KERNEL-BASED AUTOREGRESSIVE MODELING WITH A PRE-IMAGE TECHNIQUE

Maya Kallas [1,2], Paul Honeine [1], Cédric Richard [3], Clovis Francis [2] and Hassan Amoud [4]

[1] Institut Charles Delaunay (UMR CNRS 6279), LM2S, Université de Technologie de Troyes, France
[2] Laboratoire d'analyse des systèmes (LASYS), Université Libanaise, Lebanon
[3] Laboratoire Fizeau (UMR CNRS 6525), Université de Nice Sophia-Antipolis, France
[4] Azm Center for Research in Biotechnology and its Applications, Lebanese University, Lebanon

ABSTRACT

Autoregressive (AR) modeling is a very popular method for time series analysis. Being linear by nature, it obviously fails to adequately describe nonlinear systems. In this paper, we propose a kernel-based AR modeling, by combining two main concepts in kernel machines. One the one hand, we map samples to some nonlinear feature space, where an AR model is considered. We show that the model parameters can be determined without the need to exhibit the nonlinear map, by computing inner products thanks to the kernel trick. On the other hand, we propose a prediction scheme, where the prediction in the feature space is mapped back into the input space, the original samples space. For this purpose, a pre-image technique is derived to predict the future back in the input space. The efficiency of the proposed method is illustrated on real-life time-series, by comparing it to other linear and nonlinear time series prediction techniques.

Index Terms— pre-image, kernel machine, autoregressive modeling, pattern recognition, prediction

1. INTRODUCTION

Many – if not most – real-life systems are nonlinear by nature. While linear concepts can be easily tackled using simple linear algebra, they fail to adequately explain nonlinear behavior. This is the case of the autoregressive (AR) modeling for time series analysis, where each sample is given by a linear combination of a small number of previous samples. Under the assumption of a (linear) AR process, it is easy to estimate the model parameters, *i.e.*, the weights in the linear expansion, and thus predict future observations from previous ones.

One way to derive nonlinear techniques based on linear ones, is to transform the data with some nonlinear map, and apply the linear algorithm on the transformed data. This is the essence of the kernel-based machines, contributing to the proliferation of nonlinear techniques since Vapnik's Support Vector Machines (SVM) [1]. The key idea, known as the kernel trick, lies in writing a classical linear algorithm in terms of inner products of the transformed data, only to evaluate them using a (positive semi-definite) kernel, without any explicit knowledge of the mapping function. By substituting the inner product with a kernel, the data are implicitly mapped into some high dimension (even infinite-dimension for some kernels) feature space, with essentially no further computational cost. Many nonlinear techniques have been derived on this concept, such as the kernel principal component analysis, kernel Fisher discriminant analysis, and SVM novelty detection, only to name a few.

Mapping the data to the feature space is of great importance to derive nonlinear techniques based on linear ones. Nonetheless, mapping back from the feature space to the input space is also of primary interest. This is mainly because one often needs to interpret the results in the input space, *i.e.*, the signal space in signal processing. Unfortunately, it turns out that the inverse map generally does not exist and only a few elements in the feature space have a valid pre-image in the input space. This is the pre-image problem, as one seeks an approximate solution by identifying data in the input space from its counterpart in the high-dimensional feature space. Many techniques have been proposed in the literature, with a fixed-point iterative method [2], a method based on the multidimensional-scaling approach [3], or a more direct method based on the relationship between inner-products in both spaces [4] (for a recent review, see [5]).

The linear AR model is one of the most successful, flexible, and easy to use models for the analysis of time series. In this paper, we propose to extend theses advantages to the characterization of nonlinear time series. A natural extension of the linear AR modeling to nonlinear models is derived, in the light of machine learning. To this end, samples are mapped into a nonlinear feature space where, by minimizing the prediction error, the parameters of the nonlinear model are easily estimated using only kernel, without the need to exhibit the nonlinear map. Once the model parameters determined, one can apply a prediction scheme to forecast the future. However, the prediction stage still operates in the feature space. In order to get back to the input space, *i.e.,* the space of samples, we derive an appropriate pre-image technique.

This work was partly supported by the Franco-Lebanese CEDRE program.

A few attempts have been made to tackle the nonlinear AR model in the light of machine learning literature. A related work by Kumar *et al.* [6] proposes an AR model in the feature space, however, without any ability to predict. Nonlinear modeling and prediction still have not taken full advantage of recent progress in machine learning, although many efforts have been focused to develop nonlinear time series techniques, such as support vector regression [1], kernel-based Kalman filter [7], and online prediction with kernels [8].

The rest of the paper is organized as follows: In the next section, we carry out a brief description of the AR model. In Section 3, we describe the kernel-based AR model, with both parameter estimation and prediction scheme. Section 4 illustrates the efficiency of the proposed method on several time series.

2. LINEAR AUTOREGRESSIVE MODEL

The autoregressive (AR) modeling is a well-known prediction method that has been applied successfully in numerous fields. It is defined by a linear prediction formula where each sample in a time series can be predicted from previous samples. Under the assumption of an AR process of order p, a discrete time series $x_1, x_2, \ldots, x_n$ is defined by the model

$$x_i = \sum_{j=1}^{p} \alpha_{p-j+1}\, x_{i-j}, \tag{1}$$

up to some additive white noise, where the constants $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the model parameters. In other words, each sample is expressed by a linear combination of the p previous samples. To estimate the model parameters from n available samples of the time series, one often minimizes the mean square prediction error, given as

$$\sum_{i=p+1}^{n} \left(x_i - \sum_{j=1}^{p} \alpha_{p-j+1}\, x_{i-j} \right)^2 .$$

By setting to zero the derivatives of the above cost function with respect to each α_{p-j+1}, for $j = 1, 2, \ldots, p$, we get the optimal parameters[1]. Once the model parameters determined, one can predict a future value from the previous samples by using the model (1).

3. KERNEL-BASED AUTOREGRESSIVE MODELING WITH A PRE-IMAGE TECHNIQUE

The kernel-based AR modeling proposed in this paper combines, on the one hand an AR model in the feature space with parameters determined thanks to the kernel trick, and on the other hand, a pre-image technique to predict the future back into the input space. This is illustrated in Figure 1.

[1]Many methods have been proposed to determine the parameters of the AR model, including the use of the Yule-Walker equations and the forward-backward scheme. Such studies are beyond the scope of this paper.

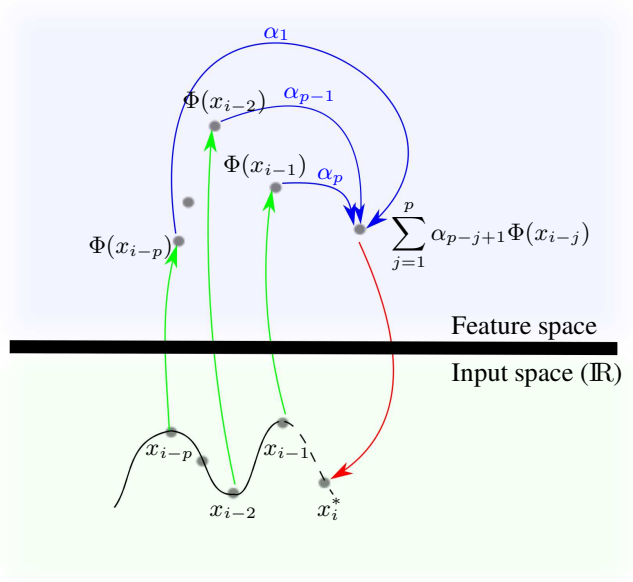

Fig. 1. Illustration of the kernel-based AR modeling. Samples are mapped from the input space into the feature space ($\rightarrow$), where we identify the AR process ($\rightarrow$). To predict back into the input space, a pre-image problem is solved ($\rightarrow$).

3.1. Autoregressive model in feature space

Let $\Phi(\cdot)$ be nonlinear function, mapping any sample x_i to an element $\Phi(x_i)$ of some feature space. For a time series with samples $x_1\ x_2\ \cdots\ x_n$, the corresponding elements in the feature space are $\Phi(x_1)\ \Phi(x_2)\ \cdots\ \Phi(x_n)$. We suppose that these elements satisfy an AR process in the feature space, namely

$$\Phi(x_i) = \sum_{j=1}^{p} \alpha_{p-j+1} \Phi(x_{i-j}), \tag{2}$$

where the predicted future is in the feature space. In matrix form, this AR model can be written as

$$\Phi(x_i) = \boldsymbol{\varphi}_i\, \boldsymbol{\alpha}$$

where $\boldsymbol{\varphi}_i$ contains the p previous samples of $\Phi(x_i)$, $\boldsymbol{\varphi}_i = [\Phi(x_{i-1})\ \Phi(x_{i-2})\ \cdots\ \Phi(x_{i-p})]$ and $\boldsymbol{\alpha} = [\alpha_p\ \alpha_{p-1}\ \cdots\ \alpha_1]^\top$ the corresponding vector of parameters.

To estimate the parameters $\boldsymbol{\alpha}$, we minimize the mean square prediction error in the feature space, between the estimated value $\sum_{j=1}^{p} \alpha_{p-j+1}\Phi(x_{i-j})$ and the real one mapped to $\Phi(x_i)$. For a sequence of n available samples, we minimize with respect to $\boldsymbol{\alpha}$ the cost function

$$\xi(\boldsymbol{\alpha}) = \sum_{i=p+1}^{n} \left\| \Phi(x_i) - \sum_{j=1}^{p} \alpha_{p-j+1}\Phi(x_{i-j}) \right\|^2 \tag{3}$$

where $\|\Phi(x_i)\|^2 = \Phi(x_i)^\top \Phi(x_i) = \kappa(x_i, x_i)$. This leads to the following expression

$$\xi(\boldsymbol{\alpha}) = \sum_{i=p+1}^{n} \big(\boldsymbol{\alpha}^\top \boldsymbol{\varphi}_i^\top \boldsymbol{\varphi}_i \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \boldsymbol{\varphi}_i^\top \Phi(x_i) + \Phi(x_i)^\top \Phi(x_i)\big).$$

By taking the derivative of this expression with respect to $\boldsymbol{\alpha}$,

$$\nabla_{\boldsymbol{\alpha}} \xi(\boldsymbol{\alpha}) = 2 \sum_{i=p+1}^{n} \big(\boldsymbol{\alpha}^\top \boldsymbol{\varphi}_i^\top \boldsymbol{\varphi}_i - \boldsymbol{\varphi}_i^\top \Phi(x_i)\big),$$

and setting it to zero, we get the optimal parameters $\boldsymbol{\alpha}$ of the nonlinear AR model with

$$\boldsymbol{\alpha} = \Big(\sum_{i=p+1}^{n} \boldsymbol{\varphi}_i^\top \boldsymbol{\varphi}_i \Big)^{-1} \sum_{i=p+1}^{n} \boldsymbol{\varphi}_i^\top \Phi(x_i)$$

In matrix form, we define the p-by-p matrix $\boldsymbol{K}$ by taking all the inner products between the p previous elements, $\boldsymbol{\varphi}_i^\top \boldsymbol{\varphi}_i$, with

$$\boldsymbol{K} = \sum_{i=p+1}^{n} \boldsymbol{\varphi}_i^\top \boldsymbol{\varphi}_i,$$

and the p-by-1 vector $\boldsymbol{k}$ corresponding to

$$\boldsymbol{k} = \sum_{i=p+1}^{n} \boldsymbol{\varphi}_i^\top \Phi(x_i).$$

This leads to a more compact form for estimating the parameters, as follows

$$\boldsymbol{\alpha} = \boldsymbol{K}^{-1} \boldsymbol{k}. \tag{4}$$

The parameter vector $\boldsymbol{\alpha}$ can be estimated using only inner products between pairs of elements in the feature space, defined by the kernel function $\kappa(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$. It is clear that $\boldsymbol{K}$ is the p-by-p matrix whose (j,k)-th entry is $\sum_{i=p+1}^{n} \kappa(x_{i-j}, x_{i-k})$, and $\boldsymbol{k}$ is the p-by-1 column vector whose j-th entry is $\sum_{i=p+1}^{n} \kappa(x_{i-j}, x_i)$. We can therefore consider any off-the-shelf (positive semi-definite) kernel to provide a nonlinear AR model. From the learning machines literature, the most used kernel function is the Gaussian radial basis function, of the form $\kappa(x_i, x_j) = \exp(\frac{-1}{2\sigma^2}\|x_i - x_j\|^2)$ where σ is the bandwidth of the kernel.

3.2. Prediction scheme using a pre-image technique

Once, we have estimated the model parameters $\boldsymbol{\alpha}$, the prediction stage consists of predicting a new value x_i, for $i > n$, from the p previous samples, $x_{i-1}, x_{i-2}, \ldots, x_{i-p}$. By applying the nonlinear AR model (2), we obtain

$$\psi_i = \sum_{j=1}^{p} \alpha_{p-j+1} \Phi(x_{i-j}), \tag{5}$$

where the prediction ψ_i lies in the feature space. However, we are more interested in the predicted sample in the original input space. Thus, we need to map back ψ_i from the feature space to the input space.

In general, the exact pre-image may not exist, and even if it exists, it may not be unique. This is referred to as the *pre-image problem*, where one identifies the best x^* in the input space whose image $\Phi(x^*)$ is as close as possible to ψ_i. This optimization problem consists of minimizing the distance between elements in the feature space, namely

$$x_i^* = \arg\min_x \frac{1}{2}\|\psi_i - \Phi(x)\|^2.$$

Many methods have been introduced in literature to solve this nonlinear optimization problem. We propose to use the iterative fixed-point method, in the same spirit as in [9]. By injecting the model (5) into the above expression, we obtain the following optimization problem for any x_i:

$$x_i^* = \arg\min_x \frac{1}{2}\Big\| \sum_{j=1}^{p} \alpha_{p-j+1}\, \Phi(x_{i-j}) - \Phi(x) \Big\|^2$$

This optimization problem can be written as

$$x_i^* = \arg\min_x J_i(x)$$

where $J_i(x)$ is the cost function defined by

$$J_i(x) = -\sum_{j=1}^{p} \alpha_{p-j+1}\, \kappa(x_{i-j}, x) + \frac{1}{2}\kappa(x, x).$$

In this expression, the term independent of x, *i.e.*, $\frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{p} \alpha_{p-k+1}\alpha_{p-j+1}\kappa(x_{i-k}, x_{i-j})$, has been removed.

To solve this problem, one may study the gradient of the cost function $J_i(x)$ with respect to x. At the optimum, the gradient with respect to x disappears, namely $\nabla_x J_i(x) = 0$. The resulting gradient is given as

$$\nabla_x J_i(x) = -\sum_{j=1}^{p} \alpha_{p-j+1} \frac{\partial \kappa(x_{i-j}, x)}{\partial x} + \frac{1}{2}\frac{\partial \kappa(x, x)}{\partial x}. \tag{6}$$

This is the general form for all kernels. This expression can be further simplified for the wide class of radial kernels, such as the Gaussian kernel. In such cases, $\kappa(x, x)$ is independent of x, thus $\partial\kappa(x,x)/\partial x$ equals to zero, and only the first term in (6) remains. Therefore, the gradient can be expressed as

$$\begin{aligned} \nabla_x J_i(x) &= -\sum_{j=1}^{p} \alpha_{p-j+1} \frac{\partial \exp(-\frac{1}{2\sigma^2}\|x_{i-j} - x\|^2)}{\partial x} \\ &= -\frac{1}{\sigma^2}\sum_{j=1}^{p} \alpha_{p-j+1}\, \kappa(x_{i-j}, x)\, (x_{i-j} - x). \end{aligned}$$

Setting this gradient to zero at the optimum x_i^*, we get the fixed-point iterative expression

$$x_i^* = \frac{\sum_{j=1}^{p} \alpha_{p-j+1}\, \kappa(x_{i-j}, x_i^*)\, x_{i-j}}{\sum_{j=1}^{p} \alpha_{p-j+1}\, \kappa(x_{i-j}, x_i^*)}.$$

This result can be interpreted as an AR model, in the same spirit as (1), although the parameters are no longer *constants*, since we have the form $x_i^* = \sum_{j=1}^{p} \beta_{p-j+1} x_{i-j}$, with

$$\beta_{p-j+1} = \Big(\sum_{i=1}^{p} \alpha_{p-i+1}\, \kappa(x_{i-i}, x_i^*) \Big)^{-1} \alpha_{p-j+1}\, \kappa(x_{i-j}, x_i^*).$$

4. EXPERIMENT

In this section, we show the relevance of the proposed method on two real-life time series[2]: the Mackey-Glass MG_{30} time series, modeling the blood cells production evolution with

$$\frac{dx(t)}{dt} = -0.1\, x(t) + \frac{0.2\, x(t-\tau)}{1 + x(t-\tau)^{10}}$$

with $\tau = 30$, and the *Laser* time series from the Santa Fe competition (dataset A). For each time series, the first 300 samples were considered to determine the model parameters, as well as estimating the best value for the order. The mean square prediction error was estimated on the next 300 samples.

To give a well-defined benchmark for comparison, we compare the proposed method to several time series prediction techniques: the linear AR model, the multilayer perceptron with a tanh activation function, the support vector regression [1] and the nonlinear Kalman filter [7]. For the kernel machines, the Gaussian kernel has been used, with the proper bandwidth value estimated within the estimation stage. For our approach, the bandwidth is set to $\sigma = 0.3$ and $\sigma = 0.015$ for the *Laser* and the MG_{30} time series, respectively. The value of order was set to $p = 6$ for both sequences. Table 1 shows the mean square prediction error, where the values of the methods multilayer perceptron, support vector regression and nonlinear Kalman filter are borrowed from [7].

5. CONCLUSION

In this paper, we proposed a kernel-based AR modeling for time series analysis and prediction, inspired by recent advances in machine learning. To this end, we bind, on the one hand an AR model in a feature space with model parameters estimated thanks to the kernel trick, and on the other hand, a prediction scheme back into the input space using a pre-image technique. Experiments on a benchmark of real-life time series confirm the relevance of the proposed method.

This strategy in combining kernel machines and AR models opens the way to the development of a range of diverse nonlinear time series techniques. As future work, we are exploring many possibilities in this direction, such as the use of the Yule-Walker equations, the extension to a vector AR model, as well as a nonlinear ARMA model. The choice of the optimal order remains an open question.

[2]The MG_{30} and the *Laser* time series are available from http://www.bme.ogi.edu/∼ericwan/data.html.

Table 1. The mean square prediction error (MSE) for several time series techniques.

	Laser	MG_{30}
multilayer perceptron	1.4326	0.0461
support vector regression	0.2595	0.0313
nonlinear Kalman filter	0.2325	0.0307
linear AR modeling	16.6956	0.0158
kernel-based AR modeling	0.0702	0.00084

6. REFERENCES

[1] V. N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, September 1998.

[2] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, Cambridge, MA, USA, 1999, pp. 536–542, MIT Press.

[3] J. T. Kwok and I. W. Tsang, "The pre-image problem in kernel methods," in *ICML*, 2003, pp. 408–415.

[4] P. Honeine and C. Richard, "Solving the pre-image problem in kernel machines: a direct method," in *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, September 2009.

[5] P. Honeine and C. Richard, "Pre-image problem in kernel-based machine learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 77–88, March 2011.

[6] R. Kumar and C. V. Jawahar, "Kernel approach to autoregressive modeling," in *The 13th National Conference on Communications (NCC)*, Kanpur, India, January 2007.

[7] L. Ralaivola and F. D'alche-Buc, "Time series filtering, smoothing and learning using the kernel kalman filter," in *Proc. IEEE International Joint Conference on Neural Networks*, 2005, vol. 3, pp. 1449–1454.

[8] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, 2009.

[9] M. Kallas, P. Honeine, C. Richard, H. Amoud, and C. Francis, "Nonlinear feature extraction using kernel principal component analysis with non-negative preimage," in *Proc. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Buenos Aires, Argentina, 2010.