1° Congresso Brasileiro de Redes Neurais

Escola Federal de Engenharia de Itajubá
Itajuba. 24 a 27 de outubro de 1994

"STOCHASTIC PARAMETER ESTIMATION NEURAL NETS SUPERVISED LEARNING APPROACH"

Atair Rios Neto
Instituto Nacional de Pesquisas Espaciais-INPE
São José dos Campos, SP

ABSTRACT

In this short paper the problem of supervised learning of a neural net is treated as one of stochastic parameter estimation. The identification of the neural net parameters in its supervised training is as usually formulated as an optimization problem. A linear pertubation scheme, reducing the problem to one of stochastic parameter estimation in each iteration is then proposed. The structure of a Kalman filtering in the resulting algorithm allows the use of all the existing results in state and parameter estimation to guarantee a good performance to the proposed procedure.

1. INTRODUCTION

The experience with adaptative state estimation and control schemes based on the Kalman filter (e.g. Jazwinki, 1970) in aerospace applications (Rios Neto and Kuga, 1985; Rios Neto and Cruz, 1990) and in parameter estimation (Orlando and Rios Neto, 1984), has naturally led the author to consider these schemes in his recent work with neural nets in control.

Looking at the problem of neural nets supervised training in the modelling of nonlinear systems as one of parameter identification, and following an approach very much influenced by the one proposed by Chen and Billings (1992), this short paper tries to give an interdisciplinary contribution. It shows how to formulate and treat this problem as one of stochastic parameter estimation, with the objective of calling the attention of researchers in the field of neurocomputing for results of the area of state and parameter estimation that can be of great utility to them.

2. NEURAL NET SUPERVISED LEARNING APPROACH

The supervised training of a neural net to learn a nonlinear continuous mapping

$$f(.): x \in D \subset R^{nI} \rightarrow y \in R^{no} \ldots (2.1)$$

can naturally be treated as a problem of estimation the connection weight parameters p in the network correspondent mapping:

$$\hat{f}(.,p): x \in D \subset R^{nI} \rightarrow y \in R^{no} \ldots (2.2)$$

such as to have $\hat{f}(x, \bar{p})$ as close as possible to $f(x)$ for $x \in D$.

A set of pairs $(x(t), y(t))$, $t=1,2,\ldots N$, given by the mapping in (2.1) is selected in order to get this approximation, and the parameters are usually determined under the condition of minimizing

$$J(p)=\tfrac{1}{2}((p-\bar{p})^T\bar{P}^{-1}(p-\bar{p}) + \sum_{t=1}^{n} (y(t)-\hat{y}(t))^T R^{-1}(t)(y(t)-\hat{y}(t))) \ldots (2.3)$$

where $\bar{p}$ is a given a priori value of p; $\hat{y}(t)= \hat{f}(x(t),p)$; $\bar{P}^{-1}$ and $R^{-1}(t)$ are weight matrices.

To solve the problem of (2.3) an iterative scheme based on linear perturbation has to be used. In a typical iteration, one usually takes:

$$J(p^k)= \tfrac{1}{2}((p^k-\bar{p})^T\bar{P}^{-1}(p^k-\bar{p}) + \sum_{t=1}^{n}(\alpha^k(y(t)-\bar{y}^k(t))-\hat{f}_p(x(t),\bar{p}^k)(p^k-\bar{p}^k))^T R^{-1}(t) (\alpha^k(y(t)-\bar{y}^k(t))-\hat{f}_p(x(t),\bar{p}^k)(p^k-\bar{p}^k)) \ldots (2.4)$$

where $k=1,2,\ldots,k_c$; $\bar{p}^1=\bar{p}$, $\bar{y}^k(t)=\hat{f}(x(t),\bar{p}^k)$; $\hat{f}_p(x(t),\bar{p}^k)$ is the matrix of first partial derivatives with respect to p; $0<\alpha^k\leq 1$ is an adjusting parameter to guarantee the hypotesis of linear perturbation.

The solution of (2.4) is formally equivalent (see, for example, Jazwinki, 1970) to the following stochastic parameter estimation problem:

$$\bar{p}= p^k+\bar{e} \ldots (2.5)$$

$$\alpha^k(y(t)-\bar{y}^k(t))= \hat{f}_p(x(t),\bar{p}^k)(p^k-\bar{p}^k)+v(t) \ldots (2.6)$$

where, $E[\bar{e}]=0$, $E[\bar{e}\bar{e}^T]= \bar{P}$, $E[v(t)]=0$, $E[v(t)v^T(t)]= R(t)$, usually diagonal; $E[.]$ is the expextation value operator; $\bar{e}$ and $v(t)$ are assumed to be gaussian distributed and not correlated; and $v(t)$ is also assumed not correlated along $t=1,2,\ldots, N$.

Expressing this problem in a more compact notation, one has:

$$\bar{p}= p^k+\bar{e} \ldots (2.7)$$

$$z^k= H^k p^k + v \ldots (2.8)$$

where in (2.6) $\hat{f}_p(x(t), \bar{p}^k)\bar{p}^k$ has been taken to the left hand side and all the values of the index t were considered to define the extended z^k, H^k and v.

Thus, in a typical iteration, for $k=1,2,\ldots, kc$, the resulting parameter estimation problem can then be solved with a Gauss Markov estimator, which in a Kalman form results as:

$$\hat{p}^k = \bar{p} + K^k(z^k - H^k\bar{p}) \quad \text{........(2.9)}$$

$$P^k = (I - K^k H^k)\bar{P} \quad \text{..........(2.10)}$$

$$K^k = \bar{P}H^{kT}(H^k\bar{P}H^{kT}+R)^{-1} \quad \text{...(2.11)}$$

where $R=E[vv^T]$, diagonal; and to reiterate one takes $\bar{p}^{k+1}=\hat{p}^k$ in (2.6) until one gets:

$$\hat{p}^{kc} = \hat{p} \quad \text{................(2.12)}$$

$$P^{kc} \cong E[(p-\hat{p})(p-\hat{p})^T] \quad \text{..(2.13)}$$

If one further wants to explore the possibilities given by this approach, one should notice that:

(i) since the components of the error v in (2.8) are uncorrelated (R diagonal), the recursive algorithm of (2.9) to (2.11) can be used to process z^k componentwise, thus avoiding the need of matrix inversion in (2.11);

(ii) if it happens that a new set of pairs is to be considered in the training of the neural net, one only has to take the values in (2.12) and (2.13) as a priori information ($\bar{p} \leftarrow \hat{p}$, $\bar{P} \leftarrow P^{kc}$) to process the new information;

(iii) there are many well developed results to guarantee a good performance to the estimator of (2.9) to (2.11), as for example the factorization methods (Bierman, 1977) to improve numerical behavior and the methods of noise adaptive estimation (Jazwinski, 1970; Rios Neto and Kuga, 1985) to guarantee both good numerical behavior and statistical consistency.

3. CONCLUSIONS

Motivated by the possibility of giving an interdisciplinary contribution, the author has used his experience in stochastic control and parameter estimation to formulate the problem of supervised training of neural nets as one of parameter stochastic estimation.

This approach leads to an algorithm which is more general and more realistic than the usual backpropagation. If a full version (as in (2.9) to (2.11)) is used one looses part of the paralel processing characteristic of the backpropagation algorithm. However, as shown in Chen and Billings (1992), for algorithms with the structure of a recursive least squares it is possible to get simplified versions still more general than backpropagation but wich preserve the paralel processing capability.

It is expected that the approach of parameter stochastic estimation as presented in its full version will be useful in off line neural nets training, and will provide simplified versions which will be equally useful in on line, real time schemes.

4. REFERENCES

BIERMAN, G.J. "Factorization Methods for Discrete Sequential Estimation", Academic Press, 1977.

CHEN, S.; BILLINGS, S.A. "Neural Networks for Nonlinear Dynamic System Modelling and Identification", Int. J. Control, 56(2), 1992.

JAZWINSKI, A.H. "Stochastic Processes and Filtering Theory", Academic Press, N.Y., 1970.

ORLANDO, V.; RIOS NETO, A. "Aplicação de Técnica de Ruído Adaptativo em Estimação de Estado à Suavização de Dados", Procs. 1º Latin American Congress of Automatic and 5º Brazilian Congress of Automatica, Campina Grande, Paraíba, Brasil, 1984.

RIOS NETO, A.; CRUZ, J.J. "Design of a Stochastic Linear Discrete Regulator Based on an Optimal State Estimation Approach", SBA: Controle e Automação, 2(3), 1990.

RIOS NETO, A.; KUGA, H.K. "Kalman Filtering State Noise Adaptative Estimation", Procs. of Second IASTED Int. Conf. in Telecommunication and Control- TELECON'85, Rio de Janeiro, RJ, Brasil, 1985.