Support Vector Regression With Kernel Combination for Missing Data Reconstruction

Luca Lorenzi, Student Member, IEEE, Grégoire Mercier, Senior Member, IEEE, and Farid Melgani, Senior Member, IEEE

Abstract—Over the past few years, the reconstruction of missing data due to the presence of clouds received an important attention. Applying region-based inpainting strategies or conventional regression methods, such as support vector (SV) machine regression, may not be the optimal way. In this letter, we propose new combinations of kernel functions with which we obtain a better reconstruction. In particular, in the regression, we add to the radiometric information, i.e., the position information of the pixels in the image. For each kind of information adopted in the regression, a specific kernel is selected and adapted. Adopting this new kernel combination in a SV regression (SVR) comes out that only few SVs are needed to reconstruct a missing area. This means that we also perform a compression in the number of values needed for a good reconstruction. We illustrate the proposed approaches through some simulations on FORMOSAT-2 multitemporal images.

Index Terms—Cloud removal, image reconstruction, missing data, support vector (SV) machine, support vector regression (SVR).

I. INTRODUCTION

The presence of clouds in remote sensing optical images may produce missing data. In general, if the application does not need to study them, as in this letter, clouds represent only an unwanted noise, which distorts the spectral response of land covers. In the last years, different works have been presented to cope with this problem. Usually, the solutions are intended to detect and to remove cloud presence for low or medium spatial resolutions and may require temporal information [1]–[3]. Note that the focus of this letter is on the reconstruction of the missing areas; their detection is thus not considered here. One of the first techniques that deal with this problem produces a cloud-free image mosaic from several cloudy optical satellite images acquired from the same area [4]. Here, the goal is to compose a reasonably cloud-free composite scene by merging together different parts of images. In a similar way, in [5], authors use a regression-tree strategy and a histogram matching method to obtain a more plausible mosaic image. Other techniques to reconstruct missing areas are based on prediction techniques. A first work was presented in [6], where a least-square linear prediction with escalator structure is implemented. In this case, the algorithm predicts each missing pixel from its temporal behavior. Furthermore, unsupervised contextual prediction has been adopted, as in [7], where the local spectrotemporal relationship is used to predict the missing data through a nonlinear regressor. In particular, the author adopts a support vector (SV) machine regressor. To better exploit available information, authors of [8] make use of the spatial and spectral correlations. The literature also reports another technique that exploits Synthetic Aperture Radar information to remove the cloud presence in optical imagery [9]. In [10], an inpainting method is introduced for removing clouds by means of the bandelet transform (a special case of the wavelet transform) and a multiscale geometrical grouping. In [11], several region-based inpainting strategies are proposed to reconstruct missing regions by propagating spectral and geometrical information from the remaining parts of the image. The underlying idea is to enrich the region (patch) search process by including local image properties, by isomorphic transformations, or to reformulate it under a multisresolution processing scheme.

In this letter, we propose to improve the reconstruction process by integrating both radiometric and position information. For each kind of information, a specific kernel is selected and adapted. Subsequently, their fusion is performed by a linear combination of the two resulting kernels. SV regression (SVR) is applied to derive the prediction function [12]. Simulations conducted on temporal images acquired with the optical high-resolution FORMOSAT-2 satellite are reported and discussed.

The remainder of this letter is organized as follows. In the next section, we will formulate the reconstruction problem. In Section III, we briefly introduce the SVR theory adopting classical kernels, and we propose new kernel combinations. The efficiency of the proposed approach is illustrated in Section IV, and Section V draws the conclusions.

II. PROBLEM FORMULATION

Let us consider multitemporal $B$-band data acquired and registered over the same geographical area by an optical sensor at two different dates, i.e., $I_b^{(1)}$ and $I_b^{(2)}$, with $b \in \{1, 2, \ldots, B\}$ being the set of spectral bands. We assume that the acquisitions of the images are temporally close to each other and they are characterized by similar spatial structures. As aforementioned,
we remark that the detection step, which is useful to find the position of the clouds, is not treated in this letter. We make the hypothesis that image \( I^{(2)} \) is obscured by the presence of a cloud. This cloudy area in image \( I^{(2)} \) is viewed as a target region \( \Omega \) and the remaining part as the source region \( \Theta \) (following the classical notation in the inpainting literature). Note that image \( I^{(1)} \) is supposed to be cloud-free. In case it is not, the proposed technique could be applied in a similar way from \( I^{(2)} \) to \( I^{(1)} \). The only restriction is that the clouds need not to be at the same positions in the image set.

The aim of this letter is to generate a new image \( Y^{(2)} \) such that, for a specific band \( b \) and for each pixel in the coordinates \((k, l)\)

\[
Y^{(2)}_b(k, l) = \begin{cases} 
I^{(2)}_b, & \text{if } (k, l) \in \Theta^{(2)} \\
\hat{f}(k, l, I^{(1)}), & \text{otherwise} 
\end{cases}
\]  

(1)

where \( \hat{f}(\cdot) \) represents a prediction function that takes into account the position coordinates \((k, l)\) of a pixel and the radiometric information of \( I^{(1)} \). The fact to take the location information \((k, l)\) in addition to the radiometric information takes its inspiration from the idea of promoting the use of SVs in the same location as the pixel to be inferred. It helps in considering the same kind of landscape in the missing data reconstruction process. Fig. 1 illustrates the reconstruction process, i.e., training and prediction steps.

III. PROPOSED SOLUTION

A. \( \varepsilon \)-Insensitive SVR

In any regression problem, one seeks a function that best links the input to the output spaces. \( \varepsilon \)-SVR is a prediction approach, which takes origin from the statistical learning theory [13], [15] and has proven to be efficient in different contexts including the remote sensing field [12]. It aims at getting function \( f(x) \) that has, at most, \( \varepsilon \) deviation from the desired target \( y \) and, at the same time, is as smooth as possible. This can be obtained by performing a nonlinear projection, mapping the data from the original \( d \)-dimensional domain to a higher dimensional feature space, i.e., \( \Phi(x) \in \mathbb{R}^{d'}(d' > d) \), in order to increase the flatness of the function and, accordingly, to approximate it in a linear way as

\[
f(x) = w^* \cdot \Phi(x) + b^* .
\]  

(2)

The optimal hyperplane defined by the weight vector \( w^* \in \mathbb{R}^{d'} \) and bias \( b^* \in \mathbb{R} \) is the one that minimizes the cost function, i.e.,

\[
\Psi(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi^*_i).
\]  

(3)

and is subject to the following constraints:

\[
\begin{align*}
&y_i - (w \cdot \Phi(x_i) + b) \leq \varepsilon + \xi_i \\
&(w \cdot \Phi(x_i) + b) - y_i \leq \varepsilon + \xi^*_i \\
&\xi_i, \xi^*_i \geq 0
\end{align*}
\]  

(4)

where \( \xi_i \) and \( \xi^*_i \) are the slack variables introduced for the samples that are not in the \( \varepsilon \)-tube, depending on whether they lie above or below the tube, respectively. Constant \( C \) represents a regularization parameter with which it is possible to obtain a compromise between the flatness (model complexity) of function \( f(x) \) and the accuracy of the regression (training error). Note that the sum in (3) takes into account all the available \( N \) training samples [14].

There exists an optimization solution that can reformulate and solve the previous problem by using the Lagrange multipliers and a quadratic programming solver, leading to the following final prediction model:

\[
f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha^*_i) \cdot K(x_i, x) + b,
\]  

(5)

where \( K(\cdot, \cdot) \) is a kernel function that characterizes the dot product behavior in a feature space defined by \( \Phi(\cdot) \), and only samples with nonzero coefficients \( \alpha_i \) and \( \alpha^*_i \) are the SVs that can lie on the \( \varepsilon \)-tube and contribute to the prediction. The function kernel \( K(\cdot, \cdot) \) must satisfy the Mercer theorem [12]. In the following subsection, we will briefly describe the most common kernel functions [16].

B. Common Kernel Functions

A simple kernel function is the linear (LIN) kernel, which is defined as

\[
K_{\text{LIN}}(x_1, x_2) = x_1^T x_2 + c
\]  

(6)

where \( c \) is a constant.

Another kernel is the nonstationary polynomial (POL) kernel, i.e.,

\[
K_{\text{POL}}(x_1, x_2) = (\beta x_1^T x_2 + c)^d
\]  

(7)

where \( \beta \) is a parameter for the slope, \( c \) is a constant, and \( d \) the degree of the polynomial.

Another common and most versatile kernel is the Gaussian function, also known as the radial basis function (RBF) kernel, i.e.,

\[
K_{\text{RBF}}(x_1, x_2) = \exp \left(-\gamma \|x_1 - x_2\|^2\right)
\]  

(8)

where \( \gamma = 1/2\sigma^2 > 0 \) is a width parameter.
Generally, all kernel parameters play an important role in the performance of the kernel and must be thus carefully tuned jointly with the regularization parameter \( C \) and the precision parameter \( \varepsilon \) (e.g., by adopting a grid search via an \( n \)-fold cross validation (CV) procedure).

C. Kernel Function

In our reconstruction problem, since we deal with heterogeneous features (i.e., position and radiometric information), the proposed idea is to design a fusion kernel derived from a combination of single kernels, each for every feature typology. From [17], if a set of kernels forms a convex cone, closed and under pointwise convergence, or in other words, if \( K_1 \) and \( K_2 \) are valid kernels and \( \beta_1, \beta_2 \geq 0 \), then the following expression is also a valid kernel:

\[
\beta_1 K_1 + \beta_2 K_2.
\]

It is possible to rewrite this expression, using only one weight constant to balance the sum, i.e.,

\[
\mu K_1 + (1 - \mu) K_2.
\]

This kernel combination opens the way to design a large number of new kernels, by linearly combining the most common kernels (e.g., LIN–LIN, LIN–POL, POL–LIN, LIN–RBF, etc.). Because of space limitations, we will just consider the combination based on two RBF kernels, empirically found to be the best one, i.e.,

\[
K_{RBF-\text{RBF}} = \mu K_{RBF1} + (1 - \mu) K_{RBF2}
\]

where \( 0 \leq \mu \leq 1 \) and each kernel has its proper kernel width \( \gamma \) value, which is evaluated with a grid search CV. Note that, we also try to add all the spatial information in a dedicate kernel; in another simulation, we add a third kernel \( K_3 \) to have a specific kernel for each kind of information (spatial coordinates, radiometric, and neighborhood). In both cases, we do not achieve interesting changes in the results.

In the following, we will refer to the method that applies this kind of kernel as “Gaussian kernel combination regression” (GKCR).

D. Feature Vector

The contextual nonlinear prediction (CNP) was originally defined to adopt an SVR for the reconstruction of cloudy regions in multitemporal images, by exploiting a “local neighborhood” [7]. In our implementation, we propose to add position information [i.e., pixel coordinates \((k, l)\) in Fig. 1] in a specific kernel \( K_1 \) (see Fig. 2). The new addition of this information provides to locally restrict the sample similarity. It induces a trend on a regular sampling in the selection of the SVs. The second kernel \( K_2 \) considers radiometric information, composing the second part of the feature vector. This part can be subdivided into two subparts. The first \( B \) values come from the spectral information extracted from image \( I^{(1)} \). The last four values come from the spatial information conveyed in a spatial neighborhood system of first order and centered on the pixel of interest in image \( I^{(1)}_b \), where \( b \) represents the spectral band that is targeted.

where \( b \) represents the band that is targeted. Note that the pixel of interest comes from \( \Theta^{(1)} \) or \( \Omega^{(1)} \) depending if the training or test step is considered, respectively.

IV. Experimental Results

A. Data Set Description

To cope with this missing-data problem, at least two images are necessary. In our simulations, we will assume that just one of them can contain cloudy regions. The test images that we used come from the optical high-resolution FORMOSAT-2 satellite [18]. The images represent part of the Arcachon basin in the south region of Aquitaine in France. Images convey four spectral bands (blue, green, red, and near infrared) and are characterized by a spatial resolution of 2 m. They were acquired in two different days. Image \( I^{(1)} \) was taken on the 24th of June 2009, whereas image \( I^{(2)} \) was taken three weeks later, on the 16th of July. The two images do not contain significant differences in the spatial structures. For our experiments, the two images have been cut out into \( 400 \times 400 \) pixel-size images containing different land cover typologies, principally grass and urban areas [see the regions of interest (ROI) A and B in Figs. 4 and 6, respectively]. We artificially obscured part of the \( I^{(2)} \) image, as shown in Figs. 4(b) and 6(b). The resulting mask makes it possible to quantify the reconstruction accuracy by comparing the reconstruction result with the true pixel values. As an accuracy measure, we adopt the well-known peak signal-to-noise ratio (PSNR) measure [19] evaluated on the images for the training and test steps.

B. Experiments

Before the parameter estimation phase, we decide to fix the number of training samples (\#TR) to adopt in the regression, in correlation with the number of missing values, i.e., the test samples (\#TS). In particular, we took three different ratios \#TR/\#TS, which are 1/5, 1/3, and 3/4, where the number of test samples \#TS = 2060. Note that the training samples are collected adopting ten-time shifted sampling grids [see an example of sampling in Figs. 4(a) and in Fig. 6(a)].

Before starting with the experiments, all different kinds of data represented in the feature vector have to be normalized in interval [0, 1]. Afterward, one of the first evaluations is to find which combination of RBF kernels returns the better result, namely, which value of \( \mu \) to adopt in (11). The answer in this case comes from empirical estimations. We apply for each value of \( \mu \) a complete grid search, finding the best parameters.
TABLE 1

<table>
<thead>
<tr>
<th>ROI A</th>
<th>ROI B</th>
</tr>
</thead>
<tbody>
<tr>
<td>μ</td>
<td>κ</td>
</tr>
<tr>
<td>0.05</td>
<td>100</td>
</tr>
<tr>
<td>0.1</td>
<td>100</td>
</tr>
<tr>
<td>0.2</td>
<td>100</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
</tr>
<tr>
<td>0.4</td>
<td>2</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>0.9</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Fig. 3. Evolution of training and test PSNR values varying the μ parameter, for the two crop regions, adopting the #TR/#TS = 1/5 ratio.

(with C ∈ {10−3, 2 · 10−3, 5 · 10−3, . . . , 102}, γi ∈ {10−4, 2 · 10−4, 5 · 10−4, . . . , 1}, and ε ∈ {0.01, 0.02, . . . , 0.1}), considering the first of the three different sampling ratios (i.e., #TR/#TS = 1/5). From an analogous problem, where an RBF kernel was adopted [see the CNP in [7], similar to our kernel with μ = 0 in (11)], we start to weight more K1, increasing the μ value. In other words, we initiate favoring radiometric information over spatial coordinates, until we reach μ = 1, where the radiometric information disappears (see this evaluation in Table I and Fig. 3). Note that, if we take into account only one of the two parts of the RBF kernels, we obtain worst reconstructions. For example, if μ = 0, we have a lack of information, and when μ = 1, the estimation does not take into account the radiometric information; it considers only the localization of the pixel. It is worth noting that the result seems to be more stable between μ = 0.1 and μ = 0.9. Similar results are obtained for ROI B and for the three different ratios, where the best result still arises with μ = 0.1.

Once that μ value is found, the corresponding SVR parameters for both kernels K1 and K2 are known, i.e., the regularization parameter C, the two width parameters γi, and the precision parameter ε. A first important result that comes out in all the experimentation is that ε does not significantly change the CV results; for this reason, we decide to fix its value for the following simulations at ε = 0.05, in order to make the grid search faster. Having fixed μ and ε values, it comes out that, generally, γ1 ≪ γ2. This last means that the location kernel K1 is less sensitive than the radiometric kernel K2, but both are still important for achieving a good reconstruction.

In the following, we qualitatively present results obtained applying #TR/#TS = 1/5 ratio, which means to adopt 395 training samples to reconstruct 2060 test values. Fig. 4(a) shows ROI A of image I(1) with the TR samples marked with yellow points, while Fig. 4(c) gives the corresponding reconstructed image, obtained with only 19 SVs (highlighted in yellow) and a high accuracy value (PSNR = 30.72 dB in average). Fig. 5 shows a zoom of the reconstruction area from ROI A. Given the fact that the missing area represents an urban region, we note that almost all the SVs come from similar regions. Similarly, Fig. 6(a) shows ROI B of image I(1) with the 395 TR samples marked with yellow points, and Fig. 6(c) shows the reconstruction of the missing area with the position of the selected SVs. In this case, only half of the SVs belong to the urban region. This result is probably due to the fact that the dominating class area in ROI B is the vegetation. It may explain the lower value of the PSNR (PSNR = 26.91 dB in average) and the need of more SVs (#SV = 34). For more visual details about the reconstructions of the two ROI images, see the zoom images in Figs. 5 and 7.

C. Comparative Analysis

In this subsection, we report a comparison with state-of-the-art reconstruction techniques, where the experimental settings
are, in general, similar as for the GKCR method. The first one is a multiresolution inpainting (MRI) technique, which does not need temporal information to reconstruct the missing region [11]. It adopts $3 \times 3$, $6 \times 6$, and $9 \times 9$ square patches. In a similar way as the GKCR, other techniques perform regression within a multitemporal reconstruction context. The first one is the contextual multiple linear prediction (CMLP), which exploits only temporal information [7]. The second one is the CNP, which exploits local-spectral information adopting in a local neighborhood of the missing area [7]. The third one is the First Spatial Model contextual spatiospectral postreconstruction, which, starting from CMLP, takes advantage of the local properties in a predefined neighborhood system (i.e., $3 \times 3$) [8]. The CMLP adopts a simple linear least-square regression, whereas the last two techniques exploit RBF kernels in an SV regression, where a similar CV step is adopted. Table II lists the results of the investigated reconstruction techniques carried on ROI A (mean and standard deviation over ten runs). The poorest results are yielded by the MRI and CMLP methods. They return unsatisfactory results, i.e., MRI due to its unsupervised nature and to the fact it does not make use of the temporal dimension of the image sequence. Even worse is the second, i.e., the CMLP, mainly because it adopts a simple linear regression. The best method is the GKCR, which exhibits the higher PSNR values. We can perceive that the accuracy results are stable in the three ratio cases, it is still possible to reach high accuracy also with a reduced complexity, namely, a smaller number of SVs, particularly if compared with the CNP method. Its underlying idea to exploit all available information (temporal, spectral, and spatial information) within a suitable kernel fusion framework has provided it with a superior capability to handle the reconstruction issue.

V. CONCLUSION

In this letter, we have proposed a new reconstruction method for missing data due to cloud covers. It integrates in the reconstruction process two types of information, i.e., the position of the missing value and the radiometric information. Their fusion performed by means of a kernel combination, together with the power of the SVR, has made it particularly promising as suggested by the experiments. The price of its superior effectiveness is however a higher (but still contained) computation time because of the larger number of free parameters to estimate compared with reference reconstruction methods.

ACKNOWLEDGMENT

The authors would like to thank C. C. Chang and C. J. Lin for supplying the software library, so-called LIBSVM, used in this letter, and Centre National d’Etudes Spatiales (CNES) for making the FORMOSAT-2 images available online.

REFERENCES


