ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Multi-view feature extraction for hyperspectral image classification

Michele Volpi[1,*], Giona Matasci[1], Mikhail Kanevski[1], Devis Tuia[2].

[1] Université de Lausanne - Centre de Recherche en Environnmenent Terrestre
UNIL-Mouline, 1015 Lausanne - Switzerland

[2] Ecole Polytechnique Fédérale de Lausanne - Laboratoire des Systèmes
d'Information Géographique - EPFL, 1015 Lausanne - Switzerland

Abstract. We study the multi-view feature extraction (MV-FE) framework for the classification of hyperspectral images acquired from airborne and spaceborne sensors. This type of data is naturally composed by distinct blocks of spectral channels, forming the hypercube. To reduce the dimensionality of the data by taking advantage of this particular structure, an unsupervised multi-view feature extraction method is applied prior to classification. First, a technique to automatically obtain the blocks, based on the global spectral correlation matrix, is applied. Then, the kernel canonical correlation analysis is performed in a multi-view setting (MV-kCCA) to find projections of the data blocks in a correlated subspace, gaining thus discriminant power. Experiments using the linear discriminant classifier (LDA) show the appropriateness of adopting a MV-FE approach.

1 Introduction

Hyperspectral images acquired by airborne and spaceborne sensors have been extensively used in Earth observation applications, thanks to the detailed information about the energy reflected or emitted by the different ground covers. The use of hyperspectral data has been successfully documented, among others, in urban monitoring, forest biomass estimation, crop and cultivation assessment, mineral and geological exploration, target and hotspot detection [1]. The number of spectral channels available for the detailed analysis of the materials is very large: the dimensionality of hyperspectral data may range from dozens to thousands variables (spectral bands) and it can prevent the successful application of standard pattern recognition techniques, in particular under small sample size situations. This is known as the curse of dimensionality [2]. To avoid these adverse effects on many learning systems, it is common to apply as a preprocessing step, one among the many existing feature extraction (FE) or dimensionality reduction (DR) techniques [3]. The benefits of FE methods such as, to name a few, principal component analysis (PCA), its kernel extension (kPCA), partial least squares and its nonlinear variants (PLS, NIPALS or kernel PLS) and canonical correlation analysis (CCA), are well documented [4].

[*] This project has been supported by the Swiss National Science Foundation grant No. 200020-144135 "kernelCD II" and PZ00P2-136827. More information: www.kernelCD.org. Email: michele.volpi@unil.ch.

The underlying idea of feature extraction is that the most of a specific data characteristic, usually coded by a statistic, can be maximally preserved while reducing the data dimensionality and removing noisy or uninteresting subspaces. This criterion defines the type of information contained in the retained (sub)space in which the data are projected. For instance, in the PCA one looks for the directions maximizing the variance of the original data, with PLS for the covariance among two sets and with CCA for a joint mapping that maximizes the empirical correlation. In parallel, the kernelization of these methods has been considered to implicitly work in a higher dimensional reproducing kernel Hilbert space (RKHS) providing nonlinear solutions in the original space.

Approaches aiming at jointly combining different views of the same examples are known as multi-view learning (MV) methods [5]. In this framework, one looks for an intrinsic combination of the disjoint feature sets to optimally solve the task at hand, such as clustering, classification or, as in our case, FE. In this work, the multi-view kernel CCA (MV-kCCA) [6, 7] is introduced for dimensionality reduction of hyperspectral data, by accounting for the multiple views that naturally compose such data. First, on the basis of the block structure of the correlation matrix among spectral bands, distinct subsets of channels are automatically selected via clustering [8]. The MV-FE step is then performed using these distinct structures as disjoints feature sets describing the common (paired) examples, the pixels. The underlying assumption is that, by separately considering blocks of strongly correlated spectral bands, the FE step can discover useful nonlinear projections for classification, hidden in the within-block variability versus the usually much higher between-blocks covariance.

Experiments on real world hyperspectral data show promising results for the MV-FE to reduce the dimensionality before linear classification using LDA [7].

2 The multi-view feature extraction system

2.1 View generation

To take advantage of the block structure of the hyperspectral data in a MV-FE context, the complete hypercube should be decomposed in disjoint sub-blocks. In the MV literature, it is common to assume that the views satisfy criteria of sufficiency, i.e. contain enough information to guarantee stable learning, consistency (labelings agree along the views) and independence, often relaxed to more realistic situations [5]. In the case of hyperspectral data, spectral bands form distinct sets depending on the sensed ground materials.

To automatically obtain a partitioning of the bands, the correlation matrix between spectral channels is clustered using k-means. Groups showing a low within cluster variations and large distance to the other clusters are automatically selected by the partitioning, thus implicitly grouping the linear dependencies between channels. For instance, this scheme has been adopted in [8] to perform active learning based on the disagreement of a committee of classifiers on the views. To automatize the process, the number of band clusters (views) to be discovered in the partitioning is given by $|\lambda_q|$, where λ_q is the set of the sorted eigenvalues of the correlation matrix explaining a given amount of variance, given by ϵ [9].

2.2 The multi-view canonical correlation analysis

The extension of the CCA to multiple sets has been first proposed in [10], and it finds nowadays different applications [6, 11]. The idea is to find automatically a series of mappings of the input blocks in a subspace maximizing the correlation among them. The standard CCA (two views, $k = 1, 2$) can thus be solved by the $\mathbf{w}_k$ maximizing

$$\rho = \frac{\mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2}{\sqrt{(\mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_1 \mathbf{w}_1)}\sqrt{(\mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{w}_2)}} = \frac{\mathbf{w}_1^T \mathbf{C}_{12} \mathbf{w}_2}{\sqrt{\mathbf{w}_1^T \mathbf{C}_{11} \mathbf{w}_1}\sqrt{\mathbf{w}_2^T \mathbf{C}_{22} \mathbf{w}_2}},$$

where $\mathbf{X}_1 \in \mathbb{R}^{n \times d_1}$, $\mathbf{X}_2 \in \mathbb{R}^{n \times d_2}$, $\mathbf{X}_1^T \mathbf{X}_2 = \mathbf{C}_{12} \in \mathbb{R}^{d_1 \times d_2}$. $\mathbf{X}_1$ and $\mathbf{X}_2$ are two different mean-centered feature sets (views) of the same n examples. Since we are interested only in the directions of $\mathbf{w}_k$, we can reformulate the problem, by setting $\mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_1 \mathbf{w}_1 = \mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{w}_2 = 1$, as [4, 7]:

$$\begin{aligned} \{\mathbf{w}_1, \mathbf{w}_2\} = \underset{\mathbf{w}_1, \mathbf{w}_2}{\arg\max} \quad & \mathbf{w}_1^T \mathbf{C}_{12} \mathbf{w}_2 \\ \text{s.t.} \quad & \mathbf{w}_1^T \mathbf{C}_{11} \mathbf{w}_1 = \mathbf{w}_2^T \mathbf{C}_{22} \mathbf{w}_2 = 1 \end{aligned}$$

and it can be solved by:

$$\begin{pmatrix} \mathbf{0} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = \rho \begin{pmatrix} \mathbf{C}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix}. \tag{1}$$

Following [6, 7], the generalized eigenproblem in Eq. (1) can be rewritten as

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = (1+\rho) \begin{pmatrix} \mathbf{C}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix}. \tag{2}$$

which provides a natural extension to $(2, \ldots, k)$ sets (views). In our case, $\mathbf{X}_k \subseteq \mathbf{X}, \forall\, k$, with $\mathbf{X}_k = \mathbf{X}$ only when considering one view. Here, k must be at least equal to 2 (standard CCA). To obtain a kernel expression, the primal formulation in Eq. (1) is replaced with its dual by plugging $\mathbf{w}_k = \mathbf{X}_k^T \boldsymbol{\alpha}_k$, and by left multiplying by $\begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix}$ [4]. Following same reasoning that brings to Eq. (2), the MV-kCCA results in [6]:

$$\begin{pmatrix} \mathbf{K}_{11}^2 & \mathbf{K}_{11}\mathbf{K}_{22} & \cdots & \mathbf{K}_{11}\mathbf{K}_{kk} \\ \mathbf{K}_{22}\mathbf{K}_{11} & \mathbf{K}_{22}^2 & \cdots & \mathbf{K}_{22}\mathbf{K}_{kk} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{kk}\mathbf{K}_{11} & \mathbf{K}_{kk}\mathbf{K}_{22} & \cdots & \mathbf{K}_{kk}^2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{pmatrix} =$$

$$\lambda \begin{pmatrix} \mathbf{K}_{11}^2 + \gamma\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22}^2 + \gamma\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{K}_{kk}^2 + \gamma\mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{pmatrix}$$

where $\mathbf{K}_{kk}$ are centered kernel matrices computed on the k-th view, and γ is a user-defined penalization to avoid overfitting of correlations when using kernels inducing very high-dimensional spaces [6]. The projected data view k, or canonical variate, is obtained as $\mathbf{Y}_k = \mathbf{K}_{kk}\boldsymbol{\alpha}_k$.

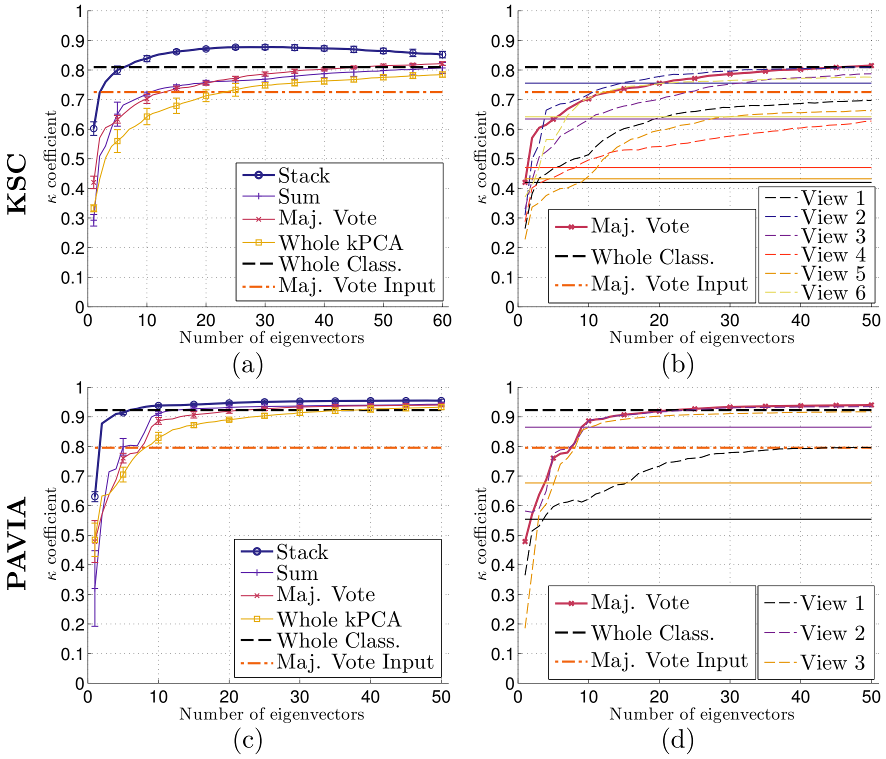
3 Experimental analysis

To test the proposed method, two airborne hyperspectral images are used. The first (KSC) is acquired in 1996 over the natural area of the Kennedy space center (FL, USA) by the AVIRIS sensor (Airborne Visible / Infra Red Image Spectrometer). The sensor acquires 224 bands of 10nm width each, from 400nm to 2500nm. The second dataset (PAVIA) is acquired by the ROSIS-03 (Reflective Optics System Imaging Spectrometer) sensor over the urban area of Pavia (Italy), in the range between 430nm and 860nm with 4nm bandpass for a total of 115 channels. In both cases, water absorption bands and low signal-to-noise ratio are manually removed, resulting in 176 and 102 bands respectively. The classification following the FE step involves different classes of land cover, 13 for KSC and 9 for PAVIA, to be classified using a linear discriminant. The use of a simple and linear classifier is preferred to evaluate the effectiveness of the nonlinear FE step. The view generation approach grouped the KSC and PAVIA datasets in 6 and 3 groups respectively, with a threshold retaining in both cases the 99.9% of the explained variance of the PCA rotation.

The MV-kCCA has been compared to standard classification and after kPCA using the whole image, 'Whole Class.' and 'Whole kPCA' respectively, and to majority voting of the independent classifications of the original spectral blocks ('Maj. Vote Input'). For the MV-kCCA, three different feature combinations are tested. The 'Stack' approach replaces each block of original spectral bands by their lower dimensional representations found by MV-kCCA, i.e. replacing $\mathbf{X}_k$ by $\mathbf{Y}_k$. Thus, the classification is carried out in a space with dimensions equal to the number of eigenvectors extracted times the number of views k. The 'Sum' approach additions, for each pixel, the projections of each block after normalization, i.e. substituting $\mathbf{X}$ by $\sum_k \mathbf{Y}_k$ [12]. The 'Maj. Voting' consists in simple unweighted voting over the independently classified projected views.

The regularization parameter of the MV-kCCA has been set empirically to $\gamma = 0.1$ for both datasets. The RBF kernel bandwidth (for the MV-kCCA and kPCA) has been fixed as the median Euclidean distance among the samples in each view. To compute kernels, 500 randomly chosen pixels among the unlabeled samples have been used in each case study. The linear discriminant has been trained with 50 example pairs per class for both datasets, while the remaining 4561 (KSC) and 147'702 (PAVIA) assess the generalization accuracy. Experiments are repeated 10 times using independent initializations.

In Fig. 1(a),(c) the average Cohen's κ coefficient of agreement is plotted with their respective standard deviations, for the KSC and PAVIA dataset respectively, versus the growing number of eigenvectors used. The advantages of replacing spectral blocks by their respective canonical variates appears imme-



(a) (b) (c) (d)

Fig. 1: (a),(c) Classification accuracy for the tested setups along with the standard deviations, for KSC and PAVIA datasets. In (b),(d) classification and voting for the independent views for KSC and PAVIA datasets. In (b),(d) the horizontal solid lines represent the accuracy of the views classified independently.

diately. Interestingly, for both datasets by using the first 6 canonical variates of each block (thus classifying the KSC and PAVIA datasets in a 36 and a 18 dimensional space), the accuracy of the 'Whole class.' is reached. The kPCA reaches the 'Whole Class.' accuracy when using more than 70 (KSC) and 37 (PAVIA) kernel principal components. For the 'Sum' approach, 60 dimension (KSC) and 15 (PAVIA) are needed. The highest accuracies for KSC and PAVIA data are obtained when stacking the projections of the views issued from 30 and 15-20 (plateau effect) eigenvectors respectively. The standard deviations of the accuracy of tested approaches are very low thanks to the stability of both the classifier and the MV-kCCA.

In Fig. 1(b),(d) accuracies for single view classification are plotted. The one for 'View 2' equals the 'Whole Class' performance for both datasets, after the projection into a (nonlinear) 40 dimensional subspace for the KSC (originally composed by 31 spectral bands) and 22 for PAVIA (originally including 36 channels). The corresponding majority voting behaves similarly. Furthermore, note that the classification of feature blocks mapped into a subspace of (approximately) 10 dimensions always outperforms the classification of the original ones. It is also worth noting that, generally, the ordering of the curves of the mapped blocks for a given dimensionality is the same as the one of the original groups.

4 Conclusions and future work

It has been illustrated that when disposing of a high dimensional hyperspectral image, the MV-kCCA is a valid alternative to single-view traditional unsupervised nonlinear FE techniques. In particular, by considering disjoint blocks of correlated features, accuracies higher than original data classification can be obtained (also when using a comprehensive lower dimensional data representation).

Future developments will consider a classification oriented MV feature extraction framework, in which the available labeled information is included to learn discriminant subspaces. Attention will be paid to particular case of the Fisher's discriminants and their kernel extension. Also, to approach the MV-FE task from a semi-supervised perspective, manifold-based regularization will be considered.

References

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sens. Env.*, 113:S110–S122, 2009.

[2] G.F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory*, 14(1):55–63, 1968.

[3] A. J. Izenman. *Modern Multivariate Statistical Techniques.* Springer Texts in Statistics. Springer, 2008.

[4] T. De Bie, N. Cristianini, and R. Rosipal. *Handbook of computational geometry for pattern recognition, computer vision, neurocomputing and robotics*, chapter Eigenproblems in pattern recognition, pages 129–171. Springer Verlag, 2004.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th annual conference on computational learning theory (COLT)*, pages 92–100, 1998.

[6] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.

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

[8] W. Di and M. M. Crawford. View generation for multiview maximum disagreement based active learning for hyperspectral classification. *IEEE Trans. Geosci. Remote Sens.*, 50(5):1942–1954, 2012.

[9] M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Trans. Neural Net.*, 13(3):780–784, 2002.

[10] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[11] M. McVicar and T. De Bie. CCA and a multi-way extension for investigating common components between audio, lyrics and tags. In *Proceedings of the 9th international symposium on computational music modeling and retrieval (CMMR)*, pages 53–68, 2012.

[12] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia. A new method of feature fusion and its application in image recognition. *Patt. Recogn.*, 38(12):2437–2448, 2005.