Long Term Time Series Prediction with Multi-Input Multi-Output Local Learning

G. Bontempi[1]

[1]Machine Learning Group, Département d'Informatique,
ULB, Université Libre de Bruxelles,
gbonte@ulb.ac.be

ESTSP, 2008

Outline

- Long term time series forecasting
- Limitation of current approaches
- Multiple output modeling
- The LL-MIMO algorithm
- Competition

Time series modelling

- A regular time series is a sequence of measurements φ^t of an observable φ at equal time intervals.
- The dynamics of the time series can be represented by a Nonlinear Auto Regressive (NAR) one-step-ahead model

$$\varphi^{t+1} = f\left(\varphi^{t-d}, \varphi^{t-d-1}, \dots, \varphi^{t-d-m+1}\right) + w(t+1)$$

  where the missing information is lumped into a noise term w, m (*dimension*) is the number of past values taken into consideration and d is the lag time.
- A NAR model can be built from training data with conventional machine learning and statistical techniques.

Long term forecasting : current approaches

1. Iterated prediction : the output returned by the one-step-ahead model

$$\varphi^{t+1} = f(\varphi^{t-d}, \varphi^{t-d-1}, \dots, \varphi^{t-d-m+1}) + w(t+1)$$

   is fed back as an input to the following 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. Examples of iterated approaches are recurrent neural networks or local learning iterated techniques.

2. Direct prediction : H different models

$$\varphi^{t+h} = f^h(\varphi^{t-d}, \varphi^{t-d-1}, \dots, \varphi^{t-d-m+1}) + w(t+h)$$

   are required to perform a *H-step-ahead* forecasting. Direct methods often require high functional complexity in order to model the iterated mapping.

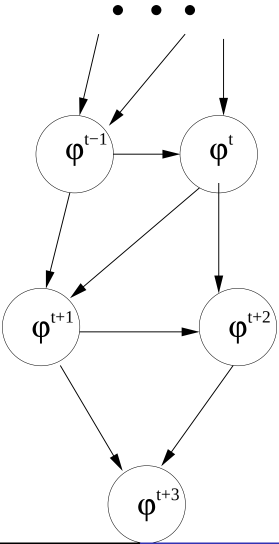
Beyond single-output modelling

1. Iterated and direct techniques for multi-step-ahead prediction share a common feature : they model from historical data a multi-input single-output mapping where the output is the variable φ^{t+1} in the iterated case and the variable φ^{t+h} in the direct case, respectively.
2. For very long term prediction, the modeling of a single-output mapping neglects the existence of stochastic dependencies between future values, (e.g. φ^{t+h} and φ^{t+h+1}) and consequently biases the prediction accuracy.
3. We propose to move from the modeling of single-output mapping to the modeling of multi-output dependencies

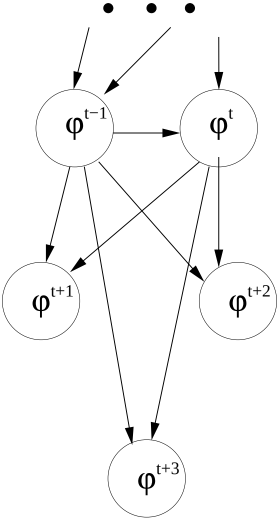
$$X = \{\varphi^{t-d}, \varphi^{t-d-1}, \dots, \varphi^{t-d-m+1}\} \rightarrow Y = \{Y^1, \dots, Y^H\} = \{\varphi^{t+1}, \dots, \varphi^{t+H}\}$$

by adopting a multi-output technique where the predicted value is no more a scalar quantity but a vector of future values of the time series.

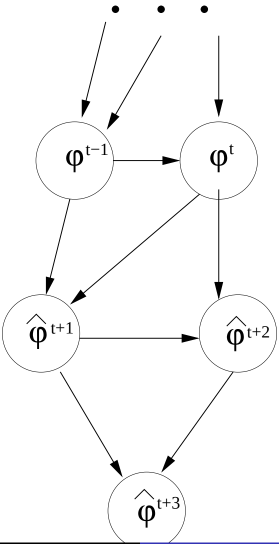
Long term dependencies ($H = 3,\ m = 2,\ d = 0$)



Direct prediction (conditional independence assumption)



Iterated prediction



Locally constant method for multi-output regression

- The time series is embedded into a dataset D_N made of N pairs (X_i, Y_i), where X_i is a temporal pattern of length m, and the vector Y_i is the consecutive temporal pattern of length H.
- Assume for simplicity that the lag $d = 0$. Let

$$\bar{X} = \{\varphi^t, \dots, \varphi^{t-m+1}\}$$

  be the lag embedding vector at time t.
- According to a metric on the space $\mathbb{R}^m$ let $[j]$ be the index of the jth closest neighbor of $\bar{X}$.
- For a given number k of neighbors the H step prediction is a vector whose hth component is the average

$$\hat{Y}_k^h = \frac{1}{k} \sum_{j=1}^{k} Y_{[j]}^h$$

  where $Y_{[j]}$ is the output vector of the jth closest neighbor of $\bar{X}$ in the training set D_N.

Leave-one-out neighbour selection

- We can associate to the estimation $\hat{Y}_h^k$ a multi-step leave-one-error

$$E_k = \frac{1}{H} \sum_{h=1}^{H} \left(e^h\right)^2$$

where e_h is the leave-one-out error of a constant model used to approximate the output at the h step.

- In case of constant model the l-o-o term is easy to derive

$$e^h = \sum_{j=1}^{k} e_j^h, \quad e_j^h = \hat{Y}_{[j]}^h - \frac{\sum_{i \neq j} \hat{Y}_{[i]}^h}{k-1} = k \frac{Y_{[j]}^h - \hat{Y}_k^h}{k-1}$$

- The optimal number of neighbors can be then defined as the number

$$k^* = \arg\min_k E_k$$

Averaging (LL-MIMO-COMB)

- A multi-output approach allows the availability of a large number of estimators once the prediction horizon H is long.
- Suppose $H = 20$ and we estimate φ^{t+10}. We can combine several long term estimators which have an horizon larger than 10 (e.g. all the predictors with horizon between 10 and 20).
- The prediction at time $t+h$ is given by

$$\hat{\varphi}^{t+h} = \frac{\sum_{j=h}^{H} \hat{Y}_{(j)}^{h}}{H-h+1},$$

where in this case the notation $\hat{Y}_{(j)}^{h}$ is used to denote the h^{th} term of the vector prediction returned by a LL-MIMO trained for an horizon $j \geq h$.

Competition training

We compared

1. a conventional iterated approach
2. a direct approach
3. a multi-output LL-MIMO approach
4. a combination of several LL-MIMO predictors (denoted by LL-MIMO-COMB)
5. a combination of the LL-MIMO and the iterated approach (denoted by LL-MIMO-IT).

Results on the training set : average NMSE

Tab.: Average NMSE of the predictions for the three time series. The bold notation stands for significantly better than all the others at 0.05 significativity level of the paired permutation test.

Test data	LL-IT	LL-DIR	LL-MIMO	LL-MIMO-COMB	LL-MIMO-IT
ESTSP1	1.016	0.239	0.240	**0.219**	0.453
ESTSP2	0.426	0.335	0.335	0.326	**0.189**
ESTSP3	1.63e-2	1.05e-2	1.04e-2	**1.02e-2**	1.12e-2

Results on the training set : average NMSE

Tab.: Minimum NMSE of the predictions for time series.

Test data	LL-IT	LL-DIR	LL-MIMO	LL-MIMO-COMB	LL-MIMO-IT
ESTSP1	0.228	0.171	0.172	0.1678	0.190
ESTSP2	0.188	0.130	0.125	0.115	0.104
ESTSP3	1.00e-2	0.96e-2	0.95e-2	0.88e-2	0.93e-2

Prediction example (series ESTSP3)

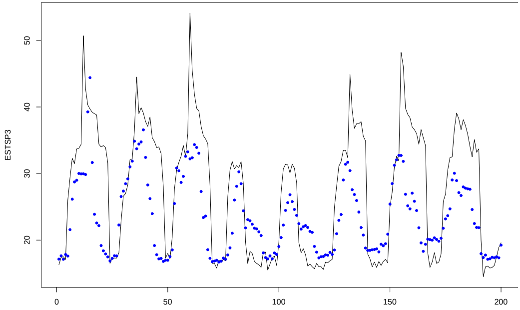


Fig.: ESTSP3 : time series (line) vs. LL-MIMO-COMB prediction (dots).

Conclusion

- Long term forecasting is a difficult problem yet to be solved.
- So far the mainstream approach consists in adapting short term prediction techniques to the long term scenario.
- This paper advocates the need of a major shift in the design of forecasting techniques for long term.
- Two main innovative principles :

1. the outcome of a long term forecasting technique is not a single value but a series itself : multiple output prediction techniques may serve to learn and preserve the stochastic dependency between sequential predicted values.
2. since the prediction is a time series, global criteria at the series level can be successfully employed to design the forecaster (future work).