ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 341-346

On the use of the wavelet decomposition for time series prediction

Skander Soltani
Laboratoire HEUDIASYC U.M.R. C.N.R.S. 6599
Université de Technologie de Compiègne
B.P. 20529, 60205 Compiègne Cedex, France
Tel: (33) 3 44 23 44 23 (ext. 4266), Fax: (33) 3 44 23 44 77
e-mail : Skander.Soltani@hds.utc.fr

Abstract. This paper deals with the problem of nonlinear time series prediction. The method uses a couple of filters to decompose iteratively the series. This scheme leads to a time series which contains the slowest dynamics and a hierarchy of detail time series which contain intermediate, up to the highest, dynamics. The new series are then used for modeling and predicting. The result obtained on the Mackey-Glass chaotic series show the efficiency of this approach.

1. Introduction

Let $X_1, X_2, \cdots, X_\ell$ be a stationary time series. Our objective is to predict the value of $X_{k+p},\ p \geq 1$ using all the observations until the instant k. For this purpose, a function (or a link) between the observations $\{X_1, X_2, \cdots, X_k\}$ and X_{k+p} is to be constructed with a principal concern in the prediction accuracy. Indeed, the optimal prediction sequence $\widehat{X}_1^*, \widehat{X}_2^*, \cdots$ minimize a criterion (the least squares in our case), i.e.

$$C_{gen} = \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} E\{(\widehat{x}_{k+p} - x_{k+p})^2 | x_k = X_k, x_{k-1} = X_{k-1}, \cdots\}. \quad (1)$$

The solution of this minimization problem is given by

$$\widehat{X}_{k+p}^* = E\{x_{k+p} | x_k = X_k, x_{k-1} = X_{k-1}, \cdots\}. \quad (2)$$

Unfortunately, this value can not be computed since the conditional probability density $P\{x_{k+p} | x_k = X_k, x_{k-1} = X_{k-1}, \cdots\}$ is unknown. The criterion in Eq. (1) is replaced by the empirical criterion given by

$$C_{emp} = \frac{1}{\ell} \sum_{i=1}^{\ell} (X_i - \widehat{X}_i)^2. \quad (3)$$

The relationship between $\widehat{X}_{k+p}$ and the sequence $X_k, X_{k-1}, \cdots$ is supposed to be nonlinear of unknown nature with the following autoregressive form

$$\widehat{X}_{k+p} = \widehat{f}(X_k, X_{k-1}, \cdots, X_{k-r+1}). \quad (4)$$

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 341-346

Where r is the model's order. This fact suggests the use of techniques like neural networks [8] or RBF [2]. In this context, two problems appear

- the model's order which is related to the curse of dimensionality,
- the estimator complexity control which is related to the underfitting and overfitting problem.

In time series prediction, a challenge is to learn fast dynamics (equivalently to high frequencies in the linear case) and, simultaneously cancel noise. This challenge is directly related to the underfitting/overfitting problem. Indeed, learning noise causes overfitting. Whereas, forgetting fast dynamics leads potentially to underfitting. Our approach to resolve this problem is based on a multiscale decomposition of the time series. The decomposition is achieved using a low-pass and a band-pass filters. The iterative application of these filters results in a trend series and a hierarchy of detail series which contain information about the system's dynamics at different scales.
The paper is organized as follows: in the §2., the principles of multiscale filtering are briefly recalled. In §3., the use of the obtained series for the prediction is discussed. The application of the method is then illustrated in §4.

2. The multiscale filtering

The multiscale decomposition uses a low-pass and a band-pass filters [3]. Applying this pair of filters to the original time series leads to a first series which contains the trend (or slower dynamics) and a second one which is the difference between the original series and the trend. The reconstruction of the original series is possible by summing up the trend and the detail series.
The nature of the application imposes the use of causal filters. In fact, at the present moment, the future value of the series is unknown. Let $(h_n),\ n \in Z$ and $(g_n),\ n \in Z$ be the impulse response of the low-pass and high-pass filters respectively. The causality and the reconstruction constraints imply

$$\begin{cases} h_n = g_n = 0, & n \geq 1, \\ h_0 + g_0 = 1, & \\ h_n = -g_n, & n \leq -1. \end{cases} \tag{5}$$

The simplest filters satisfying Eq. (5) are the Haar filters [6] given by

$$\begin{cases} h_0 = h_1 = \frac{1}{2}, \\ g_0 = -g_1 = \frac{1}{2}. \end{cases} \tag{6}$$

This decomposition scheme can be performed several iterations. At each one, it consists on decomposing the trend series of the previous iteration. Let $x_m = c_{0,m},\ m = 1, \cdots, \ell$ be the original series,and let $c_{N,m},\ d_{j,m},\ j = 1, \cdots, N,\ m = 1, \cdots, \ell$ be respectively the trend and the different detail levels obtained after N iterations. We can write then

$$x_m = c_{N,m} = c_{N,m} + \sum_{j=1}^{N} d_{j,m}, \quad m = 1, \cdots, \ell. \tag{7}$$

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 341-346

In this case,

$$\left\{\begin{array}{l} c_{N,m} = (\underbrace{h * h * \cdots * h}_{N\ times} * x)_m, \\ d_{j,m} = (\underbrace{h * h * \cdots * h}_{j-1\ times} * g * x)_m \quad j = 1, \cdots, N. \end{array}\right. \tag{8}$$

Note that at each iteration, we may use a different pair of filters. These however must satisfy the constraints given in (5). A simple application of this remark is padding with zeros the impulse response of the filters. Thus, at iteration j, the low-pass and band-pass filters, noted $h_{j,.}$ and $g_{j,.}$, are given by

$$\left\{\begin{array}{l} h_{j,0} = h_{j,2^j-1} = g_{j,0} = -g_{j,2^j-1} = \frac{1}{2}, \\ h_{j,m} = g_{j,m} = 0, \qquad m \neq 0, \cdots, 2^j - 1. \end{array}\right. \tag{9}$$

The trend and detail series are used to predict the original series. This will be sketched in the next section.

3. Time series prediction

The use of the wavelet coefficients is motivated by the easy analysis of the obtained series. In fact, the trend may be used to analyze the system's slowest dynamics. The detail series $d_{j,.}$ contain the difference between the time series $c_{j-1,.}$ and $c_{j,.}$, they inform about the importance of the intermediate dynamics. The highest detail series includes the fastest dynamics and noise. As the trend and the lowest detail series are practically noise free, the training and the complexity control of their corresponding estimators are simpler than the ones of the original series. However, if the information is totally embedded in noise in the highest detail series, one can simply put at zero the corresponding predictions to avoid the overfitting.

For each series, an estimator is constructed. The first idea is to treat separately each time series. In this case we have [1]

$$\left\{\begin{array}{l} \widehat{c}_{N,k+p} = \widehat{f}_0(c_{N,k}, c_{N,k-1}, \cdots, c_{N,k-r_0}), \\ \widehat{d}_{j,k+p} = \widehat{f}_j(d_{j,k}, d_{j,k-1}, \cdots, d_{j,k-r_j}),\ j = 1, \cdots, N. \end{array}\right. \tag{10}$$

The choice of the estimators $\widehat{f}_0, \widehat{f}_1, \cdots, \widehat{f}_N$ is related to the nature of the time series. In this paper, only multilayer perceptrons are used. Each estimator has its proper order r_j, $j = 0, \cdots, N$. This method has the major drawback of not taking into account the existing correlation between the different series. A more complex method consists in including, for each series, an information about the other series (considered as exogenous variables). This leads to the following estimators

$$\left\{\begin{array}{l} \widehat{c}_{N,k+p} = \widehat{f}_0(c_{N,k}, \cdots, c_{N,k-r}, \cdots, d_{j,k}, \cdots, d_{j,k-r}, \cdots), \\ \widehat{d}_{j,k+p} = \widehat{f}_j(d_{j,k}, \cdots, d_{j,k-r}, \cdots, c_{N,k}, \cdots, c_{N,k-r}),\ j = 1, \cdots, N. \end{array}\right. \tag{11}$$

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 341-346

The drawback of this method is that it increases the problem dimensionality. For each estimator, all the variables are took with the same order in order to simplify the problem.
The sum of the predictions is put equal to the predictions sum; i.e.

$$\widehat{x}_{k+p} = \widehat{c}_{N,k+p} + \widehat{d}_{N,k+p} + \cdots + \widehat{d}_{1,k+p}. \tag{12}$$

In this context, we have the following property:

Property 1 *If the estimator $\widehat{f}$, of order r, is obtained by minimizing the risk C_{emp} on the raw data, and if the estimators $\widehat{f}_0, \widehat{f}_1, \cdots, \widehat{f}_N$, with the same order r and the same number of neurons, are obtained simultaneously by minimizing the following risk*

$$C^w_{emp} = \frac{1}{\ell}\sum_{k=1}^{\ell}((c_{N,m+p} - \widehat{c}_{N,m+p}) + \cdots + (d_{j,m+p} - \widehat{d}_{j,m+p}) + \cdots)^2, \tag{13}$$

then, we have

$$\min C^w_{emp} \leq \min C_{emp}. \tag{14}$$

for all the estimators written as a linear or nonlinear combination of linear projections of the input variables (multilayer perceptron and RBF are within this class).

Proof: *In the above conditions, we write*

$$\widehat{f} = \sum_{i=1}^{s} w_i \varphi(\sum_{l=0}^{r-1} a_{i,l} x_{m-l} + a_{i,r}), \tag{15}$$

and

$$\left\{ \begin{array}{l} \widehat{f}_0 = \sum_{i=1}^{s} w_i^0 \varphi(\sum_{l=0}^{r-1} a^0_{0,l} c_{N,m-l} + \sum_{n=1}^{N}\sum_{l=0}^{r-1} a^0_{n,l} d_{n,m-l} + a^0_{i,r}) \\ \widehat{f}_j = \sum_{i=1}^{s} w_i^j \varphi(\sum_{l=0}^{r-1} a^j_{0,l} c_{N,m-l} + \sum_{n=1}^{N}\sum_{l=0}^{r-1} a^j_{n,l} d_{n,m-l} + a^j_{i,r}), \\ \qquad\qquad\qquad\qquad j = 1, \cdots, N. \end{array} \right. \tag{16}$$

So that, in the case where

$$a^0_{i,l} = a^j_{i,l}, \quad j = 1, \cdots, N, \tag{17}$$

all the estimators $\widehat{f}^0, \cdots, \widehat{f}_N$ are proportional to $\widehat{f}$. This implies the equivalence between C_{emp} and C^w_{emp} under the constraint (17). The C_{emp} definition domain is a subspace of the one of C^w_{emp}. Finally, we conclude that

$$\min C^w_{emp} \leq \min C_{emp}. \tag{18}$$

It is useful to note that this property is valid for linear AR models.

□

This property affirms that, under some conditions, the estimators using the wavelet coefficients fits more the data than the classical ones. However, the prediction error reduction is not guaranteed. In order to find estimators with good generalization properties, the cross-validation method is used [7]. The order r may be fixed using some knowledge about the series (e.g. the embedding space dimension in case of chaotic

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 341-346

series) or using a statistical criterion (cross-validation [7]). The simulation shows that the method using the wavelet coefficients is robust to the order misspecification; the generalization performance hardly varies with r [6].
The use of the above described approach is illustrated in the next section.

4. Application

The method has been applied to the well known Mackey-Glass chaotic series given by [4]

$$x_{k+1} = x_k + \frac{x_{k-\Delta}}{1 + [x_{k-\Delta}]^{10}}. \tag{19}$$

The objective is to compare our results with those obtained by other authors on the raw data. The parameters of the series are the following: $\Delta = 17$, the sampling rate is $\tau = 6$ (only the sample $x_0, x_\tau, x_{2\tau}, \cdots$ are used), the training and the test sets are $\ell_{train} = 500$ and $\ell_{test} = 1000$ length respectively. Two prediction times were tested: $p = 6$ and $p = 84$. The performance of the estimators is measured by the normalized error on the test set, i.e.

$$e = \frac{\sum_{k=1}^{\ell_{test}} (x_k - \widehat{x}_k)^2}{\sum_{k=1}^{\ell_{test}} (x_k - \overline{x})^2}, \quad \overline{x} = \sum_{k=1}^{\ell_{test}} x_k, \tag{20}$$

The decomposition of the series were achieved using the "padded with zeros" Haar filters over $N = 4$ levels. The model's order was fixed at $r = 4$ when $p = 6$ and at $r = 6$ for $p = 84$ (these values correspond to the embedding space dimension of the Mackey-Glass series [4]). Table 1 shows the results obtained with our method (last column), and those obtained with classical methods (neural networks, RBF, local linear polynomials, $\cdots$); see [4, 5] and the references therein for more details on these methods. Our method is shown to be the more efficient since it increases the prediction accuracy on the test set in the two cases.

	$p = 6$	$p = 84$
neural networks	0.010	0.050
local linear polynomials	0.033	0.045
standard RBF	0.011	0.158
weighted linear map	0.013	0.030
support vector machines	0.004	-
our method	0.002	0.023

Table 1: The results obtained with different methods for $p = 6$ and $p = 84$.

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 341-346

5. Conclusion

In this paper, a method for predicting nonlinear time series was presented, it is based on the multiscale filtering. The obtained series contain information on the system's dynamics at different scales. This property simplifies the learning of the series with slow dynamics. It may also be used to separate noise from relevant information. For each new time series, an estimator is constructed, it may include some information about the other series. The results obtained through the Mackey-Glass chaotic time series substantiate our approach.

References

[1] Alex Aussem and Fionn Murtagh. Combining neural network forecasts on wavelet transformed time series. *Connection Science*, 9(1):113–122, 1997.

[2] Martin Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35:335–356, 1989.

[3] Ingrid Daubechies. *Ten Lectures on Wavelets.* CBMS-NSF Regional Conference Series on Applied Mathematics, No **61**, SIAM, 1992.

[4] B. Lillekjendlie, D. Kugiumtzis, and N. Christophersen. Chaotic time series. part ii: System identification and prediction. *Modeling, identification and control*, 15(4):225–243, 1994.

[5] Sayan Mukerjee, Edgar Osuna, and Frederico Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *IEEE NNSP'97, Ameila Island, FL, USA,24-26 Sept*, 1997.

[6] Skander Soltani. *Application de la transforme en ondelettes pour la Reconnaissance des Formes.* PhD thesis, Universit de Technologie de Compigne, France, 1998.

[7] P. Vieu. Order choice in nonlinear autoregressive models. *Statistics*, 26:307–328, 1995.

[8] Andreas S. Weigand, Bernardo A. Huberman, and David E. Rumelhart. Predicting the future: A connectionist approach. *International Journal of Neural Systems*, 1(3):193–209, 1990.