = \sum_k (\mathbf{M} + \lambda I_n)_{kj}^{-1} \left\{ \sum_p \phi_{kp} t_{jp} \right\}, \tag{14}$$

where I_n is the $n \times n$ identity matrix and n is the number of RBF neurons. Gradient-descent is probably the simplest approach for attempting to find the solution of centers and widths, though, of course, it is not guaranteed to converge [23]. In the gradient-descent method the values of $\mathbf{y}_j$ and σ_j which minimize E_s are given by

$$\Delta \mathbf{y}_j = -\beta \frac{\partial E_s}{\partial \mathbf{y}_j},$$

$$\Delta \sigma_j = -\beta \frac{\partial E_s}{\partial \sigma_j}, \tag{15}$$

where β is a learning parameter which is related to the rate of convergence.

2.6. The proposed TRBF-NN

The proposed TRBF-NN also finds the approximated solution of centers and widths while reducing the number of degrees of freedom. We introduce a novel cost function E_t given by

$$E_t = \frac{1}{2}\sum_p \sum_i (Z_{ip} - t_{ip})^2 + \lambda \sum_j \frac{1}{\sigma_j^2}, \tag{16}$$

where the regularization parameter λ can be a positive number by some heuristics. By adding an inverse σ^2 penalty term to the minimization criterion, the regularization should be more accelerated than the conventional method given by Eq. (12). This cost function is designed so that the greater the widths of RBF neurons, the less the training error. The weights ω_{ij} are solved using Eq. (10), and using the gradient-descent method, the delta values of $\mathbf{y}_j$ and σ_j which minimize E_t are given by

$$\Delta \mathbf{y}_j = -\beta \frac{\partial E_t}{\partial \mathbf{y}_j} = -2\beta \sum_p \sum_i \left(\sum_k \omega_{ik}\phi_{kp} - t_{ip} \right) \times \omega_{ij}\phi_{jp} \frac{(\mathbf{x}_p - \mathbf{y}_j)}{\sigma_j^2}, \tag{17}$$

$$\Delta \sigma_j = -\beta \frac{\partial E_t}{\partial \sigma_j} = -2\beta \left(\sum_p \sum_i \left(\sum_k \omega_{ik}\phi_{kp} - t_{ip} \right) \times \omega_{ij}\phi_{jp} \frac{\|\mathbf{x}_p - \mathbf{y}_j\|^2}{\sigma_j^3} - \lambda \frac{1}{\sigma_j^3} \right), \tag{18}$$

where β is a learning parameter which is related to the rate of convergence. We evaluate the regularization performance among the conventional RBF-NNs and the TRBF-NN.

3. Results and discussion

The performances of the KL features and the TRBF-NN were evaluated through the round-robin method, where one sample is tested after the learning based on

the remaining 127 samples. Regarding the round-robin method, we should assume two conditions below:

(1) The resulting raw feature plain and KL feature plain selected based on 127 testing examples are same as those based on the entire 128 examples, respectively. Although both raw feature plain and KL feature plain should be selected for each training set using our proposed method, we approximately used the feature plains based on the entire examples to test each example.
(2) Even if a plural of ROIs are selected from one case, those ROIs are independent. Those ROIs are not independent in reality. However, most features (e.g., the number of microcalcifications, the shape of distribution, and the contrast of clustered microcalcifications) would be different from those of neighboring ROIs, because when extracting a plural of ROIs from one case, those were selected in the manner, where no more than 50% of areas did not overlap.

The LABROC1 algorithm developed by Metz was used to fit the receiver operating characteristic (ROC) curve to the continuous data from 128 outputs of each neural network. All neural networks, evaluated in this section, have two output ports as shown in Fig. 5. Both value ranges of the benign port and the malignant port are $[0,1]$. If both outputs are 0, it means that this image is classified as neither benign nor malignant. Similarly, if both outputs are 1, it means that its data is classified as both benign and malignant. However, the LABROC1 algorithm does not support such a two output model, so the two output values must be converted to be a combined output within $[0,1]$, given by

$$t_f = \begin{cases} 0.5 - 0.5 \times (t_b - t_m), & t_b \geqslant t_m, \\ 0.5 + 0.5 \times (t_m - t_b), & t_b \leqslant t_m, \end{cases} \tag{19}$$

where t_f, t_b, and t_m are the combined output, the output of the benign port, and the output of the malignant port, respectively. If the combined value t_f is 0, it means the microcalcification is classified as benign with 100% certainty by the neural network, and if 1, then the microcalcification is malignant. If t_f is 0.5, it means the neural network presents the undecided answer.

3.1. Comparison of the feature plains using the TRBF-NN

The raw feature plain, composed by u'_0 and u'_3, and the KL feature plain, composed by v'_1 and v'_2, shown in Fig. 4a and b, respectively, were compared using the TRBF-NNs. All image data points were projected from the ten-dimensional feature space to those plains. The neural network outputs for the raw features and the KL features are shown in Fig. 6a and b, respectively. The neural network parameters were determined based on the Az value by means of scanning various combinations of the parameters, i.e., the number of the RBF neurons, the regularization parameter λ, and the initial positions of the neurons. The five patterns of the initial positions were tested. The learning parameter β is empirically fixed as 0.001 in both neural networks. The Az value represents the area under the ROC curve. If the two classes are clearly separated, the Az value of the system is 1.0, and if two classes overlap each other, the Az is almost 0.5. The higher Az value represents the better performance system. The initial width $\sigma_{initial}$ of each neuron was equally given by

$$(2\sigma_{initial})^2 = \frac{16}{\text{neurons}}, \tag{20}$$

where the number 16 ($=4\times4$) is the area in which most of the data are supposed to be. We assume that most of the data are settled in $[-2,2]$, where the feature values are normalized by Eq. (1), and the area controlled by each neuron is define by the square of $2\sigma_{initial}$. The best parameters for the raw feature plain were 12 neurons and $\lambda = 1.4$. On the other hand, those for the KL feature plain were 11 neurons and $\lambda = 3.6$. Both neural networks were trained for 100 epochs, where the differences of the training errors, defined by Eq. (16), were converged in less than 0.01% of the training error. The training errors, given by Eq. (8), of the raw feature plain and the KL feature plain were 55.5 and 43.1, respectively. As seen in Fig. 6a and b, it is much easier to find the separating strategy in the KL feature plain than in the raw feature plain.

In the KL feature plain, the left part divided by the line, not indicated, with a 45° slope is mainly classified as malignant, and the right side is classified as benign. However, the line or curve which distinguishes two classes can be hard to find in the raw feature plain. The Az values of both feature plains shown in Fig. 7 endorse the human evaluation based on Fig. 6a and b. It is helpful to evaluate the system performance with the two-dimensional space. The Az for the KL feature plain is much better than that for the raw feature plain, which indicates that more than two raw features are effective for this classification task, because the combinational plain from 10 raw features gives better performance than every raw plain.

3.2. Comparison among the neural network models

The full size RBF-NN (hereafter FRBF-NN), the regularized RBF-NN defined by Eq. (12) (hereafter RRBF-NN), and the TRBF-NN were evaluated using the KL feature plain shown in Fig. 4b. The FRBF-NN has the same number of neurons as the training data (i.e., 127), and the positions of the RBF neurons are identical to those of the training data. For the FRBF-NN, the width of each neuron, which is the only parameter, was chosen

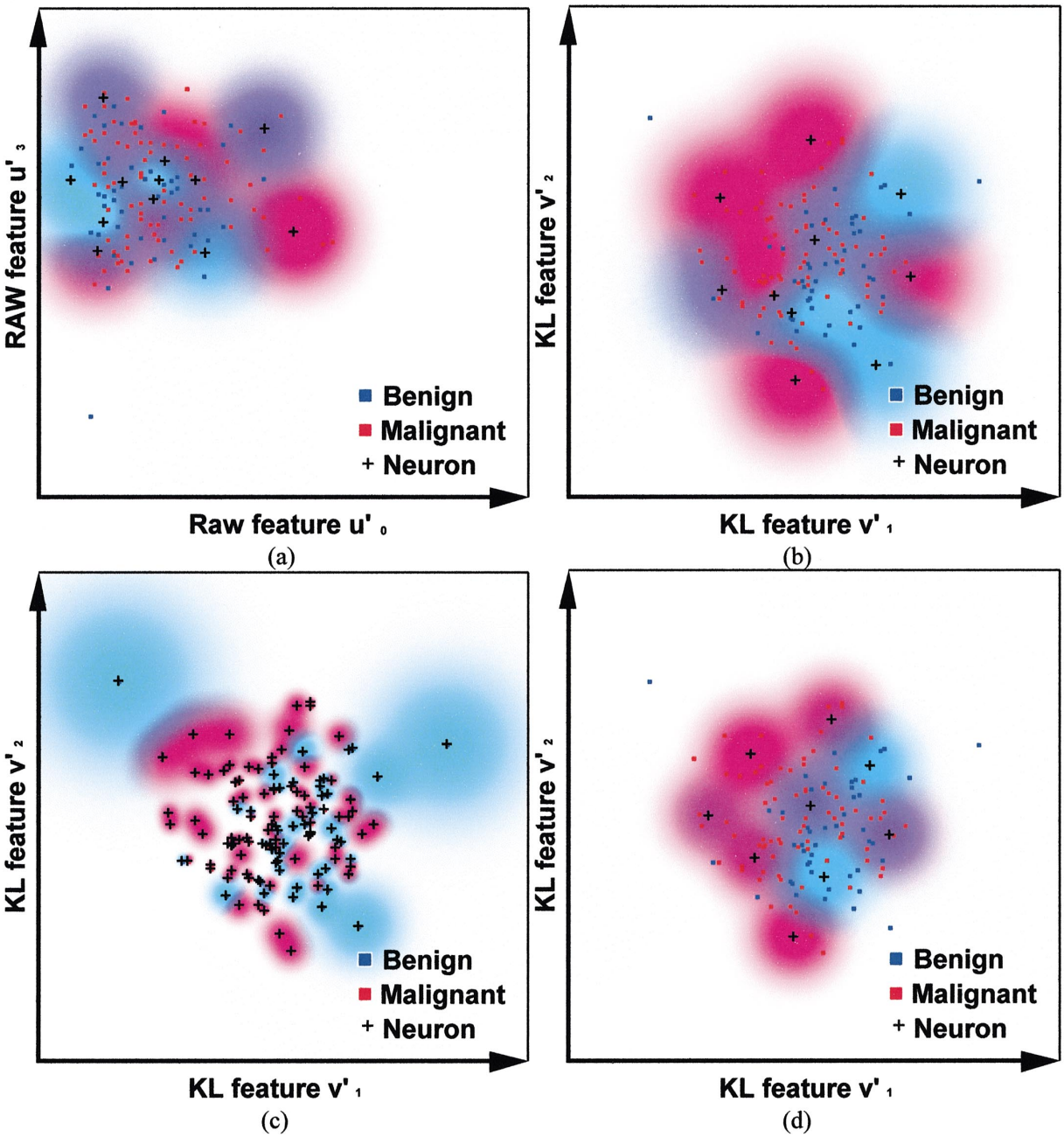

Fig. 6. The training outputs of (a) the TRBF-NN for the raw features v_1 and v_2, (b) the TRBF-NN for the KL features u'_o and u'_3, (c) the FRBF-NN for the KL features v_1 and v_2, and (d) the RRBF-NN for the KL features v_1 and v_2. Pale blue, purple and pink represent the spheres classified as benign, intermediate and malignant, respectively. Each neuron, indicated by the plus sign, has the Gaussian sphere of influence. When an image falls in the pink spheres, it is classified as malignant. If that image is actually malignant, then the neural network works correctly. Similarly, when an image falls in the purple spheres, that is classified as intermediate. White space represents indeterminate.

on the basis of the Az value by scanning the various widths. After scanning, each width was decided to be the distance to the nearest neuron divided by the constant 1.5. The network parameters for the RRBF-NN were selected based on the Az value with the same method as that of the TRBF-NN. The best parameters of the RRBF-NN were nine neurons and $\lambda = 0.2$. The outputs, which gave the best performances, of the FRBF-NN and

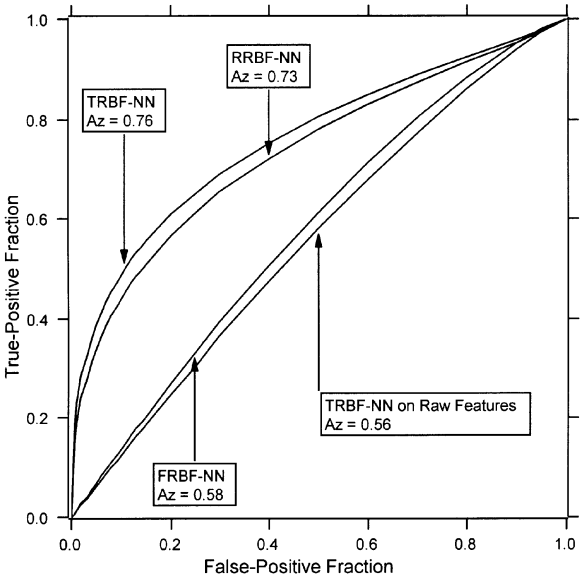

Fig. 7. The ROC analysis of the neural networks shown in Fig. 6 (a)–(d) using the Round-Robin method. Az value represents the area under the ROC curve.

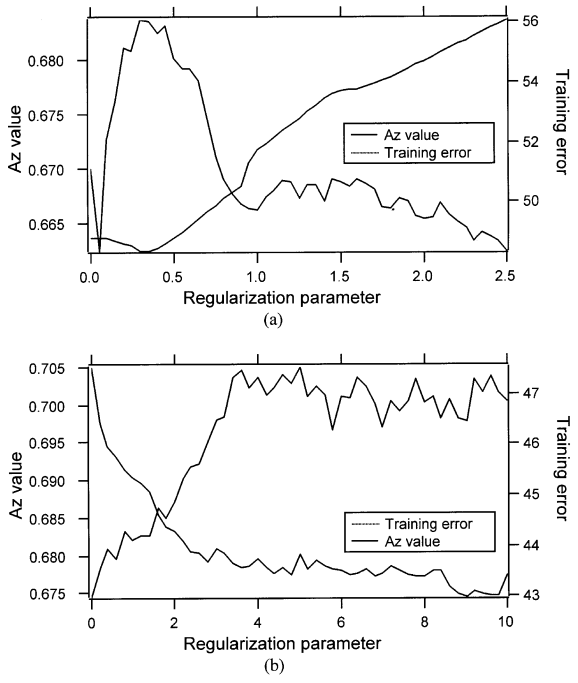

Fig. 8. (a) the Az value and (b) the training error, given by Eq. (8), for the various regularization parameter λs. Note that the value ranges of the regularization parameter are different for the two graphs.

RRBF-NN are shown in Fig. 6c and d, respectively. The RRBF-NN was trained for 100 epochs, and its training error, given by Eq. (8), was 46.0. Those of the TRBF-NN are explained in the previous section. As the outputs of the FRBF-NN are identical to the training data, the training error, given by Eq. (8), was almost zero. But no trend of the distributions could be found in the appearance of the FRBF-NN output. Therefore the Az value of the FRBF-NN is not good (Fig. 7). The different points between two outputs RRBF-NN and TRBF-NN are the appearances of the border areas between the RBF neurons. The outlines of the RBF neurons in the RRBF-NN are much clearer than those of the TRBF-NN. In the RRBF-NN, several data, which fall between the RBF neurons, are still in the white background. It means that those data are classified as neither benign nor malignant. On the other hand, in the TRBF-NN most of the data are classified as benign or malignant. The learning equation Eq. (18) for the TRBF-NN could train the neural network in a more trend-oriented manner than the RRBF-NN given by Eq. (15). The Az values for the FRBF-NN and the RRBF-NN are shown in Fig. 7. The Az value of the TRBF-NN is better than both values of the FRBF-NN and the RRBF-NN.

3.3. Comparison based on the regularization parameter λ

The regularization parameter λ controls the degree of regularization for the RRBF-NN and the TRBF-NN. But it works in different manners for the two NN models as defined in Eqs. (12) and (16). The Az value and the training error, given by Eq. (8), for the various λs are shown in Fig. 8a and b. The number of the neurons for the RRBF-NN and the TRBF-NN are nine and eleven, respectively. Those values are averaged through the five patterns of the initial neuron positions.

In the RRBF-NN, the maximum peak of the Az value and the minimum peak of the training error can be found. As assumed from Eq. (12), if the regularization parameter λ continues to grow while fixing the number of the neurons, the training phase tends to make smaller the weight power given by $\sum \omega_{ij}^2$. And if the weight power alone decreases, then the neurons cannot contribute enough to classify the data.

In other words, the controlling areas of the neurons are shrunk and the borders between the neurons become wider as shown in Fig. 6d. However, if the widths σs increase as the weight power decreases, the RRBF-NN can prevent the area controlled by the neurons from being shrunk. The learning equation, given by Eq. (15), for the widths σs cannot make them big enough to compensate for the decrease of the ω_{ij}, so the training error monotonically increases after the minimum peak. The λ value which maximizes the Az value is almost equal to the λ value which minimizes the training error. It means that as long as the optimal λ value is searched based on the training error, that λ value gives the neural network the optimal Az.

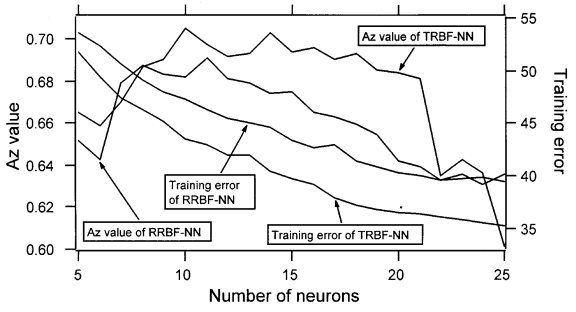

Fig. 9. The Az value and the training error, given by Eq. (8), based on the number of the neurons for the RRBF-NN and the TRBF-NN.

In the TRBF-NN output, the training error saturates for the λ values which are more than four, and the Az value also saturates for the same λ value range even with the deviations. There can be found a tolerant λ value range giving acceptable Az values and training errors. In general, both Az values and training errors of the TRBF-NN are better than those of the RRBF-NN.

3.4. Performances based on the number of the neurons

Fig. 9 shows the Az value and the training error, given by Eq. (8), based on the number of the neurons for the RRBF-NN and the TRBF-NN. Those values were also averaged through the five patterns of the initial neuron positions. While the training error decreases as the number of the neurons grow, the Az values give a broad peak with some deviations. The peak period of the TRBF-NN is wider than that of the RRBF-NN.

4. Conclusion

We proposed some novel features for the classification of microcalcifications on mammograms, and selected the effective combined features using KL transformation followed by the restricted EDM, and finally applied the proposed TRBF-NN to distinguish the benign group from the malignant group. For comparison with two trained radiologists, their Az performances are around 0.5, which means that this database is very difficult for classification, we found that the performance of the proposed system was much better than those of the radiologists. The visualization method using the two-dimensional plain was successful to help the human understand how the system works. For comparison between the RRBF-NN and the TRBF-NN, the TRBF-NN gave a better performance and was as dependent on the regularization parameter in order to get acceptable performances as the RRBF-NN. The key difference between the RRBF-NN and the TRBF-NN is that the TRBF-NN helps the widths grow more than the RRBF-NN, as defined in Eq. (16), so that the TRBF-NN is able to define the more generalized distribution than those defined by the conventional RBF-NNs. Orr reported that half the maximum distance separating pairs of input training points often gives good results [21], we adopted Eq. (20) as the initial width of the neurons in this study. Our initial widths were much smaller than those of his report. Our experiments, not mentioned in the RESULT section, represented that most converged widths were approximately twice the initial widths using the regularization parameter λ as four. This application did not endorse his report. We have already applied the TRBF-NN to the functional approximation, which has not been reported yet. The TRBF-NN also gave a good performance in that application. In the future work, we must investigate the relationship between the regularization parameter λ and the initial width values.

5. Summary

We propose an automated classification method for clustered microcalcifications associated with benign and malignant processes in digital mammograms. Our database consists of 47 benign and 81 malignant region of interest (ROI) images selected from $50\,\mu m \times 50\,\mu m$ digitized whole mammograms manually. First, we extract 10 raw features which are calculated from an original ROI image, the binarized microcalcification image, automatically made in preprocess, and two processed images based on the binarized image. These features are based on three morphological criteria: (1) number, size, and shape of the calcifications, (2) size and shape of the "cluster", and (3) contrast of microcalcifications. Second, we apply Karhunen–Loeve (KL) transformation to ten-dimensional raw feature hyper-space in order to reduce the dimension of the problem. Next, we select the best two-dimensional KL feature plain from 10-dimensional KL feature hyper-space using the restricted Euclidean distance measure. Finally, we classify them based on the two-dimensional plain using the proposed trend-oriented radial basis function neural network (TRBF-NN), and evaluate the receiver operating characteristic (ROC) performance with the round-robin method. The two-dimensional KL features were more distinguishable than the raw two-dimensional features. For comparison with two trained radiologists, their Az, the area under the ROC curve, performances were around 0.5, which means that this database is very difficult for classification. We found that the performance of the proposed system was much better than that of the two radiologists. The visualization method using the two-dimensional plain was successful to help the human understand how the system works. In our comparison of the regularized radial basis function neural network (RRBF-NN) and the TRBF-NN, the TRBF-NN gave a better performance and was

not as dependent on the regularization parameter in order to get acceptable performances as the RRBF-NN. The key difference between the RRBF-NN and the TRBF-NN is that the TRBF-NN helps the widths increase more than the RRBF-NN. The TRBF-NN was also able to define the more generalized distribution than those defined by the conventional RRBF-NN.

Acknowledgements

This work is supported in part by a US Army Grant (DAMD17-94-V-4015). The content of this manuscript does not necessarily reflect the position or the policy of the U.S. government. The authors gratefully acknowledge the constructive discussion of Drs. Akira Hasegawa, Shih-Chung Ben Lo, and Y. Chris Wu, the evaluation support of Dr. Wendelin Hayes and the editorial support of Miss Susan Kirby.

References

[1] G. Svane, E.J. Potchen, A. Sierra, E. Azavedo, Screening mammography, breast cancer diagnosis in asymptomatic women. Mosby-Year Book, St. Louis, Missouri, 1993.

[2] M. Le-Gal, G. Chavanne, D. Pellier, Diagnostic value of clustered microcalcifications discovered by mammography (apropos of 227 cases with histological verification and without a palpable breast tumor), Bull Cancer 71(1) (1984) 57–64.

[3] M. Rossenlli-Del-Turco, S. Ciatto, P. Bravetti, P. Pacini, The significance of mammographic calcifications in early breast cancer detection, Radiol. Med. 72 (1986) 7–12.

[4] S. Ciatto, L. Cataliotti, Y. Distante, Nonpalpable lesions detected with mammography: review of 512 consecutive cases, Radiology 165 (1) (1987) 99–102.

[5] E.A. Sickles, Mammographic features of 300 consecutive nonpalpable breast cancer, Am. J. Roentgenol 146 (4), (1986) 661–663.

[6] P.C. Stommer, J.L. Connolly, J.E. Meyer, J.R. Harris, Clinically occult ductal carcinoma in situ detected with mammography: analysis of 100 cases with radiologic-pathologic correlation, Radiology 172 (1) (1989) 235–241.

[7] Y.C. Wu, M.T. Freedman, A. Hasegawa, R.A. Zuurbier, S.C.B. Lo, S.K Mun, Classification of microcalcifications in radiographs of pathologic speciments for the diagnosis of breast cancer, Acad. Radiol. 2 (1995) 199–204.

[8] D.L. Thiele, C. Kimme-Smith, T.D. Johnson, M. McCombs, L.W. Bassett, Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes, Med. Phys. 23 (4) (1996) 549–555.

[9] Y. Jiang, R.M. Nishikawa, D.E. Wolverton, C.E. Metz, M.L. Giger, R.A. Schmidt, C.J. Vyborny, K. Doi, Malignant and benign clustered microcalcifications: automated feature analysis and classification, Radiology 198 (1996) 671–678.

[10] O. Tsujii, A. Hasegawa, Y.C. Wu, S.C.B. Lo, M.T. Freedman, S.K Mun, Classification of microcalcifications in digital mammograms for the diagnosis of breast cancer, SPIE Med. Imaging, 2710 (1996) 794–804.

[11] J. Suckling et al., The mammographic image analysis society digital mammogram database, Exerpa Medica, International Congress Series, Vol. 1069 (1994) 375–378.

[12] A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[13] Z. Chi, H. Yan, Feature evaluation and selection based on an entropy measure with data clustering, Opt. Eng. 34 (12) (1995) 3514–3519.

[14] D.S. Broomhead, D. Lowe, Multi-variable functional interpolation and adaptive networks, Complex Anal. Ser. B2 (1988) 205.

[15] D.E. Rumelhart, J.L. McCelland, Parallel distributed processing: explorations in the microstructure of cognition, Foundation, Vol. 1, The MIT Press, Cambridge, 1986.

[16] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-Posed Problems, Winston, Washington, 1977.

[17] C. Bishop, Improving the generalization properties of radial basis function neural networks, Neural Comp. 3 (1991) 579–588.

[18] D.J.C. Mackay, Bayesian interpolation, Neural Comp. 4 (3) (1992) 415–447.

[19] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.

[20] S. Chen, C.F.N. Cowan, P.M. Grant, Orthogonal least squares learning for radial basis function networks, IEEE Trans. Neural Networks 2 (2) (1991) 302–309.

[21] M.J.L. Orr, Regularization in the selection of radial basis function Centers, Neural Comp. 7 (1995) 606-623.

[22] T. Kohonen, Self Organization and Associative Memory, Springer, New York, 1988.

[23] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (9) (1990) 1481–1497.