

Linear Decision Fusions in Multilayer Perceptrons for Breast Cancer Diagnosis

Yunfeng Wu, Jinming Zhang, Cong Wang,
*Beijing University of Posts and
Telecommunications
Beijing 100876, P.R. China
y.wu@ieee.org*

Sin Chun Ng
*The Open University of Hong Kong
30 Good Shepherd Street, Homantin,
Kowloon, Hong Kong, P.R. China
scng@ouhk.edu.hk*

Abstract

We introduce a non-parametric linear decision fusion called Perceptron Average (PA) for breast cancer diagnosis. We concretely compare the accuracy between both two fusion strategies for breast cancer diagnosis. The PA fusion demonstrates a higher overall diagnostic accuracy versus the Weighted Average fusion, and the PA fusion method also exhibits a better capability of generalization when a casualty of training data sizes. Moreover, the PA fusion gains a larger area covered by its Receiver Operating Characteristic curve.

1. Introduction

Breast cancer diagnosis is a typical binary classification problem, i.e., to distinguish benign masses from malignant tumors in suspicious lesions. Recently fusion of Component Neural Networks (CNNs) has been utilized for early-stage breast cancer detection and diagnosis [3]. Study showed that fusion of a finite number of CNNs may significantly enhance the generalization ability and ameliorate classification accuracy of a learning system. In this paper we implemented the Perceptron Average (PA) and the Weighted Averaged (WA) fusion and numerically compare their diagnostic performance for breast cancer.

2. Methodology

When the data are statistical independent Gaussian distributed, the WA fusion (see details in [2]) minimizes the expectation of added error of a linear fusion. It is a “parametric” algorithm because its derivation is contingent on the assumption that the underlying distributions of the estimation errors on

different classes are Gaussian, which may limit its area of application. On the other hand, the perceptron convergence algorithm is “non-parametric” and it operates by concerning on errors that occur where the distributions overlap. Based on such a concept, we can train the linear fusion by the perceptron convergence algorithm to obtain the optimal weights assigned to each output of the CNNs. In the PA fusion, the bias $b^{(n)}(\mathbf{x}_t)$ over t -th input pattern at the n -th training epoch is treated as an additional weighted coefficient of fusion, driven by a fixed input equal to +1. Assuming the desired output of the linear fusion at the n -th training epoch is $D^{(n)}(\mathbf{x}_t)$, we have

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

Thus the weighted coefficients and bias are updated as:

$$\alpha_k^{(n+1)} = \alpha_k^{(n)} + [D^{(n)}(\mathbf{x}_t) - \text{sgn}(o^{(n)}(\mathbf{x}_t))] \cdot F_k^{(n)}(\mathbf{x}_t) \quad (2)$$

$$b^{(n+1)}(\mathbf{x}_t) = b^{(n)}(\mathbf{x}_t) + [D^{(n)}(\mathbf{x}_t) - \text{sgn}(o^{(n)}(\mathbf{x}_t))] \quad (3)$$

where the symbol $\text{sgn}(\bullet)$ is the signum function.

3. Experiments and Results

The database [1] we used contains 357 benign cases and 212 malignant cases with thirty real-valued input features, including the mean, standard error, and largest (mean of the three largest values) of ten cell nucleus attributes. We set up a group of six CNNs in the fusion for the diagnosis, and utilized the weight decay approach to regularize the best topology of CNNs following the risk minimization rule. Each CNN was a MLP with the same topology (30 input nodes, 3 hidden nodes, and 1 output node), trained by a number of algorithms respectively: Batch Gradient Descent algorithm, Resilient BP algorithm, Conjugate Gradient algorithm, Quasi-Newton algorithm, One Step Secant

algorithm, and Levenberg-Marquardt algorithm. To observe the generalization capability of each fusion method, we varied the training set sizes from 1% to 100% of the total available data. In Fig. 1, we can find the PA fusion gains a greater advantage of generalization when a casualty of training data sizes. Furthermore, both the two fusions have higher accuracy (99.65% for the PA fusion and 99.30% for the WA fusion) over the one (97.50%) reported in [1].

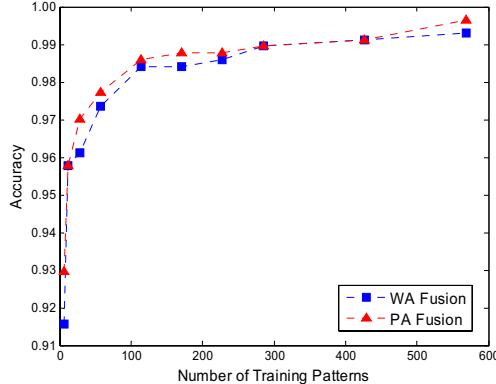


Fig. 1. Accuracy of the WA and PA fusions with a range sizes of training data sets.

Besides the overall accuracy, the provision of separate correct classification rates for each class, such as Sensitivity and Specificity, can facilitate improved analysis. The effectiveness of a classifier can be measured by a Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC). Fig. 2 shows that ROC curve covers a larger AUC (0.9951) in the PA fusion than the AUC (0.9812) in the WA fusion.

Table 1 lists the individual MSE of each CNN, and their corresponding weighted coefficients in the both PA and WA fusions. We note that the PA fusion assigns the CNN#1 with the highest individual MSE: 0.0679 the smallest weighted coefficient (0.1552), however, the WA fusion assigns a medium weighted coefficient (0.1627) to the CNN#1. On the other hand, the PA fusion assigns the largest weighted coefficient (0.1711) to the CNN#5 with the lowest individual

Table 1. Individual MSE of Each CNN and the Corresponding Weighted Coefficients in Fusions

CNN Index	Individual MSE	PA Weighted Coefficients	WA Weighted Coefficient
CNN#1	0.0679	0.1552	0.1627
CNN#2	0.0197	0.1676	0.1739
CNN#3	0.0166	0.1690	0.1675
CNN#4	0.0243	0.1666	0.1732
CNN#5	0.0084	0.1711	0.1637
CNN#6	0.0137	0.1704	0.1590

MSE: 0.0084, but the WA fusion assigns another medium weighted coefficient (0.1637) to the CNN#5. Refer to pattern recognition experience, the CNN having the lowest MSE is considered to play the most important role in the linear fusion, vice versa. The PA fusion followed this rule in the experiments but the WA fusion did not, and as a result, the PA fusion outperformed the WA fusion for breast cancer diagnosis.

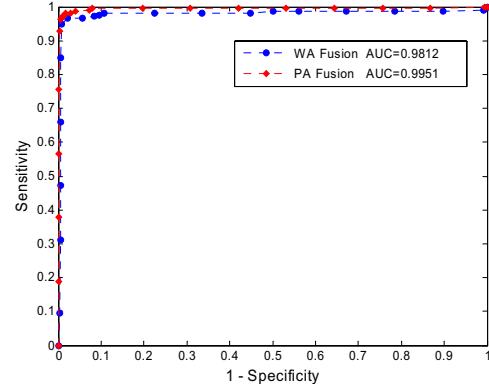


Fig. 2. ROC curves of the WA and PA fusions in the diagnosis.

4. Conclusion

Although the WA fusion minimizes the expectation of added error of a linear fusion, its derivation is contingent on the assumption that the underlying Gaussian distributions of the estimation errors on different classes. On the other hand, the non-parametric PA fusion does not require any assumption concerning the form of the underlying distributions. It may therefore work well when input patterns are generated by some nonlinear physical mechanisms whose probability distributions might be skewed or non-Gaussian. Our experiments exposed the disadvantages of the WA fusion when applied to the practical medical problems, and exhibited the higher level of competence of the PA fusion for a binary classification problem.

References

- [1] O.L. Mangasarian, W.N. Street, and W.H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, no. 4, pp. 570-577, Jul 1995.
- [2] Roli, F., and G. Giacinto, "Design of multiple classifier systems," *Hybrid Methods in Pattern Recognition*, Singapore: World Scientific Publishing, 2002.
- [3] Y.F. Wu, J.J. He, Y. Man, and J.I. Arribas, "Neural network fusion strategies for identifying breast masses," *Proc. 2004 Int'l Jt. Conf. Neural Networks*, pp. 2437-2442, 2004.