Although neural networks do offer a few advantages over some other nonlinear methods, in certain situations these advantages are difficult to utilize. However, many neural network applications in the social sciences are flawed in ways that obfuscate such effects. In this article, the neural network methodology is reviewed, some common flaws are pointed out, and a rather commonplace data set—dealing with school delinquency—is analyzed for illustrative purposes.

Comments on Neural Networks

HAEJUNG PAIK
University of Oklahoma

Many complex processes can be explained in terms of a number of simpler processes (at least according to the reductionist point of view). Equivalently, many complex systems can be modeled in terms of a number of elementary entities and the interactions between them. Perhaps the best example of such a system is the brain. For instance, most neurons in the brain perform the relatively simple task of outputting an electrical signal when the total input into the neuron exceeds a given, predetermined, threshold. However, the complex tasks performed by the brain can be faithfully modeled in terms of the interactions between such neurons (Churchland and Sejnowski 1992). The interactions take place via synaptic connections, and the electrical resistance of a connection determines the strength of the interaction between the connected neurons. Typically, a single neuron in the memory centers of the brain is connected to approximately 10,000 other neurons in a fully interconnected fashion, but certain portions of the brain have been observed to have a layered structure, for example, a vision layer taking inputs from the sensory receptors in the eyes and a motor layer that controls the eye muscles.

Although it may not be obvious at first sight, the latter (layered) network of connections is functionally similar to a nonlinear, multivariate regression model. The vision layer of neurons is analogous to the set of independent (predictor) variables, whereas the motor neurons are

SOCIOLOGICAL METHODS & RESEARCH, Vol. 28 No. 4, May 2000 425-453

analogous to the dependent (response) variables. The map between the vision and motor layers can then be thought of as a mathematical function represented by the regression model. This interpretation has been fruitful both in the neurosciences and in statistics and has given rise to a cross-disciplinary field generally referred to as neural networks (NNs). NNs have been used in a wide range of applications, including organizational processes (Carley and Svoboda 1996; Schrodt 1991), decision making (Artyushkin et al. 1990; Lenard, Alam, and Madey 1995), and communications (Woelfel 1993). Some of these studies use NNs as a dynamical model of the interactions, whereas others treat NNs as a statistical tool for estimating the interactions from sample data. This article is concerned only with the latter aspect of NNs.

NNs are often presented as a novel and assumption-free statistical method for performing regression and classification. However, recently, it has become evident that NNs are by no means assumption free, although the assumptions may be considered milder and more implicit than those of many other methods. This robustness has given rise to a rapid popularity of NNs among statistical model builders. Several excellent discussions of NNs and their relation to other statistical methods can be found in Bishop (1996), Masters (1993), and Sarle (1994b).

One may also argue that the use of NNs has been somewhat extravagant, the justification for which has been the NNs' capability to model highly nonlinear relationships and nontrivial interactions between the variables of a model. One may even expect this property to imply that an NN can outperform all other methods. Indeed, many applications have enjoyed this flexibility with great success (Collins, Ghosh, and Scofield 1988; Marzban and Stumpf 1996; Paik and Marzban 1995; Van Nelson and Neff 1990). However, for a given data set, an NN may be outperformed by a method with many restrictive and explicit assumptions. For example, if the function underlying the data is a polynomial, then polynomial regression will certainly outperform an NN (Bishop 1996). In other words, it is not true that the milder and more implicit assumptions of NNs automatically render them superior to the alternatives.

The flexible fitting property of NNs and the search for the "best" statistical method have given rise to a plethora of research articles wherein different statistical methods are compared and contrasted (Anderer et al. 1994; Garson 1991; Hardgrave, Wilson, and Walstrom 1994; Marzban, Paik, and Stumpf 1997; Paik and Marzban 1995; Wilson and Hardgrave 1995). Many such endeavors are flawed in that they neglect (at least) three important contingencies, namely, that the choice of the best method is contingent on (1) the proper implementation of the methods, (2) the measure of performance, and (3) the data.

Some of the improper implementations of NNs involve the following:

- An ad hoc value for the number of hidden nodes (one of the two quantities that determine the complexity or nonlinearity of an NN)
- The use of a single data set for estimating both the optimal number of hidden nodes and the performance of an NN
- An inappropriate choice of the error function to be minimized
- A disregard for the existence of local minima in the error function

The second contingency refers to the dependence of any empirical comparison between two models on the measure of performance employed for the comparison. Performance is a multifaceted entity, and it is entirely possible that model A may outperform model B in terms of one facet of performance but not in terms of another. Often, however, the multifaceted nature of performance is neglected in the comparison of one method with another.

As for data dependence, it must be emphasized that it is entirely possible that method A will outperform method B on one data set but not on another. This contingency is one that requires only a confession, in that it is sufficient to acknowledge that any empirically established superiority of one method over another is specific only to the particular data set (and measure) being analyzed.

In this article, an NN methodology that encompasses the abovementioned contingencies is reviewed and then illustrated in an analysis of a rather "generic" data set—one that has been employed in many analyses (Lee and Smith 1994; O'Brien and Rollefson 1995; Reardon 1996; Rees, Argys, and Brewer 1996). The aim of the article is twofold: (a) to point out some errors that are commonly made in NN applications and (b) to prove that sometimes (e.g., when the data are too noisy) the many advantages of a properly implemented NN are not easily realized. For example, it is shown that any amount of nonlinearity allowed in a correctly implemented NN causes it to overfit[1] the particular data set examined herein. This implies that the optimal NN is a linear one and, in turn, that the classification boundaries underlying the present data set are mostly linear (to within statistical errors). Additionally, a linear discriminant analysis—a model with many explicit assumptions—is undertaken and shown to perform comparably with a linear NN and superior to a nonlinear NN.

NEURAL NETWORKS

An NN generally refers to a network of elementary processing units, called neurons or nodes, interconnected via synaptic connections, or simply, weights.[2] It is the set of values assigned to these weights that determines the task the NN is to perform; and to arrive at the desired weights, the NN must be trained. In this sense, NNs are parametric models, and training is nothing more than the process of parameter estimation. There exist a wide variety of NNs for performing an equally wide range of tasks, but the way in which NNs are trained can generally be divided into (at least) two paradigms—unsupervised and supervised. The key idea in the former is self-organization, in that such an NN is designed to automatically search for salient features in the data. Such NNs are nonlinear analogs of the traditional methods for cluster analysis. Supervised NNs, on the other hand, are nonlinear analogs of regression and classification methods wherein the dependent variable is known and used in the training procedure. For such NNs, training refers to the process of varying the weights to minimize the "difference" between the output of the NN and the desired value of the dependent variable. The data set used for this process is called the training set.

A particular type of supervised NN is the so-called multilayered perceptron wherein the network has a layered architecture with the nodes on a given layer not interacting with one another. It is this type of NN that is the primary interest of the present article. In fact, it will be assumed that the nodes on a given layer interact only with those of the adjacent layers. The input layer contains the nodes that represent the independent variables, and the nodes of the output layer represent the dependent variables of the problem. An NN with one hidden layer

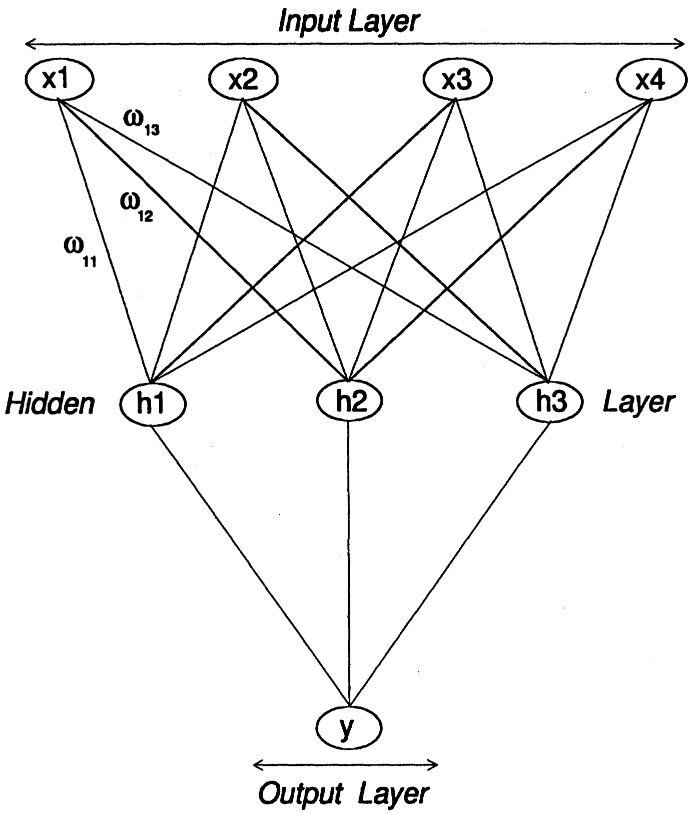


Figure 1: A Multilayered Perceptron With Four Input Nodes, Three Hidden Nodes (on one layer), and One Output Node

NOTE: Also shown are three of the weights/parameters.

containing H hidden nodes (Figure 1) can be written as a single parameterization:

$$y_k = f[\Sigma_{i=1}^{H} \omega'_{ik} f(\Sigma_{j=1}^{n} \omega_{ij} x_j - \theta_j) - \theta'_k], \qquad (1)$$

where ω, ω', θ, θ' are weights (parameters) that are to be estimated, y_k are the output nodes, and x_j are the n input variables. The function $f(x)$ is called the activation function and represents the manner in which any two nodes interact with one another, and this in turn affects the function represented by the NN as a whole (i.e., the function relating the inputs to the outputs). A linear activation function between two nodes renders the NN a linear function as a whole, regardless of the existence or the number of hidden nodes. A nonlinear activation function does not guarantee that the NN can represent any nonlinear function; for that, it is necessary to introduce hidden nodes into the NN. A commonly employed, nonlinear activation function is the so-called logistic function, $f(x) = 1/(1 + \exp^{-x})$; the logistic function varies between 0 and 1 and, therefore, allows for a probabilistic interpretation of the values taken by the nodes.

To elucidate equation (1), it is helpful to consider some simple cases. With one input node x, one output node y, and no hidden nodes, an NN is a representation of the equation $y = f(\omega x - \theta)$, where ω and θ are the parameters of the model. Clearly, if $f(x) = x$, then the NN represents nothing more than the linear regression model $y = \omega x - \theta$. If $f(x)$ is the logistic function, then the NN can represent a logistic regression model. In fact, for classification problems (as opposed to regression problems), it is sufficient to use the logistic function because it allows for the outputs of the NN to represent class-conditional probabilities (more on this below).[3]

The word *hidden* may suggest that the corresponding nodes perform some hidden (or mysterious) function. However, they are nothing more than abstract constructs whose function is to introduce and control the nonlinearity of the NN. They also allow for interactions between the independent variables. Introducing a single hidden node renders the NN nonlinear, and the representing equation takes the (unattractive) form

$$y = f[\omega' f[\omega\, x - \theta) - \theta'] = \frac{1}{1 + \exp^{-\frac{\omega'}{1+\exp^{-\omega x+\theta}} + \theta'}}, \qquad (2)$$

where ω, ω', θ, θ' are the parameters to be estimated from the data. Clearly, this y is a highly nonlinear function of x, and with an appropriate choice of the parameters, it can represent a wide variety of nonlinear functions. With more hidden nodes, the representing equation has more parameters and is even more nonlinear, and therefore can represent a wider variety of functions. In fact, it has been shown that with a sufficient number of hidden nodes, one can represent any function (Hornik, Stinchcombe, and White 1989).

Although it is true that the NN represented by equation (1) is capable of approximating "any" function to any desired accuracy, the same is also true of many traditional methods, such as spline regression, polynomial regression, and projection pursuit (Sarle 1994b). Therefore, NNs are not to be considered a panacea. If there is any advantage that NNs have over other methods, it is the way in which they handle the problem of "the curse of dimensionality" (Bishop 1996; Ripley 1996). Briefly, the number of free parameters in polynomial regression, for example, increases exponentially with the number of independent variables. By contrast, the number of free parameters in an NN grows only linearly.[4] The "explosion" of the number of free parameters in polynomial regression makes it more prone to overfitting problems. On the other hand, the drastically smaller number of parameters in the NN renders it less likely to overfit data, yet it does not prevent it from approximating "any" function (Hornik et al. 1989).

A statistical model has little utility if it does not produce probabilities. Just as logistic regression models class-conditional posterior probabilities, the output of an NN can be arranged to represent class-conditional posterior probabilities. It has been shown that if the activation function is the logistic function, and if the error function being minimized is the cross-entropy, defined as

$$S = \frac{1}{N}\Sigma_i^N \left[t_i \log\left(\frac{t_i}{y_i}\right) + (1-t_i)\log\left(\frac{1-t_i}{1-y_i}\right)\right], \qquad (3)$$

then the output of the NN is the posterior probability of class membership, given the inputs (Richard and Lippmann 1991). In this equation, y_i is the output of the NN for the ith case and $t_i = 0,1$ are the values of the corresponding dependent variable labeling the two classes. This conclusion is contingent on a training set in which the class sizes are proportional to those of the sample (i.e., the a priori probabilities); if they are not, then the outputs must be corrected for the difference (Bishop 1996). However, many NN applications artificially equalize the class sizes in the training set with no such corrections (e.g., Anderer et al. 1994; Lenard et al. 1995).

The number of hidden nodes is one quantity that gauges the complexity of the underlying function (or classification boundaries, for classification problems), that is, the nonlinearity of the function and the complexity of the interactions between the independent variables. The magnitude of the weights is another quantity that affects the complexity of an NN, but to a much lesser degree.[5] By systematically varying the number of hidden nodes, one effectively spans the space of "all" functions and "all" interactions (Geman, Biensenstock, and Doursat 1992; Hornik et al. 1989). Therefore, selecting the number of hidden nodes is tantamount to specifying or selecting the underlying model in its entirety. Consequently, the "correct" number of hidden nodes is of paramount importance in any NN development, and therefore great care must be taken to find its optimal value; an NN with a number of hidden nodes less than the optimal number can underfit the data, whereas overfitting can occur if the NN has more than the optimal number of hidden nodes. In both cases, the model's predictive capabilities are jeopardized. In spite of this, that value is selected in an ad hoc fashion in many applications (e.g., Hardgrave et al. 1994; Lenard et al. 1995; Markham and Ragsdale 1995; Warner and Misra 1996; Wilson and Hardgrave 1995).

There are a number of techniques for determining the optimal number of hidden nodes, but a popular method is bootstrapping (Efron and Tibshirani 1993). In it, the data are randomly divided into two sets: a training set for estimating the weights in the NN and a validation set for determining the optimal number of hidden nodes (or similar parameters). Note that contrary to some practices, the performance of a trained NN on the validation set is not a measure of its predictive (or generalization) capability; another data set—a test set—is required if an unbiased measure of generalization performance is desired.

An NN with zero hidden nodes is trained with the training set and its performance is gauged on the validation set. The number of hidden nodes is then incremented and the procedure repeated until the validation error begins to rise. In this way, one arrives at the number of hidden nodes that precludes overfitting the training set. However, because the validation error is used in arriving at the number of hidden nodes, this procedure leads to an NN that overfits the validation set. This is why the validation error is not an unbiased measure of generalization performance. To preclude overfitting the validation set, the original data set is divided again but with a different random partitioning into a training and a validation set, and the entire procedure is repeated again. The validation errors over the different random sets (or bootstrap trials) are then employed to compute an average interval and a confidence interval for the validation performance measures. The optimal number of hidden nodes is the value beyond which the average validation error begins to rise.

Another important issue in the training of NNs is that of the local minima of the error function. Most training algorithms (i.e., parameter estimation techniques) are iterative procedures in which randomly selected weights are slowly varied in an attempt to minimize the error function. Given that equation (1) is nonlinear in the weights, frequently the training algorithm gets trapped in a local minimum of the error function. Such an NN does not correctly represent the underlying structure of the data. The simplest way of dealing with this problem is the "brute force" way, namely, to repeat the entire training phase from a different random set of initial weights some number of times. There exist other methods for eluding local minima (Masters 1993), one of which (called simulated annealing) is employed in this article; however, these methods do not guarantee that a global minimum will be reached, and therefore it is well warranted to augment them with the brute force method.

LINEAR DISCRIMINANT ANALYSIS

As mentioned previously, NNs are not a panacea, and in fact if the NN methodology is implemented properly, it may turn out that a more restrictive model (in terms of the invoked assumptions) outperforms the NN. Two popular classification models are logistic regression and discriminant analysis. As previously mentioned, the former is implicitly implemented in an NN with zero hidden nodes and, therefore, will be treated as such in this article. A more restrictive variant of discriminant analysis is linear discriminant analysis (LDA); its linearity allows for the possibility of identifying the "best predictors" in the model. LDA has several explicit assumptions (Huberty 1994; McLachlan 1992): The data are assumed to be multivariate Gaussian (normal) and the different classes are assumed to have equal covariance matrices (homoelastic). As in logistic regression and NNs, the "output" of LDA is also the posterior probability of class membership, given the inputs.

To expose the explicit assumptions of LDA, it suffices to review the univariate case and the two-class case (labeled as 0 and 1). The basic equations of LDA are as follows: The likelihood functions for the two classes are assumed to be normal, that is,

$$L_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}},$$

where μ_i and σ_i are the mean and the standard deviation, respectively, of x in the ith class. According to Bayes's theorem, the posterior probability of belonging to class i is given by

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

where p_0, p_1 are the prior probabilities for the two classes. The discriminant function is then given by the logarithm (conventionally) of the ratio of the posterior probabilities, that is,

$$\log\frac{P_1(x)}{P_0(x)} = \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)x^2 - 2\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \left(\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2}\right) + \log\left(\frac{\sigma_0}{\sigma_1}\right) - 2\log\left(\frac{1-p_1}{p_1}\right).$$

This function is the basis of discriminant analysis. If it is negative, then x is classified as 0; otherwise, it is classified as 1. Due to its quadratic (in x) nature, the model is referred to as quadratic discriminant analysis. However, it can be seen that if $\sigma_0 = \sigma_1$ (i.e., the two classes are, or are assumed to be, equivariant), then the discriminant function is linear in x. It is this resulting equation on which LDA is based. Although only the univariate case is reviewed here, the advantage of LDA is in the case where there are several independent variables, in which case the coefficients of the linear terms can be interpreted (with some care) as a measure of the predictive strength of the corresponding independent variable.

Although the terms *training* and *validation* are not ordinarily used in non-NN circles, it can be seen that in the context of LDA, training can be thought of as the estimation of the means and the covariance matrices. Also note that the appearance of the prior probabilities in the discriminant function implies that the class sizes in the training set must be according to the priors; as in NNs, artificially equalizing the classes in the training set can rob the outputs from their probabilistic interpretation. The validation set would not be required, since there are no other parameters (such as the number of hidden nodes) to determine, although it could be employed to decide which model is optimal—the quadratic or the linear. Indeed, training can refer to the process of estimating the parameters of any model—NN, regression, discriminant analysis, and the like. The validation set can be employed (via bootstrapping) to select the optimal configuration of any of the models.

MEASURES OF PERFORMANCE

A quantity that is often selected in an ad hoc fashion is the measure of performance. A proper choice of the measure is especially important because it is entirely possible that method A will outperform method B in terms of one measure of performance but not in terms of another. In many NN applications, however, the quantity that is minimized is the mean square error, even though that choice is justified only if the probability density of the dependent variable is Gaussian (or at least continuous, or bell shaped), for only then will the parameter estimates coincide with the maximum likelihood estimates (Draper and Smith 1981). However, such NNs are often unjustifiably employed for classification problems in which the dependent variable is discrete, for example, binary, three-valued, and so on (e.g., Hardgrave et al. 1994; Lenard et al. 1995; Markham and Ragsdale 1995; Warner and Misra 1996; Wilson and Hardgrave 1995).

Another frequently neglected fact is the multifaceted nature of performance itself. Any comparison between the performance of one model and another model in terms of a scalar (one-dimensional) quantity is apt to be incomplete. Even a 2 × 2 contingency table representing the performance of the binary classification of two classes (of fixed sample size) has two degrees of freedom. Therefore, in a binary classification task, a faithful comparison would require at least two independent measures of classification performance. Many such measures are discussed in the literature (Goodman and Kruskal 1959; Hays 1973; Paik 1998).

Two types of performance measures must be distinguished: continuous and discrete. The former are computed directly from the (continuous) output of the NN, whereas the latter are computed from a contingency table formed by placing a (probability) threshold on the output. Cross-entropy (equation [3]) and mean square error are examples of the former, and percentage correct is an example of the latter.

It is important to examine the behavior of these measures in certain special situations, such as deviations from normality or small sample sizes. Such matters have been considered by Hammond and Lienert (1995) and by Parshall and Kromrey (1996). Another special, yet ubiquitous, situation arises when the class sample sizes are disproportionate (Paik 1998). For instance, the use of the commonly employed measure percentage correct is misleading if the classes are not equally represented in the data set. This is so because that measure does not take into account chance or random guessing and, as a result, even random classification can yield a 99.9-percent accuracy. Of course, this shortcoming is readily exposed if the statistical significance of the measure is considered, but often it is not.

From all the discrete measures examined in the above articles, two that appear to be relatively "healthy" were selected to gauge the performance of the models examined herein. They are Pearson's chi-square and the likelihood ratio chi-square, defined as

$$\chi^2 = \Sigma_{i=1}^{4} \frac{(C_i - E_i)^2}{E_i^2}$$

$$LR = \Sigma_{i=1}^{4} C_i \log\left(\frac{C_i}{E_i}\right),$$

where C_i are the elements of the 2 × 2 contingency table

$$C=\begin{pmatrix} C_1 & C_2 \\ C_3 & C_4 \end{pmatrix}=\begin{pmatrix} no.of\ 0s\ classifed\ as\ 0 & no.of\ 0s\ misclassified\ as\ 1 \\ no.of\ 1s\ misclassified\ as\ 0 & no.of\ 1s\ classified\ as\ 1 \end{pmatrix}$$

and E_i are the elements of the expected matrix, that is, the contingency table that would ensue upon random guessing,

$$E=\frac{1}{C_1+C_2+C_3+C_4}\begin{pmatrix} (C_1+C_2)(C_1+C_3) & (C_1+C_2)(C_2+C_4) \\ (C_3+C_4)(C_1+C_3) & (C_3+C_4)(C_2+C_4) \end{pmatrix}.$$

In what follows, both measures have been normalized so that a perfect classification of both classes (i.e., a diagonal contingency table) will yield a value of 1, whereas random classification (i.e., $C = E$) will yield a value of 0.

The continuous measure, S, is "superior" to the discrete measures χ^2 and LR, in that the output is not required to be binary. As a result, most of the present analysis was performed in terms of S. However, at the end of the analysis, χ^2 and LR were employed to assess the classification performance of both LDA and NN. Note that in contrast to S, χ^2 and LR are measures of "success," in that larger values imply better performance.

DATA

The data were taken from the first follow-up (1990) of the National Education Longitudinal Study (U.S. Department of Education, National Center for Education Statistics 1992), base year 1988. The 1990 student component collected basic background information about students' school and home environments; participation in classes and extracurricular activities; current jobs; and goals, aspirations, and opinions about themselves. This component also measured 10th-grade achievement and cognitive growth between 1988 and 1990 in the subject areas of mathematics, science, reading, and social studies. The 20,706 subjects were all 10th-grade students in the United States during the 1989-1990 school year. The sampling was done in a two-stage sampling process, distributed across 1,500 schools, involving the selection of a core group of students who were in the 8th-grade sample in 1988. Based on prior literature (Evans et al. 1996; Kendall-Tackett 1996; Simons et al. 1991; Watts and Wright 1990), 71 variables were selected as the independent variables (see the appendix) and the dependent variable was in-school suspension; 15,906 students were never suspended from school, and 2,169 were suspended one or more times.

Some amount of preprocessing of the data is almost always necessary, and even beneficial, before any analysis—LDA and NN alike:

- All observations (students) with any missing data were neglected.
- All categorical independent variables were discarded.[6]
- All the independent variables were standardized (*z* scores).

The Pearson correlation coefficients, *r*, between the 71 independent variables and the dependent variable are plotted in Figure 2. The numbers on the *x*-axis refer to the independent variables as enumerated in the appendix. The height of each bar in the graph is a measure of the linear correlation between the independent variable and the dependent variable. The utility of this figure is in allowing for the selection of the input variables that are most (linearly) correlated with in-school suspension. For example, it can be seen that the five variables numbered 17, 48, 50, 57, and 59 have the highest linear correlation with in-school suspension.

METHOD

The popularity of NNs has reached a level at which some well-known statistical packages now include NN routines (Sarle 1994a). The present project employed an NN that was developed by the National Severe Storms Laboratory for tornado detection (Marzban and Stumpf 1996), and the remainder of the analysis was performed in SAS (1989).

The NN was trained and validated on (1) all 71 variables; (2) all 71 variables, with equal-class representation; and (3) 5 of the original variables. The first experiment is the most simplistic, in that no further

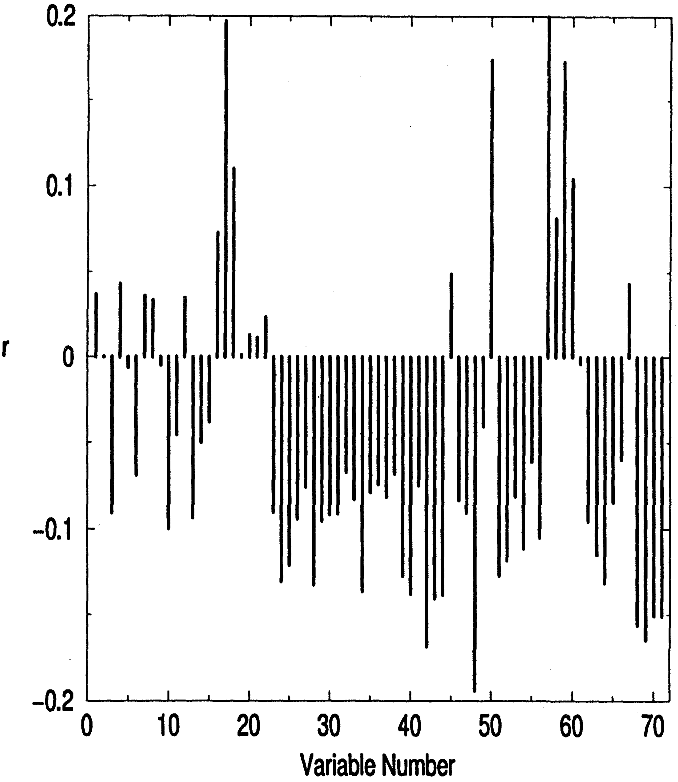


Figure 2: Pearson's Correlation Coefficient, *r*, Between the 71 Independent Variables and In-School Suspension

preprocessing of the data is performed. In experiment 2, the two classes were artificially equalized by including an equal number of 1s and 0s in the training set (but not in the validation set). This balancing of the classes is believed to enhance the validation performance of an NN (Masters 1993). As in LDA, it is important to emphasize that changing the class sizes in the training set robs the output from being interpreted as a posterior probability; that interpretation is valid only if the classes in the training set are represented according to their sample a priori probabilities. However, in this experiment, the output of the NN was transformed according to Bayes's theorem in order to recover the probabilistic interpretation (Bishop 1996). In experiment 3, to reduce the variance in the data, only 5 of the original variables that are mutually uncorrelated (Pearson's $r < .10$), yet most correlated with the dependent variable (Pearson's $r \sim .20$), were considered; these are the 5 variables with the longest bars in Figure 2. Finally, an LDA was performed.

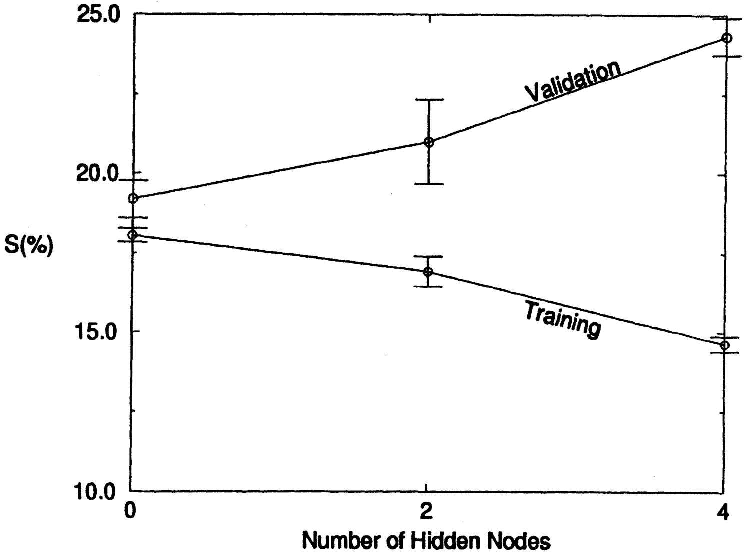
After the preprocessing, the remaining 18,075 cases were randomly partitioned into a training set (12,000) and a validation set (6,075), four times, for bootstrapping. The mean interval and the 90-percent confidence interval of the validation performance measure, *S*, over the four different validation sets were then computed. Finally, the two measures of classification performance—χ^2 and *LR*—were also computed; because these measures are discrete, their computation calls for a threshold placed on the output. The threshold was varied in .01 increments, and the validation performance measures were computed at each increment. In this way, one can identify the optimal value of the probability threshold and the corresponding value of the performance measure.

Further details of the training method can be found in Paik and Marzban (1995). For the technical reader, suffice it to say that simulated annealing and the brute force method (described above) were both employed to deal with local minima and that the training algorithm was the conjugate gradient method. When the improvement in the error function was less than .00001, training was halted (the stopping criterion) and then reinitiated with an entirely new set of random weights.

RESULTS

As mentioned previously, performance was gauged in terms of one measure of error, *S*, and two measures of classification success, χ^2 and *LR*. The results of the three different experiments with the NN are presented in Figures 3 and 4; the error bars on the various curves are the 90-percent confidence intervals. LDA is compared with NN in Figure 5; these measures have been computed from the validation data set.

(text continues on p. 444)



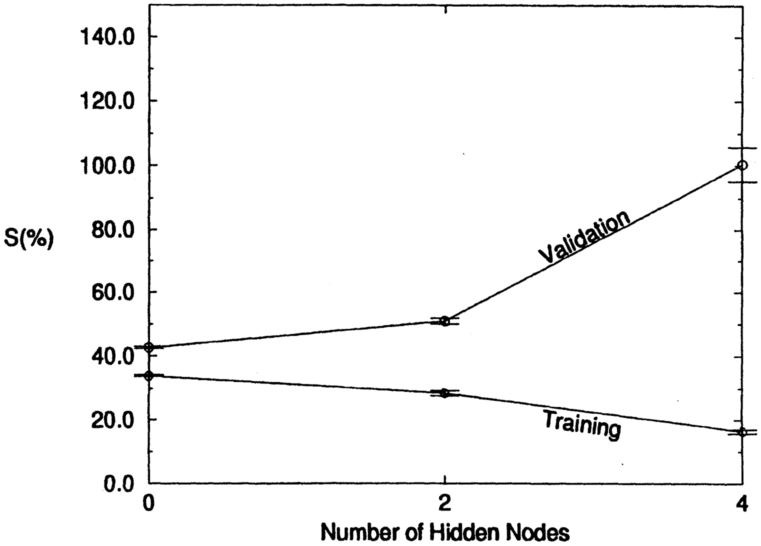
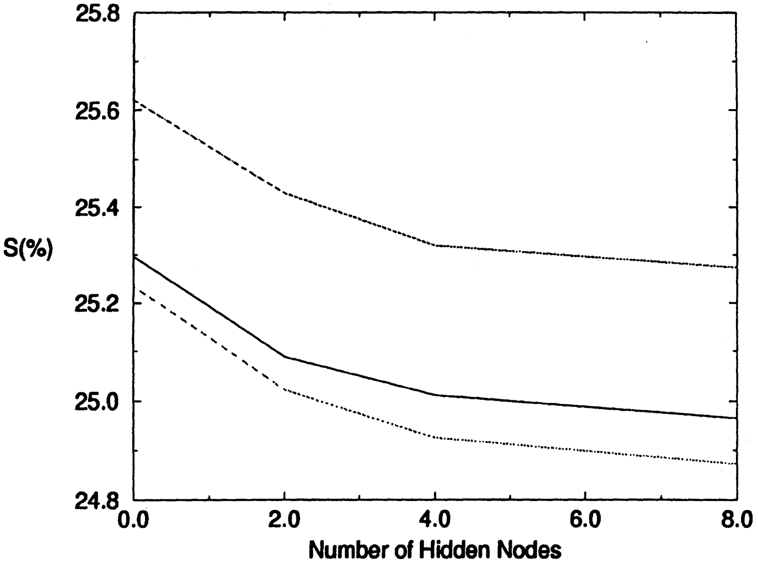


Figure 3: Cross-Entropy (*S*) as a Function of the Number of Hidden Nodes for Experiments 1 and 2, Averaged Over the Bootstrap Trials; Also Shown Are the 95-Percent Confidence Intervals



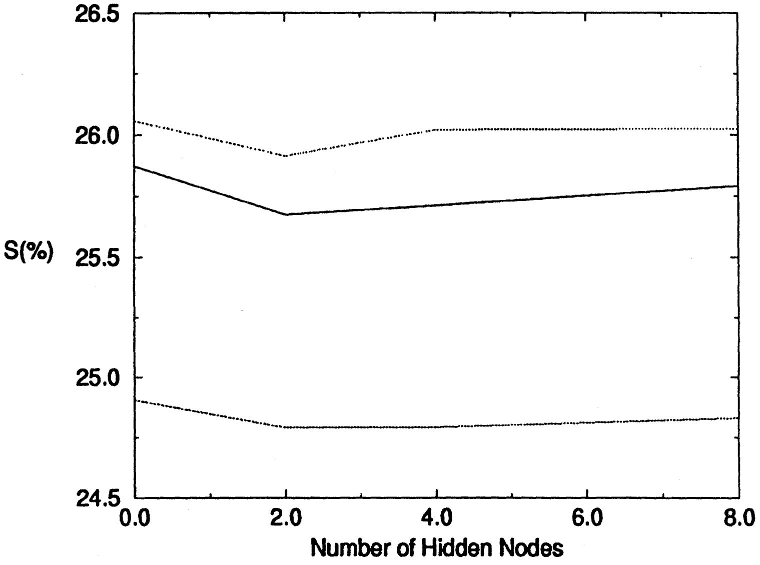


Figure 4: The Training (top) and Validation (bottom) Cross-Entropy (*S*) as a Function of the Number of Hidden Nodes in Experiment 3 for Three Bootstrap Trials

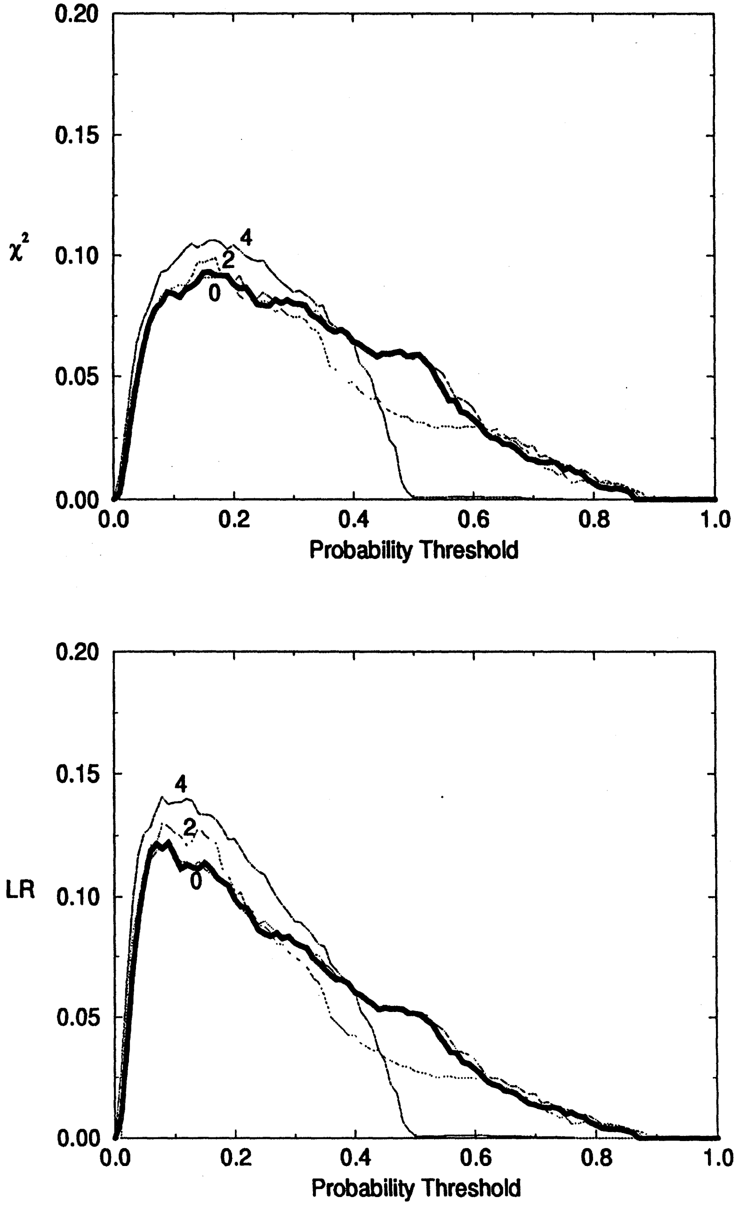


Figure 5: Training Classification Measures χ^2 (top) and *LR* (bottom) as a Function of Probability Threshold for Linear Discriminant Analysis and Neural Network With Zero, Two, and Four Hidden Nodes

From Figure 3, it is evident that the optimal number of hidden nodes is zero in experiments 1 and 2. Here, while all nonlinear NNs (i.e., with nonzero hidden nodes) have lower errors on the training set, the validation errors are higher. In other words, the nonlinear NNs overfit the data, implying that the underlying classification boundaries are mainly linear, and this is true regardless of the class sizes in the training set. Because an NN with zero hidden nodes—a linear NN—minimizing *S*, and with a logistic activation function, is equivalent to logistic regression, in experiments 1 and 2 the nonlinear NN has little to offer.

The nonlinear structure of the data is captured in the last experiment. Figure 4 shows the training and validation errors for a range of the number of hidden nodes and three of the bootstrap trials. It can be seen that for each trial, the optimal number of hidden nodes is two (Figure 4). Of course, the performance of this nonlinear NN is far below that of the linear ones in the first two experiments, but that is simply because the five-input-node NN has been deprived of the information in the larger number of input variables of the first two experiments. However, accompanying that information is more variance, and so with less variance in the last data set, the NN procedure does allow for the identification of nonlinear underlying relations.

The graphs in Figures 5 and 6 display the discrete measures of classification success for different values of the probability threshold. The measures in Figure 5 are computed from the training set, and the measures in Figure 6 are for the validation set; both figures pertain to the first experiment only because the performance of the NN in this experiment is superior to that in the other experiments. The dark curve is for LDA; the other curves are for the NNs with the corresponding number of hidden nodes. It can be seen that whereas the training performance appears to improve with more hidden nodes, the validation performance is reduced; this is true for both χ^2 and *LR*. As such, the nonlinear NNs overfit the data and have nothing to offer.

Furthermore, LDA performs comparably to the linear NN (i.e., with zero hidden nodes, or logistic regression). This is somewhat surprising given the data's violation of the explicit assumptions of normality and homoelasticity invoked in LDA. However, the robustness of LDA under violations of its assumptions is an attribute that is well known

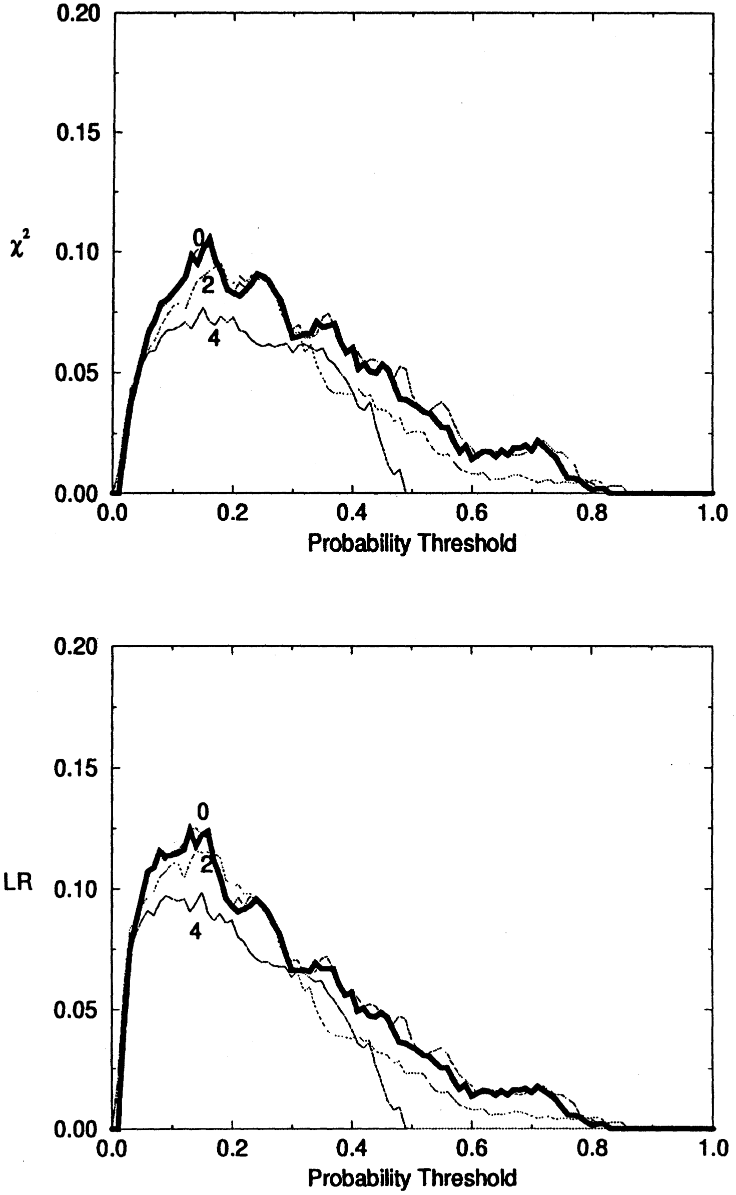


Figure 6: Validation Classification Measures χ^2 (top) and *LR* (bottom) as a Function of Probability Threshold for Linear Discriminant Analysis and Neural Network With Zero, Two, and Four Hidden Nodes

(Lachenbruch 1975). This example clearly illustrates how a model with many explicit assumptions may perform comparably to an NN.

CONCLUSION AND DISCUSSION

The findings illustrate that when appropriately implemented, the flexibility of NNs cannot always be used, and far simpler models can sometimes perform comparably to NNs. In many applications, the number of hidden nodes and the measure of performance are two quantities that are frequently selected inappropriately, or in an ad hoc fashion. With these two quantities appropriately determined, a rather typical social science data set is employed to show that the advantages of NNs, and their flexibility, do not automatically render them superior to other models in terms of three measures of performance and three different training procedures. The reason for the inability of the NN to outperform LDA is traced to the predominantly linear structure of the underlying relations when the data are too noisy, not the disproportionate class representation in the training set. As the amount of variance is reduced through the reduction of the number of input variables, the NN captures the underlying nonlinearities. However, the overall performance of the nonlinear NN is hindered due to the smaller number of input variables. As such, the advantages of NNs are impossible to realize in the particular data set examined here.

A question arises as to the conditions that may hinder the NN in identifying any nonlinearities in the data. It is possible that the underlying function or decision boundary is in fact linear. Under certain conditions, it is even possible to identify these situations. For example, as discussed in the LDA section, if the distribution of the independent variables is multivariate normal, and if the classes are equivariant, then the optimal (or Bayes's) boundary is in fact linear. Then, no classifier can outperform a linear one. Similarly, if the assumption of equivariance is significantly violated, then it can still be said that no classifier can outperform a quadratic one. In both cases, an NN is simply not required. What if the underlying linearity (or nonlinearity) cannot be established in such a fashion? In that case, consideration should certainly be given to the development of an NN model. However, as illustrated in this article, this should not discourage the examination of simpler models, since it is possible that the variance in the data is so large that the existence or the statistical significance of any nonlinearity cannot be established.

Further discussion of some flawed practices is offered. For instance, one common practice is to assess performance of an algorithm from the same data set (training set) that is used for estimating the parameters of the model (e.g., Anson and Sagy 1995; Cherry 1993; Christensen and Duncan 1995; Dannehl and Groth 1992; Famularo et al. 1992). The performance of any parametric algorithm, including NNs, on the training set is positively biased. Indeed, as described in this article, the performance of NNs on the validation set is also positively biased. A "third" test set is required for an unbiased assessment.

As already noted, an NN with zero hidden nodes and a logistic activation function is nothing but logistic regression if and only if cross-entropy is minimized. This is so because logistic regression models posterior probabilities, but an NN will model posterior probabilities only if cross-entropy is minimized. Indeed, it is the minimization of cross-entropy that yields the maximum likelihood parameter estimates (Bishop 1996). In spite of this, many NN applications to classification problems incorrectly minimize the mean square error (e.g., Hardgrave et al. 1994; Lenard et al. 1995; Markham and Ragsdale 1995; Warner and Misra 1996; Wilson and Hardgrave 1995).

As mentioned previously, the nonlinearity of the NN leads to the existence of local minima in the error function that is minimized. Frequently, however, NN applications neglect to consider this problem (e.g., Hardgrave et al. 1994; Lenard et al. 1995; Markham and Ragsdale 1995; Warner and Misra 1996; Wilson and Hardgrave 1995). The consequence of this practice is an NN whose performance has been compromised in at least two ways. First, an NN does not faithfully represent the underlying relations simply by virtue of being in a local minimum. Second, the optimal number of hidden nodes arrived at via bootstrapping may be incorrect. The latter may occur when one (or more) of the various NNs with a different number of hidden nodes is trapped in a local minimum, leading the bootstrapping procedure to identify the "wrong" number of hidden nodes.

Another issue worth mentioning is the absence, in this article, of any expression of classification performance in terms of the percentage of correctly classified cases. It can be shown that this measure is ill behaved when the class-conditional sample sizes are disproportionate (Paik 1998). The percentage of correctly classified cases can be written as $(C_1 + C_4)/(C_1 + C_2 + C_3 + C_4)$, where C_i are the elements of the 2×2 contingency table. Note that this expression approaches 100 percent if C_1 is much larger than the other three elements of the contingency table, that is, when one class is much larger than the other. Consequently, this measure overestimates the performance of the classifier, and for no skill-related reasons at all. In spite of this pathology, the percentage of correctly classified cases is a commonly employed measure of performance (e.g., Anderer et al. 1994; Azari et al. 1993; Bernard, McGrath, and Houston 1993; Boone 1991; Warner and Misra 1996).

Finally, a comment about the explanatory capabilities of NNs is in order. NNs have been referred to as black boxes, in that it is difficult or impossible to ascertain the rule that a trained NN represents. In other words, even if an NN performs superbly in predicting some phenomenon, it is quite difficult to decompose the function that the NN represents in terms of simpler, more "palatable" effects. Even determining the predictive strength of a given input is a complex task. There are many reasons for this opaqueness, some of which can be attributed not only to NNs but to any nonlinear model. One of the reasons is that the weights of an NN are almost entirely uninterpretable. First, in the presence of, say, hidden nodes (H), every input node has H weights connecting it to the hidden nodes. This is in contrast to the single weight emerging from an input node of an NN with no hidden nodes, or the single regression coefficient accompanying an independent variable in multiple regression. Under certain conditions, these single weights may be interpreted as the predictive strength of the corresponding variable, but what are we to do with many (H) weights associated with any given variable? To make matters worse, the values assigned to these weights (via training) depend greatly on the particular global minimum of the error function in which the NN has landed. In other words, the weights vary greatly from one minimum to another, whereas the overall performance of the NN in the same minima may be comparable. All of these problems can be traced back to the nonlinear nature of the activation function; in fact, any nonlinear model will suffer from the same problems. However, as mentioned above, there are problems that render the weight meaningless even in linear models. The most notorious of these is the presence of any collinearity among the independent variables. It is well known that such multicollinearity can render the weights uninterpretable even in linear multiple regression. In short, there are many good reasons for referring to NNs as black boxes, although many of the reasons are not peculiar to NNs and apply equally to many other (even linear) models.

APPENDIX
Definition of Independent Variables

1. Students get along well with teachers. 2. There is real school spirit. 3. Rules for behavior are strict at school. 4. Discipline is fair at school. 5. Students friendly with other racial groups. 6. Other students often disrupt class. 7. The teaching is good at school. 8. Teachers interested in student. 9. When student works hard teachers praise effort. 10. In class often feel put down by teachers. 11. Often feel put down by students in class. 12. Most teachers listen to me. 13. It doesn't feel safe at this school. 14. Disruptions impede my learning. 15. Misbehaving students often get away with it. 16. Had something stolen at school. 17. Someone offered to sell me drugs at school. 18. Someone threatened to hurt me at school. 19. It's okay to work hard for good grades. 20. It's okay to ask challenging questions. 21. It's okay to solve problems using new ideas. 22. It's okay to help students with school work. 23. It's okay to be late for school. 24. It's okay to cut a couple of classes. 25. It's okay to skip school a whole day. 26. It's okay to cheat on tests. 27. It's okay to copy someone's homework. 28. It's okay to get into physical fights. 29. It's okay to belong to gangs. 30. It's okay to make racist remarks. 31. It's okay to make sexist remarks. 32. It's okay to steal belongings from school. 33. It's okay to destroy school property. 34. It's okay to smoke on school grounds. 35. It's okay to drink alcohol at school. 36. It's okay to use drugs at school. 37. It's okay to bring weapons to school. 38. It's okay to abuse teachers. 39. It's okay to talk back to teachers. 40. It's okay to disobey school rules. 41. Time spent on homework in school. 42. Time spent on homework out of school. 43. How important are good grades to me. 44. Time spent on extracurricular activities. 45. Visit friends at local hangout. 46. How far in school father wants me to go. 47. How far in school mother wants me to go. 48. How far in school I think I will go. 49. Number of close friends now friends in eighth grade. 50. Number of close friends who dropped out. 51. Among friends, how important is to attend classes regularly. 52. Among friends, how important is to study. 53. Among friends, how important is to play sports. 54. Among friends, how important is to get good grades. 55. Among friends, how important is to be popular with students. 56. Among friends, how important is to finish high school. 57. How many cigarettes do you smoke a day. 58. Last 12 months number of times you drank alcohol. 59. Last 12 months number of times you used marijuana. 60. Last 12 months number of times you took cocaine. 61. I think of myself as a religious person. 62. Sex. 63. Socioeconomic status. 64. Parent's highest education level. 65. Locus of control. 66. Self-concept. 67. Entire school enrollment. 68. Reading standardized score. 69. Math standardized score. 70. Science standardized score. 71. History/geography standardized score.

NOTES

1. A model is said to overfit a data set if it is dominated by the statistical fluctuations in the data rather than the underlying function. Intuitively, an overfitted model "wiggles" more than it should. A more precise definition will be given later. For now, suffice it to say that a model that overfits a given data set has little or no predictive capabilities.

2. Two points are worth mentioning. First, technically, a neuron and a node are different entities. The former refers to a biological unit, whereas the latter is an abstract representation thereof. If modeling the neuron is the task at hand, then the distinction is an important one. In the present context, however, the distinction is irrelevant, and so the two terms will be used interchangeably. Second, the various components of an NN do not have unique names; nodes and weights, and so on, are the common terminology in the statistical applications of NNs.

3. Three variants of this layered structure are worth mentioning. First, the activation functions for the two layers, that is, the two *f*s in equation (1), may be different. For example, in a regression problem the dependent variable does not necessarily lie in the range 0-1, whereas the logistic activation function is restricted to that range. Therefore, the activation function for the nodes in the hidden layer is taken to be the logistic function, but that of the output layer is taken to be a linear function of the form $f(x) = ax + b$. Second, it is possible to connect the input nodes not only to the hidden nodes but also directly to the output nodes. This allows for explicit linear and nonlinear terms in the model. Finally, it is possible to have more than one hidden layer. Although it has been argued that one hidden layer is sufficient for learning "all" functions (Bishop 1996), an additional hidden layer can sometimes perform some preprocessing of the inputs, such as transforming the inputs to z scores, if the user has not already done so.

4. Consider an NN as shown in Figure 1. There exist two sets of parameters (1) between the input and the hidden layer and (2) between the hidden and the output layer. However, only the former involves input nodes (i.e., the independent variables), wherein there are $n \times H$ parameters. Therefore, the number of parameters grows linearly with n.

5. The reason the magnitude of the weights can affect the nonlinearity of the NN is that the logistic function $1/(1 + \exp(-\omega x))$ is highly nonlinear for large values of ω but linear for small values.

6. A proper inclusion of categorical variables requires the use of extra dummy variables (Draper and Smith 1981). However, to keep the NN analysis as simple as possible, categorical variables were not considered as inputs.

REFERENCES

Anderer, P., B. Saletu, B. Klöppel, H. V. Semlitsch, and H. Werner. 1994. "Discrimination Between Demented Patients and Normals Based on Topographic EEG Slow Wave Activity: Comparison Between Z Statistics, Discriminant Analysis and Artificial Neural Network Classifiers." *Electroencephalography and Clinical Neurophysiology* 91:108-17.

Anson, O. and S. Sagy. 1995. "Marital Violence: Comparing Women in Violent and Nonviolent Unions." *Human Relations* 48:285-305.

Artyushkin, V. F., A. V. Belyayev, Y. M. Sandler, and V. M. Sergeyev. 1990. "Neural Network Ensembles as Models of Interdependence on Collective Behavior." *Mathematical Social Sciences* 19:167-77.

Azari, N. P., P. Pietrini, B. Horwitz, and K. D. Pettigrew. 1993. "Individual Differences in Cerebral Metabolic Patterns During Pharmacotherapy in Obsessive-Compulsive Disorder: A Multiple Regression/Discriminant Analysis of Positron Emission Tomographic Data." *Biological Psychiatry* 34:798-809.

Bernard, L. C., M. J. McGrath, and W. Houston. 1993. "Discriminating Between Simulated Malingering and Closed Head Injury on the Wechsler Memory Scale—Revised." *Archives of Clinical Neuropsychology* 8:539-51.

Bishop, Christopher M. 1996. *Neural Networks for Pattern Recognition.* Oxford, UK: Clarendon.

Boone, S. L. 1991. "Aggression in African-American Boys: A Discriminant Analysis." *Genetic, Social, and General Psychology Monographs* 117:203-28.

Carley, Kathleen M. and David M. Svoboda. 1996. "Modeling Organizational Adaptation as a Simulated Annealing Process." *Sociological Methods & Research* 25:138-68.

Cherry, A. 1993. "Combining Cluster and Discriminant Analysis to Develop a Social Bond Typology of Runaway Youth." *Research on Social Work Practice* 3:175-90.

Christensen, L. and K. Duncan. 1995. "Distinguishing Depressed From Nondepressed Individuals Using Energy and Psychosocial Variables." *Journal of Consulting and Clinical Psychology* 63:495-98.

Churchland, Patricia S. and Terrence J. Sejnowski. 1992. *The Computational Brain.* Cambridge, MA: MIT Press.

Collins, E., S. Ghosh, and S. Scofield. 1988. *Risk Analysis: Darpa Neural Network Study.* Fairfax, VA: AFCEA International Press.

Dannehl, C. R. and A. J. Groth. 1992. "Communist and Non-Communist Europe: Functional Differentiation, 1970-1985." *Social Indicators Research* 27:59-87.

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

Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap.* London: Chapman & Hall.

Evans, T. David, Francis T. Cullen, V. S. Burton, Jr., R. G. Dunaway, G. L. Payne, and S. R. Kethineni. 1996. "Religion, Social Bonds, and Delinquency." *Deviant Behavior: An Interdisciplinary Journal* 17:43-70.

Famularo, R., T. Fenton, R. Kinscherff, R. Barnum, S. Bolduc, and D. Bunschaft. 1992. "Differences in Neuropsychological and Academic Achievement Between Adolescent Delinquents and Status Offenders." *American Journal of Psychiatry* 149:1252-57.

Garson, David G. 1991. "A Comparison of Neural Network and Expert Systems Algorithms With Common Multivariate Procedures for Analysis of Social Science Data." *Social Science Computer Review* 9:399-434.

Geman, S., E. Biensenstock, and R. Doursat. 1992. "Neural Networks and the Bias/Variance Dilemma." *Neural Computation* 4:1-58.

Goodman, L. A. and W. Kruskal. 1959. "Measures of Association for Cross Classifications." *Journal of the American Statistical Association* 54:123-63.

Hammond, S. M. and G. A. Lienert. 1995. "Modified Phi Correlation for the Multivariate Analysis of Ordinally Scaled Variables." *Educational and Psychological Measurement* 55:225-36.

Hardgrave, B. C., R. L. Wilson, and K. A. Walstrom. 1994. "Predicting Graduate Student Success: A Comparison of Neural Networks and Traditional Techniques." *Computers and Operation Research* 21:249-64.

Hays, W. L. 1973. *Statistics for the Social Sciences.* 2d ed. New York: Holt, Rinehart & Winston.

Hornik, K., M. Stinchcombe, and H. White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 4:251-57.

Huberty, C. A. 1994. *Applied Discriminant Analysis.* New York: John Wiley & Sons.

Kendall-Tackett, K. A. 1996. "The Effects of Neglect on Academic Achievement and Disciplinary Problems: A Developmental Perspective." *Child Abuse and Neglect* 20:161-69.

Lachenbruch, P. A. 1975. *Discriminant Analysis*. New York: Hafner Press.

Lee, V. E. and J. B. Smith. 1994. "High School Restructuring and Student Achievement: A New Study Finds Strong Links." Issue Report No. 7. (ERIC Document Reproduction Service No. ED 326-565)

Lenard, Mary Jane, Pervaiz Alam, and Gregory R. Madey. 1995. "The Application of Neural Networks and a Qualitative Response Model to the Auditor's Going Concern Uncertainty Decision." *Decision Sciences* 26:209-27.

Markham, Ina S. and Cliff T. Ragsdale. 1995. "Combining Neural Networks and Statistical Predictions to Solve the Classification Problem in Discriminant Analysis." *Decision Sciences* 26:229-42.

Marzban, Caren, Haejung Paik, and Gregory Stumpf. 1997. "Neural Networks vs. Gaussian Discriminant Analysis." *AI Applications* 11:1-10.

Marzban, Caren and Gregory Stumpf. 1996. "A Neural Network for Tornado Prediction Based on Doppler Radar-Derived Attributes." *Journal of Applied Meteorology* 35:617-26.

Masters, Timothy. 1993. *Practical Neural Network Recipes in C++*. San Diego, CA: Academic Press.

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

O'Brien, E. and M. Rollefson. 1995. "Extracurricular Participation and Student Engagement. Education Policy Issues: Statistical Perspectives." (ERIC Document Reproduction Service No. ED 384-097)

Paik, Haejung. 1998. "The Effect of Prior Probability on Skill in Two-Group Discriminant Analysis." *Quality and Quantity* 32:1-11.

Paik, Haejung and Caren Marzban. 1995. "Predicting Television Extreme Viewers and Nonviewers: A Neural Network Analysis." *Human Communication Research* 22:284-306.

Parshall, C. G. and J. D. Kromrey. 1996. "Tests of Independence in Contingency Tables With Small Samples: A Comparison of Statistical Power." *Educational and Psychological Measurement* 56:26-44.

Reardon, S. F. 1996. "Eighth Grade Minimum Competency Testing and Early High School Dropout Patterns." Presented at the annual meeting of the American Educational Research Association, April 8-12, New York. (ERIC Document Reproduction Service No. ED 400-273)

Rees, D. I., L. M. Argys, and D. J. Brewer. 1996. "Tracking in the United States: Descriptive Statistics From NELS." *Economics of Education Review* 15:83-89.

Richard, M. D. and R. P. Lippmann. 1991. "Neural Network Classifiers Estimate Bayesian A-Posteriori Probabilities." *Neural Computation* 3:461-83.

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

Sarle, Waren S. 1994a. "Neural Network Implementation in SAS Software." Pp. 1551-73 in *SAS Institute Inc. Proceedings of the Nineteenth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

———. 1994b. "Neural Networks and Statistical Models." Pp. 1538-50 in *SAS Institute Inc. Proceedings of the Nineteenth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc. Available: ftp.sas.com/pub/neural/neural1.ps

SAS Institute Inc. 1989. *SAS/STAT® User's Guide, Version 6*. 4th ed., vol. 1. Cary, NC: SAS Institute Inc.

Schrodt, Philip A. 1991. "Prediction of Interstate Conflict Outcomes Using a Neural Network." *Social Science Computer Review* 9:359-80.

Simons, R. L., L. B. Whitbeck, R. D. Conger, and K. J. Conger. 1991. "Parenting Factors, Social Skills, and Value Commitments as Precursors to School Failure, Involvement With Deviant Peers, and Delinquent Behavior." *Journal of Youth and Adolescence* 20:645-64.

U.S. Department of Education, National Center for Education Statistics. 1992. *National Education Longitudinal Study, 1988: First Follow-Up (1990)* [Student Data] [Computer File]. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement [Producer]. Ann Arbor, MI: Inter-University Consortium on Political and Social Research [Distributor].

Van Nelson, C. and Kathryn J. Neff. 1990. "Comparing and Contrasting Neural Net Solutions to Classical Statistical Solutions." Presented at the annual meeting of the Midwestern Educational Research Association, October 19, Chicago. (ERIC Document Reproduction Service No. ED 326-577)

Warner, B. and M. Misra. 1996. "Understanding Neural Networks as Statistical Tools." *American Statistician* 50:284-94.

Watts, W. D. and L. S. Wright. 1990. "The Relationship of Alcohol, Tobacco, Marijuana, and Other Illegal Drug Use to Delinquency Among Mexican-American, Black, and White Adolescent Males." *Adolescence* 25:171-81.

Wilson, Rick L. and Bill C. Hardgrave. 1995. "Predicting Graduate Student Success in an MBA Program Regression Versus Classification." *Educational and Psychological Measurement* 55:186-95.

Woelfel, Joseph. 1993. "Artificial Neural Networks in Policy Research: A Current Assessment." *Journal of Communication* 43:63-80.

Haejung Paik is an assistant professor in the Department of Communication at the University of Oklahoma. She received a Ph.D. in mass communications from Syracuse University in 1991. Her research places special emphasis on the application of various statistical methods to the study of complex, nonlinear systems.