Chapter 12

Time series prediction

Amaury Lendasse, Yongnan Ji, Nima Reyhani, Jin Hao, Antti Sorjamaa

12.1 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 a time series prediction challenge and the creation of a new benchmark in the field "The Cats Benchmark" (`http://www.cis.hut.fi/ lendasse/competition/competition.html`).

In the reporting period 2004 - 2004, TSP research has been established as a new project in the laboratory. Nevertheless, TSP research has already been extended to a new direction: "Chemometry".

This Chapter starts by introducing some theoretical advances undertaken during the reporting period, including the presentation and the results of the CATS Benchmark. Also the problem of input selection for TSP is reported. The applications range includes Chemometry.

12.2 The CATS benchmark

Amaury Lendasse

In the CATS competition [2], the goal was the prediction of 100 missing values of the time series; they are grouped in 5 sets of 20 successive values. The prediction methods have then to be applied several times, allowing a better comparison of the performances. Twenty-four papers and predictions were submitted to the competition (organized during IJCNN'04). Seventeen papers were accepted according to the quality of the prediction and the quality of the paper itself. Seven papers have been accepted for oral presentation and ten for poster presentation.

This series is represented in Fig. 12.1. This artificial time series is given with 5,000 data, among which 100 are missing. The missing values are divided in 5 blocks:

- elements 981 to 1,000;
- elements 1,981 to 2,000;
- elements 2,981 to 3,000;
- elements 3,981 to 4,000;
- elements 4,981 to 5,000.

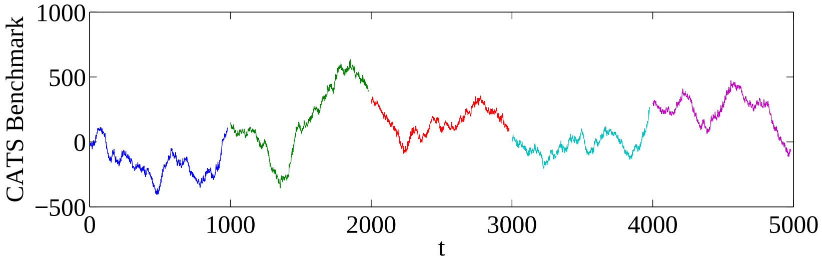

Figure 12.1: The CATS Benchmark.

The Mean Square Error is computed on the 100 missing values. The 24 methods that were submitted to the competition are very different and give very dissimilar results. The Error is in a range between 408 and 1714. It is important to notice that some methods are very good for the prediction of the eighty first values but very bad for the last 20 ones. The method that has been used by the winner of the competition is divided in two parts: the first sub-method provides the short-term prediction and the second sub-method provides the long-term one. Both sub-methods are linear, but according to the author better results could be obtained if the first sub-method was nonlinear. According to this author, the key of a good prediction is this division between two subproblems.

12.3 Methodology for long-term prediction of time series

Amaury Lendasse, Yongnan Ji, Nima Reyhani, Jin Hao, 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 12.1).

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{12.1}$$

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: the Direct and the Recursive Prediction Strategies [5].

12.4 Input selection strategies

Amaury Lendasse, Yongnan Ji, Nima Reyhani, Jin Hao, Antti Sorjamaa

Input selection is an essential pre-processing stage to guarantee high accuracy, efficiency and scalability in problems such as machine learning, especially when the number of observations is relatively small compared to the number of inputs. It has been the subject in many application domains like pattern recognition, process identification, time series modeling and econometrics. Problems that occur due to poor selection of input variables are:

- If the input dimensionality is too large, the 'curse of dimensionality' problem may happen. Moreover, the computational complexity and memory requirements of the learning model increase. Additional unrelated inputs lead to poor models (lack of generalization).
- Understanding complex models (too many inputs) is more difficult than simple models (less inputs), which can provide comparable good performances.

In the TSP group, two input selection methods based on different criteria have been studied: Mutual Information and Nonparametric Noise Estimator [3, 4, 6, 7, 8].

12.5 Chemometry

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 two concept from information theory: the measure of mutual information [1] and the nonparametric noise estimation.

References

[1] 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.

[2] Amaury Lendasse, Erkki Oja, Olli Simula, and Michel Verleysen, Time Series Prediction Competition: The CATS Benchmark, *Proceedings of IJCNN 2004*, International Joint Conference on Neural Networks, Budapest (Hungary), vol. II, 25-29 July 2004, pages 1615–1620.

[3] Antti Sorjamaa, Jin Hao, and Amaury Lendasse. Mutual Information and k-Nearest Neighbors approximator for Time Series Predictions. In *Lecture Notes in Computer Science, Proceedings of ICANN 2005*, Publisher: Springer-Verlag GmbH, volume 3697, pages 553–558, Warsaw, Poland, September 2005.

[4] Amaury Lendasse, Yongnan Ji, Nima Reyhani, and Michel Verleysen. LS-SVM Hyperparameter Selection with a Nonparametric Noise Estimator. In *Lecture Notes in Computer Science, Proceedings of ICANN 2005*, Publisher: Springer-Verlag GmbH, volume 3697, pages 625–630, Warsaw, Poland, September 2005.

[5] Yongnan Ji, Jin Hao, Nima Reyhani, and Amaury Lendasse. Direct and Recursive Prediction of Time Series Using Mutual Information Selection. In *Lecture Notes in Computer Science, Proceedings of IWANN 2005*, Publisher: Springer-Verlag GmbH, volume 3512, pages 1010–1017, July 2005.

[6] Antti Sorjamaa, Nima Reyhani, and Amaury Lendasse. Input and Structure Selection for k-NN Approximator. In *Lecture Notes in Computer Science, Proceedings of IWANN 2005*, Publisher: Springer-Verlag GmbH, volume 3512, pages 985–991, July 2005.

[7] Antti Sorjamaa, Amaury Lendasse, and Michel Verleysen. Pruned Lazy Learning Models for Time Series Prediction. In *Proceedings of ESANN 2005*, pages 509–514, Bruges, Belgium, April 2005.

[8] Nima Reyhani, Jin Hao, Yongnan Ji, and Amaury Lendasse. Mutual Information and Gamma Test for Input Selection. In *Proceedings of ESANN 2005*, pages 503–508, Bruges, Belgium, April 2005.