SEMISUPERVISED NONLINEAR FEATURE EXTRACTION FOR IMAGE CLASSIFICATION

Emma Izquierdo-Verdiguier[1], *Luis Gómez-Chova*[1], *Lorenzo Bruzzone*[2] *and Gustavo Camps-Valls*[1]

[1]Image Processing Laboratory (IPL). Universitat de València, Spain.
{izverem,chovago,gcamps}@uv.es, http://isp.uv.es
[2]Remote Sensing Laboratory, Dept. of Information Engineering and Computer Science. University of Trento, Italy.
lorenzo.bruzzone@ing.unitn.it, http://rslab.disi.unitn.it

ABSTRACT

Feature extraction is of paramount importance for an accurate classification of remote sensing images. Techniques based on data transformations are widely used in this context. However, linear feature extraction algorithms, such as the principal component analysis and partial least squares, can address this problem in a suboptimal way because the data relations are often nonlinear. Kernel methods may alleviate this problem *only* when the structure of the data manifold is properly captured. However, this is difficult to achieve when small-size training sets are available. In these cases, exploiting the information contained in unlabeled samples together with the available training data can significantly improve data description by defining an effective semisupervised nonlinear feature extraction strategy. We present a novel semisupervised Kernel Partial Least Squares (KPLS) algorithm for non-linear feature extraction. The method relies on combining two kernel functions: the standard RBF kernel using labeled information and a *generative* kernel directly learned by clustering the data. The effectiveness of the proposed method is successfully illustrated in multi- and hyper-spectral remote sensing image classification: accuracy improvements between $+15-20\%$ over standard PCA and $+10\%$ over advanced kernel PCA and KPLS for both images is obtained. Matlab code is available at http://isp.uv.es for the interested readers.

Index Terms— Classification, feature extraction, kernel methods, partial least squares (PLS), generative kernels

1. INTRODUCTION

Feature extraction consists in identifying the most discriminative variables for data classification. These variable are often associated with the most relevant directions in the data distribution. The family of multivariate analysis methods for feature extraction is commonly used to reduce the data dimensionality by projecting points onto the most relevant directions. Principal component analysis (PCA) [1] and partial least squares (PLS) [2] are two of the most common linear feature extraction methods in remote sensing data analysis. However, when the features and the target variables are nonlinearly related, linear methods cannot properly describe the data distribution. Different non-linear versions of PCA and PLS have been developed, which can address non-linear problems either by local approaches [3], neural networks [4], or kernel-based algorithms [5].

In the last decade, kernel methods have attracted the interest of the remote sensing community because they allow one to develop nonlinear models from linear ones in a very easy and intuitive way [6]. Essentially, kernel methods project the input data to a high dimensional Hilbert space, and define a linear method therein. The model is nonlinear with respect the input space. Interestingly, there is no need to work explicitly with the mapped data, but one computes the nonlinear relations between data via a kernel (similarity) function implicitly. Kernel methods have in general good performance in the case of high dimensional problems and low number of training examples. This is the approach used in kernel principal components analysis (KPCA) [5] and kernel partial least squares (KPLS) [7]. The main difference between KPCA and KPLS is that while KPCA finds the projections that maximize the variance of the input data in the feature space, KPLS extracts projections that account for both the projected input and target data (labels). In this paper, we focus on the KPLS method, which proved to be effective and can extract nonlinear features aligned with the class labels. These features are then used in canonical linear classification or regression.

Extracting nonlinear features by KPLS is a very complex problem in the common situation in remote sensing where relatively few labeled data points are available. Including the information conveyed by unlabeled data via *semisupervised learning* can potentially improve the feature extraction task. The semisupervised framework has recently attracted a considerable amount of theoretical [8] as well as remote sensing applied research [6]. In this paper, we present a new semisupervised KPLS method for nonlinear feature extraction. Our approach considers to modify the kernel similarity function via a semisupervised kernel defined on the basis of clustering the analyzed image. Specifically, we propose to combine a standard supervised kernel with a Gaussian mixture model (GMM) clustering algorithm. While the supervised kernel exploits the information conveyed by the labeled samples, the cluster kernel accounts for the structure of the data manifold. The proposed semisupervised KPLS (SS-KPLS) method is successfully tested in very high resolution and hyperspectral image classification scenarios.

The paper is outlined as follows. Section 2 reviews the standard formulation of KPLS and highlights the problems encountered when dealing with very few labeled samples. This motivates the introduction of the proposed method in Section 3. Section 4 presents the data set and the experimental results. Finally, Section 5 concludes this paper.

This paper has been partially supported by the Spanish Ministry for Science and Innovation under projects AYA2008-05965-C04 and CSD2007-18.

978-1-4673-1159-5/12/$31.00 ©2012 IEEE

2. KERNEL PARTIAL LEAST SQUARES

Notationally, we are given a set of l training data pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{l}$, with $\mathbf{x}_i \in \mathbb{R}^N$, $\mathbf{y}_i \in \mathbb{R}^M$. By using matrix notation we can write, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_l]^\top$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_l]^\top$, where superscript $^\top$ denotes matrix or vector transposition. For classification problems, $Y_{ij} = 1$ if sample $\mathbf{x}_i$ belongs to class j and $Y_{ij} = 0$ otherwise. We denote by $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ the centered versions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Note that, centering removes the mean of every variable in the corresponding matrix.

KPLS is the nonlinear kernel-based extension of PLS [9], which is based on maximizing the variance between the projected data in a proper Hilbert space $\mathcal{H}$ and the target data matrix $\tilde{\mathbf{Y}}$ (i.e. the labels):

$$\text{KPLS:} \quad \begin{aligned} \mathbf{U}, \mathbf{V} = \underset{\mathbf{U},\mathbf{V}}{\arg\max} \;\; & \text{Tr}\{(\tilde{\mathbf{\Phi}}\mathbf{U})^\top \tilde{\mathbf{Y}}\mathbf{V}\} \\ \text{subject to: } & \mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}, \end{aligned} \tag{1}$$

where matrix $\tilde{\mathbf{\Phi}}^\top$ contains the mapped data centered in the Hilbert space[1], and $\mathbf{U}$ and $\mathbf{V}$ are the projection matrices to be estimated for the data and the labels, respectively.

To solve this problem we use the representer's theorem, which states that all projection vectors (the columns of $\mathbf{U}$) can be approximated as a linear combination of the training data, i.e. $\mathbf{U} = \tilde{\mathbf{\Phi}}^\top\mathbf{A}$, where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{n_p}]$ and $\boldsymbol{\alpha}_i$ is an l-length column vector containing the coefficients for the ith projection vector. Introducing this expression into (1), the maximization problem becomes:

$$\text{KPLS (2):} \quad \begin{aligned} \mathbf{A}, \mathbf{V} = \underset{\mathbf{A},\mathbf{V}}{\arg\max} \;\; & \text{Tr}\{\mathbf{A}^\top\mathbf{K}\tilde{\mathbf{Y}}\mathbf{V}\} \\ \text{subject to: } & \mathbf{A}^\top\mathbf{K}\mathbf{A} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}, \end{aligned} \tag{2}$$

where we have defined the symmetric centered kernel matrix $\mathbf{K} = \tilde{\mathbf{\Phi}}\tilde{\mathbf{\Phi}}^\top$ containing the inner products between pairs of points in feature spaces, $K(\mathbf{x}_i, \mathbf{x}_j) = \langle\tilde{\boldsymbol{\phi}}(\mathbf{x}_i), \tilde{\boldsymbol{\phi}}(\mathbf{x}_j)\rangle$. The solution to this problem can be obtained from the singular value decomposition of $\mathbf{K}\tilde{\mathbf{Y}}$. Alternatively, the problem can be efficiently solved using the following two-steps iterative procedure (see [5, Sec. 6.7] for more details):

1. Find the largest singular value of $\mathbf{K}\tilde{\mathbf{Y}}$, and the associated vector directions: $\{\boldsymbol{\alpha}_i, \mathbf{v}_i\}$.
2. Deflate the kernel matrix and labeled vector using:

$$\mathbf{K} \leftarrow \left[\mathbf{I} - \frac{\mathbf{K}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top\mathbf{K}}{\boldsymbol{\alpha}_i^\top\mathbf{K}\mathbf{K}\boldsymbol{\alpha}_i}\right]\mathbf{K}\left[\mathbf{I} - \frac{\mathbf{K}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top\mathbf{K}}{\boldsymbol{\alpha}_i^\top\mathbf{K}\mathbf{K}\boldsymbol{\alpha}_i}\right] \tag{3}$$

$$\mathbf{Y} = \mathbf{Y} - \mathbf{K}\boldsymbol{\alpha}_i\mathbf{Y}\frac{\mathbf{K}\boldsymbol{\alpha}_i}{\|\mathbf{K}\boldsymbol{\alpha}\|_2^2} \tag{4}$$

This deflation procedure allows us to extract more features than classes. For a more detailed description as well as implementation details, the reader is referred to [5, 6].

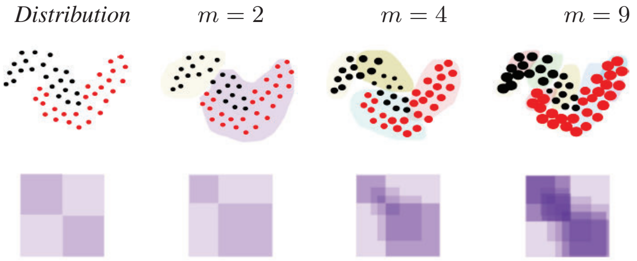

Fig. 1. Illustration of the cluster kernel construction. The method clusters data with GMM clustering for $m = \{2, 4, 9\}$, and acumulates the similarities in a multiscale way. Samples classified in the same clusters should belong to the same class. The multiscale cluster kernel (right kernel) is a better estimation of the optimal ideal kernel $K = \mathbf{y}\mathbf{y}^\top$ (left kernel).

3. PROPOSED SEMISUPERVISED KPLS

The underlying idea of the proposed semisupervised KPLS (SS-KPLS) is to modify the KPLS kernel (similarity) function $K(\mathbf{x}_i, \mathbf{x}_j)$ to account for the distribution of unlabeled pixels. To this aim, we propose the Gaussian mixture model (GMM) cluster kernel, which consists in combining a kernel on labeled data with a kernel computed from clustering unlabeled data. The multiscale cluster kernel is obtained as follows:

1. Compute the *supervised* kernel function:

$$K_s(\mathbf{x}_i, \mathbf{x}_j) = \langle\boldsymbol{\phi}_s(\mathbf{x}_i), \boldsymbol{\phi}_s(\mathbf{x}_j)\rangle \tag{5}$$

2. Run t times the Gaussian mixture model (GMM) clustering algorithm with different initializations and with different number of clusters q. This results in $q{\cdot}t$ cluster assignments where each sample $\mathbf{x}_i$ has its corresponding posterior probability vector $\boldsymbol{\pi}_i \in \mathbb{R}^m$.

3. Build a *cluster* kernel K_c based upon the fraction of times that $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned to the same cluster:

$$K_c(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{Z}\sum_{p=1}^{t}\sum_{m=2}^{q}\boldsymbol{\pi}_i^\top\boldsymbol{\pi}_j, \tag{6}$$

where Z is a normalization factor. An illustrative toy example of the multiscale cluster kernel construction is shown in Fig. 1.

4. Define the final kernel function K as the weighted sum of the supervised and the cluster kernels:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \beta K_s(\mathbf{x}_i, \mathbf{x}_j) + (1 - \beta)K_c(\mathbf{x}_i, \mathbf{x}_j), \tag{7}$$

where $\beta \in [0, 1]$ is a scalar parameter.

5. Plug K into the standard KPLS solver (see Section 2).

Note that the proposed kernel in (6) is a valid kernel because it corresponds to a summation of inner products in tq-dimensional spaces, $\boldsymbol{\phi}_c(\mathbf{x}_i) = \bigcup_{p=1,m=2}^{t,q}\boldsymbol{\pi}_i$, where operator $\bigcup$ represents vector concatenation. The summation of kernels done in step 4, Eq. (7), leads also to valid Mercer's kernels, as

[1] Centering in feature space can be done implicitly via the simple kernel matrix operation $\mathbf{K} \leftarrow \mathbf{HKH}$, where $H_{ij} = \delta_{ij} - \frac{1}{l}$, δ represents the Kronecker symbol, and $\delta_{i,j} = 1$ if $i = j$, and zero otherwise.

it corresponds to the concatenation of feature vectors in the Hilbert space, $\boldsymbol{\phi}(\mathbf{x}_i) = \{\sqrt{\beta}\cdot\boldsymbol{\phi}_s(\mathbf{x}_i)^\top, \sqrt{1-\beta}\cdot\boldsymbol{\phi}_c(\mathbf{x}_i)^\top\}^\top$.

The new averaged kernel accounts for similarities at small and large scales in the manifold between the samples by using both labeled and unlabeled data. Note that finding a proper kernel is equivalent to learn metric relations in the manifold which are defined here through a generative model learned from the data. The proposed kernel generalizes previous approaches based on multiscale cluster kernels. For example, the kernel in (6) reduces to the approach in [10] when only the cluster assignment with maximum posterior probability is considered.

4. EXPERIMENTAL RESULTS

This section presents the results obtained by applying the proposed SS-KPLS technique to remote sensing multispectral and hyperspectral image classification. The next section details the data used in the experiments. Then, we focus our attention on the accuracy and robustness of the proposed algorithm in terms of the number of extracted nonlinear features. Finally, we analyze the eigenspectrum, structure, and information content of the derived kernels.

4.1. Data

The first image dataset consists of 4 spectral bands acquired on a residential neighborhood of the city of Zürich by the QuickBird satellite in 2002. The portion of the image analyzed has size (329×347) pixels. The original image has been pansharpened using a Bayesian data fusion method to attain a spatial resolution of 0.6 m. Nine classes of interest have been defined by photointerpretation. According to the good results obtained in previous studies [11], a total of 18 spatial features extracted using morphological opening and closing have been added to the spectral bands, resulting in a final 22-dimensional vector.

The second image was acquired by the DAIS7915 sensor over the city of Pavia (Italy), and constitutes a challenging 9-class urban classification problem dominated by structural features and relatively high spatial resolution (5-meter pixels). Following previous works on classification of this image, we took into account only 40 spectral bands in the range [0.5, 1.76] μm, and thus skipped thermal and middle infrared bands above 1958 nm.

4.2. Experimental setup

For our experiments, we used only 4 labeled samples per class to illustrate the robustness of the proposed method to challenging ill-posed classification problems. In order to define the $(q \cdot t)$ cluster centers and the pixel posterior probabilities for each of them, $\boldsymbol{\pi}_i$, we used 190 unlabeled samples *per* class for both images. In all cases, we used $t = q = 20$ and the parameter β was tuned between $[0, 1]$ in steps of 0.05 for each number of extracted features with the proposed algorithm. Once the mixture models are computed and stored, the data are assigned to the most probable Gaussian mode and K_c is constructed accordingly. The same assignment is used for predicting the class membership of an unknown test pixel. A 3-fold cross-validation procedure was run to find the optimal σ parameter, which was varied between $[0.5, 2] \times s$, where s represents the median distance between all labeled data.

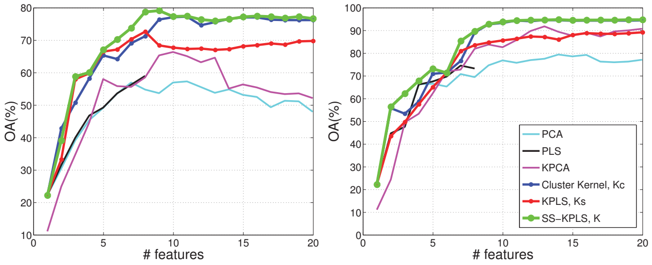

Fig. 2. Overall accuracy as a function of extracted nonlinear features for the Zürich image (left) and Pavia image (right).

Once the projections are obtained for all methods, the discriminative power of the features was tested using a simple linear model followed by a "winner-takes-all" activation function , i.e. $\hat{\mathbf{y}} = \text{w.t.a.}[\mathbf{W}^\top\tilde{\boldsymbol{\phi}}(\mathbf{x})]$, where $\mathbf{W}$ is the optimal regression matrix given by $\mathbf{W} = \tilde{\boldsymbol{\Phi}}^\dagger\tilde{\mathbf{Y}}$. For testing the models, the overall accuracy OA[%] and the estimated Cohen's kappa statistic κ are computed over a total of 1,710 test randomly chosen samples in both images. We also provide the classification maps and the accuracies obtained in the whole scenes. Matlab code and demos are available for the interested reader in http://isp.uv.es.

4.3. Results and discussion

We evaluated the accuracy of several methods for a varying number of extracted features: 1) unsupervised linear, PCA, and its nonlinear version, KPCA; 2) supervised feature extraction algorithms (PLS and its nonlinear version KPLS); and 3) the different kernels involved in SS-KPLS. Note that the proposed SS-KPLS generalizes both the supervised KPLS (when $\beta = 1$) and a fully unsupervised feature extraction (for $\beta = 0$).

Results are shown in Fig. 2. In general, nonlinear kernel methods (KPCA, KPLS and variants) outperform linear approaches (PCA and PLS). The proposed SS-KPLS outperforms the standard KPLS and the cluster kernel. The (unsupervised) generative cluster kernel proposed here outperforms the supervised kernel with an increasing number of features. When a higher number of nonlinear features is extracted, all curves become stable, but the proposed SS-KPLS largely outperforms the standard PCA between +15-20% and the more advanced KPCA or KPLS by about +10%. The behaviour of PCA and KPCA in the Zürich image is worth analyzing because higher accuracy is not obtained with higher number of extracted features, revealing a kind of overfitting problem. This effect has been recently reported in the literature [12]. This is not the case of the proposed *unsupervised* kernel K_c. These results are confirmed by the visual inspection of the classification maps shown in Fig. **??**, where the SS-KPLS shows a clear and consistent gain over KPLS of about +6% for the Pavia dataset.

Figure 4 shows the eigenvalues of the best kernels for the Pavia image. The eigendecomposition of the semisuper-

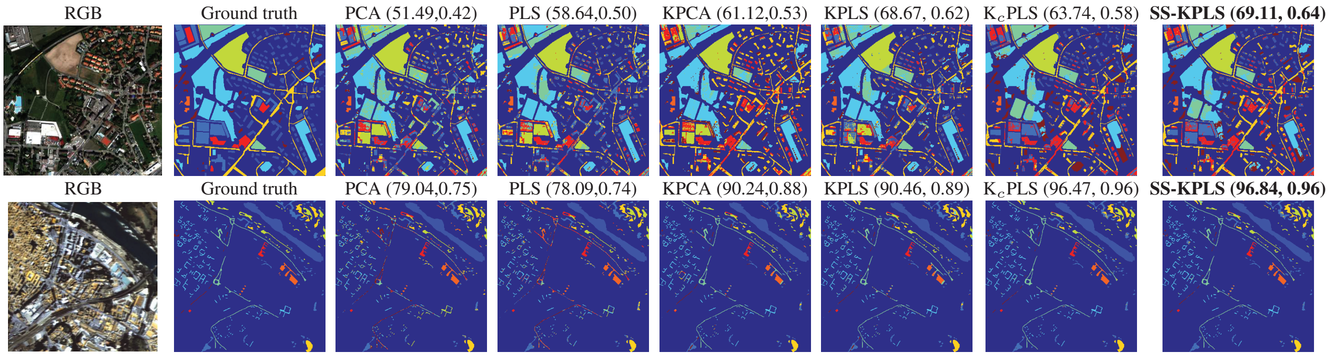

Fig. 3. Left to right: RGB composite, ground truth and three classification maps along with the overall accuracy and kappa for the Zürich (top) and the Pavia (bottom) images for 20 extracted features.

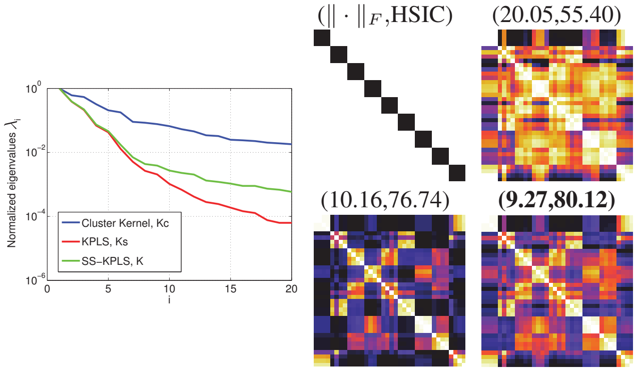

Fig. 4. Left: Normalized eigenvalues for all kernels used in the Pavia dataset. Right: ideal and used kernels, along quantitative measures of error $\|\cdot\|_F$ and dependence (HSIC).

vised kernel shows a tradeoff between the supervised and the unsupervised kernels, as expected. It is worth noting that the unsupervised spectrum (blue line) shows a slower decay because the kernel is indeed quite blocky and sparse. On the other hand, the supervised kernel shows a heavier tail. The introduction of the cluster kernel can be casted as an extra regularization of the supervised kernel. The right plots present the used kernels and their similarity to the ideal one, $\mathbf{K}_{\text{ideal}} = \mathbf{y}\mathbf{y}^\top$. Two quantitative measures are given: the Frobenius norm of the difference of these two kernels, $\|\cdot\|_F$, and the Hilbert-Schmidt Independence Criterion (HSIC) between them [13]. The proposed semisupervised kernel aligns well with the ideal kernel (lower error, higher dependence), and takes advantage of the sharper structure learned by the cluster kernel.

5. CONCLUSIONS

This paper proposed a novel nonlinear feature extraction technique for remote sensing image classification. The method is specifically devised for addressing critical ill-posed problems where the number of training samples available is relatively small, and thus using unlabeled samples in a semisupervised framework can significantly improve the representation of the data. Note that these problems are common in operational applications of remote sensing. Good results were obtained on both multispectral and hyperspectral data sets considered in our experiments, where the proposed method largely outperformed supervised and unsupervised linear and nonlinear literature approaches. Future work will consider the direct use of the generative cluster kernel in unsupervised image segmentation, and the study of the induced metric space by the kernels.

6. REFERENCES

[1] I. T. Jollife, *Principal Component Analysis*, Springer, 1986.

[2] R. Rosipal and N. Krämer, “Overview and recent advances in partial least squares,” in *Subspace, Latent Structure and Feature Selection*. 2006, vol. 3940 of *LNCS*, pp. 34–51, Springer.

[3] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, December 2000.

[4] Mark A. Kramer, “Nonlinear principal component analysis using autoassociative neural networks,” *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[5] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[6] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*, John Wiley and Sons, 2009.

[7] R. Rosipal and L.J. Trejo, “Kernel partial least squares regression in reproducing kernel Hilbert space,” *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, March 2002.

[8] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, 1st edition, 2006.

[9] J. Arenas-García and G. Camps-Valls, “Efficient kernel orthonormalized PLS for remote sensing applications,” *IEEE Trans. Geosc. Rem. Sens.*, vol. 46, pp. 2872 –2881, Oct 2008.

[10] D. Tuia and G. Camps-Valls, “Urban image classification with semisupervised multiscale cluster kernels,” *IEEE JSTARS*, vol. 4, pp. 65–74, Mar 2011.

[11] D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls, “Multisource composite kernels for urban-image classification,” *IEEE Geosc. Rem. Sens. Lett.*, vol. 7, pp. 88–92, 2010.

[12] Mikio L. Braun, Joachim Buhmann, and Klaus-Robert Müller, “On relevant dimensions in kernel feature spaces,” *Journal of Machine Learning Research*, vol. 9, pp. 1875–1908, Aug 2008.

[13] G. Camps-Valls, J. Mooij, and B. Schölkopf, “Remote sensing feature selection by kernel dependence measures,” *IEEE Geosc. Rem. Sens. Lett.*, vol. 7, no. 3, pp. 587–591, 2010.