Transfer Component Analysis for Domain Adaptation in Image Classification

Giona Matasci[a], Michele Volpi[a], Devis Tuia[b,c], Mikhail Kanevski[a]

[a] IGAR, University of Lausanne, Bâtiment Amphipôle, 1015 Lausanne (Switzerland);
{giona.matasci, michele.volpi, mikhail.kanevski}@unil.ch
[b] IPL, University of València, 46100 Burjassot-València (Spain);
[c] LASIG, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne (Switzerland);
devis.tuia@epfl.ch

ABSTRACT

This contribution studies a feature extraction technique aiming at reducing differences between domains in image classification. The purpose is to find a common feature space between labeled samples issued from a source image and test samples belonging to a related target image. The presented approach, Transfer Component Analysis, finds a transformation matrix performing a joint mapping of the two domains by minimizing a probability distribution distance measure, the Maximum Mean Discrepancy criterion. When predicting on a target image, such a projection allows to apply a supervised classifier trained exclusively on labeled source pixels mapped in this common latent subspace. Promising results are observed on a urban scene captured by a hyperspectral image. The experiments reveal improvements with respect to a standard classification model built on the original source image and other feature extraction techniques.

Keywords: Domain adaptation, Feature extraction, Transfer Component Analysis, Image classification.

1. INTRODUCTION

In remote sensing image classification the ground truth collection process can be very demanding. Therefore, when classifying series of similar images with the supervised learning paradigm, the possibility to reuse labeled samples from a first acquisition is very appealing. Particularly, the ability to adapt a classifier built on an image, the *source domain*, to a new scene without needing any (or needing little) labeled data from the second image, the *target domain*, is of remarkable interest.[1] In the pattern recognition/machine learning community, this field of investigation is known as *domain adaptation* (DA).[2]

Similarly, such an approach can be applied when dealing with partial ground truth data covering a small and moderately representative subset of the image only. In a remote sensing study involving images covering a large surface, it is often impossible to acquire reference data uniformly over the whole considered region. Therefore, the collected class signatures are suffering a *sample selection bias*, meaning that the complete true statistical distribution of the classes is not adequately sampled and, consequently, can not be suitably modeled.

The shifts in the statistical distribution of the ground materials between the acquisitions (or in different subparts of an image) can be due to differences in illumination conditions, in the phenological state of the vegetative cover, in the shadowing effects caused by satellite view angles or solar elevations, etc.

In the recent literature, several contributions tackled the problem using active learning techniques.[3–5] By the addition of new samples to an initial training set built only on the source image, the dataset shift is corrected with an intelligently designed sampling scheme on the target image.

However, the *partially unsupervised* classification scheme, where the model for the target classification is exclusively built using labeled examples from the source image, is encountered much more frequently when

Further author information: Giona Matasci, ☎ +41 21 692 3538

dealing with concrete remote sensing studies. An accurate land cover-mapping not needing any sampling effort on the newly acquired image constitutes a notable achievement. Within this framework, we find approaches that aimed at iteratively adding pseudo-labeled target samples to the training set while removing the initial source examples[6] and other strategies exploiting unlabeled target samples to build more robust cluster similarities across domains.[7] More specifically, concerning dimensionality reduction problems, Bruzzone *et al.* (2009)[8] proved that a thoughtful selection of the features of a hyperspectral image, based on spatial invariance, could improve the accuracy when classifying testing regions using training data extracted from spatially disjoint areas.

Nevertheless, little attention has been paid to *feature extraction* (FE) techniques as tools to reduce the distribution change between source and target domains. In the present paper we study the possibility to project both domains to a common feature space minimizing the difference between them. After the proper joint mapping of the samples belonging to the two domains based on these newly extracted components, a model trained exclusively on the source image can be used for the predictions on the target image. In such a setting, we propose hereafter a method which is especially designed for DA. Originally proposed by Pan *et al.* (2011),[9] this approach named *Transfer Component Analysis* (TCA) is founded on the minimization of the distance between probability distributions of the two domains as measured by the *Maximum Mean Discrepancy* (MMD) criterion.[10] Such a measure is based on the evaluation of the distance between the means of the samples of the different domains when mapped in a common Reproducing Kernel Hilbert Space (RKHS).

The rest of the present paper is organized as follows. Section 2 outlines the general framework for DA based on the extraction of features from the original variables. In Sect. 3.1 we present and analyse the distribution divergence measure on which TCA is based, while Sect. 3.2 details the TCA technique in its unsupervised version. Next, in Sect. 4, we describe the ROSIS hyperspectral image of Pavia (Italy) and the associated design of the experiments. Section 5 reports and discusses the results obtained on the cited dataset and, finally, Sect. 6 concludes the paper.

2. DOMAIN ADAPTATION VIA FEATURE EXTRACTION

When dealing with data points issued from different but related distributions, a family of approaches designed to transfer the knowledge from the one domain to the other is the feature-representation-transfer ensemble of methods.[11] The underlying rationale consists of finding a proper common description of the source and target datasets minimizing the differences between these two domains while keeping the main data properties. Once the samples are projected in the same subspace created with the extracted features, a classifier is trained in the source domain, where labeled examples exist, and then inference is performed directly in the projected target domain. In remote sensing image classification, the ultimate scope is to extract a set of features that possess both a good discrimination capability of the ground cover classes as well as a sound spatial invariance.[8]

Let $\mathcal{D}_S = \{X_S, Y_S\} = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_s}$ be the n_s labeled source training data and $X_T = \{\mathbf{x}_{T_j}\}_{j=1}^{n_t}$ the n_t unlabeled target data, with samples $\mathbf{x}_{S_i}, \mathbf{x}_{T_j} \in \mathbb{R}^d \, \forall \, i, j$. In such a setting, the goal is to predict labels $y_{T_j} \in \Omega = \{\omega_c\}_{c=1}^{C}$ (with the same set of C classes as for y_{S_i}) making use exclusively of the labeled data belonging to $\mathcal{D}_S$ in the modeling phase. To this end, we look for a common mapping ϕ of the samples of both domains: $X_S \to \phi(X_S) = X_S^*$, $X_T \to \phi(X_T) = X_T^*$. Such a transformation is intended to reduce the divergence between marginal probability distributions $P(X_S)$ and $P(X_T)$ so that $P(X_S^*) \approx P(X_T^*)$.

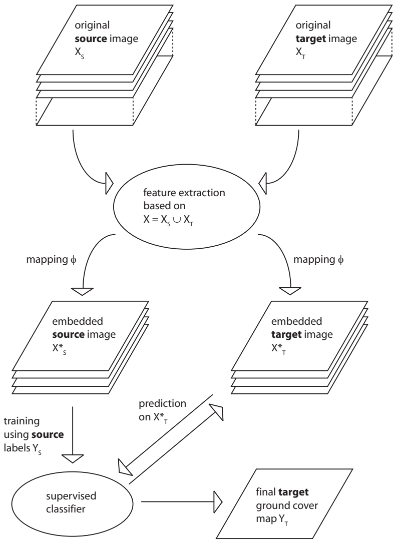


Figure 1. The process of knowledge transfer from a labeled source image to an unlabeled target image using a feature extraction approach. The final purpose of the procedure is to provide an accurate land cover map for the target image.

More concretely, exploiting X_T, one is allowed to build a stacked input set $X = X_S \cup X_T$. Using X we aim at finding a transformation matrix $\mathbf{W}$ able to execute the joint mapping ϕ of the original data X obtaining thus the embedded counterparts X_S^* and X_T^*. Different mapping matrices $\mathbf{W}$ can be obtained by applying standard FE methods such as *Principal Component Analysis* (PCA) or *Kernel Principal Component Analysis* (KPCA),[12] etc. Then, the user usually embeds the data in a m-dimensional space with $m \ll d$.

In the next stage, we learn a classifier on the mapped training set $\{(\mathbf{x}_{S_i}^*, y_{S_i})\}_{i=1}^{n_s}$ and then apply it without any adjustment to predict class labels for the mapped target points $\{\mathbf{x}_{T_j}^*\}_{j=1}^{n_t}$. As a matter of fact, the adaptation from the source to the target domain has already taken place during the FE phase. Therefore, the collection of labeled reference samples on the target image is not necessary.

Figure 1 illustrates the concept of DA using FE techniques.

3. TRANSFER COMPONENT ANALYSIS

3.1 Differences between distributions: the Maximum Mean Discrepancy

The FE technique studied in this contribution is intended to provide a common embedding of the considered domains such that the differences in the statistical distribution of the samples are reduced. As it will be shown in the next section, the proper optimal mapping in a DA setting is the one minimizing the shift in the data distributions. Hence, we first need to quantify the importance of the dataset shift occurred between source and target domains with an objective and robust measure of distribution divergence.

Many distances exist to evaluate the difference between probability distributions: Kullback-Leibler divergence,[13] Jensen-Shannon divergence, Bhattacharyya distance,[14] etc. However, these methods are affected by the data dimensionality, with the necessary probability density estimations becoming infeasible in high-dimensional spaces. Borgwardt *et al.* (2006)[10] propose a new indicator for comparing distributions based on the difference of the mean of the distributions computed in a common RKHS. This non-parametric measure, called *Maximum Mean Discrepancy*, is easily calculated no matter the number of variables describing the examples.

Following the notation of Sect. 2, the empirical estimate of the MMD between the distribution of a given source dataset X_S and that of a related target dataset X_T is given by

$$\mathrm{MMD}(X_S, X_T) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(\mathbf{x}_{S_i}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(\mathbf{x}_{T_j}) \right\|_{\mathcal{H}}^2 , \tag{1}$$

where $\|\cdot\|_{\mathcal{H}}^2$ is the squared norm computed in a RKHS. This quantity is nothing but the squared distance between sample means in the feature space and approaches zero when the two distributions tend to be exactly the same. Taking advantage of the well-known kernel trick one can rewrite (1) as

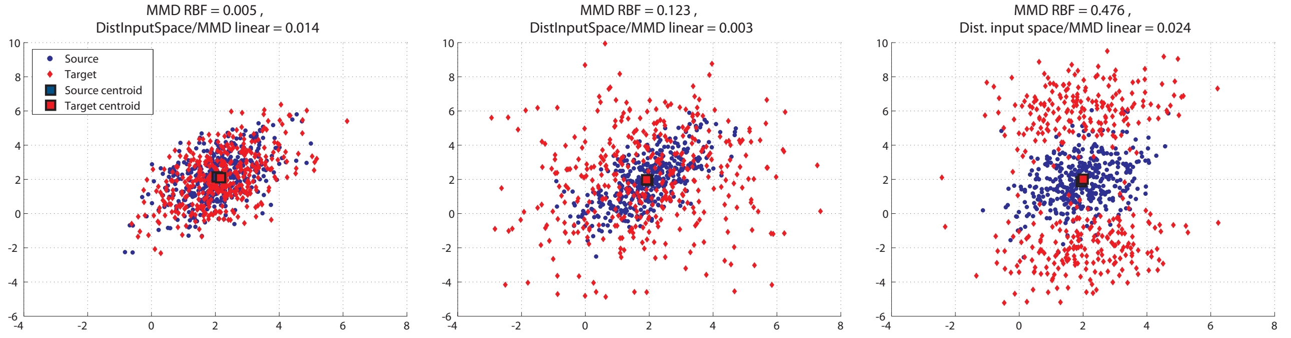
$$\mathrm{MMD}(X_S, X_T) = \mathrm{Tr}(\mathbf{KL}) , \tag{2}$$

where

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{S,S} & \mathbf{K}_{S,T} \\ \mathbf{K}_{T,S} & \mathbf{K}_{T,T} \end{pmatrix} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)} , \tag{3}$$

with $\mathbf{K}_{S,S}, \mathbf{K}_{T,T}, \mathbf{K}_{S,T}, \mathbf{K}_{T,S}$ being the kernel matrices (of elements $K_{i,j} = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$) obtained from the data of the source domain, target domain and cross domains, respectively. Moreover, $L_{i,j} = 1/n_s^2$ if $\mathbf{x}_i, \mathbf{x}_j \in X_S$, else if $\mathbf{x}_i, \mathbf{x}_j \in X_T$ we have $L_{i,j} = 1/n_t^2$ and, otherwise, $L_{i,j} = -1/n_s n_t$.

The use of the kernel trick, with its non-linear mapping of the data in a higher-dimensional feature space, is key for the MMD. Figure 2 presents the effectiveness of the presented indicator in detecting distribution shifts on toy datasets. We compare the MMD in its non-linear kernel version (with a Gaussian RBF kernel) with the simple difference of the means of the two distributions in the input space: in these examples we can appreciate the capability of MMD RBF to properly capture the difference in the shape of the data. In fact, even though the means of the distributions basically remain the same (overlapping blue and red centroids in the plots of Fig. 2), as the divergence between the blue and red data points becomes more marked, the MMD RBF indicator increases (value of 0.005 when distributions almost exactly overlap, 0.123 when X_T displays a higher variance and 0.476 when $P(X_T)$ is bimodal).



(a) (b) (c)

Figure 2. Dataset shift assessment for 3 toy datasets, representing source (in blue) and target (in red) domains combinations of increasing divergence ((a): no shift, (b): increase of the variance, (c): appearance of two modes). All experiments compare the MMD computed using a Gaussian RBF kernel with $\sigma = 2$ to the MMD with a linear kernel, *i.e.* the distance between the means in the input space (distance between centroids in 2-D).

3.2 Transfer Component Analysis

In order to reduce the differences between two distributions, it is very intuitive to seek a shared feature representation across domains. Such a representation is intended to mitigate the shift occurred between the source and target datasets. To this end, we present the TCA technique[9] (in its unsupervised form, *i.e.* not requiring any labeled target data) designed to extract meaningful *transfer components* from the original data belonging to different but related domains. As mentioned in Sect. 2, the purpose of FE algorithms in a DA setting is to find a mapping function ϕ, practically a transformation matrix $\mathbf{W}$, whose aim is to:

a) obtain a reduced distance between the probability distributions of $\phi(X_S)$ and $\phi(X_T)$

b) preserve the main properties of the original data X_S and X_T.

Starting from the kernel matrix $\mathbf{K}$ of (3) built on the stacked source and target sets, it is possible to use an embedding matrix $\mathbf{W} \in \mathbb{R}^{(n_s+n_t)\times m}$ (with $m \ll n_s + n_t$) to compute the kernel matrix between mapped samples as $\tilde{\mathbf{K}} = \mathbf{K}\mathbf{W}\mathbf{W}^\top\mathbf{K}$. Afterwards, to obtain the MMD measure for the mapped samples, we are allowed to rewrite (2) as

$$\mathrm{MMD}(X_S^*, X_T^*) = \mathrm{Tr}((\mathbf{K}\mathbf{W}\mathbf{W}^\top\mathbf{K})\mathbf{L}) = \mathrm{Tr}(\mathbf{W}^\top\mathbf{K}\mathbf{L}\mathbf{K}\mathbf{W})\,. \tag{4}$$

The objective stated in a) is thus achieved by minimizing (4) w.r.t. $\mathbf{W}$.

On the other hand, goal b) requires ϕ not to harm the target supervised learning task by deforming too much the input space. Hence, we look for a matrix $\mathbf{W}$ able to preserve (and maximize) the initial data variance in the newly created subspace, whose covariance matrix $\tilde{\mathbf{\Sigma}}$ is given by

$$\tilde{\mathbf{\Sigma}} = \mathbf{W}^\top\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W}\,, \tag{5}$$

where $\mathbf{H} = \mathbf{I} - (1/(n_s+n_t)\mathbf{1}\mathbf{1}^\top) \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)}$ is the centering matrix. Thus, the following constraint will be integrated in the optimization problem: $\tilde{\mathbf{\Sigma}} = \mathbf{I}_m$, where $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ is the identity matrix.

The final kernel learning problem is then set up as

$$\begin{aligned} \min_{\mathbf{W}} \quad & \mathrm{Tr}(\mathbf{W}^\top\mathbf{K}\mathbf{L}\mathbf{K}\mathbf{W}) + \mu\mathrm{Tr}(\mathbf{W}^\top\mathbf{W}) \\ \text{s.t.} \quad & \tilde{\mathbf{\Sigma}} = \mathbf{I}_m\,, \end{aligned} \tag{6}$$

where μ is a tradeoff parameter tuning the influence of the regularization term $\mathrm{Tr}(\mathbf{W}^\top\mathbf{W})$ controlling the complexity of $\mathbf{W}$. Such an optimization problem can be reformulated as a trace maximization problem yielding the following solution: the mapping matrix $\mathbf{W}$ is obtained by performing the eigendecomposition of

$$\mathbf{M} = (\mathbf{K}\mathbf{L}\mathbf{K} + \mu\mathbf{I})^{-1}\mathbf{K}\mathbf{H}\mathbf{K}\,, \tag{7}$$

and keeping the m eigenvectors associated with the m largest eigenvalues eig($\mathbf{M}$). For the detailed development refer to (Pan *et al.*, 2011).[9]

Once $\mathbf{W}$ is available, one is allowed to readily compute the m coordinates (the m uncorrelated *transfer components*) of the mapped samples as $\mathbf{X}^* = \mathbf{KW}$. In this latent subspace where distribution differences are reduced it is now possible to train a supervised classifier on the mapped source labeled samples and subsequently use it to classify the target image embedded in the same subspace.

4. DATA AND EXPERIMENTAL SETUP

Experiments have been carried out on a 1.3 m spatial resolution hyperspectral image acquired by the ROSIS-03 optical sensor over the city of Pavia (Italy).[15] The region of the spectrum covered by the 102 retained bands (13 noisy bands have been removed from the initial set of 115 channels) ranged from 0.43 to 0.86 μm. Given the urban setting, the goal is to discriminate between 4 classes: "buildings", "roads", "shadows" and "vegetation" (the class "water" has not been considered in the experiments). The image, presented in Fig. 3, shows a noticeable variation over space of the signatures of the ground cover classes (different materials constituting roofs and roads, different types of vegetative cover, etc.). In this context, we were allowed to consider subsets of the image as separate domains.

Therefore, to assess the ability of different FE techniques to transfer the knowledge acquired on a given source image to a related target image, we partitioned the hyperspectral scene into several spatially disjoint subsets. Figure 3 illustrates which parts of the ROSIS image were taken as target and source domains. To analyse different DA scenarios, we defined two source sub-images. The statistical distribution of the pixels issued from the sub-region named "Source A" is assumed to possess a moderate divergence from that of the pixels belonging to the sub-image identified as "Target" (empirical MMD RBF computed on standardized data equal to 0.0054). This divergence is thought to be higher when considering the "Source B" sub-image (MMD RBF = 0.0085), also due to the small spatial extent of the region covered by this subset. As preprocessing step, the histograms of the sub-images have been matched separately for each one of these source/target domains combinations.

Using *Linear Discriminant Analysis* (LDA) as base classifier, we compared the performances of 3 FE techniques. Namely, we report the classification accuracies on the target test set of a LDA model built with source samples described by features extracted by PCA (`PCA_FE`), by KPCA (`KPCA_FE`) and by the proposed TCA approach (`TCA_FE`). After the FE step, LDA models have been trained with source samples embedded in a space (common to the target samples) of increasing dimension (1 to 18 features). Furthermore, for comparison sake, models trained with samples belonging to the original target (`Target_orig`) and source (`Source_orig`) input space described by the 102 initial bands (no mapping) have also been tested.

Classification performances have been evaluated on a test set counting 14'047 pixels issued from the "Target" sub-image. As training sets to be exploited by the `Source_orig` model and by the 3 FE methods, we took into account 300 pixels per class belonging to the source regions of the Pavia image. We then have resorted to

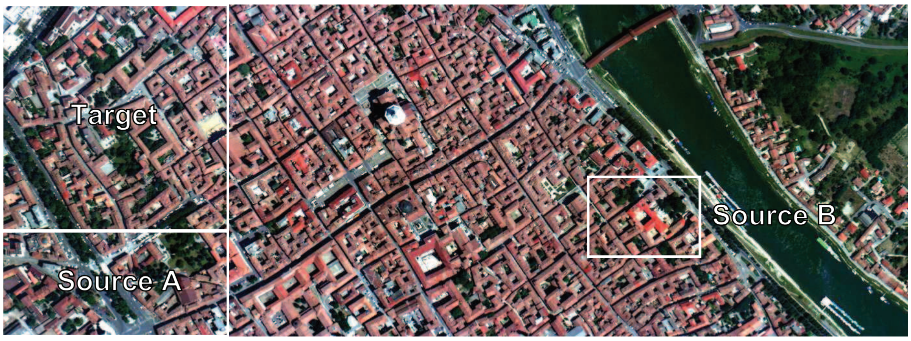


Figure 3. ROSIS image of Pavia used for the experiments (true color visualization). The sub-region identified with "Target" has been used as the target image. Sub-regions "Source A" and "Source B" have been considered as source images displaying a moderate and a large shift with respect to "Target", respectively.

the same amount of unlabeled examples coming from the target image to find the corresponding transformation matrices **W**. Lastly, 500 pixels/class were considered to build the model setting reference accuracies in the target domain (Target_orig method). Ten independent experiments with randomly selected pixels have been carried out to validate the approaches considered.

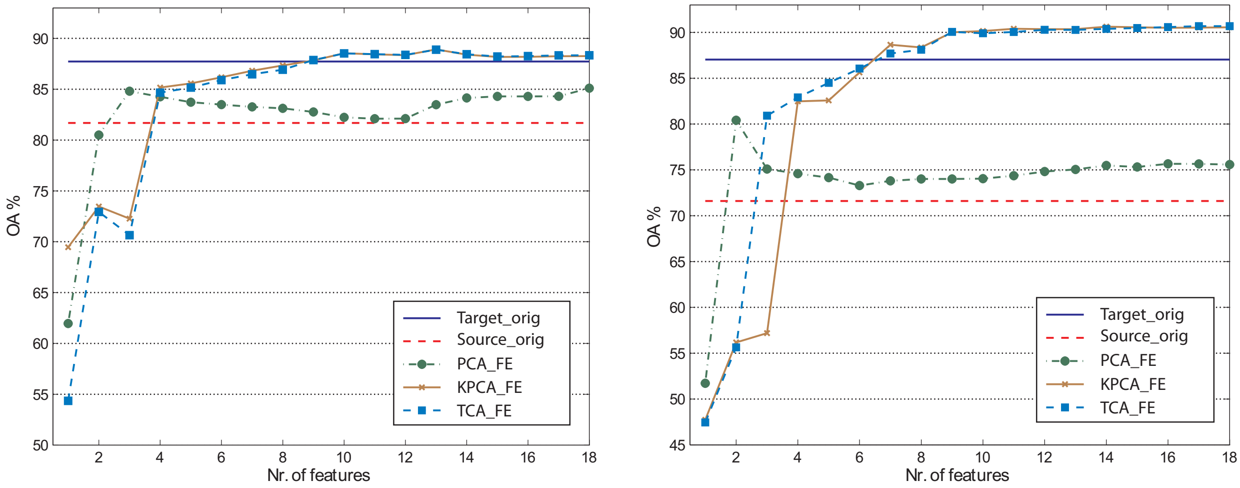
The selection of the σ parameter of the Gaussian RBF kernel, chosen both for the KPCA_FE and the TCA_FE mapping methods, has been carried out using the *Kernel Alignment* heuristic.[16] As done in previous studies,[17] we selected the parameter maximizing the alignment between the kernel matrix encoding sample similarities and the ideal kernel constructed using the associated class labels. Such a procedure only involved pixels belonging to the source domain, those whose labels were known. For the TCA_FE technique, after several tests having highlighted a low model sensitivity to its variation, we set the tradeoff parameter μ as equal to the standard value of 1.[9]

5. RESULTS AND DISCUSSION

Figure 4 illustrates the *Overall Accuracy* (OA) of the LDA model on the "Target" image when trained using an increasing number of components (1 to 18) extracted with the 3 FE methods.

In the moderate shift setting involving image "Source A" (Fig. 4(a)), the Source_orig model achieves an average OA of 81.68%, approaching the reference performance of the same-domain model Target_orig (OA = 87.73%). The PCA-based extraction considering the first 3 components is suitable to adequately describe the ground cover classes when source and target pixels are embedded in a shared subspace were the domain differences are reduced (OA = 84.81%). On the other hand, if we only consider components with the 3 largest eigenvalues, the two kernel-based FE methods provide less informative features (and consequent inferior classification accuracies). However, already from the inclusion of the 4th component on, the performance increases for both KPCA_FE and TCA_FE. With a similar evolution, such methods are able to reach and even exceed the Target_orig performance with LDA models built using 9 or more features.

In the second scenario considering image "Source B" as source domain, differences in methods performances are emphasized (Fig. 4(b)). Due to the larger divergence affecting the marginal probabilities of X_S and X_T,



(a) Training sets issued from the "Source A" sub-image: moderate dataset shift.

(b) Training sets issued from the "Source B" sub-image: large dataset shift.

Figure 4. Average LDA classification accuracies (OA %) on the "Target" sub-image for the 5 compared methods over 10 experiments as a function of the number of extracted features included in the model. Target_orig (solid blue line): model only using original target samples, Source_orig (dashed red line): model only using original source samples, PCA_FE (dashed green curve with dots): model using source samples with features extracted by PCA, KPCA_FE (solid brown curve with crosses): model using source samples with features extracted by KPCA, TCA_FE (dashed light blue curve with squares): model using source samples with features extracted by TCA.

the Source_orig model fails in appropriately mapping the land cover in the target domain (OA = 71.60%). The PCA and KPCA dimensionality reduction methods display here accuracy curves which are close to those of Fig. 4(a). However, because of the larger shift, we notice lower starting points for both methods and a sharper decrease in accuracy with the inclusion of noisy features (from 3 features on) for PCA. The TCA technique presented in this paper (TCA_FE curve), reveals good generalization abilities if 3 or more features are retained. TCA properly reduces the domain differences attaining notable accuracy levels (> 90.50% OA) when predicting on the target image. We also remark that this ability is exhibited by KPCA as well.

Figure 5 represents the RGB visualizations of the first 18 extracted components, taken 3 at a time in a decreasing order of importance (w.r.t. the eigenvalues), for the "Target" portion of the Pavia scene. The sub-image possessing a large shift ("Source B") was used as source for the FE step. A good correspondence can be found between the ability of the different groups of features in (visually) discriminating the land cover classes and the behavior of the associated classification models. Concerning the PCA extraction method, one can remark noisy features starting to appear already in the second RGB image (components 4, 5 and 6). This can be directly related to the decrease in classification accuracy described in the preceding paragraph, which is due to a lack of useful information contained in these features. On the other hand, one can observe the fairly neat RGB images involving features produced by KPCA and TCA. These extracted variables seem to be able to capture, in a non-linear fashion, the underlying structure of the data valid across domains. They display a certain richness in information, which is missing in those linearly combined by PCA (whose accuracy stagnates at a lower level from the inclusion of 3$^{\text{rd}}$ feature on). A supervised classifier will be greatly helped if asked to make use of such components for a given ground cover mapping task.

As general considerations, one can retain that the extraction of a set of features that is shared across domains induces a significant gain in the quality of the classification maps on the target image. In the large shift setting, one can notice up to 18% OA gains for the TCA and KPCA techniques over the source model using unmapped samples. This can be explained by the fact that all the considered FE procedures jointly use examples drawn from both the domains of interest, and this intrinsically reduces the existing dataset shift. Kernel-based non-linear FE methods proved very efficient in reducing distribution divergences, outperforming their linear counterparts. Even when the labeled pixels carrying the knowledge regarding the land cover signatures were issued from a source domain exhibiting a large divergence with respect to the target domain, a proper joint embedding of the

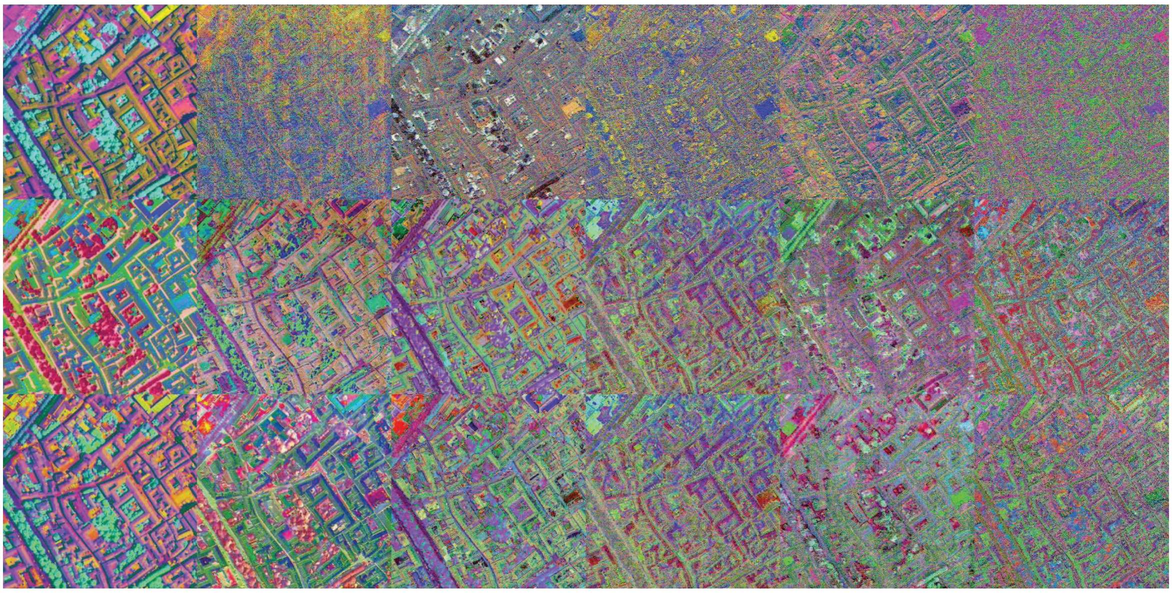


Figure 5. Visualization of the first 18 components respectively extracted by the 3 FE methods using samples from the "Source B" image. The columns represent the RGB combinations (3 components at a time) of the features sorted by decreasing eigenvalues. Top row: PCA, middle row: KPCA, bottom row: TCA.

samples allowed to effectively reuse the already acquired class labels. Thanks to the information conveyed by smartly built features, no labeling efforts on the target image were then needed in order to keep a satisfying classification performance. Moreover, note that, bearing in mind the high initial number of features (102 spectral bands), the dimensionality of the problem at hand is sharply reduced, no matter which FE technique is applied.

However, in this DA setting involving a single hyperspectral image, no pronounced differences have been observed between the presented TCA technique and KPCA. The reason behind such a similar evolution of the curves probably lies in the light distribution shift induced by the experimental setting considered. Moreover, one of the objectives of TCA, besides the minimization of the MMD distance between domains, is that of maximally preserving the data variance (constraint given by Eq. (5)), objective that is also pursued by KPCA. The specific aim of the transfer components, *i.e.* the reduction of distribution divergences, is perhaps not particularly required for a correct modeling of the various subregions of the considered scene.

6. CONCLUSIONS

With this contribution, we put forward the use of FE techniques as powerful tools to reduce distribution divergences between remote sensing images. We provided a preliminary study of some methods able to generalize the predictive abilities of existing classification models to newly acquired target images. Considering the increasing frequency and spatial extent of the acquisitions of remote sensing images we observed in the recent years, the possibility to intelligently reuse the information provided by previously labeled pixels is strongly needed.

In particular, the TCA method we presented, with its MMD minimization scope, proved potential in extracting features mitigating domain differences while being discriminant for image classification. LDA models trained on source images using such variables showed classification accuracies even exceeding those of models built on the target image constituted by the original spectral bands. Although more work (namely with other types of images and more challenging DA settings) is needed to thoroughly validate the usefulness of the considered approach, encouraging results have been obtained with these experiments on a hyperspectral dataset.

ACKNOWLEDGMENTS

This work has been supported by the Swiss National Science Foundation with grants no. 200021-126505, PBLAP2-127713/1 and PZ00P2-136827. The authors would like to acknowledge Prof. Paolo Gamba from the University of Pavia for providing the ROSIS image.

REFERENCES

[1] Bruzzone, L. and Prieto, D., "Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **39**(2), 456–460 (2001).

[2] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., [*Dataset Shift in Machine Learning*], MIT Press (2009).

[3] Tuia, D., Pasolli, E., and Emery, W. J., "Using active learning to adapt remote sensing image classifiers," *Remote Sensing of Environment* **115**(9), 2232–2242 (2011).

[4] Matasci, G., Tuia, D., and Kanevski, M., "Domain Separation For Efficient Adaptive Active Learning," in [*Proc. IEEE IGARSS 2011, Vancouver, Canada*], (2011).

[5] Persello, C. and Bruzzone, L., "A Novel Active Learning Strategy for Domain Adaptation in the Classification of Remote Sensing Images," in [*Proc. IEEE IGARSS 2011, Vancouver, Canada*], (2011).

[6] Bruzzone, L. and Marconcini, M., "Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy," *IEEE Trans. Geosci. Remote Sens.* **47**(4), 1108–1122 (2009).

[7] Gomez-Chova, L., Camps-Valls, G., Bruzzone, L., and Calpe-Maravilla, J., "Mean Map Kernel Methods for Semisupervised Cloud Classification," *IEEE Trans. Geosci. Remote Sens.* **48**(1), 207–220 (2010).

[8] Bruzzone, L. and Persello, C., "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.* **47**(9), 3180–3191 (2009).

[9] Pan, S. J., Tsang, I., Kwok, J. T., and Yang, Q., “Domain Adaptation via Transfer Component Analysis,” *IEEE Trans. Neural Netw.* **22**(2), 199–210 (2011).

[10] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J., “Integrating structured biological data by Kernel Maximum Mean Discrepancy,” *Bioinformatics* **22**(14), e49–e57 (2006).

[11] Pan, S. J. and Yang, Q., “A Survey on Transfer Learning,” *IEEE Trans. Knowl. Data Eng.* **22**(10), 1345–1359 (2010).

[12] Schölkopf, B., Smola, A., and Müller, K.-R., “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation* **10**(5), 1299–1319 (1998).

[13] Kullback, S. and Leibler, R. A., “On information and sufficiency,” *The Annals of Mathematical Statistics* **22**(1), 79–86 (1951).

[14] Bhattacharyya, A., “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bulletin of the Calcutta Mathematical Society* **35**, 99–109 (1943).

[15] Licciardi, G., Pacifici, F., Tuia, D., Prasad, S., West, T., Giacco, F., Thiel, C., Inglada, J., Christophe, E., Chanussot, J., and Gamba, P., “Decision Fusion for the Classification of Hyperspectral Data: Outcome of the 2008 GRS-S Data Fusion Contest,” *IEEE Trans. Geosci. Remote Sens.* **47**(11), 3857–3865 (2009).

[16] Cristianini, N., Kandola, J., Elisseeff, A., and Shawe-Taylor, J., “On kernel target alignment,” tech. rep., Roy. Holloway College, Univ. London, London, U.K., NeuroCOLT, 2001-087 (2001).

[17] Tuia, D., Camps-Valls, G., Matasci, G., and Kanevski, M., “Learning Relevant Image Features With Multiple-Kernel Classification,” *IEEE Trans. Geosci. Remote Sens.* **48**(10), 3780–3791 (2010).