Remote Sensing Feature Selection by Kernel Dependence Measures

Gustavo Camps-Valls, *Senior Member, IEEE*, Joris Mooij, and Bernhard Schölkopf

***Abstract*—This letter introduces a nonlinear measure of independence between random variables for remote sensing supervised feature selection. The so-called Hilbert–Schmidt independence criterion (HSIC) is a kernel method for evaluating statistical dependence and it is based on computing the Hilbert–Schmidt norm of the cross-covariance operator of mapped samples in the corresponding Hilbert spaces. The HSIC empirical estimator is easy to compute and has good theoretical and practical properties. Rather than using this estimate for maximizing the dependence between the selected features and the class labels, we propose the more sensitive criterion of minimizing the associated HSIC p-value. Results in multispectral, hyperspectral, and SAR data feature selection for classification show the good performance of the proposed approach.**

***Index Terms*—Dependence estimation, feature selection, image classification, kernel methods, support vector machine (SVM).**

I. Introduction

REDUCING the dimensionality of the data while keeping the most of its expressive power is the goal of feature selection, and a great many methods have been proposed, either filters, wrappers, or embedded [1], [2]. Filters use an indirect measure of the quality of the selected features, e.g., the correlation between each input feature (spectral channel) and the observed output (class label). In wrapper methods, a fitness criterion between the observed and predicted outputs by a classifier is directly optimized. Filters converge much faster than wrapper methods, and select features independently of the subsequent classifier, which facilitates feature interpretation. Note that often the user is much more interested in *understanding* the relative relevance of the considered features than in minimizing a classification error. Filter methods have been extensively studied in remote sensing. In [3], a feature-selection procedure was proposed to combine spectral channels, while a canonical correlation method was applied in [4] to sensors with overlapping bands. Strategies to constrain the search space were deployed in [5]. The main problems with most of the filter methods are the following: 1) The relationship between sets of features and the class labels is not jointly considered, as they usually use pairwise feature–label measurements, and 2) they do assume either linear dependences (e.g., by using Pearson's correlation) or *ad hoc* criteria of class separability (e.g., mean class differences). These problems reduce their usefulness for classification and have motivated the recent interest in wrappers and embedded methods. For example, recursive feature elimination [6] or genetic algorithms [7] optimizing the accuracy of support vector machines (SVM) have been successfully used in remote sensing data analysis. The main problem of these methods is their excessive computational cost, particularly high in the case of hyperspectral images, which often requires including heuristics in the procedure. Besides, these feature-selection methods may suffer from overfitting when working with a small number of training samples, as it is typically the case in remote sensing data classification.

Here we introduce a filter approach based on measuring the nonlinear dependence between features and class labels. The so-called Hilbert–Schmidt independence criterion (HSIC) is a kernel method for evaluating statistical dependence between random variables. It is based on computing the Hilbert–Schmidt norm of the cross-covariance operator of mapped samples in the corresponding Hilbert spaces [8]. The so-called backward HSIC (BAHSIC) procedure iteratively removes the feature that is least important by maximizing the dependence of the remaining features on the class labels. The method was originally presented in [9] for general-purpose applications and further used in a bioinformatics application [10]. Here, we analyze its capabilities in the specific context of remote sensing image classification. Additionally, unlike the original method [9], we propose to minimize the p-value associated to the empirical HSIC test rather than maximizing the HSIC value.

The main problems of both filter and wrapper methods described before are alleviated with the use of HSIC: 1) Good robustness capabilities to high dimensionality and low number of training samples are typically observed for kernel methods, particularly in remote sensing data processing [11]; 2) the criterion is not restricted to estimate pairwise dependences but captures higher order relations between features; and 3) the proposed method is very simple to implement and provides good and generally interpretable results. Unlike most of the feature-selection methods, the proposed HSIC criterion can be *directly* applied to binary, multiclass, or even regression problems by choosing appropriate kernels. Finally, it can be demonstrated that the proposed method contains other feature-selection procedures as particular cases.

II. Kernel Methods for Measuring Independence

This section presents a criterion for measuring general forms of dependence between random variables.

Manuscript received October 20, 2009; revised January 8, 2010. Date of publication April 1, 2010; date of current version April 29, 2010. This work was supported in part by the Spanish Ministry of Education and Science under Projects TEC2009-13696, AYA2008-05965-C04-03, and CONSOLIDER/CSD2007-00018.

G. Camps-Valls is with the Image Processing Laboratory, Universitat de València, 46980 Paterna, Spain (e-mail: gustavo.camps@uv.es).

J. Mooij and B. Schölkopf are with the Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany (e-mail: joris.mooij@tuebingen.mpg.de; bernhard.schoelkopf@tuebingen.mpg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2010.2041896

A. Linear Dependence Between Random Variables

Let us consider two spaces $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, on which we jointly sample observation pairs $(\mathbf{x}, \mathbf{y})$ from distribution $\mathbb{P}_{\mathbf{xy}}$. The covariance matrix can be defined as

$$\mathcal{C}_{\mathbf{xy}} = \mathbb{E}_{\mathbf{xy}}(\mathbf{x}\mathbf{y}^\top) - \mathbb{E}_{\mathbf{x}}(\mathbf{x})\mathbb{E}_{\mathbf{y}}(\mathbf{y}^\top) \tag{1}$$

where $\mathbb{E}_{\mathbf{xy}}$ is the expectation with respect to $\mathbb{P}_{\mathbf{xy}}$, $\mathbb{E}_{\mathbf{x}}$ is the expectation with respect to the marginal distribution $\mathbb{P}_{\mathbf{x}}$ (here and in the following, we assume that all these quantities exist), and $\mathbf{y}^\top$ is the transpose of $\mathbf{y}$. The covariance matrix encodes all first-order dependences between the random variables. A statistic that efficiently summarizes the content of this matrix is its Hilbert–Schmidt norm. The square of this norm is equivalent to the squared sum of its eigenvalues γ_i, $\|\mathcal{C}_{\mathbf{xy}}\|^2_{\mathrm{HS}} = \sum_i \gamma_i^2$. This quantity is zero if and only if there exists no first-order dependence between $\mathbf{x}$ and $\mathbf{y}$, but is limited. Note that the Hilbert–Schmidt norm is limited to the detection of first-order relations, and thus, more complex (higher order) effects cannot be captured.

B. Measuring Dependence With Kernels

The nonlinear extension of the notion of covariance was proposed in [8] and [12]. Essentially, let us define a (possibly nonlinear) mapping $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$ such that the inner product between features is given by a positive definite (p.d.) kernel function $k_x(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle$. The feature space $\mathcal{F}$ has the structure of a reproducing kernel Hilbert space (RKHS). Let us now denote another feature map $\boldsymbol{\psi} : \mathcal{Y} \to \mathcal{G}$ with associated p.d. kernel function $k_y(\mathbf{y}, \mathbf{y}') = \langle \boldsymbol{\psi}(\mathbf{y}), \boldsymbol{\psi}(\mathbf{y}') \rangle$. Then, it is possible to define a cross-covariance operator between these feature maps, similar to the covariance matrix in (1). The cross-covariance operator is a linear operator $\mathcal{C}_{\mathbf{xy}} : \mathcal{G} \to \mathcal{F}$ such that

$$\mathcal{C}_{\mathbf{xy}} = \mathbb{E}_{\mathbf{xy}} \left[(\phi(\mathbf{x}) - \mu_{\mathbf{x}}) \otimes (\psi(\mathbf{y}) - \mu_{\mathbf{y}}) \right] \tag{2}$$

where $\otimes$ is the tensor product, $\mu_x = \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]$, and $\mu_y = \mathbb{E}_{\mathbf{y}}[\boldsymbol{\psi}(\mathbf{y})]$. See more details in [13] and [14]. The squared norm of the cross-covariance operator, $\|\mathcal{C}_{\mathbf{xy}}\|^2_{\mathrm{HS}}$ is called the HSIC and can be expressed in terms of kernels [8]

$$\begin{aligned} \mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{\mathbf{xy}}) &= \|\mathcal{C}_{\mathbf{xy}}\|^2_{\mathrm{HS}} \\ &= \mathbb{E}_{\mathbf{xx'yy'}} \left[k_x(\mathbf{x}, \mathbf{x}') k_y(\mathbf{y}, \mathbf{y}') \right] \\ &\quad + \mathbb{E}_{\mathbf{xx'}} \left[k_x(\mathbf{x}, \mathbf{x}') \right] \mathbb{E}_{\mathbf{yy'}} \left[k_y(\mathbf{y}, \mathbf{y}') \right] \\ &\quad - 2\mathbb{E}_{\mathbf{xy}} \left[\mathbb{E}_{\mathbf{x}'} \left[k_x(\mathbf{x}, \mathbf{x}') \right] \mathbb{E}_{\mathbf{y}'} \left[k_y(\mathbf{y}, \mathbf{y}') \right] \right] \end{aligned}$$

where $\mathbb{E}_{\mathbf{xx'yy'}}$ is the expectation over both $(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}_{\mathbf{xy}}$ and an additional pair of variables $(\mathbf{x}', \mathbf{y}') \sim \mathbb{P}_{\mathbf{xy}}$ drawn independently according to the same law.

Now, given a sample data set $Z = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\}$ of size m drawn from $\mathbb{P}_{\mathbf{xy}}$, an empirical estimator of HSIC is [8]

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{\mathbf{xy}}) = \frac{1}{m^2} \mathrm{Tr}(\mathbf{K}_x \mathbf{H} \mathbf{K}_y \mathbf{H}) \tag{3}$$

where Tr is the trace, $\mathbf{K}_x$ and $\mathbf{K}_y$ are the kernel matrices for the data $\mathbf{x}$ and the labels $\mathbf{y}$, respectively, and $H_{ij} = \delta_{ij} - (1/m)$ centers the data and the label features in $\mathcal{F}$ and $\mathcal{G}$. Here, δ represents the Kronecker symbol, where $\delta_{i,j} = 1$ if $i = j$, and zero otherwise.

Note that the actual HSIC is the Hilbert–Schmidt norm of an operator mapping between potentially infinite dimensional spaces, and thus, would give rise to an infinitely large matrix. However, due to the kernelization, the empirical HSIC only depends on computable matrices of size $m \times m$.

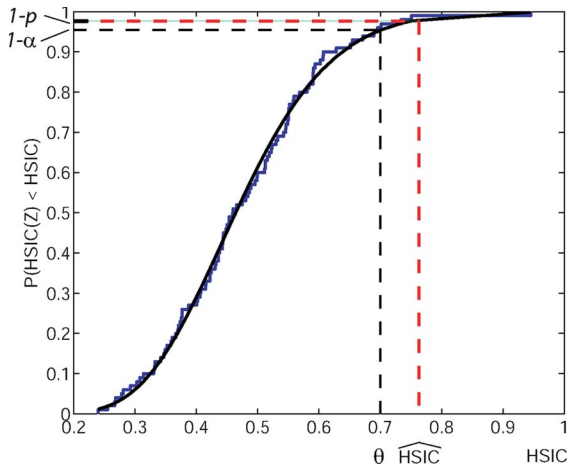

Fig. 1. CDF of HSIC under H_0 for $m = 100$ samples obtained empirically using (blue line) 1000 independent draws of HSIC and approximated using (black line) the two-parameter Gamma distribution of (4). We also indicate the threshold for the HSIC test θ, which corresponds to the inverse cdf of the significance level $\alpha = 0.05$, and the p-value which corresponds to $\widehat{\mathrm{HSIC}}$.

C. Setting the HSIC Decision Threshold

In [12, Ch. 4], several statistical tests of independence based on the empirical HSIC estimator (3) are revised. The test should discern between the null hypothesis $\mathrm{H}_0 : \mathbb{P}_{\mathbf{xy}} = \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$ (factorization means independence) and the alternative hypothesis $\mathrm{H}_1 : \mathbb{P}_{\mathbf{xy}} \neq \mathbb{P}_{\mathbf{x}}\mathbb{P}_{\mathbf{y}}$. This is done by comparing the test statistic HSIC with a given threshold.[1] Among the possibilities to define such a threshold over the HSIC estimate, a reasonable one is to approximate the null distribution as a two-parameter gamma distribution, as suggested in [15]

$$\widehat{\mathrm{HSIC}} \sim \frac{x^{a-1} e^{-x/b}}{m b^a \Gamma(a)} \tag{4}$$

where $a = \mathbb{E}[\widehat{\mathrm{HSIC}}]^2 / \mathbb{V}[\widehat{\mathrm{HSIC}}]$ and $b = \mathbb{V}[\widehat{\mathrm{HSIC}}] / \mathbb{E}[\widehat{\mathrm{HSIC}}]$, whose detailed expressions can be found in [12, Th. 43 and 44]. Then, the threshold θ is computed through the inverse cumulative density function (cdf) of the $1 - \alpha$ value, where α is the adopted significance level (typically, $\alpha = 0.05$ or $\alpha = 0.01$). Two random variables are then considered dependent if $\widehat{\mathrm{HSIC}} \geq \theta$, and independent otherwise. Alternatively, as proposed here, one can directly compute the HSIC p-value from the HSIC estimate and its cdf to test dependence (see Fig. 1). The p-value represents the probability of obtaining a result at least as extreme as the actually observed, assuming that the null hypothesis is true.

D. HSIC for Feature Selection

In [9], a method for using HSIC as a criterion for feature selection was introduced. The method is summarized in

[1]The null hypothesis H_0 is accepted if the test is lower than the threshold. The Type I error is defined as the probability of rejecting H_0 based on the observed sample, despite $\mathbf{x}$ and $\mathbf{y}$ being independent. Conversely, the Type II error is the probability of accepting the null hypothesis when the underlying variables are dependent. The level α of a test is an upper bound on the Type I error and is used to set the test threshold.

Algorithm 1. HSIC is zero if the spectral bands and class labels are fully independent. Since we search for the most dependent features on the labels, a possibility is to select features that maximize HSIC. The method starts with all features S and iteratively removes the feature that is least dependent on the class labels (step 3a in Algorithm 1). Note that each time a feature is removed, the kernel parameter must be estimated and HSIC and its p-value computed. Thus, the BAHSIC method iteratively generates a feature list S^* in increasing level of importance. After ranking the features, one only has to pick up a number of highly relevant features. Although forward selection is computationally more efficient, backward elimination provides better features in general, since the features are assessed within the context of all others.

Algorithm 1 Feature selection with HSIC-based criteria. The proposed BAHSIC$_p$ method replaces the maximization of HSIC in step 3a with the minimization of the HSIC p-value in 3b.

Input: Full set of features S, training data set $\{\mathbf{x}_i, y_i\}$
Output: Ranking of features, S^*
1: $S^* \leftarrow \emptyset$
repeat
  for each $j \in S$ **do**
    2: Estimate kernel width σ_j from data $\mathbf{x}_i(S \setminus \{j\})$
  end for
  3a: $j \leftarrow \arg\max_j \text{HSIC}(\sigma_j, S \setminus \{j\})$
  3b: $j \leftarrow \arg\min_j \text{HSIC}_p(\sigma_j, S \setminus \{j\})$
  4: $S \leftarrow S \setminus \{j\}$
  5: $S^* \leftarrow \langle S^*, j \rangle$
until $S = \emptyset$

In this letter, we propose to alternatively select features by minimizing the HSIC p-value rather than maximizing the HSIC value (step 3b in Algorithm 1). We observed that the p-value is a more sensitive criterion for guiding the selection. Hence, we call the method BAHSIC$_p$.

The algorithm relies on computing HSIC in each iteration, and thus, one could possibly optimize the kernel parameters (step 2 in Algorithm 1). In our experiments, however, we estimated the kernel width from the data as the median distance among all training samples. This allows a very fast convergence of the method, even if suboptimal.

E. Selection of the Kernels

Note that the proposed method has two kernels: the input data kernel k_x and the output label kernel k_y [cf., (3)]. Concerning the data kernel, one can resort to common kernels such as the polynomial or the radial basis function (RBF), $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$, $\sigma \in \mathbb{R}^+$. In the context of dependence estimation, however, it is important to note that from [8, Th. 4], if $\mathcal{F}$ and $\mathcal{G}$ are RKHSs with *universal* kernels k_x and k_y [16], then $\text{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{\mathbf{xy}}) = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent. This is why the RBF or Laplacian kernels (which are universal) are preferred in this specific kernel method over polynomial kernels.

The output (labels) kernel may be defined depending on the task (classification, regression, sorting, or visualization) and the known relations in the output. This is certainly a nice property of the method. For the binary ($\{0, 1\}$-valued) case, the kernel function can be defined as $k_y(\mathbf{y}, \mathbf{y}') = (1/m_+)\mathbf{y}^\top\mathbf{y}$, where m_+ and m_- are the relative number of samples for each class. This corresponds to making the criterion independent of the specific class sizes. For the case of k classes, several possibilities arise. In this letter, we adjusted the inner product between classes and defined $k_y(\mathbf{y}, \mathbf{y}') = \langle \boldsymbol{\psi}(\mathbf{y}), \boldsymbol{\psi}(\mathbf{y}') \rangle$, where $\boldsymbol{\psi}(\mathbf{y}) = 1_{\mathbf{y}}(m/m_{\mathbf{y}}(m - m_{\mathbf{y}})) - \mathbf{z}$, $\mathbf{z} = [(m - m_1)^{-1}, \ldots, (m - m_k)^{-1}]^\top$, and $1_{\mathbf{y}}$ represents the $\mathbf{y}$-th unit vector in $\mathbb{R}^k$ (see [9] for details).

F. Computational Issues

Note that the HSIC empirical estimate is very simple to implement as it basically requires computing the trace of the centered kernel matrices [cf., (3)]. In fact, $\text{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{\mathbf{xy}})$ can be computed in $\mathcal{O}(m^2)$ time, while many other kernel methods cost at least $\mathcal{O}(m^3)$. Nevertheless, this may be a problem when working with many labeled data points, which fortunately is not typically the case in remote sensing data classification.

G. Relation to Other Feature-Selection Methods

An important aspect of BAHSIC is its generality. It can be demonstrated that the method contains other feature-selection methods as particular cases [9]. For instance, BAHSIC reduces to the standard Pearson's correlation when using a linear kernel in both input and output spaces. Similarly, with the proper normalization of the data and use of the linear kernel, BAHSIC reduces to the t-test criterion, B-statistic, or the shrunken centroid methods for feature selection. Nevertheless, other methods such as the mutual information (MI) cannot be seen as examples of HSIC.

III. Data Collection

Several images are considered in the experiments, for both binary and multiclass problems:

1) *Naples 99*. This data set consists of images from ERS2 synthetic aperture radar (SAR) and Landsat TM sensors acquired in 1999 over Naples (Italy). The problem is binary classification: detection of urban versus nonurban areas. The available features were the seven Landsat bands, two SAR backscattering intensities (0–35 days), and the SAR interferometric coherence. We used all seven Landsat TM spectral bands and appended two SAR features: the computed coherence Co and a spatially filtered version of the coherence FCo, which was specially designed to increase the urban-area discrimination [17].
2) *FC1*. The Flightline C1 data is a 12-band multispectral image taken over Tippecanoe County, Indiana (U.S.) by the M7 scanner in June 1966 [18]. The image is 949 $\times$ 220 pixels and contains ten classes, mainly crop types. A ground survey of 70 847 pixels has been used.
3) *Salinas*. This 224-band Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) hyperspectral image was acquired over an agricultural area of California, U.S. A total of 16 crop classes were labeled. This is a high-resolution scene with pixels of 3.7 m. The high number of spectrally similar subclasses makes the classification problem very complex.
4) *Pavia*. This is an image acquired by the DAIS7915 sensor over the city of Pavia (Italy), and constitutes a challenging nine-class urban classification problem dominated by

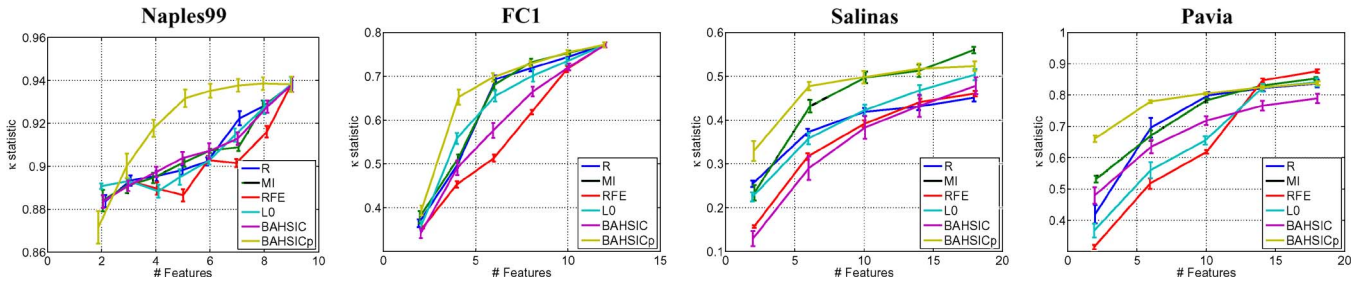

Fig. 2. Kappa statistics in the test set for different numbers of selected features by several algorithms and data sets. The average results over ten realizations are shown, and the error bars indicate the 95% confidence intervals for the mean value.

directional features and relatively high spatial resolution (5-m pixels). Following previous works on classification of this image, we took into account only 40 spectral bands of reflective energy in the range [0.5, 1.76] μm, and thus skipped thermal IR bands and middle IR bands above 1958 nm.

Note that the selected images cover the most significant remote sensing situations and sensors: multispectral (Landsat, FC1), hyperspectral (AVIRIS, DAIS), ERS2 SAR data, and high-resolution imagery (FC1, DAIS).

IV. Experimental Results

A. Experimental Setup

In all cases, we compare the performance of BAHSIC and BAHSIC$_p$ with standard and advanced feature-selection methods: Pearson's correlation R; MI; SVM recursive feature elimination (RFE) [19], and the L_0-norm approach [20]. Note also that due to the method's generality, we are implicitly comparing performance of other methods (cf., Section II-G). The SVM-RFE algorithm analyzes the relevance of input variables by estimating changes in the cost function, $\Delta J_u = \|\mathbf{w}\| - \|\mathbf{w}_u\|$, where $\mathbf{w}$ represents the SVM weight vector in the RKHS for the complete set of input variables and $\mathbf{w}_u$ denotes the SVM weight vector when variable u is removed. The L_0-norm approach iteratively seeks for the maximum classification accuracy of a linear SVM while essentially restricting $\|\mathbf{w}\|_0 < r$, where r is the desired number of features.

In the experiments, we fixed different numbers of selected features and run all the methods with 100 randomly selected samples only. The selected features are then used for SVM classification with an RBF kernel. SVMs are trained following the standard tenfold cross-validation method on the same 100 training examples. The kernel width was tuned in the range $\sigma = \{10^{-3}, \ldots, 10^{3}\}$ and the SVM regularization parameter was varied in $C = \{10^{-1}, \ldots, 10^{3}\}$. The procedure was repeated ten times with different selected samples. Fig. 2 shows the kappa statistic κ [21] in the test set (whole labeled data in the images) as a function of the number of selected features. We show the averaged results and the 95% confidence intervals over the ten realizations.

B. Binary Classifications

We used the Naples95 data set described in Section III for binary classification. The numerical results are shown in Fig. 2 for all the considered feature-selection methods and different numbers of selected features. Note that the proposed HSIC$_p$ selection consistently yields very good results, particularly when more than two features are selected. This is because the BAHSIC$_p$ method selects the coherence and the spatially filtered coherence features most of the times (see Fig. 3). This selection matches that previously reported with other methods in [17]. Only BAHSIC and MI select the FCo feature as one of the top five features. It is also observed that BAHSIC$_p$ focuses on the combination of the thermal and the shortwave IR (SWIR) bands and pays less attention to the visible spectral bands (as the rest of the methods do). This selection is quite consistent in this particular problem: SWIR bands are, in general, useful to detect soil types and soil disturbance since moisture is an important characteristic of soil structure.

C. Multiclass Problems

Results in Fig. 2 show that in all the considered multiclass scenarios, a noticeable gain is obtained by using the p-value to guide HSIC in feature selection. The gain is particularly important when few features are selected. For instance, by selecting the best two features of each method only, average HSIC$_p$ gain in overall accuracy is around 8%. For a moderate number of selected features, the HSIC$_p$ performance saturates and generally outperforms the rest of the methods. Only for a higher number of features, MI shows a slight improvement. It must be noted that RFE performs poorly in all databases, which was already reported elsewhere [22].

The numerical results can be explained in physical terms by looking at the frequency with which the methods select the top five features (see Fig. 3). The most important observation is that, in all data sets, the BAHSIC$_p$ method covers the spectrum more uniformly than the others. This typically results in non-redundant selected features. This behavior is also observed for BAHSIC and MI, which often perform very well (see Fig. 2). The poor numerical results yielded by RFE is clearly due to the low level of feature diversity accounted in the selection. The L_0 method follows a similar behavior in "Pavia" and "Salinas" images.

In the agricultural area of the "FC1" image, BAHSIC$_p$ (and similarly, BAHSIC) more or less uniformly covers the whole spectral range but more focused on selecting features that account for cellular pigment (carotenoids) absorption and chlorophyll absorption (500–700 nm). In the complex Salinas scenario, BAHSIC$_p$ (and also BAHSIC and MI) again selects the bands characterizing leaf pigments but now in combination with some spectral bands around 1 μm (related to plant cell structure) and bands close to 2.2 μm to take into consideration

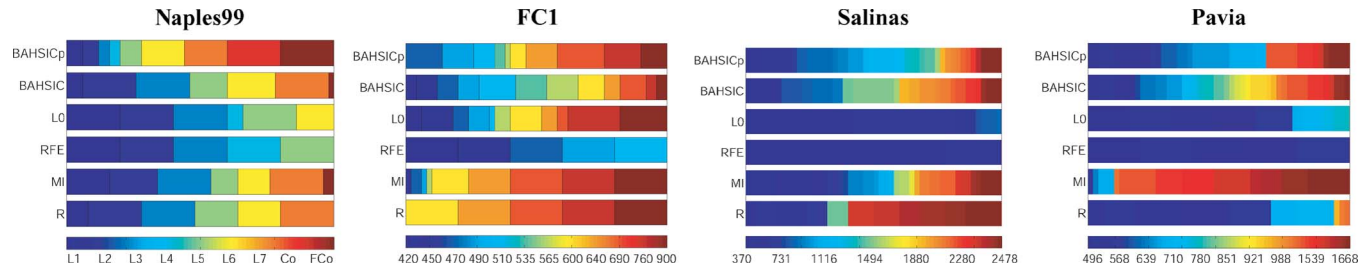

Fig. 3. Average frequency (over ten realizations) in the selection of the top five features by each method and image data set. The bigger the extension of a color in the horizontal bars, the more frequent is the selection of those particular features by a method. The bottom color bars indicate either the features (Naples99) or the wavelength (in nanometers).

soil moisture and leaf water content. Finally, for the case of the Pavia image, BAHSIC$_p$ uniformly covers the whole spectrum, trying to adapt to the high diversity of natural and man-made covers present in the scene. Interestingly, MI finds bands in the range 1000–1600 nm highly dependent on the class labels, suggesting a biased selection toward the detection of the majoritary classes (water and trees) in the image.

V. Discussion and Conclusion

This letter has presented a kernel method for remote sensing feature selection based on measuring nonlinear dependence between the spectral bands and the class labels. The method has very good theoretical and practical properties. We proposed to remove features by minimizing the HSIC p-value rather than maximizing the HSIC value itself between features and class labels. We tested the method in multisource, multispectral, and hyperspectral image-classification problems, both under urban and agricultural areas. The proposed method outperformed other standard and advanced methods, both filter and wrappers. The good numerical performance was complemented with the interpretability of the results. Kernel-based methods are particularly well suited for cases with relatively low number of labeled samples per dimension. For higher numbers of labeled samples, other methods for feature selection, such as filtering using MI, can be more efficient while yielding comparable performance. Further research is required to investigate the influence of the kernel width on the BAHSIC$_p$ ranked features and to adapt the HSIC for nonindependent identically distributed samples which is the case in image processing.

Acknowledgment

The authors would like to thank Dr. D. Landgrebe, Dr. C. Biehl, Dr. A. Gualtieri, and Dr. P. Gamba for kindly providing the data sets used in this work.

References

[1] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1/2, pp. 245–271, Dec. 1997.
[2] R. Kohavi and G. H. John, "Wrappers for features subset selection," *Artif. Intell.*, vol. 97, no. 1/2, pp. 273–324, Dec. 1997.
[3] S. B. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 484–495, Feb. 2007.
[4] B. Paskaleva, M. M. Hayat, Z. Wang, J. S. Tyo, and S. Krishna, "Canonical correlation feature selection for sensors with overlapping bands: Theory and application," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 3346–3358, Oct. 2008.
[5] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, Jul. 2001.
[6] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 674–679, Oct. 2007.
[7] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, Nov. 2006.
[8] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *Proc. Algorithmic Learn. Theory*, S. Jain and W.-S. Lee, Eds., 2005, pp. 63–77.
[9] L. Song, A. J. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo, "Supervised feature selection via dependence estimation," in *Proc. Int. Conf. Mach. Learn.*, C. Sammut and Z. Ghahramani, Eds., 2007, pp. 823–830.
[10] L. Song, J. Bedo, K. M. Borgwardt, A. Gretton, and A. J. Smola, "Gene selection via the BAHSIC family of algorithms," *Bioinformatics (ISMB)*, vol. 23, no. 13, pp. i490–i498, Jul. 2007.
[11] G. Camps-Valls and L. Bruzzone, Eds., *Kernel Methods for Remote Sensing Data Analysis*. London, U.K.: Wiley, Nov. 2009.
[12] L. Song, "Learning via Hilbert space embedding of distributions," Ph.D. dissertation, School Inf. Technol., Univ. Sydney, Sydney, Australia, 2008.
[13] C. Baker, "Joint measures and cross-covariance operators," *Trans. Amer. Math. Soc.*, vol. 186, pp. 273–289, Dec. 1973.
[14] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces," *J. Mach. Learn. Res.*, vol. 5, pp. 73–99, 2004.
[15] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions Extraction: Foundations and Applications*, 2nd ed. Hoboken, NJ: Wiley, 1994.
[16] I. Steinwart, "On the influence of the kernel on the consistency of support vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 67–93, 2001.
[17] L. Gómez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila-Francés, and G. Camps-Valls, "Urban monitoring using multitemporal SAR and multispectral data," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 234–243, Mar. 2006.
[18] Q. Jackson and D. A. Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 12, pp. 2664–2679, Dec. 2001.
[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002.
[20] J. Weston, A. Elisseeff, B. Scholkopf, M. Tipping, and P. Kaelbling, "Use of the zero-norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, 2003.
[21] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–663, May 2004.
[22] A. Statnikov, D. Hardin, and C. Aliferis, "Using SVM weight-based methods to identify causally relevant and non-causally relevant variables," in *Proc. NIPS. Workshop Causality Feature Selection*, 2006, pp. 129–150.