Chapter 16

Time series prediction

Amaury Lendasse, Francesco Corona, Antti Sorjamaa, Elia Liitiäinen, Tuomas Kärnä, Yu Qi, Emil Eirola, Yoan Miché, Yongnang Ji, Olli Simula

16.1 Introduction

Amaury Lendasse

What is Time series prediction? Time series prediction (TSP) is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; 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 following: how can one analyse and use the past to predict the future? Many techniques exist: linear methods such as ARX, ARMA, etc., and nonlinear ones such as artificial neural networks. In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict the future values. The common difficulty to all the methods is the determination of sufficient and necessary information for an accurate prediction.

A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information.

Our contributions in TSP research. The TSP group is a new research group. It has been created in 2004. A notable achievement has been the organization of the first European Symposium on Time Series Prediction (ESTSP'07) on February 2007 in Helsinki. (`http://www.estsp.org`, [1]). For this symposium, a time series competition has been organized and a benchmark has been created.

In the reporting period 2006 - 2007, TSP research has been established as a new project in the laboratory. Nevertheless, TSP research has already been extended to a new direction: “Chemoinformatics”.

This Chapter starts by introducing some theoretical advances undertaken during the reporting period, including the presentation of the ESTSP´07 competition. Also the problem of input selection for TSP is reported. The applications range includes Chemoinformatics.

16.2 European Symposium on Time Series Prediction

Amaury Lendasse and Antti Sorjamaa

Time series forecasting is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. ESTSP 2007 was a unique opportunity for researcher from Statistics, Neural Networks, Machine Learning, Control and Econometrics to share their knowledge in the field of Time Series Prediction.

The common point to their problems is the following: how can one analyse and use the past to predict the future?

Many techniques exist for the approximation of the underlying process of a time series: linear methods such as ARX, ARMA, etc. , and nonlinear ones such as artificial neural networks. In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict the future values. The common difficulty to all the methods is the determination of sufficient and necessary information for an accurate prediction.

A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information to predict the future values.

Papers were presented orally (single track).

The following is a non-exhaustive list of machine learning, computational intelligence and artificial neural networks topics covered during the ESTSP conferences:

- Short-term prediction
- Long-term prediction
- Econometrics
- Nonlinear models for Time Series Prediction
- Time Series Analysis
- Prediction of non-stationary Time Series
- System Identification
- System Identification for control
- Feature (variable or input) Selection for Time Series
- Selection of Exogenous (external) variables

The goal of the competition is the prediction of the 50 next values (or more) of the time series. The evaluation of the performance was done using the MSE obtained from the prediction of both the 15 and the 50 next values.

So far, there are now 74 values available and the results can be found in http://www.cis.hut.fi/projects/tsp/ESTSP/. In the following figure the predictions of all the competition participants are plotted in blue. In red and in the table below are shown the real values so far.

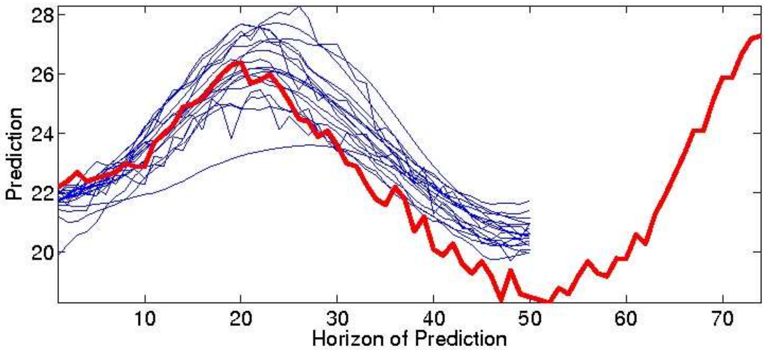

Figure 16.1: The ESTSP Benchmark.

16.3 Methodology for long-term prediction of time series

Amaury Lendasse, Yu Qi, Yoan Miché and Antti Sorjamaa

The time series prediction problem is the prediction of future values based on the previous values and the current value of the time series (see Equation 16.1).

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{16.1}$$

The previous values and the current value of the time series are used as inputs for the prediction model. One-step ahead prediction is needed in general and is referred as Short-Term Prediction. But when multi-step ahead predictions are needed, it is called Long-Term Prediction problem.

Unlike the Short-Term time series prediction, the Long-Term Prediction is typically faced with growing uncertainties arising from various sources. For instance, the accumulation of errors and the lack of information make the prediction more difficult. In Long-Term Prediction, performing multiple steps ahead prediction, there are several alternatives to build models. Two variants of prediction strategies are studied and compared [2]: the Direct (see Equation 16.2) and the Recursive Prediction Strategies (see Equation 16.1).

$$\hat{y}_{t+k} = f_k(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{16.2}$$

16.4 Nonparametric noise estimation

Elia Liitiäinen, Francesco Corona, Emil Eirola and Amaury Lendasse

The residual variance estimation problem (or Nonparametric noise Estimation) is well-known in machine learning and statistics under various contexts. Residual variance estimation can be viewed as the problem of estimating the variance of the part of the output that cannot be modelled with the given set of input variables. This type of information is valuable and gives elegant methods to do model selection. While there exist numerous applications of residual variance estimators to supervised learning, time series analysis and machine learning, it seems that a rigorous and general framework for analysis is still missing. For example, in some publications the theoretical model assumes additive noise and independent identically distributed (iid) variables. The principal objective of our work is to define such a general framework for residual variance estimation by extending its formulation to the non-iid case. The model is chosen to be realistic from the point of view of supervised learning. Secondly, we view two well-known residual variance estimators, the Delta test and the Gamma test in the general setting and we discuss their convergence properties. Based on the theoretical achievements, our general approach seems to open new directions for future research and it appears of fundamental nature [3]. We have also applied NNE for time series prediction [4].

16.5 Chemoinformatics

Francesco Corona, Elia Liitiäinen, Tuomas Kärnä and Amaury Lendasse

Many analytical problems related to spectrometry require predicting a quantitative variable through a set of measured spectral data. For example, one can try to predict a chemical component concentration in a product through its measured infrared spectrum. In recent years, the importance of such problems in various fields including the pharmaceutical, food and textile industries have grown dramatically. The chemical analysis by spectrophotometry rests on the fast acquisition of a great number of spectral data (several hundred, even several thousands).

In spectrometric problems, one is often faced with databases having more variables (spectra components) than samples; and almost all models use at least as many parameters as the number of input variables. These two problems, colinearity and risk of overfitting, already exist in linear models. However, their effect may be even more dramatic when nonlinear models are used (there are usually more parameters than in linear models, and the risk of overfitting is higher). In such high-dimensional problems, it is thus necessary to use a smaller set of variables than the initial one. We have proposed methods to select spectral variables by using concepts from information theory:

- the measure of mutual information [5].
- the measure of topological relevance on the Self-Organizing Map [6]
- the Functional Data Analysis (FDA) [7]
- Nonparametric Noise Estimation [8]

One particular application has been studied in the field of Oil Production.

In this industrial application, there has been applied process data from Neste Oil Oyj. The aim has been to get new empirical modelling tools, which are based on information technology. The outcome has been emphasized on tools, which are suitable in fast data mining from large data sets. The test cases have included:

- Analysis of instrumental data, on-line monitoring data and quality data
- Non-linear processes
- Identification of delays between stages in industrial processes
- Robust variable selection methods

Analysis of instrumental data, on-line monitoring data and quality data The case has been progressed using a real process data set having 13000 on-line samples (time points) and over a thousand variables. The variables contained different blocks: Z (NIR), X (Process variables) and Y (Quality of end product).

References

[1] Amaury Lendasse. European Symposium on Time Series Prediction, ESTSP'07, Amaury Lendasse editor, ISBN 978-951-22-8601-0.

[2] Amaury Sorjamaa, Jin Hao, Nima Reyhani, Yongnang Ji and Amaury Lendasse, Methodology for Long-term Prediction of Time Series Neurocomputing (70), October 16-18, 2007, pp. 2861-2869

[3] E. Liitiäinen, Francesco Corona and Amaury Lendasse, Non-parametric Residual Variance Estimation in Supervised Learning IWANN 2007, International Work-Conference on Artificial Neural Networks, San Sebastian (Spain), June 20-22, Springer-Verlag, Lecture Notes in Computer Science, 2007, pp. 63-71.

[4] Elia Liitiäinen and A. Lendasse, Variable Scaling for Time Series Prediction: Application to the ESTSP'07 and the NN3 Forecasting Competitions IJCNN 2007, International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007.

[5] Fabrice Rossi, Amaury Lendasse, Damien François, Vincent Wertz and Michel Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometrics and Intelligent Laboratory Systems, Volume 80, Issue 2, 15 February 2006, pages 215-226.

[6] Francesco Corona, Satu-Pia Reinikainen, Kari Aaljoki, Anniki Perkkio, Elia Liitiäinen, Roberto Baratti and Amaury Lendasse and Olli Simula. Wavelength selection using the measure of topological relevance on the Self-Organizing Map, Journal of Chemometrics, submitted and accepted in 2007.

[7] Tuomas Kärnä and Amaury Lendasse, Gaussian fitting based FDA for chemometrics, IWANN'07, International Work-Conference on Artificial Neural Networks, San Sebastian, Spain, June 20-22 , 86–193, 2007.

[8] Amaury Lendasse and Francesco Corona Optimal Linear Projection based on Noise Variance Estimation - Application to Spectrometric Modeling SSC10, 10th Scandinavian Symposium on Chemometrics, Lappeenranta (Finland) June 11-15, 2007