Bounding the Performance of Neural Network Estimators, Given Only a Set of Training Data

Weibo Liang[1], Michael T. Manry[1], Qiang Yu[1], Steven J. Apollo[2], Michael S. Dawson[1], and Adrian K. Fung[1]

[1]Department of Electrical Engineering
University of Texas at Arlington
Arlington, Texas 76019

[2]Lockheed Fort Worth Company, Mail Zone 2615
P.O. Box 748
Fort Worth, Texas 76101

Abstract

In this paper, we use a neural network method for obtaining a stochastic Cramer-Rao bounds on estimates, given only the training data. The Cramer-Rao bounds can be used (1) to help determine when neural net training should be stopped, (2) to re-order the network inputs according to their contributions to the bounds, and (3) to eliminate less useful inputs. The convergence of the modelling procedure is shown. Examples are provided to illustrate the method.

I. Introduction

In previous papers [1,2], it has been shown that the multilayer perceptron (MLP) approximates the minimum mean-square estimator (mmse) [3], and that the MLP's training error is bounded by the Cramer-Rao MAP bound for the random parameter case, which is also called the stochastic Cramer-Rao bound. The bounds can also be used to help determine when neural net training should be stopped. The calculation of Cramer-Rao bounds for a given training data set requires a statistical signal model of the inputs as functions of the desired outputs. The lack of such signal models in most cases has prevented the widespread application of estimation theory, and specifically Cramer-Rao bounds, to neural networks.

In this paper, we propose a method for obtaining a statistical signal model, given only a set of training data. The convergence of the modelling procedure is shown. A method is proposed for re-ordering the sequence of input features according to their contributions to the Cramer-Rao MAP bounds. Examples are provided to illustrate the method.

II. Review

Let $(\mathbf{x}_p,\boldsymbol{\theta}_p)$ $p = 1, 2, \ldots, N_v$ represent the training set for a MLP. Here, $\mathbf{x}_p$ represents the pth example of the random input vector $\mathbf{x}$ and $\boldsymbol{\theta}_p$ represents the pth example of the random parameter vector $\boldsymbol{\theta}$. Our ultimate goal is to design an MLP which almost optimally estimates $\boldsymbol{\theta}$ from $\mathbf{x}$. In a previous paper [1], we motivated minimum mean square estimation via the MLP by showing that the training error for the MLP is minimized when the MLP outputs equal $E[\boldsymbol{\theta}|\mathbf{x}]$. This quantity, which is denoted by $\boldsymbol{\theta}_{MMS}$, is the minimum mean-square estimate of $\boldsymbol{\theta}$.

In MAP estimation [3], rather than estimating $E[\boldsymbol{\theta}|x]$ directly, one tries to maximize the conditional density $p_{\theta|\mathbf{x}}$ evaluated at $\mathbf{x}$ equal to the observation x. The MAP and MMS estimates are equivalent when $p_{\theta|\mathbf{x}}$ has its maximum at $\boldsymbol{\theta}_{MMS}$ [3]. Note that $p_{\theta|\mathbf{x}}$ can be expanded as

$$p_{\theta|\mathbf{x}}(\theta|x) = \frac{p_{\mathbf{x}|\theta}(x|\theta)p_{\theta}(\theta)}{p_{\mathbf{x}}(x)} \qquad (1)$$

The denominator of equation (1) can be ignored as it is a constant that depends only on the observation. For computational convenience we take the log of both sides of equation (1) to yield the log-likelihood function (LLF)

$$\Lambda^{MAP} = \Lambda^{MLE} + \Lambda^{AP} \qquad (2)$$

where $\Lambda^{MAP} = \ell n(p_{\theta|\mathbf{x}}(\theta|\mathbf{x}))$, $\Lambda^{MLE} = \ell n(p_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$, and $\Lambda^{AP} = \ell n(p_{\theta}(\theta))$. Equation (2) is maximized when

$$\nabla_{\theta}\Lambda^{MAP}(\mathbf{x},\ \theta = \theta_{MAP}) = 0\,, \qquad (3)$$

which is referred to as the MAP equation. The superscripts

1058-6393/95 $4.00 © 1995 IEEE

MLE and *AP* respectively stand for maximum likelihood estimation and a-priori.

Elements of the Fisher information matrix [3] (FIM), $\mathbf{J}_\theta^{MAP}$ are defined as

$$J_{ij}=E_\theta[E_x[\frac{\partial\Lambda^{MLE}}{\partial\theta_i}\frac{\partial\Lambda^{MLE}}{\partial\theta_j}]]+E_\theta[\frac{\partial\Lambda^{AP}}{\partial\theta_i}\frac{\partial\Lambda^{AP}}{\partial\theta_j}] \qquad (4)$$

In order to calculate the log-likelihood functions and the Cramer-Rao lower bounds on the variance of the parameter estimates, a statistical model of the input vector **x** is required. This model consists of a deterministic expression for the signal vector s in terms of the parameter vector θ, the joint probability density of the additive noise vector **n**.

III. Signal modelling from data

The problem of estimating a signal model from data has arisen and been partially solved many times in the past. In each case, severe constraints are placed on the model and data in order to make the problem solvable [4-6]. Given the training patterns we want to find the signal component model and noise pdf. We make the following assumptions.

(A1) The exact signal model is $\mathbf{x}_p = \mathbf{s}_p + \mathbf{n}_p$ where $\mathbf{x}_p$ is the noisy input vector, of dimension N, for the pth pattern, θ_p is the desired output vector of dimension M for the pth pattern, $\mathbf{s}_p$ denotes the noiseless signal component of the input vector $\mathbf{x}_p$, and $\mathbf{n}_p$ denotes the noise component of dimension the input vector $\mathbf{x}_p$.

(A2) The elements $\theta_p(k)$ of θ_p are statistically independent.

(A3) The noise vector **n** has independent elements with a jointly Gaussian pdf.

(A4) An expression exists for the signal component $\mathbf{s}_p$ in terms of the M given parameters.

The signal model above can be rewritten as $\mathbf{x}_p = \mathbf{s'}_p + \mathbf{n'}_p$ where $\mathbf{s}_p'$ and $\mathbf{n}_p'$ denote approximations to $\mathbf{s}_p$ and $\mathbf{n}_p$ respectively. The calculation of $\mathbf{s}_p'$ and $\mathbf{n}_p'$ are described separately.

Assume that the nth element of the approximate model $\mathbf{s}_p'$ are well approximated as

$$s_p'(n) = \sum_{k=1}^{L} a_{nk} \cdot T_p(k)$$

where a_{nk} denotes the coefficient of $T_p(k)$ in the approximation to s(n), and where $T_p(k)$ is the kth basis function calculated from the desired pth output vector θ_p. $T_p(k)$ can represent a multinomial function of parameter vector θ in a functional link network, or a hidden unit output in a MLP. The error between $x_p(n)$ and its model is measured as

$$E_p(n) = \frac{1}{N_v}\sum_{p=1}^{N_v}[x_p(n) - s_p'(n)]^2$$

The model is determined from the noisy data by setting the following partial derivative equal to zero.

$$\frac{\partial E_p(n)}{\partial a_{nk}} = \frac{-2}{N_v}\sum_{p=1}^{N_v}[x_p(n) - s_p'(n)]\frac{\partial s_p'(n)}{\partial a_{nk}}$$

$$= \frac{-2}{N_v}\sum_{p=1}^{N_v}[s_p(n) - s_p'(n)]\frac{\partial s_p'(n)}{\partial a_{nk}} + e_{nk},$$

$$e_{nk} = \frac{2}{N_v}\sum_{p=1}^{N_v} n_p(n)\frac{\partial s_p'(n)}{\partial a_{nk}}$$

Using the facts that

$$\frac{\partial s_p'(n)}{\partial a_{nk}} = T_p(k), \qquad E[n_p(n)n_q(n)] = \sigma_n^2 \cdot \delta(p-q)$$

the mean-square of the noise term e_{nk} is evaluated as

$$E[e_{nk}^2] = \frac{4}{N_v^2}\sum_{p}^{N_v}\sum_{q=1}^{N_v} E[n_p(n)n_q(n)]T_p(k)T_q(k)$$

$$= \frac{4\sigma_n^2}{N_v^2}\sum_{p=1}^{N_v}(T_p(k))^2$$

$$= \frac{4\sigma_n^2 \cdot E_T(k)}{N_v}$$

where $E_T(k)$ is the average energy of the kth basis function. Note that the mean-square error goes to zero in the limit as the number of training vectors increases.

Given a model $\mathbf{s}_p'$ for the signal component, we model the mean vector and covariance matrices of the noise component as

$$m_n' = \frac{1}{N_v}\sum_{p=1}^{N_v} x_p - s_p', \qquad (5)$$

$$C_{nn}' = \frac{1}{N_v}\sum_{p=1}^{N_v}(x_p - s_p' - m_n')(x_p - s_p' - m_n')^T$$

IV. Optimal subsetting of inputs

Elements of the FIM are numerically calculated as

$$J_{ij}^{MAP}(k) = \frac{1}{N_v}\sum_{p=1}^{N_v}\left[\frac{\partial s_p(k)}{\partial \theta_i}\cdot d(i,j)\cdot\frac{\partial s_p(k)}{\partial \theta_j}\right]$$

$$(i \neq j)$$

$$J_{ii}^{MAP}(k) = \frac{1}{N_v}\sum_{p=1}^{N_v}\left[\frac{\partial s_p(k)}{\partial \theta_i}\cdot d(i,i)\cdot\frac{\partial s_p(k)}{\partial \theta_i}\right] + b(i,i)$$

$$(1 \le i, j \le N, \qquad 1 \le k \le N_{out})$$

where b(i,j) denotes an element of the inverse matrix C_θ, d(i,j) denotes an element of the matrix C'^{-1}_{nn}, N denotes the number of inputs and N_{out} denotes the number of outputs. However, it is often desirable to find a good subset of the elements of x to process. This not only reduces the number of multiplications necessary for calculating the parameter vector θ but also reduces the time necessary for training the MLP. Our general approach is to find out which feature reduces the Cramer-Rao bounds the most, when it is added to the feature subset.

For simplicity, let's assume that C'_{nn} in equation (5) is a diagonal matrix. This is a good assumption when the elements of x correspond to transform coefficients. Also assume that we want to enlarge our set of inputs from k to (k+1) elements. The MAP FIM element for (k+1) input features can be written as

$$J_{ij}^{MAP}(k+1) = \sum_{p=1}^{N_v}\left[\sum_{l=1}^{k}\frac{\partial s_p(l)}{\partial \theta_i}\cdot\frac{\partial s_p(l)}{\partial \theta_j}\cdot d(l,l)\right.$$

$$\left.+\frac{\partial s_p(k+1)}{\partial \theta_i}\cdot\frac{\partial s_p(k+1)}{\partial \theta_j}\cdot d(k+1,k+1)\right] + b(i,j)$$

The difference between the MAP FIM elements for the cases of k and (k+1) coefficients is

$$J_{ij}^{MAP}(k+1) - J_{ij}^{MAP}(k) = \sum_{p=1}^{N_v}\left[\frac{\partial s_p(k+1)}{\partial \theta_i}\cdot\frac{\partial s_p(k+1)}{\partial \theta_j}\cdot d(k+1,k+1)\right] \quad (6)$$

which is easily calculated and stored ahead of time.

For the one parameter case, the inputs are easily optimally ordered. We simply add elements to our set which cause the largest decrease in the MAP bound, using the FIM changes generated by equation (6). When we have multiple parameters, it is necessary to define an objective function to help in the re-ordering process. We have chosen to use the weighted sum

$$O(k) \equiv \sum_{i=1}^{M} w_i \cdot B_{cr}(k,i)$$

where $B_{cr}(k,i)$ is the MAP bound on the ith parameter for the case of k inputs, w_i is a positive weight, and M is the number of parameters to be estimated. Now, the (k+1)th re-ordered input is that whose addition to our set causes the largest decrease in O(k+1).

V. Experimental Results

As a first example, we generated a signal Z(n) with the signal model

$$Z(n) = A\cdot e^{-\frac{n}{\tau}}\sin(\omega\cdot n) + n(n) \quad (7)$$

where n(n) was zero-mean Gaussian white noise with a standard derivation of .1. The time variable n varied from 0 to 127. The random parameters A and τ had uniform probability densities which extend from 1 to 2 and 10 to 20 respectively. The frequency ω had a value of .2 radian.

First, we calculated the Cramer-Rao MAP bounds on the variance of the A estimate for DFT feature data and optimally ordered DFT features. In the function O(k), $w_1=.1$ (for amplitude A) and $w_2=.9$. The optimal sequence of the row numbers for the DFT transformation matrix is 5,7,2,3,1,6,4,8. We used the original and re-ordered DFT transformation matrices to generate training data sets with 5,000 patterns each. Next we found MLP signal models from the training data and calculated the Cramer-Rao MAP bounds of the original feature sequence and the optimal feature sequence. For comparison, we found the same bounds using knowledge of the exact signal model. Using the training data, we also trained MLP estimators having topologies of the form M-20-10-2. In other words, there were M input transform coefficients where M varies between 2 and 8. The networks had 20 units in the first hidden layer and 10 units in the second hidden layer. The bounds and MLP testing error for A are shown in Fig. 1. The bounds and MLP testing error for τ are shown in Fig. 2.

The total bound, which is the sum of the bound for estimates of A and τ, is shown in Fig.3. Comparing the three figures, we can see that the bounds for the estimate of τ contribute much more to the total bound than does that of A.

As a second example, we use a set of remote sensing data with 16 inputs, 3 outputs and 10,000 patterns. We calculated the optimal feature sequence and its total MAP bound (sum of the 3 bounds), along with MAP bounds for the original feature sequence. We also trained MLP estimators for each subset of features and obtained the corresponding testing errors. The network topology used here is M-20-7-3, where M is the number of input features used. From the following table, we see that the bounds of the optimally ordered features are less than the bounds of the naturally ordered features for this set. Also, the MLP testing error is always greater than the sum of the bounds, as is to be expected.

VI. Conclusions

In this paper, we have developed a method for bounding the training error of MLP estimators, given training data but no statistical signal model. A method for finding the required signal model from the data has been presented and analyzed for convergence. The bounds have been calculated for two examples. In one example, the bounds were close to those found when the signal model is known beforehand. In the second example, the bounds provided a lower limit to MLP testing error.

More work remains before the methods presented here become easy to apply. For example, the topology of the MLPs was only guessed at in our simulations. A technique for estimating MLP topologies from training data would speed up the modelling, bounding and estimator design processes.

Acknowledgement

This work was funded by NASA under Grant NAGW-3091, by the NSF under grant IRI-9216545, by EPRI under grant RP 8030-09, and by a grant from the state of Texas.

References

[1] Q. Yu, S.J. Apollo, and M.T. Manry, "MAP Estimation and the Multilayer Perceptron," *Proceedings of the 1993 IEEE Workshop on Neural Networks for Signal Processing*, Linthicum Heights, Maryland, Sept. 6-9, 1993, pp. 30-39.

[2] S.J. Apollo, M.T. Manry, L.S. Allen, and W.D. Lyle, "Optimality of transforms for parameter estimation," *Conference Record of the Twenty-Sixth Annual Asilomar Conference on Signals, Systems, and Computers*, Oct. 1992, vol. 1, pp. 294-298.

[3] H. L. Van Trees, *Detection, Estimation, and Modulation Theory - Part I*, New York, NY: John Wiley and Sons, 1968.

[4] E.R. Cole, "The removal of unknown image blurs by homomorphic filtering," Ph.D. Dissertation, Department of Electrical Engineering, University of Utah, Salt Lake City, 1973.

[5] K.T. Knox, "Image retrieval from astonomical speckle patterns," *J. Opt. Soc. Am.*, vol. 66, November 1976, pp. 1236-1239.

[6] H.C. Luc and M.T. Manry, "Efficient calculation of signal and noise variances," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 1002-1004, August 1986.

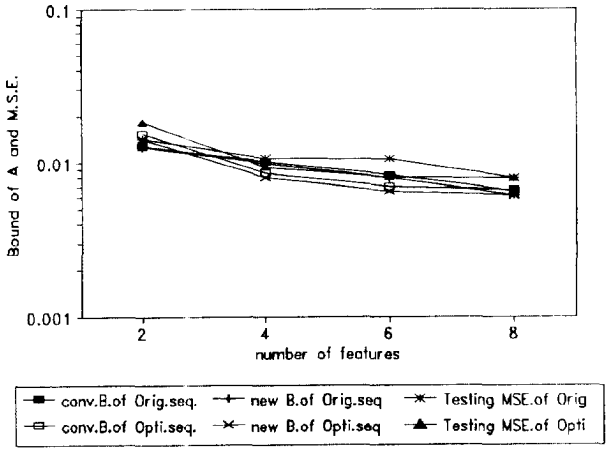

Figure 1. Bound and MLP Training MSE for A

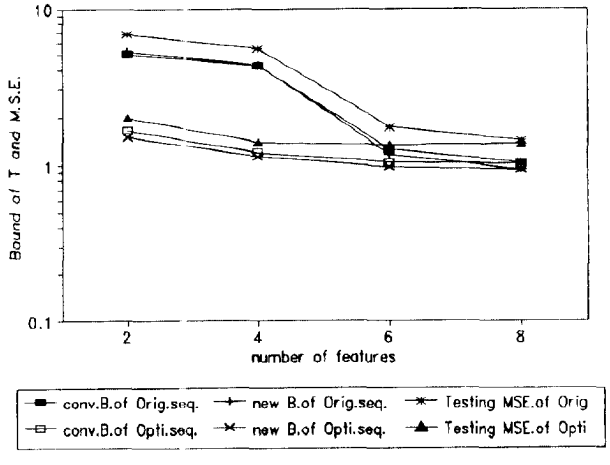

Figure 2. Bound and Training MSE for τ

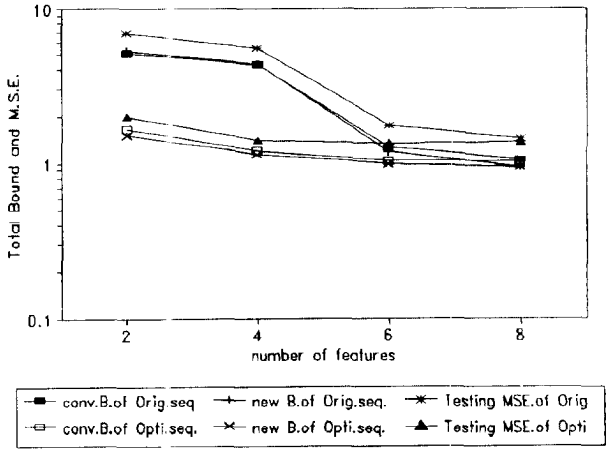

Figure 3. Total bound and MLP Training Error

Original Feature Sequence	MAP Bound of Original Feature Sequence	Optimal Feature Sequence	MAP Bound of Optimal Feature Sequence	Neural Network Testing M.S.E.
1	.2878582	8	.0323686	13.898120
2	.1432576	7	.0124901	10.590990
3	.0652140	5	.0094344	2.694916
4	.0563197	6	.0080898	1.789253
5	.0230884	12	.0073818	1.752958
6	.0156218	15	.0068454	1.037593
7	.0103243	13	.0064807	.968820
8	0068213	2	.0061556	.793245
9	.0068212	1	.0058699	.687170
10	.0065856	11	.0056195	.675128
11	.0062822	14	.0053922	.579369
12	.0058518	10	.0052263	.545192
13	.0055823	3	.0051335	.542235
14	.0053559	4	.0050716	.537011
15	.0050716	16	.0050141	.530456
16	.0050141	9	.0050141	.528107

Table Total of MAP bounds and neural net training error for a remote sensing problem