# Design and Evaluation of Neural Classifiers Application to Skin Lesion Classification*

Mads Hintz-Madsen, Lars Kai Hansen and Jan Larsen
CONNECT, Electronics Institute, build. 349
Technical University of Denmark,
DK-2800 Lyngby, Denmark
emails: hintz, lkhansen, jlarsen@ei.dtu.dk

Eric Olesen and Krzysztof T. Drzewiecki
Dept. of Reconstructive Surgery S,
The National Hospital of Denmark,
Rigshospitalet, Blegdamsvej 9,
DK-2100 Copenhagen, Denmark

**Abstract**

**We address design and evaluation of neural classifiers for the problem of skin lesion classification. By using Gauss Newton optimization for the entropic cost function in conjunction with pruning by Optimal Brain Damage and a new test error estimate, we show that this scheme is capable of optimizing the architecture of neural classifiers. Furthermore, error-reject tradeoff theory indicates, that the resulting neural classifiers for the skin lesion classification problem are near-optimal.**

## 1 INTRODUCTION

Melanoma is the most lethal of skin cancers. However, patients may be saved from this life threatening cancer if their lesion is detected at an early stage. Computer imaging may assist and improve the detection of such early lesions. The "State of the art" in this field was recently reviewed in an editorial in the journal "Computerized Medical Imaging and Graphics" [1]. Although applied to the problem of skin lesion classification, the main objective of this paper is to introduce and apply a new methodology for optimization of neural classifiers. The methods applied may be considered as an extension of

*To appear: 1995 IEEE Workshop on Neural Networks for Signal Processing (NNSP'95)

the *Designer Net* time series processing tool [8, 9] to the realm of classifiers. In particular we derive the Hessian for the so-called *entropic* cost function and apply the Hessian for Gauss Newton second order optimization and for estimation of weight saliency for use in Optimal Brain Damage pruning. A key ingredient of the proposed method is a new test error estimate for the entropic cost function [2]. The test error estimate enables us to select the optimal network within a nested family of pruned networks.

## 2 NEURAL CLASSIFIERS

The aim in classification is to model the probability of classification, $P(y|\mathbf{x})$, of a given input vector $\mathbf{x}$ where $y$ is the class label. In the context of skin lesion classification the input vector for the classifier is formed from $n_I$ feature measurements on a given skin lesion. If provided with a training set, $D$, consisting of $p$ input-output pairs[1]: $(\mathbf{x}_\alpha, y_\alpha)$, where $\mathbf{x} \in \mathcal{R}^{n_I}$ and $y = \pm 1$, the likelihood of the neural classifier, $F_u(\mathbf{x})$, with parameters (weights) $u$ is given by [7],

$$P(D|u) = \prod_{\alpha=1}^{p} \left( \frac{1 + F_u(\mathbf{x}_\alpha)}{2} \right)^{\frac{1+y_\alpha}{2}} \left( \frac{1 - F_u(\mathbf{x}_\alpha)}{2} \right)^{\frac{1-y_\alpha}{2}} \tag{1}$$

Hence, for the well trained network, $\frac{1}{2}(1+F_u(\mathbf{x})) \sim P(y|\mathbf{x})$. Training is based on minimization of the negative log-likelihood:

$$E(u) = -\log P(D|u) = \sum_{\alpha=1}^{p} \epsilon\left(\mathbf{x}_\alpha, y_\alpha, u\right) \tag{2}$$

with the error measure given by

$$\epsilon\left(\mathbf{x}_\alpha, y_\alpha, u\right) = -\frac{1+y_\alpha}{2} \log\left( \frac{1 + F_u(\mathbf{x}_\alpha)}{2} \right) - \frac{1-y_\alpha}{2} \log\left( \frac{1 - F_u(\mathbf{x}_\alpha)}{2} \right) \tag{3}$$

The cost function (2) is in turn recognized as the entropic cost function (see, e.g., [6]). In order to eliminate overfitting, and for numerical convenience, we often augment the cost function by a weight decay term corresponding to minimizing instead the negative log-posterior,

$$C(u) = E(u) - \log P(u). \tag{4}$$

The log-prior, $-\log P(u)$, is conveniently chosen to be a quadratic form in the weight parameters, e.g, representing a simple weight decay.

[1] We discuss binary classification for convenience.

### 2.1 Generalization

For a given network $u$ the generalization or test error may be defined as,

$$\mathrm{E}_{\mathrm{test}}(u) = \int D\mathbf{x} dy P(\mathbf{x}, y) \epsilon\left(\mathbf{x}, y, u\right). \tag{5}$$

Here, the "true" underlying distribution of examples, $P(\mathbf{x}, y)$, need not be a singular measure, - we do allow noisy classifications, i.e., contradicting labels for the same input, corresponding to $|F_u(\mathbf{x})| < 1$. Since the generalization error involves an average over all possible patterns it is not observable, but may be estimated by invoking additional statistical assumptions. The training set error is given by

$$\mathrm{E}_{\mathrm{train}}(u) = \frac{1}{p}\sum_{\alpha=1}^{p} \epsilon\left(\mathbf{x}_\alpha, y_\alpha, u\right), \tag{6}$$

hence, the average entropic cost on the training set. In the limit, $p \to \infty$, $\mathrm{E}_{\mathrm{train}}(u) \to \mathrm{E}_{\mathrm{test}}(u)$; asymptotic theory quantifies this limiting behavior.

While the above quantifies the generalization of a single classifier we are, in fact, interested in the typical behavior of the test error. Therefore we compute the training set averaged quantities:

$$\langle \mathrm{E}_{\mathrm{test}} \rangle = \int Du P_p(u) \mathrm{E}_{\mathrm{test}}(u) \tag{7}$$

$$\langle \mathrm{E}_{\mathrm{train}} \rangle = \int Du P_p(u) \mathrm{E}_{\mathrm{train}}(u). \tag{8}$$

Here, $P_p(u)$, is the distribution of optimal networks obtained by minimizing the cost function based on randomly selected samples of size $p$ from the example distribution $P(\mathbf{x}, y)$. $P_p(u)$ could be thought of as the distribution of network parameters in an ideal cross validation ensemble.

It is possible to show that $P_p(u)$ is asymptotically *normal* with $P_p(u) \sim N(u^*, \mathbf{\Sigma})$, where $u^*$ are the parameters that minimize the regularized test error, $\mathrm{E}_{\mathrm{test}}(u) - \frac{1}{p}\log P(u)$ [2].

If we now choose the particular network architecture:

$$F_u(\mathbf{x}) = \tanh\left(\phi_u(\mathbf{x})\right), \tag{9}$$

$$\phi_u(\mathbf{x}) = \sum_{j=0}^{n_H} W_j h_j(\mathbf{x}), \tag{10}$$

$$h_j(\mathbf{x}) = \tanh\left(\sum_{k=0}^{n_I} w_{j,k} x_k\right), \tag{11}$$

with $n_H$ hidden units, $n_I$ input units, and parameters $u = (w, W)$, the covariance matrix has the particularly simple form,

$$\begin{aligned} \mathbf{\Sigma} &= (\mathbf{H} + \mathbf{R})^{-1} && (12) \\ \mathbf{H}_{ii'} &\approx \int D\mathbf{x}dyP(\mathbf{x}, y)\left(1 - F_{u^*}^2(\mathbf{x})\right) \frac{\partial \phi_{u^*}(\mathbf{x})}{\partial u_i} \frac{\partial \phi_{u^*}(\mathbf{x})}{\partial u_{i'}} && (13) \end{aligned}$$

where we further simplified the Hessian using the Levenberg-Marquart approximation [10]. $\mathbf{R}$ is the second derivative matrix of the regularization term. The quantity $\mathbf{H}_{ii'}$ may be approximated by the Hessian of the entropic cost function,

$$\mathbf{H}_{ii'} \approx \mathbf{H}_{ii'}^p \equiv \frac{1}{p} \sum_{\alpha=1}^{p} \left(1 - F_{u(D)}^2(\mathbf{x}_\alpha)\right) \frac{\partial \phi_{u(D)}(\mathbf{x}_\alpha)}{\partial u_i} \frac{\partial \phi_{u(D)}(\mathbf{x}_\alpha)}{\partial u_{i'}} \qquad (14)$$

where $u(D)$ are the best weights found for the actual training set $D$.

With the asymptotic form of the cross validation ensemble distribution we are in a position to compute the averaged quantities in (7) and (8), and find,

$$\begin{aligned} \langle \mathrm{E}_{\mathrm{test}} \rangle &= \epsilon_0 + \frac{N_{\mathrm{eff}}}{2p} && (15) \\ \langle \mathrm{E}_{\mathrm{train}} \rangle &= \epsilon_0 - \frac{N_{\mathrm{eff}}}{2p}. && (16) \end{aligned}$$

where the effective number of parameters is $N_{\mathrm{eff}} = \mathrm{Tr}[\mathbf{H\Sigma}]$, and the asymptotic test error given by the average entropic cost of the teacher parameters is,

$$\epsilon_0 = \int D\mathbf{x}dyP(\mathbf{x}, y)\epsilon\left(\mathbf{x}, y, u^*\right). \qquad (17)$$

One may interpret $\epsilon_0$ as a “noise level” for the classifier; if the classifier is “crisp”, i.e., if $|F_{u^*}(\mathbf{x})| \approx 1$ for almost all inputs, $\mathbf{x}$, $\epsilon_0 \approx 0$. On the other hand, if the classifier is “fuzzy”, i.e., $|F_{u^*}(\mathbf{x})| \approx 0$ for almost all inputs, $\epsilon_0 \approx \log 2$.

In a practical situation one only has access to a single training set, and the two averages may be combined to provide a test error estimate,

$$\widehat{\langle \mathrm{E}_{\mathrm{test}} \rangle} = \mathrm{E}_{\mathrm{train}}(u(D)) + \frac{N_{\mathrm{eff}}}{p} \qquad (18)$$

where the average training error is estimated using the actual training error. The estimator (18) may be used to select the optimal network, e.g., among a family of pruned networks, hence, be used as a *pruning stop criterion* similarly to the criterion previously developed for regression type problems in [8, 9].

Figure 1: Skin lesion showing the characteristical large variations in texture and coloring associated with melanoma (original is in color).

### 2.2 Pruning

The architecture (11) may be subjected to pruning by *Optimal Brain Damage* (OBD), developed by Le Cun and coworkers [10]. In OBD the parameters of a network are ranked for pruning according to their importance for the training error. If the importance is estimated using a second order expansion of the training error around its minimum, we find the weight *saliency*:

$$s_i = \frac{1}{2}\mathbf{H}_{ii}^p u_i^2(D), \tag{19}$$

where the Hessian is given as in (14).

## 3 EXPERIMENTAL

We evaluated the optimizing scheme and the classifier theory on the skin lesion classification problem. This classification involves the classes: the malignant, the premalignant and the benign skin lesions. However we focused on the classification problem of separating malignant lesions from benign. An example of a malignant lesion is shown in figure 1 (original is in color). Generic characteristics of melanoma are large variations in coloring, absence or presence of certain texture features and irregular boundaries. The samples

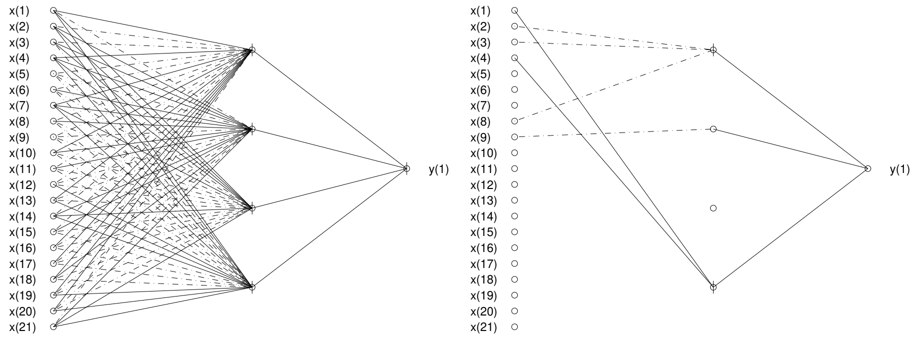


Figure 2: Left: Fully connected network for classification of skin lesions, the 21 inputs represent color and texture features. Right: Architecture of the optimal pruned network. Note that only a few of the available inputs are used by the net. The used features include first and second order texture statistics.

used in the present study were all taken from a photographic library of skin lesions, that were collected at the Dept. of Recontructive and Plastic Surgery at the National Hospital of Denmark, and which all had been considered potentially malignant. Hence, the sampling of the benign group is rather biased. In previous studies [3, 4] it is unclear how the sampling of the classes has been done, making comparisons difficult. The data set consisting of a total of 160 images was split into a training set of 120 images and a test set of the remaining 40 images each containing an even split of the two classes. As input to the network a group of 21 features incorporating color and texture statistics were selected. Boundary features were not incorporated, since they were found not to contribute significantly to the classification of the two classes. We believe that this might be a result of the biased sampling, since most of our benign samples, in fact, have irregular boundaries.

A fully connected network with 4 hidden units was chosen initially, see the left panel of figure 2. In figure 3 we show the development of training and test errors during training of the fully connected network; the weight decay parameter was set to 0.8. Next we pruned the network iteratively according to the OBD saliency ranking, pruning 5% of the remaining weights per iteration. After each pruning session the remaining weights were retrained for 30 epochs. For the resulting nested family of networks training errors, test errors and *estimated* test errors were computed. We also computed the sample standard deviation of the test errors. The development of these error measures during pruning are shown in figure 4. The estimated test error was used to stop the pruning and the architecture of the selected network is shown in the right panel of figure 2. The pruning process is successful in identifying a much smaller network with better test performance than the fully connected network. Furthermore, the variance of the test error of the networks in the

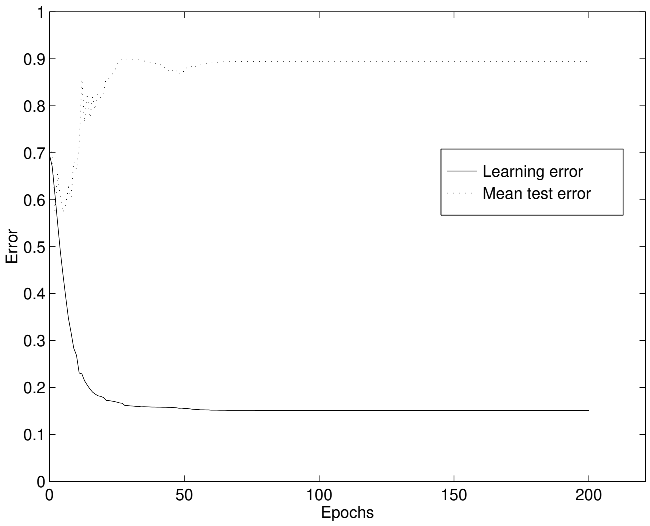


Figure 3: Development of the entropic test and training errors for the fully connected network. Note that the test error shows significant overtraining.

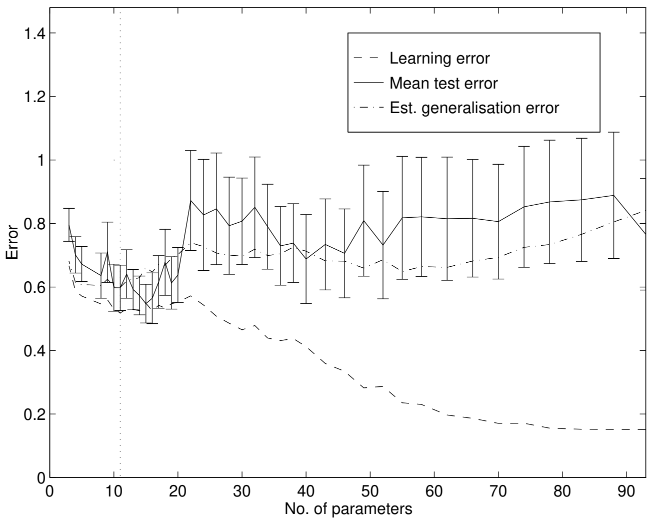


Figure 4: Development of training error, test error, and estimated test error. The vertical dotted line indicates the location of the optimal network according to the estimated test error. The error bars on the test set errors indicate sample standard deviations. Note the close correspondence between the estimated and empirical test errors.

vicinity of the optimal network is significantly lower; hence we can be more confident in the properties of these networks. Although the entropic test error is indeed smaller for the pruned network than for the fully connected network, it is of interest to see what the *classification* error of the two networks are. Following Bayes decision theory we selected the class label according to the sign of the network output when converting probabilities to classifications. In this way we found, that the pruned network classified 74% of the lesions correctly on the training set and 66% on the independent test set. The fully connected network classified 98% correctly on the training set and 66% on the test set, ie. when converted to classifications the performances of the two networks are similar.

For comparison we have performed a *k-Nearest Neighbor* (k-N-N) analysis of the data sets. Within k-N-N a pattern is classified according to a simple majority vote among its $k$ nearest neighbors (using the simple Euclidean metric). The training error may be computed from the training set by including the actual pattern in the vote. A *leave-one-out* "validation" error on the training set may be computed by *excluding* the actual pattern from the neighbor vote. Finally, the test patterns may be classified by voting among the $k$ nearest neighbors found among the training patterns. Using the validation error we found that $k = 3$ was optimal for this data set. The training error for the $3-$N-N scheme was found to be 83%, while the test error was 63%, ie., the network classifiers have slightly better performance than the k-N-N standard algorithm.

Since the neural classifier is trained to produce classification probabilities (and not only Bayes classifications), we can inspect the error-reject trade-off induced by a reject threshold on the probability (rejecting decisions for which $|F_u(\mathbf{x})| < \tau$). The error-reject trade-off was recently discussed in [5]. Denoting the classification error rate, at reject rate $R$, by $E(R)$, it was argued that near-optimal binary classifiers should obey the relation (for low reject rates),

$$E(R) = E(0) - (1/2 - E(0)) \cdot R \tag{20}$$

Since $E(0) \approx 0.35$ for the present system, we expect the slope of the trade-off curve (the efficiency of the reject mechanism) to be $\gamma \equiv 1/2 - E(0) \approx 0.15$. This is indeed confirmed by the actual error-reject trade-off curve presented in figure 5.

## 4 CONCLUSION

We have developed a methodology for design and evaluation of neural classifiers. The approach was applied to the problem of skin lesion classification. The new test error estimator for classifiers was shown to be able to produce

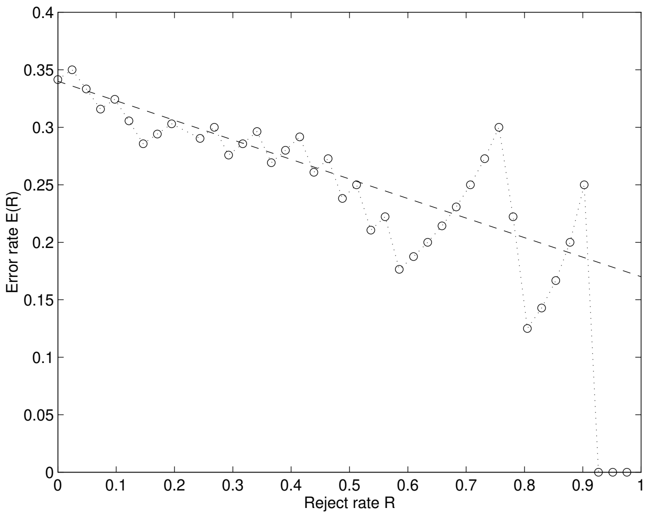


Figure 5: The error rate $E(R)$ versus reject rate $R$ computed on the test set examples. The initial slope of the trade-off curve is approximately given by $1/2 - E(0)$ in line with a recent theoretical analysis of probability driven reject mechanisms in near-optimal binary classifiers.

valid estimates of the empirical test error and could be used to select optimal networks among a family of pruned networks. The optimal network for the skin lesion classification problem based its classification on texture statistics. Currently, the aim is to establish more empirical data for validation of the neural classifier design approach and to compare our classification results with other recent neural network approaches for solving the skin lesion classification problem.

## ACKNOWLEDGMENT

This research is supported by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center. MHM wishes to acknowledge a generous donation from Novo Nordisk A/S Academic Affairs. JL thanks the Radio-Parts Foundation for financial support.

## REFERENCES


[1] W.V. Stoecker, R.H. Moss, F. Ercal and S.E. Umbaugh: "Editorial: Digital Imaging in Dermatology". Computerized Medical Imaging and Graphics **16**, 145-150, (1992).

[2] L.K. Hansen et al.: "Asymptotic generalization in neural classifiers". In preparation (1995).

[3] A. Kjoelen, S.E. Umbaugh, R.H. Moss and W.V. Stoecker: "Artificial Intelligence Applied to Detection of Melanoma". IEEE Engineering in Medicine and Biology, 15th Annual Conference, Vol. 2, 604-605, (1993).

[4] F. Ercal, A. Chawla, W.V. Stoecker, H-C. Lee and R.H. Moss: "Neural Network Diagnosis of Malignant Melanoma from Color Images". IEEE Transactions on Biomedical Engineering, Vol. 41, No. 9, September (1994).

[5] L.K. Hansen, Chr. Liisberg, and P. Salamon: "The Error Reject Trade-off". Submitted for publication. Available by anonymous ftp: eivind.ei.du.dk/dist/hansen.reject.ps.Z.

[6] J. Hertz, A. Krogh and R.G. Palmer: "Introduction to the Theory of Neural Computation", Addison Wesley, New York (1991).

[7] D. MacKay: "The Evidence Framework Applied to Classifier Networks". Neural Computation **4**, 720-736, (1992).

[8] C. Svarer, L.K. Hansen and J. Larsen: "On Design and Evaluation of Tapped-Delay Neural Network Architectures" The 1993 IEEE Int. Conference on Neural Networks, San Francisco. Eds. H.R. Berenji et al., 45-51, (1993).

[9] C. Svarer, L.K. Hansen, J. Larsen and C.E. Rasmussen: "Designer Networks for Time Series Processing". The 1993 IEEE Workshop on Neural Networks for Signal Processing (NNSP'93) Baltimore. Eds. C.A. Kamm et al., 78-87, (1993).

[10] Y.L. Cun, J.S. Denker and S.A. Solla: "Optimal Brain Damage" In Advances in Neural Information Processing Systems II (Denver 1989), ed. D.S. Touretzky, 396-404. San Mateo: Morgan Kaufmann, (1989).