ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Determination of the Mahalanobis Matrix using Nonparametric Noise Estimations

A. Lendasse[1], F. Corona[1,2], J. Hao[1], N. Reyhani[1] and M. Verleysen[3] *

1- Helsinki University of Technology - Neural Network Research Centre
P.O. Box 5400, FI-02015 HUT - Espoo, Finland

2- Università di Cagliari - Dept. of Chemical Engineering and Materials
Piazza d'Armi 1, I-9123 - Cagliari, Italy

3- Université Catholique de Louvain - Machine Learning Group
Place du Levant 3, B-1348 - Louvain-la-Neuve, Belgium

Abstract. In this paper, the problem of an optimal transformation of the input space for function approximation problems is addressed. The transformation is defined determining the Mahalanobis matrix that minimizes the variance of noise. To compute variance of the noise, a nonparametric estimator called the Delta Test paradigm is used. The proposed approach is illlustrated on two different benchmarks.

1 Introduction

In this paper, the problem of an optimal transformation of the input space for function approximation problems is addressed. In the general context of function approximation from data, we have observations $\{\mathbf{x}_i, y_i\}_{i=1}^N \in \mathbb{R}^d \times \mathbb{R}$ and the problem consists of reproducing the underlying functionality $y_i = f(\mathbf{x}_i) + \epsilon_i$ between the inputs and the output variables. A formally analogous problem is the analysis of a time series where, having at disposal the observed temporal evolution $\{x_t\}_{t=1}^N \in \mathbb{R}$ of some input variable, we want to describe its future output approximating the object: $x_{t>N} = f(x_{t<N}) + \epsilon_t$.

These two common tasks in machine learning share the necessity to select only a subset of the input variables that is truly relevant for modeling the system that generated the observations (for an exhaustive review, cfr. [8]). In addition, since the final model should not achieve an accuracy smaller in magnitude than level of noise in the observations, an independent model to estimate the noise (the terms ϵ_i and ϵ_t mentioned above) can also be defined.

In this work, we propose the application of a nonparametric noise estimator known as the Delta Test paradigm [4] in order to determine an optimal transformation of the input subspace that enforces appropriate variable selection. In our approach, the critical step of variable selection is approached as a subprolem in the general framework of scaling the inputs. The transformation is performed

*A. Lendasse, J. Hao and N. Reyhani acknowledge the support from the Academy of Finland, through project *New Information Processing Principles, 44886*. A. Lendasse acknowledges the support from the IST Programme of the European Community, through the *PASCAL Network of Excellence, IST-2002-506778*. M. Verleysen is Research Director of the Belgian FNRS.

determining the Mahalanobis matrix [6] that satisfies the optimality criterion of minimizing the variance of the noise between the input and the ouput variables. As such, the illustrated methodology is suitable for function approximators that make explicit use of the metric properties of the embedding space of the observations: e.g., k-Nearest Neighbors (k-NN), Self-Organizing Maps (SOM), Support Vector Machines (SVM) and Radial Basis Functions (RBF) models.

The paper is organized as follows. In Section 2, we illustrate the adopted methodology and briefly recall the basic properties of the considerd algorithms. Section 3 supports the presentation reporting the preliminary results of the proposed technique with two benchmarks in the domain of nonlinear regression and time series prediction using k-NN function approximators.

2 Methodology and Algorithms

Input 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.

2.1 Transforming the Input Space with Mahalanobis Matrices

The Mahalanobis distance $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$ of two d-dimensional observations $\{\mathbf{x}_i, \mathbf{x}_j\}$ is a proximity (or 'similarity') measure defined on the dependencies between the embedding dimensions [6]. Formally, $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$ extends the traditional Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j) = [(\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j)]^{1/2}$ transforming the observations' subspace by means of a $(d \times d)$ full-rank matrix $\mathbf{M}$:

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} \tag{1}$$

From Equation 1, it is evident that: i) if $\mathbf{M} = \mathbf{I}$ then the original Euclidean metric is retained, and ii) if $\mathbf{M}$ is a $(d \times d)$ diagonal matrix then the original space is simply rescaled according to the diagonal elements. Matrix $\mathbf{M}$ is also symmetric and semi-definite positive, by definition. Moreover, the Mahalanobis matrix $\mathbf{M}$ can be factorized as:

$$\mathbf{M} = \mathbf{S}^\top \mathbf{S} \tag{2}$$

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

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{(\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{x}_j)^\top (\mathbf{S}\mathbf{x}_i - \mathbf{S}\mathbf{x}_j)} \quad (3)$$

where, by restricting $\mathbf{S}$ to be a non-square ($s \times d$, 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'.

2.2 Nonparametric Noise Estimation using the Delta Test

The Delta Test [4] is a nonparametric paradigm for estimating the variance of noise (ϵ, in the following). In its basic formulation, the Delta Test exploits the 'similarity' in the behaviour of the noise of two closeby observations. In details, for a given d-dimensional observation $\mathbf{x}_i$ and some close neighbor $\mathbf{x}'_i$, the expected Mean Square Error on the corresponding outputs (y_i and y'_i) will estimate the variance of the noise as the distance $\delta(\mathbf{x}'_i, \mathbf{x}_i)$ tends to zero:

$$\mathrm{var}(\epsilon) \leftarrow \mathrm{E}\Big\langle \frac{1}{2}(y'_i - y_i)^2 \Big| \|\mathbf{x}'_i - \mathbf{x}_i\| < \delta \Big\rangle, \text{for } \delta \to 0 \quad (4)$$

Despite this approach appears to be promising in its formulation, it fails when the size of the data is small if compared to the complexity of the underlying function and noise distribution. An improvement of the Delta Test can be achieved exploiting the k-Nearest Neighbors distances between the observations in the input subspace and the corresponding outputs. This leads to an approach called here Nonparametric Noise Estimation. According to [4], an estimate for the variance of the noise is represented by the intercept of a linear regression line between the average of k nearest distances in the input subspace and the corresponding average of k nearest distances in the output subspace. Formally, the nearest distances can be espressed as:

$$\delta(k) = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_{NN(\mathbf{x}_i,k)} - \mathbf{x}_i \right\|^2 \text{ and } \gamma(k) = \frac{1}{2N} \sum_{i=1}^{N} \left\| y_{NN(\mathbf{x}_i,k)} - y_i \right\|^2 \quad (5)$$

where, the index $NN(\mathbf{x}_i, k)$ refers to the k-*th* neighbor of $\mathbf{x}_i$. A proof (also referred to as Gamma Test) based on a generalization of the Chybechov inequality and the properties of the k-Nearest Neighbors is reported in [3]. The proof depends on the assumptions: i) the Jacobian and the Hessian of $y = f(\mathbf{x})$ exist, ii) the 3-rd to 4-th moments of the noise distribution exist, and iii) the noise is independent from the inputs.

In [3], it is demonstrated that the variance of noise is estimated as the intercept of the vertical line $\delta(k) = 0$ and the regression line between $\gamma(k)$ and $\delta(k)$, where $k \geq 1$. In this paper, we applied the Delta Test that assumes $k = 1$.

2.3 Function Approximation using k-Nearest Neighbors

The k-Nearest Neighbors function approximator [2] is a simple, but powerful method. In its basic formulation, it assumes that observations that are closeby (or, again, 'similar') in the input space have corresponding outputs that are also close. The Euclidean distance is a natural measure typically used to assess the proximity. For a given input observation $\mathbf{x}_i$, we estimate the output $\hat{y}_i$ by averaging in the neighborhood, so that:

$$\hat{y}_i = \frac{1}{k} \sum_{j=1}^{k} y_{P(j)} \tag{6}$$

where, $P(j)$ is the index number of the j-th nearest neighbor $\mathbf{x}_i$ and k is the number of neighbors used in the estimation. We use the same neighborhood size for every observation, a global k, which is to be determined beforehand. As for the validation of k, the Leave-One-Out resampling method is adopted [1].

3 Simulation Results

3.1 The Stereopsis Regression Problem

The Stereopsis dataset [5] provides a benchmark for regression problems. The number of input variables is equal to 4, and we used 192 samples for training, and 300 for testing the model. The output of the SRP is depicted in Figure 1.

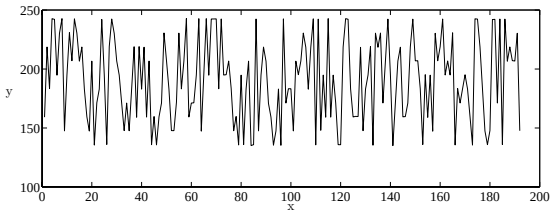

Fig. 1: Learning data set of the Stereopsis regression problem.

First, the k-NN regressor is built in the original input subspace. The $NMSE$ obtained on the test set ($NMSE_{test}$) is equal to 0.16.

Secondly, the diagonal Mahalanobis matrix (equivalent to scaling) is optimized using the Genetic Algorithms Toolbox from Matlab. The achieved $NMSE_{test}$ is reduced to 0.05.

Finally, a Mahalanobis matrix equivalent to a projection into a 1-dimensional ($s = 1$) subspace is found to be optimal: the corresponding $NMSE_{test}$ is 0.01.

Figure 2a represents the output plotted against the 1-dimensional projection of the original input space. Figure 2b presents also the corresponding estimates on the test set obtained with the k-NN approximator.

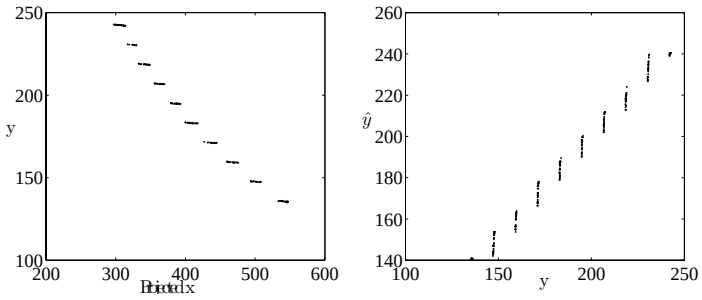

Fig. 2: Output of the Stereopsis test dataset: a, left) plotted against the optimal projection of the inputs, b, right) plotted against the predictions using k-NN.

3.2 The SantaFeA Time Series

As for the time series prediction, the Santa Fe Laser Dataset [7] was used (Figure 3). It consist of a 10000 temporal observations of a 1-dimensional quantity from a far-infrared laser in chaotic regime. 1000 observations are used for training, and 9000 for testing.

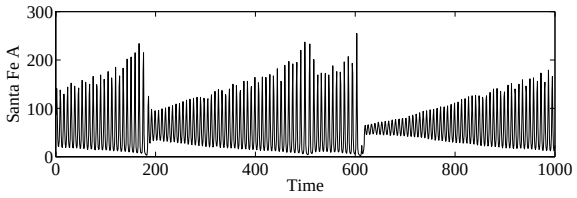

Fig. 3: SantaFeA times series, 1000 values training set.

First, a k-NN is built using a 12-dimensional regressor input. The $NMSE_{test}$ that is achieved is 0.088. Secondly, the optimal diagonal Mahalanobis matrix (equivalent to scaling) is determined. The achieved $NMSE_{test}$ is then 0.027. Finally, a Mahalanobis matrix equivalent to a projection into a 5-dimensional ($s = 5$) subspace is found to be optimal: the achieved $NMSE_{test}$ is 0.026.

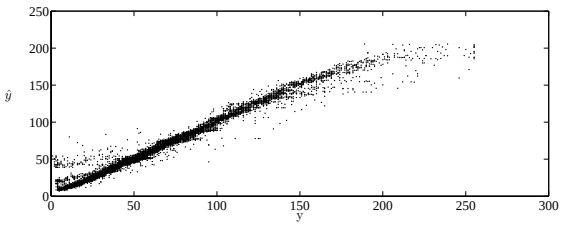

Fig. 4: Prediction of the test set of SantaFeA using the projected inputs.

Figure 4 represents the $NMSE_{test}$ using the projected inputs. In Figure 5 a recursive prediction of the first 100 values of the test set is also represented.

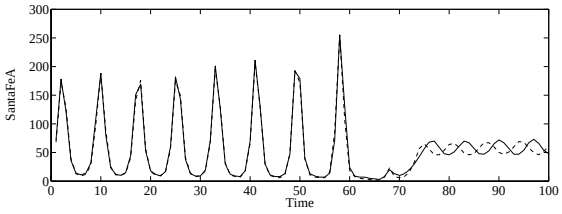

Fig. 5: Recursive prediction on the test set of SantaFeA.

4 Conclusions

In this paper, a nonparametric noise estimator is used to optimizing the Mahalanobis matrix for both scaling and projection of the input subspaces for function approximation. The method has been tested on two different benchmarks using the k-NN approximator.

The results obtained on both benchmarks show that the accuracy of the estimations is improved by scaling the inputs (the diagonal case). Furthermore, if also the projection is performed, even better results are achieved. The computational time of the optimal scaling was found to be approximately 1 minute. Conversely, optimizing the projection is demanding (around 100 times longer). This is due to the larger number of variables and the optimization technique.

As further work, alternative optimization methods will be investigated as well as data with higher dimensionality and non homogeneous characteristics. Also different nonlinear models, like LS-SVM and RBF network, will be investigated.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[2] L. Devroye, "The Uniform Convergence of Nearest Neighbor Regression Function Estimators and their Application in Optimization", *IEEE Trans. Inf. Theory*, 24(2), 1978.

[3] D. Evans and A. J. Jones, "A Proof of the Gamma Test", *Proc. R. Soc Lond. A*, 458:2759-2799, 2002.

[4] A. J. Jones, "New Tools in Non-linear Modeling and Prediction", *Computational Management Science*, 1:109-149, 2004.

[5] PASCAL Challenges Workshop, Evaluating Predictive Uncertainty Challenge: `http://predict.kyb.tuebingen.mpg.de/pages/dataset_regression.php`.

[6] G. Taguchi and R. Jugulum. *The Mahalanobis-Taguchi Strategy*, John Willey & Sons, New York, 2002.

[7] Times Series Prediction, Forecasting the Future and Understanding the Past: `http://www-psych.stanford.edu/ andreas/time-series/santafe.html`.

[8] M. Verleysen, The Curse of Dimensionality in Data Mining. In *proceedings of the International Work-Conference on Artificial Neural Networks* (IWANN 2005), pages 758-770, June 08-10, Barcelona (Spain), 2005.