An Information Criterion for Optimal Neural Network Selection

David B. Fogel, *Member, IEEE*

***Abstract*—Neural networks have been used to resolve a variety of classification problems. The computational properties of many of the possible network designs have been analyzed, but the decision as to which of several competing network architectures is "best" for a given problem remains subjective. A relationship between optimal network design and statistical model identification is described. A derivative of Akaike's information criterion (AIC) is given. This modification yields an information statistic which can be used to objectively select a "best" network for binary classification problems. The technique can be extended to problems with an arbitrary number of classes.**

I. Introduction

NEURAL networks are broadly applicable to classification problems. While there has been rigorous analysis of the properties of neural networks having diverse designs including feedforward connections, feedback connections, a variety of nonlinear processing nodes, and several suitable training procedures, the choice of an optimal network design for a given problem has remained an art. It is always the case that given sufficient layers and nodes, more than one network architecture can be trained to "solve" a given classification problem to within a given degree of error if given sufficient time. The designer is then left with the decision as to which network is "best," that is, which network will most fully achieve the user's purpose when placed in actual practice.

This requires a trade-off. The number of hidden nodes and layers must be sufficient to provide the discriminating capability required for the given application; however, if too many units and layers are employed the network becomes "brittle," capable of classifying the training data completely but incapable of generalizing between input patterns that are minor variations of the training samples. Further, choosing an excessively large number of hidden units may significantly increase the required training time for certain learning algorithms. The computational cost per training epoch may become prohibitive.

Gutierrez *et al.* [1] examined a procedure for estimating the required number of hidden units for fully connected networks consisting of an input layer, a hidden layer, and an output layer in which each node is connected to every node at the next layer. There were no connections between nodes in the same layer. The problem of binary mapping was addressed. A function of the number of "conflicts," sets of input–output relationships that require incompatible weight set solutions when the responses of an output unit are learned on a single hidden layer perceptron, was used as a estimate of the minimum number of hidden nodes required to perform the desired classifications. Gutierrez *et al.* noted that their estimate is a "heuristic estimation procedure developed through experimentation . . ." and remarked that their procedure may overestimate the number of hidden units by as much as 10%, but rarely underestimates this number. As noted above, it may not be desirable to have the network completely classify the training data precisely. The problem of selecting the best network from a collection of several contending networks was not addressed.

Kung and Hwang [2] devised an algebraic analysis for the optimal number of hidden units in a two-layer network using back-propagation learning. Given M totally "irregular" training patterns, the optimal number of hidden units is suggested to be approximately equal to $M - 1$. For training patterns that have inherent regularities in the input–output patterns, the optimal number of hidden units may be much lower. Again, while the algebraic projection analysis may yield a suitable number of hidden units to provide for complete and efficient training, the problem of choosing a suitable criterion for the trade-off between the number of weights being adapted and the overall goodness-of-fit of the network remained unanswered.

The problem of choosing the optimal number of nodes and layers is analogous to choosing an optimal subset of regressor variables in statistical model building. In optimal subset selection it is desired to include those independent variables that are relevant in explaining a significant percentage of the variation of the dependent variable, but it is recognized that the inclusion of a sufficient number of extraneous variables can completely "explain" the available data while yielding no insight into the underlying process. It is desired that the model include as many regressors as possible so that the "information content" in these factors can influence the predicted value of the dependent variable; yet it is also desired that the model include as few regressors as possible because the variance of the model's predictions increases as the number of regressors increases [3].

There have been several attempts to define appropriate

Manuscript received August 29, 1990; revised February 28, 1991.

The author is with the Orincon Corporation, 9363 Towne Centre Drive, San Diego, CA 92121.

IEEE Log Number 9144671.

1045-9227/91/0900-0490$01.00 © 1991 IEEE

criteria for optimal subset selection. One of the more commonly accepted methods is Mallows's C_p statistic [4]–[6],

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p$$

where $SS_E(p)$ is the sum of the squared errors between the observed value of the dependent variable and the value generated by the model having p independent variables, n represents the number of observations, and $\hat{\sigma}^2$ is an unbiased estimate of the variance of an error term. The expected value of C_p is p if the fitted model has negligible bias, and is greater otherwise. Generally, small values of C_p that are close to p are desired. While this statistic has a good theoretical development, it requires the subjective judgment on the use of an estimator $\hat{\sigma}^2$.

Barron [7] developed a criterion for polynomial network selection based on the sum of the training squared error (TSE) and an overfit penalty. Termed the PSE for predicted squared error, the statistic takes the form

$$\text{PSE} = \text{TSE} + 2\sigma_p^2 \left(\frac{k}{n}\right)$$

where σ_p^2 is a prior estimate of the true error variance that does not depend on a particular model being considered, k is the number of coefficients in the model, and n represents the number of samples. The network that achieves the least PSE is chosen to be "best."

Akaike introduced an information criterion (AIC) based on maximizing the mean log-likelihood of a model [8]. This statistic requires no subjective judgment to calculate and is broadly applicable to statistical identification problems in prediction, signal detection, classification, and other areas. A modification of this criterion can be used for optimal neural network selection. In this paper, a review of the AIC is given, along with a system identification example. This is followed by the application of an information criterion to neural network selection. A suitable pattern classification example is described. Finally, some theoretical and computational issues of the procedure are considered.

II. Background

A. Akaike's Information Criterion for Optimal Model Selection

The quantity

$$I(g(x); f(x)) = \int g(x) \ln \left(\frac{g(x)}{f(x)}\right) dx \qquad (1)$$

describes the mean information for discrimination between $g(x)$ and $f(x)$ [9]. If

$$S(g(x); f(x)) = \int g(x) \ln (f(x))\, dx \qquad (2)$$

then

$$I(g(x); f(x)) = S(g(x); g(x)) - S(g(x); f(x)). \qquad (3)$$

Consider $f(x)$ to be an approximation to some true density $g(x)$. $I(g(x); f(x))$ is positive for $f(x) \neq g(x)$ and zero for $f(x) = g(x)$. The greater the value of $I(g(x); f(x))$, the less "resemblance" there is between the two distributions. Maximizing $S(g(x); f(x))$ defines a resonable criterion for defining a best approximate model $f(x)$.

Akaike proposed an information criterion based on a weighted function of the fit of a maximum log-likelihood model and the number of independent parameters that are to be adjusted to maximize the likelihood [8]. The maximum likelihood estimate (MLE) is, under regularity conditions, asymptotically efficient [10]. Thus the likelihood function is a sensitive criterion of the deviation of the estimated model parameters from the true values. Let x_1 through x_N be N independent observations of a random variable with probability density function given by $g(x)$. If a parametric family of density functions is given by $f(x|\theta)$ with a vector parameter θ, the mean log-likelihood is

$$\left(\frac{1}{N}\right) \sum_{i=1}^{N} \ln (f(x_i|\theta)). \qquad (4)$$

As N approaches infinity, the mean log-likelihood tends to

$$S(g(x); f(x|\theta)) = \int g(x) \ln (f(x|\theta))\, dx \qquad (5)$$

with probability 1, where the existence of the integral is assumed.

An important characteristic of (5) is that its natural estimate (4) can be obtained without *a priori* knowledge of $g(x)$. In problems of plant identification, several families of $f(x|\theta)$ are typically given and it is required to select the best approximation. Each family may have a different form $f(x|\theta)$, or may have similar forms but restrictions on θ. When only one family $f(x|\theta)$ is given, maximizing the estimate (4) or (5) with respect to θ leads to the MLE $\hat{\theta}$.

Let $g(x) = f(x|\theta_0)$, where θ_0 is the true parameter set. Denote $I(g(x); f(x))$ and $S(g(x); f(x|\theta))$ by $I(\theta_0; \theta)$ and $S(\theta_0; \theta)$, respectively. When θ is sufficiently close to θ_0, $I(\theta_0; \theta)$ can be approximated [9] by

$$I(\theta_0; \theta_0 + d\theta) = (\tfrac{1}{2})\ \|d\theta\|_{J^2}$$

with

$$\|d\theta\|_{J^2} = (d\theta)' J (d\theta)$$

where the prime denotes a transpose and J is the Fisher information matrix,

$$J_{ij} = E\left\{\left(\frac{\partial \ln f(X|\theta)}{\partial \theta_i}\right)\left(\frac{\partial \ln f(X|\theta)}{\partial \theta_j}\right)\right\}.$$

Here J_{ij} denotes the (ij)th element of J, and θ_i is the ith component of θ. When the MLE $\hat{\theta}$ of θ_0 is close to θ_0, the variation of $S(\theta_0, \theta)$, that is, the deviation of $f(x|\theta)$ from the true distribution $f(x|\theta_0)$, will be measured by $(1/2)\|\theta - \theta_0\|_{J^2}$. Suppose the variation of θ for maximizing the likelihood is restricted to a lower p-dimensional subspace Θ of θ that does not include θ_0; that is, let the MLE $\hat{\theta}$ of θ_0 be restricted in Θ. Suppose $\boldsymbol{\Theta}$, which is also restricted in Θ, yields the maximum of $S(\theta_0; \theta)$ under the restriction. If $\boldsymbol{\Theta}$ is sufficiently close to θ_0, the distribution of $N\|\hat{\theta} - \boldsymbol{\Theta}\|_{J^2}$ for sufficiently large N is approximated under certain regularity conditions by a χ^2 distribution with the degree of freedom being equal to the dimension of the restricted parameter space [9], [11]. Therefore

$$E_\infty 2NI(\theta_0; \hat{\theta}) = N\|\boldsymbol{\Theta} - \theta_0\|_{J^2} + p \qquad (6)$$

where E_∞ denotes the mean of the approximate distribution and p is the dimension of $\boldsymbol{\Theta}$ or the number of parameters independently adjusted for the maximization of the likelihood [8]. For identification, it will be natural to choose the model that gives the minimum $EI(\theta_0; \hat{\theta})$. The relation (6) is based on the asymptotic distribution of $\sqrt{N}(\hat{\theta} - \theta)$ being approximated by a Gaussian distribution with mean zero and variance matrix J^{-1} [12]–[15]. If

$$2\left(\sum_{i=1}^{N} \ln f(x_i|\theta_0) - \sum_{i=1}^{N} \ln f(x_i|\hat{\theta})\right) \qquad (7)$$

is used as an estimate of $N\|\boldsymbol{\Theta} - \theta_0\|_{J^2}$, it requires a correction for the downward bias introduced by replacing $\boldsymbol{\Theta}$ with $\hat{\theta}$. This correction can be made by adding p to (7) [8]. Only a comparison of the values of the estimates of $EI(\theta_0; \hat{\theta})$ for various models is necessary for identification. Therefore, it is not necessary to utilize the term in (7) that includes θ_0.

Akaike [8] defined the information criterion AIC of θ to be

$$\text{AIC}(\hat{\theta}) = (-2)\ln(\text{maximum likelihood}) + 2p \qquad (8)$$

where "ln (maximum likelihood)" is the natural log of the likelihood when $f(x|\hat{\theta})$ is maximized, and p is the number of independently adjusted parameters to get $\hat{\theta}$. When there are several specifications of $f(x|\theta)$ corresponding to several models, the model yielding the minimum value of the AIC should be chosen as being "best."

B. An AIC Example

The AIC has been used in a variety of studies (e.g., [8], and [16]–[18]). To give an example of how the AIC can be used for model identification, let y be a function of a third-degree polynomial. Specifically, let

$$y_i = a_0 + a_1x_i + a_2x_i^2 + a_3x_i^3 + \epsilon_i$$

where the error terms ϵ_i are independent and identically distributed as $N(0, 1)$. Let $a_0 = 1$, $a_1 = -1$, $a_2 = -1$, and $a_3 = 1$. Suppose 200 samples of y are taken with x stepping at 0.01 increments over the interval (0, 2). The identification problem is to discover the underlying model that generated y, assuming $\epsilon \sim N(0, \sigma^2)$. Suppose attention is restricted to polynomials of degrees 1 through 5. Fig. 1 shows the true relationship, the noise sample, and the observed data and indicates the maximum likelihood (least-squares) fit to the data for each model along with its corresponding AIC score. This score was computed using the formula

$$\text{AIC}(p) = N\ln(SS_E/N) + 2(p + 2) \qquad (9)$$

where $N = 200$, $SS_E = \Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2$, and p is the degree of the model. The quantity $(p + 2)$ is used because there are p independent variable terms being estimated, along with the mean a_0 and the variance of the error term σ^2. The relation (9) is a simplified version of the complete AIC. The first term in (9) is the only term from the log-likelihood function that varies with each model and is therefore sufficient. As would be expected, the AIC is minimized for $p = 3$; thus, the third-degree model is shown to possess the "best" fit to the data.

As noted by Akaike [8], when models are specified by a successive increase of restrictions on the parameter θ of $f(x|\theta)$, minimizing the AIC takes the form of repeated applications of conventional log-likelihood ratio tests of goodness of fit with automatically adjusted levels of significance. As the underlying variance of the data increases for a fixed sample size, experimental evidence suggests the sensitivity of the AIC diminishes. Increasing the sample size increases the sensitivity of the procedure.

III. Applying an Information Criterion to Neural Network Selection

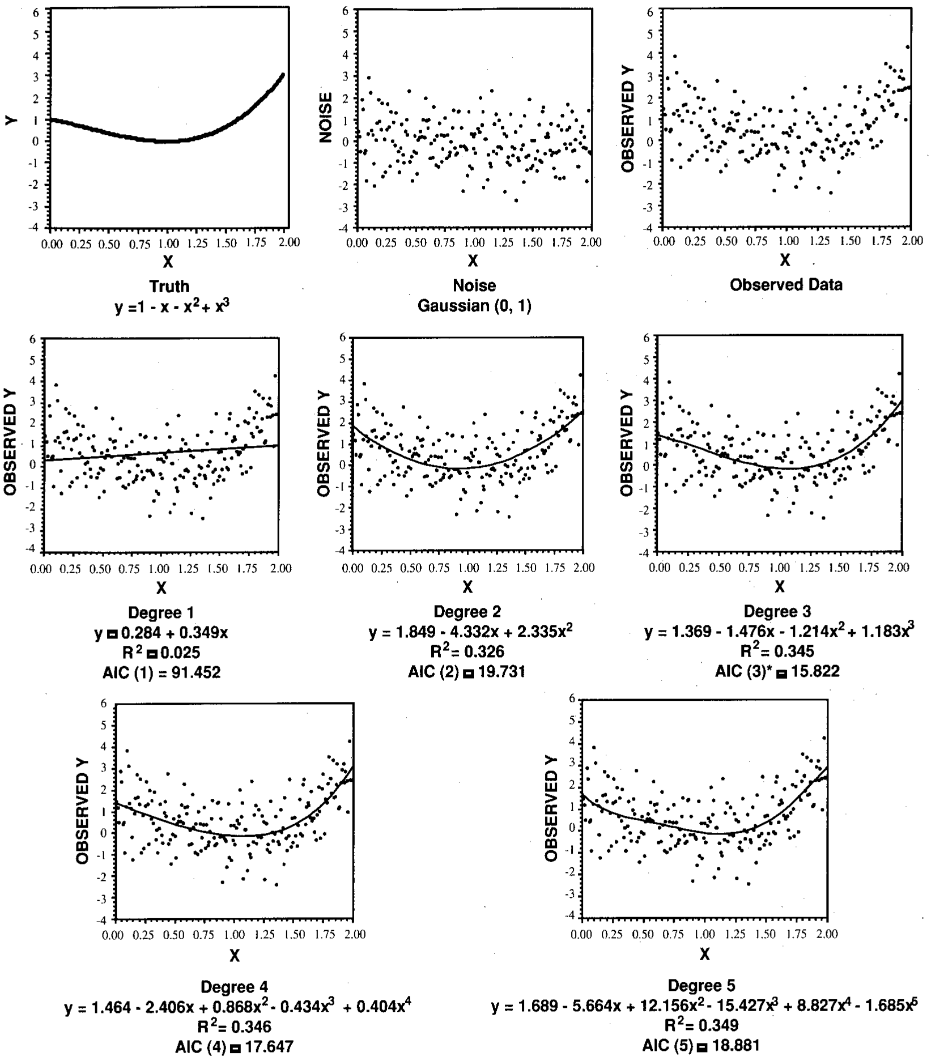
Consider the general neural network architecture given in Fig. 2. Let k represent the number of layers, n_k the number of nodes in the kth layer, and $-\beta_{ik}$ the output, that is, the weighted sum of the incoming signal strengths offset by a bias term, from the ith node in the kth layer. The output of each node, $-\beta_{ik}$, is passed through a standard nonlinear sigmoid filter $(1 + \exp(\beta_{ik}))^{-1}$. Suppose a binary decision is to be made about each pattern presented to the network. The ith training pattern, T_i, is assigned a corresponding network target output $t_i \in \{0, 1\}$ designating the desired classification. The network is then adapted using a suitable training algorithm, for example back-propagation or simulated annealing [19], or evolutionary programming [20] to minimize the sum of the squared differences between each t_i and each actual network output a_i,

$$\sum e_i^2 \qquad (10)$$

where

$$e_i = (t_i - a_i).$$

The final trained weights and biases of the network can be viewed as random variables. Each time the training process is restarted, it is likely that a different set of weights and biases will be discovered, although considerably different sets of weights and biases could lead to comparable classification. While minimization of (10) is a reasonable criterion, it presents the problem of overfitting the training data (i.e., using a greater number of weights than is suitable) to drive the value obtained in (10) to zero. If the patterns presented to the network are



Truth
y =1 - x - x^2+ x^3

Noise
Gaussian (0, 1)

Observed Data

Degree 1
y ■ 0.284 + 0.349x
R^2 ■ 0.025
AIC (1) = 91.452

Degree 2
y = 1.849 - 4.332x + 2.335x^2
R^2= 0.326
AIC (2) ■ 19.731

Degree 3
y = 1.369 - 1.476x - 1.214x^2+ 1.183x^3
R^2= 0.345
AIC (3)* ■ 15.822

Degree 4
y = 1.464 - 2.406x + 0.868x^2- 0.434x^3 + 0.404x^4
R^2= 0.346
AIC (4) ■ 17.647

Degree 5
y = 1.689 - 5.664x + 12.156x^2- 15.427x^3 + 8.827x^4 - 1.685x^5
R^2= 0.349
AIC (5) ■ 18.881

Fig. 1. An AIC example.

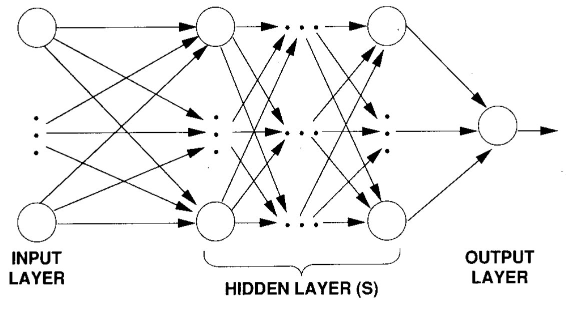


Fig. 2. General feedforward neural network.

themselves random variables, or have random components, overfitting the data can lead to excellent classification of the training data but poor performance on new data drawn independent of the training data.

Akaike's procedure can be adapted to provide a maximum likelihood criterion for the selection of an optimal neural network. Consider there to be several competing networks. Each is capable of classifying the input training patterns T_i (typically vectors of components that are random variables) with some degree of error. That is, no e_i is equal to zero. Denote ${}_t\beta_{1ki}$ to be the negative output of the final node at layer k for the ith pattern having a target value of $t \in \{0, 1\}$ before passing through the nonlinear sigmoid filter. By the central limit theorem, if there are sufficient nodes n_{k-1}, the asymptotic distributions of ${}_0\beta_{1ki}$ and ${}_1\beta_{1ki}$ are Gaussian with means $\mu_{\beta 0}$ and $\mu_{\beta 1}$ and variances $\sigma^2_{\beta 0}$ and $\sigma^2_{\beta 1}$, respectively. The distribution of the output of filter can be determined through a series of transformations [21].

Let β_t represent ${}_t\beta_{1k}$. Since $\beta_t \sim N(\mu_{\beta t}, \sigma^2_{\beta t})$, the distribution of $\beta'_t = \exp(\beta_t)$ is lognormal, that is,

$$f(\beta'_t) = (\sigma_{\beta t}\beta'_t(2\pi)^{1/2})^{-1} \exp\left(-\frac{(\ln \beta'_t - \mu_{\beta t})^2}{2\sigma^2_{\beta t}}\right).$$

Proceeding with the transformations, the resulting distribution of $a = (1 + f(\beta_i'))^{-1}$ is

$$f(a) = \left[\left(\frac{1-a}{a}\right)\sigma_{\beta t}\, a^2(2\pi)^{1/2}\right]^{-1} \cdot \exp\left(\frac{-\left(\ln\left(\frac{1-a}{a}\right) - \mu_{\beta t}\right)^2}{2\sigma_{\beta t}^2}\right).$$

Contruct a random variable

$$r_i = \begin{cases} 1 - a_i & \text{if } t_i = 1 \\ a_i & \text{if } t_i = 0 \end{cases}$$

where a_i is a sample from $a \sim f(a)$. Let $f(r_1)$ describe the density of the residual error r_i when $t_i = 1$. Let $f(r_0)$ describe the density of the residual error r_i when $t_i = 0$. It follows that

$$f(r_1) = \left[\left(\frac{r_1}{1-r_1}\right)\sigma_{\beta 1}(1 - r_1)^2(2\pi)^{1/2}\right]^{-1} \cdot \exp\left(\frac{-\left(\ln\left(\frac{r_1}{1-r_1}\right) - \mu_{\beta 1}\right)^2}{2\sigma_{\beta 1}^2}\right)$$

and

$$f(r_0) = \left[\left(\frac{1-r_0}{r_0}\right)\sigma_{\beta 0} r_0^2(2\pi)^{1/2}\right]^{-1} \cdot \exp\left(\frac{-\left(\ln\left(\frac{1-r_0}{r_0}\right) - \mu_{\beta 0}\right)^2}{2\sigma_{\beta 0}^2}\right).$$

Let there be P total patterns, with p_1 patterns having $t_i = 1$ and p_0 patterns having $t_i = 0$. If the training patterns T_i are independent, the likelihood function describing the joint density is

$$L(\mu_{\beta 1}, \mu_{\beta 0}, \sigma_{\beta 0}^2, \sigma_{\beta 1}^2) = \prod^{p_1} f(r_1) \prod^{p_0} f(r_0).$$

The natural logarithm of the likelihood function is defined by

$$\begin{aligned} \ln\left(\prod^{p_1} f(r_1)\prod^{p_0} f(r_0)\right) &= -\frac{p_1}{2}\ln 2\pi\sigma_{\beta 1}^2 - \frac{1}{2\sigma_{\beta 1}^2}\sum_{i=1}^{p_1}\left(\ln\frac{r_{1i}}{1-r_{1i}} - \mu_{\beta 1}\right)^2 \\ &\quad + \sum_{i=1}^{p_1}\ln\left(\frac{1}{r_{1i}(1-r_{1i})}\right) + \frac{-p_0}{2}\ln 2\pi\sigma_{\beta 0}^2 \\ &\quad - \frac{1}{2\sigma_{\beta 0}^2}\sum_{j=1}^{p_0}\left(\ln\frac{1-r_{0j}}{r_{0j}} - \mu_{\beta 0}\right)^2 \\ &\quad + \sum_{j=1}^{p_0}\ln\left(\frac{1}{r_{0j}(1-r_{0j})}\right). \end{aligned} \tag{11}$$

The MLE's of $\mu_{\beta 1}$, $\mu_{\beta 0}$, $\sigma_{\beta 1}^2$, and $\sigma_{\beta 0}^2$ maximize this statistic:

$$\begin{aligned} \hat{\mu}_{\beta 1} &= \frac{1}{p_1}\sum_{i=1}^{p_1} {}_1\beta_{1ki} \\ \hat{\mu}_{\beta 0} &= \frac{1}{p_0}\sum_{i=1}^{p_0} {}_0\beta_{1ki} \\ \hat{\sigma}_{\beta 1}^2 &= \frac{1}{p_1}\sum_{i=1}^{p_1} ({}_1\beta_{1ki} - \hat{\mu}_{\beta 1})^2 \\ \hat{\sigma}_{\beta 0}^2 &= \frac{1}{p_0}\sum_{i=1}^{p_0} ({}_0\beta_{1ki} - \hat{\mu}_{\beta 0})^2. \end{aligned} \tag{12}$$

The value determined by (11) is an intermediate quantity (IQ), which must be incremented by the number of parameters being adapted. The final information statistic (FIS) is

$$\text{FIS} = -\text{IQ} + w$$

where w is the number of independently chosen adapted weights and bias terms. To select the "best" network, choose the network that minimizes the FIS.

Note that as any residual r_{1i} or r_{0j} approaches zero, the IQ approaches infinity. That is, when there is precisely no error between the target output and the actual output for any given input pattern, the log-likelihood becomes infinitely large. Note also that as any residual r_{1i} or r_{0j} approaches 1, the IQ also approaches infinity. This should be considered because the sole goal of any suitable training algorithm is to minimize a function of the residuals (typically the sum of the squared residuals). It may be that the best solution is to drive some of the residuals to 0 at the expense of driving others to 1.

IV. An Application of the FIS to Neural Network Selection

To illustrate how the FIS may be used for neural network selection, the following classification problem was examined. Consider there to be two populations of interest, the first distributed $N(0, 1)$ and the second distributed $N(1, 1)$. Ten independent samples are all drawn from either of the populations, the objective being to properly identify the appropriate underlying distribution.

As an initial attempt at classification, a fully connected feedforward network consisting of ten input nodes, ten hidden nodes, and one final output classifying node was used ($k = 3$). The output of each node was found by passing $-\beta_{ik}$ through a standard nonlinear sigmoid filter $(1 + \exp(\beta_{ik}))^{-1}$. The ith training pattern, T_i, was assigned a corresponding network target output $t_i \in \{0, 1\}$ designating the desired classification. These target outputs were chosen to be identical to the mean of the population from which the ten independent samples were drawn. To train the network, an evolutionary programming algorithm [20] was used to minimize the sum of the squared differences between each t_i and each actual network output a_i as in (10).

The evolutionary algorithm trains by evolving an optimal population of weight and bias vectors. Each vector (organism) is assigned a score from (10). This is defined to be its fitness value. Each "parent" vector is mutated randomly in accordance with a Gaussian distribution having zero mean and variance proportional to its error score. These mutated parents, or "offspring," are then put in competition with the parents for survival to the next generation. As the process iterates, superior vectors evolve. For a more detailed description of evolutionary programming, see [20] and [22].

The population size was set at ten parent vectors, and the limit of evolution was chosen to be 1000 generations. A training set of 100 patterns, 50 from each population, was determined randomly and then fixed for the extent of the evolution. Fig. 3 shows the optimization in the mean squared error per pattern for the best evolved network. After 1000 generations the best network achieved a mean squared error per pattern of 0.0146 and correctly classified 98 of the 100 patterns.

Before calculating the FIS for this network, the distribution of the output of the final node before passing through the sigmoid filter, $-\beta_{13}$, was checked for normality. This assumption is required to formulate the appropriate likelihood function to be maximized in the FIS. Figs. 4 and 5 indicate the cumulative standardized distribution of the values of $-\beta_{13}$ for each input pattern vector from the $N(0, 1)$ and $N(1, 1)$ populations in relation to the cumulative distribution of a standardized Gaussian. The closer this relationship is to linear, the more closely each distribution resembles a Gaussian. A Kolmogorov–Smirnov procedure was utilized on each set of 50 values. Neither set yielded sufficient evidence to reject the hypothesis that the data were normally distributed ($P > 0.8$ for the $N(0, 1)$ and $P > 0.45$ for the $N(1, 1)$).

With the presumption of normality, the residuals for each input pattern were tabulated by class. The MLE's of $\mu_{\beta 1}$, $\mu_{\beta 0}$, $\sigma^2_{\beta 1}$, and $\sigma^2_{\beta 0}$ were calculated and the IQ was then calculated from (11). This value was determined to be 210.379. The FIS = $-$IQ + w, where w is the number of independently adjustable weights and bias terms, 121 in this case. Therefore, the FIS for this ten-hidden-node network is -89.379.

For comparison, an eight-hidden-node network was trained using the same training samples and evolutionary algorithm. Again, there were ten input nodes and one final classifying node. Fig. 6 indicates the optimization of the best eight-hidden-node network as the evolution was conducted. After 1000 generations, the final best evolved network possessed a mean squared error per sample of 0.0129 and classified 99 of the 100 training vectors correctly.

Figs. 7 and 8 indicate the relationships between the standardized distribution of the values of $-\beta_{13}$ for each input pattern vector from the two populations in relation to the cumulative distribution of a standardized Gaussian. Again, a Kolmogorov–Smirnov procedure was utilized on each set of 50 values. Neither set yielded sufficient evidence to reject the hypothesis that the data were normally distributed ($P > 0.44$ for the $N(0, 1)$ and $P > 0.76$ for the $N(1, 1)$).

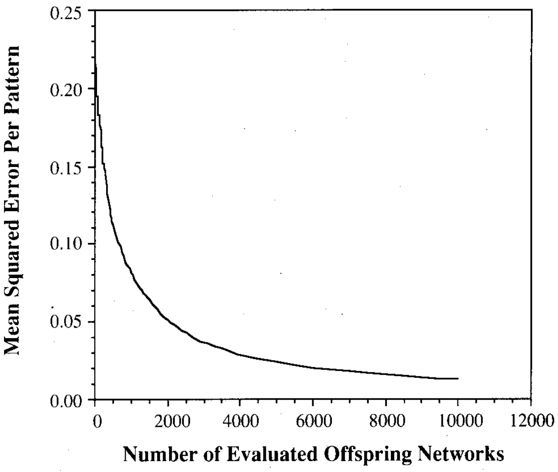


Fig. 3. Evolutionary optimization of best network; ten hidden nodes.

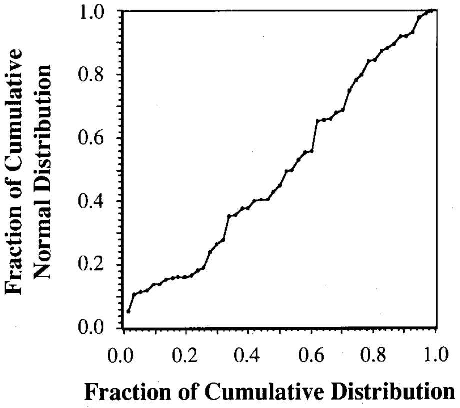


Fig. 4. Cumulative distribution of $-\beta_{13}$ for $N(0, 1)$; ten hidden nodes.

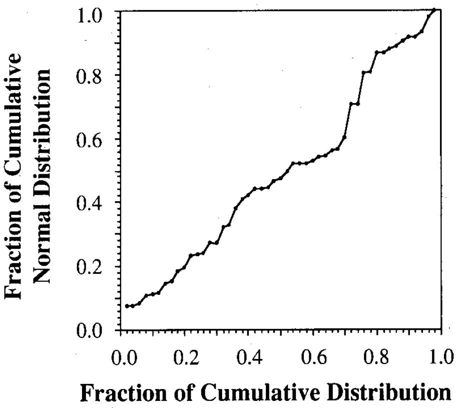


Fig. 5. Cumulative distribution of $-\beta_{13}$ for $N(1, 1)$; ten hidden nodes.

The residuals for each input pattern were tabulated by class. The MLE's of $\mu_{\beta 1}$, $\mu_{\beta 0}$, $\sigma^2_{\beta 1}$, and $\sigma^2_{\beta 0}$ were calculated and the IQ value was found to be 234.040. For this

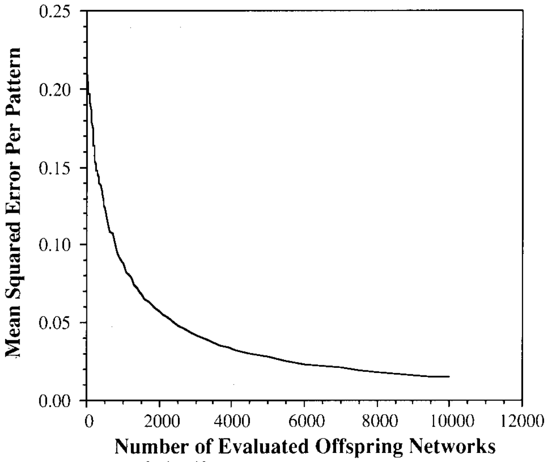


Fig. 6. Evolutionary optimization of best network; eight hidden nodes.

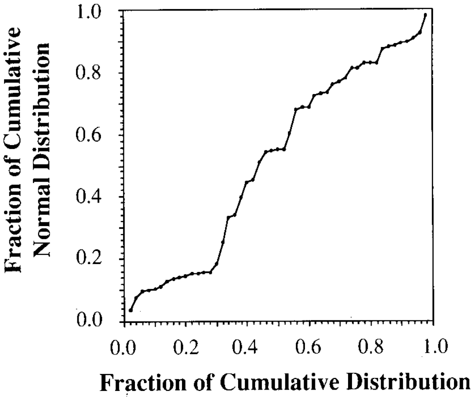


Fig. 7. Cumulative distribution of $-\beta_{13}$ for $N(0, 1)$; eight hidden nodes.

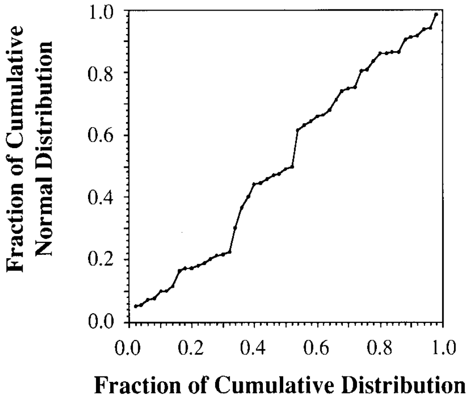


Fig. 8. Cumulative distribution of $-\beta_{13}$ for $N(1, 1)$; eight hidden nodes.

TABLE I
Comparing the Best Evolved Networks on Ten Sets of 400 Patterns

	8 Nodes		10 Nodes		Difference	
Trial #	MSE	%	MSE	%	MSE	%
1	0.0447	0.9375	0.0566	0.9300	−0.0120	0.0075
2	0.0549	0.9200	0.0640	0.9125	−0.0091	0.0075
3	0.0438	0.9450	0.0637	0.9150	−0.0199	0.0300
4	0.0616	0.9150	0.0764	0.8875	−0.0148	0.0275
5	0.0487	0.9425	0.0586	0.9200	−0.0098	0.0225
6	0.0462	0.9350	0.0567	0.9250	−0.0105	0.0100
7	0.0744	0.9025	0.0846	0.8825	−0.0102	0.0200
8	0.0623	0.9150	0.0750	0.9025	−0.0127	0.0125
9	0.0593	0.9150	0.0745	0.8925	−0.0152	0.0225
10	0.0566	0.9350	0.0726	0.8950	−0.0160	0.0400

MSE = mean squared error per pattern; % = fraction of patterns classified correctly.

Difference signifies the difference in performance between the eight-hidden-node and ten-hidden-node networks.

network architecture, there are 97 independently adjustable weights and bias terms, yielding an FIS for the eight-hidden-node network of −137.040. Therefore, the FIS procedure selected the eight-hidden-node network to be superior to the ten-hidden-node network.

A validation of this result was discovered by testing the final best evolved network for each architecture on new data taken independent from the training sample. A paired t-test experimental design was implemented. Ten trials were generated where each trial consisted of a new set of 400 total input pattern vectors, 200 from each population. The performance of the best evolved network on each trial is tabulated in Table I. The eight-hidden-node network outperformed the ten-hidden-node network in every trial. The one-sided t-tests of the difference in mean squared error per pattern and overall percent correct were both statistically significant, $P < 0.001$ in both cases.

Note that since the expected value of the sum of ten samples from a $N(0, 1)$ or a $N(1, 1)$ are zero and 10, respectively, a rational classification could be made simply by summing the components of each input vector and then comparing this value to 5. If the sum is greater than 5, a $N(1, 1)$ sample is more likely. Thus, a single perceptron could be viewed as a reasonable solution to this problem. The FIS result above conveys no assumption of the optimality of the eight-hidden-node network in comparison with all other possible designs; the result does indicate that the best evolved eight-hidden-node network should outperform the best evolved ten-hidden-node network.

V. Conclusions

The FIS provides an appropriate criterion for the comparison of neural networks. The use of a mean log-likelihood statistic has been demonstrated to be useful in a variety of statistical identification problems. It is natural to carry this procedure over to the optimal selection of a "best" neural network.

Other criteria may be similarly employed. If the underlying variance of the data is known, the AIC procedure [8] is equivalent to the PSE [7]. Schwarz [23], Rissanen [24], and Barron [25] have proposed other metrics based on different weightings of the log-likelihood of a model and the number of independently selected parameters. As observed by Barron [7], the choice of the appropriate metric is something of a philosophical question, but there are noted similarities in the form of these criteria.

Caution must be used when utilizing the FIS. The statistic is based on the assumption that the distribution of the negative weighted sum of the inputs to the final classifying node before passing through the nonlinear sigmoid filter is Gaussian. This assumption may be approximated for large networks, but may not hold well for certain small networks. If the density function in question can be estimated, then the FIS procedure can still apply, although the final statistic for the FIS will differ from that in (11). Further, the log-likelihood function was derived by assuming that the input patterns are statistically independent. If they are not, the joint density is no longer simply the product of the individual densities. The degree of correlation between the input patterns can aid in determining whether the assumption of independent patterns can be relaxed. The effects of statistically dependent training samples must be determined empirically.

There are possible computational difficulties in calculating the FIS. Should any value of the negative output of the final node before passing through the sigmoid filter be very large in magnitude, a computing machine may determine the value of the output from the filter to be identically equal to zero or 1. Thus, one or more residuals may be calculated as being exactly zero, causing the IQ to become indeterminable. This problem was not encountered in the FIS example given above. Moreover, the magnitude of the residuals is likely to be small for a well-trained appropriate network. Since the IQ requires calculating functions involving the inverse of these residuals, numerical roundoff errors could be present.

When applied correctly, the use of the FIS allows for virtually unrestricted training techniques. The method by which the final weights and biases are obtained does not enter into the FIS, although the criterion used for training must be reasonable. Moreover, the design of the network is almost unrestricted, provided that the distribution of the output of the final node can be determined or approximated.

While the FIS technique described here is limited to binary classification applications, the technique can be extended to multiple output binary classifications. If there are several output units, each having binary target values, a joint density function can be constructed for each set of outputs over all of the input patterns. This allows for the selection of a most appropriate network for problems involving an arbitrary number of classes.

Acknowledgment

The author would like to thank L. J. Fogel, T. W. Brotherton, D. L. Klamer, V. Samant, J. T. Rickard, V. W. Porto, A. V. Sebald, and J. W. Atmar for their helpful comments and suggestions. Thanks are also owed to the referees for their insightful reviews.

References

[1] M. Gutierrez, J. Wang, and R. Grondin, "Estimating hidden unit number for two-layer perceptrons," in *Proc. Int. Joint Conf. Neural Networks* (Washington, DC), 1989, pp. 677-681.

[2] S. Y. Kung and J. N. Hwang, "An algebraic projection analysis for optimal hidden units size and learning rate in back-propagation learning," in *Proc. IEEE Int. Conf. Neural Networks* (San Diego, CA), 1988, pp. 363-370.

[3] D. C. Montgomery and E. A. Peck, *Introduction to Linear Regression Analysis*. New York: Wiley, 1982.

[4] C. L. Mallows, "Choosing variables in a linear regression: A graphical aid," presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, KS, 1964.

[5] C. L. Mallows, "Choosing a subset regression," presented at the Joint Statistical Meetings, Los Angeles, CA, 1966.

[6] C. L. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, pp. 661-675, 1973.

[7] A. R. Barron, "Predicted squared error: A criterion of automatic model selection," in *Self-Organizing Methods in Modeling*, S. J. Farlow, Ed. New York: Marcel Dekker, 1984.

[8] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC 19, no. 6, pp.716-723, 1974.

[9] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[10] H. Cramer, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press, 1946.

[11] P. J. Huber, "The behavior of maximum likelihood estimate under nonstandard conditions," in *Proc. 5th Berkeley Symp. Math. Statistics and Prob.*, 1967, pp. 221-223.

[12] H. Cramer, *Random Variables and Probability Distributions* (Cambridge Tracts in Mathematics, no. 36) Cambridge: Cambridge University Press, 1937.

[13] H. Cramer, *The Elements of Probability Theory and Some of its Applications*. New York: Wiley, 1955.

[14] D. A. S. Fraser, *Nonparametric Methods in Statistics*. New York: Wiley, 1957.

[15] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968 (reprint).

[16] H. Akaike, "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting," *Biometrika*, vol. 66, no. 2, pp. 237-242, 1979.

[17] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, pp. 117-126, 1976.

[18] Q. Zhang, K. M. Wong, P. C. Yip, and J. P. Reilly, "Statistical analysis of the performance of information theoretic criteria in the detection of the number of signals in array processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1557-1567, 1989.

[19] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, vol. 1. Cambridge, MA: MIT Press, 1986.

[20] D. B. Fogel, L. J. Fogel, and V. W. Porto, "Evolving neural networks," *Biol. Cybern.*, vol. 63, pp. 487-493, 1990.

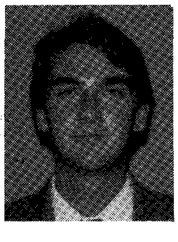
[21] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984.

[22] D. B. Fogel, "An evolutionary approach to the traveling salesman problem," *Biol. Cybern.*, vol. 60, pp. 134-144, 1988.

[23] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461-464, 1977.

[24] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465-471, 1978.

[25] A. R. Barron, "Complexity approach to estimating the order of a model," Electrical Engineering 378B Final Report, Information Systems Laboratory, Stanford University, 1982.



David B. Fogel (M'89) was born in San Diego, CA. He received the B.S. degree in mathematical sciences from the University of California at Santa Barbara and the M.S. degree in engineering sciences (system sciences) from the University of California at San Diego (UCSD). He is currently pursuing a Ph.D. in the Department of Applied Mechanics and Engineering Sciences at UCSD.

Mr. Fogel is a senior principal engineer at the Orincon Corporation, where he directs studies in evolutionary programming. His work has involved neural network design and training, sonar signal processing, combinatorial optimization, and automatic control systems. He has published several papers, and his book *System Identification through Simulated Evolution: A Machine Learning Approach to Modeling* will be published by Ginn Press.