Lazy learning for iterated time-series prediction

Gianluca Bontempi, Mauro Birattari and Hugues Bersini

Iridia - Université Libre de Bruxelles
1050 Bruxelles - Belgium

Abstract- Lazy learning is a memory-based technique which query-by-query selects a local model representation by assessing and comparing different alternatives in cross-validation. The paper investigates the effectiveness of a lazy learning method for time series prediction. The paper contribution is a formulation of the PRESS statistic for iterated prediction in time series forecasting. Lazy learning is generally used for function estimation purposes which do not take temporal behavior into account. Our technique extends the idea of local estimation to the problem of long-horizon prediction. Experimental applications of the techniques to the a time series prediction are presented.

I. Introduction

Modeling from data has been the object of several disciplines from nonlinear regression to machine learning and system identification. In the literature dealing with this problem, two main opposing paradigms have emerged: local memory-based versus global methods.

Global modeling builds a single functional model of the dataset. This has traditionally been the approach taken in neural network modeling and other form of nonlinear statistical regression. The available dataset is used by a learning algorithm to produce a model of the mapping and then the dataset is discarded and only the model is kept.

Local memory-based algorithms defer processing of the dataset until they receive an explicit request for information (e.g. prediction or local modeling). A database of observed input-output data is always kept and the estimate for a new operating point is derived from an interpolation based on a neighborhood of the query point. Local techniques are an old idea in classification, regression and time series prediction. The idea of local approximators as alternative to global models originated in non-parametric statistics to be later rediscovered and developed in the machine learning field. Recent work on *lazy learning* [1] gave a new impetus to the adoption of local techniques for modeling and control problems [3].

Despite the differences in the learning procedure, global and local modeling share a common issue: how to select the model structure which will have the best generalization performance given a set of noisy data. This is the well-known *bias/variance* dilemma [5] which although in different formulations, reappears each time one estimates a model from data. When no a priori assumptions can be made about the unknown process underlying the data, a common approach is to assess the quality of an estimated model using data which are independent from those used for training. In many practical situations, however, the exiguity of data discourages the method of

keeping aside part of the data set for validation purposes. A valid alternative is the adoption of *cross-validation* techniques [11]. These procedures allow to use an high proportion of the available data to train the model, while also making use of all data points in evaluating the cross-validation error. Unfortunately, this approach has the disadvantage that the training process has to be repeated as many times as the number of partitions of the training set. In the case of global nonlinear models, the consequence is a relevant increase in the amount of processing time.

The local paradigm can represent a possible solution to this dilemma. Local memory based modeling adopts a query-based approach where the linear identification procedure focuses only on a neighborhood of the point where an estimation is required (*local weighted regression*). This allows the adoption of enhanced linear statistical procedures to validate the local approximator. One example is the PRESS statistic [7]: this statistics returns the leave-one-out cross-validation error of a linear model at the same computational cost of the linear regression. As a consequence, the performance of a local linear model can be easily assessed with no additional computational burden. It is worthy noting how this also means that local learning returns along with each predicted value an estimation of its standard error. This property is more relevant if compared with the intrinsic difficulty of extracting the same information from other nonlinear approximators (e.g. neural networks). In this paper we will show how the combination of the PRESS validation method with linear regression techniques can be effective for time-series prediction.

We present a formulation of the PRESS statistic for iterated prediction in time series estimation. In [4] we proposed a query-based method to select the order of the model and the number of neighbors for a single step prediction. Here we extend the lazy learning idea to iterated long-time horizon prediction. We will show how the PRESS formula can be easily extended to longer horizon than a simple one-step prediction preserving its property of computational efficiency. This technique makes of lazy learning an effective alternative to recurrent neural networks approaches [9] typically based on time consuming tuning procedures (e.g. back propagation through time).

II. Iterated PRESS for time series prediction

A time series is a sequence of measurements φ^t of an observable φ at equal time intervals. The Takens theorem [8] implies that for a wide class of deterministic systems, there exists a *diffeomorphism* (one-to-one differential mapping) between a finite window of the time series $\varphi^{t-1}, \varphi^{t-2}, \ldots, \varphi^{t-m}$ (*lag vector*) and the state of the dynamic system underlying the series. This means that in theory it exists a multi-input single-output mapping (*delay coordinate embedding*) $f : R^m -> R$

$$\varphi^{t+1} = f(\varphi^t, \varphi^{t-1}, \ldots, \varphi^{t-m+1}) \quad (1)$$

where m (*dimension*) is the number of past values taken into consideration. Let $\mathbf{\Phi}$ be a matrix of dimensionality $[N \times m]$, and $\mathbf{y}$ a vector of dimensionality $[N \times 1]$, where $\mathbf{\Phi}(i)$ is a generic lag vector and $\mathbf{y}(i)$ is the corresponding next value of the series. A model of the mapping (1) can be used for two objectives: *single step prediction* and *iterated prediction*. In the first case, the m previous values of the series are assumed to be available and the problem is equivalent to a problem of function estimation. This model returns a *1-step ahead* prediction.

In the case of iterated prediction, the predicted output is fed back as input for the next prediction. Hence, the inputs consist

of predicted values as opposed to actual observations of the original time series. A prediction iterated for h times returns a *h-step ahead* forecasting.

However, iterated prediction is not the only way to make *h-step ahead* forecasting. Weigend [2] classifies the models for h-step ahead prediction, according to two features: the horizon of the model prediction and the horizon of the training criterion. He enumerates three cases:

1. the model predicts one-step ahead (Eq. 1) and the parameters of the model are optimized to minimize the error on one-step ahead forecast (one-step ahead criterion)

2. the model predicts one step ahead but the parameters of the model are optimized to minimize the error on the iterated h-step ahead forecast (h-step ahead criterion)

3. the model makes a direct forecast at time $t+h$:

$$\varphi^{t+h} = f^h(\varphi^t, \varphi^{t-1}, \ldots, \varphi^{t-m+1}) \quad (2)$$

Methods of class 1 have a low performance in long horizons task. This is due to the fact that they are essentially models which are tuned on a one-step criterion which is not capable of taking temporal behavior into account. Methods like recurrent neural networks [6] are an example of class 2. Their recurrent architecture and the associated training algorithm (temporal backpropagation) are more able to handle the time-dependent nature of the data. Direct methods of class 3 often require high functional complexity in order to emulate the system. An example of combination of local techniques of type 1 and 3 is provided by Sauer [10] who uses both the approaches to improve the iterated prediction of his architecture.

In this section we present a lazy technique of type 2 which extends the neighbor selection method to iterated prediction tasks. The lazy model still returns an one-step ahead prediction but the choice of neighbors is no more done on the basis of a one step ahead cross-validation but on the basis of an iterated formulation of the PRESS statistic (h-step ahead criterion).

To explain the idea, let us limit to a simple case of a one step ahead lazy model with an h-step ahead criterion for model selection where $h = 2$. We are at time t and we want to predict the value $\hat{\varphi}_{t+1}$ with an iterated method. The standard lazy idea consists in searching for the optimal number of neighbors in order to approximate locally the dynamics f with a linear model. A single step criterion chooses the model which is expected to be the best generalizer at time $t+1$. Here, instead, we choose the model that, if iterated up to time $t+2$, would have the best generalization.

Let us now compute an iterated formulation of the PRESS statistic for $h = 2$. Note that an iterated prediction from time t to time $t+1$ is the composition of two mappings: $\varphi^{t+1} = f(\varphi^t, \varphi^{t-1}, \ldots, \varphi^{t-m+1})$ and $\varphi^{t+2} = f(\hat{\varphi}^{t+1}, \varphi^t, \ldots, \varphi^{t-m+2})$. Henceforth, we will refer to a configuration with two mappings $z = g(y)$ and $y = f(x)$, and a set of N points $\{(x_i, y_i, z_i)\}_{i=1}^N$.

We denote with $e^{cv}_{xy}(i)$ the cross validation residual of the linear model estimated on the data set $\{(x_i, y_i)\}$ and with $e^{cv}_{yz}(i)$ the cross validation residual of the linear model estimated on the data set $\{(y_i, z_i)\}$. In a single step task (e.g. the mapping $X \rightarrow Y$) the cross-validation residual for the $i-th$ sample ($e^{cv}_{xy}(i)$) is the difference between the real value of the mapping (y_i) and the prediction of model estimated with the i-th point aside (x_i). In a iterated task (e.g. the mapping $X \rightarrow Z$) the cross-validation residual for the $i-th$ sample (denoted by $e^{it}_{xz}(i)$) is the differ-

ence between the real value of the mapping (z_i) and the the prediction obtained by cumulating the predictions of the two models estimated with the i-th points (x_i and y_i) aside.

We illustrate the idea with Fig. 1. Here we have 4 samples and we want to estimate the iterated cross-validation residual in $x_i = x_3$. Let $R^{cv}_{yz}(y_i)$ be the value in y_i of the regression estimated on the training set $\{(y_i, z_i)\}$ with the sample y_i aside and $R^{cv}_{xy}(x_i)$ the analogous for the xy mapping. By definition $e^{cv}_{xy}(x_3)$ is the difference between y_3 and $R^{cv}_{xy}(x_3)$, and $e^{cv}_{yz}(y_3)$ is the difference between z_3 and $R^{cv}_{yz}(y_3)$. The figure shows how the sequence of predictions with the points x_3 and y_3 aside leads from A to C through B. Then, the iterated cross-validation residual $e^{it}_{xz}(i)$ is the difference between z_3 and the regression value $R^{yz}(y_3 - e^{cv}_{xz}(x_3))$ computed in the data point y_3 shifted of the error $e^{cv}_{xz}(x_3)$ occurred at the previous step.

The analytical expression of the iterated PRESS for the composition of the two mappings is computed in Appendix VI. Here we report the value for our simplified case:

$$e^{it}_{xz}(i) = \frac{e_{yz}(i)}{1 - h_{ii}} + \\ + \frac{e^{cv}_{xy}(i)\left[(1 - h_{ii})\beta - (\mathbf{Y}^T\mathbf{Y})^{-1} y_i e_{yz}(i)\right]}{1 - h_{ii}} \tag{3}$$

where $e_{yz}(i)$ is the residual, h_{ii} is the diagonal element of the HAT matrix and β denotes the least squares parameters of the model fitted on the dataset $\{(y_i, z_i)\}$.

In the case of time series prediction we have not a single-input single-output case but a more complex mapping. In this case the analytical expression of the iterated PRESS (see Appendix VI) for an horizon of h steps is

$$e^{it}_h(i) = \frac{e(i) + \delta_i\left[(1 - h_{ii})\beta - (\Phi^T\Phi)^{-1}\varphi_i e(i)\right]}{1 - h_{ii}} \tag{4}$$

where $e(i)$ is the residual, h_{ii} is the diagonal element of the HAT matrix and β denotes the least squares parameters of the regression of $\mathbf{y}$ on Φ. The symbol δ_i denotes the $i-th$ row of the matrix δ of dimension $[N, m]$ having as generic element $\delta(i, j)$ the iterated cross validation error of the i-th point at the j-th previous time step. For instance, in the case of an horizon $h = 2$ only the first column of δ differs from zero. The formula (4) was obtained for a non weighted regression. The extension to a weighted regression requires some slight modifications.

We remark that the vector of iterated cross-validation errors in Eq. 4 returns at time t the expected generalization of the iteration of a local model from time t up to time $t + h$. At each time t this statistic returns a richer information than the simple one-step PRESS statistics and/or the direct h-step PRESS. This allows a more reliable local model selection for iterated prediction.

III. Predicting a chaotic time series

The iterated PRESS approach has been applied to the prediction of a "real-world" data set, recorded from a far-infrared laser in chaotic state. This series was used in the *Santa Fe Time Series Prediction and Analysis Competition* [13] as data set A. The training set is a series of 1000 values, the test set is made of 10000 samples: the task is to predict the continuation for 100 steps, starting from different points. We adopt a lazy learning method where the selection of neighbors is made according to the iterated PRESS with horizon $h = 2$. The number of neighbors is limited to range from 4 to 8. Table 1 summarizes our results compared with those of

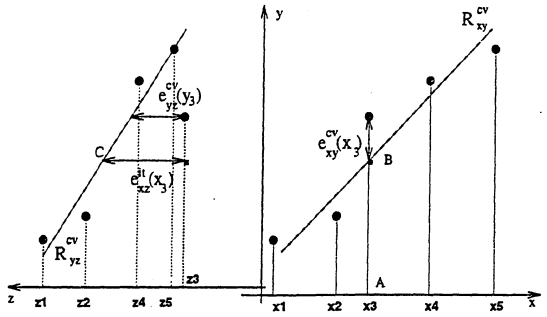

Figure 1: Iterated Press

Table 1:

NMSE	Non iterated PRESS	Iterated PRESS	Sauer	Wan
1001-1100	0.350	0.029	0.077	0.055
2180-2280	0.035	0.028	0.174	0.065
3870-3970	0.003	0.003	0.183	0.487
4000-4100	0.022	0.030	0.006	0.023
5180-5280	0.001	0.001	0.111	0.160

Sauer [10] and Wan [12] with different starting points.

Sauer used a combination of iterated and direct local linear models with a fixed number of neighbors (4) on a data set obtained by interpolating the original one. Wan used a recurrent network (FIR-network) with one input unit, two layers of 12 hidden units each, and one output unit. The results show how our lazy iterated approach outperforms non lazy methods and it is largely better than the non iterated lazy in the interval $1000-1100$ which was the object of the competition.

IV. Predicting for the time series competition

The iterated PRESS approach has been applied also to the prediction of a time series in the competition of the *International Workshop on Advanced Black-box techniques for nonlinear modeling*. The available data set consists of 2000 values and the task is to predict the next 200 samples. We adopt a lazy learning method where the selection of neighbors is made according to the iterated PRESS with horizon $h = 2$. The number of neighbors is limited to range from 4 to 8. The authors submitted to the competition two continuations obtained by considering two delay embeddings having different orders ($m = 20$ and $m = 24$ in the Eq. (1)). These orders were selected as they resulted the most promising ones according to a training and test procedure on the available 2000 points.

V. Conclusions

A large amount of literature in nonlinear modeling focuses on the definition of complex architectures having nice properties of nonlinear approximation. Unfortunately,

tuning these models with a limited amount of data requires a large amount of time and often leads to poor generalization. This paper aims to demonstrate how the definition of effective validation procedures can largely compensate the simplicity of the approximator structure. We showed how local models whose parameters have been designed by optimizing the generalization performance can outperform complex neural architectures.

VI. Appendix

The PRESS residual [7] is given by

$$e^{cv}(i) = y_i - \varphi_i^T \beta_{-i} =$$
$$= y_i - \varphi_i^T \left[\mathbf{P} + \frac{\mathbf{P}\varphi_i\varphi_i^T\mathbf{P}}{1-h_{ii}}\right]\Phi_{-i}^T y_{-i} =$$
$$= \frac{e(i)}{1-h_{ii}} \quad (5)$$

where $\mathbf{P} = (\Phi^T\Phi)^{-1}$, φ_i is a vector of dimensionality $[m,1]$, $e(i)$ is the residual and β_{-i} denotes the vector $[m,1]$ of least-squares coefficients computed with the i-th data point aside. The iterated PRESS residual is the value of the regression β_{-i} in the i-th point φ_i shifted of the vector $[m,1]$ of errors (δ_i^T) cumulated in the previous steps. Then we have:

$$e^{it}(i) = y_i - (\varphi_i - \delta_i^T)^T \beta_{-i} =$$
$$= y_i - \varphi_i^T \left[\mathbf{P} + \frac{\mathbf{P}\varphi_i\varphi_i^T\mathbf{P}}{1-h_{ii}}\right]\Phi_{-i}^T y_{-i} +$$
$$+ \delta_i \left[\mathbf{P} + \frac{\mathbf{P}\varphi_i\varphi_i^T\mathbf{P}}{1-h_{ii}}\right]\Phi_{-i}^T y_{-i} =$$
$$= \frac{e(i) + \delta_i\left[(1-h_{ii}) + \mathbf{P}\varphi_i\varphi_i^T\right]\mathbf{P}\Phi_{-i}^T y_{-i}}{1-h_{ii}} =$$
$$= \frac{e(i) + \delta_i\left[(1-h_{ii}) + \mathbf{P}\varphi_i\varphi_i^T\right]\mathbf{P}(\Phi^T y - \varphi_i y_i)}{1-h_{ii}} =$$
$$= \frac{e(i) + \delta_i\left[(1-h_{ii})\mathbf{P}\Phi^T y - (1-h_{ii})\mathbf{P}\varphi_i y_i\right]}{1-h_{ii}} +$$
$$+ \frac{\delta_i\left[\mathbf{P}\varphi_i\varphi_i^T\mathbf{P}\Phi^T y - \mathbf{P}\varphi_i\varphi_i^T\mathbf{P}\varphi_i y_i\right]}{1-h_{ii}} =$$
$$= \frac{e(i) + \delta_i\left[(1-h_{ii})\mathbf{P}\Phi^T y - \mathbf{P}\varphi_i y_i + \mathbf{P}\varphi_i\hat{y}_i\right]}{1-h_{ii}}$$
$$= \frac{e(i) + \delta_i\left[(1-h_{ii})\beta - \mathbf{P}\varphi_i e(i)\right]}{1-h_{ii}} \quad (6)(7)$$

where the following equivalences hold: $\Phi_{-i}^T y_{-i} + \varphi_i y_i = \Phi^T y$ (in Eq. 6), $\varphi_i^T\mathbf{P}\Phi^T y = \hat{y}_i$, $\varphi_i^T\mathbf{P}\varphi_i y_i = h_{ii}$ and $\beta = \mathbf{P}\Phi^T y$ (in Eq. 7).

References

[1] D.W. Aha. Editorial. *Artificial Intelligence Review*, 11(1–5):1–6, 1997.

[2] A.S.Weigend. Time series analysis and prediction. In M.C. Mozer Paul Smolensky and D.E. Rumelhart, editors, *Mathematical Perspectives on Neural Networks*, chapter 12, pages 395–449. Lawrence Erlbaum Associates, 1996.

[3] C.G. Atkeson, A.W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1–5):11–73, 1997.

[4] H. Bersini, M. Birattari, and G. Bontempi. Adaptive memory-based regression methods. In *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks*, 1998. to appear.

[5] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[6] M.C. Mozer. Neural net architectures for temporal sequence processing. In A.S. Weigend and N.A. Gershenfeld, editors, *Time Series Prediction: forecasting the future and understanding the past*, pages 175–193. Addison Wesley, Harlow, UK, 1994.

[7] R.H. Myers. *Classical and Modern Regression with Applications*. PWS-KENT, Boston, MA, 1990.

[8] N. H. Packard, J.P. Crutchfeld, J.D. Farmer, and R.S. Shaw. Geometry from a time series. *Physical Review Letters*, 45(9):712–716, 1980.

[9] B. Pearlmutter. Gradient calculations for dynamic recurrent neural networks. *IEEE Transactions on Neural Networks*, 6(5):1212–1228, 1995.

[10] T. Sauer. Time series prediction by using delay coordinate embedding. In A.S. Weigend and N.A. Gershenfeld, editors, *Time Series Prediction: forecasting the future and understanding the past*, pages 175–193. Addison Wesley, Harlow, UK, 1994.

[11] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36(1):111–147, 1974.

[12] E.A. Wan. Time series prediction using a connectionist network with internal delay lines. In A.S. Weigend and N.A. Gershenfeld, editors, *Time Series Prediction: forecasting the future and understanding the past*, pages 195–217. Addison Wesley, Harlow, UK, 1994.

[13] A.S. Weigend and N.A. Gershenfeld. *Time Series Prediction: forecasting the future and understanding the past.* Addison Wesley, Harlow, UK, 1994.