ELSEVIER

Neurocomputing 48 (2002) 299–311

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Forecasting electricity consumption using nonlinear projection and self-organizing maps

A. Lendasse[a,*], J. Lee[b], V. Wertz[a], M. Verleysen[b]

[a] *Université catholique de Louvain, CESAME, 4 av. G. Lemaître, B-1348 Louvain-la-Neuve, Belgium*

[b] *Université catholique de Louvain, Electricity Department, 3 pl. du Levant, B-1348 Louvain-la-Neuve, Belgium*

Received 27 November 2000; accepted 6 June 2001

Abstract

A general-purpose useful parameter in time series forecasting is the regressor size, corresponding to the minimum number of variables necessary to forecast the future values of the time series. If the models are nonlinear, the choice of this regressor becomes very difficult. We present a quasi-automatic method using a nonlinear projection named curvilinear component analysis to build this regressor. The size of this regressor will be determined by the estimation of the intrinsic dimension of an over-sized regressor. This method will be applied to electric consumption of Poland using systematic cross-validation. The nonlinear model used for the prediction is a Kohonen map (self-organizing map). © 2002 Published by Elsevier Science B.V.

Keywords: Time series prediction; Nonlinear projection; Curvilinear component analysis; Self-organizing map; Electricity consumption

1. Introduction

Time series forecasting is a great challenge in many fields. In finance, one forecasts stock exchange courses or indices of stock markets; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. The common point to their problems is the

* Corresponding author.

E-mail addresses: lendasse@auto.ucl.ac.be (A. Lendasse), lee@dice.ucl.ac.be (J. Lee), wertz@auto.ucl.ac.be (V. Wertz), verleysen@dice.ucl.ac.be (M. Verleysen).

0925-2312/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S0925-2312(01)00646-4

following: how can one analyse and use the past to forecast the future? Many techniques exist, the linear methods such as ARX, ARMA, etc. [1,9], and the nonlinear methods such as artificial neural networks [13]. In general, these methods try to build a model of the process that is to be predicted. This model connects the last values of the series to these future values. The common difficulty to all methods is the determination of sufficient and necessary information for a good prediction. If the information is insufficient, the forecasting will be poor. On the contrary, if information is useless or redundant, modelling will be difficult or even skewed.

In this paper, we will describe an original method for the determination of the information that is useful for a good forecasting. The size of the regressor (vector including the past values of the series) will be determined by the estimation of the intrinsic dimension of an over-sized regressor [11,6]. The optimal regressor will be obtained by the nonlinear projection of this initial over-sized regressor [8]. For this nonlinear projection, we will use a method named curvilinear component analysis [5].

We will also briefly present a model of nonlinear forecasting using the Kohonen maps [7]. Finally, we will illustrate the presented methods with a traditional example of time series, the forecasting of the electrical consumption in a country. The data that we will use correspond to the electrical load in Poland [3]. A systematic cross-validation methodology is presented that prevents the overfitting of the learning data.

2. General method for forecasting

In this section, we will briefly describe a general forecasting method (without exogeneous variables) [10]. We note the series y_t, with t varying between 1 and N. A standard model that collects the dynamics of the process is

$$\hat{y}_{t+1} = f(y_t, y_{t-1}, \ldots, y_{t-n}, \vartheta), \tag{1}$$

where ϑ is the set of parameters that makes it possible for the model f to approximate the series as well as possible. For example, in a multi-layer perceptron (MLP), ϑ is the set of synaptic weights [14]. The vector y_t to y_{t-n} is called regressor.

It is obvious that the choice of the regressor and thus of n is capital. If this choice is badly done, the model will be vague or possibly skewed. In the best case, the model will be correct but the determination of ϑ will be very difficult. Several methods exist to choose the regressor. For example, one can use the optimal regressor obtained from a linear model, but using a linear criterion in a nonlinear context is far from being optimal. One can also use pruning methods [2], but these usually require extensive computations and are in most cases limited to a particular model.

Generally, a model is parameterized by a given number of parameters say M. In our Eq. (1), M would be the size of ϑ. Usually, when M is small, the model is not complex enough to capture the dynamics of the true system and in case of time series prediction, the prediction will not be accurate. On the contrary, if M is taken too large, the parameters try to capture also the noise contained in the learning data. This is the overfitting phenomenon. The prediction of the learning data will hence be very accurate (even the noise is "correctly" predicted) but on validation data (the generalization step) the model will be inaccurate.

The goal is thus to determine the optimal number M of parameters. With this aim in view, the data y_t will be randomly divided into a learning set and a validation set. Two different mean squared error are calculated, the learning mean squared error ($LMSE$):

$$LMSE = \frac{\sum_{t=1}^{N_1} (\hat{y}_t - y_t)^2}{N_1} \tag{2}$$

where N_1 is the dimension of the learning set, and the validation mean squared error ($VMSE$)

$$VMSE = \frac{\sum_{t=1}^{N_2} (\hat{y}_t - y_t)^2}{N_2} \tag{3}$$

with N_2 being the dimension of the validation set.

The optimal number of parameters M^* is a compromise between an accurate model (minimizing the $VMSE$) and a parsimonious model (with a little number of parameters). Unfortunately, this optimum M^* depends on the choice of the learning and the validation sets. A solution to this problem is the cross-validation method. In this method, the splitting between learning and validation sets is repeated several times. For each division, the $LMSE$ and the $VMSE$ are computed with respect to M, finally the $LMSE$ and the $VMSE$ are averaged to get an optimal M^* independent of the choice of the learning and validation sets.

3. Determination of the regressor using CCA

The method that we will present to determine the best regressor is different from the classical ones. Indeed, we will not pick up variables among the past values of the series to build the best regressor, but we will construct it from a nonlinear projection.

Let us build a regressor of large dimension, which will contain too many information

$$Y_t = [y_t, y_{t-1}, \ldots, y_{t-n}]. \tag{4}$$

We thus created a set of regressors in an n-dimensional space, in which information is redundant. The fact that the real or intrinsic dimension d of the set of regressors

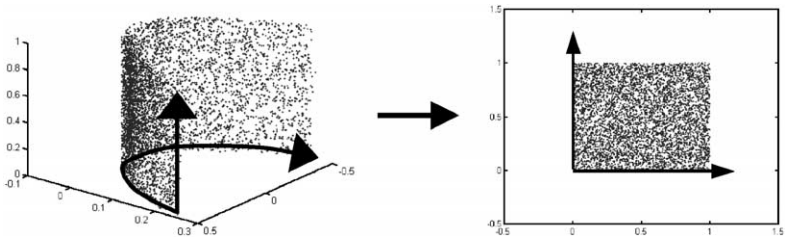

Fig. 1. Projection carried out by CCA from R^3 to R^2.

is lower than n expresses this redundancy of information. The Y_t data form a d-dimensional surface in space R^n. We are thus going to construct a new regressor of dimension d that stores all the information contained in the initial regressor. For this purpose, a projection can be used. Various techniques of projection exist to project from an n-dimensional space to a d-dimensional one. For example, principal component analysis (PCA) may be used, but PCA is probably not judicious here because it is a linear projection. An interesting alternative to PCA is the curvilinear component analysis (CCA) that is one of its nonlinear extensions. Fig. 1 shows an example of projection on the horseshoe distribution carried out by CCA.

CCA is a nonlinear mapping from an n-dimensional space to a d-dimensional space. This mapping is obtained by minimizing

$$E = \frac{1}{2} \sum_i \sum_{j \neq i} (X_{ij} - Y_{ij})^2 F(Y_{ij}, \lambda_y) \tag{5}$$

where X_{ij} is the Euclidean distance between two inputs (i and j), Y_{ij} the Euclidean distance between two outputs (i and j) and

$$F(Y_{ij}, \lambda_y) = \begin{cases} 1 & \text{if } Y_{ij} \leqslant \lambda_y, \\ 0 & \text{if } Y_{ij} > \lambda_y. \end{cases} \tag{6}$$

Details on the minimization of E using a gradient descent can be found in [4,5].

We can thus summarize the global method as follows: the regressor

$$Y_t = [y_t, y_{t-1}, \ldots, y_{t-n}] \tag{7}$$

is projected using CCA to

$$Z_t = [z_1, z_2, \ldots, z_d]. \tag{8}$$

The forecasting model is then built,

$$\hat{y}_{t+1} = f(z_1, z_2, \ldots, z_d, \vartheta). \tag{9}$$

The methodology is illustrated in Fig. 2.

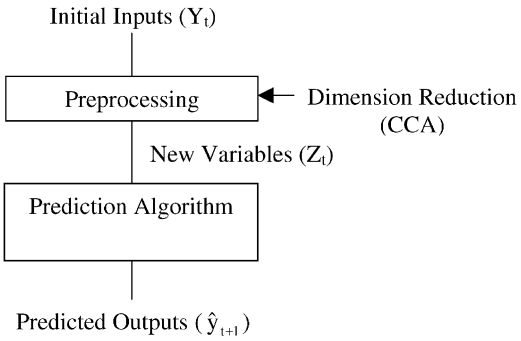

Fig. 2. The two steps of the methodology.

4. Intrinsic dimension

First, it is important to evaluate the projection dimension, i.e. the dimension of the space of the new variables. If the estimation of this dimension is too small, information will be lost in the projection. If it is too large, the usefulness of the method is lost. To evaluate this dimension, the concept of intrinsic dimension is used. The intrinsic dimension is the effective number of degrees of freedom of a set, i.e. the number of independent variables. The definition of the intrinsic dimension (named d) of a set of points Y ($Y \in R^n$) is

$$d = \lim_{r \to 0} \frac{\ln(C_m(r))}{\ln(r)} \tag{10}$$

with

$$C_m(r) = \lim_{N \to \infty} \frac{2}{N(N-1)} \sum_{1 < i < j < N} I(\|Y_j - Y_i\| \leqslant r) \tag{11}$$

and

$$I(\lambda) = 1 \quad \text{iff condition } \lambda \text{ holds, } 0 \text{ otherwise.}$$

More details on the intrinsic dimension can be found in [11]. This concept is presented here with the well-known horseshoe distribution (Fig. 2): for this data set, the intrinsic dimension is equal to two as two degrees of freedom are sufficient to uniquely determine any data in the set, although the data live in R^3. The computation of the intrinsic dimension is explained in [6], but its determination remains very difficult to apply, not to say approximate, for high-dimensional data sets. Therefore, the intrinsic dimension will be only considered here as a rough approximation of the dimension that should be used for the projection.

5. Toy example

Fig. 3 shows an artificial series used to test the presented method. This series is built as the sum of two sine waves: $\sin(wt)$ and $\sin(2wt)$ with noise. The continuous time series is then discretized and considered as a dynamical system.

The initial regressor that we will choose is of dimension 3. Fig. 4 shows the regressor at each time step of the series.

It is clearly visible in the figure that the intrinsic dimension of the series is 1. We will thus project the regressor to R using CCA. In Fig. 5, we represent y_{t+1} with respect to the new regressor Z.

This series can be easily modelled using a RBF network (radial basis function network) with five Gaussian kernels [12]. The VMSE obtained is 2.4. In comparison, the VMSE obtained with the initial regressor is 11.50 with a linear model and 2.1 using a RBF with 25 Gaussian kernels.

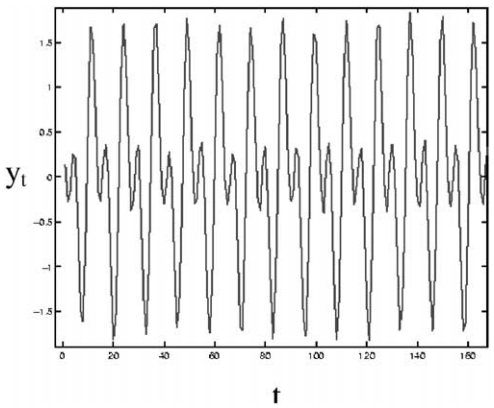

Fig. 3. An artificial time series.

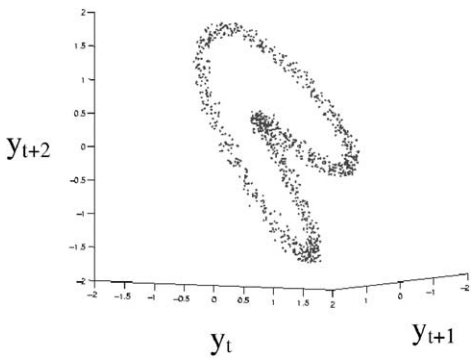

Fig. 4. The initial regressor at every step of the series.

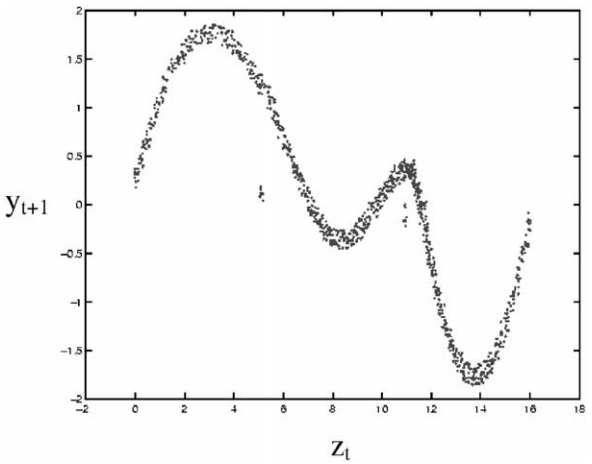

Fig. 5. y_{t+1} with respect to the new regressor Z.

6. Forecasting time series using Kohonen maps

The model of forecasting that we will use is based on Kohonen self-organizing maps (SOM) [7]. Assume that an optimal regressor Z_t has been built using CCA. We will concatenate this regressor with the next value y_{t+1} in a new vector named x_t

$$x_t = [z_1, z_2, \ldots, z_d, y_{t+1}]. \tag{12}$$

Then, we will quantify the x_t distribution by a SOM whose centroids will be noted as C_i. These centroids are thus made of two parts, the first part C_{i1} corresponding to the regressors and the second part C_{i2} corresponding to the predictions. These centroids form our model. Indeed, at each time t, the forecasting will be calculated in the following way. First, the regressor Z_t is built using CCA. Then, the centroid C_i whose part C_{i1} is closest from Z_t is selected. Finally, the forecasting is part C_{i2} of this centroid:

$$\hat{y}_{t+1} = C_{i2} \overset{i}{\leftarrow} \min_i \|z_t - C_{i1}\|. \tag{13}$$

7. Application to electricity consumption

The series studied in this paper represents the daily electrical consumption in Poland [3] during 2700 days. The standardized series is shown in Fig. 6. The quasi-sinusoidal seasonal variation is clearly visible. If we look at a few weeks scale (Fig. 7), it may be seen that the electrical consumption has the shape of saw teeth, where the maximum occurs during the weekdays.

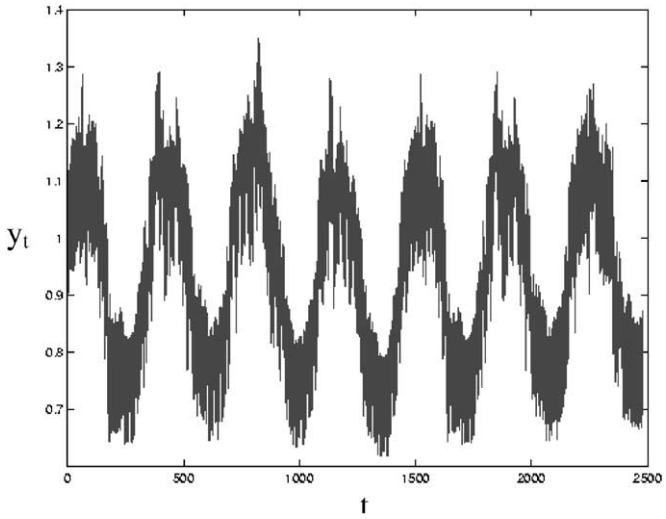

Fig. 6. Electricity consumption in Poland.

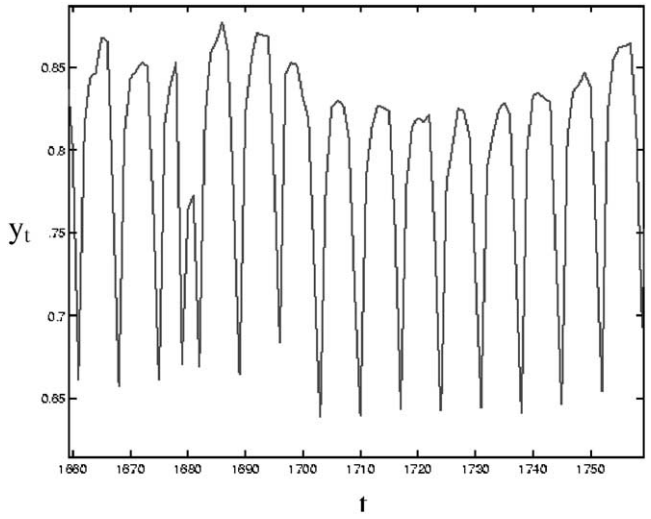

Fig. 7. Electricity consumption in Poland on a few weeks scales.

7.1. Linear model

For the sake of comparison, and also to determine an initial (too large) size of the regressor that we will use in a nonlinear model, we first adjust a linear model of the form

$$\hat{y}_{t+1} = a_0 + a_1 y_t + a_2 y_{t-1} + \cdots + a_{n+1} y_{t-n}. \tag{14}$$

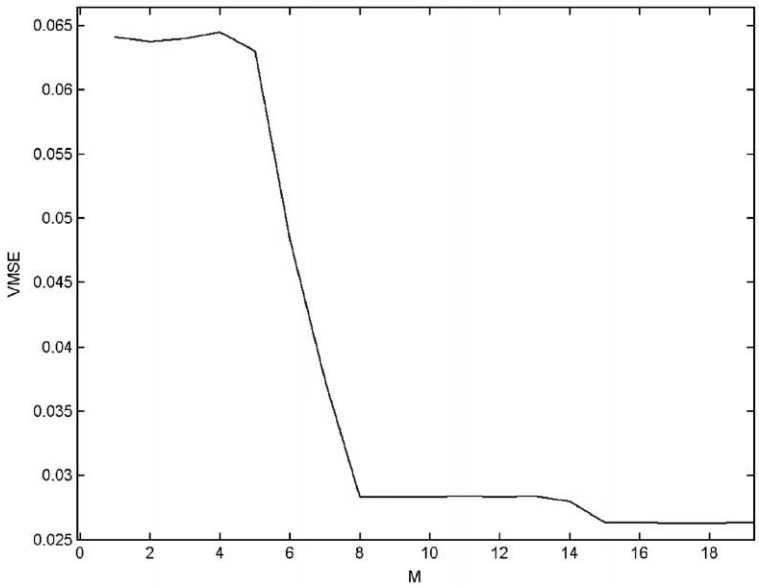

Fig. 8. Averaged VMSE w.r.t. the size of the regressor.

In such a linear model, the number of a_i parameters is the size of the regressor. To determine this size, we will use the cross-validation methodology presented in Section 3. The averaged $VMSE$ with respect to the size of the regressor is shown in Fig. 8.

The cross-validation is realized on a large number of splittings between learning and validation sets (1000 splittings); for each division two third of the data are used for training and one third is kept for validation. A good choice for the number M of parameters $(a_0, a_1, \ldots, a_{n+1})$ is 15, i.e. 14 days in the regressor. The regressor covers two weeks most probably because some saturdays are public holidays in Poland, and some others are not. The $VMSE$ is 0.0263. This value is very low; we conclude that the real dynamics of this times series is nearly linear.

7.2. Initial nonlinear model

The second step is the determination of an optimal nonlinear model. First, the regressor obtained with the linear model is maintained. The number of centroids in the Kohonen map is again determined by cross-validation on the $VMSE$. One hundred splittings between learning and validation sets are used. The result is presented in Fig. 9.

The optimum number of centroids is about 1500. For this optimum, the $VMSE$ is 0.0179 i.e. 30% less than that with the linear model. However, the use of a cross-validation procedure increases the computational load. Moreover, the large

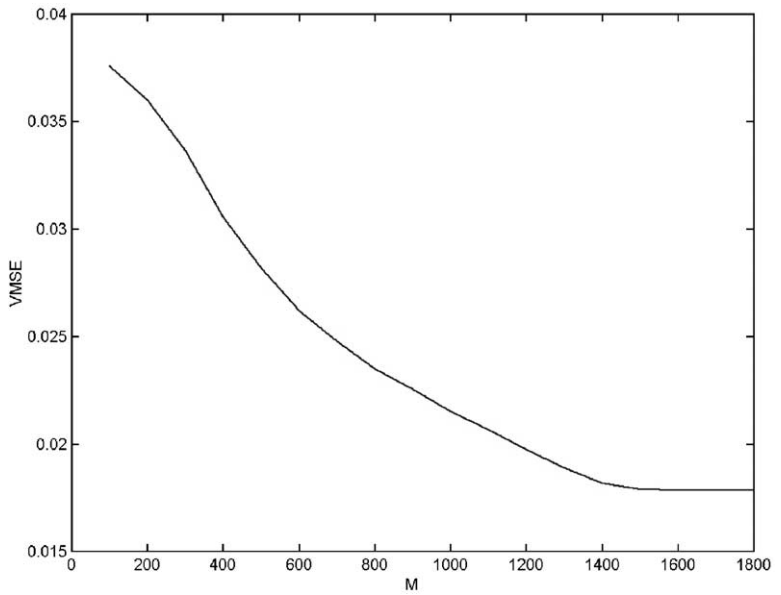

Fig. 9. Averaged VMSE w.r.t. the number of centroids (initial nonlinear model).

dimension of the regressor (and thus the information redundancy) leads to difficulties in the learning phase of the Kohonen map. The number of parameters seems very large but the use of the cross-validation technique guarantees that we do not overfit the data. In general, in local models techniques, it is know low noisy data lead to a great number of parameters.

7.3. Intrinsic dimension and nonlinear projection

The intrinsic dimension of this time series is computed using the technique defined in [6]. The dimension found is equal to 6. The computation of an intrinsic dimension is never accurate, but the results will show that accuracy is not crucial at this stage. The curvilinear component analysis is performed from a 14-D space to a six-dimensional space and in this new input space (Z_t) a Kohonen map is built. The CCA projection is repeated, and for each projection different nonlinear models (Kohonen maps) are built. The results obtained by this new cross-validation procedure are shown in Fig. 10.

The optimum number of centroids is about 1500. For this optimum, the *VMSE* is 0.0178. This error is nearly the same than the error obtained without projection, but in this case, the dimension of the input space is smaller. The information contained in the regressor is kept, noise is reduced and the convergence of the Kohonen map is easier.

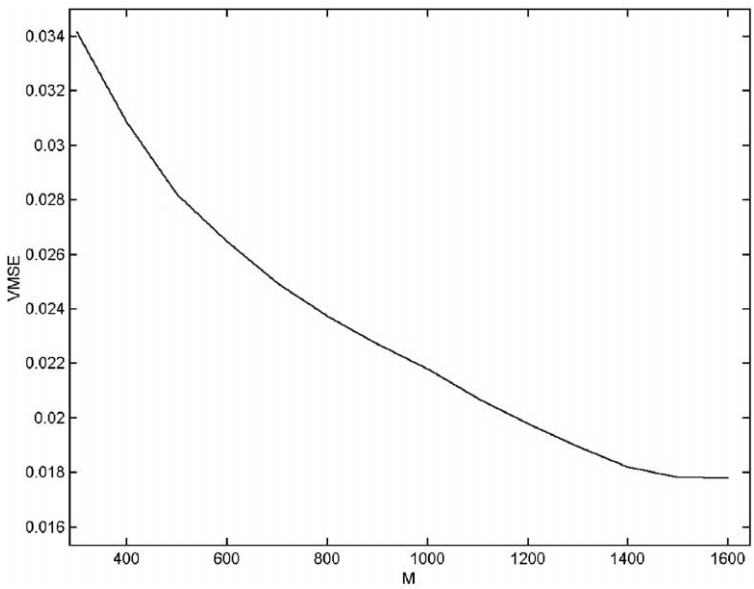

Fig. 10. Averaged VMSE w.r.t. the number of centroids (nonlinear model after projection).

7.4. Discussion

Why are the results not better? We chose on purpose an approximation technique (Kohonen map) that is probably not the best model that could be used for this series. However, with this technique, the problems of local minima in the learning phase are reduced. With a MLP or a RBF network, these problems would interfere with the determination of the optimal parameters.

Finally, the nonlinear projection obtained with CCA is not perfect and could add some undesirable noise in the new regressor; improvements in the CCA could improve the results too.

8. Conclusion

In this paper, we presented a methodology based on a nonlinear projection to build a good regressor in the problem of time series prediction. The intrinsic dimension is used to estimate the information contained in the series. This estimation is not very accurate but sufficient to get good results. The methodology shows that knowledge resulting from linear models on the same series is useful. Moreover, the use of systematic cross-validations avoids local minima and varying errors. This makes the method robust.

We used Kohonen maps as nonlinear approximators, because they are less subject to problems of local minima than other nonlinear models as MLP and RBF.

Kohonen maps may be good approximators, but other models could be used to improve the results. However, the goal of this study was not to obtain the best possible results, but to assess a new methodology including nonlinear projection.

Acknowledgements

Michel Verleysen is a Research Associate of the Belgian National Fund for Scientific Research (FNRS). The work of John Lee was realized with the support of the "Ministère de la Région wallonne", under the "Programme de Formation et d'Impulsion à la Recherche Scientifique et Technologique". Part of the results presented in this paper has been funded by the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors.

References

[1] G.E.P. Box, G. Jenkins, Time Series Analysis: Forecasting and Control, Cambridge University Press, Cambridge, 1976.
[2] M. Cottrell, B.Y. Girard, M. Mangeas, C. Muller, Neural modeling for time series: a statistical stepwise method for weight elimination, IEEE Trans. Neural Networks 6 (6) (1995) 1355–1364.
[3] M. Cottrell, B. Girard, P. Rousset, Forecasting of curves using a Kohonen classification, J. Forecasting 17 (1998) 429–439.
[4] P. Demartines, J. Hérault, Vector quantization and projection, in: A. Prieto, J. Mira, J. Cabestany (Eds.), International Workshop on Artificial Neural Networks, Lecture Notes in Computer Sciences, Vol. 686, Springer, Berlin, 1993, pp. 328–333.
[5] P. Demartines, J. Herault, Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets, IEEE Trans. Neural Networks 8 (1) (1997) 148–154.
[6] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D 56 (1983) 189–208.
[7] T. Kohonen, Self-organising Maps, Springer Series in Information Sciences, Springer, Berlin, 1995.
[8] A. Lendasse, E. de Bodt, V. Wertz, M. Verleysen, Nonlinear financial time series forecasting—application to the Bel20 stock market index, European J. Econom. Social Systems 14 (1) (2000) 81–91.
[9] L. Ljung, System Identification—Theory for User, Prentice-Hall, Englewood Cliffs, NJ, 1987.
[10] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.Y. Glorennec, H. Hjalmarsson, A. Juditsky, Non linear black box modeling in system identification: an unified overview, Automatica 33 (1997) 1691–1724.
[11] F. Takens, On the Numerical Determination of the Dimension of an Attractor, Lecture Notes in Mathematics, Vol. 1125, Springer, Berlin, 1985, pp. 99–106.
[12] M.Verleysen, K. Hlavackova, An optimised RBF network for approximation of functions, Proceedings of European Symposium on Artificial Neural Networks, Belgium, 1994.
[13] A.S. Weigend, N.A. Gershenfeld, Times Series Prediction: Forecasting the future and Understanding the Past, Addison-Wesley, Reading, MA, 1994.
[14] P. Werbos, Beyond regression: new tools for prediction and analysis in the behavioural sciences, Ph.D. Thesis, Harvard University, 1974.

Amaury Lendasse was born in 1972 in Tournai, Belgium. He received the M.S. degree in mechanical engineering from the Université catholique de Louvain (Belgium) in 1996 and the M.S. in control from the same University in 1997. Currently he is a working towards the Ph.D. degree in the Centre for Systems Engineering and Applied Mechanics (CESAME) in the Université catholique de Louvain and he is a teaching assistant in the Applied Mathematics department. He is author or co-author of about 11 scientific papers in international journals or communications to conferences with reviewing committee. His research concerns time series prediction, Kohonen maps, nonlinear projections and nonlinear approximators.

John A. Lee was born in 1976 in Brussels, Belgium. He received the M.S. degree in computer engineering from the Université catholique de Louvain (Belgium) in 1999. Currently, he is working towards the Ph.D. degree in the Microelectronics Laboratory of the Université catholique de Louvain. He is author or co-author of about 5 scientific papers in international journals or communications to conferences with reviewing committee. His research concerns signal processing and neural networks, with a specialisation in the area of nonlinear projection algorithms.

Vincent Wertz was born in Liège in 1955. He obtained his engineering degree in applied mathematics in 1978 and a Ph.D. in engineering in 1982, both from Université catholique de Louvain (Belgium). Then, he held positions first as a researcher then as a professor in the Automatic Control Laboratory of the same university. He is author or co-author of more than 100 scientific papers in international journals and books or communications to conferences with reviewing committee. His main research interests are in identification and control, both from a theoretical and a practical viewpoint. Recently, he has also been interested in teaching and learning issues for the undergraduate curriculum.

Michel Verleysen was born in 1965 in Belgium. He received the M.S. and Ph.D. degrees in electrical engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. In 1992, he was an Invited Professor at the Swiss E.P.F.L. (Ecole Polytechnique Fédérale de Lausanne). He is now a research associate of the Belgian F.N.R.S. (Fonds National de la Recherche Scientifique) and Lecturer at the Université catholique de Louvain. He is editor-in-chief of the Neural Processing Letters journal and chairman of the annual ESANN conference (European Symposium on Artificial Neural Networks). He is author or co-author of about 80 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the co-author of the scientific popularization book on artificial neural networks in the series "Que Sais-Je?", in French. His research interests artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, and electronic implementations of neural and biomedical systems.