Unsupervised hyperspectral band selection via multi-feature information-maximization clustering

Marco Bevilacqua, Yannick Berthoumieu

UNSUPERVISED HYPERSPECTRAL BAND SELECTION VIA MULTI-FEATURE INFORMATION-MAXIMIZATION CLUSTERING

Marco Bevilacqua, Yannick Berthoumieu

Université de Bordeaux, Bordeaux INP, Laboratoire IMS, UMR 5218, F-33405 Talence, France.

ABSTRACT

This paper presents a new approach for unsupervised band selection in the context of hyperspectral imaging. The hyperspectral band selection (HBS) task is considered as a clustering problem: bands are clustered in the image space; one representative image is then kept for each cluster, to be part of the set of selected bands. The proposed clustering method falls into the family of information-maximization clustering, where mutual information between data features and cluster assignments is maximized. Inspired by a clustering method of this family, we adapt it to the HBS problem and extend it to the case of multiple image features. A pixel selection step is also integrated to reduce the spatial support of the feature vectors, thus mitigating the curse of dimensionality. Experiments with different standard data sets show that the bands selected with our algorithm lead to higher classification performance, in comparison with other state-of-the-art HBS methods.

Index Terms— Hyperspectral imaging, band selection, dimensionality reduction, clustering, classification

1. INTRODUCTION

Hyperspectral imaging offers the possibility of acquiring what is named a *hyperspectral cube*, i.e. a collection of images corresponding to different spectral bands. Typically, the number of images composing a cube is in the range of one or two hundreds, i.e. each pixel of a hyperspectral image can be seen as a high-dimensional vector. As a consequence, the computational complexity when processing and analyzing such data can be very high. Moreover, a certain information redundancy, due to highly correlated bands, may even decrease the accuracy of algorithms exploiting hyperspectral data. Therefore, methods that efficiently select a significant subset of spectral bands are welcome. An extra advantage of such methods is that the selection of few bands would enable the design of less expensive ad-hoc multispectral systems (3 to 10 bands).

Hyperspectral band selection (HBS) methods are typically divided into unsupervised and supervised methods, the latter making use of labeled samples. Unsupervised HBS methods are more robust, in the sense that they offer a way to select a chosen number of bands given any input cube (regardless of the presence and quality of labeled samples). Unsupervised HBS methods are further classified into two main categories: ranking-based and clustering-based methods. The former ones aim at ranking the bands according to a certain metric that quantifies their importance (e.g. [1, 2, 3, 4, 5]). Clustering-based HBS methods [6, 7], by following an optimization strategy, group images into clusters: the bands to select are then collected by choosing a representative image for each cluster. Clustering-based HBS methods are generally showed to be able to reach higher performance, and are typically designed by adapting well-known clustering methods to the hyperspectral case. HBS methods leveraging graph theory (e.g. [8]) can also be included in this category. While there are a few HBS methods relying on information theory principles [9, 6], as far as we know clustering methods fully based on probabilistic modeling have not been studied in this context. In particular, a family of these methods follows the so called *infomax principle* [10, 11], which is related to the maximization of the mutual information between the unknown cluster labels and the data vectors. The key idea is that a good labeling of the images should be in some way informative about the underlying high-dimensional data structure. This is particularly desirable in the case of hyperspectral data, since the clustering should have a certain "spectral consistency".

Recently a new clustering method relying on the infomax principle has been introduced in [12, 13]. The authors propose a kernel model for the posterior class probability and use a variant of Mutual Information (MI) called *squared-loss MI* (SMI), which allows the solution to be computed analytically. The resulting method is referred to as SMI-based clustering (SMIC). In this paper, we propose a new HBS algorithm that builds on SMIC. Besides the adaptation of SMIC to the band selection problem, we propose a new kernel model for the posterior class probability, which involves multiple features. Moreover, we introduce a dedicated pixel selection step, which reduces the number of dimensions to be considered for distance computation. The remainder of the paper is organized as follows. In Section 2 we specify the notation we use and briefly recall the clustering framework for HBS. Section 3 details our proposed clustering-based approach, which relies on the modeling of a multi-feature posterior class probability and an optimization stage via SMI maximization. In Section 4, a quantitative assessment is performed to compare our algorithm with other state-of-the-art HBS methods in terms of classification performance.

2. HYPERSPECTRAL BAND SELECTION VIA CLUSTERING

We denote a hyperspectral cube as a 3-dimensional matrix $\mathbf{H} \in \mathbb{R}^{M \times N \times L}$, i.e. it is composed by L images of dimension $M \times N$. The matrix $\mathbf{H}$ can be further rearranged by stacking all the pixels of one image to form a column of a unique matrix $\mathbf{X} \in \mathbb{R}^{D \times L}$, where $D = M \times N$. HBS consists then in finding a subset of bands $\mathbb{B}$ of cardinality $|\mathbb{B}| = K < L$.

A family of unsupervised HBS methods is based on clustering approaches. All images composing the cube (the columns of the matrix $\mathbf{X}$) are considered as individual points and clustered in the shared D-dimensional space. Clustering-based HBS methods consist of two steps.

This study has been carried out with the financial support from the PharmaSense project funded by the French Program "Fonds Unique Interministériel - Nouvelle Aquitaine".

1. Clustering: all images are clustered in the image space, i.e. each image $\mathbf{x}_i$ is given a label $y_i \in \{1, \ldots, K\}$.
2. Band selection: for each cluster, a unique representant (an image, corresponding to a specific spectral band) is kept.

Clustering-based HBS methods are generally based on existing clustering algorithms, e.g. affinity propagation [7] and Ward's method for agglomerative hierarchical clustering [6]. Once the clusters are formed, the band selection step often consists in picking the cluster centroids. While these methods already allow to select subsets of bands that enable good performance in further tasks (e.g. classification), they lack of a probabilistic interpretation. In the next section we present a clustering-based HBS method that relies on probabilistic modeling.

3. PROPOSED APPROACH

In probabilistic model-based clustering, y is seen as a hidden random variable, which reflects the cluster membership for every point in the data set. After optimizing over the parameters of the model, it is possible to compute a *posterior class probability* $p(y|\mathbf{x})$, which expresses for a data point its probability of belonging to a certain cluster. In the HBS problem we can exploit the computed probabilities to decide on the bands to select. For the k-th cluster, we retain as a representative band the image $\mathbf{x}_i$ exhibiting the highest probability of having the label $y_i = k$ assigned. The selection of all bands, by repeating this procedure for all clusters ($k = 1, \ldots, K$), is given by the following expression:

$$\mathbb{B} = \left\{ \underset{\mathbf{x}_i \in \mathbb{R}^D}{\arg\max}\, p(y_i = k|\mathbf{x}_i) \right\}_{k=1}^{K} . \tag{1}$$

In Section 3.1 we present a posterior class probability model for hyperspectral image clustering. The model includes the use of multiple image features and an internal pixel selection step (Algorithm 1). The parameters of the model are then inferred by maximizing the information between data features and cluster assignments (Section 3.2). Band selection can finally be performed by the rule in (1).

3.1. Posterior class probability model

Most of the probabilistic model-based clustering are *generative models*, where the probability of the data $p(\mathbf{x})$ is defined as a mixture of processes. The parameters of this mixture are generally inferred by maximizing the marginal likelihood $p(\mathbf{x}|y)$. The class labels are then assigned according to the posterior class probability computed as $p(y|\mathbf{x}) \propto p(\mathbf{x})p(\mathbf{x}|y)$. However, when dealing with high-dimensional vectors, this approach basically constrains to use mixtures of Gaussian processes (for their easy tractability), by typically requiring a very large number of components [10]. Instead, we choose to directly model the posterior class probability. At this regard, kernel estimation methods have been proven to effectively estimate densities even when few data points are available [14]. We then choose to model the posterior class probability as a linear combination of kernel functions evaluated at each data point:

$$p(y|\mathbf{x}; \boldsymbol{\alpha}) = \sum_{i=1}^{L} \alpha_{y,i} \mathcal{K}(\mathbf{x}, \mathbf{x}_i) , \tag{2}$$

where $\mathcal{K}$ is a smoothing kernel function, and $\boldsymbol{\alpha}_y$ is a L-dimensional parameter vector characteristic of the cluster labeled as y. We then have a set of K parameter vectors $\{\boldsymbol{\alpha}_y\}_{y=1}^{K}$ that have to be determined in order to quantify the posterior probability (2).

A first contribution to this model relates to the expression of the kernel function $\mathcal{K}$, which implies a prior *pixel selection* step. The function used is a sparse version of the Gaussian local-scaling kernel [15], which scales distances according to each local neighborhood:

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \exp\left(-\dfrac{d_{\mathcal{D}_i}(\mathbf{x}_i, \mathbf{x}_j) d_{\mathcal{D}_j}(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma_i \sigma_j} \right) & d_{\mathcal{D}_i}(\mathbf{x}_i, \mathbf{x}_j) \leq \sigma_i \quad \vee \\ & d_{\mathcal{D}_j}(\mathbf{x}_i, \mathbf{x}_j) \leq \sigma_j \\ 0 & \text{otherwise} \end{cases} , \tag{3}$$

where $\sigma_i = d_{\mathcal{D}_i}(\mathbf{x}_i, \mathbf{x}_i^P)$ and $\sigma_j = d_{\mathcal{D}_j}(\mathbf{x}_j, \mathbf{x}_j^P)$ are the distances from $\mathbf{x}_i$ and $\mathbf{x}_i$ to their P-*th* neighbors ($\mathbf{x}_i^P$ and $\mathbf{x}_j^P$ respectively). In fact, as also stated in [13], P represents a crucial parameter in the definition of the kernel. The pixel selection appears in the expression of the inter-image distances, which are intended as average segmental distances over a reduced support (specific to the reference image considered):

$$d_{\mathcal{D}_i}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k \in \mathcal{D}_i} |x_{ik} - x_{jk}|}{|\mathcal{D}_i|} , \tag{4}$$

where $\mathcal{D}_i$ is the reduced support (the subset of selected pixels) specific to the image $\mathbf{x}_i$ and the double subscript notation indicates a particular pixel (i.e. x_{ij} is the j-th pixel of the image $\mathbf{x}_i$). The idea of performing pixel selection comes from the following remark: when clustering high-dimensional data, only a fraction of all dimensions are locally pertinent [16]. In our case, we decide to select the pixels which are the most stable "locally" (within neighboring bands), i.e. their values vary the least. For a given image $\mathbf{x}_i$, we consider a set of spectrally neighboring bands and compute for each image pixel a score z_{ij} measuring local changes. The pixels related to the D_{tgt} lowest scores are finally retained, where D_{tgt} is a desired target number of dimensions. The pixel selection procedure is reported in Algorithm 1. The algorithm takes as parameters D_{tgt} and Δ (the radius of the local neighborhood), and returns all selected supports $\{D_i\}_{i=1}^{L}$ to be used in (3).

Algorithm 1 Find the best local supports for kernel computation.

procedure PIXELSELECTION($\mathbf{X} \in \mathbb{R}^{D \times L}$, D_{tgt}, Δ)
  for $i = 1, \ldots, n$ **do**
    $\mathcal{N}_i = \{\mathbf{x}_j\}_{|i-j|<\Delta,\ i \neq j,\ 1 \leq j \leq L}$ ▷ neighboring bands
    $\mathbf{z}_i = \dfrac{1}{|\mathcal{N}_i|} \displaystyle\sum_{\mathbf{x}_j \in \mathcal{N}_i} |\mathbf{x}_i - \mathbf{x}_j|$
    $\mathcal{D}_i \leftarrow D_{\text{tgt}}$ first indexes (lowest values of $\mathbf{z}_i$)
  end for
end procedure

A second contribution to the model (2) relates to the extension to multiple image features. For a given image $\mathbf{x}_i$, besides considering the simple reflectance intensity (the collection of all pixel intensities, which we can refer to as $\mathbf{x}_i^r \in \mathbb{R}^D$), we can also consider spatial and spectral features. These features are intended to account for spatial and spectral "texture" and can be derived from gradient-type operations. If we consider per-pixel features, for an image we then have two extra feature vectors $\mathbf{x}_i^s, \mathbf{x}_i^w \in \mathbb{R}^D$, referring to spatial and spectral features respectively. Overall, we have that a hyperspectral cube yields three matrices ($\mathbf{X}^r, \mathbf{X}^s, \mathbf{X}^w \in \mathbb{R}^{D \times L}$), related to the three types of features considered. To combine these multiple features in the posterior kernel model (2), we adopt the *generalized composite kernel* (GCK) approach [17]. For a pair of images, instead of considering a convex combination of kernels, we simply concatenate the kernel outputs computed individually w.r.t. three feature vectors:

$$\mathcal{K}_C(\mathbf{x}_i, \mathbf{x}_j) := \left[\mathcal{K}(\mathbf{x}_i^r, \mathbf{x}_j^r), \mathcal{K}(\mathbf{x}_i^s, \mathbf{x}_j^s), \mathcal{K}(\mathbf{x}_i^w, \mathbf{x}_j^w) \right]^{\mathsf{T}} . \tag{5}$$

If we consider kernel distance matrices storing all inter-image distances for all features, $\mathbf{K}_r, \mathbf{K}_s, \mathbf{K}_w \in \mathbb{R}^{L\times L}$, this is equivalent to vertically concatenating the three matrices to form a unique kernel matrix

$$\mathbf{K}_{\mathrm{C}} = [\mathbf{K}_r, \mathbf{K}_s, \mathbf{K}_w]^{\mathsf{T}} \in \mathbb{R}^{3L\times L} \ . \tag{6}$$

With the multi-feature composite kernel, the posterior class probability model in (2) formally becomes:

$$p(y|\mathbf{x};\boldsymbol{\alpha}) = \sum_{i=1}^{L} \alpha_{y,i} \mathcal{K}_C(\mathbf{x},\mathbf{x}_i) \ , \tag{7}$$

where now $\alpha_{y,i} \in \mathbb{R}^3$, i.e. each per-cluster parameter vector to estimate has $3L$ values ($\boldsymbol{\alpha}_y \in \mathbb{R}^{3L}$).

Unlike kernel computation via convex combination, the GCK approach does not require extra mixing parameters, but an optimal combination is indirectly found by optimizing the parameters of (7).

3.2. Squared-Loss Mutual Information Maximization

As a strategy to find the parameters of the model (7), we choose to maximize the information between cluster labels and feature vectors. The metric used is the squared-loss mutual information (SMI) [13]:

$$\mathrm{SMI} := \frac{1}{2}\int \sum_{y=1}^{K} p(\mathbf{x})p(y) \left(\frac{p(\mathbf{x},y)}{p(\mathbf{x})p(y)} - 1\right)^2 \mathrm{d}\mathbf{x} \ . \tag{8}$$

After some manipulations of (8), we obtain:

$$\begin{aligned} \mathrm{SMI} = & \ \frac{1}{2}\int \sum_{y=1}^{K} \frac{1}{p(y)} \left[p(y|\mathbf{x})\right]^2 p(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \frac{1}{2} \\ \simeq & \ \frac{1}{2n} \sum_{j=1}^{L}\sum_{y=1}^{K} \frac{1}{p(y)} \left[p(y|\mathbf{x}_j)\right]^2 - \frac{1}{2} \end{aligned} \ , \tag{9}$$

where $p(y)$, if unknown, can be assumed to follow the uniform distribution (i.e. $p(y) = 1/K \ \forall y$). If we now plug in (9) the posterior model of (7), we have:

$$\mathrm{SMI} \simeq \frac{1}{2L} \sum_{y=1}^{K} \frac{1}{p(y)} \boldsymbol{\alpha}_y^{\mathsf{T}} \mathbf{K}_{\mathrm{C}} \mathbf{K}_{\mathrm{C}}^{\mathsf{T}} \boldsymbol{\alpha}_y - \frac{1}{2} \ , \tag{10}$$

where $\mathbf{K}_{\mathrm{C}}$ is the composite kernel matrix defined in (6).

With the expression found for the SMI, we can define the optimization problem to find the parameter vectors of the model as follows:

$$\{\hat{\boldsymbol{\alpha}}_y\}_{y=1}^{K} = \underset{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K}{\arg\max} \sum_{y=1}^{K} \boldsymbol{\alpha}_y^{\mathsf{T}} \mathbf{K}_{\mathrm{C}} \mathbf{K}_{\mathrm{C}}^{\mathsf{T}} \boldsymbol{\alpha}_y \ . \tag{11}$$

Every product $\boldsymbol{\alpha}_y^{\mathsf{T}} \mathbf{K}_{\mathrm{C}} \mathbf{K}_{\mathrm{C}}^{\mathsf{T}} \boldsymbol{\alpha}_y^{\mathsf{T}}$, under the constraint $\|\boldsymbol{\alpha}_y\| = 1$, is the so-called Rayleigh quotient [18]. Finding K vectors that maximize the sum in (11) corresponds then to taking the first K normalized eigenvectors of the square matrix $\mathbf{K}_{\mathrm{C}}\mathbf{K}_{\mathrm{C}}^{\mathsf{T}}$, $\{\boldsymbol{\phi}_y\}_{y=1}^{K}$. As in [13], we impose the eigenvectors to be mutually orthogonal ($\boldsymbol{\phi}_y^{\mathsf{T}}\boldsymbol{\phi}_z = 0$) and, the sign being arbitrary, to sum up to a positive number.

After deriving the parameters of the model, we can finally compute the posterior class probabilities for each image. From (7), after rounding up possible negative outputs and normalizing, we obtain:

$$p(y|\mathbf{x}_i) = \frac{\max\left(0, \boldsymbol{\phi}_y^{\mathsf{T}}(\mathbf{K}_{\mathrm{C}})_i\right)}{\max\left(\mathbf{0}_L, \boldsymbol{\Phi}^{\mathsf{T}}(\mathbf{K}_{\mathrm{C}})_i\right)^{\mathsf{T}} \mathbf{1}_L} \tag{12}$$

where $(\mathbf{K}_{\mathrm{C}})_i$ indicates the i-th column of $\mathbf{K}_{\mathrm{C}}$ and $\boldsymbol{\Phi}$ is the matrix with all eigenvectors as columns. Band selection is then possible by following the criterion reported in (1).

4. EXPERIMENTAL RESULTS

In this section we evaluate our algorithm in terms of classification performance. In particular, we compare with:

- a variance-based ranking approach (the K bands reporting the highest variance values are selected) [1];
- the well-known K-means clustering method [19];
- the LCMV-BCM (Linearly constrained minimum variance with band correlation minimization) ranking-based method [2];
- the recent *E-FDPC* algorithm proposed in [5], a hybrid clustering-ranking method that defines cluster centers as local maxima of the data point density (empirically estimated);
- the case, indicated as "FULL", where band selection is not performed, but all bands are considered.

As for our proposed algorithm, we consider both a complete version including additional spatial and spectral features, and a simplified version considering only reflectance intensity features. We denote them as "PROPOSED (multi feat.)" and "PROPOSED (only int.)", respectively. For the spatial characterization, we consider single-scale *Laplacian of Gaussian (LoG)* features, where only one Gaussian filter with size 5×5 and standard deviation $\sigma = 0.5$ is applied. As for the spectral features we instead consider an average of the finite differences, computed on a local portion of the spectrum. For the n-th pixel of the image x_i, we have:

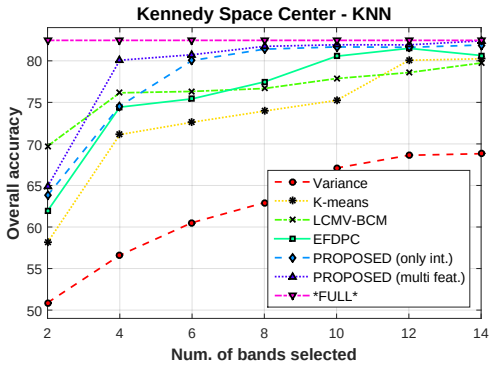
$$x_{i,n}^{w} := \frac{1}{2\tau} \sum_{\substack{-\tau\leq k\leq\tau \\ k\neq 0}} |x_{i,n} - x_{i+k,n}| \ , \tag{13}$$

with $\tau = 2$ taken in our experiments to define the local spectrum.

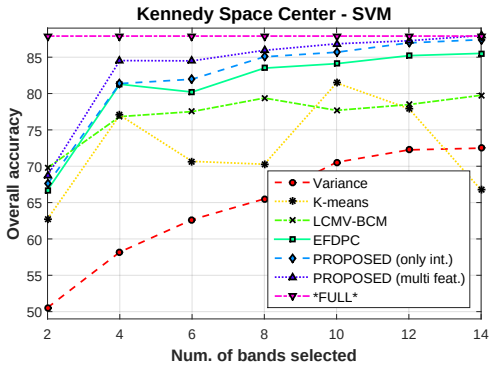
The posterior kernel model presented in Section 3.1 needs 3 parameters to be chosen. The first one is the reference neighborhood taken to determine the scaling distances σ_i and σ_j in (3). In [13], the reference number of P is chosen after estimation of the squared mutual information (SMI) via an external algorithm. We instead choose P to be proportional to the ratio between number of bands and number of clusters, i.e. the average number of images per cluster: $T = \beta\frac{L}{K}$. In the experiments we take a constant $\beta \in [0.5, 0.6]$. Other two parameters are related to the pixel selection procedure of Algorithm 1. In the experiments we choose Δ (the radius of the local neighborhood considered) equal to 3, and D_{tgt} (the target number of selected pixels) equal to 1000.

The proposed approach is compared to the state-of-the-art methods listed above with two popular benchmark data sets[1] : the *Pavia University* data set, consisting in hyperspectral data of resolution $610 \times 340 \times 103$ acquired with the ROSIS-03 sensor, and the *Kennedy Space Center* data set, consisting in hyperspectral data of resolution $512 \times 614 \times 176$ acquired with the AVIRIS sensor. The data sets are provided with ground-truth pixel-wise classification maps. We divide the image pixels into training and testing sets. We then consider the case where the complete spectral signature is used and the one where the spectral vectors have been reduced after HBS, performed with any of the listed methods, including ours. The performance is measured in terms of *overall accuracy* for different numbers of selected bands. The classifiers considered are *K Nearest Neighbors (KNN)* with 5 neighbors considered, and *Support Vector Machines (SVM)* with Gaussian radial basis kernel. The results of such tests are reported in Fig. 1 and Fig. 2, for the *Kennedy Space Center* and *Pavia University* datasets respectively.

[1]The two data sets used can be publicly downloaded at the following *URL*: `http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes`.

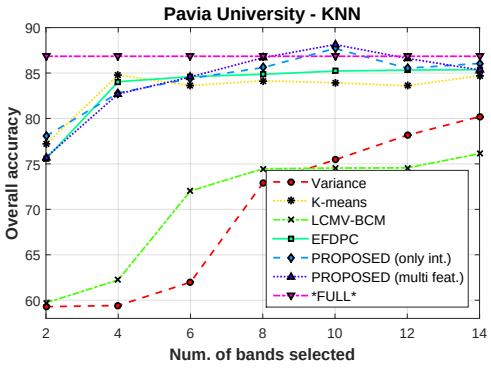


(a) KNN

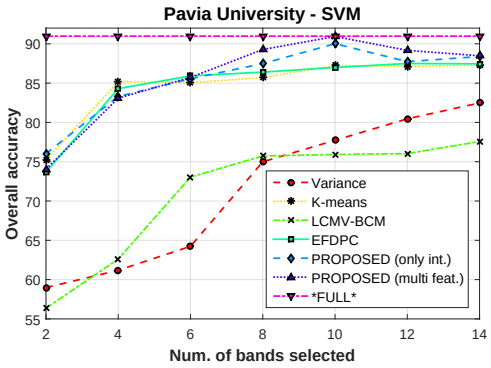


(b) SVM

Fig. 1. Performance (overall accuracy) on the *Kennedy Space Center* data set versus number of bands selected, for different band selection methods and two classifiers: KNN (a) and SVM (b).



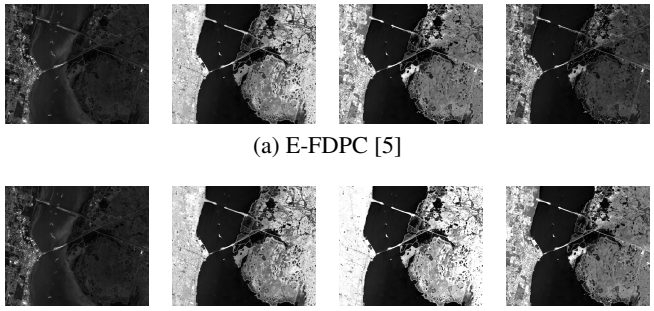
(a) KNN



(b) SVM

Fig. 2. Performance (overall accuracy) on the *Pavia University* data set versus number of bands selected, for different band selection methods and two classifiers: KNN (a) and SVM (b).

As the curves, show, our approach generally outperforms the other state-of-the-art-methods. Starting from a certain value of K, the subset of bands selected with our method leads to a higher classification accuracy. The gap is particularly significant w.r.t. simple ranking-based methods such as [1, 2], which perform poorly. In comparison to the full-spectrum case, the performance is satisfactory too: $K \simeq 10$ selected bands are generally sufficient to obtain a similar performance. Moreover, we observe that the introduction of additional spatial and spectral features brings an extra, more or less appreciable, gain, with respect to the only use of intensity features. Fig. 3 reports an example of selected bands ($K = 4$) from the *Kennedy Space Center* data set, with the *E-FDPC* methods and with our proposed method.



(a) E-FDPC [5]

(b) Proposed (multiple features)

Fig. 3. Example of selected bands ($K = 4$) from the *Kennedy Space Center* data set, with the *E-FDPC* method (bands # 23, 51, 121, 147) and our proposed method (bands # 22, 52, 75, 120).

5. CONCLUSION

In this paper we presented a new clustering-based algorithm for unsupervised hyperspectral band selection (HBS). As a novelty in the HBS context, we adopt a probabilistic model-based clustering approach, where the posterior class probability is directly modeled. The parameters of the model are inferred by maximizing the information between data features and cluster assignments. A canonical posterior model is enriched by considering, besides intensity features, additional spatial and spectral features, which are integrated via a generalized composite kernel approach. Prior to the kernel matrix computation, an ad-hoc pixel selection step is also proposed, to reduce the image support for distance computation. The algorithm is evaluated, in terms of classification performance, with standard data sets, by varying the number of selected bands. Results prove the effectiveness of the proposed approach, which outperforms other state-of-the-art methods, and the general usefulness of the HBS task. 10 bands are in fact generally sufficient to match the performance with full-spectrum data. Future work will concern the study of different image features and optimization strategies, while remaining in the framework of probabilistic model-based clustering.

6. REFERENCES

[1] Claudio Conese and Fabio Maselli, "Selection of optimum bands from TM scenes through mutual information analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 48, no. 3, pp. 2–11, June 1993.

[2] Chein-I Chang and Su Wang, "Constrained Band Selection for Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1575–1585, June 2006.

[3] Baofeng Guo, Steve R Gunn, RI Damper, and JDB Nelson, "Band Selection for Hyperspectral Image Classification Using Mutual Information," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 4, pp. 522–526, Oct. 2006.

[4] Qian Du and He Yang, "Similarity-Based Unsupervised Band Selection for Hyperspectral Image Analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 4, pp. 564–568, Oct. 2008.

[5] Sen Jia, Guihua Tang, Jiasong Zhu, and Qingquan Li, "A Novel Ranking-Based Clustering Approach for Hyperspectral Band Selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 88–102, Jan. 2016.

[6] Aldolfo Martínez-Usó, Filiberto Pla, José Martínez Sotoca, and Pedro García-Sevilla, "Clustering-Based Hyperspectral Band Selection Using Information Measures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.

[7] Sen Jia, Zhen Ji, Yuntao Qian, and Linlin Shen, "Unsupervised Band Selection for Hyperspectral Imagery Classification Without Manual Band Removal," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 531–543, Apr. 2012.

[8] Guokang Zhu, Yuancheng Huang, Jingsheng Lei, Zhongqin Bi, and Feifei Xu, "Unsupervised Hyperspectral Band Selection by Dominant Set Extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 227–239, Jan. 2016.

[9] Claude Cariou, Kacem Chehdi, and Steven Le Moan, "BandClust: An Unsupervised Band Reduction Method for Hyperspectral Remote Sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 565–569, May 2011.

[10] Felix V. Agakov and David Barber, "Kernelized Infomax Clustering," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 18, pp. 17–24. MIT Press, 2006.

[11] Andreas Krause, Pietro Perona, and Ryan G. Gomes, "Discriminative Clustering by Regularized Information Maximization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 23, pp. 775–783. Curran Associates, Inc., 2010.

[12] Masashi Sugiyama, Makoto Yamada, Manabu Kimura, and Hirotaka Hachiya, "On Information-Maximization Clustering: Tuning Parameter Selection and Analytic Solution," in *28th International Conference on Machine Learning*, 2011, pp. 65–72.

[13] Masashi Sugiyama, Gang Niu, Makoto Yamada, Manabu Kimura, and Hir Hachiya, "Information-Maximization Clustering Based on Squared-Loss Mutual Information," *Neural Computation*, vol. 26, no. 1, pp. 84–131, Jan. 2014.

[14] David W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, 2015.

[15] Lihi Zelnik-Manor and Pietro Perona, "Self-Tuning Spectral Clustering," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 17, pp. 1601–1608. MIT Press, 2004.

[16] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park, "Fast Algorithms for Projected Clustering," in *ACM SIGMoD Record*. ACM, jun 1999, vol. 28, pp. 61–72.

[17] Jun Li, Prashanth Reddy Marpu, Antonio Plaza, José M Bioucas-Dias, and Jon Atli Benediktsson, "Generalized Composite Kernel Framework for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 9, pp. 4816–4829, sep 2013.

[18] Roger A. Horn and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[19] Stuart P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.