Breast Cancer Diagnosis Using Neural-Based Linear Fusion Strategies

Yunfeng Wu[1], Cong Wang[1], S.C. Ng[2], Anant Madabhushi[3], and Yixin Zhong[1]

[1] School of Information Engineering, Beijing University of Posts and Telecommunications,
Xi Tu Cheng Road 10 Haidian District, Beijing 100876, China
y.wu@ieee.org
[2] School of Science and Technology, The Open University of Hong Kong, Hong Kong
[3] Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08854, USA

Abstract. Breast cancer is one of the leading causes of mortality among women, and the early diagnosis is of significant clinical importance. In this paper, we describe several linear fusion strategies, in particular the Majority Vote, Simple Average, Weighted Average, and Perceptron Average, which are used to combine a group of component multilayer perceptrons with optimal architecture for the classification of breast lesions. In our experiments, we utilize the criteria of mean squared error, absolute classification error, relative error ratio, and Receiver Operating Characteristic (ROC) curve to concretely evaluate and compare the performances of the four fusion strategies. The experimental results demonstrate that the Weighted Average and Perceptron Average strategies can achieve better diagnostic performance compared to the Majority Vote and Simple Average methods.

1 Introduction

Breast cancer is one of the leading forms of cancer diagnosed among women in the United States [9]. The latest surveillance investigation indicates this type of cancer accounts for an estimated 32% incidence rate and an estimated 15% mortality rate in 2005, ranking second only to lung carcinoma [9]. The most common and palpable signs of cancer are lumps or masses detected in the breast, and the benign masses are frequent in a majority of cases [12]. Studies have shown that early diagnosis by means of breast imaging, including digital mammography, ultrasound imaging, and magnetic resonance imaging (MRI), could help prognosis and increase therapeutic options [4]. In this paper, we are considering the binary classification problem of distinguishing benign or malignant breast lesions. In order to improve the biopsy yield ratio, techniques and systems are being developed for computer-aided diagnosis, to effectively assist radiologists and physicians in screening and diagnosis [1].

Recently, artificial neural networks have been applied to classifying mammographic masses for early-stage breast cancer detection and diagnosis [21], which would help reduce the number of unnecessary surgical biopsies. Artificial neural networks, with the properties of experience-based learning and generalization ability, are regarded as one of the emerging computational technologies for solving complex problems that might not have a tractable solution provided by traditional methods.

I. King et al. (Eds.): ICONIP 2006, Part III, LNCS 4234, pp. 165 – 175, 2006.

However, when given a complex data set, different neural classifiers typically provide diverse generalizations by determining different boundaries. The variety of performance would be dramatically influenced by a number of factors, including different network architectures, learning styles (supervised or unsupervised), network architecture (the number of layers and hidden nodes involved, type of activation functions, and degree of connectivity), training parameters (weights initialization, learning rates, and training epochs), and so forth.

Previous research showed that an ensemble of neural networks may significantly improve the generalization capability of an intelligent system [10], [18]. Other than solely toiling over the training data toward an expected generalization, a group of Component Neural Networks (CNNs) could work collectively with given fusion strategies to ameliorate the classification capability, and then hopefully solve an entire complex problem. The ensembles of neural networks can be divided into two main categories: Generative and Nongenerative methods [11]. The Generative methods generate a series of CNNs whose training sets are determined by the performance of former ones (e.g. Boosting [5]), or based on the bootstrap sampling data sets (e.g. Bagging [2]). The Nongenerative ensembles combine their well-devised CNNs to comprehend the entire problem and drive a comprehensive decision with the fusion strategies. The research focus has recently been shifted from practical applications of ensembles towards investigating why ensembles and fusion strategies may work so well and in which situations some methods may outperform the others [14], [23]. In the following sections, we will focus on the Nongenerative ensemble methods in the context of distinguishing between malignant and benign breast lesions.

The rest of this paper is organized as follows. Section 2 and Section 3 describe the optimal Multilayer Perceptron (MLP) architecture selection and several linear fusion strategies applied in our experiments. Section 4 presents the empirical results of breast cancer diagnosis. Section 5 discusses some technical details of linear fusion strategies. Conclusion and directions for the future work are presented in Section 6.

2 Optimal MLP Architecture Selection Based on Regularization

The implementation of the optimal MLP architecture selection in our work contains two steps: First, search the minimum risks associated with a series of MLP structures based on parameter regularization and cross-validation; later select the optimal MLP architecture according to the dynamics of the minimum-risk ranking.

Interpreted as a nonlinear system, a MLP maps the input features **x**, $\mathbf{x} \in \Re^{P \times N}$ by following the rule: $\mathbf{O}(\mathbf{x}, \mathbf{w})$, $\mathbf{O} \in \Re^{P \times M}$. Referring to Hornik *et al.* [8], we consider the MLP with *N* input nodes, *K* hidden nodes in only one hidden layer, and *M* output nodes (herein denoted as (*N*-*K*-*M*) architecture.) Let $w_{k,m}$ be the weight between *m*-th output node and *k*-th hidden node, and $\overline{w}_{n,k}$ be the weight between the *k*-th hidden node and *n*-th input node. The MLP architecture is selected by minimizing a scalar risk function $R(\mathbf{w}, \boldsymbol{\lambda})$, which is the sum of a performance-loss function $E(\mathbf{w})$, and a complexity-cost function $C(\mathbf{w})$ parameterized by a linear regularization vector $\boldsymbol{\lambda}$, i.e.,

$$R(\mathbf{w}, \boldsymbol{\lambda}) = E(\mathbf{w}) + \boldsymbol{\lambda}^{\mathrm{T}} C(\mathbf{w}) \tag{1}$$

where the parameter $\boldsymbol{\lambda}$ represents the relative importance of the complexity-cost in respect of the performance-loss.

For regression and signal processing problems, the loss function is normally measured by mean squared errors between the expected targets $\mathbf{t}_p$ and the estimated outputs over training patterns, i.e.,

$$E(\mathbf{w})=\frac{1}{P}\sum_{p=1}^{P}\left[\mathbf{t}_p-\hat{\mathbf{O}}(\mathbf{x}_p,\mathbf{w})\right]^2=\frac{1}{P}\sum_{p=1}^{P}\left[\mathbf{t}_p-\hat{\mathbf{O}}(\mathbf{x}_p,\mathbf{w})\right]^{\mathrm{T}}\cdot\left[\mathbf{t}_p-\hat{\mathbf{O}}(\mathbf{x}_p,\mathbf{w})\right]=\frac{1}{P}\sum_{p=1}^{P}e_p(\mathbf{w})^{\mathrm{T}}\cdot e_p(\mathbf{w}) \tag{2}$$

where $e_p(\mathbf{w})$ denotes the error between the expected targets and estimated outputs.

There are some complexity regularization methods, well-known as Weight Decay [7] and Weight Elimination [19]. Here we only consider the Weight Decay proposed by Hinton *et al.* [7]. In the Weight Decay, the complexity cost is defined as squared norm of the synaptic weights, including the input-to-hidden and hidden-to-output weights. Thus the regularization term in the risk function is

$$\begin{aligned}\boldsymbol{\lambda}^{\mathrm{T}}\cdot C(\mathbf{w})&=\boldsymbol{\lambda}^{\mathrm{T}}\cdot\|\mathbf{w}\|^2=\lambda_{NK}\left\|\overline{\mathbf{w}}_{N,K}\right\|^2+\lambda_{KM}\left\|\mathbf{w}_{K,M}\right\|^2\\&=[\lambda_{NK},\ \lambda_{KM}]\cdot\left[\left\|\overline{\mathbf{w}}_{N,K}\right\|^2,\left\|\mathbf{w}_{K,M}\right\|^2\right]^{\mathrm{T}}=[\lambda_{NK},\ \lambda_{KM}]\cdot\left[\overline{\mathbf{w}}_{N,K}^{\mathrm{T}}\cdot\overline{\mathbf{w}}_{N,K},\mathbf{w}_{K,M}^{\mathrm{T}}\cdot\mathbf{w}_{K,M}\right]^{\mathrm{T}}\end{aligned} \tag{3}$$

For architecture selection purpose, the Cross-Validation approach [3], [16] is employed to validate the optimal network architecture with the best-performance parameter estimates. Normally, data for regression and classification problems may involve a training set and a testing set, and in the L-fold cross-validation method, all the available training set of P patterns would be randomly split into L disjoint subsets of approximately equal size, i.e. $P=\bigcup_{l=1}^{L}P_{\mathrm{V}}^{l}$ and $\forall i\neq j: P_{\mathrm{V}}^{i}\bigcap P_{\mathrm{V}}^{j}=\varnothing$. Training and validation are repeated for a total of L trials, in the l-th iteration using the subset $P\setminus P_{\mathrm{V}}^{l}$ for training and the other subset P_{V}^{l} for validation. The performance-loss of L-fold cross-validation is estimated by the average of validation mean squared errors:

$$\Gamma_{\mathrm{V}}(\hat{\mathbf{w}})=\frac{1}{L}\sum_{l=1}^{L}E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l}) \tag{4}$$

$$E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l})=\frac{1}{P_{\mathrm{V}}^{l}}\sum_{j\in P_v^l}\left[\mathbf{t}_j-\hat{\mathbf{O}}(\mathbf{x}_j^{l},\hat{\mathbf{w}}^{l})\right]^2=\frac{1}{P_{\mathrm{V}}^{l}}\sum_{j\in P_v^l}e_j(\hat{\mathbf{w}}^{l})^{\mathrm{T}}\cdot e_j(\hat{\mathbf{w}}^{l}) \tag{5}$$

Using the second order information during regularization [3], the parameter vector $\boldsymbol{\lambda}$ would converge through the gradient descent path of the network risk:

$$\boldsymbol{\lambda}^{(i+1)}=\boldsymbol{\lambda}^{(i)}-\eta\cdot\nabla_{\boldsymbol{\lambda}}\Gamma_{\mathrm{V}}(\hat{\mathbf{w}}) \tag{6}$$

where is $\eta>0$ is the convergence update rate. Note that during the i-th iteration the synaptic weights $\hat{\mathbf{w}}$ (or to be explicitly written as $\hat{\mathbf{w}}(\boldsymbol{\lambda})$) is an implicit function of $\boldsymbol{\lambda}^{(i)}$, since $\boldsymbol{\lambda}$ could only be optimized after the settlement of synaptic weights $\hat{\mathbf{w}}$. In case of linear regularization, the gradient of the cross-validation error is

$$\nabla_{\boldsymbol{\lambda}}\Gamma_{\mathrm{V}}(\hat{\mathbf{w}})=\frac{1}{L}\sum_{l=1}^{L}\frac{\partial E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l})}{\partial\boldsymbol{\lambda}} \tag{7}$$

Following the differential chain rule, the gradient vector of the cross-validation error can be derived:

$$\frac{\partial E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l})}{\partial \boldsymbol{\lambda}} = \frac{\partial \mathbf{w}^{l^{\mathrm{T}}}}{\partial \boldsymbol{\lambda}} \cdot \left(\hat{\mathbf{w}}^{l}(\boldsymbol{\lambda})\right) \cdot \frac{\partial E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l})}{\partial \mathbf{w}} \tag{8}$$

where $\partial \mathbf{w}^{l^{\mathrm{T}}} / \partial \boldsymbol{\lambda}$ is the derivative matrix of synaptic weights. To get this derivative matrix, we consider the Taylor expansion of scalar risk function around $\boldsymbol{\lambda}^{(i)}$:

$$\frac{\partial R(\mathbf{w}, \boldsymbol{\lambda})}{\partial \mathbf{w}} = \frac{\partial R(\mathbf{w}, \boldsymbol{\lambda}^{(i)})}{\partial \mathbf{w}} + \frac{\partial^{2} R(\mathbf{w}, \boldsymbol{\lambda}^{(i)})}{\partial \mathbf{w} \partial \boldsymbol{\lambda}^{\mathrm{T}}} \cdot \left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(i)}\right) + o\left(\left|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(i)}\right|\right) \tag{9}$$

where $o\left(\left|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(i)}\right|\right)$ represents a high-order small value which could be ignored when estimated. Note that when regularization parameter vector $\boldsymbol{\lambda}$ meets the optimal scene (i.e. both the gradient of the cross-validation error and network risk cannot be updated further), derived from (9), we have

$$\frac{\partial^{2} R(\mathbf{w}, \boldsymbol{\lambda}^{(i)})}{\partial \mathbf{w} \partial \boldsymbol{\lambda}^{\mathrm{T}}} = 0 \tag{10}$$

Combining (1), (3), (4), and (7), we may develop

$$\frac{\partial \mathbf{w}^{l^{\mathrm{T}}}}{\partial \boldsymbol{\lambda}} \cdot \left(\hat{\mathbf{w}}^{l}(\boldsymbol{\lambda})\right) = -\frac{\partial C(\hat{\mathbf{w}}^{l})}{\partial \mathbf{w}^{\mathrm{T}}} \cdot \left[\frac{\partial^{2} R(\mathbf{w}^{l}, \boldsymbol{\lambda})}{\partial \mathbf{w} \partial \mathbf{w}^{\mathrm{T}}}\right]^{-1} \tag{11}$$

Finally, substituting (11) into (8) gives

$$\frac{\partial E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l})}{\partial \boldsymbol{\lambda}} = -\frac{\partial C(\hat{\mathbf{w}}^{l})}{\partial \mathbf{w}^{\mathrm{T}}} \cdot \left[\frac{\partial^{2} R(\hat{\mathbf{w}}^{l}, \boldsymbol{\lambda})}{\partial \mathbf{w} \partial \mathbf{w}^{\mathrm{T}}}\right]^{-1} \cdot \frac{\partial E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l})}{\partial \mathbf{w}} = -2\hat{\mathbf{w}}^{l^{\mathrm{T}}} \cdot H_{\mathrm{V}}^{-1}(\hat{\mathbf{w}}^{l}) \cdot \frac{\partial E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l})}{\partial \mathbf{w}} \tag{12}$$

where $H_{\mathrm{V}}(\hat{\mathbf{w}}^{l}) = \partial^{2} R(\hat{\mathbf{w}}^{l}, \boldsymbol{\lambda}) / \partial \mathbf{w} \partial \mathbf{w}^{\mathrm{T}}$ is the Hessian matrix of the risk function, and $\partial E_{\mathrm{V}}^{l}(\hat{\mathbf{w}}^{l}) / \partial \mathbf{w}$ could be estimated during training over the validation subset.

3 Linear Fusion Strategies for Combining Neural Classifiers

There are several Non-generative neural networks fusion strategies that have proved to be effective in improving the classification performance [10], [22]. In general, they can be differentiated into two styles: Fixed and Trained rules [15]. Fixed rules, e.g. Majority Vote (MV) [18] and Simple Average (SA) [14], do not need any training phase in the fusion. Trained rules, on the other hand, like Weighted Average (WA) [15] and Perceptron Average (PA) [22], require a learning phase to initialize and adjust fusion parameters. For the MV fusion, the class which receives the largest number votes among the CNNs is chosen as the consensus or majority decision. For the SA and WA fusions, the CNNs are linearly combined to form an overall decision. In this investigation, we use the both fixed and trained fusion methods to effectively improve performance of multiple classifier systems for breast cancer diagnosis.

Simple Average (SA). For the SA fusion, the outputs of the independently trained CNNs are assumed to be scalar-valued and then linearly combined with the equal fusion coefficients to form an overall output. Assume a fusion combines the outputs of total K CNNs, with normalized fusion coefficients α_k, and we have

$$F(\mathbf{x}_p) = \sum_{k=1}^{K} \alpha_k \cdot \mathbf{O}_k(\mathbf{x}_p, \hat{\mathbf{w}}), \quad \alpha_k \geq 0 \tag{13}$$

$$\sum_{k=1}^{K} \alpha_k = 1, \quad k = 1, 2, \ldots, K \tag{14}$$

where $\mathbf{O}_k(\mathbf{x}_p, \hat{\mathbf{w}})$ denotes the output of the k-th CNN for a given p-th input pattern vector $\mathbf{x}_p$. For the SA fusion, the fusion coefficients are $\alpha_k = 1/K$.

Weighted Average (WA). In the WA fusion [17], for a one-dimensional input x_p, the estimation of the *a posteriori* probability of the i-th class from the output of the k-th CNN is denoted as $\hat{p}_k^i(\mathbf{x}_p)$. According to Roli *et al.* [15], it can be expressed as

$$\hat{p}_k^i(\mathbf{x}_p) = p_k^i(\mathbf{x}_p) + \varepsilon_k^i(\mathbf{x}_p) \tag{15}$$

where $p_k^i(\mathbf{x}_p)$ is the *a posteriori* probability of the i-th class, and $\varepsilon_k^i(\mathbf{x}_p)$ denotes the estimation error. Assume that the class boundaries provided from the approximate *a posteriori* probabilities are close to the optimal Bayes boundaries [17]. According to Tumer *et al.* [17], if the estimation errors $\varepsilon_k^i(\mathbf{x}_p)$ on different classes are independent and identically distributed (i.i.d.) variables with zero mean and variance σ_ε^2, the expectation of the added errors (the error in addition to the Bayesian one) can be expressed as $E^{add} = \sigma_\varepsilon^2 / s$, where s is a constant term depending only on the values of probability density functions at the optimal decision boundary. Using (14) and (15), under the hypothesis that the output of the network approximates the posterior probabilities of the classes, the *a posteriori* probability of the linear fusion is

$$\hat{p}_{ave}^i(\mathbf{x}_p) = p_{ave}^i(\mathbf{x}_p) + \sum_{k=1}^{K} \alpha_k \cdot \varepsilon_k^i(\mathbf{x}_p) = p_{ave}^i(\mathbf{x}_p) + \bar{\varepsilon}^i(\mathbf{x}_p) \tag{16}$$

where $\bar{\varepsilon}^i(\mathbf{x}_p)$ denotes the estimation fusion. In the case of uncorrelated estimation errors, the expectation E_{ave}^{add} of the added error of the WA fusion is [15]

$$E_{ave}^{add} = \sum_{k=1}^{K} E_k^{add} \cdot \alpha_k^2 \tag{17}$$

Considering (13), the fusion coefficients that minimize E_{ave}^{add} are [17]

$$\alpha_j = \left(\sum_{k=1}^{K} 1/E_k^{add} \right)^{-1} \cdot \left(1/E_j^{add} \right) \tag{18}$$

In other words, the optimal fusion coefficients are inversely proportional to the expectation errors of each CNN.

Perceptron Average (PA). When the data are statistical independent Gaussian distributed, the operation of the Bayes classifier reduces to a linear classifier, which is equivalent to the perceptron having exponential family activation functions [6]. Note that the WA fusion is "parametric," because its derivation is contingent on the assumption that the underlying distributions of the estimation errors $\varepsilon_k(\mathrm{x}_p)$ on different classes are Gaussian, which may limit its area of applications. On the other hand, the perceptron convergence algorithm is "non-parametric" in the sense that it makes no assumptions concerning the form of the underlying distributions [6]. It operates by concentrating on errors that occur where the distributions overlap. It may therefore work well when the input patterns are generated by some nonlinear physical mechanisms whose distributions might be heavily skewed and non-Gaussian. With such a concept, we may utilize the perceptron convergence algorithm to train the linear fusion to obtain the optimal fusion coefficients assigned to each output of the CNNs. In the PA fusion, the bias $b^{(n)}(\mathrm{x}_p)$ over the p-th input pattern at the n-th training epoch is treated as an additional coefficient driven by a fixed input equal to +1. Let $D^{(n)}(\mathrm{x}_p)$ denote the desired fusion output at the n-th training epoch, we have:

$$D^{(n)}(\mathbf{x}_p) = \begin{cases} +1 & \text{if } \mathbf{x}_p \text{ belongs to } \textit{malignant} \\ -1 & \text{if } \mathbf{x}_p \text{ belongs to } \textit{benign} \end{cases} \tag{19}$$

Thus, the fusion coefficients and bias are updated by following the rule:

$$\alpha_k^{(n+1)} = \alpha_k^{(n)} + \left[D^{(n)}(\mathbf{x}_p) - \mathrm{sgn}\left(F^{(n)}(\mathbf{x}_p) \right) \right] \cdot \mathbf{O}_k^{(n)}(\mathbf{x}_p, \hat{\mathbf{w}}) \tag{20}$$

$$b^{(n+1)}(\mathbf{x}_p) = b^{(n)}(\mathbf{x}_p) + \left[D^{(n)}(\mathbf{x}_p) - \mathrm{sgn}\left(F^{(n)}(\mathbf{x}_p) \right) \right] \tag{21}$$

4 Experimental Results

4.1 Data Description

The data set applied in our experiments was obtained from the Wisconsin Diagnostic Breast Cancer Database described by Mangasarian *et al.* [13]. The data set contains 569 instances (357 benign cases and 212 malignant cases) with thirty real-valued input features, including the mean, standard error, and "worst" or largest (mean of the three largest values) of ten cell nucleus attributes (i.e. radius, texture, perimeter, area, smoothness, compactness, concavity, concave, points, symmetry, fractal dimension). In the experiments, we split the whole data set into two sets: training set and testing set, each involving 200 instances and 369 instances, respectively. And we divided the thirty input features into three parts: Mean, Standard Error, and Largest Deviation features of the ten cell nucleus attributes correspondingly sent to three CNNs (labelled CNN-1, CNN-2, CNN-3) which to be independently trained by the Resilient Back-propagation, Scaled Conjugate Gradient, and Levernberg-Marquardt algorithms. All the input features were normalized to zero mean and unity standard deviation in order to accelerate the backpropagation learning process. And the MLP performance was validated with the 10-fold cross validation.

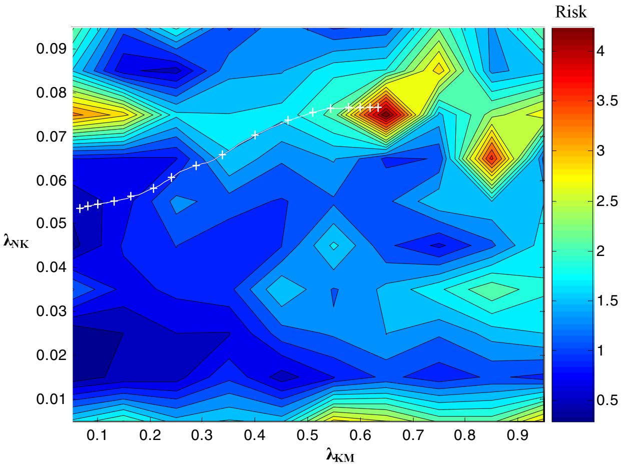


Fig. 1. The convergence of regularization parameter vector $\lambda = [\lambda_{NK}, \lambda_{KM}]^T$ through the gradient descent path of the network risk (the track points are depicted as "+"). The current MLP architecture is (*10-4-1*) and trained by the Resilient BP algorithm.

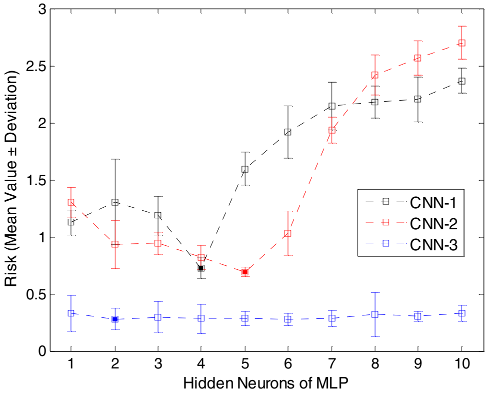


Fig. 2. MLP risk dynamics curves and MSE performance independently carried out by different algorithms for the CNNs in breast cancer diagnosis

4.2 Results of Optimal MLP Architecture Selection

Referring to (1), the risk exported from a MLP would be jointly affected by the synaptic weights and regularization parameter vector λ. In order to achieve the optimal architecture for a particular task, we first relax the number of the hidden nodes at the range from 1 to 10, and then search the most appropriate parameter vector λ which minimizes the risk associated with each certain network structure. In this case, we will have a series of 10 dynamic networks which reach the minimum risks within their own structures. On comparison of the risk dynamics, the optimal MLP architecture could be found. In Fig. 1, we can observe that the regularization parameter vector λ converges through the gradient descent risk path on the network (*10-4-1*) for CNN-1. By locating the regularization parameter vector λ after convergence (referring to (6)), we will find the minimum risk on each network structure. Later, the optimal MLP architecture can be obtained according to the risk dynamics curve varying from a series of hidden nodes (see Fig. 2). The optimal structures for CNN-2 and CNN-3 are (*10-5-1*) and (*10-2-1*), respectively. Their regularization convergence situations are quite similar to the one of CNN-1, and are not illustrated again.

4.3 Results of Breast Lesion Classification Via Linear Fusion Strategies

Table 1 and Table 2 show the weighted coefficients and diagnostic results of the four fusion strategies. Note that absolute errors are the misclassification cases in percentage terms, and for relative error ratios, the averaged error of the CNNs is regarded as 1.0000, and the reported error of each fusion is in fact the ratio over that of the averaged value of the CNNs. In Table 2, we find that MV and SA are at the same degree, when WA and PA are at a lower level in terms of relative error ratio.

Table 1. Normalized fusion coefficients assigned to the CNNs in different fusion strategies

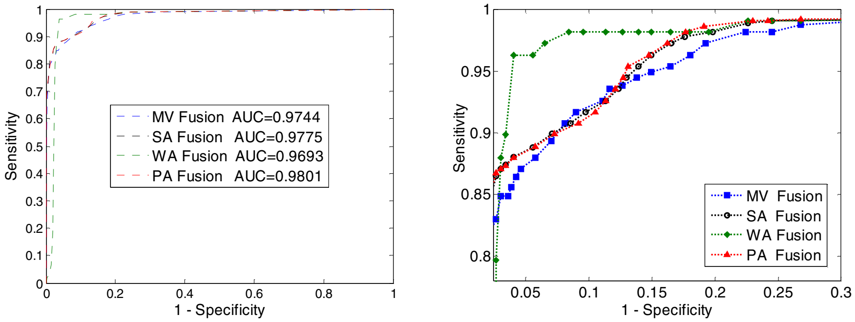
Fusion Strategy	Weighted Coefficients		
	CNN-1	CNN-2	CNN-3
MV Fusion	N/A	N/A	N/A
SA Fusion	0.3333	0.3333	0.3333
WA Fusion	0.3727	0.2037	0.4236
PA Fusion	0.3401	0.2081	0.4518

Table 2. Diagnostic performances of different fusion strategies

	MSE	Absolute Error (%)	Relative Error Ratio
CNN averaged	0.3577	8.9431	1.0000
MV Fusion	0.3111	7.7778	0.8697
SA Fusion	0.2883	7.2087	0.8061
WA Fusion	0.2060	5.1491	0.5758
PA Fusion	0.1951	4.8780	0.5454

Measures of overall error of classification as percentage provide limited indications in a medical diagnostic method. Especially in breast cancer diagnosis, a misidentification between benign mass and malignant tumor has their different costs [4]. The provision of separate correct classification rates for each class, such as Sensitivity/True Positive (TP) rate (the percentage of cancer correctly diagnosed) and Specificity (the percentage of benign lesions correctly diagnosed), can facilitate improved analysis. A Receiver Operating Characteristic (ROC) curve is a plot of operating points showing the possible tradeoff between the classifier's TP rate versus its False Positive (FP) rate

(1-Specificity) [20]. A summary measure of effectiveness of classifier is given by the ROC Area Under Curve (AUC). Here we show two zoomed ROC plots in Figures 3 (a) and (b), because if we move out the range of 0.03 to 0.3 in the horizontal axis, all four fusion ROCs tend to converge with no apparent significant differences. It is clear from Fig. 3 that the PA fusion's ROC covers a larger area (AUC = 0.9801) than the second ranking one of the SA fusion (AUC = 0.9775). It is interesting that the WA fusion strategy just covers the smallest area under the ROC, and the further discussion is provided in Section 5.



(a) (b)

Fig. 3. ROC curves of all four fusion strategies in breast cancer diagnosis: (a) panoramic curves, and (b) zoomed curves from the range of 0.03 to 0.3 in the horizontal axis

5 Discussion

For the MV fusion, assuming that only two classes are considered, and we restrict the choice of the number of CNNs (*N*) to an odd number. The MV fusion will assign the wrong class to input vector $\mathbf{x}$ if at least $\frac{N+1}{2}$ CNNs incorrectly vote for it. It is therefore possible that the consensus decision might be worse than that of the best individual CNN. So the decision of MV fusion might not be always superior to all the individual CNNs.

The SA fusion is widely used due to its simplicity and effectiveness, which has been demonstrated in several experimental studies. However, it might suffer from individual classifiers whose performances are significantly diverse. In our experiments, the SA fusion was poor at fusing the individual CNNs, i.e., the relative error ratio of the SA fusion is 0.2303 and 0.2607 above those in the WA and PA fusions, respectively (see Table 2).

For the WA fusion, we note in Fig. 3 (a) that the ROC curve of the WA fusion ascends slowly (even behind the MV and SA fusions) from 0 to 0.02 in the horizontal axis. We believe that the preliminary assumption of Gaussian distributions for the estimation errors on different classes in the WA fusion results in this phenomenon in the ROC curve, especially when a casualty of training data sizes.

The PA fusion can achieve the lowest absolute error and relative error ratio in our experiments, but it is vastly inferior to the WA when a moderate FP rate is tolerable. This could be the direction for us to improve the PA fusion in the future work.

6 Conclusion

In this paper, we presented the MLP architecture selection method based on parameter regularization and cross-validation, and four linear fusion strategies for combining the component MLP classifiers. The numerical experiments reveals the pitfalls of the MV, SA, and WA fusion strategies in solving the classification of breast lesions, and also exhibits the advantages of the PA fusion strategy, which achieves the lowest absolute error and relative error ratio, and has the top ranking AUC in its ROC versus the other linear fusion strategies. The development of new adaptive weighted average algorithm and the nonlinear fusion strategies will be the next step of our work.

Acknowledgment

This work was supported in part by National Science Foundation of China under the Grant No. 60575034, and the 2005 Innovation Research Funds from Graduate School, Beijing University of Posts and Telecommunications.

References

1. Baker, J.A., Rosen, E.L., Lo, J.Y., Gimenez, E.I., Walsh, R., Soo, M.S.: Computer-aided Detection (CAD) in Screening Mammography: Sensitivity of Commercial CAD Systems for Detecting Architectural Distortion. American J. Roentgenology **181** (2003) 1083–1088
2. Breiman, L.: Bagging Predictors. Machine Learning **24** (1996) 123-140
3. Chen D., Hagan M.: Optimal Use of Regularization and Cross-Validation in Neural Network Modeling. Proc. the 1999 Int’l Joint Conf. on Neural Networks (1999) 1275–1280
4. Donegan, W.L., Spratt, J.S., Orsini, A. (ed.): Cancer of the Breast. 5th edn. Elsevier Science, Amsterdam, Netherlands (2002)
5. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. Proc. the 15th Int’l Conf. on Machine Learning (1996) 148–156
6. Haykin, S.: Neural Networks: A Comprehensive Foundation. 2nd edn. Prentice Hall PTR, Englewood Cliffs, NJ, USA (1998)
7. Hinton, G.E.: Connectionist Learning Procedures. Artificial Intelligence **40** (1989) 185-234
8. Hornik, K.M., Stinchcombe, M., White, H.: Multilayer Feedforward Networks are Universal Approximators. Neural Networks **2** (1989) 359–366
9. Jemal, A., Murray, T., Ward, E., Tiwari, R.C., Ghafoor, A., Feuer, E.J., Thun, M.J.: Cancer Statistics, 2005. CA: A Cancer Journal for Clinicians **55** (2005) 10–30
10. Kuncheva, L.I.: A Theoretical Study on Six Classifier Fusion Strategies. IEEE Transactions on Pattern Analysis and Machine Learning **24** (2002) 281–286
11. Madabhushi, A., Feldman, M., Metaxas, D., Tomasezweski, J., Chute, D.: Automated Segmentation of Prostatic Adenocarcinoma from High Resolution MR by Optimally Combining 3D Texture Features. IEEE Transactions on Medical Imaging **24** (2005) 1611–1625

12. Madabhushi, A., Metaxas, D.: Combining, Low, High and Empirical Domain Knowledge for Automated Segmentation of Ultrasonic Breast Lesions. IEEE Transactions on Medical Imaging **22** (2003) 155–169
13. Mangasarian, O.L., Street, W.N., Wolberg, W.H.: Breast Cancer Diagnosis and Prognosis via Linear Programming. Operations Research **43** (1995) 570–577
14. Roli, F., Giacinto, G.: Design of Multiple Classifier Systems. In: Bunke, H., Kandel, A. (eds.): Hybrid Methods in Pattern Recognition. World Scientific Publishing (2002)
15. Roli, F., Fumera, G., Kittler, J.: Fixed and Trained Combiners for Fusion of Unbalanced Pattern Classifiers. Proc. the 5th Int'l Conf. on Information Fusion (2002) 278–284
16. Stone M.: Cross-validatory Choice and Assessment of Statistical Predictions. Journal of Royal Statistics Society **B36** (1974) 111–133
17. Tumer, K., Ghosh, J.: Analysis of Decision Boundaries in Linearly Combined Neural Classifiers. Pattern Recognition **29** (1996) 341–348
18. Wanas, N.M., Kamel, M.S.: Decision Fusion in Neural Network Ensembles. Proc. of the 2001 Int'l Jt. Conf. on Neural Networks **4** (2001) 2952–2957
19. Weigend, A.S., Rumelhart, D.E., Huberman, B.A.: Generalization by Weight-Elimination with Application to Forecasting. Advances in Neural Information Processing Systems **3** (1991) 875–882
20. Woods, K., Bowyer, K.W.: Generating ROC Curves for Artificial Neural Networks. IEEE Transactions on Medical Imaging **16** (1997) 329–337
21. Wu, Y., He, J., Man, Y., Arribas, J.I.: Neural Network Fusion Strategies for Identifying Breast Masses. Proc. the 2004 Int'l Jt. Conf. on Neural Networks **3** (2004) 2437–2442
22. Wu, Y., Zhang, J., Wang, C., Ng, S.C.: Linear Decision Fusions in Multilayer Perceptrons for Breast Cancer Diagnosis. Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (2005) 699–700
23. Zhou, Z.H., Wu, J, Tang, W.: Ensembling Neural Networks: Many Could be Better Than All. Artificial Intelligence **137** (2002) 239–263