On Using Feedforward Neural Networks for Clinical Diagnostic Tasks

Georg Dorffner
Austrian Research Institute for Artificial Intelligence
Schottengasse 3
A-1010 Vienna, Austria
Tel: +43–1–53532810, Fax: +43–1–5320652,
email: georg@ai.univie.ac.at

and
Dept. of Medical Cybernetics and Artificial Intelligence
University of Vienna

Gerold Porenta
Dept. of Cardiology
2nd Clinic of Internal Medicine
University of Vienna Medical School
Waehriger Guertel 18-20, A-1090 Vienna
Tel: +43–1–40400, Fax: +43–1–4081148
email: porenta@vm.akh–wien.ac.at

On Using Feedforward Neural Networks for Clinical Diagnostic Tasks

Georg Dorffner
Austrian Research Institute for Artificial Intelligence
and
Dept. of Medical Cybernetics and Artificial Intelligence
University of Vienna
georg@ai.univie.ac.at

Gerold Porenta
Dept. of Cardiology
2nd Clinic of Internal Medicine
University of Vienna Medical School
porenta@vm.akh–wien.ac.at

Abstract

In this paper we present an extensive comparison between several feedforward neural network types in the context of a clinical diagnostic task, namely the detection of coronary artery disease (CAD) using planar thallium-201 dipyridamole stress-redistribution scintigrams. We introduce results from well-known (e.g. multilayer perceptrons or MLPs, and radial basis function networks or RBFNs) as well as novel neural network techniques (e.g. conic section function networks) which demonstrate promising new routes for future applications of neural networks in medicine, and elsewhere. In particular we show that initializations of MLPs and conic section function networks—which can learn to behave more like an MLP or more like an RBFN—can lead to much improved results in rather difficult diagnostic tasks.

Keywords: Feedforward neural networks, neural network initialization, multilayer perceptrons, radial basis function networks, conic section function networks; thallium scintigraphy, angiography, clinical diagnosis and decision making.

1 Introduction

In this paper we present an extensive comparison between several feedforward neural network types in the context of a clinical diagnostic task, namely the detection of coronary artery disease (CAD) using planar thallium-201 dipyridamole stress-redistribution scintigrams. We demonstrate that given the peculiar aspects of many clinical applications—such as small and unbalanced training sets—the widely used multilayer perceptrons (MLP) and the backpropagation learning rule need not be the method of choice to arrive at optimum re-

sults. We introduce results from well-known (e.g. radial basis function networks, or RBFN) as well as novel neural network techniques (e.g. conic section function networks, or CSFN) which demonstrate promising new routes for future applications of neural networks in medicine, and elsewhere. In particular we show that initializations of MLPs and conic section function networks—which can learn to behave more like an MLP or more like an RBFN—lead to much improved results in the rather difficult task of detecting significant CAD from planar thallium images using coronary angiography as the reference method. We also discuss the viability of MLP initialization for screening a multitude of different possible training sets and thus reducing the training time during the design of an optimal network for generalization.

2 Detecting Coronary Artery Disease with Neural Networks

Since 1988, in a cooperation between the Austrian Research Institute for Artificial Intelligence and the Dept. of Cardiology at the University of Vienna, we have been investigating the possible use of neural networks in the assessment of coronary artery disease (CAD) [15, 16]. The reference method to confirm or exclude CAD is coronary angiography, which is an evasive procedure with an associated rate of complication including rare cases of death. Among a variety of non-invasive tests myocardial thallium scintigraphy has proven useful in the detection of CAD [13]. Diagnostic statements of clinical relevance in CAD concern the presence of disease (yes/no), the localization (vascular territory) and the extent of disease (single or multivessel).

The very nature of the interpretation of scintigrams lends itself nicely to an automated solution employing neural networks. In particular, data from the segmental analysis of thallium-201 scintigraphy was chosen as input to a three-layer perceptron, while the output was designed to correspond to one of the three diagnostic classifications as mentioned above. The process of diagnosis in this setting is a one-shot mapping from the input data to the output. Thus the involved knowledge (in the sense of artificial intelligence) or structures have to be considered as low-level information, in contrast to knowledge sources that many medical expert systems have to deal with. At the same time this technique appears easily generalizable to other similar problems of medical image classification and interpretation.

Planar thallium–201 scintigraphy generates images of the myocardial tracer uptake after pharmacological stress with dipyridamole and at rest in standard views of the heart (i.e. the projections anterior, LAO 45 degrees, LAO 70 degrees).

The original scintigrams are stored on computer file as 64 x 64 8-bit pixel matrices. The complete encoding of such an image would require a very large number of input units, which is associated with several limitations. First, large input patterns require extreme digital storage space. Secondly, the number of input units increases the number of connections and thus the training time. Thirdly, training networks with large numbers of connections would require a similarly large number of training cases, as the network has to achieve optimization of a process involving as many degrees of freedom as connections. In our case, a maximum of

159 patient records were available, far too few to match the number of connections in a full–fletched network taking entire images as input. For these reasons, a method of reducing the size of a single input pattern was desired from the beginning. The method we employed is based on segmental analysis of the circumferential profile and generates normalized count activities in five anatomical segments per planar view. In addition, segmental washout rates were also obtained. All results were achieved using this reduced input representation of 45 values. This shows that while analysis of images without preprocessing may exceed available resources, in many cases efficient solutions can be achieved by data reduction algorithms preserving the relevant information.

For all the 159 data sets an expert diagnosis was available, while for a subgroup for 81 patients also results from angiography were obtained. Two series of experiments were conducted using either the expert diagnosis of the results from angiography as target output. 30 cases of the former population, and 20 cases of the latter one were taken for training (leaving the rest as a test set for evaluating the generalization performance). Alternatively, the diagnosis from expert readings of the scintgrams or from angiography were used as target output. The prototype network consisted of 45 input units (corresponding to segmental thallium uptake and washout values), 5 to 30 hidden units, and 1 to 3 output units (depending on the type of diagnosis). Different variations of the backpropagation learning rule [19] were applied. The goal for the initial studies was to find the network with the best performance on the test set. The test set was varied in order to achieve this goal, while keeping it as small as possible to increase the confidence level of the results. Results were 87 to 92 % for the expert output and 61 to 85 % for the angiography results.

In a subsequent study (described in [16]), the goal was to evaluate the robustness of the method over different conditions. Five different training sets were selected from the original populations and the average results over all runs were plotted. Three different outputs were used: either one output unit indicating the presence of CAD, or two units representing the location of the disease, or two units representing the extent of disease. The results showed that the network could predict the expert's readings far better than the angiography results. While for the former the results indicated that the diagnostic performance of the neural network appeared potentially promising, the network trained on the angiography data did not achieve a level that could be usefully applied in clinical practice. Thus the goal of the present work was to improve the results on predicting the angiography readings. Although this is a very specific goal, this application can serve as a testbed for similarly structured medical applications. Therefore, in the next section we briefly review some aspects common to many diagnostic tasks in clinical medicine.

3 Clinical Diagnosis and Supervised Learning

Clinical decision making and diagnosis frequently involves the identification of normal and abnormal patterns obtained from diagnostic tests. In a task of classifying patterns representing clues about a disease based on supervised neural networks, the goal is to find an optimum network which is likely to yield reproducible and accurate results in future routine use. Given

a set of pathological and normal cases, the task is therefore to use a small fraction of that data for training, and to keep a large fraction for evaluation (test set). It is generally hypothesized that the network with the best prediction rate on the test set will also lead to high prediction rates when presented with novel data (not available during the network design). Thus, as it was done in the above-cited work, several subsets of the sample set have to be selected for several training runs to identify an optimum network. In general, the larger the test set is in relation to the training set, the more reliable the results get.

The application described above exemplifies several aspects of clinical decision making that may impede this kind of application of supervised neural networks.

– diagnostic tests may be associated with risks to the patient, such as morbidity and mortality for invasive tests (e.g. angiography) or radiation burden (scintigraphy). Consequently, very often only a rather small number of training samples are available.

– for the same reasons, plus ethical reasons prohibiting the application of invasive techniques to healthy subjects, training sets are often unbalanced. For instance, in the case of angiography, the number of normal cases is considerably smaller than the number of CAD cases (15 vs. 66). Furthermore, normal reference cases may include a systematic bias, since angiography is only conducted if there is sufficient evidence suggesting CAD.

– for the routine clinical use of diagnostic systems the indication of confidence values, i.e. values expressing the probability of a diagnosis being correct, are of critical importance. For neural networks this can be achieved by an indication of how many percent of cases were classified correctly at a certain threshold value of activation at the output units. To arrive at reliable confidence values, many experiments with large test sets are advisable. This, again, puts restrictions on the size of the training sets.

Similar problems have also been encountered in other clinical applications conducted by the Austrian Research Institute for Artificial Intelligence, such as the classifcation of EEG data ([5, 17]) or the detection of alarms in CTG signals in laboring women. The research reported here aimed at exploring several methods and alternative learning rules to deal with such problems, all in the context of the application of detecting CAD from myocardial scintigrams.

4 Methodology of the Conducted Experiments

To study and compare the different neural network types, several series of experiments have been conducted. For all these experiments, the focus of the study was on the prediction of the presence of CAD with angiography as standard. As mentioned above, from a total data set of 81 cases always 20 (15 pathological/5 normal) were included in the training set and 61 (52/9) in the test set. To cross-evaluate the consistency of each method and to conduct a wide-spread search for the optimum network, five different training sets of the same size and distribution and with minimal overlap were chosen (henceforth labeled ‘ang20–0 through ‘ang20–4’). Each rule with each parameter setting was trained on all five sets, and tested with each of the five corresponding test sets, respectively.

The diagnostic performance of clinical tests is often specified as a pair of sensitivity (percentage of correctly classified pathological cases) and specificity (percentage of correctly classified normal cases) values. Thes two values are independent on each other, in that increasing one of them will decrease the other. Therefore, for a more accurate descriptio, so-called ROC curves that plot sensitivity over 1–specificty are given to assess the interdependence of the two values. Since there is only one output unit, sensitivity and specificity can be altered for a given network by shifting the threshold that discriminates between normal and abnormal (decides whether the output is interpreted as 1 or 0). Theoretically, any threshold in the interval]0..1[(excluding the values 0 and 1) could be meaningful since the criterion is optimal separation given certain constraints (such as a certain specificty). To evaluate each network, the threshold was varied between 0.01 and 0.99 at steps of 0.01. For any given specificity value (corresponding to every number between 0 and 9 of normal cases being misclassified), the maximum achievable sensitivity value (the minimum number of pathological cases misclassified) was recorded. For the purpose of the present study, we plotted sensitivity (on the y-axis) over specificity values (on the x-axis) instead of ROC curves, either as polygone curves or as bar graphs. Only specificity values above 40 % (above 4 correctly classified normals) were considered, since results below that are of limited utility.

If single values could not be computed (meaning that shifting the threshold by 0.01 resulted in a jump of false positives by two) then they were linearly interpolated. If values at specificities above 80 % (0 or 1 false positives) could not be computed then 0 % was assumed (extrapolation would have been too error-prone). Thus, results in this area are not always reliable. Fig. 1 shows such a curve for the best multilayer perceptron . It demonstrates that sensitivity

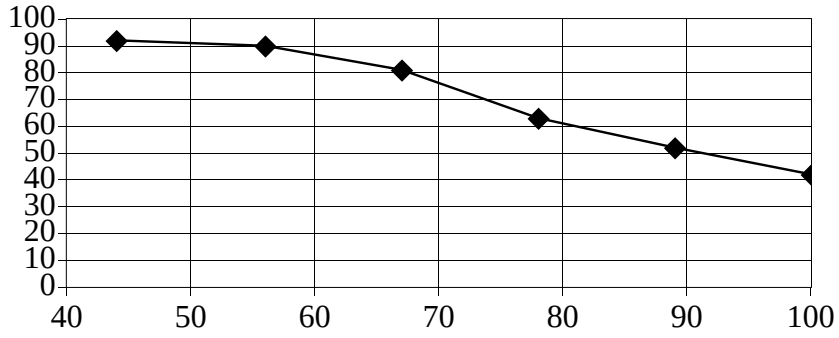

Figure 1*: A sensitivity over specificity curve for the best network trained on the angiography standard using multilayer perceptrons and backpropagation.*

critically drops below 80 % when specificity is raised to just 67 % or more. Thus the results are of little use in practical application. Alternatively, bar graphs are used to show the same results.

Wherever possible, 15 hidden units were used (since this was the value chosen to be optimum in the previous studies – [16]).

5 Multilayer Perceptrons and Radial Basis Function Networks

In the original long-term study ([16]) only multilayer perceptrons using the backpropagation learning rule were used. It is probably fair to say that this type of network belongs to the most widely applied ones, in medicine (e.g. [3], [8], [9], [11], [20]) and elsewhere. As is widely known, backpropagation is extremely slow and requires a large number of training steps. In the case of the angiography data the number of training steps (each step comprising one cycle through backpropagation based on a randomly chosen training pair) ranged from 1400 to 2100. Compared to other applications this is rather low but still required long training times given the large number of runs that had to be performed during the design of the optimum network.

Another fact about backpropagation is that it works best if training sets are balanced, i.e. if the numbers of positive and negative cases are about equal. This was not possible to achieve with the angiography data, since only 14 out of 81 cases were normal, but any training set smaller than 20 was not sufficient for acceptable results. Thus the training sets were chosen so as to keep the ratio of pathological and normal cases about equal for training and test sets. No duplication of cases (a possible way of achieving balance) was conducted so as not to bias the network toward individual cases. Consequently, the results achieved are probably influenced by the negative impact of the unbalanced training sets. An indication of this are the rather low specificity values for high sensitivity (as in fig.1), suggesting a bias toward the more numerous pathological cases.

An alternative network, which is becoming increasingly popular for practical applications, is the radial basis function network ([1],[2]). Such networks have even been applied to a similar case of detection of CAD in thallium scintigrams ([18]). The difference to multilayer perceptrons is best described when considering the so-called decision regions each unit in either network can apply to the space of input patterns. MLP units divide the space into two half-spaces by hyperplanes (reflected in the linear term of the weighted sum), while RBFN units cut out hyperspherical regions around training points (reflected by the quadratical term of the Euclidean distance and by the Gaussian activation function; for an overview see [12] and [7]). Theoretically, an RBFN can be built with only positive examples (or at least with fewer negative than positive examples), since the decision regions are closed and can be set on the basis of single training points, while MLPs need pairs of positive and negative samples to set the hyperplanes.

What training time is concerned, RBFNs have a decisive advantage over MLPs. Since the first layer (the “hidden layer”) is initialized by setting the centers of the hyperspheres according to the training samples, only the connections from the hidden units to the output units have to be trained. Since this part is a linear associator, the simple delta rule ([23]) is sufficient, which on average needs fewer cycles and less CPU time to achieve good results.

The biggest disadvantage of RBFNs is that they can be too restrictive on the basis of their closed decision regions. Given a pattern of input values one unit becomes most sensitive only to a certain small range of values, everything above and below counts as negative (see, also,

[5]). In other words, patterns are classified according to a nearest-neighbor criterion given several prototypes (represented by several hidden units). It depends on the application whether this is appropriate. In the case of scinitgrams at the input it might not be, at least not totally. There, the input data is derived from concentrations of thallium-201 over several regions of the heart. Over large parts of the input, the criterion for CAD is certainly whether this concentration is very high or very low, not whether it is at a certain value (and neither above nor below). In this sense, aspects of the distruibution of the input data, overlooked in most applications, critically decide upon the appropriateness of either MLP or RBFN. In many cases, however, it is very difficult or impossible to determine the data distribution and thus decide upon which type of network (MLP or RBFN) is most appropriate.

Fig.2 shows the results of applying an RBFN as described in [2] to the angiography data (black bars), compared to the previous results from backpropagation (white bars). Since

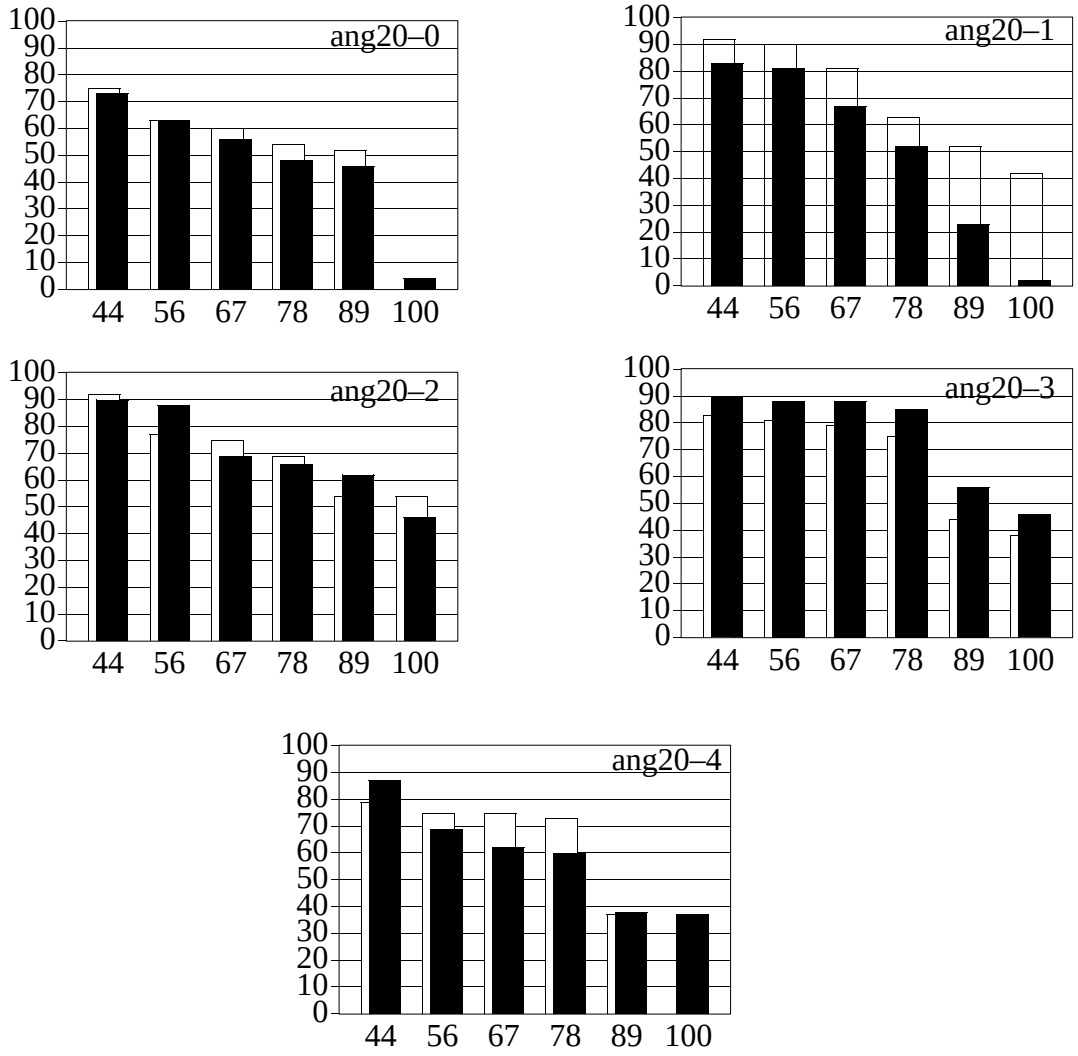

Figure 2: *The results from radial basis function networks (black bars) compared to multilayer perceptrons with backpropagation (white bars), depicted as sensitivity over specificity curves, for all five training sets (evaluated over the corresponding test sets).*

there are 20 training vectors, a full-fletched RBFN consists of 20 hidden units (black bars). Backpropagation was done with a learning rate of 0.1 and a momentum factor of 0.9 (see [19]) until all training cases were learned with a tolerance of 0.3. The delta rule to train the hidden-to-output units of the RBFN was used with a learning rate of 0.2 for 2000 steps. For both learning rules, training pairs were presented randomly.

The results show that the RBFN with 20 hidden units was consistently better in only one of the five cases, while in the other four it tended to be inferior. This indicates that indeed this method might be too restrictive unless some favorable training set is chosen.

6 Initializing Multilayer Perceptrons

As described in the previous section, a decisive advantage of RBFNs is the ease with which they can be trained. Weights between the input and hidden layer are directly set through an initialization procedure, namely identical to the input values of one training sample. In the subsequent training of the hidden-to-output weights with the delta rule, convergence is guaranteed and the network cannot get stuck in local minima. As a result, a single training run is sufficient to guarantee an optimum result for a current parameter setting (which, too, is usually smaller than for the MLP), while an MLP usually has to be trained several times with different inital (random) weight configurations in order to rule out local minima.

To make an MLP profit from similar effects, we developed a simple method of initializing the input-to-hidden layer weights (independently also reported in [21]). The basic idea is a direct analogy to the initialization of RBFNs, namely setting the decision borders (the surface dividing the decision regions) according to training samples. For an MLP unit, given a pair of a positive and a negative sample, this easily translates to setting a hyperplane which optimally divides the two points. This is the case if the hyperplane dissects the line connecting the two points halfway between the points, while being perpendicular to that line (see fig. 3 for the three-dimensional case). If the two points are cluster centers, this kind of separation is called a Voronoi tesselation. Through solving the corresponding set of equations (defining the hyperplane) the following formulas for calculating the weights and the thresholds are easily derived (see [7] and [21]

$$w_{ij} = x_{ji}^{+} - x_{ji}^{-} \qquad (1)$$

$$\theta_j = \frac{1}{2}\sum_{i=1}^{n}\left(x_{ji}^{+\,2} - x_{ji}^{-\,2}\right)$$

$$y_j = \sum_{i=1}^{n} w_{ij}x_i - \theta_j$$

where x_{ji}^{+} and x_{ji}^{-} are the i-th input values of the j-th positive and negative sample, respectively (assuming that there be one hidden unit j for each pair). y_j is the net input of the j-th unit before applying a sigmoid transfer function. This methodology nicely reflects the abovementioned need for negative samples in an MLP. It is clear that among the large number of

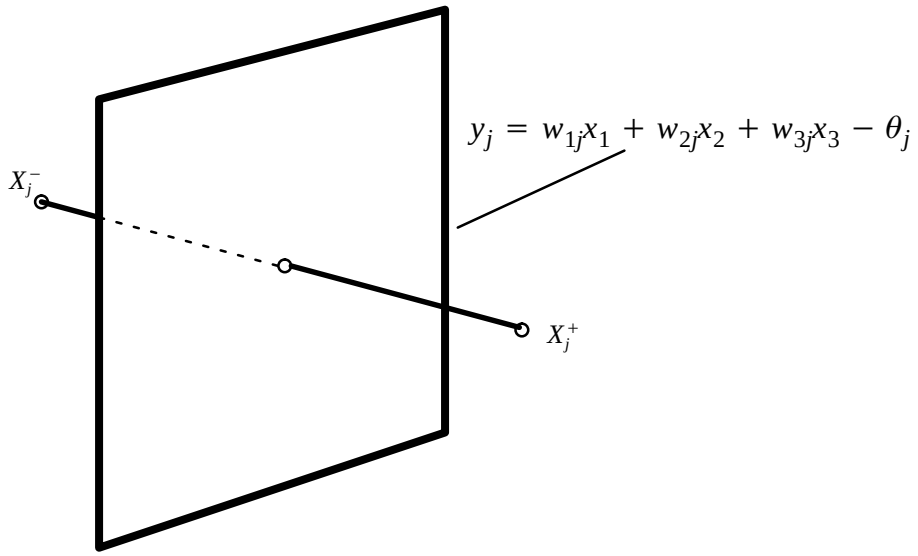

***Figure 3**: Setting a hyperplane, corresponding to the weighted sum in the propagation rule, such that it optimally separates a positive (X+) and a negative training sample (X–).*

possible pairs, in practical applications it is advisable to choose only a subset. It is also clear that an optimum outcome of such an initialization procedure can only be expected if not arbitrary data points are chosen, but cluster centers derived from cluster analysis (e.g. from LVQ, [10]). But, as the results below demonstrate, cluster analysis need not be necessary. To improve results, among the hidden units there should be detectors of both positive and negative cases. Negative detectors are easily derived by switching x_{ji}^+ and x_{ji}^- in the above formulas. Setting the hidden-to-output weights can be done the same way as in the RBFN, i.e. by applying the delta rule.

Fig. 4 shows the rather surprising results of applying the initialization procedure based on randomly chosen pairs of positive and negative samples of the angiography data, plus the delta rule for 200 steps with a learning rate of 0.2 (grey bars; backpropagation is depicted as white bars, as before). To make immediate comparsion with the backpropagation results possible, a constant number of 15 hidden units was chosen for all runs. In many cases (ang20-0, 2 and 4) initialization plus delta rule (the direct analogy to RBFNs) lead to improved results over backpropagation. Especially in areas of specificity between 60 and 80 % the curves are considerably flatter, thus permitting high sensitivity values. These improvements raise the neural network method, applied to scintigrams plus output from angiography to a level that might be clinically useful.

Another important application of the initialization method can be derived from these experiments. The results show that the goodness of the classification after simple initialization (i.e. all weights are initialized, no delta rule) strongly correlates with the goodness of results of the final trained networks (delta rule, and original backpropagation). In other words, the in-

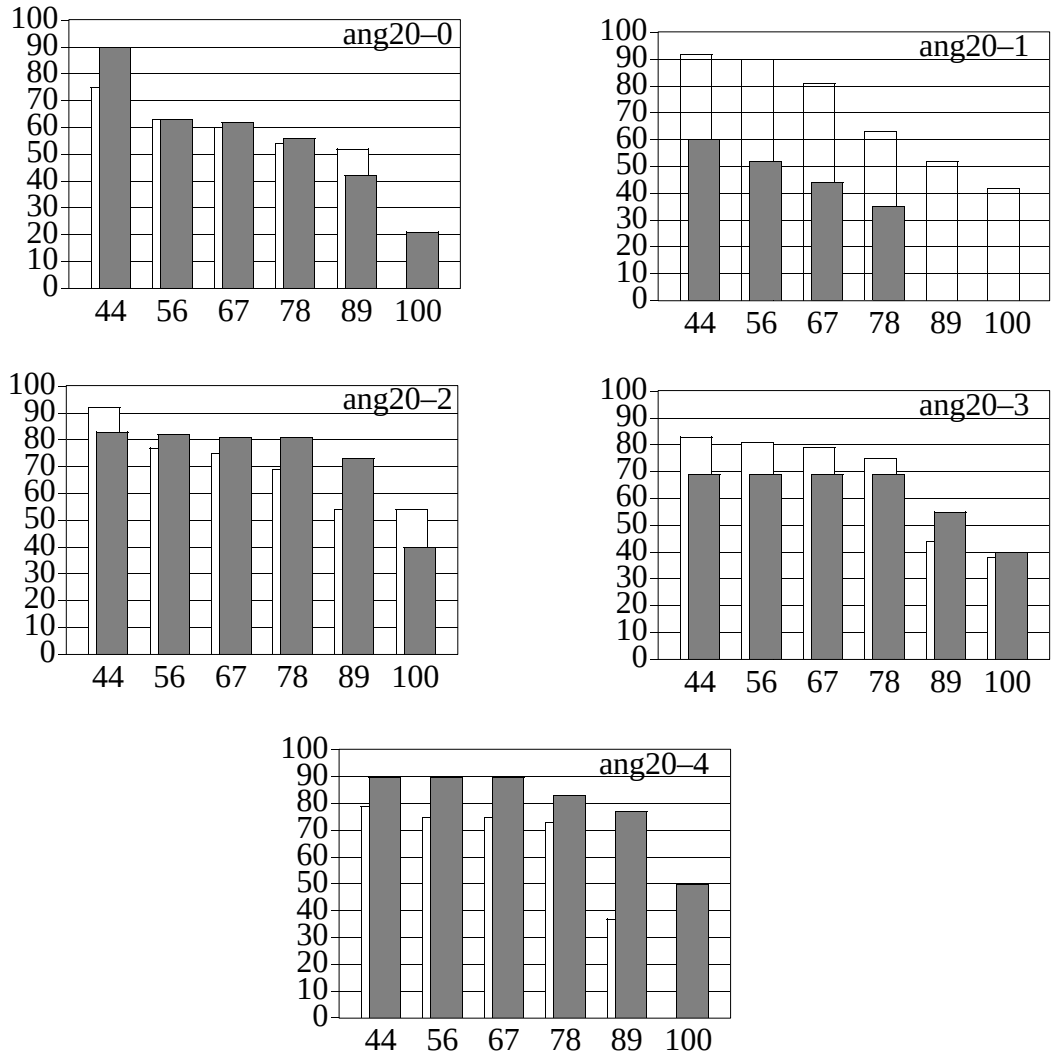

Figure 4: *The results from MLP initialization plus delta rule (dark grey bars) compared to multilayer perceptrons with backpropagation (white bars), depicted as sensitivity over specificity curves, for all five training sets (evaluated over the corresponding test sets).*

itializtion can serve as a good indicator for which training set might be optimal for maximizing the final prediction rate (as defined above). Thus it can be used for screening a large number of training sets so as to choose a small number of promising candidates for the actual gradient descent training. Valuable time in the design of the network solution can be gained this way.

7 Conic Section Function Networks (CSFN)

To combine the advantages of MLPs and RBFNs, a unification of the two has been developed, called Conic Section Function Networks (CSFN; see [7]). It is based on the observa-

tion that both hyperplane and hypersphere are special cases of generalized conic section functions. In the n-dimensional space of input patterns, each decision border is seen as the intersection of an n+1-dimensional hypercone with that space (fig. 5 shows an example for n=2 in a frontal view). According to this, the net input of a CSFN unit is given through the

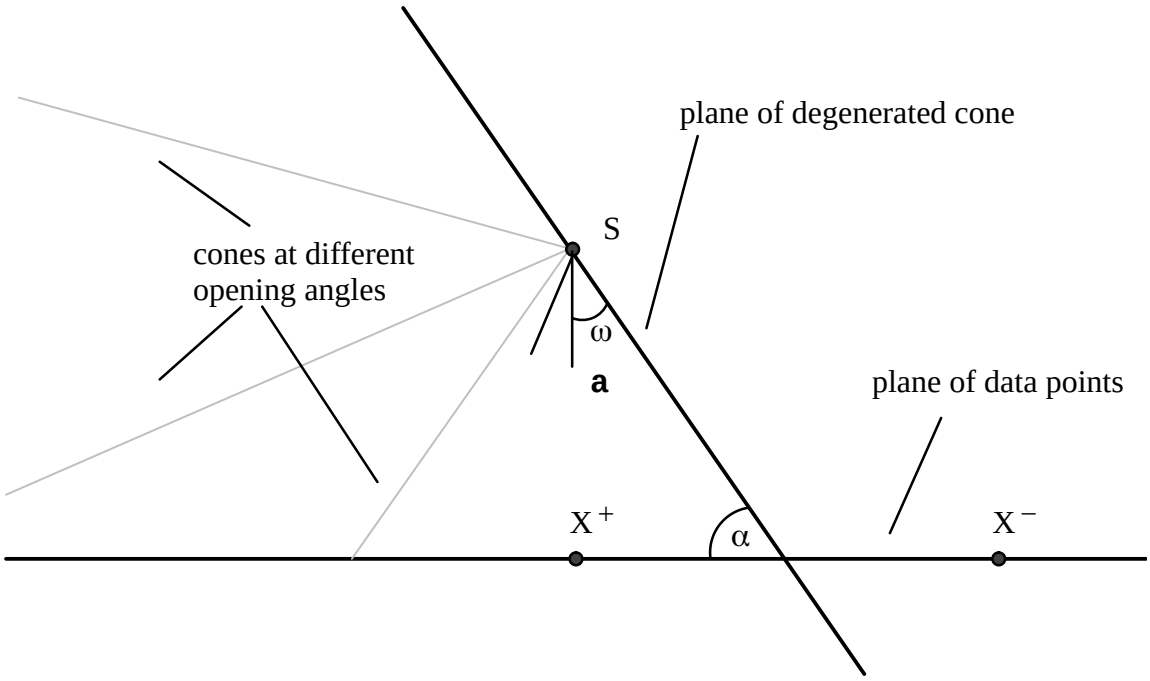

Figure 5*: A three-dimensional cone in a frontal view, intersecting a two-dimensional input space forming a circle, ellipse, or (being degenerated as a plane) a straight line. The tip of the cone and the angle α are kept constant, while the opening angle ω is altered.*

following equation (for a derivation see [7]):

$$y_j = \sum_{i=1}^{n+1}(x_i - s_i)a_{ij} - \sqrt{\sum_{i=1}^{n+1}\left(x_i - s_{ij}\right)^2}\cos\omega_j \qquad (2)$$

$$x_{n+1} \equiv 0$$

where $2\omega_j$ is the opening angle of the cone j (hidden unit j). One can easily see that this formula conains two major parts, roughly corresponding to an MLP (weighted sum) and an RBF term (Euclidean distance). For this we have to equate the *a* factors with the weights in an MLP, and the *s* factors with the weights, or better still, *offsets* or *center coordinates*, of an RBFN. Contrary to work in [12, 4] the two sets of degrees-of-freedom are not identical but separate. This results in a doubling of resources by introducing two degrees-of-freedom per connection, certainly one major drawback of the CSFN.

More precisely, all a_{ij} are given through

$$a_{ij} = w_{ij}\Big(- \cos(\omega_j + \alpha_j)\Big), \;\; 1 \leq i \leq n \qquad (3)$$

$$a_{n+1,j} = \sin\Big(\omega_j + \alpha_j\Big)$$

The parameters s_{ij} are set identical to the coordinates of a positive class sample like in the RBFN (or in the above-described initializiation of MLPs). $s_{i,n+1}$ determines α and can be chosen such that $\alpha = \frac{\pi}{4}$ (45 degrees).

The most attractive feature of CSFNs is that there exists one major parameter (ω) the variation of which can gradually shift a CSFN unit from being an MLP unit ($\omega = \frac{\pi}{2}$, separating the input space by a hyperplane. Here the cone degenrates to a being a plane) to being an RBFN unit ($\omega + \alpha = \frac{\pi}{2}$), separating the space by a hypersphere), where α remains constant.

This one parameter can rather easily be adapted by one of several rules (also called "cone-folding"), for instance, a gradient descent rule maximizing the separability of each unit. This rule is applied as follows. First the network is initialized, for instance, by the method of setting hyperplanes according to eq. (1). Then an input pattern is presented and activations are propagated to the hidden layer. Finally ω is adapted according to the following equation

$$\Delta\omega_j = \begin{cases} -\eta\delta \;\; \text{... if } \bar{y}_j > 0 \text{ and } w_{jk} > 0 \\ +\eta\delta \;\; \text{... if } \bar{y}_j < 0 \text{ and } w_{jk} < 0 \\ 0 \;\; \text{... otherwise} \end{cases} \qquad (4)$$

$$\delta_j = \bar{y}_j \sin(\omega + \alpha) + s_{i,n+1} \cos(\omega + \alpha) + \bar{\bar{y}}_j \sin\omega_j$$

where $\bar{y}_j$ and $\bar{\bar{y}}_j$ are the two major parts of the net input (equation (2))

$$\bar{y}_j = \sum_{i=1}^{n} (x_i - s_{ij}) w_{ij} \qquad (5)$$

$$\bar{\bar{y}}_j = \sqrt{\sum_{i=1}^{n+1} \Big(x_i - s_{ij}\Big)^2}$$

such that

$$y_j = \bar{y}(-\cos(\omega + \alpha)) + s_{i,n+1} \sin(\omega + \alpha) - \bar{\bar{y}} \cos\omega_j \qquad (6)$$

and w_{jk} is the weight from hidden unit j to the output unit (either also initialized or acquired through learning, indicating whether the unit is a positive or negative detector). A derivation can, again, be found in [7]. If many closed regions exist in the input data, then some of the hidden units, starting off as MLP units, will learn to set ω close to $\frac{\pi}{2}$ and thus form RBFs. Thus the above problem of the unknown appropriateness of the type of decision region is largely solved automatically. The reverse approach, starting as RBFN and letting some units open their regions is possible, too, but was not tested in this project. A possible negative effect can be overfitting of the training data, which can be controled by a parameter limiting the maximum change of ω.

The effect of adapting ω can be expected to be larger if not single data but cluster centers are used for initialization. Therefore, for a second round of experiments a cluster analysis, done by the LVQ method ([10], [21], p.85ff) was applied to all five training sets, resulting in 3 to 5 positive and 2 to 3 negative clusters (positive and negative samples were clustered separately).

Fig.6 shows the effect of applying learning of ω, both with single data (left column) and cluster center initialization (right column). Dark grey bars depict the results of initialization, light grey bars depict the results after cone-folding. Eq. (4) was applied twice with each training pair (which empirically turned out to be an optimum number in this application), leading to an additional training run of only 40 cycles. Subsequently, the delta rule was applied for 2000 cycles, during which training pairs were presented randomly. The graphs show that this type of gradient descent learning on ω lead to consistent, albeit sometimes small, improvements, especially in the interesting areas of specificity of 60 to 80 %. As expected, improvements were larger after the cluster initialization. Notable exceptions are training set 3, and to some extent 4.

8 Discussion

In a comparative summary, figs. 7 and 8 show the results of multilayer perceptrons (MLPs) with backpropagation (white, diamonds), radial basis function networks (RBFNs; black, circles), conic section function networks (CSFNs) with initialization (dark grey, crosses), and with cone folding (gradient descent learning over ; light grey, triangles). Comparisons for both initialization by single data points and by cluster centers are shown. Initialization of RBFNs by cluster centers has also been tried but lead to inferior results (and was thus omitted). The graphs show that MLPs are consistently outperformed, with the best results due to CSFNs. Interestingly, RBFNs, CSFNs with single data initialization and CSFNs with cluster center initialization seem to complement each other, in that in all five training sets one of them produces the best results. This appears to indicate a high dependency on the chosen training set, which is not surprising given the small number of training samples. It also indicates that by using all of the described neural network methods, satisfying results can still be achieved in most cases.

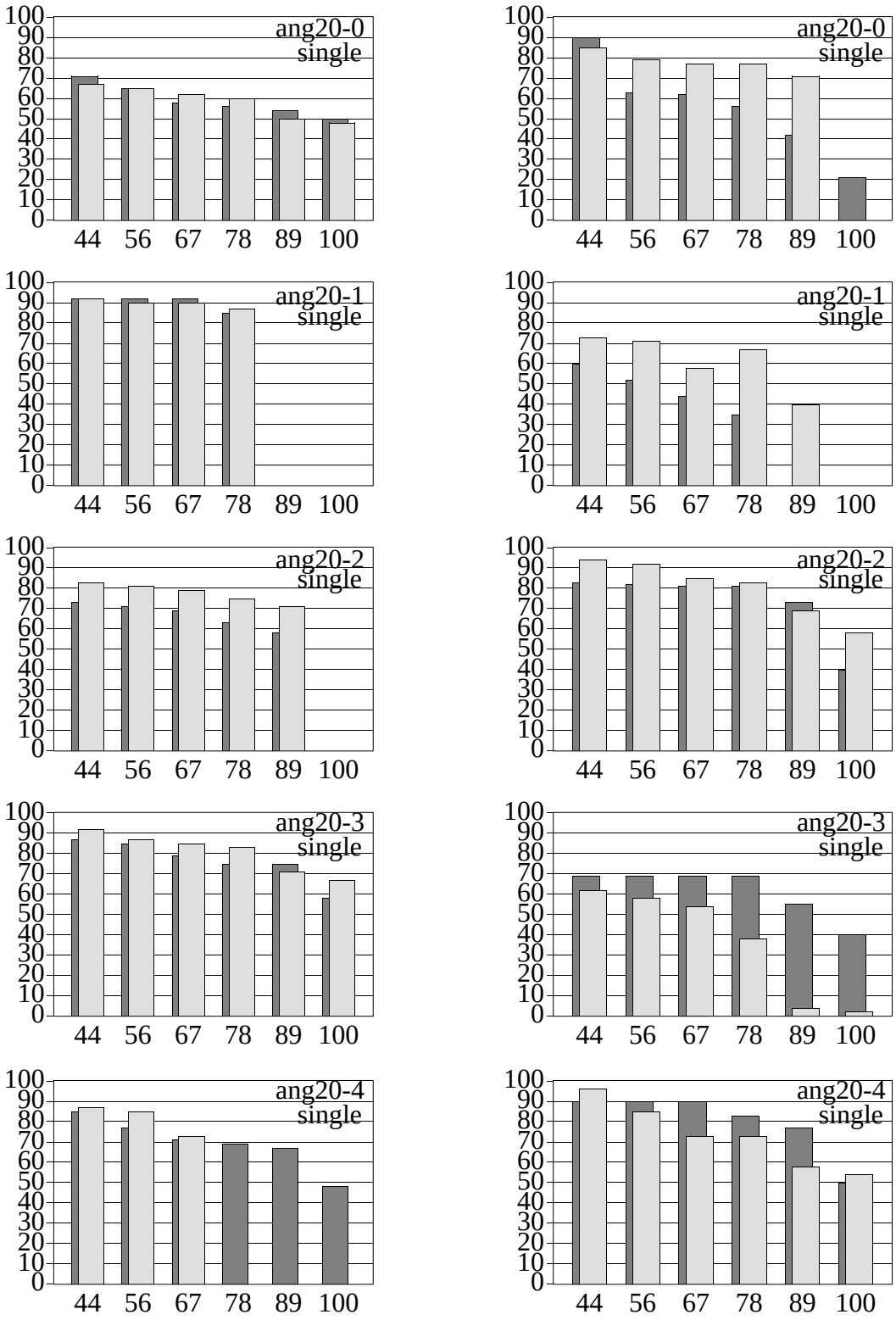

Figure 6: *The results from gradient descent learning over the major CSFN parameter—or 'cone-folding'—(light grey bars) compared to the results after MLP initialization (dark grey bars), depicted as sensitivity over specificity curves, for all five training sets (evaluated over the corresponding test sets).*

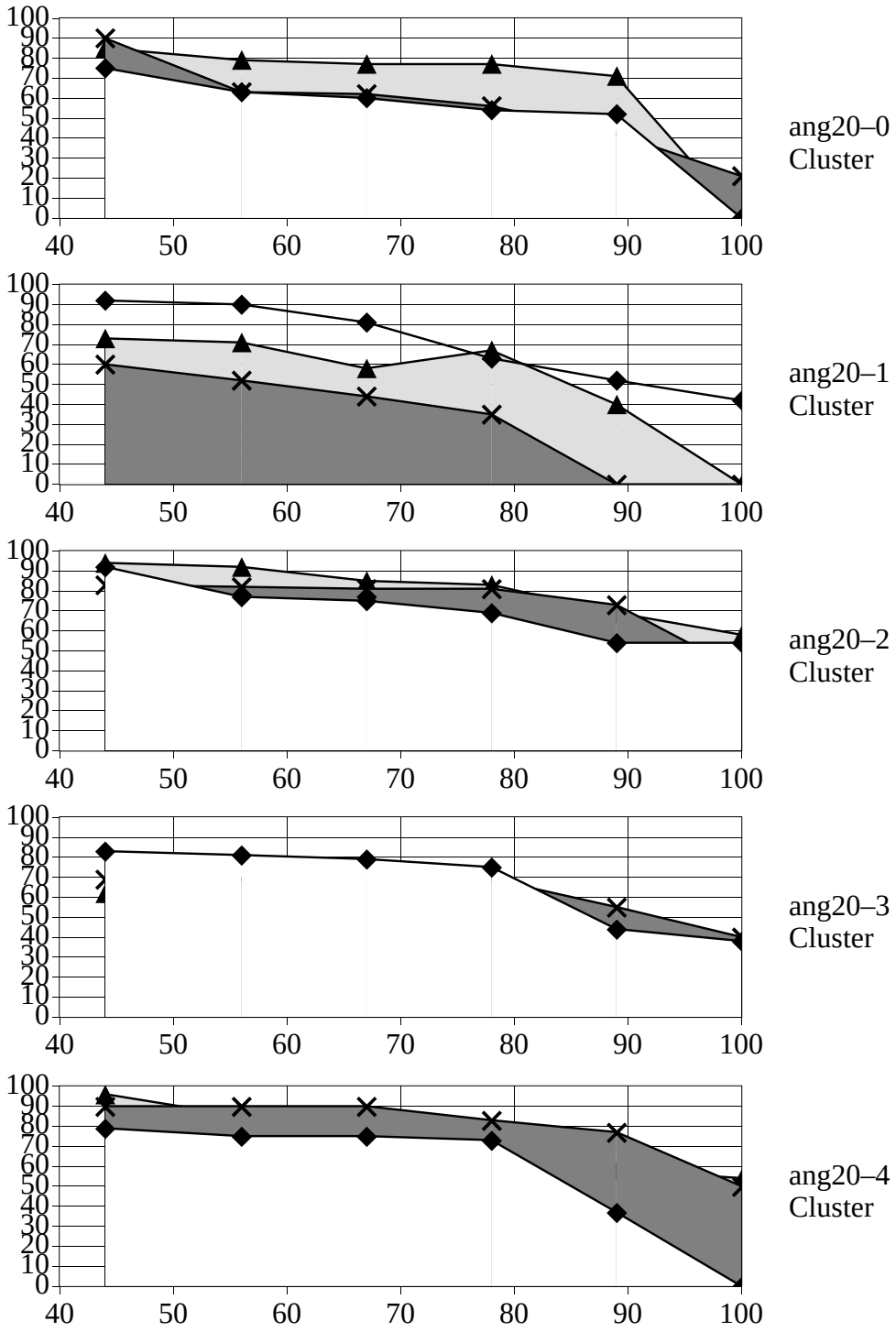

Figure 7: *The comparative results from backpropagation (white area, diamonds), MLP initialization plus delta rule (dark grey area, crosses) and cone-folding (light grey, triangles) for initialization with cluster centers, depicted as sensitivity over specificity curves, for all five training sets (evaluated over the corresponding test sets). Since for training set ang20–1 backpropagation leads to better results the areas are depicted in reverse.*

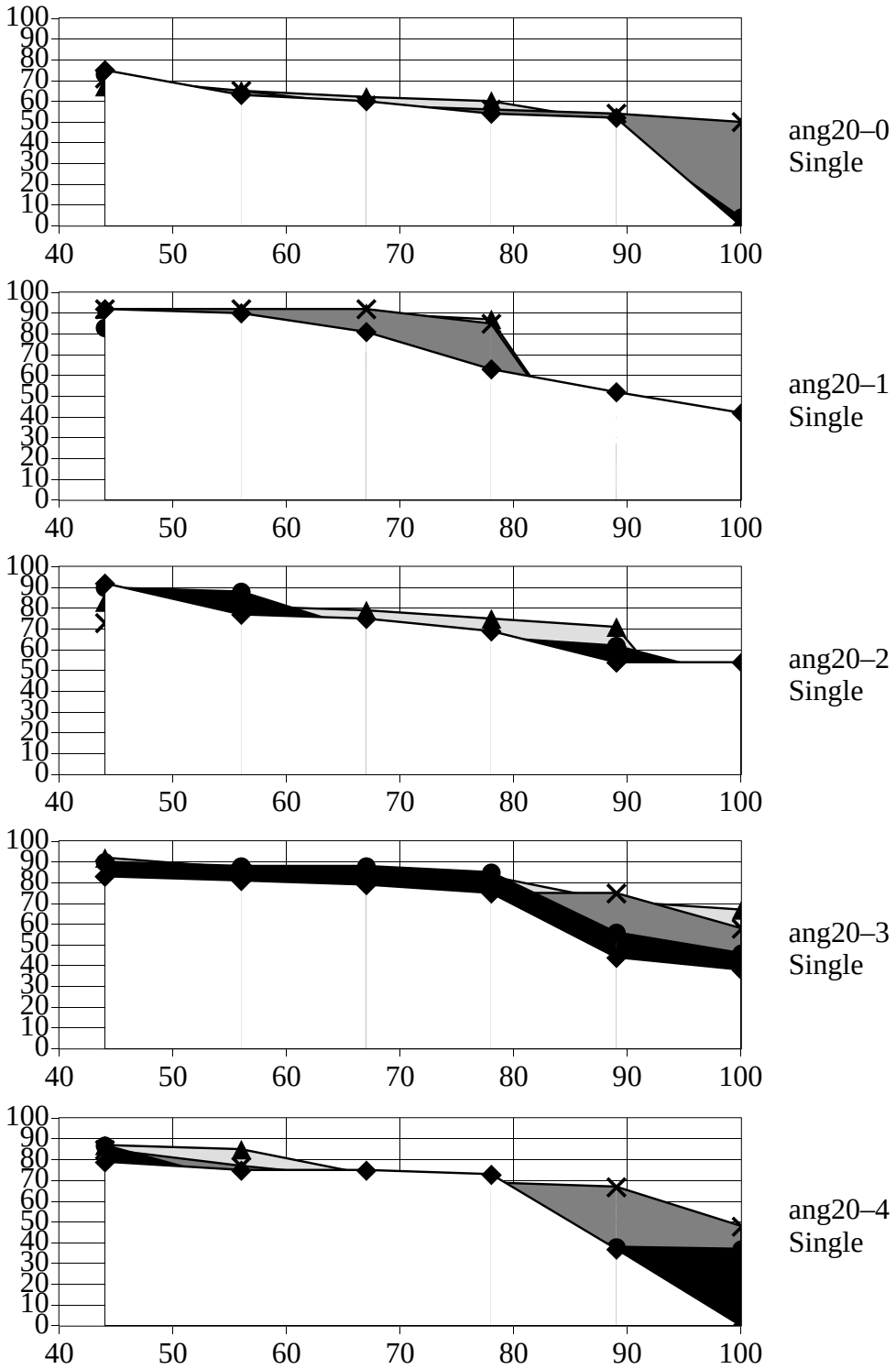

***Figure 8**: The comparative results from backpropagation (white area, diamonds), MLP initialization plus delta rule (dark grey area, crosses) and cone-folding (light grey, triangles) for initialization with single (randomly chosen) data point pairs, depicted as sensitivity over specificity curves, for all five training sets (evaluated over the corresponding test sets).*

The results confirm the initial observation that clean RBFNs overall might be inappropriate for this given task. The MLP initializations show consistently better results than RBFNs (with the exception of one training set). The further improvements of CSFNs show that the optimal solutions lie in between the two extremes MLP or RBFN (open or closed decision regions), but closer to the MLP type. Therefore this application demonstrates the ability of CSFNs to help solve the problem of appropriate decision regions automatically. From the comparative results, the following strategy for future similar applications can be derived: Use initialization to quickly scan over several training sets and identify the most promising candidates. Then use cone-folding to improve or fine-tune the results. Where neither of them leads to acceptable results, RBFNs or MLPs (in that order) can still be tried.

Coming back to the aspects of clinical decision making with supervised neural networks (section 3), the results suggest improved solutions for these kinds of application. In particular,

– MLP initialization and cone-folding, complementing conventional techinques like RBFNs, have proven capable of achieving useful results even in the case of extremely small training sets and a large variation between several randomly chosen ones.

– RBFNs and CSFNs have furthermore proven to be less sensitive to unbalanced training sets. As figs. 7 and 8 depict, large sensitivity can be paired with relatively large specificty, suggesting less bias toward the more numerous pathological cases.

When looking at the best results (e.g. training sets ang20-0 and 4 for the cluster method, fig.7) one can clearly see the essential improvements over the previous results from backpropagation (fig.1) Sensitivity considerably drops below 80 % only at specificity values of 80 or 85 %. The identification of patients suffering from significant CAD is clinically important to guide in proper patient management. In particular, the identification of a subgroup of patients that may benefit from interventional procedures such as percutaneous transluminal angioplasty or coronary bypass surgery to restore adequate nutritional blood flow to the myocardium is a primary diagnostic concern in clinical cardiology. Currently, coronary angiography is the method of choice to define precisely coronary anatomy prior to interventional procedures. However, the quest for noninvasive techniques that may permit to select accurately patients at risk is still a major research topic. While the prototype of the neural network as utilized in a previous study did not meet the strict requirements of diagnostic accuracy to be of clinical utility, these new techniques presented in the present study may overcome some limitations so that applicability in a clinical setting may ultimately be achieved. However, further research especially with respect to different scintigraphic techniques such as SPECT imaging and an extension of the case data base is needed before clinical utility can be reassessed.

9 Conclusion

In this paper we have presented results from experiments with several different types of feedforward neural network, applied to the problem of detecting coronary artery disease (CAD)

in planar thallium scintigrams, while using readings from angiography as the reference. By introducing the novel technique of initializing multilayer perceptrons (MLP) and the novel conic section function network (CSFN) we have succeeded in improving the results of this particular task so as to make it more clinically useful.

The experiments have shown that MLPs with backpropagation can be among the poorest methods in the presented collection of learning rules. However, we do not want to simply bash that kind of popular neural network model—many discussions with similar criticism have sprung up in neural network literature since backpropagation has become popular—but wanted to contribute to a more general discussion around practical applications of neural networks in medicine (and in other domains). We have discussed two extremes of a spectrum of neural network classifiers with respect to their propagation rules (and decision borders), and have introduced and applied a novel neural network type that can place itself on any point between the two extremes. We have also pointed out the use of some quick methods of setting weights in a network before any time-consuming gradient-descent learning is necessary. We have done all this in the light of some peculiar aspects of clinical diagnostic tasks—such as limited and unbalanced training sets, or the need for large evaluation sets to extract confidence values. In this sense, this work can be viewed as a contribution to improved medical applications of neural networks in the future.

Acknowledgments

This work has been done as part of the EC-funded ESPRIT-II project 5433 “NEUFODI—Neural Networks for Forecasting and Diagnosis Applications.” The Austrian contribution to this project has been funded by the Austrian Industrial Research Promotion Fund under grant no. 2/282. I wish to express my gratefulness to Prof. Trappl for encouragment and inspiration. I further thank my colleagues, namely Herbert Wiklicky, Erich Prem, Claudia Ulbricht and Guenter Linhart, for fruitful discussions and many valuable hints for this research. Dr. Heinz Sochor, head of the Working Group of Cardiovascular Nuclear Medicine at the Dept. of Cardiology provided ample support for this research project. The clinical data processing for this work was funded by Informatica Ges.m.b.H.

References

[1] S.M. Botros, C.G. Atkeson: Generalization Properties of Radial Basis Functions, in Lippmann R.P., et al.(eds.), Advances in Neural Information Processing 3, Morgan Kaufmann, San Mateo, CA, pp.707− 713, 1991.

[2] D.S. Broomhead, D. Lowe: Multivariable Functional Interpolation and Adaptive Networks, Complex Systems, 2,321−355, 1988.

[3] D. Coffey, G. Banks: A Connectionist Visual Field Analyzer, in Kingsland L.C. (ed.): Symposium on Computer Applications in Medical Care, IEEE Computer Society Press, 1989.

[4] T. Denoeux, R. Lengelle: Initializing Backpropagation Networks with Prototypes, Neural Networks, (to appear), 1993.

[5] G. Dorffner: EuclidNet − A Multilayer Neural Network using the Euclidian Distance as Propagation Rule, To appear in: Proceedings of the International Conference on Artificial Neural Networks (ICANN), Brighton, 1992.

[6] G. Dorffner, P. Rappelsberger, A. Flexer: Using Selforganizing Feature Maps to Classify EEG Coherence Maps, to appear in: Proceedings of the International Conf. on Artificial Neural Networks (ICANN), Amsterdam; Springer Verlag, 1993.

[7] G. Dorffner: Conic Section Function Networks − A Unification of MLPs and RBFNs, forthcoming (submitted for publication, available as technical report from the first author).

[8] H.Fujita, T. Katafuchi, T.Uehara, T. Nishimura: Application of Artificial Neural Network to Computer−Aided Diagnosis of Coronary Artery Disease in Myocardial SPECT Bull's−eye Images, Journal of Nuclear Medicine, 33(2)272−276, 1992.

[9] L. Hutton, V. Sigillito, Johannes R.S.: An Interaction between Auxiliary Knowledge and Hidden Nodes on Time to Convergence, in Kingsland L.C. (ed.): Symposium on Computer Applications in Medical Care, IEEE Computer Society Press, 1989.

[10] T. Kohonen: Self−Organization and Associative Memory, Springer-Verlag, Berlin, Heidelberg, New York, 1984.

[11] A. Manduca, P. Christy, R. Ehman: Neural Network Diagnosis of Avascular Necrosis from Magnetic Resonance Images, in Moody J.E., Hanson S.J., Lippmann R.P. (eds.): Advances in Neural Information Processing Systems 4, Morgan Kaufmann, 1992.

[12] M. Maruyama, F. Girosi, T. Poggio: A Connection between GRBF and MLP, Massachusetts Institute of Technology, Cambridge, MA, AI Memo No. 1291, 1992.

[13] C.G. Myron: Thallium−201 Two−dimensional Imaging. In:M.J.Gelfand, S.R.Thomas (Eds.), Effective Use of Computers in Nuclear Medicine, Mc Graw−Hill, New York, 1988, pp 75−108.

[14] J.C. Platt: Learning by Combining Memorization and Gradient Descent, in Lippmann R.P., et al.(eds.), Advances in Neural Information Processing 3, Morgan Kaufmann, San Mateo, CA, pp.714− 720, 1991.

[15] G. Porenta, G. Dorffner, J. Schedlmayer, H. Sochor: Parallel Distributed Processing as a Decision Support Approach in the Analysis of Thallium−201 Scinitgrams. In: Proc.Computers in Cardiology 1988, Washington D.C., IEEE.

[16] G. Porenta, G. Dorffner, S. Kundrat, P. Petta, J. Duit, H. Sochor: Computer Interpretation of Planar Thallium-201 Dipyridamole Stress-Redistribution Scintigrams using Artificial Neurla Networks, forthcoming (submitted for publication, available as technical report from the authors).

[17] P. Rappelsberger, G. Dorffner, A. Flexer: Classification of EEG coherence maps of cognitive processes; to appear in: Löffler (ed.): Central Nervous System Monitoring, Verlag Wilhelm Maudrich Wien-München-Bern, 1993.

[18] C. Rosenberg, J. Erel, H. Atlan: A Neural Network that Learns to Interpret Myocardial Planar Thallium Scintigrams, Neural Computation 5(3), 1993.

[19] D.E. Rumelhart, G.E. Hinton, R.J. Williams: Learning Internal Representations by Error Propagation, ICS Report 8506, also in: Rumelhart,McClelland(eds), Parallel Distributed Processing, Vol I, MIT Press, 1986.

[20] W. Schoner, M. Berger, G. Holzmueller, A. Neiss, H. Ulmer: Diagnosing functional disorders of the cervical spine using backpropagation networks − preliminary results, in Lun K.C. (ed.) MEDINFO 92, Elsevier Science Publishers (North Holland) 1992.

[21] P.K. Simpson: Artificial Neural Systems, Foundations, Paradigms, Applications, and Implementations, Pergamon Press, 1990.

[21] S.G. Smyth: Designining Multilayer Perceptrons from Nearest-Neighbor Systems, IEEE Transactions on Neural Networks, 3(2)329−333, 1992.

[23] B. Widrow, M.E. Hoff: Adaptive switching circuits, IRE WESCON Convention Record, New York:IRE, pp.96−104, 1960.