RELEVANT AND INVARIANT FEATURE SELECTION OF HYPERSPECTRAL IMAGES FOR DOMAIN GENERALIZATION

Claudio Persello and Lorenzo Bruzzone

Department of Information Engineering and Computer Science, University of Trento
Via Sommarive, 5 I-38123, Povo, Trento, Italy,
e-mail: claudio.persello@disi.unitn.it, lorenzo.bruzzone@ing.unitn.it

ABSTRACT

This paper presents a novel feature selection method for the analysis of hyperspectral images. The proposed method aims at selecting a subset of the original features that are both 1) relevant for the considered problem (i.e., preserve the functional relationship between input and output variables), and 2) invariant (stable) across different domains (i.e., minimize the data set shift among different domains). Domains can be associated with images collected on different areas or on the same area at different times. We propose a novel measure of domain stability, which evaluates the distance of the conditional distributions between the source and target domain. Such a measure is defined on the basis of kernel embeddings of conditional distributions and can be applied to both classification and regression problems. Experimental results show the effectiveness of the proposed method in selecting features with high generalization capabilities on the target domain.

Index Terms— Feature Selection, Kernel methods, Image Classification, Hyperspectral Data, Remote Sensing.

1. INTRODUCTION

Hyperspectral remote sensing images, due to their capability to precisely characterize the spectral signature of the different materials on the ground, represent a very rich source of information for environmental monitoring, land cover mapping and analysis. However, the automatic analysis and classification of hyperspectral data by supervised algorithms is a complex task. One of the main challenges is to deal with the typical small ratio between the number of available training samples and the number of features. In such conditions, the associated classification problems become difficult and can lead to the Hughes phenomenon. Support Vector Machines (SVMs) and kernel-based classification/regression algorithms have shown to be robust to the high dimensionality of the feature space. Nevertheless, also these algorithms may significantly benefit from feature reduction methods.

Another important problem in the analysis of hyperspectral images is the non-stationary behavior of the spectral signatures of the land-cover classes. This is due to possible differences in the image acquisition conditions (e.g., illumination and viewing angle), ground conditions (e.g., soil moisture and topography), or in the phenological stages of vegetation that may affect the observed spectral signatures of the land-cover classes. In order to tackle this issue, domain adaptation methods have been employed, which aim at adapting the classifier trained on an initial source domain (e.g., an hyperspectral image or part of it) to correctly classify a target domain, i.e., a disjoint area of the same image, or another image that is acquired on a different geographical area or at a different time.

In this paper, we present a feature selection technique that jointly addresses both the problems of dimensionality reduction and feature non-stationarity. The goal is to develop a method for selecting a subset of the features (original bands and/or features extracted from them) that are both 1) relevant (e.g., discriminant) for the considered problem, and 2) invariant (stable) among different domains. To this end, we propose a novel feature stability measure based on kernel embeddings of conditional distributions. This measure allows us to asses the invariance of features among different domains and to select the most stable ones.

2. PROPOSED FEATURE SELECTION METHOD

The proposed method explicitly considers in the criterion function of the feature-selection process both 1) the feature *relevance* $\mathcal{R}$, i.e., the functional dependence between the output variable Y (e.g., the information classes) and the input features X, and 2) the *invariance* Θ, i.e, the stability of the selected features across different domains. In an earlier work [1], we addressed such a problem in the context of Gaussian Maximum Likelihood classifiers. Here, we address this problem by proposing a novel feature selection technique that is based on kernel methods and can be adopted both for

The work of Dr. Claudio Persello is supported by the Autonomous Province of Trento and the European Community in the framework of the project "Trentino - PCOFUND-GA-2008-226070 (call 3 - post-doc 2010 Outgoing)"

classification and regression problems. Moreover, we propose a novel domain stability measure that can evaluate the amount of different types of data-set shift. For simplicity, we consider here problems with only two domains. However, the proposed method can be easily extended to deal with multiple domains.

Both the relevance and invariance terms are evaluated considering kernel-based dependence estimators designed by embedding the sample distributions in a reproducing kernel Hilbert space (RKHS). Using kernel embeddings, we can asses the two terms without explicitly estimating the distributions on the two domains. The feature relevance $\mathcal{R}$ is evaluated considering the *Hilbert-Schmidt Independence Criterion (HSIC)* as a measure of dependence between X and Y [2, 3]. The feature invariance Θ is evaluated by means of a stability measure, which asseses the difference in the conditional probabilities $P(X|Y)$ between source and target.

The final subset of features is selected by jointly optimizing the two terms of the criterion function, i.e., feature relevance $\mathcal{R}$ and invariance Θ. This is done by defining a multi-objective search strategy for deriving the subsets of features that exhibit the best trade-off between the two concurrent objectives. The selected subset of features improves the robustness and generalization capability of the classification/regression system to make predictions on the target domain.

2.1. Feature Relevance

The aim of the feature relevance term $\mathcal{R}$ is to estimate the functional dependence between X and Y by using a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. The linear dependence between X and Y can be assessed by considering the cross-covariance matrix. A statistic that effectively represents the content of this matrix is the Hilbert-Schmidt norm. In order to capture also higher-order dependence, an extension of the notion of covariance was proposed in [4] adopting a non-linear mapping for both X and Y to an RKHS. The square of the Hilbert-Schmidt norm of such an extended cross-covariance operator defines the HSIC measure. The empirical estimation of HSIC is used as relevance term $\mathcal{R}$ in the proposed feature selection method:

$$\mathcal{R} = \mathrm{HSIC}(X, Y) = \frac{1}{n^2}\mathrm{Tr}(\mathbf{KHLH}), \tag{1}$$

where Tr is the trace, $\mathbf{K}$ and $\mathbf{L}$ are the kernel matrices for the samples $\mathbf{x}_i$ and their labels y_i, respectively, and $\mathbf{H}_{ij} = \delta_{ij} - (1/n)$ is the centering matrix. Here, δ represents the Kronecker symbol, i.e., $\delta_{ij} = 1$ if $i = j$ and zero otherwise. In our experiments we adopted a Gaussian Radial Basis Function (RBF) for the input kernel $\mathbf{K}$ and a linear function for the output kernel $\mathbf{L}$. The design of the output kernel is further discussed in Section 2.3. Fig. 1 shows an example of input and output kernel matrices. Using characteristic kernels (e.g., Gaussian or Laplacian kernels), it can be shown that HSIC asymptotically approaches zero if and only if X and Y are independent. Conversely, large values of HSIC are associated with strong dependence.

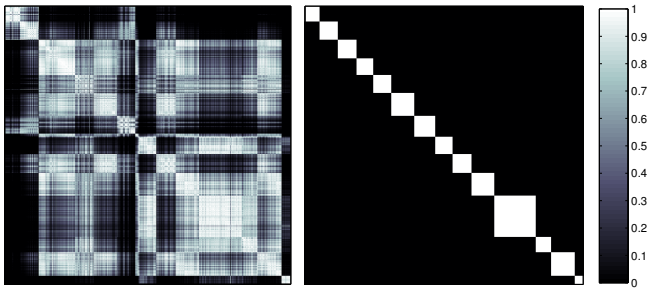

Fig. 1. Example of input and output kernel matrices

2.2. Domain Stability Measure

The feature invariance is assessed by considering the availability of two training sets, one defined on the source domain $\mathcal{D}^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$, and one on the target domain $\mathcal{D}^t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$. We define a stability measure which evaluates the shift between the conditional distributions $P^s(X^s|Y^s)$ and $P^t(X^t|Y^t)$ by means of the their kernel embeddings. In this way, we can compute this distance in a non-parametric fashion and without explicitly estimating the two conditional distributions. The stability term is defined as:

$$\Theta(\mathcal{D}^s, \mathcal{D}^t) = \sum_{(y_i^s, y_j^t) \in Y^s \times Y^t} \|\mu_{X^s|y_i^s} - \mu_{X^t|y_j^t}\|^2 \tag{2}$$

where $\mu_{X^s|y_i^s}$ and $\mu_{X^t|y_j^t}$ are the kernel embeddings of the conditional distributions on the source and target domains, respectively. The invariance term is obtained by summing up the squared norm of the difference between the embeddings of the conditional probabilities for all possible couples (y_i^s, y_j^t) of conditioning variables. Given a generic data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, the empirical estimate of the kernel conditional embedding is [5]:

$$\hat{\mu}_{X|y} = \hat{\mathcal{C}}_{X|Y}\phi(y) = \Upsilon(\mathbf{L} + \lambda\mathbf{I})^{-1}\Phi^{\mathrm{T}}\phi(y)$$

where $\Phi := (\phi(y_1), \ldots, \phi(y_n))$ and $\Upsilon := (\varphi(\mathbf{x}_1), \ldots, \varphi(\mathbf{x}_n))$ are the feature mappings of the variables X and Y, $\mathbf{K} = \Psi^{\mathrm{T}}\Psi$ and $\mathbf{L} = \Phi^{\mathrm{T}}\Phi$ are the input and output kernel matrices, respectively, and λ is a regularization parameter. $\mathbf{I}$ is the identity matrix of appropriate dimension. Furthermore, we have that:

$$\|\hat{\mu}_{X^s|y_i^s} - \hat{\mu}_{X^t|y_j^t}\|^2 = \langle \hat{\mu}_{X^s|y_i^s} - \hat{\mu}_{X^t|y_j^t}, \hat{\mu}_{X^s|y_i^s} - \hat{\mu}_{X^t|y_j^t} \rangle =$$
$$\langle \hat{\mu}_{X^s|y_i^s}, \hat{\mu}_{X^s|y_i^s} \rangle + \langle \hat{\mu}_{X^t|y_j^t}, \hat{\mu}_{X^t|y_j^t} \rangle - 2\langle \hat{\mu}_{X^s|y_i^s}, \hat{\mu}_{X^t|y_j^t} \rangle$$

Let us denote with $\mathbf{L}^s$, $\mathbf{L}^t$, $\mathbf{L}^{st}$ the source, target, and cross-domain output kernel matrices, respectively. We define $\mathbf{K}^s$, $\mathbf{K}^t$, $\mathbf{K}^{st}$ the source, target, and cross-domain input kernel matrices, respectively. Moreover, we define $\mathbf{\Omega}^s := (\mathbf{L}^s + \lambda\mathbf{I})^{-1}$ and $\mathbf{\Omega}^t := (\mathbf{L}^t + \lambda\mathbf{I})^{-1}$. We derive the following equations:

$$\begin{aligned}\langle \hat{\mu}_{X^s|y_i^s}, \hat{\mu}_{X^s|y_i^s} \rangle &= \mathbf{L}^s_{i\cdot}\mathbf{\Omega}^s\mathbf{K}^s\mathbf{\Omega}^s\mathbf{L}^s_{\cdot i} \\ \langle \hat{\mu}_{X^t|y_j^t}, \hat{\mu}_{X^t|y_j^t} \rangle &= \mathbf{L}^t_{j\cdot}\mathbf{\Omega}^t\mathbf{K}^t\mathbf{\Omega}^t\mathbf{L}^t_{\cdot j} \\ \langle \hat{\mu}_{X^s|y_i^s}, \hat{\mu}_{X^t|y_j^t} \rangle &= \mathbf{L}^s_{i\cdot}\mathbf{\Omega}^s\mathbf{K}^{st}\mathbf{\Omega}^t\mathbf{L}^t_{\cdot j}\end{aligned}$$

where $\mathbf{L}_{i\cdot}$ and $\mathbf{L}_{\cdot j}$ denote row i and column j of matrix $\mathbf{L}$, respectively. We can finally obtain the empirical estimate of the data set shift measure as:

$$\begin{aligned}\hat{\Theta}(\mathcal{D}^s, \mathcal{D}^t) = &\sum_{i,j} \mathbf{L}^s \odot \mathbf{\Omega}^s\mathbf{K}^s\mathbf{\Omega}^s + \sum_{i,j} \mathbf{L}^t \odot \mathbf{\Omega}^t\mathbf{K}^t\mathbf{\Omega}^t \\ &- 2\sum_{i,j} \mathbf{L}^{st} \odot \mathbf{\Omega}^s\mathbf{K}^{st}\mathbf{\Omega}^t\end{aligned} \tag{3}$$

where $\odot$ is the Hadamar (element-wise) matrix product.

2.3. Output Kernel Design

The output kernel is a measure of similarity between output variables, and should be designed according to the specific analysis task, i.e., classification or regression. In regression problems, the RBF kernel is usually an effective choice. In the context of classification problems, one possible definition is $\mathbf{L}_{ij} = 1$ if $y_i = y_j$ and zero otherwise. In this particular case, the proposed invariance measure $\hat{\Theta}$ tends to the sum of the *Maximum Mean Discrepancy (MMD)* [4] between samples of the same class that belong to the source and target domain, respectively, i.e., $\hat{\Theta} \rightarrow \sum_{i=1}^{C} mmd(X^s_{\omega_i}, X^t_{\omega_i})$ when λ tends to zero, where $\{\omega_1, \cdots, \omega_C\}$ is the set of information classes. With such a definition of the output kernel, the proposed stability measure evaluates the deviation of the distribution of the classes from the source to the target domain. However, the distance among different classes between source and target is not taken into account.

In many real problems, it is important to capture the case when the shift of one or more class distributions causes overlap with different classes on the source domain. This type of behavior would significantly reduce the classification accuracy and the feature selection method should remove the features that may result in this type of confusion. An explanatory example for a three-class classification problem is reported in Fig. 2. Panel a) reports the case where the data-set shift does not lead to classes overlap. Panel b) reports an example where the class distributions are affected by a similar amount of shift, however the direction of the shift makes some classes overlap with different classes in the source domain. The proposed $\hat{\Theta}$ measure is able to capture this type of instability by designing the output kernel (for instance) as $\mathbf{L}_{ij} = 1$ if $y_i = y_j$ and -1 otherwise.

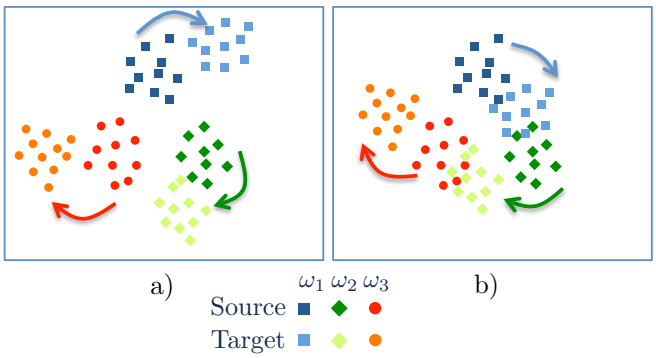

Fig. 2. Examples of different data-set shifts for a three-classes classification problem.

2.4. Multi-objective Search Strategy

The final feature subset Λ of size m is selected by jointly optimizing the two considered terms, i.e., by solving the following multi-objective problem: $\min_{|\Lambda|=m}(-\mathcal{R}, \hat{\Theta})$. Solving this problem consists in finding the set of Pareto optimal solutions (also called Pareto front). In grater detail, a solution is said to be Pareto optimal if there are no other solutions that would improve an objective without simultaneously worsen one of the others. The estimation of the Pareto front can be achieved with different multi-objective optimization algorithms (e.g., multi-objective evolutionary algorithms). In our experiments we adopted a modification of the genetic multi-objective strategy NSGA-II [1, 6].

3. EXPERIMENTAL RESULTS

We assessed the effectiveness of the proposed feature selection method in the context of classification problems. We carried out several experiments on a hyperspectral image acquired by the Hyperion sensor of the EO-1 satellite in an area of the Okavango Delta, Botswana. For details on this data set, we refer the reader to [7]. The labeled reference samples were collected on two different and spatially disjoint areas, thus representing two different domains. The samples taken on the first area (considered as source domain) were partitioned into a training set $\mathcal{D}^s$ and a test set $\mathcal{T}^s$ by a random sampling. Samples taken on the second area (target domain) were used to derive a training set $\mathcal{D}^t$ and test set $\mathcal{T}^t$ according to the same procedure. The estimated Pareto front for the selection of $m = 10$ features is reported in Fig. 3. In panel a), the color of the points indicates the Overall Accuracy (OA) obtained on the source-domain test set $\mathcal{T}^s$ using an SVM classifier trained using $\mathcal{D}^s$ (according to the reported color scale bar). In panel b), the color indicates the OA obtained by the SVM classifier on the target-domain test set $\mathcal{T}^t$. The results show that for the classification on the source domain, the solutions with higher relevance $\mathcal{R}$ result in better accuracies (note that the x-axis reports $-\mathcal{R}$). This behavior reveals that HSIC

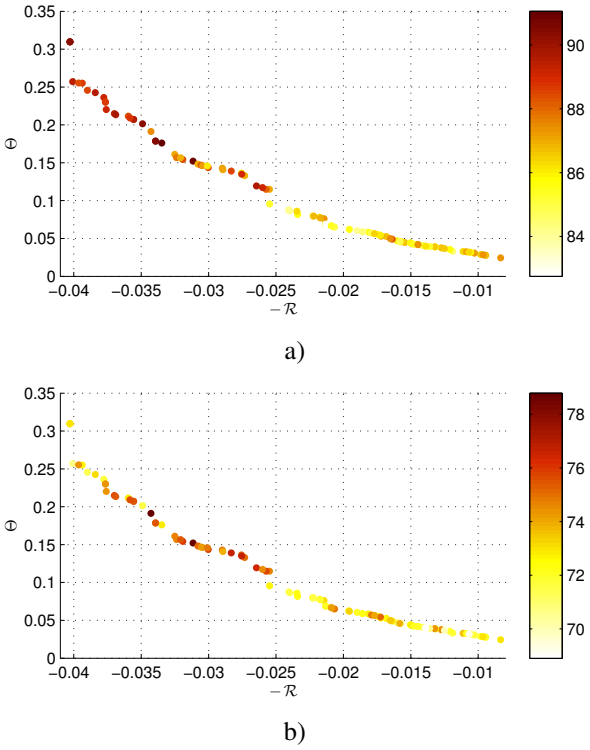

Fig. 3. Pareto front estimated by the multi-objective genetic algorithm for the selection of ten features. The color indicates the overall accuracy on a) source test set $\mathcal{T}^s$ and b) target test set $\mathcal{T}^t$ according to the reported color scale bar.

is actually a good indicator of feature relevance for the classification with SVM. However, the relevance only is not enough for selecting features that are stable for the classification on a different domain. We observe that the most accurate solutions for the classification of the spatially disjoint domain $\mathcal{T}^t$ are those that exhibits a good tradeoff between the relevance and invariance terms. This confirms the importance of the invariance term and that the proposed Θ measure is able to capture the information of feature stability. In order to select the subset of features that leads to good generalization capabilities on different domains, tradeoff solutions between the two competing objectives should be identified. We finally selected the subset of features that maximizes the OA on $\mathcal{D}^t$, resulting in an OA of 90.8% on the source domain ($\mathcal{T}^s$) and 78.8% on the target ($\mathcal{T}^t$). The set of features selected according to the optimization of $\mathcal{R}$ only resulted in an OA of 90.3% on the source ($\mathcal{T}^s$) and 73.0% on the target ($\mathcal{T}^t$). This result shows that the features selected by the proposed method can significantly improve the generalization capability on the target domain. Fig. 4 reports the Producer Accuracies obtained on the target domain ($\mathcal{T}^t$) by the standard and the proposed method.

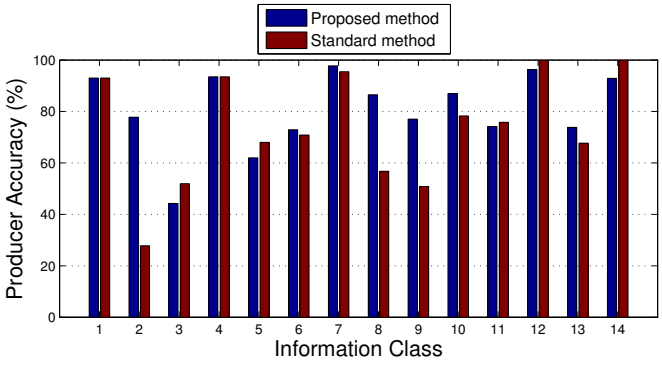

Fig. 4. Producer accuracies on the target domain.

4. CONCLUSION

In this paper, we have presented a novel feature-selection approach for the analysis and classification of hyperspectral images. The criterion function of the feature-selection method is based on the evaluation of both a feature relevance term and a novel domain invariance measure. Both measures are based on techniques for kernel embedding of distributions. Experimental results confirm that the proposed technique is able to select feature subsets that lead to augmented generalization capability of the classification system on the target domain.

5. REFERENCES

[1] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. on Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, 2009.

[2] L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo, "Supervised feature selection via dependence estimation," in *ICML*, 2007.

[3] G. Camps-Valls, J. Mooij, and B. Schölkopf, "Remote sensing feature selection by kernel dependence measures," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 3, 2010.

[4] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *ALT*, 2005.

[5] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *ICML*, 2009.

[6] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.

[7] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, 2005.