Neural Networks vs. Gaussian Discriminant Analysis

Caren Marzban[1,2,3]
Haejung Paik[4]
and
Gregory J. Stumpf[1,2]

[1] National Severe Storms Laboratory, Norman, OK 73069
[2] Cooperative Institute for Mesoscale and Meteorological Studies,
University of Oklahoma, Norman, OK 73019
[3] Department of Physics, University of Oklahoma, Norman, OK 73019
[4] Department of Communication, University of Oklahoma, Norman, OK 73019

Abstract

A classification task is chosen to compare the performance of a feed-forward neural network with that of a gaussian discriminant analysis, both in a Bayesian framework. The data set is taken from the National Severe Storms Laboratory's Mesocyclone Detection Algorithm, and the two classes of interest consist of circulations that are tornadic and those that are not. Two measures of performance and two methods of classification are considered. It is shown that a neural network whose outputs have been transformed to posterior probabilities outperforms a neural network without such a transformation, and also outperforms discriminant analysis, regardless of the two measures of performance considered herein.

1 Introduction

In a series of two recent papers (Marzban and Stumpf, 1995, 1997) the development of several Neural Networks (NN) for the diagnosis of tornados and damaging wind has been outlined. In the first paper, using 6 storm days (3,258 circulations) a comparison was made between an NN and Linear Discriminant Analysis [1] (LDA) in predicting whether or not a circulation detected by the National Severe Storms Laboratory's Mesocyclone Detection Algorithm (MDA) is tornadic. The comparison was made in a non-probabilistic scheme. In the second paper, a Bayesian probabilistic framework was set up within which an NN was developed for the prediction of damaging wind. Both data sets were composed of 6 storm days.

Since that time, more data have become available and revised NNs have been developed. One such NN is based on 22 storm days (26,058 circulations) and is for the diagnosis of tornados. In this article, the performance of this NN will be compared to that of Discriminant Analysis (DA) (both Linear and Quadratic) in a Bayesian framework, in terms of two performance measures. Thus the primary difference between this paper and the first of the aforementioned papers is in 1) the larger data set, 2) the probabilistic framework, and 3) a "better" measure of performance, in that chance is taken into account, and it is "equitable" (Gandin and Murphy, 1992), all attributes of the present article. The input variables are the same 23 that were employed in the earlier works, derived from Doppler radar velocity data, and the outputs represent the existence/nonexistence of tornados based on ground observation.

It is worth pointing out that since both of the measures considered here are one-

[1] Throughout this article, discriminant analysis refers to gaussian discriminant analysis. Other nonparametric forms will not be discussed.

dimensional (scalar) quantities, in contrast to the inherently multi-dimensional nature of forecast quality (Marzban and Stumpf, 1997; Murphy, 1991, 1996), any conclusion regarding the relative performance of the various methods is limited to these measures only. In other words, it is entirely possible that whereas an NN may outperform DA in terms of some measure of performance, the opposite may occur when performance is gauged in terms of some other measure. For example, Marzban and Stumpf (1997) gauged the performance of an NN (for a different problem) in terms of 15 different measures, and found that the NN outperformed logistic regression in terms of all the measures, except for one - a measure of discrimination put forth by Murphy, Brown and Chen (1989). That the "winner" depends on the choice of measure may seem disturbing at first; however, it is a simple consequence of attempting to gauge a multidimensional quantity with a one-dimensional measure. For such reasons, it is imperative to select a particular measure of performance based on some other criterion, before any statements are made about the superiority of one algorithm over another. Several criteria were considered in Marzban and Stumpf (1997); they all require that a measure be well-defined in a rare-event situation, and that it should not be prone to "hedging" (Murphy and Epstein, 1967). The two measures considered herein satisfy both criteria, but it should be emphasized that they are still only *scalar* measures, and as such do not represent all the dimensions of performance quality.

2 The Mesocyclone Detection Algorithm (MDA)

The National Severe Storms Laboratory's MDA uses pattern recognition to detect rotation signatures (hereafter called "circulations") in Doppler weather radar data. First, azimuthal shear pattern vectors are built; these are vectors of decreasing radial velocity along adjacent radar sample volumes at constant range from the radar. These pattern vectors must meet minimum strength requirements (velocity differences ranging from 10 m/s at ranges within 100 km of the radar, decreasing to 6 m/s at ranges greater than 200 km from the radar). Next, the pattern vectors are combined into 2D "features," based on spatial proximity. "Core regions" of azimuthal shear are extracted from the 2D feature such that their aspect ratio does not exceed 2. The strength of a 2D feature is determined by measuring the rotational velocity and shear using the maximum inbound and outbound radial velocities. Next, a "3D circulation" is defined when there is a continuous depth of 2D features meeting the above strength criteria, which can be vertically associated (the details of vertical association are omitted for brevity). Some 3D circulations can be associated in time with another detection from a previous volume scan (usually 5 or 6 minutes in the past) (the details of time association are also omitted for brevity).

One of the many attributes computed by the MDA is a quantity called the "strength rank." It is a non-dimensional number related to the range-dependent strength parameters (rotational velocity and shear) of a circulation. Each 2D feature used to build a 3D detection has a strength rank. The strength rank of the 3D detection is the value at which a continuous depth ($> 3\ km$, base $< 5\ km$ above ground level) of 2D features used to make up this 3D detection are greater than or equal to this value. In turn, another quantity - called MSI - is defined as the vertically integrated strength rank ($\times 1000$) for all 2D features within the 3D detection. This quantity is thought to be a good predictor of tornados. Its distribution both for nontornadic (0) and tornadic (1) circulations is shown in Figure 1. A classification rule, based on this quantity, can be devised by, first, placing a threshold on it, above which circulations are classified as tornadic, and then, vary the threshold itself to optimize the performance of the algorithm. The values of the performance measures based on this MSI classification rule will be presented in the Conclusions Section.

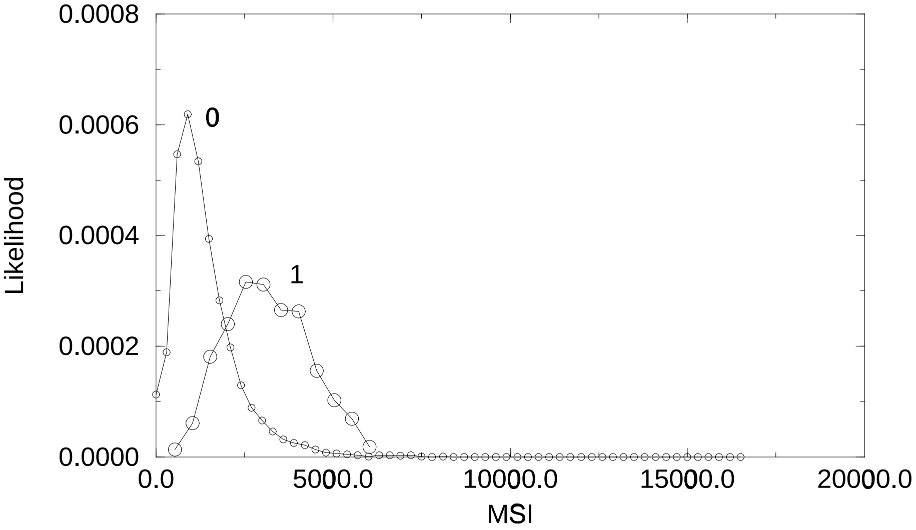

Figure 1: The distribution of MSI for nontornadic (0) and tornadic (1) circulations. The two distributions have been scaled to have equal areas.

The 23 attributes based on Doppler velocity data used for both the NN and DA were 1. Base (m), 2. Depth (m), 3. "Strength rank" (0-9), 4. Low-altitude diameter (m), 5. Maximum diameter (m), 6. Height of maximum diameter (m), 7. Low-altitude rotational velocity (m/s), 8. Maximum rotational velocity (m/s), 9. Height of maximum rotational velocity (m), 10. Low-altitude shear (m/s/km), 11. Maximum shear (m/s/km), 12. Height of maximum shear (m), 13. Low-altitude gate-to-gate velocity difference (m/s), 14. Maximum gate-to-gate velocity difference (m/s), 15. Height of maximum gate-to-gate velocity difference (m), 16. Core base (m), 17. Core depth (m), 18. Age (s), 19. Strength index weighted by avg density of integrated layer (MSI), 20. Strength index "rank", 21. Relative depth (%), 22. Low-altitude convergence (m/s), and 23. Mid-altitude convergence (m/s).

3 NN and DA: Theoretical Considerations

Details of DA can be found in (McLachlan, 1992). In its simplest form, the data is assumed to be multivariate normal, and classification is made based on whether or not an observation is less than or greater than some threshold value according to the posterior probability of each class. Specifically, the Likelihood of an observation, x, given that it belongs to the i^{th} class, is assumed to be

$$L_i(x) \sim \frac{1}{\sqrt{det\ V_i}}\ exp^{-\frac{1}{2}(x-\mu_i)^T\ V_i^{-1}\ (x-\mu_i)}\ ,$$

where μ_i is the vector of the means and V_i is the covariance matrix for the i^{th} class (here $i = 0, 1$), all estimated from a training data set. Then, an observation, x, is classified into the class with the larger posterior probability, $P_i(x)$, which in turn is derived from Bayes' theorem:

$$P_i(x) = \frac{p_i L_i(x)}{p_0 L_0(x) + p_1 L_1(x)},$$

where p_0, p_1 are the prior probabilities for the two groups, discussed below. Of course, $p_0 + p_1 = 1$ and $P_0 + P_1 = 1$. It is easy to show (Marzban, 1997; McLachlan, 1992) that the decision criterion (or the discrimination function, $\log(P_1(x)/P_0(x))$), is quadratic in the quantity x; such an analysis is referred to as a Quadratic Discriminant Analysis (QDA). However, if $V_0 = V_1$, then the decision criterion is linear in x, leading to a Linear Discriminant Analysis (LDA). The main advantage of the latter is in allowing for the interpretation of the linear coefficients as "predictive strengths" of the corresponding variables.

By contrast, training an NN involves the minimization of some error function. As such, it is a generalized regression model. Often, and in this application, the mean-square error is chosen as the error function:

$$E(\omega) = \frac{1}{N}\sum_{i=0}^{N}(t_i - p_i(\omega))^2,$$

where the ω_{ij} are the weights (parameters) of the network to be estimated during training, N is the number of cases in the training set, and t_i and p_i are the target (actual) and the predicted values of a single output node. For a larger number of output nodes, one introduces another $\sum$ that ranges over the number of output nodes. This choice of the error function is motivated by the well-known fact (Bishop, 1996; Draper and Smith, 1981) that if the distributions of the dependent variables are normal (gaussian), then least-square estimates are equal to maximum-likelihood estimates.

The following are some specifics of the present application: The 23 input variables were linearly scaled to lie in the range 0.1 to 0.9; the training algorithm was Conjugate Gradient, and it was halted when no further decrease in E occurred; Simulated Annealing was employed both to avoid and to escape the local minima of the error function; the activation function for all the layers was taken to be the logistic function, $f(x) = \frac{1}{1+\exp(-x)}$; the optimal number of hidden nodes was determined by a bootstrapping method in which trained NNs with different number of hidden nodes were tested on several (here 5) training and validation data sets. For each of the 5 random, independent partitions of the data set into a training set and a validation set, 100 random initial weights were considered, again to improve the chances of finding a global (or a sufficiently deep, local) minimum.

A comment is in order regarding the determination of the optimal number of hidden nodes. In this application, as in (Marzban and Stumpf 1995), several networks with a variety of number of hidden nodes were tested, and the one that yielded the highest performance on the validation set was selected. This precludes any overfitting to the training set. It may be objected that this method of finding the optimal number of hidden nodes, though not overfitting the training set, may overfit the validation set instead. However, such an outcome is precluded since several (here 5) randomly selected validation sets (and training sets) were considered. Additionally, these outcomes were averaged, and 90% confidence limits were placed on them.

The training and the validation sets were selected from a total of 25,939 circulations detected by the MDA. The number of tornadic circulations (i.e., events, or "1"s) was 784, and the remaining 25,155 circulations were nontornadic, making for a ratio of 784/25,155 = 0.031 . The 784 "1"s were randomly divided into two groups of 500 and 284 cases to be used in the training set and the validation set, respectively (and this was repeated 5 times). The ratio, 0.031 was employed to select 9155 ($\sim$ 284/0.031) "0"s in the validation set in order to maintain the same climatological ratio of the class sizes in the validation set as in the entire data set.

Two classification schemes are possible: In one, a classification can be made directly from the 2 output nodes of the network in what is referred to as the "winner-takes-all" method. In this method, there are as many output nodes as the number of classes (i.e. 2), and the output with the higher activation designates the class. This discrete method of classification is quite common and leads to one type of classification. We shall label these results as "NN".

A second type of classification can be made by considering the conditional probability of a class, given the outputs. In this method, these likelihoods can be converted to event posterior probabilities using Bayes' theorem. In order to compare the two types of classification in terms of the measures considered herein, it is necessary to reduce the posterior probabilities to dichotomous (0/1) outcomes. This reduction can be implemented by placing a threshold on the probabilities and then varying this threshold in order to optimize some measure of performance. However, given that classification is based on whether $P_1 > P_0$ or $P_0 > P_1$

(recall $P_0 + P_1 = 1$), the natural threshold for a probability is 50%. The threshold is fixed at this value, and instead, the prior probability p_1 is varied in order to optimize some measure of performance (Marzban, 1997). These results will be labeled as "NN_p".

The transformation of the two output nodes to posterior probabilities is done in three steps: First, the two output nodes, σ_{left}, σ_{right}, are combined into a single fictitious output node, σ:

$$\sigma = f(\beta(\sigma_{left} - \sigma_{right})),$$

where f is the logistic function, and β is a parameter that measures the strength of the mixing; we have found that the ultimate results are quite insensitive to the exact value of the β parameter. Second, estimates of the likelihoods $L_i(\sigma)$ are arrived at by a method proposed by Parzen (1962). There, an estimator is shown to be

$$L(\sigma) = \frac{1}{n\lambda} \sum_j \exp^{-[(\sigma - \sigma_j)/\lambda]^2},$$

where σ_j are a random sample of size n. Note that this choice does not imply that the Likelihoods themselves are gaussian. Here, the random sample σ_j is drawn from the training data. In other words, n is the size of the training set, and the σ_j are the values of the single fictitious output node that result from exposing the trained NN to the training data. The parameter λ is a "smoothing parameter" that is to be fixed; the final results are insensitive to the specific value of λ as well. Parzen's method is simply one of many methods for superimposing (or fitting) a density function - in this case, $L_i(\sigma)$ - onto a histogram of σ; see Figures 2a and 2b. Third, and finally, these likelihood functions are then employed to obtain posterior event probabilities, $P_1(\sigma)$, via Bayes' theorem.

The final issue is that of the prior probability p_1 ($= 1 - p_0$). It is often argued that it should be estimated from the ratio $N_1/(N_0 + N_1)$, where N_1 and N_0 are the tornadic and

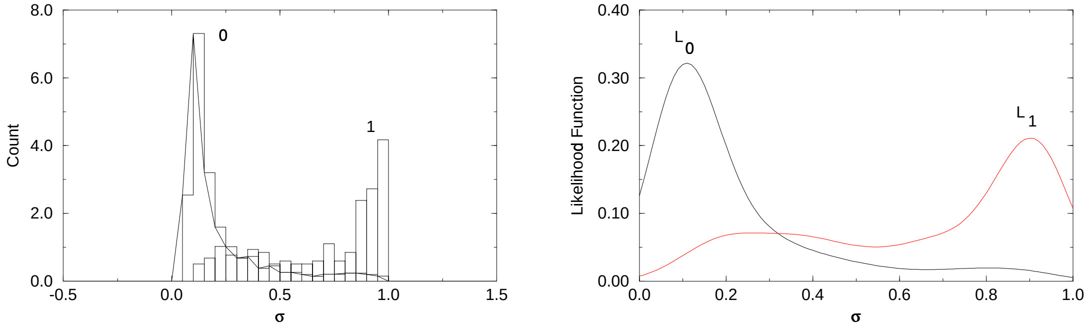

Figure 2: a) The distribution of activations, and b) Parzen's fit at $\lambda = 0.1$.

nontornadic sample sizes, respectively. However, it is shown in (Marzban, 1997) that this choice is often not the one that optimizes a measure of performance. There, a gaussian model was developed and it was shown that different measures of performance are optimized at different values of p_1. As a result, in this article, the values of the measures are computed for $p_1 = 0.01$, and all other p_1 in the range 0.1 to 0.9, in 0.1 increments.

4 Measures of Performance

Many measures of performance have been examined (Brooks and Doswell, 1996; Doswell, et al., 1990; Gandin and Murphy, 1992; Marzban and Stumpf, 1997; Murphy, 1996, 1988; Murphy and Winkler, 1992; Murphy and Winkler, 1987), and many represent different aspects of performance. However, some behave pathologically in rare-event situations. One that appears to be "most healthy" is the Heidke Skill Statistic (HSS), while another "not-so-healthy" measure (Murphy and Epstein, 1967) is the Critical Success Index (CSI); for more detail, see (Marzban and Stumpf, 1997). In this article, only these two measures will be considered - HSS because of its proper behavior in the rare-event situation, and CSI because of its popularity in meteorological circles.

The performance of any classifier can be encapsulated in a Contingency Table (otherwise known as the Confusion Matrix), or in short the C-table,

$$\text{C-table} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \text{\# of 0's classified as 0's} & \text{\# of 0's classified as 1's} \\ \text{\# of 1's classified as 0's} & \text{\# of 1's classified as 1's} \end{pmatrix} .$$

Note that the total number of nonevents is given by $N_0 = a + b$, that of events is $N_1 = c + d$, and $N = N_0 + N_1$. As mentioned above, the two measures employed for the present analysis are CSI and HSS, and they are defined as

$$\text{CSI} = \frac{d}{b+c+d} \quad \text{and} \quad \text{HSS} = \frac{2(ad - bc)}{(a+b)(b+d)+(a+c)(c+d)} ,$$

with a, b, c, d being the elements of the *validation* C-table.

Although it may not be apparent from these expressions, CSI does not take into account any non-skill related factors such as bias or random guessing, but HSS does; it is easy to show that HSS = 0 when classification is done randomly, or when it is done persistently (i.e. classifying everything as either event or as nonevent).

5 Results

Figures 3a and 3b show the values of CSI and HSS, respectively, as a function of p_1. The values of the measures are the averages of the respective measures over the 5 random validation sets, and the error bars are the 90% confidence intervals. It can be seen that regardless of the two measures of performance, at their respective optimal value of p_1, the NN based on the probabilistic classification rule (NN_p) outperforms linear discriminant analysis (LDA), which in turn outperforms quadratic discriminant analysis (QDA). The horizontal lines correspond to the NN with the non-probabilistic classification rule [2] (NN) which itself outperforms the rule based on MSI.

6 Discussion

The performances of the various methods can be "explained:" For instance, it is not surprising that the performance of the NN is higher than that of MSI, because one of the 23 inputs of NN was MSI itself. Similarly, based on the fewer and less explicit assumptions in NNs, it is not surprising that NN_p (at its optimum) is higher than all the other non-NN methods. As for why NN_p outperforms NN, this can be understood if one notes the rare-event nature of the data; since there are many more nontornadic circulations than tornadic ones, an NN trained on such disproportionate classes will have a tendency to classify more observations as nontornadic than tornadic. Such a bias can reduce the overall performance of a non-probabilistic NN. A somewhat surprising result is the superior performance of LDA over QDA, in spite of the additional assumption of homoelasticity (i.e. equality of the covariance matrices) in the former. The reason for this is the relatively small sample size, especially that of the tornadic circulations, because without the assumption of homoelasticity the covariance matrices have many more parameters that must be estimated from a small sample. Needless to say, these conclusions are specific to the problem at hand (i.e. the data set); larger data sets, including different types of storms are currently under investigation.

Some general remarks are in order. All methods of statistical analysis have inherent assumptions that limit their applicability, and NNs are no exception, although the assumptions made in an NN analysis are often milder and more implicit. For instance, in DA, distributions are explicitly assumed to be gaussian (normal). In LDA, additionally, the co-

[2] The optimal number of hidden nodes was found to be 6, all on one hidden layer.

variance matrices for the various classes are assumed to be equal (homoelasticity). There exist nonparametric variations of DA that do not employ either assumption, but there are still "kernels" and "smoothing parameters" with which one must reckon in estimating the distributions (Silverman, 1986). Another nonparametric method with similar assumptions is Classification and Regression Trees (CART); see Burrows (1991). In Multiple Linear Regression the underlying relation is assumed to be linear in the parameters. As the names suggest, in Polynomial Regression, the function is assumed to be some prespecified polynomial, and Logistic Regression assumes the logistic function $f(x) = \frac{1}{1+\exp(-x)}$. All of these models incorporate explicit assumptions whose violation may jeopardize the predictive capability of the model.

NNs are thought to be robust in regards to the a priori distribution of the data (Masters, 1993); however the primary advantage of NNs is in that no explicit assumptions are made regarding the underlying function; by varying the number of hidden nodes systematically, one effectively parametrizes the space of all functions (Hornik, et al., 1989). On the other hand, there is evidence to suggest that although all functions may be representable by NNs, not all are equally learnable (Carnevalli and Patarnello, 1987; Ferran and Perazzo, 1990; Parisi 1992). Consequently, it may still be argued that certain assumptions are made in an NN analysis as well, although most are of an implicit nature.

Finally, it is worth pointing out that, in principle, for a proper NN development, one must have three independent data sets: a training set, a validation set, and a test set. The validation set may actually be used during the training phase in order to monitor the performance of the NN, but the test set is to be kept completely out of the training phase. As in regression methods, the performance of an NN on the training set itself is optimistically biased. The bias can be reduced by evaluating the NN on the validation set, and even further reduced by testing the NN on the test set. However, the price one pays in this process is in the smaller sample size (per set) and increased variance. For this reason, in this application, no test set was constructed.

This is a specific instance of the bias/variance dilemma in NNs (Geman, et al., 1992; Bishop, 1996; Ripley, 1996). The scope of this dilemma goes beyond the realm of NNs, but the flexibility of NNs renders the dilemma of particular concern. It is well known that the mean square error can be decomposed into a term that represents bias and another that represents variance. The former is a measure of the average (over all data sets) difference between the NN and the underlying function, while the latter is a measure of the extent to which the NN depends on the particular choice of data set. It is evident that ideally one would like to minimize bias while minimizing variance simultaneously. Since a sufficiently complex NN - say, in terms of the number of hidden layers and hidden nodes - is capable of representing any function to arbitrary accuracy, it follows that for finite data sets such a complex NN may minimize bias, but it also has maximum variance. Conversely, an NN with no hidden nodes minimizes variance but only at the cost of maximizing bias. Geman, et al. (1992) show that one way to minimize both bias and variance is by cross-validation; in cross-validation, some number of the training cases are left out to be used for validation. Then these cases are placed back into the training phase and another set is kept out for validation, etc. This has been the procedure adopted in the present article.

7 Conclusions

In summary, it is found that a neural network whose outputs are translated into posterior probabilities, with the prior probabilities set appropriately, outperforms one that employs the winner-takes-all classification scheme; it also outperforms the two versions of discriminant analysis - linear and quadratic. These results are statistically significant and hold true regardless of whether performance is gauged in terms of the Critical Success Index or the Heidke Skill Statistic.

Some points are worth emphasizing: Given that performance is a multifaceted entity, the comparison of any two algorithms is a multifaceted endeavor. It is rare for one algorithm to outperform another in terms of *all* the dimensions of performance. It is much more ubiquitous to find an algorithm that excells in only a subset of all the performance measures. For this reason, it is important to decide what the measure of performance is to be, before any comparisons are made. Of course, there exist a variety of multi-dimensional "measures" that can represent several aspects of performance at once, e.g. reliability diagrams. However, in spite of the diagnostic advantages of such "measures", their multidimensional nature may preclude them as candidate measures for determining the winner of some contest, or the better of two algorithms.

Therefore, the conclusions of this article are specific to the measures CSI and HSS. [3] This is the main reason why it is possible to increase the performance of an NN (in terms of CSI or HSS) by varying the prior probabilities. In other words, it is entirely possible that a given value of p_1 may maximize HSS, but be suboptimal for some other measure of performance.

Acknowledgements

We would like to thank Mike Eilts for a careful reading of the manuscript.

References

Bishop, C. M., 1996: *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford,

[3] Needless to say, the conclusions are specific to the data set as well.

pp 482.

Brooks, and H. E., C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288-303.

Burrows, W. R., 1991: Objective guidance for 0-24-hour and 24-48-hour mesoscale forecasts of lake-effect snow using CART. *Wea. Forecasting*, **6**, 357-378.

Carnevalli,P. and S. Patarnello, 1987: Exhaustive Thermodynamical Analysis of Boolean Learning Networks. *Europhys. Lett.*, **4**, 1199-1204.

Doswell III, C. A., R. Davies-Jones, and D. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576-585.

Draper, N. R. and H. Smith, 1981: *Applied Regression Analysis.* John Wiley and Sons, New York, 709 pp.

Ferran, E. A., and R. P. J. Perazzo, 1990: Inferential entropy of feed-forward neural networks. *Physical Review A*, **42**, 6219-6226.

Gandin, L. S., and A. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361-370.

Geman, S., E. Biensenstock, and R. Doursat, 1992: Neural networks and the bias/variance dilemma. *Neural Computation*, **4**, 1-58.

Hornik, K., M. Stinchcombe, and H. White, 1989: Multilayer feedforward networks are universal approximators. *Neural Networks*, **4:2**, 251-257.

Marzban, C., 1997: The effect of Bayesian prior probability on skill in GAUssian models. *Journal of Applied Meteorology*, accepted for publication.

Marzban, C., and G. Stumpf, 1995: A neural network for tornado prediction based on Doppler radar-derived attributes. *Journal of Applied Meteorology*, **35**, 617-626.

Marzban, C., and G. Stumpf, 1997: Measures of Skill: Application to a neural network for damaging wind prediction. To appear in *Wea. Forecasting.*

Masters, T., 1993: *Practical Neural Network Recipes in C++.* Academic Press, 493 pp.

McLachlan, G. J., 1992: *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley and Sons, Inc., New York. 526 pp.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417-2424.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590-1601.

Murphy, A. H., B. G. Brown, and Y-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting*, **4**, 485-501.

Murphy, A. H., and E. S. Epstein, 1967: A note on probabilistic forecasts and "hedging", *Journal of Applied Meteorology*, **6**, 1002-1004.

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

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

Murphy, A. H., 1996: The Finley affair. *Wea. Forecasting*, **11**, 3-20.

Parisi, G., 1992: On the classification of learning rules. *Network*, **3**, 259-265.

Parzen, E., 1962: On estimation of a probability density function and mode. *Ann. Math. Statistics*, **33**, 1065-1076.

Ripley, B. D., 1996: *Pattern Recognition and Neural Networks.* Cambridge: University Press.

Silverman, B. W., 1986: *Density estimation for statistics and data analysis.* New York: Chapman and Hall.

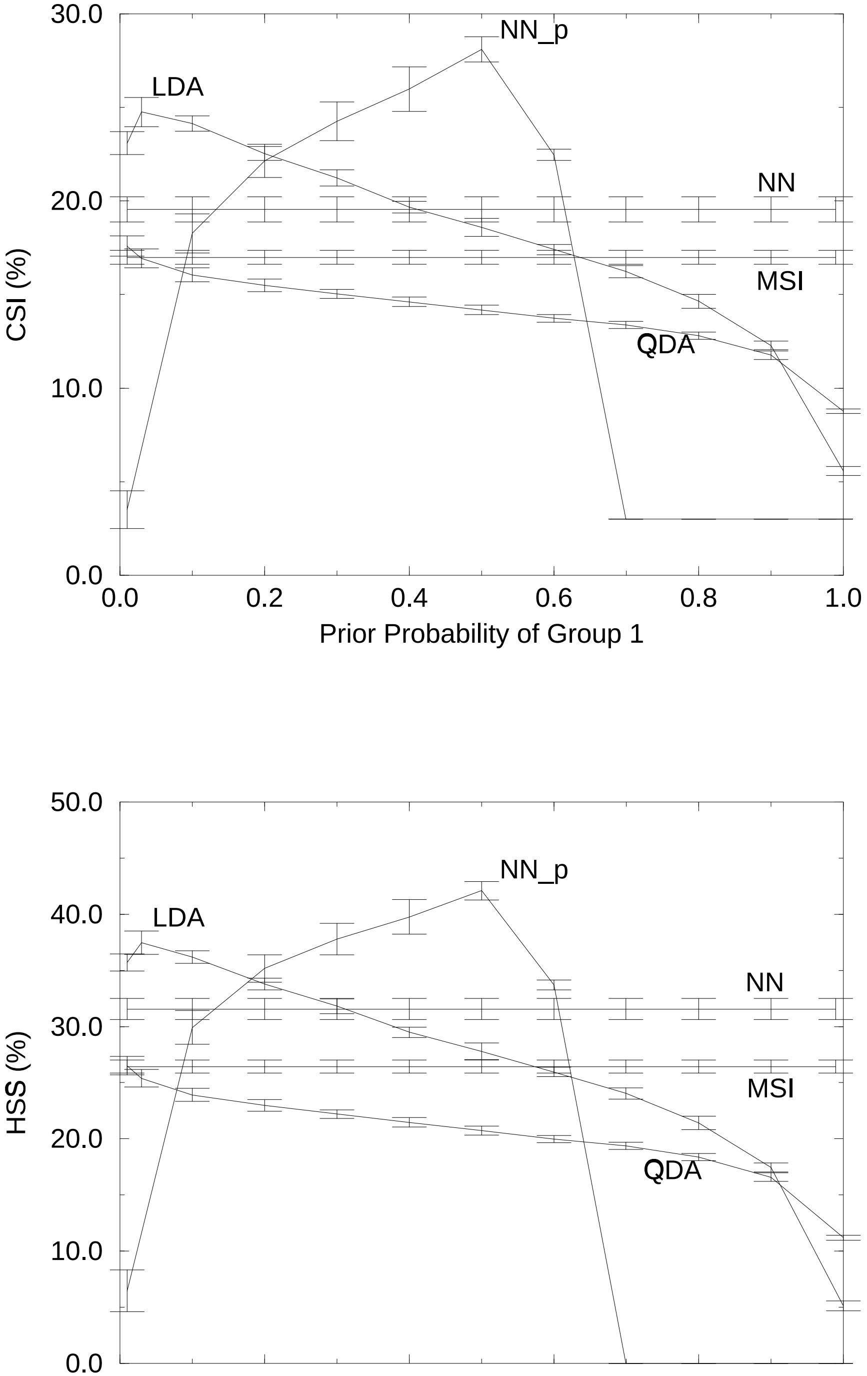

Figure 3: a) CSI and b) HSS as a function of p_1, for the NN with the outputs transformed to posterior probabilities (NN_p), the NN with a winner-takes-all classification criterion (NN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and the MSI-based rule (MSI).