A Neural Network for Tornado Diagnosis: Managing Local Minima

Caren Marzban *
National Severe Storms Laboratory, Norman, OK 73069
Cooperative Institute for Mesoscale and Meteorological Studies, and
Department of Physics, University of Oklahoma, Norman, OK 73019

Abstract

There exist radar-based algorithms designed to detect circulations in the atmosphere. Not all detected circulations, however, are associated with tornados on the ground. Outlined herein, is the development of a multi-layered perceptron designed to classify the two types of circulations - nontornadic and tornadic - based on various attributes of the circulations. Special emphasis is placed on the role of local minima in determining the optimal architecture via bootstrapping, and on the performance of the network in terms of probabilistic measures.

1 Introduction

A great deal of effort is required to determine the optimal architecture of a Multi-layered Perceptron (MLP) designed to perform a specific task. That issue is important to consider because the nonlinearities inherent in a MLP can allow it to overfit data, leading to poor generalization.

The nonlinearity of a MLP is determined primarily by two quantities - the number of hidden nodes, and the magnitude of the weights. This can be seen as follows: If the magnitude of the weights is restricted to be "small" (relative to some scale), then most activation functions (e.g. logistic, tanh, etc.) are linear in the range of allowed weights.

*E-mail: marzban@nssl.noaa.gov; http://www.nhn.ou.edu/~marzban

As a result, regardless of the number of hidden units, the MLP will represent nothing more than linear regression or a linear classifier. On the other hand, if the weights are allowed to take "large" values, then most activation functions become highly nonlinear, and consequently, even a small number of hidden nodes can render the MLP highly nonlinear. Such nonlinearities may allow the MLP to fit features in the data that are driven by noise or statistical fluctuations. As such, an MLP can overfit the (training) data by exhibiting superb performance on it, but have no generalization capability.

From a statistical point of view, the problem of the optimal architecture is one of "model selection," and a variety of methods exist for that purpose. Certain proposals have been made that even avoid the entire issue of the optimal architecture (Buntine and Weigend, 1991; MacKay, 1996; Neal 1996), and one of these was examined in a meteorological context in (Marzban, 1998). However, it turns out that in these methods the question of optimal architecture is replaced by the difficult task of evaluating the distribution of certain (hyper) parameters. Although such methods are most likely the most promising methods for model selection, their implementation is quite involved and technically demanding. Simpler alternatives include prunning techniques (Mao, Mohiuddin, and Jain 1994; Hassibi and Stork 1993; Le Cun, Denker, and Henderson 1990) wherein one begins with a large network and systematically removes some of the less important weights. These methods make certain restrictive approximations, such as diagonal hessian, quadratic error functions, and measures of weight importance, that must be handled with care. The simplest approach which also avoids all such difficulties, if the problem is not too computationally expensive (e.g., if the number of weights is not prohibitively large), is the brute force approach of training MLPs with $0, 1, 2, ...H$ hidden nodes (one one layer) and selecting the one that meets the performance criterion. The MLP for tornado prediction as considered in this article is sufficiently small for this brute force method to be viable.

Many of these approaches to model selection involve re-sampling of data. A well-known and relatively simple example is bootstrapping (Efron and Tibshirani 1993). In its simplest form, one repeatedly trains a MLP with subsamples of the data - called the

training sets. The architecture with the lowest average on the remaining (validation) sets is asserted to be the optimal architecture. Of course, the performance of the resulting MLP on both the training and the validation sets is an optimistically biased estimate of generalization performance, and for this reason, yet another data set - a test set - is required for estimating the generalization performance of the MLP.

An implicit assumption in all re-sampling-based procedures is that in each re-sampling trial the MLP converges to a global minimum, or at least to a "deep" local minimum. Although there exist a variety of techniques for finding deeper, and deeper, local minima, it is impossible to prove that the deepest minimum found is in fact a global minimum; in practice, given sufficient time, a training algorithm can usually find a deeper minimum. As such, in the context of model selection, there are (at least) two sources of variability - one due to finite sampling at each trial, and one due to local minima. Therefore, it is important to examine model selection with particular attention paid to both.

An arena in which neural networks have had significant success in improving predictions is that of tornado prediction (Marzban and Stumpf 1996, 1998a,b; Marzban, Paik, and Stumpf 1997). Part of the success is due to MLPs' ability to represent nonlinear relations; however, a large portion of the success is due the development of better means for identifying circulations in the atmosphere (Stumpf et al. 1998, Mitchell et al. 1998). In particular, Doppler radar provides the ability to identify such circulations, because it can identify regions of the atmosphere that move towards or away from the radar. Indeed, it is possible to identify adjacent regions with opposing movements satisfying certain conditions that in essence define a circulation. Figure 1 shows an instance from May 3, 1999, over Oklahoma City, Oklahoma, U.S.A., during which two such circulations were detected. However, not all circulations satisfy the necessary atmospheric conditions to form tornados.

The aim of this article is to outline the development of a MLP that assesses the probability that a given circulation may be tornadic. The outline places special em-

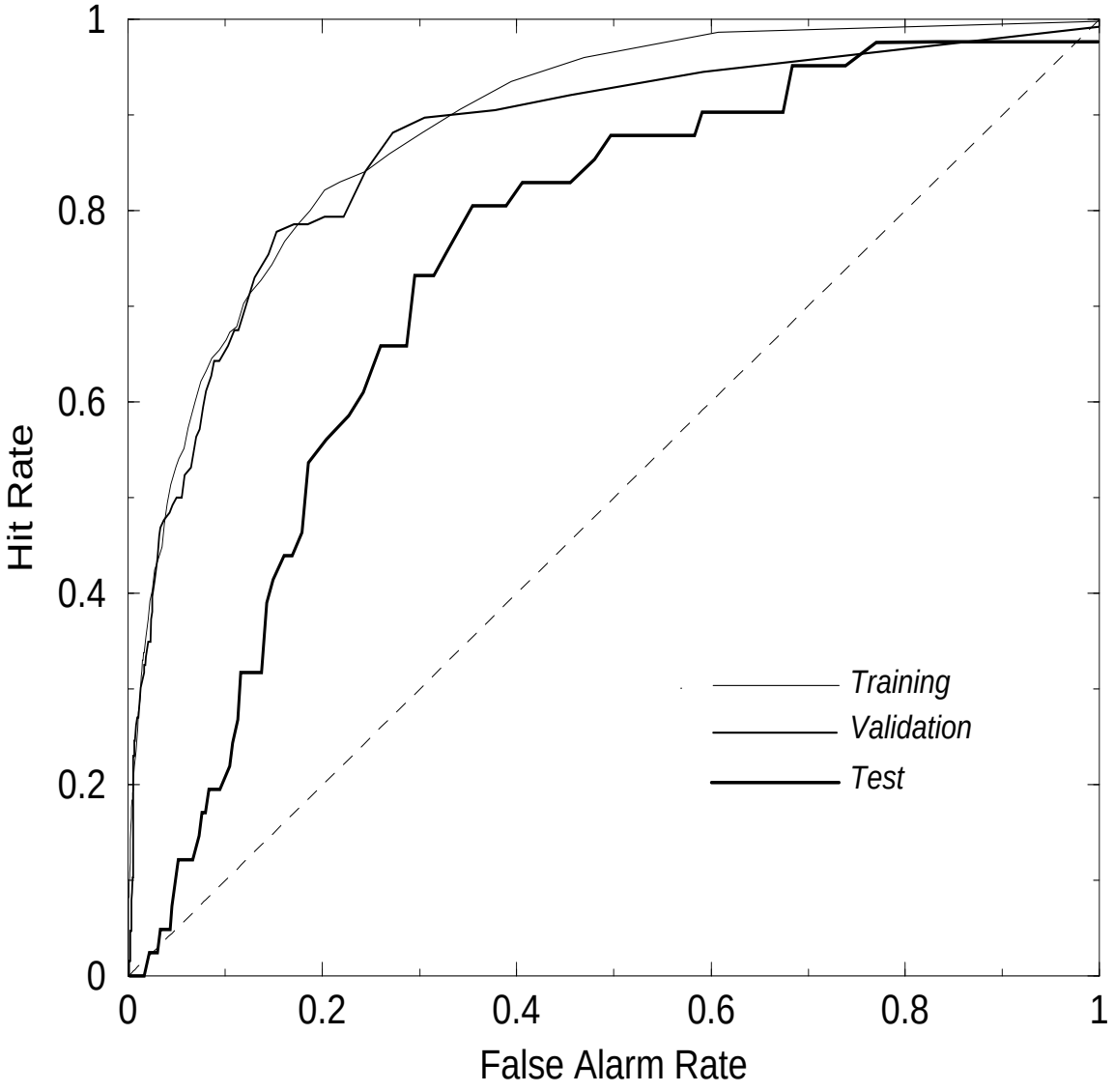

Figure 1: An example of the output of a circulation detection algorithm, with two detected circulations shown as circles. Different shades of gray correspond to different directions of air movement as measured by Doppler radar.

phasis on the effect of local minima in determining the optimal number of hidden nodes via re-sampling methods, in particular, bootstrapping. The MLP produces posterior probabilities, and so, its performance is assessed in terms of a number of categorical and probabilistic measures.

2 Data

The National Severe Storms Laboratory has developed two algorithms for detecting circulations in the atmosphere based on information constructed from Doppler radar (Stumpf et al. 1998, Mitchell et al. 1998). Another information that is also available is the existence/nonexistence of actual tornados on the ground, at a given point in space and time. This situation is suitable for the development of a MLP as a statistical model of the relationship between the various attributes of a detected circulation and the existence of tornados. The two algorithms differ in many respects, the most important of which is that one is designed for detecting larger, storm-scale, circulations (Stumpf et al. 1998), while the other is designed to detect smaller, but more-intense circulations (Mitchell et al. 1998); the physics underlying the two circulations is believed to be different. A set of MLPs have already been developed for the former (Marzban and Stumpf 1996, 1998a,b; Marzban, Paik, and Stumpf 1997). In this article, the development of a MLP for the diagnosis of tornados in the latter algorithm is outlined.

A number of statistical properties of the circulation attributes have been presented in (Marzban, Mitchell, Stumpf 1999). Here, we mention that the class-conditional frequency distributions are all highly nongaussian, and non-homoelastic (i.e., the class-conditional variances are unequal), rendering the problem suitable for a MLP. The 21 attributes and a brief description of each are as follows: 1) The height of the lowest point of a circulation; 2) The vertical extension of a circulation; 3) Low-level gate-to-gate velocity difference [1]; 4) maximum gate-to-gate velocity difference; 5) height

[1] Gate-to-gate refers to the difference in radar velocity bins which are azimuthally adjacent and constant in range.

of maximum gate-to-gate velocity difference; 6) low-level shear; 7) maximum shear; 8) height of maximum shear; 9) altitude weighted, vertically integrated low-altitude gate-to-gate velocity difference; 10-18) the corresponding time-trends; 19) convective available potential energy; 20) storm-relative environmental helicity; and 21) range from the radar. Although the range itself is not expected to be a predictor, it is expected to affect the values of the other attributes. For further details see (Mitchell, et al. 1998).

The data set examined contains 29 days of storm activity which constitutes $N_0 = 5348$ nontornadic and $N_1 = 512$ tornadic circulations. Whether or not a circulation is labeled as tornadic (or not) is based on ground-based observations. As this project neared completion, an independent data set containing 6 additional days of storm activity (943 circulations) became available, allowing for an unbiased estimation of the MLP's performance.

Some pre-processing of the data is necessary prior to the MLP development. All 21 attributes are transformed into z-scores (i.e., mean of zero and a standard deviation of 1). Outliers for a given attribute are identified, and then removed, by a visual examination of the estimated class-conditional frequency distributions. The preprocessing leaves 5791 circulations which are randomly divided into 4344 cases (about 2/3) for training and 1447 cases for validation. A similar preprocessing of the test set leads to 941 circulations. A number of other linear transformations of the inputs were also examined - principal components analysis, and "whitening" (Bishop 1996); however, the performance of these MLPs was not found to be significantly different from the one with only z-transformed inputs.

3 The Method

The input nodes, numbering 21, were the various attributes mentioned above. Although, collinearity among the inputs is not a serious detriment to the performance of

MLPs designed only for prediction/detection (in contrast to an MLP designed for rule extraction), the highly collinear inputs (numbering 2 pairs) were initially excluded from the MLP. However, it was found that the performance of the MLPs with the collinear inputs included was statistically equivalent to that of the MLP with no collinear inputs. For this reason (and to be conservative) all 21 attributes were eventually employed as inputs.

The output nodes actually numbered 2 - one for tornado (target 0 or 1), and the other for damaging wind (target 0 or 1). A single output node corresponding to the former would have sufficed for tornado prediction; however, it has been conjectured that the existence of a second output node representing a quantity that is closely related to the first output node may actually aid the MLP in surmising the true underlying relations (Mitchell 1996). No attempt was made to substantiate this conjecture; however, the second output node was included since its presence is not expected to adversely affect the performance of the MLP in predicting tornados.

The error function being minimized during training was cross-entropy

$$S = -\frac{1}{2N} \sum_{i}^{N} \sum_{j}^{2} \left[t_i^j \log \frac{y_i^j}{t_i^j} + (1 - t_i^j) \log \frac{1 - y_i^j}{1 - t_i^j} \right], \tag{1}$$

where t_i^j are the target values, and y_i^j are the outputs for the i^{th} case and the j^{th} output node. It has been shown (Richard and Lippmann 1991) that the minimization of this error function allows for the interpretation of the output nodes as posterior probabilities, if the activation function of all the layers is the logistic function $f(x) = 1/[1 + exp(-x)]$, and if the output nodes are coded as 0 or 1 for the two classes. Consequently, the first (second) output node of the present MLP is the probability of tornado (damaging wind), given the values of the 21 inputs.

A sequence of networks with $H = 0, 2, 4, 8$, and 16, hidden nodes (on one layer) were trained and validated on 4 subsamples of the data. The training algorithm was the conjugate gradient method. When a local minimum was reached, simulated annealing was employed to attempt an escape. If a better minimum was found, then conjugate

gradient was employed again, otherwise the entire training phase resumed from a new random set of initial weights. This procedure is due to Masters (1993). The total number of times that this complete reinitialization was allowed was 43, and so 43 local minima were visited. The one with the lowest value of cross-entropy over the training set can be considered to be the "global" minimum within the set of local minima visited. Of course, there is no guarantee that the deepest local minimum is anywhere as deep as any global minimum. However, a global minimum is not even necessary for a well-performing MLP; a sufficiently deep minimum will usually suffice.

To get a grasp on "sufficiently deep," it is instructive to examine the frequency distribution of all the local minima visited. For example if the frequency distribution of the local minima (i.e., that of the training errors) is generally bell-shaped, then the "width" of the distribution sets a scale for the range of possible minima; for a given training algorithm, error values far beyond the extreme ends of the distribution will be rare, and difficult to reach. So, if the true global minimum exists somewhere beyond these extreme regions, then that training algorithm will most likely not find it. In other words, from a practical point of view, any of the minima in the lower-end of the distribution can be considered as a "global" minimum (given the data and the training algorithm).

Furthermore, the frequency distribution of the local minima can make evident the difference between the deepest local minimum and the most visited local minimum (i.e., the mode of the distribution). For example, if the distribution is highly peaked and narrow, then even though the most-visited local minimum is a strong attractor, the small difference between its error function and that of the deepest minimum makes it less important to find deeper minima. On the other hand, if the distribution is relatively flat, with a wide range, then it is important to assure that a MLP occupies the deepest possible local minimum.

The frequency distribution of the errors at the local minima can serve one other function. Recall that although a "local minimum" refers to the value of the error func-

tion evaluated over the training set. Given the weights of the MLP at a local minimum, one can also compute the error function over the validation set. As such, every local minimum is associated with two values of the error function - of the training set, and of the validation set. Tracking the movement and the width of these distributions (as compared to tracking just a local minimum), as a function of the number of hidden nodes (H), offers a "safer" way of arriving at the optimal number of hidden nodes, H_c. This can be seen as follows. As H increases, the distribution of the training errors will generally shift to lower values; while that of the validation errors will initially shift to lower values, at H_c it will begin to shift to higher values, indicating that the training set is being overfit.

Finally, the correlation between the points in the two distributions can also anticipate overfitting. As will be shown below, for $H < H_c$, lower training errors are generally associated with lower validation errors at the local minima. On the other hand, for $H > H_c$, the correlation between the two errors reverses direction. Therefore, by monitoring the correlation between the training and the validation errors over the local minima, one can decide on "how far" H is from H_c. All of this can be made self-evident in a scatter-plot of the training and validation errors over the local minima (a diagram which we shall refer to as a "tv-diagram.")

4 Performance

There are at least two facets of performance that must be distinguished. An MLP can be treated as a classifier in which case its performance is best expressed in terms of a contingency table (i.e., confusion matrix), and scalar measures derived therefrom. This is the manner in which most MLPs are treated. On the other hand, it is possible to treat an MLP as a device for modeling posterior probability (Richard and Lippman 1991). A concise reference to probabilistic forecasting is (Dawid 1986), but a great deal of the subject developed in its meteorological guise is due to Murphy and Winkler (1987, 1992). In the meteorological arena preference is given to probabilistic forecasts,

and the problem of tornado prediction is no exception, but both facets of performance will be considered below.

To form a contingency table, one must reduce probabilities into a dichotomous quantity by introducing a threshold on the probabilities; any probability higher (lower) than this threshold would be classified as a tornado (non-tornado). The 4 elements of the 2×2 table are C_1 (C_4), the number of correctly classified nontornados (tornados), and C_2 (C_3), the number of incorrectly classified nontornados (tornados). Although the table has 4 elements, there are only two degree of freedom if the class-conditional sample sizes are fixed, because $N_0 = C_1 + C_2$, and $N_1 = C_3 + C_4$. The table, in turn, can be reduced to a host of scalar measures of performance, but in order to preclude any loss of information (due to the reduction from two to one degree of freedom) two scalar measures should be considered. A convention is to consider the hit rate and the false alarm ratio:

$$\text{Hit Rate} = \frac{C_4}{C_3 + C_4} \quad , \quad \text{False Alarm Rate} = \frac{C_2}{C_1 + C_2} \quad .$$

A simultaneous representation of these two measures is conveniently given in a ROC diagram (Masters 1993) which is simply a parametric plot of the hit rate versus the false alarm ratio, as the probability threshold varies from 0 to 1. It is easy to show that a classifier with no ability to discriminate between two classes yields a diagonal line of slope one and intercept zero; otherwise, the ROC curve lies in the region above the diagonal line. The area under the curve is often taken as a single (scalar) measure of the classifier's performance, and so, a perfect classifier would have an area of 1 under its ROC curve, while a random classifier would have an area of 0.5 (i.e., the area under the diagonal line). The virtue of the ROC diagram is in its 2-dimensional expression of performance, thereby maintaining the full dimensionality of the problem. The expression of performance in terms of ROC diagrams has the additional advantage that no specific value of the threshold must be specified in determining the optimal architecture of the MLP. A specific choice of the threshold calls for knowledge of the costs of misclassification which are user dependent.

A 2-dimensional representation of the quality of the produced forecasts is their class-conditional distribution, i.e., a histogram of the probabilities for tornadic and non-tornadic cases, separately. Such a histogram can display not only to what extent the classifier discriminates between the two classes, but also how it treats each class, individually. Since, a perfect classifier would have no overlap between the class-conditional histograms, the area of the overlap could serve as a scalar measure of performance; however, it should be acknowledged that such a reduction again neglects the 2-dimensional nature of the problem.

Another facet of probabilistic forecasts is their reliability, a concept best displayed in a reliability diagram, where the observed frequency of tornados is plotted against the MLP-produced probabilities. Reliability refers to the criterion that if the MLP produces, say, a 20% probability for some circulations, then 20% of such circulations should be truly tornadic. Then a perfectly reliable set of probabilities would yield a diagonal line of slope 1. Again, it is possible to compute a scalar quantity (e.g., the L2 distance between the reliability plot and the diagonal line) that would distill the diagram into a single measure, as long as one acknowledges that such a reduction again leads to loss of information.

Richard and Lippman (1991) suggest that for consistency the average of the posterior probability over all cases, $\bar{p} = \frac{1}{N}\sum_{n=1}^{N} y(x^n)$, should approximate the class prior probability, $p = N_1/(N_0 + N_1)$. Then a measure like $D = p\log(p/\bar{p})$ would serve as another scalar measure of the quality (consistency) of the modeled posterior probabilities.

5 Results

As mentioned above, 43 local minima were visited during the training phase. Each MLP trapped in a local minimum is associated with two values of cross-entropy, one for the training set and another for the validation set.

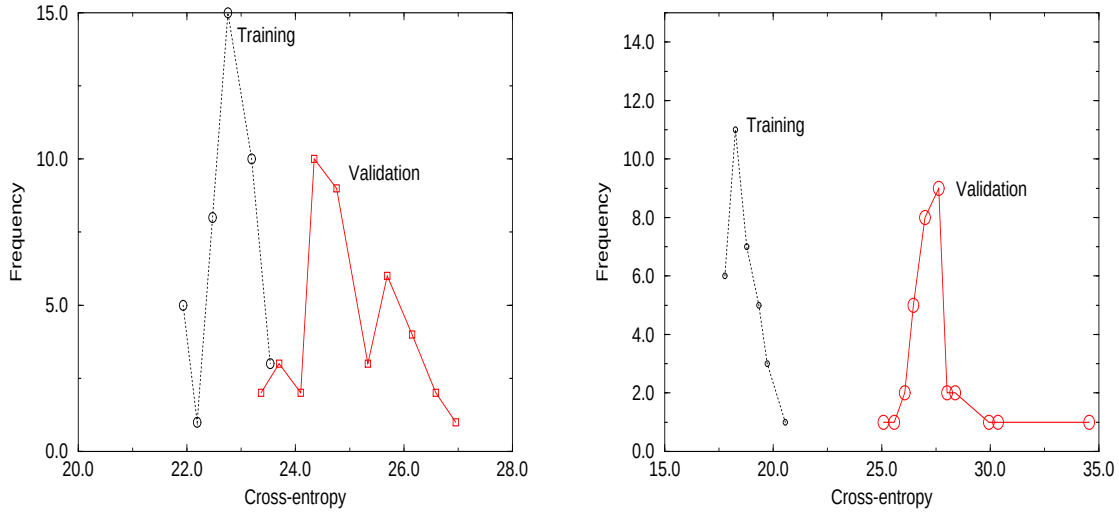

Figure 2: The distribution of 43 local minima of a MLP with $H = 4$ (left), and $H = 16$ (right), for one subsample training and validation sets.

Figure 2 shows the frequency plot of the 43 values of training and validation cross-entropy for MLPs with $H = 4$ and $H = 16$. It can be seen that the distributions are somewhat bell-shaped. Therefore, the global minimum, or more accurately the deepest local minimum is quite distinct from the most-visited local minimum. The latter is the local minimum which is most likely to be visited for a given bootstrap trial. Consequently, simple bootstrapping with no regard to whether or not the MLP is in a "global" minimum can lead to a false conclusion regarding the optimal number of hidden nodes. Also note that the training and validation curves are farther apart for the MLP with 16 hidden nodes than that with 4. This is a sign that the 16-hidden-node MLP is overfitting the training set, because lower training errors are associated with higher validation errors.

Figure 3 shows scatterplots of training and validation errors (i.e., tv-diagrams). It can be seen that the validation and the training errors for the 43 minima are correlated, as expected. For a network with 4 hidden nodes (Figure 3, left) the Pearson's correlation coefficient is $r = 0.53$. (The probability that a random sample would produce this value of r is 0.0004.) This implies that deeper minima of a MLP with $H = 4$ tend to

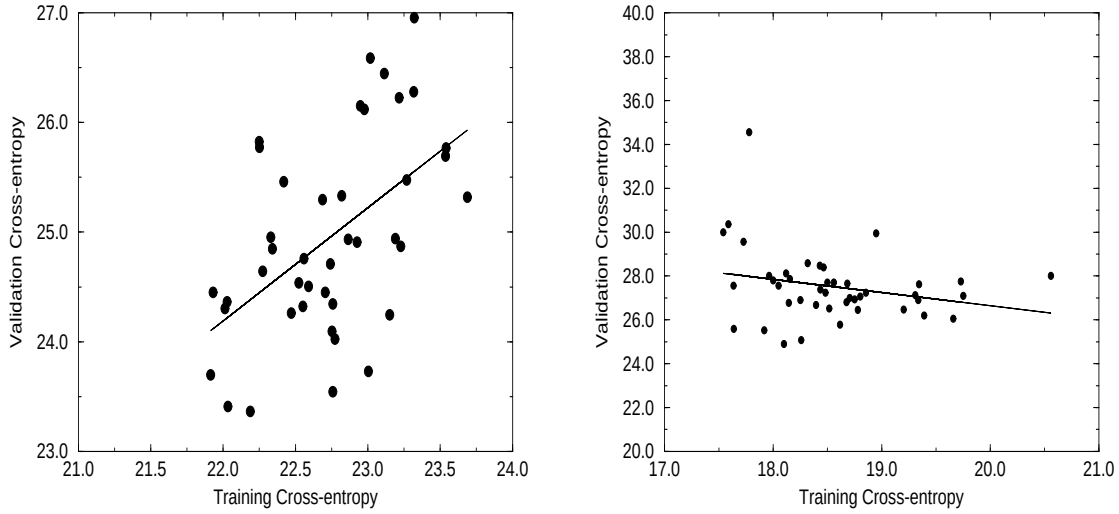

Figure 3: The tv-diagrams for the 43 local minima of a network with $H = 4$ (left) and $H = 16$ (right), for one subsample training and validation sets. Also shown, are the respective regression lines highlighting the correlation between the training and validation errors at the various local minima.

fit the training set better (i.e., without overfitting) than the shallower minima. This pattern, as well as the sign of r, reverses for larger number of hidden nodes, when the MLP overfits the training set. Figure 3 (right) shows the tv-diagram for an MLP with 16 hidden nodes; the value of r in this case is -0.25. This illustrates how the sign and the magnitude of r in a tv-diagram can anticipate the optimal number of hidden nodes.

Figure 4 shows the tv-diagram for MLPs with 0, 2, 4, 8, and 16 hidden nodes, for one bootstrap trial. It can be seen that for this subsample, the optimal number of hidden nodes is 4, because the deepest local minimum reached with 4 hidden nodes has the lowest validation error. Of course, more bootstrap subsamples must be examined in a similar way to confirm the optimal number of hidden nodes. Otherwise one has arrived at an MLP that, although does not overfit the training set, it does overfit the validation set. As such, it will have no generalization capability. In the present case, four bootstrap subsamples were examined, confirming that 4 is the optimal number of hidden nodes for this problem.

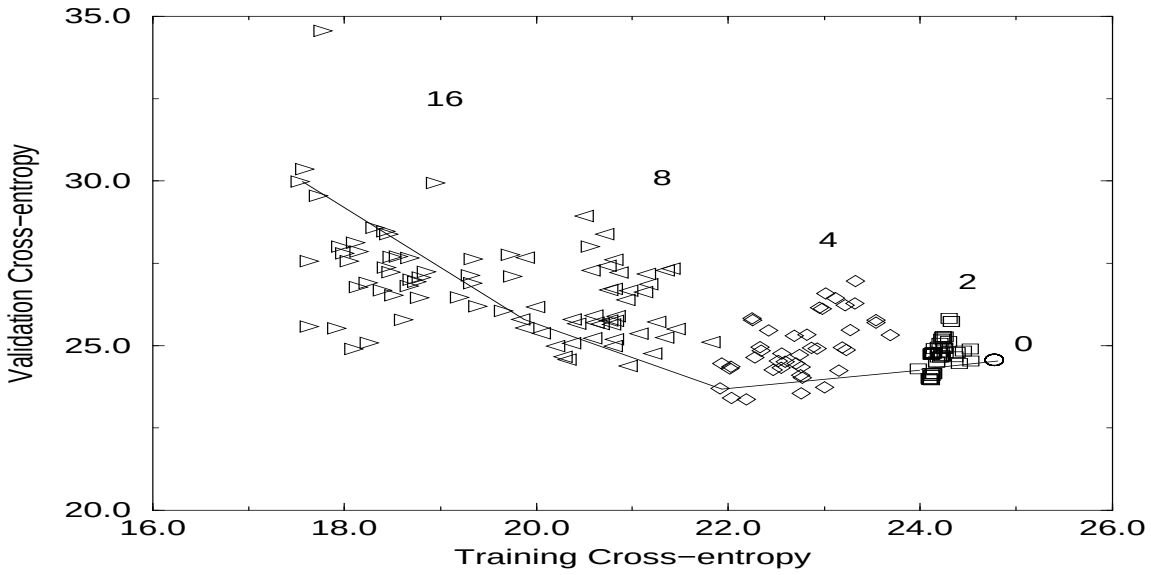

Figure 4: The tv-diagram for H=0 (circle), H=2 (square), H=4 (diamond), H=8 (right triangle), and H=16 (left triangle) hidden nodes, for one bootstrap trial. The solid line connects the "global" minimum (i.e., the point with the lowest training cross-entropy) for each H.

As discussed previously, the performance of the MLP can be considered both in terms of the forecast probabilities and the skill of the MLP as a 2-class classifier. Even though the latter calls for the introduction of an additional parameter (the probability threshold), it is the most common means of expressing performance. To that end, the ROC diagram provides a concise representation.

Figure 5 shows the ROC curves for the training, validation, and test sets as the threshold is varied from 0 to 1. Evidently, although the performance of the MLP on the training and validation sets is comparable, that of the test set is considerably lower. That is not surprising in that the former are optimistically biased, and only the latter represents an unbiased measure of generalization performance. The areas under the three curves are 0.89, 0.87, and 0.73, respectively. Recall that the area under the ROC curve is bounded by 1.0 and 0.5, corresponding to a perfect classifier, and a classifier with no ability to discriminate, respectively.

The performance of the MLP is most naturally expressed directly in terms of the predicted probabilities. This also avoids the introduction of the threshold. As

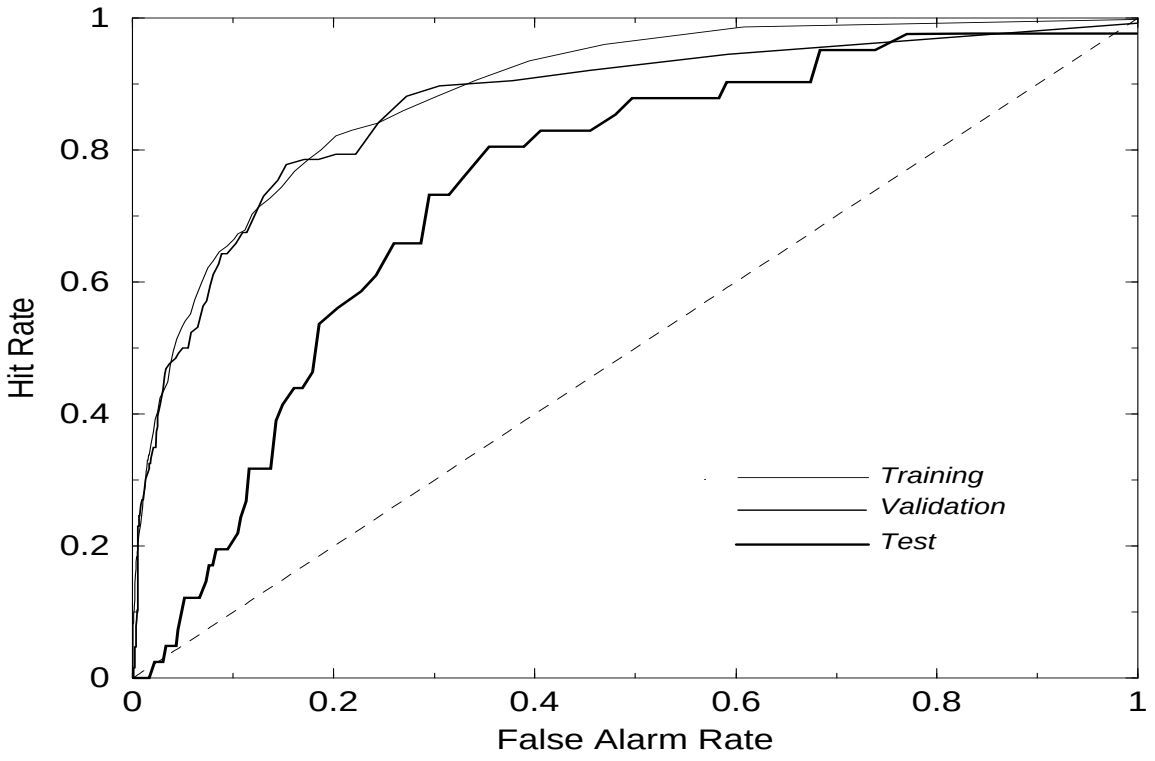

Figure 5: The ROC plots of the MLP for the training, validation, and test set.

mentioned previously, one diagram for displaying the quality of the probabilities is the class-conditional probability density (i.e., histogram) of the MLP output. Figure 6 shows that distribution for the nontornadic and tornadic circulations, separately, for the training (left), validation (middle), and test (right) sets. It can be seen that the MLP has a higher affinity for the identification of nontornadic circulations. As for the tornadic circulations, the MLP produces a relatively flat distribution. Loosely speaking, the discriminatory ability of the MLP derives primarily from the ability to identify the nontornadic circulations. The figure corresponding to the test set shows a somewhat stronger ability to identify tornadic circulations, but that is likely due to the small sample size of the test set (6 days) as compared to the size of the training and validation sets (29 days). A larger and more representative test set would be expected to yield a pattern more similar to that of the training and validation sets.

The area of the overlapping region of the two distributions constitutes a scalar measure of performance. 7%, and 7.2% of the training and the validation set, respec-

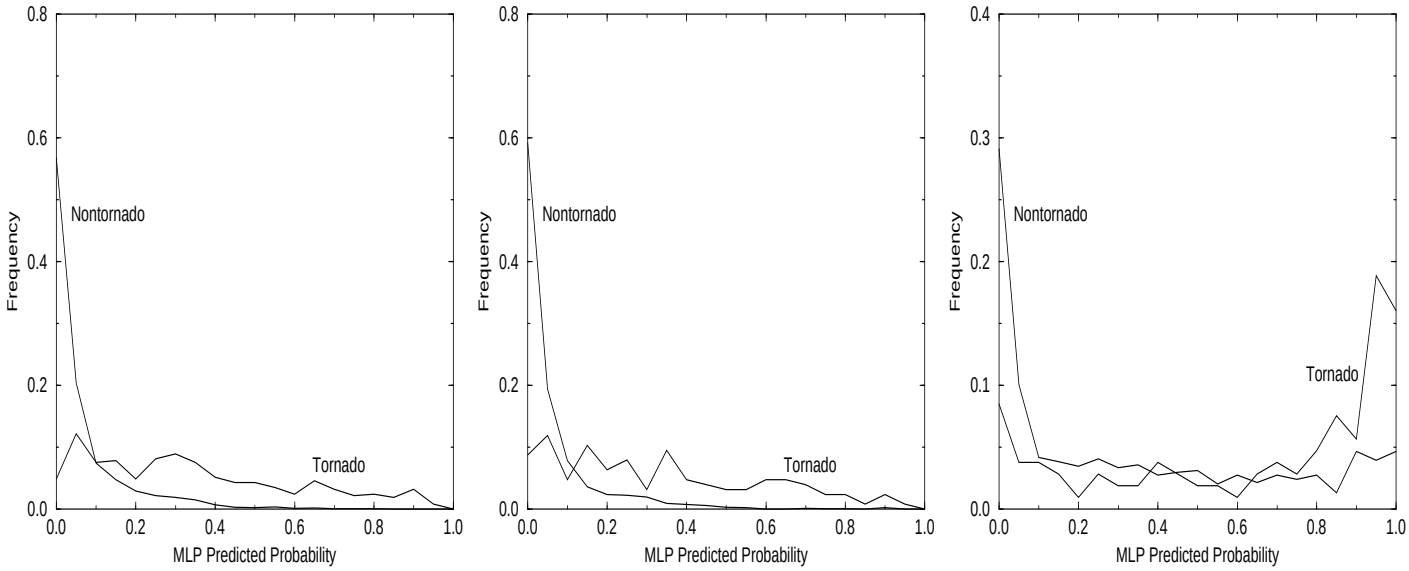

Figure 6: The class-conditional frequencies of the MLP predicted probabilities, for the training (left), validation (middle), and test (right) sets.

tively, fall into the overlap region. The same quantity for the test set is 11.2%. In other words, the overlapping areas are 0.070, 0.072, and 0.112 for the three sets. A perfect classifier would have an area of 0, and a random classifier would have an area of 1. Clearly, based on all three sets, the MLP's discrimination performance is on the perfect side.

Figure 7 shows the reliability diagrams as formed from the training (left), validation (middle), and the test set (right). It can be seen that when the MLP produces a probability of, say 15%, this is matched by an observed relative frequency of 15%. The same is true for other produced probabilities, with the exception of probabilities around 65% and 95%. The error bars on the plots represent 1 standard deviation of the bootstrap distribution. Therefore, even the exceptions can be considered reliable within, say, 2 standard deviations. Thus, the produced probabilities are statistically reliable. The test set shows a slightly different pattern; for instance, the MLP produces no probabilities higher than 85%, but such differences can again be traced to the test set's relatively small size.

The L2 norm between the reliability curve and the diagonal line of the diagram

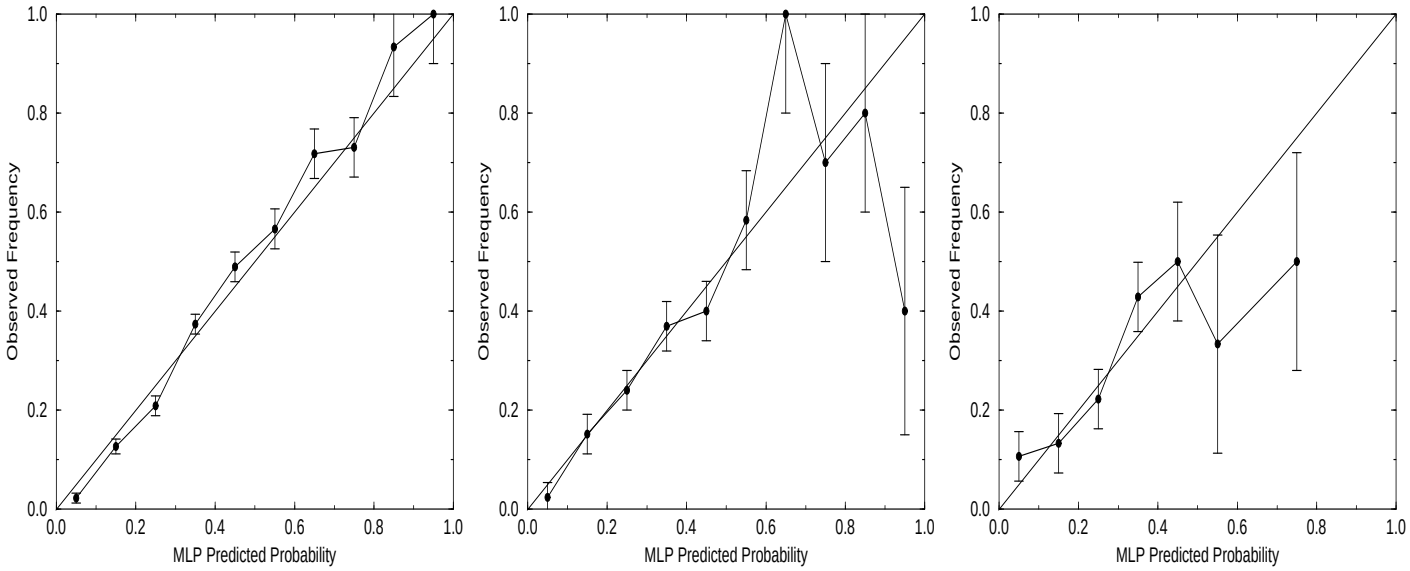

Figure 7: The reliability plots of an MLP with 4 hidden nodes for the training (left), validation (middle), and test (right) sets.

(i.e., the mean square error between the produced probabilities and the observed probabilities) serves as a scalar measure of reliability. These norms are 0.002, 0.04, and 0.05, for the training, validation, and test set, respectively. In theory, the most reliable norm is 0 and the least reliable norm is $0.33 = (1/3)$ [2]. So, according to this scalar measure, the predicted probabilities are well on the side of perfect reliability.

Finally, the scalar measure, $D = p \log(p/\bar{p})$, offers a test of consistency. The prior probability of tornadic circulation is $p = N_1/(N_0 + N_1) = 0.087$, and the values of $\bar{p}$ for the training, validation, and test sets are 0.085, 0.082, and .060, respectively. The corresponding values of the measure D are 0.002, 0.005, and .04. Recall that a perfect classifier has $p = \bar{p}$ (i.e., $D = 0$), and the upper bound to D is ∞. All of this confirms that the MLP is faithfully modeling the probabilities.

[2]Assuming that the worse-case MLP produces a reliability curve that has a slope of -1 and an intercept of 1 (i.e., the "other" diagonal), then the L2 distance between the two diagonals - 1/3 - is the worse-case norm.

6 Summary

The development of a multilayered perceptron for the prediction of tornados is outlined with special emphasis placed on the role, in bootstrapping, of the local minima of the error function. It is argued that if local minima are not taken into account, then the number of hidden nodes as computed in a bootstrap may be suboptimal. Finally, the performance of the optimal architecture is gauged in terms of categorical and probabilistic measures. It is shown that the optimum MLP produces a set of consistent, discriminatory, and reliable probabilities.

7 Acknowledgments

Gregory Stumpf and E. DeWayne Mitchell of the National Severe Storms Laboratory are acknowledged for providing the data without which a neural network for tornado prediction would not have existed.

8 References

1. Bishop, C. M., 1996: *Neural networks for pattern recognition.* Clarendon Press, Oxford, pp. 482.

2. Buntine, W. L., and Weigend, A. S. 1991: Bayesian back-propagation, Complex Systems, **5**, 603-643.

3. Dawid, A. P. 1986: Probability Forecasting. In Encyclopedia of Statistical Sciences, eds. S. Kotz, N. L. Johnson, and C. B. Read, pp. 210-218. New York, Wiley.

4. Efron, B., and R. J. Tibshirani, 1993: *An introduction to the bootstrap.* Chapman & Hall, London.

5. Hassibi, B., and D. Stork, 1993: Second order derivatives for prunning: optimal brain surgeon. Advances in Neural Information Processing Systems, Vol. 5, pages 164-171.

6. Le Cun, Y., J. Denker, and D. Henderson, 1990: Optimal brain damage. Advances in Neural Information Processing Systems, Vol. 2, pages 598-605.

7. MacKay, D. J. C., 1996: Bayesian methods for back-propagation networks, in *Models of Neural Networks III*, E. Domany, J. L. van Hemmen, K. Schulten (Eds.), Springer-Verlag, New York, physics of neural network series, pp. 309.

8. Mao, J., K. Mohiuddin, and A. Jain, 1994: Parsimonious network design and feature selection through node prunning. Proc. 12th Int. Conference on Pattern Recognition, pages 622-624.

9. Masters, T., 1993: *Practical neural network recipes in C++.* Academic Press, 493 pp.

10. Marzban, C., E. D. Mitchell, and G. Stumpf, 1999: The Notion of "Best Predictors:" An Application to Tornado Prediction. *Wea. Forecasting*, **14**, 1007-1016.

11. Marzban, C., 1998: Bayesian inference in neural networks. Proceedings of the 78th Annual Meeting of the American Meteorological Society, Phoenix, Arizona.

12. Marzban, C., and G. Stumpf, 1998a: A neural network for damaging wind prediction. *Wea. Forecasting*, **13**, 151-163.

13. Marzban, C., and G. Stumpf, 1998b: A neural network for tornado and/or damaging wind prediction based on Doppler radar-derived attributes. *Microcomputer Applications*, **17**, 21-28.

14. Marzban, C., H. Paik, and G. Stumpf, 1997: Neural networks vs. gaussian discriminant analysis. *AI applications*, **11**, 49-58.

15. Marzban, C., and G. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *Journal of Applied Meteorology*, **35**, 617-626.

16. Mitchell, E. D., S.V. Vasiloff, G. J. Stumpf, A. Witt, M. D. Eilts, J.T. Johnson, and K. W. Thomas, 1998: The National Severe Storms Laboratory Tornado Detection Algorithm, *Wea. and Forecasting*, **13**, 352-366.

17. Mitchell, Tom, 1996: What have we learned about learning? Keynote address, proceedings of the 13th National Conference on Artificial Intelligence, Portland, Oregon.

18. Murphy, A. H., and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7**, 435-455.

19. Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

20. Neal, R. M., 1996: *Bayesian learning for neural networks.* Cambridge, Cambridge University Press, pp. 183.

21. Richard, M. D., and R. P. Lippmann, 1991: Neural network classifiers estimate Bayesian a-posteriori probabilities, *Neural Computation*, **3**, 461-83.

22. Stumpf, G., A. Witt, E. D. Mitchell, P. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory Mesocyclone Detection Algorithm for the WSR-88D. *Weather and Forecasting*, **13**, 304-326.