Variable Scaling for Time Series Prediction

Francesco Corona and Amaury Lendasse [*]

Helsinki University of Technology - Laboratory of Computer and Information Science
P.O. Box 5400, 02015 HUT - Finland

Abstract. In this paper, variable selection and variable scaling are used in order to select the best regressor for the problem of time series prediction. Direct prediction methodology is used instead of the classic recursive methodology. Least Squares Support Vector Machines (LS-SVM) are used in order to avoid local minimal in the training phase of the model. The global methodology is applied to the time series competition dataset.

1 Introduction

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. 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. [1], and nonlinear ones such as artificial neural networks [2]. 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 [2].

In this paper, a global methodology to perform direct prediction is presented. It includes variable selection and variable scaling. The variable selection criterion is based on a Nonparametric Noise Estimation (NNE) performed by Delta Test.

In this paper, Least Squares Support Vector Machines (LS-SVM) are used as nonlinear models in order to avoid local minima problems [3].

Section 2 presents the prediction strategy for the Long-Term Prediction of Time Series. In Section 3 Delta Test is introduced. Section 4 introduces the variable selection and scaling selection. The prediction model LS-SVM is briefly summarized in Section 5 and experimental results are shown in Section 6 using the competition dataset.

[*] Part the work of F. Corona and A. Lendasse is supported by the project of New Information Processing Principles, 44886, of the Academy of Finland. The work of A. Lendasse is supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

2 Time Series Prediction

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 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 to as Short-Term Prediction. But when multi-step ahead predictions are needed, it is called a 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 step ahead prediction, there are several alternatives to build models. In the following sections, two variants of prediction strategies are introduced and compared: the Direct and the Recursive Prediction Strategies.

2.1 Recursive Prediction Strategy

To predict several steps ahead values of a time series, Recursive Strategy seems to be the most intuitive and simple method. It uses the predicted values as known data to predict the next ones. In more detail, the model can be constructed by first making one-step ahead prediction:

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

where M denotes the number inputs. The regressor of the model is defined as the vector of inputs: $y_t, y_{t-1}, ..., y_{t-M+1}$. It is possible to use also exogenous variables as inputs in the regressor, but they are not considered here in order to simplify the notation. Nevertheless, the presented global methodology can also be used with exogenous variables.

To predict the next value, the same model is used:

$$\hat{y}_{t+2} = f_1(\hat{y}_{t+1}, y_t, y_{t-1}, ..., y_{t-M+2}). \tag{2}$$

In Equation 2, the predicted value of $\hat{y}_{t+1}$ is used instead of the true value, which is unknown. Then, for the H-steps ahead prediction, $\hat{y}_{t+2}$ to $\hat{y}_{t+H}$ are predicted iteratively. So, when the regressor length M is larger than H, there are $M - H$ real data in the regressor to predict the H^{th} step. But when H exceeds M, all the inputs are the predicted values. The use of the predicted values as inputs deteriorates the accuracy of the prediction.

2.2 Direct Prediction Strategy

Another strategy for the Long-Term Prediction is the Direct Strategy. For the H-steps ahead prediction, the model is

$$\hat{y}_{t+h} = f_h(y_t, y_{t-1}, ..., y_{t-M+1}) \text{ with } 1 \leq h \leq H. \tag{3}$$

This strategy estimates H direct models between the regressor (which does not contain any predicted values) and the H outputs. The errors in the predicted values are not accumulated in the next prediction. When all the values, from $\hat{y}_{t+1}$ to $\hat{y}_{t+H}$, need to be predicted, H different models must be built. The direct strategy increases the complexity of the prediction, but more accurate results are achieved.

3 Nonparametric Noise Estimator using the Delta Test

Delta Test (DT) is a technique for estimating the variance of the noise, or the mean square error (MSE), that can be achieved without overfitting [4]. The evaluation of the NNE is done using the DT estimation introduced by Stefansson in [5].

Given N input-output pairs: $(x_i, y_i) \in \mathbb{R}^M \times \mathbb{R}$, the relationship between x_i and y_i can be expressed as:

$$y_i = f(x_i) + r_i, \tag{4}$$

where f is the unknown function and r is the noise. The Delta Test estimates the variance of the noise r.

The DT is useful for evaluating the nonlinear correlation between two random variables, namely, input and output pairs. The DT has been introduced for model selection but also for variable selection: the set of inputs that minimizes the DT is the one that is selected. Indeed, according to the GT, the selected set of variables is the one that represents the relationship between variables and output in the most deterministic way.

DT is based on hypotheses coming from the continuity of the regression function. If two points x and x' are close in the input space, the continuity of regression function implies the outputs $f(x)$ and $f(x')$ will be close enough in the output space. Alternatively, if the corresponding output values are not close in the output space, this is due to the influence of the noise.

Let us denote the first nearest neighbor of the point x_i in the set $\{x_1, \ldots, x_N\}$ by x_{NN}. Then the delta test, δ is defined as:

$$\delta = \frac{1}{2N} \sum_{i=1}^{N} \left| y_{NN(i)} - y_i \right|^2, \tag{5}$$

where $y_{NN(i)}$ is the output of $x_{NN(i)}$. For the proof of the convergence of the Delta Test, see [4].

4 Variable and Scaling Selection

Variable scaling is a usual preprocessing step in both function approximation and time series analysis. In scaling, weights are used to reflect the relevance of the input variables to the output to be estimated. That is, scaling seeks for redundant inputs and assigns them low weights to reduce the corresponding influence on the learning process. In such a context, it is clear that variable selection is a particular case of scaling: by weighting irrelevant variables by zero we are, indeed, enforcing selection. For the sake of brevity, only the main concepts referring to the regression problem are presented here. Nevertheless, the extension to time series analysis is trivial.

4.1 Scaling the Input Space with Mahalanobis Matrices

The Mahalanobis distance $d_M(x_i, x_j)$ of two d-dimensional observations x_i, x_j is a proximity (or 'similarity') measure defined on the dependencies between the embedding dimensions. Formally, $d_M(x_i, x_j)$ extends the traditional Euclidean distance $d(x_i, x_j) = [(x_i - x_j)^T (x_i - x_j)]^{1/2}$ transforming the observations subspace by means of a $(d \times d)$ full-rank matrix M:

$$d(x_i, x_j) = [(x_i - x_j)^T M (x_i - x_j)]^{1/2}, \tag{6}$$

From the previous equation, it is evident that: i) if $M = I$ then the original Euclidean metric is retained, and ii) if M is a $(d \times d)$ diagonal matrix then the original space is simply rescaled according to the diagonal elements. Matrix M is also symmetric and semi-definite positive, by definition. Moreover, the Mahalanobis matrix M can be factorized as:

$$M = S^T S, \tag{7}$$

with a matrix S that can linearly map the observations into the subspace spanned by the eigenvectors of the transformation. The learned metric in the projection subspace is still the Euclidean distance, that is:

$$d(x_i, x_j) = [(x_i - x_j)^T M (x_i - x_j)]^{1/2} = [(Sx_i - Sx_j)^T (Sx_i - Sx_j)]^{1/2}, \tag{8}$$

where, by restricting S to be a non-square ($s * d$, with $s < d$) matrix, the transformation performs both a reduction of the dimensionality and the scaling of the original input subspace. The resulting subspace has an induced global metric of lower rank suitable for reducing the 'curse of dimensionality'.

In this paper, we use a diagonal matrix M that is optimized in order to minimize the delta test estimation in the scaled space define by S. Details about the optimization method are given the the experiments section.

5 Nonlinear Models

In this paper, Least Squares Support Vector Machines (LS-SVM) are used as nonlinear models [3], which are defined in their primal weight space by [6, 7]

$$\hat{y} = \omega^T \varphi(\mathbf{x}) + b, \tag{9}$$

where $\varphi(\mathbf{x})$ is a function, which maps the input space into a higher-dimensional feature space, $\mathbf{x}$ is the vector of inputs. ω and b are the parameters of the model. The optimization problem can be formulated as

$$\min_{\omega,b,e} J(\omega,e) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2, \quad (10)$$

$$\text{subject to} \quad y_i = \omega^T\varphi(\mathbf{x}_i) + b + e_i, i = 1, ..., N, \quad (11)$$

and the solution is

$$h(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b. \quad (12)$$

In the above equations, i refers to the index of a sample and $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel function defined as the dot product between the $\varphi(\mathbf{x})^T$ and $\varphi(\mathbf{x})$. Training methods for the estimation of the ω and b parameters can be found in [6].

6 Experimental Results

In this paper, the ESTSP2007 competition dataset is used as an example. It includes a total of 875 values. The dataset is shown in Figure 1.

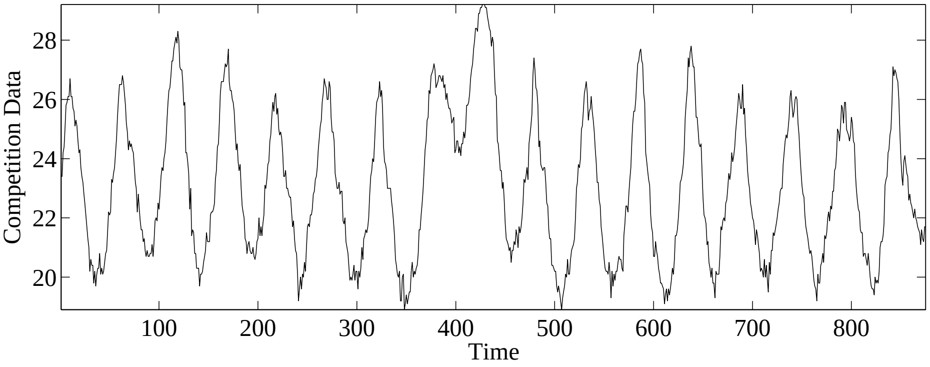

Fig. 1: Competition dataset.

In order to test the methodology, the dataset is divided into two sets, a small learning set and the global learning set. The small learning set consists of 465 first values and the global learning set consists in the 875 values. The regressor size is set to 10 after many trial and error experiments. The small learning set is used in order to evaluate the performances of the methodology.

The variable scaling is selected in order to minimize the Delta Test estimation. Because the DT is not continuous with respect to the scaling factors, a forward-backward optimization is used. The variable scaling coefficients are selected between a set of discrete values: [0 0.1 0.2 ... 0.9 1]. This discretization provides satisfactory results and reduces computational time.

The variable scaling is performed for each of the 50 prediction models from equation 3 used in direct prediction methodology. The estimation of the NNE (using Delta Test) are shown in Figure 2.

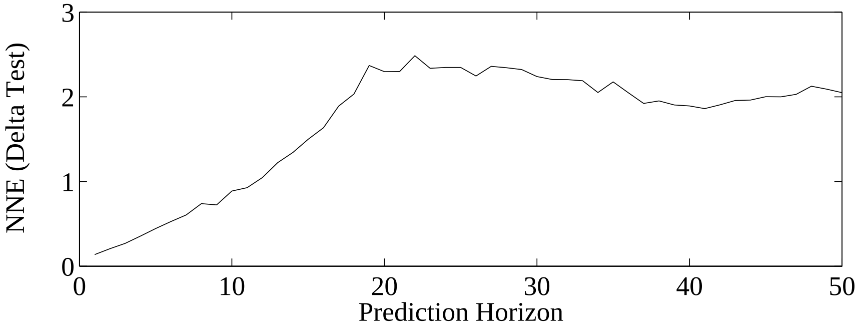

Fig. 2: Estimation of the NNE (using Delta Test) with respect to the horizon of prediction.

The result of the 50 step-ahead prediction is represented in figure 3.

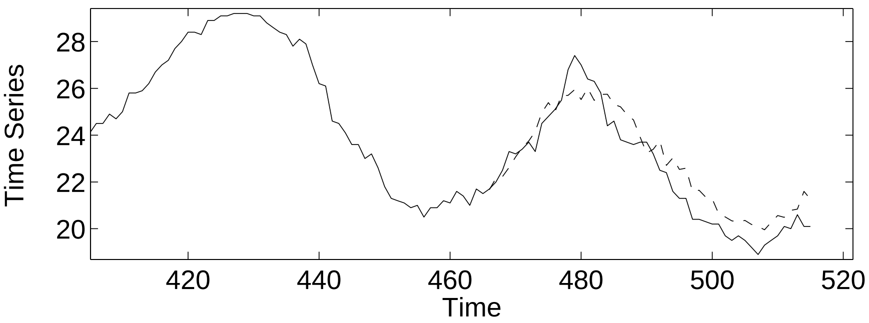

Fig. 3: Comparison between the time series (solid line) and the prediction (dashed line)

Then, the same methodology is used with the global learning set in order to predict the competition values. The estimation of the NNE (using Delta Test) are shown in Figure 4.

The result of the 50 step-ahead prediction is represented in figure 5.

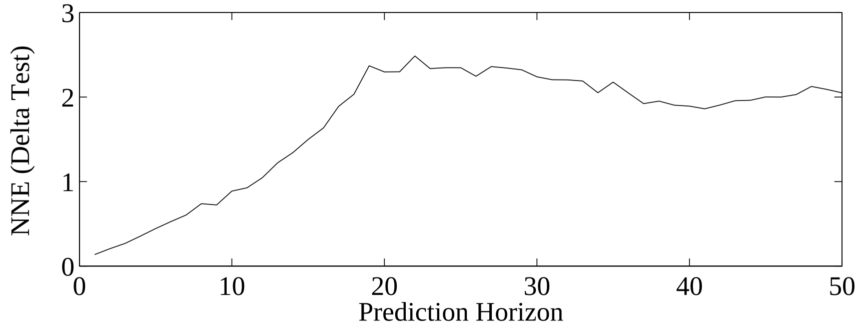

Fig. 4: Estimation of the NNE (using Delta Test) with respect to the horizon of prediction.

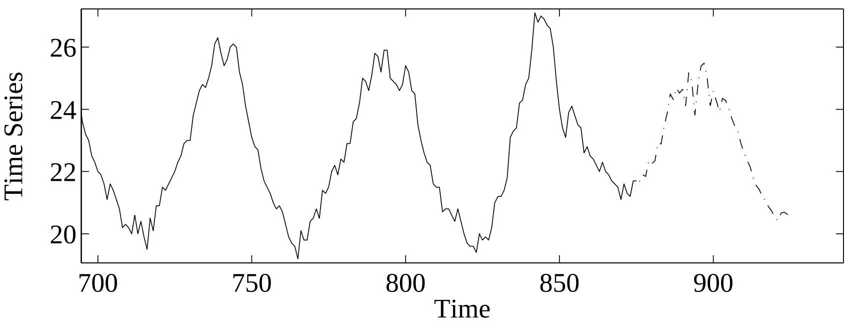

Fig. 5: Prediction of 50 next values of the competition dataset. The real values are presented by the solid line and the dashed one presents the prediction.

7 Conclusion

In this paper, we have presented a methodology for the longterm prediction of time series.

This methodology uses direct prediction methodology. This increases the computational time but improves the quality of the results.

In order to perform the variable scaling, Delta Test estimation is used. The scaling that minimized the NNE is selected. To reduce the computational time, a discrete scaling is used and a forward-backward optimization is selected.

Further research will be done to improve the minimization of the NNE estimation. Other experiments will be performed in the fields of time series prediction and function approximation.

References

[1] L. Ljung. *System identification theory for User.* Prentice-Hall, Englewood CliPs, NJ, 1987.

[2] A.S. Weigend and N.A. Gershenfeld. *Times Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley Publishing Company, 1994.

[3] Johan A K Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific Publishing Co., Pte, Ltd. (Singapore), November 2002.

[4] A. J. Jones. New tools in non-linear modeling and prediction. *Computational Management Science*, 1:109–149, 2004.

[5] Adalbjörn Stefansson, N. Koncar, and Antonia J. Jones. A note on the gamma test. *Neural Computing & Applications*, (5(3)):131–133, 1997.

[6] Available from http://www.esat.kuleuven.ac.be/sista/lssvmlab/.

[7] Johan A. K. Suykens, Jos De Brabanter, L. Lukas, and Joos Vandewalle. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48:85–105, 2002.