Neural Networks
24.7–5

Optimization and Evaluation of a Neural-Network Classifier for PET Scans of Memory-Disorder Subjects

J. Shane Kippenhan[1,2], Warren W. Barker[2],
Shlomo Pascal[2], Ranjan Duara[2], Joachim Nagel[1]

[1] Dept. of Biomedical Engineering, University of Miami, Coral Gables, FL
[2] Wien Center for Memory Disorders, Mt. Sinai Medical Ctr., Miami Beach, FL

ABSTRACT

Back-propagation neural networks were used to classify PET scans as either normal or abnormal, with abnormal subjects defined as subjects who had previously been clinically diagnosed with memory disorders. Numerous neural network experiments were performed in order to achieve optimization with respect to number of hidden units and training duration. Optimizations and performance evaluations were based on ROC analysis, in which the area under the ROC curve was the figure of merit. The neural network's performance was better than that of discriminant analysis, and comparable to the expert's performance, despite the low resolution image data, which consisted of one value per brain lobe, provided to the network.

INTRODUCTION

Quantitative approaches to the analysis and/or classification of Positron Emission Tomography (PET) scans usually involve a region-of-interest (ROI) analysis, in which regional metabolic function in the brain is evaluated [1]. Pattern recognition studies are then performed on these data. Various pattern recognition techniques, including the back-propagation neural network [2], have been applied to the classification of normal and abnormal PET scans based on ROI data. Neural networks appear to perform better than standard statistical methods like discriminant analysis [3].

In the literature describing various recent applications of neural networks, there appears to be relatively little standardization in neural-network training. Networks are often trained to satisfy particular "convergence criteria", which essentially specify how well the hypersurfaces defined by the network are able to separate the different classes comprising the training set. A more important consideration in most circumstances, however, is the ability of a network to generalize and identify previously-unseen patterns, an issue which involves the number of training patterns, the dimensionality of these patterns, the architecture of the network (e.g., the number of hidden units) and number of training iterations. Complex networks trained on high-dimensional patterns for an excessive number of iterations may tend to "memorize" their training sets, and learn criteria that are not generally applicable to populations of given pattern classes. Evaluation of a network's ability to generalize is accomplished by cross-validation studies, that is, testing trained networks on new and independent data sets. The question then arises: what is the most appropriate figure of merit for performance evaluation?

The ROC (Relative-Operating-Characteristic) method of analysis has recently come to be recognized as an objective and comprehensive way to evaluate diagnostic systems, since it measures a diagnostic system's performance independent of decision biases and prior probabilities [4]. The ROC curve represents a system's performance at several different settings of the particular decision criteria. The area under the curve is the "only performance measure available that is uninfluenced by decision biases and prior probabilities, and it places the performances of diverse systems on a common, easily interpreted scale" [4].

The method of optimization to be presented here is based on cross-validation studies, and employs the area under the ROC curve as the figure of merit. Neural network performances were evaluated for subject groups with different levels of dementia, and performances were compared to the performances of an expert human reader (RD) and to those of discriminant analysis.

OPTIMIZATION

A back-propagation neural network with one hidden layer was applied to three separate groups of subjects, each group containing two classes: normal and abnormal (clinical diagnoses were used as reference standards). The "abnormal" class was represented by, in order of decreasing dementia severity, "Late Probable Alzheimer's Disease (AD)" subjects, "Probable AD" subjects and "Possible AD" subjects. Neural-network classification performances were evaluated for different combinations of number of hidden units and training duration. Training and testing was performed twenty times for each combination. Each training session started with a new set of randomized weights. All training was conducted with a learning rate of 0.7, and a momentum constant of 0.9 [2]. Figure 1 depicts the results of experiments which tested performance on the "Possible AD" group.

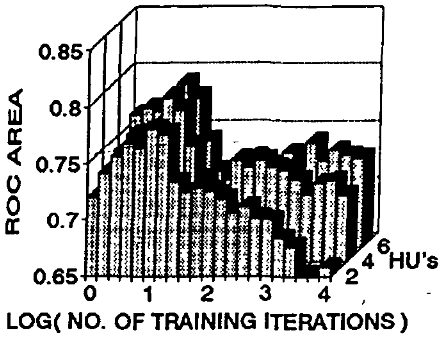

Figure 1: Cross-validation performance for "Possible AD" subject group over a range of training parameter values. Standard Error (SE) ranged from 0.003 to 0.01. Lowest SE's occurred around 10 iterations.

Figure 1 indicates that the best performance was obtained by training for just ten iterations, and that a network with six hidden units performed no better at this point than one with four hidden units. Further iterations resulted in lower ROC areas, indicating that the network was being "overtrained". These trends were typical of experiments with other groups. It can be seen that the best cross-validation performance was achieved after relatively few iterations. The network attains its greatest generalization ability before it has converged to any significant degree.

CH3068-4/91/0000-1472 $01.00 © 1991 IEEE

The claim that this combination is "optimum" for this group is best substantiated by examining the network's performance with regard to its own training set. Figures 2 and 3 indicate that, after long training periods, the networks learn their training sets very well. At this point, RMS output error is very low, and there is no impetus for further change in cross-validation performance.

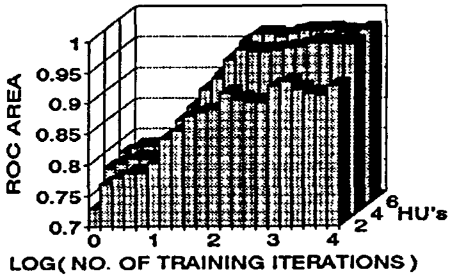

Figure 2: Results of testing on network's **training** set over a range of training-parameter values.

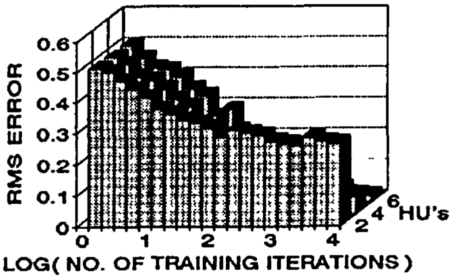

Figure 3: Network RMS output error for training set over a range of training-parameter values.

EVALUATION

For each subject group, the best performance over the given range of hidden units and training duration was chosen, and an average performance (for the 20 experiments that were performed at this "optimum" combination) was calculated.

For the "Late Probable AD" group, two hidden units proved sufficient, while four were required for the other two groups. Figure 4 depicts the ROC curve for the "Late Probable" group.

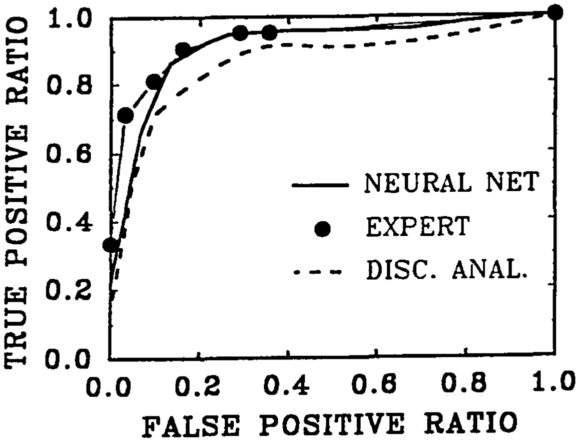

Figure 4: ROC curve for "Late Probable AD" group.

Complete results are summarized in the table below:

Method Group	Expert	Network	Discr. Anal.
Late Probable AD **vs.** **Age-Equiv Normal**	**0.93**	**0.92**	**0.87**
Probable AD **vs.** **Age-Equiv Normal**	**0.89**	**0.85**	**0.78**
Possible AD **vs.** **Age-Equiv Normal**	**0.81**	**0.81**	**0.74**

Table: Classification performance of various classification methods. Each value represents the area under the ROC curve for a given classification method.

Direct comparisons between network performance and discriminant analysis performance are quite valid, since each method was given pattern information in the same format: each pattern was represented by a group of eight values, one for each lobe of the brain. In comparing network performance to the expert's performance, the necessarily low-resolution "view" that the neural network had of each PET study certainly represented a significant handicap. Another factor which may have influenced the results is the fact that, from a quantitative-classifier point of view, the expert's "testing set" was not independent, since the expert was clinically familiar with all of the tested subjects. Part of the expert's training included subjects in the testing set.

REFERENCES

[1] Haxby, JV. Resting state regional cerebral metabolism in dementia of the Alzheimer type. In: **Positron Emission Tomography in Dementia.** R. Duara, ed. Wiley-Liss, pp. 93-116, 1990.

[2] Rumelhart, DE, Hinton, GE and RJ Williams. Learning internal representations by error propagation, In: **Parallel Distributed Processing**, Vol.1, Rumelhart, D.E. and J.L. McClelland and the PDP Research Group, MIT Press, pp. 318-364, 1986.

[3] Kippenhan, JS and JH Nagel. Diagnosis and modelling of Alzheimer's disease through neural network analyses of PET studies,**Proc 12th Ann Int Conf of IEEE/EMBS, 12**, pp. 1449-50, 1990.

[4] Swets, JA. Measuring the accuracy of diagnostic systems. **Science**, 240, pp. 1285-1293, 1988

Correspondence:
Jonathan Shane Kippenhan
Department of Biomedical Engineering
P.O. Box 248294
University of Miami
Coral Gables, FL 33124

CH3068-4/91/0000-1473 $01.00 © 1991 IEEE