Discriminative Transfer Joint Matching for Domain Adaptation in Hyperspectral Image Classification

Jiangtao Peng®, Weiwei Sun®, Li Ma®, and Qian Du®, Fellow, IEEE

Abstract—Domain adaptation, which aims at learning an accurate classifier for a new domain (target domain) using labeled information from an old domain (source domain), has shown promising value in remote sensing fields yet still been a challenging problem. In this letter, we focus on knowledge transfer between hyperspectral remotely sensed images in the context of land-cover classification under unsupervised setting where labeled samples are available only for the source image. Specifically, a discriminative transfer joint matching (DTJM) method is proposed, which matches source and target features in the kernel principal component analysis space by minimizing the empirical maximum mean discrepancy, performs instance reweighting by imposing an \( \ell_{2,1} \)-norm on the embedding matrix, and preserves the local manifold structure of data from different domains and meanwhile maximizes the dependence between the embedding and labels. The proposed approach is compared with some state-of-the-art feature extraction techniques with and without using label information of source data. Experimental results on two benchmark hyperspectral data sets show the effectiveness of the proposed DTJM.

Index Terms—Discriminative transfer joint matching (DTJM), domain adaptation, hyperspectral image classification, manifold.

I. INTRODUCTION

WITH the development of advanced sensors, the amount of remote sensing images available has increased significantly [1]. The exponential growth of hyperspectral images has created a compelling demand for automatic analysis and classification of image scenes [2]–[5]. However, automatic analysis and classification of hyperspectral images is a complex task without rich-labeled pixels. The acquisition of labeled pixels is usually time-consuming and infeasible for some nature limitations [6]. For this situation, classification is particularly difficult. When image scenes are similar, domain adaptation or transfer learning techniques allow us to use labeled pixels in the source domain to classify the scenes with limited or no labeled pixels (target domain). This classification problem can be referred to as cross-domain classification [1], [6]–[10].

In cross-domain problems, the source and target pixels may have a spectral shift, i.e., pixels belonging to the same land-cover class on two different scenes may vary in spectral distribution. The differences between source and target probability distributions are usually caused by biased sampling in the spatial domain or changes in the acquisition conditions [1]. In order to transfer the knowledge from the source domain to target domain, it is required to reduce the distribution difference in different domains. Many methods have been proposed to solve this problem. Subspace alignment (SA) is a simple and effective method which first projects the source and target data into subspaces and then learns a mapping function to align the source subspace with the target one [11], [12]. Geodesic flow kernel (GFK) method learns domain-invariant features by integrating an infinite number of subspaces [13]. These subspaces lie on the geodesic flow from the source subspace to the target one, and can be used to characterize changes in geometric and statistical properties from the source to the target domain. Usually, the source and target subspaces in SA and GFK are learned by the principal component analysis (PCA). Correlation alignment (CORAL) method minimizes domain shift by aligning the covariance of source and target distributions [14]. Transfer component analysis (TCA) aims at finding a common embedding of the data from the two domains by minimizing the empirical maximum mean discrepancy (MMD) in the kernel-PCA (KPCA) space [15]. Considering that there may exist source instances irrelevant to the target instances in the TCA subspace, transfer joint matching (TJM) jointly performs feature matching and instance reweighting through \( \ell_{2,1} \)-structural regularized KPCA [16]. As the label of source instances is available, many methods use the source labels to improve the domain adaptation. GFK method can be extended to GFK1 which uses partial least squares (PLS) to initialize the source subspace [13]. PLS is a supervised dimension reduction method which considers both sample and label information. Semisupervised TCA (SSTCA) maximizes the label dependence using the source labels and preserves the local geometry of data from different domains [8], [15].

In this letter, we embed the label information of source data into the TJM and propose a discriminative TJM (DTJM)
method for domain adaptation of hyperspectral images. The proposed DTJM matches source and target features in the KPCA embedding space by minimizing the empirical MMD, performs instance reweighting by imposing an $\ell_2,1$-norm on the embedding matrix, and preserves the local manifold structure of data from different domains and meanwhile maximizes the dependence between the embedding and labels. Based on the embedding function of DTJM, the source and target data are projected into aligned subspaces and classification is performed in the low-dimensional space.

II. TRANSFER JOINT MATCHING

Given a labeled source domain $D_s = \{(x_1, y_1), \ldots, (x_{n_s}, y_{n_s})\}$ with $n_s$ labeled samples and an unlabeled target domain $D_{t} = \{x_{n_s+1}, \ldots, x_{n_s+n_t}\}$ with $n_t$ samples under consideration, the TJM aims to learn a new feature representation to reduce the domain difference by jointly matching feature distributions in the KPCA space and reweighting source samples through $\ell_2,1$-structured regularization on the transformation matrix [16].

Let $X = [x_1, \ldots, x_{n_s+n_t}] \in \mathbb{R}^{D \times (n_s+n_t)}$ and denote the feature mapping function as $\phi : x \rightarrow \phi(x)$. Then, $\phi(X) = [\phi(x_1), \ldots, \phi(x_{n_s+n_t})]$ and kernel Gram matrix $K = (\phi(x))^T \phi(x) \in \mathbb{R}^{(n_s+n_t) \times (n_s+n_t)}$. KPCA learns a transformation matrix $A \in \mathbb{R}^{(n_s+n_t) \times k}$ by solving the following optimization problem:

$$\max_{A^T KHKA} \text{tr}(A^T K HKA)$$

(1)

where $I$ is an identity matrix and $H = I - (1/(n_s+n_t))11^T$ is the centering matrix with $1 = [1, \ldots, 1]^T$. The subspace embedding of KPCA is: $Z = A^T K$.

TJM employs the empirical MMD to match features in the reproducing kernel Hilbert space. The MMD is computed as

$$\left|\frac{1}{n_s} \sum_{i=1}^{n_s} A^T K_{ij} - \frac{1}{n_t} \sum_{j=n_s+1}^{n_s+n_t} A^T K_{ij}\right|^2_{\ell_2} = \text{tr}(A^T KMKA)$$

(2)

where $M$ is the MMD matrix with entries $M_{ij} = 1/n_i^2$ if $x_i, x_j \in D_s$; $M_{ij} = 1/n_t^2$ if $x_i, x_j \in D_t$; otherwise, $M_{ij} = -1/(n_s n_t)$.

Considering the fact that there may exist source instances irrelevant to target instances even in the feature-matching subspace, an instance reweighting strategy is employed to further minimize the distribution difference by imposing the $\ell_2,1$-norm structured sparsity regularizer on the transformation matrix $A$ as

$$\|A_s\|_{2,1} + \|A_t\|_{F}^2$$

(3)

where $A_s$ and $A_t$ are the submatrices of $A$ corresponding to the first $n_s$ rows and the last $n_t$ rows, respectively.

By jointly matching the feature distributions and reweighting the source instances, TJM optimization problem can be formulated as

$$\min_{A^T KHKA = I} \text{tr}(A^T KMKA) + \lambda \left(\|A_s\|_{2,1} + \|A_t\|_{F}^2\right)$$

(4)

where $\lambda$ is a regularization parameter.

III. DISCRIMINATIVE TRANSFER JOINT MATCHING

As the label of source instances is available, we encode label information into the TJM to maximize the dependence between the embedding and labels [15]. For this purpose, a kernel matrix is defined on labels as

$$\tilde{K}_y = \gamma K_L + (1 - \gamma)I$$

(5)

where $K_L \in \mathbb{R}^{(n_s+n_t) \times (n_s+n_t)}$ is only defined on the source domain with elements $[K_L]_{ij} = 1$ if $y_i = y_j$ and $i, j \leq n_s$ whereas $[K_L]_{ij} = 0$ otherwise. The first term on the right side of the (5) is used to maximize label dependence on the labeled data, while the second term serves to maximize the variance on both the source and target domain data [15]. $\gamma \geq 0$ is a tradeoff parameter to balance the two terms, and is empirically set as 0.5 in the experiments.

Based on the Hilbert–Schmidt independence criterion [8], [17], the dependence between the embedding and labels can be measured as

$$\text{tr}(A^T KH\tilde{K}_y HKA)$$

(6)

Similar to SSTCA, a manifold regularizer is employed to preserve the local geometrical structure in both domains [15]. To this end, we construct a data graph with affinity matrix $W = \{w_{ij}\}_{i,j=1}^{n_s+n_t}$ where $w_{ij} = \exp(-\|x_i - x_j\|^2/(2\alpha^2))$ if $x_i$ and $x_j$ are the $k$ nearest neighbors (NNs), and $w_{ij} = 0$ otherwise. In order to preserve the local manifold structure in the KPCA-embedding space

$$\sum_{i,j} w_{ij}\|\{A^T K_L\}_{ij} - \{A^T K_L\}_{ij}\|^2, = \text{tr}(A^T KLKA)$$

(7)

should be minimized where $[A^T K_L]_{ij}$ denotes the $i$th row of $A^T K_L$, and the graph Laplacian matrix is $L = D - W$ with $D$ being a diagonal matrix with entries $d_{ii} = \sum_{j=1}^{n_s+n_t} w_{ij}$.

By minimizing the MMD, maximizing the label dependence and preserving the local manifold structure, the optimization objective of DTJM can be formulated as

$$\min_{A} \{\text{tr}(A^T KMKA) + \mu \text{tr}(A^T KLKA) + \lambda (\|A_s\|_{2,1} + \|A_t\|_{F}^2)\}$$

s.t. $\text{tr}(A^T KHKA) = I$.

(8)

Denote $\Psi = \text{diag}(\psi_1, \ldots, \psi_k) \in \mathbb{R}^{k \times k}$ as the Lagrange multiplier, the Lagrange function for problem (8) is

$L = \text{tr}(A^T K(M + \mu L)KA) + \lambda (\|A_s\|_{2,1} + \|A_t\|_{F}^2) + \text{tr}(I - A^T KHKA)\Psi)$.

(9)

Setting $(\partial L/\partial A) = 0$, it can obtain a generalized eigenvector problem

$$(K(M + \mu L)K + \lambda G)A = KHKA\Psi$$

(10)

where $G$ is a diagonal subgradient matrix with entries $G_{ii} = (1/(2\|A_i\| + \epsilon))$ if $x_i \in D_s$, and $G_{ii} = 1$ if $x_i \in D_t$ [16]. An alternating optimization strategy is used to update $A$ and $G$ iteratively. Finally, $A$ is composed of the $k$ smallest eigenvectors of (10).
source scene and the $q$th class in the target scene. Clearly, the SSI value is between 0.5 and 1. The larger the value of SSI is, the larger is the degree of spectral shift. For the two hyperspectral data sets, the SSI values are 0.6543 and 0.6634, respectively. This demonstrates that both two data sets exist spectral shift, and hence, the domain adaptation is needed.

### C. Comparative Methods and Parameter Setting

The proposed DTJM method is compared with the following benchmarks.

1) **No Adaptation (NA):** No adaptation that directly classifies target samples based on the model learned from source samples.
2) **SA [11]:** Subspace alignment that performs alignment between PCA-based source and target subspaces.
3) **CORAL [14]:** CORAL that aligns the covariance of source and target distributions.
4) **GFK [13]:** GFK method that integrates an infinite number of subspaces that lie on the geodesic flow from the source subspace to the target one. Here, PCA is used to initialize the source and target subspaces.
5) **TCA [15]:** TCA that learns kernel-PCA components to minimize MMD.
6) **TJM [16]:** TJM that jointly performs feature matching and instance reweighting through $\ell_{2,1}$-structural regularized KPCA.
7) **GFK1 [13]:** GFK method that uses PLS and PCA to initialize the source and target subspaces, respectively.
8) **SSTCA [15]:** SSTCA that maximizes the label dependence using the source labels and preserves the local geometry using the data from different domains.

The one NN and support vector machine (SVM) are chosen as base classifiers. NN has no parameter to tune. For simplicity, SVM with the linear kernel is used and the penalty parameter $C$ is tuned in the range $\{1, 10, 100, 1000, 10000\}$. NN and SVM are trained on the labeled aligned source data and tested on the unlabeled aligned target data. The classification performance is assessed on the test set by the overall accuracy (OA) and the kappa coefficient ($\kappa$).

For subspace learning methods, the dimensionality of subspace is set as 20. For TCA, the regularization parameter $\lambda = 1$ is used [15]. For TJM and DTJM, the regularization parameter $\lambda$ and the number of iterations $T$ are set as: $\lambda = 1, T = 10$, as recommended in [16]. For SSTCA and DTJM, the number of NNs $k$ and the manifold regularization parameter $\mu$ are set as: $k = 20$ and $\mu = 200$.

### D. Experiments on Botswana

Fig. 3 shows the scatterplots after TJM and DTJM feature extraction for source and target data in the space formed by the first two components. Compared with TJM, DTJM-based projected data have a good local manifold structure and forms well-separated clusters. Table II provides the classification confusion matrix and the class accuracy (CA) of TJM and DTJM with NN classifier. It can be seen that DTJM shows much better results than TJM. Taking Class 1 as an example, TJM makes confusion among Classes 1, 2, 5, 6 while DTJM...
PENG et al.: DTJM FOR DOMAIN ADAPTATION IN HYPERSPECTRAL IMAGE CLASSIFICATION

Fig. 3. Scatterplots after (a) TJM and (b) DTJM feature extraction in the first versus second component spaces for the Botswana data set. (●) Source data. (□) Target data. The six classes are shown in different colors.

## TABLE II
### CONFUSION MATRIX AND CA OF TJM AND DTJM FOR THE BOTSWANA DATA SET

<table>
<thead>
<tr>
<th>Class predicted by TJM</th>
<th>True Class</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>118</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class predicted by DTJM</th>
<th>True Class</th>
<th>CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>112</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>79</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

## TABLE III
### CLASSIFICATION ACCURACIES FOR THE BOTSWANA DATA SET

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>0.7123</td>
<td>0.6490</td>
</tr>
<tr>
<td>SA</td>
<td>0.7884</td>
<td>0.7439</td>
</tr>
<tr>
<td>CORAL</td>
<td>0.7690</td>
<td>0.7205</td>
</tr>
<tr>
<td>GFK</td>
<td>0.7884</td>
<td>0.7439</td>
</tr>
<tr>
<td>TCA</td>
<td>0.7510</td>
<td>0.6987</td>
</tr>
<tr>
<td>TJM</td>
<td>0.7593</td>
<td>0.7087</td>
</tr>
<tr>
<td>GFK1</td>
<td>0.8022</td>
<td>0.7611</td>
</tr>
<tr>
<td>SSTCA</td>
<td>0.8091</td>
<td>0.7688</td>
</tr>
<tr>
<td>DTJM</td>
<td>0.8230</td>
<td>0.7857</td>
</tr>
</tbody>
</table>

1) When NN classifier is used, NA shows poor result and domain adaptation methods dramatically improve the NA which demonstrates that there exist distribution differences between source and target domains and the domain adaptation methods can effectively reduce such differences. In particular, SA and GFK provide relatively better results than CORAL, TCA, and TJM.

2) When SVM is used, SA, CORAL, GFK, TCA, and TJM show similar results as NA. This is because the powerful SVM classifier can alleviate distribution differences to a certain extent. In general, the results on SVM are better than those on NN.

3) By using the source labeled samples, GFK1 largely improves GFK. Similarly, SSTCA and DTJM improve the corresponding TCA and TJM by maximizing the label dependence and preserving the local manifold structure of data from different domains. The label of source samples can be used to improve the domain adaptation.

4) The highest OA and $\kappa$ coefficient obtained by NN and SVM classifiers demonstrate that DTJM is much more effective for domain adaptation.

### E. Experiments on Pavia City

The OAs and $\kappa$ coefficients on the target data for Pavia city data set are recorded in Table IV. No matter which method is used, NA shows very poor results due to the effect of sample selection bias problem. Here, the distribution differences between domains cannot be corrected by classifiers. By aligning the source and target data using domain adaptation methods, the classification performance is dramatically improved. In general, DTJM shows better results than other methods, especially combined with the SVM classifier.

The classification maps of different methods with SVM classifier are shown in Fig. 4. Without the domain adaptation, NA-SVM almost misclassifies all samples in the class “Building” even though this class has majority of samples, as shown in Fig. 4(b). Due to the distribution discrepancy, the model trained on the source data cannot be directly used to classify the target data. By performing feature transforms, the domain adaptation methods dramatically improve the results on the original data. However, on all maps, the “buildings” and “roads” classes are misclassified because their spectral characteristics are very similar [8].

mainly misclassifies Class 1 to Class 2. In general, DTJM can enhance class separability by maximizing label dependence.

The NN and SVM classification performance of different domain adaptation methods are given in Table III. The first six methods perform feature extraction only based on the unsupervised source and target samples $X$, while the last three methods (i.e., GFK1, SSTCA, and DTJM) use the samples $X$ and the label of source samples. Based on the results, we can obtain the following conclusions.
V. CONCLUSION

In this letter, we have proposed a DTJM method for domain adaptation of hyperspectral images. To match features in the KPCA space, the proposed DTJM minimizes the empirical MMD, maximizes the label dependence, preserves the local manifold structure, and meanwhile performs instance reweighting. Experimental results on two benchmark hyperspectral data sets have demonstrated its effectiveness and superiority.

REFERENCES