Quaderni di Statistica
Vol.1, 1999

Neural Networks as series estimators: a statistical interpretation of the hidden layer size

Francesco Giordano, Cira Perna
Dipartimento di Scienze Economiche, Università degli Studi di Salerno
Centro di Specializzazione e Ricerche, Portici (NA)
Email: giordano@unisa.it; perna@unisa.it

Summary In this paper it is shown that feed forward neural networks can be considered as series estimators. This allows to give a statistical interpretation to the parameters and to estimate them using standard statistical techniques. In particular, it is proved that the hidden layer size is related to a smoothing parameter and, therefore, it can be estimated minimising the mean squared error of a particular neural estimator, as proposed in a previous paper (Perna and Giordano, 1999).

Keywords Feed-forward Neural Networks, Series Estimators, Non linear Time Series,

1. Introduction

Let $\{Y_t\}$, t=1, ..., n be a time series generated according to the regression model:

$$Y_t=f(\mathbf{X}_t)+e_t \qquad t=1, ..., n \qquad (1)$$

where f is a non linear continuous function, $\mathbf{X}_t = (X_{1t}, X_{2t}, ...,X_{dt})$ is a vector of d non stochastic explanatory variables defined on a compact $\aleph \subset \Re^d$, and the $\{e_t\}$ are zero mean random variables with constant variance σ^2.

Neural networks can be used to estimate the function f because of their approximation properties. Many authors have demonstrated that, under general regularity conditions, a sufficiently complex single hidden layer feed-forward network can approximate any member of a class of functions to any degree of accuracy (Hornik *et al.*, 1989; Barron, 1993). The complexity of a single hidden layer feed-forward network is measured by the number of units in the hidden layer.

In this paper we prove that neural networks can be considered as series estimators. In this context, the parameters can be interpreted and determined using standard statistical techniques. In particular, we show that the hidden layer size can be considered as a smoothing parameter. This justifies the approach proposed in Perna and Giordano (1999), where this parameter is estimated on the minimisation of the mean squared error of a particular neural network

The paper is organised as follows. In the next section the feed-forward neural networks are introduced and the architecture used in the analysis is presented. In section 3, after a brief review of classical Fourier series estimators, we prove that Neural Networks are series estimators. In section 4, using the relationship between series estimators and a particular kernel estimator, we interpret the hidden layer size as a smoothing parameter and illustrate a procedure for the determination of this parameter based on the approach proposed in (Perna and Giordano, 1999). Finally, in section 5 some concluding remarks are presented

2. Feed-forward Neural Networks

An artificial Neural Network is composed by a multilayer of processing units. In the feed forward architecture each unit in the hidden layers receives information from the previous layer, processes it through a weighted sum of the input and, using a non linear activation function, transfers the result to the next layer units.

In the application of Neural Networks in time series analysis, a single hidden feed-forward neural network with one output unit is usually considered. This has the form:

$$Y_t^* = g\left(\sum_{k=1}^{m} c_k \phi\left(\sum_{j=1}^{d} a_{kj} X_{jt} + a_k \right) + c_0 \right) \qquad (2)$$

where c_k (k=1, ..., m) represents the weight of the connection between the k-th hidden unit and the output unit; a_{kj} is the connection between the *j-th* input unit and the *k-th* hidden layer unit. The parameters c_0 and a_1,..., a_m are the bias terms of the output unit and of the m units of the hidden layer. We can suppose that these constants are zero.

In (2) the functions g and ϕ represent the activation functions of the two layers; the first one concerns the output layer; the second is relative to the hidden layer. The function g can be chosen to be the identity function while the function ϕ must be a non linear function. It is almost always taken to be a sigmoidal function that is a bounded measurable function on $\Re$ for which $\phi(z) \to 1$ as $z \to \infty$ and $\phi(z) \to 0$ as $z \to -\infty$. Barron, (1993) has shown that feed forward networks with one layer of sigmoidal nonlinearities achieve integrated squared error of the order $O(1/m)$.

In this paper we have supposed that g is the identity function, as it usually happens in regression and in time series analysis, and ϕ is the standard Normal cumulative distribution function:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$$

Under these hypotheses the model (2) can be written as:

$$Y_t^* = \sum_{k=1}^{m} c_k \phi\left(\sum_{j=1}^{d} a_{kj} X_{jt} \right) \qquad (3)$$

Set:

$$\theta = \left(c_1, ..., c_m, \boldsymbol{a}_1', ..., \boldsymbol{a}_m' \right)$$

where $\boldsymbol{a}_k = (a_{k1}, \ldots, a_{kd})$ is the vector of the connection weights between the d input units and the j-th hidden layer unit.
In order to highlight the dependence of Y_t^* on the unknown parameters we can write:

$$Y_t^* = Y_t^*(\theta)$$

If we fix m and d, we can estimate the vector θ minimising the mean squared error function. The parameters are determined recursively from output to input by using a chain rule procedure known as backpropagation (Haykin, 1994; Lachtemacher and Fuller, 1995). Therefore if it is:

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{t=1}^{n} \frac{1}{2} \left(Y_t - Y_t^*(\theta) \right)^2$$

we obtain:

$$\hat{Y}_t = Y^*{}_t(\hat{\theta}) = \sum_{k=1}^{m} \hat{c}_k \phi \left(\sum_{j=1}^{d} \hat{a}_{kj} X_{jt} \right).$$

3. Feed forward neural networks as series estimators

To gain insight into the relationship between neural networks and series estimators it is necessary to review some definitions and basic results on classical Fourier series estimators.

Consider the model (1) and, without loosing in generality, suppose that d=1. The compact set $\aleph$ is assumed to be the interval $[0,1]$.
The *generalised Fourier series estimator* of the function $f(.)$ is defined as:

$$f_s^*(x) = \sum_{k=1}^{s} \beta_k Z_k(x) \qquad (4)$$

where the sequence of functions $Z_k(x)$ is a complete orthonormal system for $L_2[0, 1]$.

It can be shown (Eubank, 1988) that $\sum_{k=1}^{s} \beta_k Z_k(x)$ converges to the function $f(x)$, that is:

$$\left\| f(x) - f_s^*(x) \right\|^2 \to 0 \text{ as } s \to \infty.$$

Consider now the neural network estimator defined in (3); this can be rewritten as:

$$Y_t^* = \sum_{k=1}^{m} c_k Z_k(t)$$

where

$$Z_k(t) = \phi\left(\sum_{j=1}^{d} a_{kj} X_{jt} \right).$$

The sequence of functions $Z_k(t)$ is a complete system for $L_2[a, b]$ but is not orthonormal. Nevertheless, it can be shown (Barron, 1993) that $\forall f(.) \in C^2[a, b] \subset L^2[a, b]$ it is:

$$\left\| f(x) - Y_t^*(x) \right\|^2 \to 0 \text{ as } m \to \infty.$$

Thus neural networks can be considered as series estimators.

4. The interpretation and derivation of the hidden layer size

Using the results of the previous section, we show that the hidden layer size, that governs the approximation of the neural network to the unknown function, is related to a smoothing parameter. This justifies

the approach, proposed by Perna and Giordano (1999) who suggest estimating this parameter through the minimisation of the mean squared error of a particular neural network

At first, let us consider the estimator (4).

A complete orthonormal system for $L_2[0, 1]$ is provided by the complex exponential functions defined as:

$$Z_k(x) = \exp[(2\pi ikx)]$$

where $i^2 = -1$.

A useful expression (Eubank, 1988) for $f_s^*(x)$ is:

$$f_s^*(x) = \frac{1}{n}\sum_{r=1}^{n} Y_r \sum_{j=-s}^{s} \exp[2\pi ij(x - X_r)] =$$

$$= \frac{1}{n}\sum_{r=1}^{n} Y_r K_s(x - X_r)$$

where $K_s(x)$ is the Dirichelet kernel defined as:

$$K_s(x) = \frac{\operatorname{sen}[\pi(2s+1)x]}{\operatorname{sen}(\pi x)}.$$

Thus the generalised Fourier series estimator can be considered as a Priestley–Chao estimator with a particular kernel function. The parameter s is related to the number of elements in the bandwidth window and therefore it is:

$$s \cong 1/h$$

where h is the smoothing parameter

To determine the asymptotic optimal parameter h we can use the classical approach to the bandwidth selection problem based on the minimisation of the mean squared error. In this case it is:

$$h_{opt} = \left[\frac{\sigma^2 c_K}{n d_K^2 \int_{\chi} (f''(u))^2 du} \right]^{1/5}$$

where:

$$c_K = \int_{\chi} K(u)^2 du ; \qquad d_K = \int_{\chi} u^2 K(u) du ;$$

The latter expression permits to obtain the rate of convergence for s that is $O(n^{1/5})$.

For what concerns neural networks, they are series estimators and the parameter m, the hidden layer size, can be assimilated to the parameter s. In this case, since the functions $Z_k(x)$ are not orthonormal the relationship between m and h can be formulated only in this way:

$$m \equiv \tau(1/h)$$

where $\tau(.)$ is an appropriately chosen function.

Thus it is natural to make some assumptions similar to those usually formulated for kernel smoothing.
We suppose:

$$m=m(n);\ m(n)\to\infty;\ [m(n)]^2/n \to 0 \qquad \text{as } n\to\infty \qquad (5)$$

Thus, even in this case, this parameter can be estimated using the classical approach based on the minimisation of a particular mean squared error.
As proposed in Perna and Giordano (1999) we can consider the integrated mean squared error defined by:

$$IMSQ(\tilde{Y}_t) = \int MSE(\tilde{Y}_t) d\mathbf{X} = \text{IVAR}(\tilde{Y}_t) + \int \left(E(\tilde{Y}_t) - f(\mathbf{X}) \right)^2 d\mathbf{X} ,$$

where

$$\tilde{Y}_t = \sum_{k=1}^{m} \hat{c}_k \phi\left(\sum_{j=1}^{d} a_{kj} X_{jt} \right)$$

It can be shown (Giordano and Perna,1998; 1999) that :

$$IVAR(\tilde{Y}_t) \leq m^2 \frac{\sigma^2}{n} \frac{1}{\phi^2(x_0) b_d}$$

where

$$x_0 = \begin{cases} \min(a_k \cdot \mathbf{1}) Sup\mathbf{X} & \text{if } \min(a_k \cdot \mathbf{1}) < 0 \\ \min(a_k \cdot \mathbf{1}) Inf\mathbf{X} & \text{if } \min(a_k \cdot \mathbf{1}) > 0 \end{cases}$$

$\mathbf{1}$=(1, ..., 1) and b_d is the volume of the d-dimensional cube.
Using the Barron's approximation (Barron, 1993) it is:

$$\int \left(E(\tilde{Y}_t) - f(\mathbf{X}) \right)^2 d\mathbf{X} \leq \frac{c_f}{m}$$

where $c_f=(2rC)^2$, r is the radius of the compact set $\aleph$ and $C = \int_{\Re^d} |w| \left| \tilde{f}(w) \right| dw$ with $\tilde{f}(w)$ the Fourier transform of the function f.

An approximation for m can be obtained minimising the function:

$$\lambda(m) = \frac{L_0 \sigma^2}{n} m^2 + \frac{c_f}{m}$$

where:

$$L_0 = \frac{1}{\phi^2(x_0)b_d}$$

As shown in Perna and Giordano (1999) it is

$$m^* = \left(\frac{c_f}{2L_0 v^2} n \right)^{1/3}$$

5.Concluding remarks

The value m^* verifies the conditions imposed in (4). It is a function of c_f, L_0 and v^2.

The dependence on c_f, derives from the integrability of the squared partial derivatives of order two of the unknown function; therefore if the function is not sufficiently smooth this quantity increases. Since c_f is the bias component of the criterion IMSE, when it increases also m^* increases to guarantee a good fit of the estimated values to the observed ones.

For what concerns v^2, it is clear that a high value of this parameter implies a great perturbation in the data. This condition implies a decrease of m^* to guarantee the consistency of the neural estimators. Therefore, in this case even a very large hidden size does not produce a good fit to the data.

Finally, the relationship with h produces a dependence of m^* on the activation function ϕ.

It is interesting to compare the value m^* with $1/h_{opt}$ which is related to the number of elements in the bandwidth window. The quantity in the numerator of m^* depends only on the unknown function f, as it is for the denominator h_{opt} . The denominator of m^* includes two different components: a variance component and a quantity depending on the activation function. Similarly, the numerator of h_{opt} is given by

a combination of the same variance component in m^* and a second quantity depending only on the kernel function.

This implies that the activation function of the hidden layer is equivalent, in this context, to the kernel function.

Acknowledgments: The paper is supported by MURST98 "Modelli statistici per l'analisi delle serie temporali".
The work is joint responsability of the authors; F Giordano wrote sections 1, 2, 3 and C. Perna wrote sections 4, 5.

References

Barron, A.R. (1993) Universal Approximation Bounds for Superpositions of a Sigmoidal Function, *IEEE Transactions on Information Theory*, 39, 3, 930-945.

Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*, M. Dekker, New York.

Giordano, F; Perna C. (1998) Proprietà asintotiche degli stimatori neurali nella regressione non parametrica, *Atti della XXXIX Riunione Scientifica S.I.S.*, 2, 235-242

Giordano, F; Perna C. (1999) Large-sample properties of Neural Estimators in a Regression Model with φ-mixing errors, *Atti della Riunione Scientifica CLADAG99 "Classificazione ed Analisi dei Dati"*, 89-92

Haykin, S (1994) *Neural Networks: a comprehensive foundation*, Macmillan, New-York.

Hornik, K.; Stinchcombe, M.; White, H. (1989) Multy-Layer Feedforward Networks Are Universal Approximators, *Neural Networks*, 2, 359-366.

Lachtermacher, G.; Fuller, J.D. (1995) Backpropagation in Time-series Forecasting, *J. of Forecasting*, 14, 881-393.

Perna C. Giordano F. (1999) The Hidden Layer Size in Feed-forward Neural Networks: A Statistical Point of view, *Atti del Convegno SCO99, "Modelli complessi e Metodi Computazionali Intensivi per la Stima e la Previsione"*, 95- 100