 Supervised Learning with Kernel methods

TAOUALI OKBA, ELAISSI ILYES, GARNA TAREK AND MESSAOUD HASSANI
Unité de Recherche ATSI
Ecole Nationale d’Ingénieurs de Monastir
Rue Ibn El Jazzar 5019 Monastir ; Tel : +(216) 73 500511, Fax : +(216) 73 500 514,
Tunisia
taoualiokba@yahoo.fr, ilyes.elaissi@yahoo.fr, tarek.garna@enim.rnu.tn,
hassani.messaoud@enim.rnu.tn

Abstract: -This paper proposes a comparative study of three identification kernel methods of nonlinear systems modelled in Reproducing Kernel Hilbert Space (RKHS), where the model output results from a linear combination of kernel functions. The coefficients of this combination are the model parameters, the number of which equals the number of observations used in learning phase. Theses methods are support vector machines (SVM), regularization networks (RN) and kernel Principal Component Analysis (KPCA). The performances of each method in terms of generalization ability and computing time were evaluated on numerical simulations.

Key-Words: - Identification, RKHS, SLT, SVM, RN,

1 Introduction
The last few years has registered the birth of a new modelling technique of nonlinear systems developed on a particular Hilbert Space, known as Reproducing Kernel Hilbert Space (RKHS) which uses the statistical learning theory (SLT) to provide an RKHS model as a linear combination of the kernels forming this space [4], [6], [7], [12], [13] and [14].

The developed models are an attractive alternative to other modeling techniques based on Volterra series, neural networks, … . Indeed, the solution of optimization problem in space RKHS is a global minimum contrary to that provided by neural networks. The solution is obtained by solving a quadratic optimization problem by using the learning algorithms such as support vector machines (SVM) [1], regularization networks (RN) [2] and Kernel Analysis Principal Component KPCA [9]. These algorithms known as kernel methods construct RKHS models on the principle of structural risk minimization (SRM). The number of parameters of these models depends only on the number of observations and not on the structure model as in the conventional modeling approaches.

The paper is organized as follows. In section 2 we remind the presentation of the RKHS space. Section 3 is devoted to the modelling in RKHS. The SVM, RN and KPCA methods are presented in the section 4 and then tested to identify a benchmark in the fifth section [11].

2 Reproducing Kernel Hilbert Space
Let $E \subset \mathbb{R}^d$ an input space and $L^2(E)$ the Hilbert space of square integrable functions defined on $E$. Let $k : E \times E \to \mathbb{R}$ be a continuous positive definite kernel. It is proved [15] that it exists a sequence of an orthonormal eigen functions $(\psi_1, \psi_2, ..., \psi_l)$ in $L^2(E)$ (where $l$ can be infinite) and a sequence of corresponding real positive eigenvalues $(\sigma_1, \sigma_2, ..., \sigma_l)$ so that the kernel $k$ is defined as:

$$k(x, t) = \sum_{j=1}^{l} \sigma_j \psi_j(x) \psi_j(t) ; \quad x, t \in E$$ (1)

Let $F_k \subset L^2(E)$ be a Hilbert space associated to the kernel $k$ and defined by:

$$F_k = \left\{ f \in L^2(E) \mid f = \sum_{i=1}^{l} w_i \phi_i, \quad \text{and} \quad \sum_{j=1}^{l} \frac{w_j^2}{\sigma_j} < +\infty \right\}$$ (2)

Where $\phi_i = \sqrt{\sigma_i} \psi_i ; \quad i = 1, ..., l$. The scalar product in the space $F_k$ is given by:

$$< f, g >_{F_k} = \sum_{i=1}^{l} w_i \phi_i \cdot \sum_{j=1}^{l} z_j \phi_j >_{E} = \sum_{i=1}^{l} w_i z_i$$ (3)

The kernel $k$ is said to be a reproducing kernel of the Hilbert space $F_k$ if and only if the following conditions are satisfied.
\[ \forall x \in E, \ k(x,.) \in F_k \]
\[ \forall x \in E \text{ and } \forall f \in F_k, \ k(x,.) \leq \rho_k = f(x) \]  
\[ (4) \]

Where \( k(x,.) \) means \( k(x, x') \) \( \forall x, x' \in E \)

\( F_k \) is called reproducing kernel Hilbert space (RKHS) with kernel \( k \) and dimension \( l \). Moreover, for any RKHS, there exists only one positive definite kernel and vice versa [3].

Among the possible reproducing kernels, we mention the Radial Basis function (RBF) defined as:

\[ k(x,t) = \exp\left(- \frac{||x-t||^2}{2\mu^2}\right) ; \forall x, t \in E \]  
\[ (5) \]

with \( \mu \) a fixed parameter.

Let’s define the application \( \Phi \) :

\[ \Phi : E \rightarrow \mathbb{R}^l \]

\[ x \mapsto \Phi(x) = \begin{pmatrix} \phi_1(x) \\ . \\ . \\ \phi_l(x) \end{pmatrix} \]  
\[ (6) \]

Where \( \phi_i \) are given in (2). The kernel trick [8] is so that:

\[ k(x,x') = \langle \Phi(x), \Phi(x') \rangle \]  
\[ (7) \]

3 Modeling in RKHS

The use of RKHS models in identifying systems giving only the observations of their behaviour is of a great interest. Consider a set of observations \( D = \{x^{(i)}, y^{(i)}\}_i \)

known as learning set, with \( x^{(i)} \in \mathbb{R}^n, y^{(i)} \in \mathbb{R} \) respectively the system input and output. According to the statistical learning theory (SLT) [1], the identification problem in the RKHS \( F_k \) can be formulated as a minimization of the regularized empirical risk which consists in finding the function \( f^* \in F_k \) such that:

\[ f^* = \sum_{j=1}^{M} w^*_j \phi_j \]

\[ = \arg \min_{f \in F_k} \frac{1}{M} \sum_{i=1}^{M} V\left( y^{(i)}, f\left(x^{(i)}\right) \right)^2 + \lambda \|f\|_2 \]  
\[ (8) \]

Where \( \lambda \) is a regularization parameter chosen in order to ensure a generalization ability to the solution \( f^* \) and \( V(.) \) is a cost function. According to the representer theorem [5], the solution of the problem (8) is a combination of \( M \) kernel functions applied on the \( M \) observations.

\[ f^*(x) = \sum_{i=1}^{M} a_i k(x^{(i)}, x) \]  
\[ (9) \]

The number of parameters is equal to the observation number used in the learning phase and contained in the learning set.

4 Kernel methods

4.1 Support Vector machine (SVM)

We now discuss Support Vector Machines (SVM) [1], [2] in its regression form that corresponds to the minimization of problem (8) and where the cost function \( V \) is:

\[ V = ||y_i - f(x_i)||_2^2 \]  
\[ (10) \]

Where \( \xi \) width of tube

The coefficients \( \{a_i\} \) in the relation (9) can be determined by solving a Quadratic Programming (QP) problem with linear constraints.

4.2 Regularization networks (RN)

In this case the cost function is:

\[ V\left( y^{(i)}, f\left(x^{(i)}\right) \right) = \left( y^{(i)} - f\left(x^{(i)}\right) \right)^2 \]  
\[ (11) \]

and the solution provides the parameters \( \{a_i\} \) which verify:

\[ a_i = \sum_{j=1}^{M} \left( K + \lambda N I_N \right)^{-1}_{i,j} y^{(i)} \]  
\[ (12) \]

In a matrix form, the relation (12) is:

\[ (K + \lambda M I_M)^{-1} Y = A \]  
\[ (13) \]

where: \( K \) is the Gram matrix that satisfy:
4.3 Kernel Principal Component Analysis (KPCA)

As the dimension \( l \) of the RKHS spaces is high (possibly infinite), the determination of the coefficients \( w_i \) of the relation (2) is problematic. Using the Principal Component Analysis (PCA) [10] method to identify the RKHS model is very interesting. The KPCA method doesn't target the input space such as in the linear case but it treats the transformed data by the application \( \Phi \) defined in (6) and satisfying (7).

The RKHS model of this system is given by the relation (9). The Gram matrix \( K \) associated to the kernel \( k \) is an \( M \)-dimensional square matrix, so that:

\[
K_{i,j} = k\left(x^{(i)}, x^{(j)}\right)
\]

We assume that the transformed data \( \{\Phi(x^{(i)})\}_{i=1,...,M} \in \mathbb{R}^l \) are centred [8], [9]. The approximation covariance matrix \( C_{\phi} \) of the transformed data is symmetrical, \( l \)-dimensional and it is written as following:

\[
C_{\phi} = \frac{1}{M} \sum_{i=1}^{M} \Phi(x^{(i)}) \Phi(x^{(i)})^T, \quad C_{\phi} \in \mathbb{R}^{l \times l}
\]

Let \( l' \) the number of the eigenvectors \( \{\theta_j\}_{j=1}^{l'} \) of \( C_{\phi} \) corresponding to the non zeros positive eigenvalues \( \{\lambda_j\}_{j=1}^{l'} \). It is proved in [9] that the number \( l' \) is less or equal to \( M \). Due to the large size \( l \) of \( C_{\phi} \), the calculus of \( \{\theta_j\}_{j=1}^{l'} \) can be difficult. The KPCA method shows that these \( \{\theta_j\}_{j=1}^{l'} \) are related to the eigenvectors \( \{\beta_j\}_{j=1}^{l'} \) of the gram matrix \( K \) according to [9]:

\[
\theta_j = \sum_{i=1}^{M} \beta_{ji} \Phi(x^{(i)}), \quad j=1,...,l'
\]

where \( \{\beta_{ji}\}_{j=1,...,l'} \) are the components of \( \{\beta_j\}_{j=1}^{l'} \) associated to their nonzero eigenvalues \( \mu_i > ... > \mu_j \) of the Gram matrix \( K \).

The principle of the KPCA method consists in organizing the eigenvectors \( \{\beta_j\}_{j=1}^{l'} \) in the decreasing order of their corresponding eigenvalues \( \{\mu_j\}_{j=1}^{l'} \). The principal components are the \( p \) first vectors \( \{\theta_j\}_{j=1}^{p} \) associated to the highest eigenvalues and are often sufficient to describe the structure of the data [8], [4]. The number \( p \) satisfies the Inertia Percentage criterion IPC given by:

\[
p^* = \arg\left(\text{IPC} \geq 99\right)
\]

where

\[
\text{IPC} = \frac{\sum_{j=1}^{p} \mu_j}{\sum_{j=1}^{l} \mu_j} \times 100
\]

The principal components \( \theta_j, j=1,...,p \) are unit vectors if the corresponding vectors \( \beta_j \) satisfy:

\[
\langle \beta_j, \beta_i \rangle = \frac{1}{\mu_j}, \quad j=1,...,l'
\]

Let \( F_{\text{kprca}} \) the space spanned by the \( p \) principal components \( \{\theta_j\}_{j=1}^{p} \). The projection \( \Phi(x) \in \mathbb{R}^l \) of \( \Phi(x) \in \mathbb{R}^l \) on \( F_{\text{kprca}} \) is given by:

\[
\Phi(x)_j = \langle \theta_j, \Phi(x) \rangle, \quad j=1,...,p
\]

and from (17), we have:

\[
\Phi(x)_j = \sum_{i=1}^{M} \beta_{ji} \Phi(x^{(i)}), \quad \Phi(x)_j = \sum_{i=1}^{M} \beta_{ji} \Phi(x^{(i)})\Phi(x)
\]

and according to (7), we have:

\[
\Phi(x)_j = \sum_{i=1}^{M} \beta_{ji} k\left(x^{(i)}, x\right)
\]

Similarly to the linear PCA [10], the model proposed by the Kernel PCA [4] is given by:

\[
\hat{y}_{\text{kprca}} = \sum_{i=1}^{M} w_i \Phi(x)_i
\]
to mention that the number of the parameters \( \{w_j\}_{j=1,...,p} \)
of the model (24) is less than \( l \) given by the relation (8). To identify the model parameter \( w_j \) we solve the following problem:

\[
\min_{w \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^{M} \left( y^{(i)} - w^T \Phi(x^{(i)}) \right)^2 ; \quad w = (w_1, ..., w_p)^T
\]  

For any new input data \( x^{(\text{new})} \in E \), the corresponding KPCA model output is:

\[
y^{(\text{new})} = \sum_{j=1}^{p} w_j \Phi(x^{(\text{new})})
\]  

According to (23), the relation (26) is given by:

\[
y^{(\text{new})} = \sum_{j=1}^{p} w_j \beta_{j,i} k(x^{(i)}, x^{(\text{new})})
\]  

Thus:

\[
y^{(\text{new})} = \sum_{i=1}^{M} \sum_{j=1}^{p} w_j \beta_{j,i} k(x^{(i)}, x^{(\text{new})})
\]  

Let \( a_i = \sum_{j=1}^{p} w_j \beta_{j,i} \), relation (28) yields the model (9).

\[
y^{(\text{new})} = \sum_{i=1}^{M} a_i k(x^{(i)}, x^{(\text{new})})
\]  

5 Comparative study

To illustrate the efficiency of all methods, we proceed to their validation on nonlinear dynamic system used as a benchmark [11] and given by:

\[
y(i) = \left(0.8 - 0.5 \exp\left(-y(i-1)^2\right)\right) y(i-1) \\
- \left(0.3 + 0.9 \exp\left(-y(i-1)^2\right)\right) y(i-2) + 0.1 \sin(\pi y(i-1)) + e(i)
\]  

Where:

\( e(i) \) is a Gaussian noise with variance 0.4 and unit mean so that the Signal to Noise Ratio (SNR) is equal to 5. The input vector \( x \in \mathbb{R}^p \) of RKHS model is:

\[
x(k) = \begin{bmatrix} y(k-1) & y(k-2) \end{bmatrix}^T
\]  

To build the RKHS model we use the RBF kernel (Radial Basis Function) defined as:

\[
k(x, y) = \exp\left(-\frac{\|x - y\|^2}{100}\right)
\]  

We have used 80 observations in the identification phase and 150 other observations in the validation one.

In Table 1 we present the performances of the three kernel method carried out in the same condition. The first performance is the generalization ability evaluated by the Normalized Means Square Error (NMSE) in the identification phase and the validation one and the second concerns the compute time. For the simulation the clock frequency is set to 3 GHz.

From Table 1 we see that the compute time of both methods KPCA and RN are significantly less than the time needed by the SVM method (7.53 seconds). This is due to the fact that the SVM method solves an optimization problem with constraints.

However that the SVM method generalizes better than the other two methods RN and KPCA as the corresponding NMSE in learning and validation phase is significantly smaller than other techniques.

We also note that three principal components is enough to build an RKHS model from the system which leads to a less complex model.

Table I

| Performances of kernel methods SVM, RN and KPCA |
|-----------------|------|------|------|
| **SVM** | **RN** | **KPCA** |
| Compute time (seconds) | 7.53 s | 0.9204 s | 0.5460 s |
| NMSE identification | \( 8.88 \times 10^{-11} \) | \( 9.58 \times 10^{-5} \) | \( 3 \times 10^{-5} \) |
| NMSE validation | \( 2.32 \times 10^{-5} \) | \( 1.38 \times 10^{-4} \) | \( 8.710^{-3} \) |

In figure 1 we draw the RKHS model outputs using SVM, RN and KPCA methods and the output process in the identification phase. We notice the concordance between the three model outputs and that of the process output.
6 Conclusion

In this paper we have presented three kernel methods SVM, RN and KPCA to identify the nonlinear system in the RKHS space. A comparative study between the three methods has been achieved and shows the effectiveness of the SVM method in terms of generalization ability.

References:


[4] Ilyes. A. Modélisation, Identification et Commande Prédictive des systèmes non linéaires par utilisations des espaces RKHS, thèse de doctorat de l’université de Tunis, ENIT, 4-4-2009, Tunisie


