ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Time Series Input Selection using Multiple Kernel Learning

Loris Foresti, Devis Tuia, Vadim Timonin and Mikhail Kanevski

Institute of Geomatics and Analysis of Risk - University of Lausanne
Amphipôle, University of Lausanne, 1015 Lausanne - Switzerland

Abstract. In this paper we study the relevance of multiple kernel learning (MKL) for the automatic selection of time series inputs. Recently, MKL has gained great attention in the machine learning community due to its flexibility in modelling complex patterns and performing feature selection. In general, MKL constructs the kernel as a weighted linear combination of basis kernels, exploiting different sources of information. An efficient algorithm wrapping a Support Vector Regression model for optimizing the MKL weights, named *SimpleMKL*, is used for the analysis. In this sense, MKL performs feature selection by discarding inputs/kernels with low or null weights. The approach proposed is tested with simulated linear and nonlinear time series (AutoRegressive, Henon and Lorenz series).

1 Introduction

The analysis and modelling of time series is nowadays an important topic of research both from the theoretical and applied viewpoints. The well-known Taken's embedding theorem [1] states that, for a sufficiently long time series of a chaotic system, it is possible to reconstruct the underlying dynamics in the state space with a time delay embedding. Therefore and in order to develop efficient forecasting models, it is necessary to find a correct embedding of the time series to reconstruct its trajectory in the state space. The task consists in selecting two parameters. The first one is the lag (delay) τ, that is commonly estimated by means of mutual information and autocorrelation functions [1, 2]. Second is the embedding dimension d_e, which unfolds the time series and thus facilitates the use of a forecasting model; this dimension can be estimated using false nearest neighbors and correlation dimension, among many methods [1].

Consider a m-dimensional state vector built with time delay embedding

$$\mathbf{x} = [x_t, x_{t-\tau}, x_{t-2\tau}, ..., x_{t-m\tau}] \tag{1}$$

where $x_{t-\tau}$ is the value of the series at time $t-\tau$. To predict the value at $t+1$, i.e., $y = x_{t+1}$, we need a function $f(\mathbf{x})$ which depends on the lagged values of Eq. (1). This problem can be stated as a (non)linear multiple regression problem $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$, which can be solved using Support Vector Regression (SVR) as shown in [3, 4].

Selecting the optimal τ and d_e is a difficult task. An alternative way is to select from a sufficiently large $\mathbf{x}$ vector the inputs that are relevant for prediction. In this case, the prediction performance and, more importantly, the

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

interpretability of the results will depend strongly on the correct selection of inputs characterized by delay and embedding. In the literature, input selection has been carried out using filters such as $\delta-$ and $\gamma-$tests (see [5], [6]).

In this paper we propose to perform input selection by means of multiple kernel learning (MKL, [7]). More specifically, we adapt and apply the recently proposed *SimpleMKL* algorithm [8]. MKL framework allows to dedicate kernels to model inputs and to find their optimal weighted combination through gradient optimization. In the context of continuous valued functions, MKL wraps a Support Vector Regression (SVR) solver for optimizing both the SVR coefficients and the kernel weights. In this context, the definition of the optimal embedding is reduced to a feature selection problem since the inputs associated with kernels receiving a zero weight are discarded from the final solution. In spatial context the problem of feature selection using MKL was already considered in [9].

2 Multiple Kernel Learning for regression

Kernel methods such as Support Vector Regression have proved to yield robust solutions for nonlinear regression estimation. SVR minimizes a robust cost function comprising empirical risk and model's complexity. Such approach allows building models with high generalization performance. Trained with a suitable kernel function $K(\mathbf{x}, \mathbf{x}')$ accounting for the similarity between data points $\mathbf{x}$ and $\mathbf{x}'$, SVR achieves nonlinear function estimation.

For complex patterns choosing the right kernel function for SVR is a crucial task. Standard closed-form kernels may encode a rigid representation of the data and do not provide interpretable solutions in terms of the importance of the input features. A more flexible kernel can be considered by using a convex weighted combination of basis kernels as it is proposed in [10]:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M} d_m K_m(\mathbf{x}, \mathbf{x}') \quad \text{with} \quad d_m \geq 0 \quad \text{and} \quad \sum_{m=1}^{M} d_m = 1 \qquad (2)$$

where d_m are kernel weights subjected to the sum-to-one and positivity constraints. Multiple kernel learning is the simultaneous optimization of the vectors of the kernel weights $\mathbf{d} = [d_1, ..., d_M]^\top$ and of the SVR coefficients $\boldsymbol{\alpha}$. This corresponds to the search of a function $f(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x})$, with $f_m \in \mathcal{H}_m$ and for which the RKHS of the final kernel $\mathcal{H}$ is the sum of the subspaces $\mathcal{H}_m$ (see the derivation in [8]). The basis kernels K_m can be dedicated to single or groups of features, in order to reveal their importance in the final mixture by their respective weight d_m.

2.1 Simple Multiple Kernel Learning

Even if attractive in its original formulation, MKL becomes quickly intractable with the increase of the number of both the kernels and samples considered. Therefore, several algorithms have been proposed to solve the MKL problem

efficiently [8, 11]. In this contribution, we consider *SimpleMKL* algorithm proposed in [8].

SimpleMKL wraps a standard SVR solver using as kernel function the weighted combination described in Eq. (2). Analytical differentiation of the SVR dual function with respect to d_m provides the gradient used in an iterative optimization scheme till convergence. Positivity and equality constraints over d_m's are granted by the use of a reduced gradient. Sparsity of the solution is a result of l_1-norm regularization of the d_m coefficients.

3 Experiments

Time series analysis touches at several real life problems like environmental monitoring or financial data analysis. However, the present paper mainly focuses on the relevancy of MKL for time series inputs selection in a methodological sense. Therefore, only three simulated and benchmark series are considered: AutoRegressive (AR, Figure 1a), Henon map (Figure 1b) and Lorenz system (Figure 1c). These series cover both linear and chaotic/nonlinear time series. The three series are generated as follows:

AR process	Henon map	Lorenz system
$x_{t+1} = 0.33x_t + 0.33x_{t-4}$	$x_{t+1} = 1 - 1.4x_t^2 + y_t$	$\dot{x}_t = 10(y_t - x_t)$
$+0.33x_{t-8} + S_{t+1}$	$y_{t+1} = 0.3x_t$	$\dot{y}_t = 28x_t - y_t - x_t z_t$
with $S_{t+1} \sim N(0, 1)$		$\dot{z}_t = x_t y_t - \frac{8}{3} z_t$

Lorenz system was numerically integrated using the Runge-Kutta 4 method and with a step size of 0.01. The x-component is used to reconstruct the dynamics of both Henon and Lorenz. For the three series, 10000 data points were extracted. Training, validation (selection of C, σ and ϵ hyper-parameters by extensive grid search) and testing (generalization performance) sets were chosen randomly. Training and validation subsets used in the experiments range from 10 to 200 data points. Testing set is composed of 1000 measurements. Experiments with added white Gaussian noise of 1%, 5% and 10% are also considered. Dependency of the results with respect to sampling is assessed using 5 different training-validation splits. In all the experiments linear and nonlinear versions of SVR and MKL were used: SVR and MKL with linear kernel (SVR_{lin} and MKL_{lin}), SVR and MKL with RBF kernel (SVR_{rbf} and MKL_{rbf}).

3.1 Linear time series

Figures 2a-b show the final weights of the MKL_{lin} model when applied to the AR process with varying noise levels (Fig. 2a) and varying training set sizes (Fig. 2b). The selection of model inputs is correct (x_t, x_{t-4}, x_{t-8}) and stabilizes from 50 training points on. Moreover, it shows robustness against noise. More importantly, not only the selection is correct but also the relative weight values follow the AR coefficients (x_t, x_{t-4} and x_{t-8} contribute with equal weights). Regarding numerical results, the performances of MKL_{rbf} are similar to the

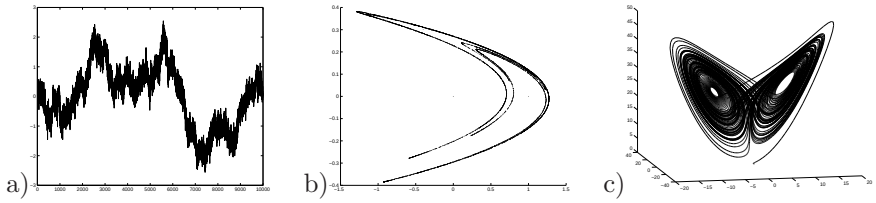

Fig. 1: a) AR series; b) Henon and c) Lorenz series in the phase space

ones observed for MKL_{lin}. Such behavior was expected, since there is no intrinsic nonlinearity in the AR series. Actually linear/nonlinear kernel is applied as a control of the problem's nonlinearity. Experiments on AR series with linear and cubic trends gave similar results.

3.2 Nonlinear time series

In this section MKL is applied to select the inputs from sufficiently large state vectors built with different time delay embeddings. Figure 2c shows the weights of the MKL_{rbf} model for the Henon time series. The inputs x_t and x_{t-1} are selected with weights equal to 0.9 and 0.1 respectively. Notice that x_{t-1} gains importance with respect to x_t as the training set size increases. MKL gives some hints about the relevance of the inputs and about the embedding dimension which in this case is equal to 2. Figure 2d shows testing performances of the 4 models considered against training size. Linear models fail in describing the patterns and are clearly outperformed by nonlinear ones. On the other hand, the MKL_{rbf} error decreases faster compared to SVR_{rbf} because of its filtering abilities.

Figure 3 illustrates the results for the Lorenz time series. Finding the embedding of the Lorenz system is more challenging since the intrinsic dimension of the data is higher and τ is different from the trivial value of 1. From mutual information criterion (computed with several bin sizes) optimal τ is 17 (Figure 3a). A similarity between the MI curve and the weights of MKL_{rbf} (with τ set to 5 to highlight its feature selection skills) is visible in Figure 3b where a local minimum on the weights occurs at $\tau \cong 15$.

Figures 3c, 3d and 3e show testing performances for SVR_{rbf} and MKL_{rbf} with increasing dimension m of the state vector. Dimensions are iteratively added one by one from x_t to $x_{t-m\tau}$. SVR performance progressively decreases as irrelevant inputs are added to the model, while MKL is robust to their introduction. However, for $\tau = 17$ and especially for $m = 3$, SVR testing error is lower than MKL. A possible interpretation is that the kernel combination proposed in this paper neglects dependency between inputs. Such dependencies can be considered using *cross-kernels* dedicated to all possible binary, ternary, etc combinations of inputs. The number of kernels and corresponding weights are inevitably increased making the optimization problem intractable. A natural solution would be to apply MKL as an exploratory tool for selecting relevant

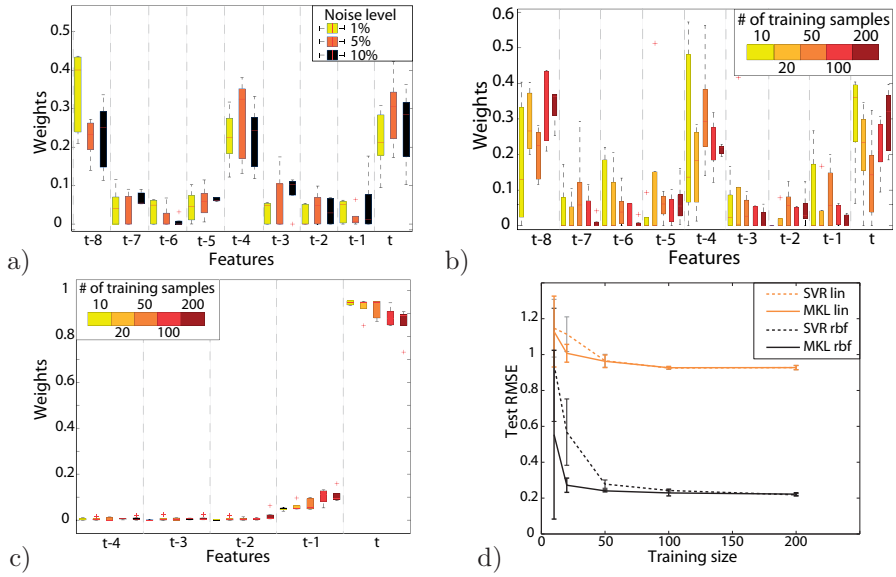

Fig. 2: Top: AR series; MKL_{lin} weights with a) varying noise levels (training size = 100) and b) varying training size (noise level = 1%). Bottom: Henon map; c) MKL_{rbf} weights with 1% of noise; d) Testing RMSE of the 4 models considered with varying training size and with 10% of noise

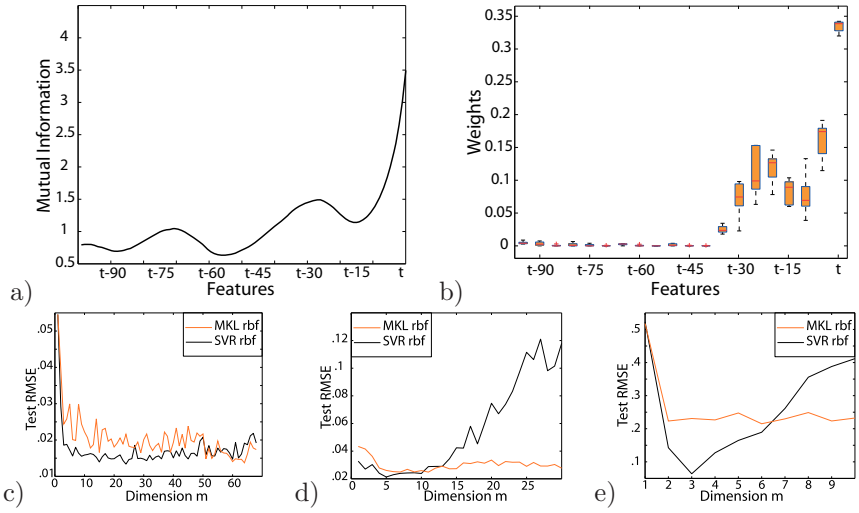

Fig. 3: Top: Average mutual information (a) and MKL_{RBF} weights (b) with $\tau = 5$ for the Lorenz time series. Bottom: Testing RMSE of SVR_{rbf} and MKL_{rbf} with increasing space dimensionality; c) $\tau = 1$, d) $\tau = 5$ e) $\tau = 17$ (from mutual information)

inputs and to use the common SVR as final predictive model.

4 Conclusions

This paper studied the efficient use of MKL for the input selection in linear and nonlinear time series. More precisely, the *SimpleMKL* algorithm was used as a data exploratory tool for selecting relevant inputs from a large input space. The flexibility of the approach allows separating the input space and provides more interpretable models. Future developments will include the study of cross-kernels accounting for input dependencies and strategies to learn them automatically. Uncertainty about MKL predictions may be addressed by applying stochastic optimization of kernel parameters (σ's for RBF kernel). Analysis of real univariate and multivariate time series is in progress.

4.1 Acknowledgments

The research is supported by the Swiss National Science Foundation projects No 200020-121835/1, PBLAP2-127713/1 and 20021-126505.

References

[1] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis (2^{nd} edition)*. Cambridge University Press, 2004.

[2] S. Geoffroy and M. Verleysen. High-dimensional delay selection for regression models with mutual information and distance-to-diagonal criteria. *Neurocomputing*, 70:1265–1275, 2007.

[3] K.R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *Proc. of the 7^{th} International Conference on Artificial Neural Networks*, pages 999–1004, 1997.

[4] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Proc. of the IEEE Workshop on Neural Networks for Signal Processing VII*, 1997.

[5] E. Eirola, E. Liitiäinen, A. Lendasse, F. Corona, and M. Verleysen. Using the delta test for variable selection. In *Proc. of the 16^{th} European Symposium on Artificial Neural Networks*, pages 25–30, 2008.

[6] A.J. Jones. New tools in non-linear modelling and prediction. *Computational Management Science*, 1(2):109–149, 2004.

[7] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[8] A. Rakotomamonjy, F.R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Reserarch*, 9:2491–2521, 2008.

[9] L. Foresti, D. Tuia, A. Pozdnoukhov, and M. Kanevski. Multiple kernel learning of environmental data. case study: analysis and mapping of wind fields. In *Proc. of the 19^{th} International Conference on Artificial Neural Networks*, volume 2, pages 933–943, 2009.

[10] G.R.G. Lanckriet, T. De Bie, N. Cristianini, M.I. Jordan, and W.S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[11] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.