IET Computer Vision

Special Issue: Visual Domain Adaptation and Generalisation

IET Journals
The Institution of Engineering and Technology

Regularised transfer learning for hyperspectral image classification

ISSN 1751-9632
Received on 29th March 2018
Revised 5th May 2018
Accepted on 4th June 2018
E-First on 3rd August 2018
doi: 10.1049/iet-cvi.2018.5145
www.ietdl.org

Qian Shi[1], Yipeng Zhang[2] ✉, Xiaoping Liu[1], Kefei Zhao[3]

[1]*School of Geography and Planning, Sun Yat-sen University, Guangzhou, Guangdong, People's Republic of China*
[2]*School of Computer Engineering, Syracuse University, Syracuse, New York, USA*
[3]*School of Management, Guangdong University of Technology, Guangzhou, Guangdong, People's Republic of China*
✉ *E-mail: yzhan139@syr.edu*

Abstract: This study presents a transfer learning method for addressing the insufficient sample problem in hyperspectral image classification. In order to find common feature representation for both the source domain and target domain, we introduce a regularisation based on Bregman divergence into the objective function of the subspace learning algorithm, which can minimise the Bregman divergence between the distribution of training samples in the source domain and the test samples in the target domain. Hyperspectral image with biased sampling is used to evaluate the effectiveness of the proposed method. The results show that the proposed method can achieve a higher classification accuracy than traditional subspace learning methods under the condition of biased sampling.

1 Introduction

The development of hyperspectral sensor on recently launched satellites gives new chances for the application of remote sensing data [1–3]. The most robust and excellent methods for automatically classifying the remote sensing imagery are based on supervised learning methods, which need a large amount of training samples to train the classifier [4–6]. However, every time we get a new remote sensing image for classification, a new set of training samples has to be labelled [7–10]. If these training samples are available from a previous classification map and can be used to train the classification algorithms in a new task, the manually labelling costs can be dramatically reduced. However, the data distribution between the source domain and the target domain may be different, so the classifiers trained based on the samples in source domain may be not suitable to predict new samples in the target domain [11, 12]. Recently, transfer learning has been accounted as an effective machine learning method to address the above problems by learning the knowledge gained from the source domain (the image with high quality of training samples) and applying the knowledge in training sample set to the target domain (the image to be classified) [13, 14]. There is an assumption that the source domain and target domain should share similar information, for example, they have the same set of land cover classes, and the distributions of related land cover classes are correlated, however, they are not exactly the same [15–17].

The most crucial problem for sample transfer is how to model the difference between the probability distributions of the source and target domain, which is the premise to reduce this difference. Intuitively, finding a perfect representation for two domains is the direct way [18, 19]. First, the designed feature representation method could model and reduce the difference in probability distributions between the source domain and target domain [20–22], at the same time, preserving pivot features of the original data, which share some properties with target domain data [23–25]. von Bünau *et al*. [26, 27] proposed to find a stationary subspace to reduce the distributions difference between the source domain and target domain in a latent space. Pan *et al.* [28] solved this problem based on maximum mean discrepancy embedding (MMDE), which built a new subspace to measure the distribution distance across source domain and target domain in a reproducing kernel Hilbert space. However, MMDE could not process the out-of-sample situation. Pan *et al.* [29, 30] introduced transfer component analysis (TCA) for domain adaptation. This method tries to produce a group of common transfer components for two domains, in which the difference in data distributions of the different domains can be reduced. Sun *et al.* [31] introduced multiple-kernel support vector machines (SVMs) for domain adaptation, which map the data in the two domains to the high-dimensional space constructed by the multiple kernels, by which the difference between the two domains can be minimised.

In this paper, we focus on learning the subspace from the two domains, in which the difference between the two domains can be minimised by reducing the Bregman divergence between the two domains. By integrating a regularisation term based on Bregman divergence, we introduce a novel subspace learning framework [32], which can be flexibly incorporated into other traditional subspace learning methods.

To evaluate the effectiveness of the regularisation based on Bregman divergence for the transfer learning task, we test several supervised subspaces learning algorithms by adding this regularisation to the objective functions, including Fisher linear discriminant analysis (FLDA) [33], marginal Fisher analysis (MFA) [34] and discriminative locality alignment (DLA) [35, 36]. The experimental results prove the effectiveness of this framework for the classification task.

The rest of this paper is presented as follows. Section 2 introduces the regularisation based on Bregman divergence for transfer subspace learning (TSL). Section 3 presents different subspace learning implementations by utilising the TSL framework. In Section 4, the biased sampling strategies are used to test the effectiveness of the proposed algorithm based on the training examples described in Section 3. Finally, Section 5 concludes the whole paper.

2 TSL framework

For a supervised subspace learning methods, given l labelled samples $L = \{(x_1, z_1), \ldots, (x_l, z_l)\}$ and u test samples $U = \{x_{l+1}, \ldots, x_{l+u}\}$, in which z_i represents the label for sample x_i. We assume that they are both sampled from a high-dimensional data space R^D. Supervised dimension reduction algorithms find a low-dimensional subspace R^d, in which samples from the different classes can be well separated. The subspace learning algorithms try

to approximate the transformation $\boldsymbol{W}$ from space R^D to R^d, wherein $\boldsymbol{W} \in R^{D\times d}$

$$\boldsymbol{W} = \arg\min_{\boldsymbol{W}\in R^{D\times d}} F(W) \tag{1}$$

To constraint elements in transformation matrix $\boldsymbol{W}$ are orthogonal to each other, which is $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = I$. The objective function $F(W)$ is used to minimise the objective function in the reduced subspace according to different purposes. For example, FLDA, most typical supervised dimension reduction method, finds a subspace to minimise the trace ratio between within-class and between-class scatter matrix.

However, in practical applications, the distribution between the source domain and the target domain is always different; the subspace of source domain is not in accordance with target domain. Thus, we introduce the regularisation to the objective function, which is used to reduce the distribution difference between the source domain and target domain. Bregman divergence $D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})$ is used to measure the distribution difference between the datasets sampled from the different domains in a reduced subspace $\boldsymbol{W}$, in which P_L and P_U represent the probability of labelled sample set and unlabelled sample set. By adding this regularisation into framework (1), we can get the objective function of a TSL framework

$$\boldsymbol{W} = \arg\min F(W) + \lambda D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}}) \tag{2}$$

with orthogonal constraints $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = I$.

In (2), $F(W)$ is the objective function of subspace learning represented in (1), $D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})$ is the Bregman divergence distance between P_{L} and P_{U} in the reduced subspace $\boldsymbol{W}$, and λ is the parameter that keeps the balance between the first term and the second term. By using the quadratic divergence to represent $D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})$

$$\begin{aligned} D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}}) &= \int (p_{\mathrm{L}}(y) - p_{\mathrm{U}}(y))^2 \,\mathrm{d}y \\ &= \int (p_{\mathrm{L}}(y)^2 - 2p_{\mathrm{L}}(y)p_{\mathrm{U}}(y) + p_{\mathrm{U}}(y)^2) \,\mathrm{d}y \end{aligned} \tag{3}$$

To represent the distribution P_{L} and P_{U} in W, we utilise the kernel density estimation technique to estimate the probability density, which can be represented by $p(y) = (1/n)\sum_{i=1}^{n} G_\Sigma(y - y_i)$. Here, $G_\Sigma(y)$ is the d-dimensional Gaussian kernel, in which the covariance matrix Σ and n is the number of samples. By introducing the discrete distributions to (3), we have

$$\begin{aligned} D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}}) = &\int \left(\frac{1}{l}\sum_{i=1}^{l} G_{\Sigma_1}(y - y_i)\right)^2 \mathrm{d}y \\ &+ \int \left(\frac{1}{u}\sum_{j=l+1}^{l+u} G_{\Sigma_2}(y - y_i)\right)^2 \mathrm{d}y \\ &- \int \frac{2}{lu}\sum_{i=1}^{l}\sum_{j=l+1}^{l+u} G_{\Sigma_1}(y - y_i)G_{\Sigma_2}(y - y_i)\,\mathrm{d}y \end{aligned} \tag{4}$$

For two different Gaussian kernels, we have the equation $\int G_{\Sigma_1}(y - y_i)G_{\Sigma_2}(y - y_i)\,\mathrm{d}y = G_{\Sigma_1+\Sigma_2}(y_s - y_t)$, and thus we can eliminate the third term in (4). Equation (4) can be represented by

$$\begin{aligned} D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}}) = &\frac{1}{l^2}\sum_{s=1}^{l}\sum_{t=1}^{l} G_{\Sigma_{11}}(y_t - y_s) \\ &+ \frac{1}{u^2}\sum_{s=l+1}^{l+u}\sum_{t=l+1}^{l+u} G_{\Sigma_{22}}(y_t - y_s) \\ &- \frac{2}{lu}\sum_{s=1}^{l}\sum_{t=l+1}^{l+u} G_{\Sigma_{12}}(y_t - y_s) \end{aligned} \tag{5}$$

where $\Sigma_{11} = \Sigma_1 + \Sigma_1$, $\Sigma_{12} = \Sigma_1 + \Sigma_2$ and $\Sigma_{22} = \Sigma_2 + \Sigma_2$. Equation (5) shows that $D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})$ is a data-dependent function with respect to W.

To obtain the optimal linear subspace W in (2), the gradient descent technique is used to iteratively optimise (2) with respect to $\boldsymbol{W}$. For the iteration $k + 1$, the update rule for solving $\boldsymbol{W}$ is

$$W_{k+1} = W_k - \eta(k)\left(\frac{\partial F(W)}{\partial W} + \lambda\sum_{i=1}^{l+u}\frac{\partial D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})}{\partial y_i}\frac{\partial y_i}{\partial W}\right) \tag{6}$$

where $\eta(k)$ represents the gradient step size for kth iteration.

According to the representation in (5), which is generated from the regularisation based on Bregman divergence (4), the derivative of $D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})$ with respect to $\boldsymbol{W}$ is

$$\begin{aligned} &\sum_{i=1}^{l+u}\frac{D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})}{\partial y_i}\frac{\partial y_i}{\partial W} \\ &= \sum_{i=1}^{l}\frac{D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})}{\partial y_i}x_i^{\mathrm{T}} + \sum_{i=l+1}^{l+u}\frac{D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})}{\partial y_i}x_i^{\mathrm{T}} \\ &= \frac{2}{l^2}\sum_{i=1}^{l}\sum_{t=1}^{l} G_{\Sigma_{11}}(y_i - y_t)(\Sigma_{11})^{-1}(y_i - y_t)x_i^{\mathrm{T}} \\ &\quad - \frac{2}{lu}\sum_{i=1}^{l}\sum_{t=l+1}^{l+u} G_{\Sigma_{12}}(y_i - y_t)(\Sigma_{12})^{-1}(y_i - y_t)x_i^{\mathrm{T}} \\ &\quad + \frac{2}{u^2}\sum_{i=l+1}^{l+u}\sum_{t=l+1}^{l+u} G_{\Sigma_{22}}(y_i - y_t)(\Sigma_{22})^{-1}(y_i - y_t)x_i^{\mathrm{T}} \\ &\quad - \frac{2}{lu}\sum_{i=1}^{l+u}\sum_{t=l+1}^{l} G_{\Sigma_{12}}(y_i - y_t)(\Sigma_{12})^{-1}(y_i - y_t)x_i^{\mathrm{T}} \end{aligned} \tag{7}$$

The terms $F(W)$ and $D(W)$ jointly determine whether objective function defined in (2) is convex. First, the convexity of $F(W)$ depends on a specific traditional subspace learning method, at the same time, the convexity of $D(W)$ depends on the how to calculate the probability distribution of the training and test sets. As a result, the convexity of the objective function for the proposed TSL framework is problem dependent. Since (2) cannot be convex, it is necessary to set a good initialisation as the starting point of the optimisation. The experiments in this paper set the linear discriminant analysis (LDA) projection matrix as the initialisation scheme. For the parameter in the optimisation step, the appropriate learning rate $\eta(k)$ is one critical element. In the proposed framework, we empirically set $\eta(k) = \eta(0)/k$, and the k is the iterations. Thus, in the early iterations, the $\eta(k)$ is large, and the searching step is large, which is because the optimal solution of the objective function is far away from the W in current state. In the last iterations, as the $\eta(k)$ is more smaller, the speed of finding optimum value slows down. To find optimal solution at a fine scale, the small step sizes are used for updating W in last few iterations.

3 Implementation of the Bregman divergence based TSL

In this section, we combine the traditional subspace learning algorithms with introduced Bregman regularisation to construct common subspace for test sample set and training sample set. We

will represent the first term $\partial_W F(W)$ for each subspace learning method.

3.1 Transferred FLDA

LDA is the most typical subspace reduction method which is used to maximise the between-class scatter, meanwhile, minimise the within-class scatter based on training sample set. Thus, we maximise the trace ration between the between-class and within-class scatter matrix, which can be formulated as

$$F(W) = \mathrm{tr}^{-1}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{W})\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{W}) \tag{8}$$

where $\boldsymbol{S}_B$ is the between-class scatter matrix, and $\boldsymbol{S}_W$ is the within-class scatter matrix.

The derivative of $F(W)$ with respect to $\boldsymbol{W}$ is given by

$$\begin{aligned}\frac{\partial_W F(W)}{\partial W} &= 2\mathrm{tr}^{-1}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{W})\boldsymbol{S}_{\mathrm{W}}\boldsymbol{W} \\ &\quad -2\mathrm{tr}^{-2}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{W})\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{W})\boldsymbol{S}_{\mathrm{B}}\boldsymbol{W}\end{aligned} \tag{9}$$

In the second term, tr^{-2} is the square of the inverse of $\mathrm{tr}(X)$. By substituting (8) and (7) into (2), subject to the constraint of $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = I$, we can obtain the solution of $\boldsymbol{W}$ in TLDA with an iterative way.

3.2 Transferred MFA

The purpose of MFA is to maximise the difference of different classes and minimise the difference of same class. This purpose can be realised by constructing two graphs, which are intrinsic graph and the penalty graph. First, intrinsic graph is constructed by the similar relationship between nearest neighbours belonging to same class, which is used to describe the intraclass compactness. Meanwhile, the interclass separability is described by the penalty graph, in which the adjacency relationship of the interclass marginal samples is shown as follows:

$$\begin{aligned}F(W) &= \frac{\sum_{i=1}^{l}\sum_{j=1}^{l}((\boldsymbol{W}^{\mathrm{T}}x_i - \boldsymbol{W}^{\mathrm{T}}x_i)^{\mathrm{T}}(\boldsymbol{W}^{\mathrm{T}}x_i - \boldsymbol{W}^{\mathrm{T}}x_i))E_{ij}^{C}}{\sum_{i=1}^{l}\sum_{j=1}^{l}((\boldsymbol{W}^{\mathrm{T}}x_i - \boldsymbol{W}^{\mathrm{T}}x_i)^{\mathrm{T}}(\boldsymbol{W}^{\mathrm{T}}x_i - \boldsymbol{W}^{\mathrm{T}}x_i))E_{ij}^{P}} \\ &= \frac{\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}X(\boldsymbol{D}^{C} - E^{C})X^{\mathrm{T}}\boldsymbol{W})}{\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}X(\boldsymbol{D}^{P} - E^{P})X^{\mathrm{T}}\boldsymbol{W})}\end{aligned} \tag{10}$$

where $\boldsymbol{D}^C$ and $\boldsymbol{D}^P$ are diagonal matrices of weight matrix $\boldsymbol{W}$. The ith entry of $\boldsymbol{D}^C$ and $\boldsymbol{D}^P$ is that $D_{ii}^{C} = \sum_{j=1}^{l} E_{ij}^{C}$ and $D_{ii}^{P} = \sum_{j=1}^{l} E_{ij}^{P}$. The implementation of $F(W)$ in MFA is

$$\begin{aligned}\frac{\partial F(W)}{\partial W} &= \frac{2X(\boldsymbol{D}^{C} - E^{C})X^{\mathrm{T}}\boldsymbol{W}}{\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}X(\boldsymbol{D}^{P} - E^{P})X^{\mathrm{T}}\boldsymbol{W})} \\ &\quad - \frac{2\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}X(\boldsymbol{D}^{C} - E^{C})X^{\mathrm{T}}\boldsymbol{W})X(\boldsymbol{D}^{P} - E^{P})X^{\mathrm{T}}\boldsymbol{W}}{\mathrm{tr}^{2}(\boldsymbol{W}^{\mathrm{T}}X(\boldsymbol{D}^{P} - E^{P})X^{\mathrm{T}}\boldsymbol{W})}\end{aligned} \tag{11}$$

3.3 Transferred DLA

DLA is a supervised manifold learning method. For a given sample x_i, in a local patch, DLA finds k_1 nearest samples $x_{i_1}, \ldots, x_{i_{k_1}}$ with same class label, while finds k_1 nearest samples $x_i^1, \ldots, x_i^{k_2}$ with different class label. Therefore, the local patch is combining two kinds of nearest neighbours $X_i = [x_i, x_{i_1}, \ldots, x_{i_{k_1}}, x_i^1, \ldots, x_i^{k_2}]$.

In the part optimisation, the ith patch is formulated by

$$\arg\min_{Y_i} \mathrm{tr}(Y_i m_i L_i Y_i^{\mathrm{T}}) \tag{12}$$

where

$$L_i = \begin{bmatrix} \sum_{j=1}^{k_1+k_2}(w_i)_j & -w_i^{\mathrm{T}} \\ -w_i & \mathrm{diag}(w_i) \end{bmatrix},$$

$$w_i = [1, \ldots 1 \quad -\beta, \ldots -\beta] \text{ and } m_i = \exp\left(-\frac{1}{(n_i + \delta)t}\right).$$

m_i represents the weight of each sample x_i.

The global optimisation is formulated as the summation of part optimisation over all patches:

$$\arg\min_{Y} \sum_{i=1}^{N} \mathrm{tr}(Y_i(\boldsymbol{S}_i m_i L_i \boldsymbol{S}_i^{\mathrm{T}})Y_i^{\mathrm{T}}) = \arg\min_{Y} \mathrm{tr}(YLY^{\mathrm{T}}) \tag{13}$$

where $\boldsymbol{S}_i$ is the selection matrix, which selects the most nearest samples from the whole dataset. Finally, with $Y = \boldsymbol{W}^{\mathrm{T}}X$, the subspace is obtained by minimising the function $F(W)$

$$F(W) = \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}XLX^{\mathrm{T}}\boldsymbol{W}) \quad \mathrm{s.t.} \quad \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = I \tag{14}$$

The derivative of $F(W)$ is obtained with respect to

$$\frac{\partial F(W)}{\partial W} = (XLX^{\mathrm{T}} + (XLX^{\mathrm{T}})^{\mathrm{T}})\boldsymbol{W} \quad \mathrm{s.t.} \quad \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = I \tag{15}$$

4 Experiments

In this experiment, we use two hyperspectral datasets to test our proposed method. First, the Indian Pines hyperspectral dataset is most common hyperspectral imagery, which was acquired by National Aeronautics and Space Administration's AVIRIS sensor. We called this dataset as Indian Pines dataset for short, as shown in Fig. 1*a*. AVIRIS dataset contains 145 × 145 pixels. For each pixel, there are 220 spectral bands covering the range of 375–2200 nm. The spatial resolution is ∼20 m. There are 16 agriculture classes in AVIRIS dataset.

The second dataset was acquired by the HYDICE sensor, as shown in Fig. 1*b*. This dataset contains 1280 × 307 pixels which cover the region of Washington DC (WDC) Mall. There are five land-cover classes, which are roof, street, path, grass and trees. All these classes are contained in source domain area, which can be shown in green rectangle in the figure.

To investigate the effect of the sample selection bias problem in remote sensing data, the sampling strategies used in [37] are also used for performance validation in this paper. It is assumed that selecting a sample set by randomly sampling over the whole image is unbiased sampling. We also assume that the way that sampling training samples in the small local patches and testing on the whole image is biased sampling. Thus, we cut a part from the original hyperspectral image, which is represented by yellow rectangles in Fig. 1. This segment area is seen as the source domain, from which training sample set is selected; and the rest of image is accounted as the target domain. To reduce the complexity of the computation, few samples are selected from the target domain to compute $D_W(P_{\mathrm{L}} \parallel P_{\mathrm{U}})$. According to [17], in order to select the most representative unlabelled samples to be labelled for the target domain, we first segment the whole image, and then select the root point in each segment patch to represent the whole image. For the Indian Pines dataset, only ten classes are contained in the yellow regions; thus, we test these ten land cover classes in the target domain. These particular classes are: 2, 3, 4, 5, 6, 9, 10, 11, 12 and 15.

There are two parameters that need to be given. The first one is λ, which is used to balance the discriminant term $F(W)$ and the Bregman distance D_W. The second parameter is the learning rate $\eta(k)$. We experimentally set these to $\lambda = 0.5$ and $\eta(k) = 1$.

Classic SVM is used as the classifier. A Gaussian kernel is used as the default kernel. A grid search is used to find the optimal parameters for SVM classifier. Specifically, $C \in \{0.1, 1, 10, 100, 1000\}$ and $\lambda \in \{0.002, 0.008, 0.032, 0.128, 0.512\}$. Tenfold cross-validation is used for this grid search.

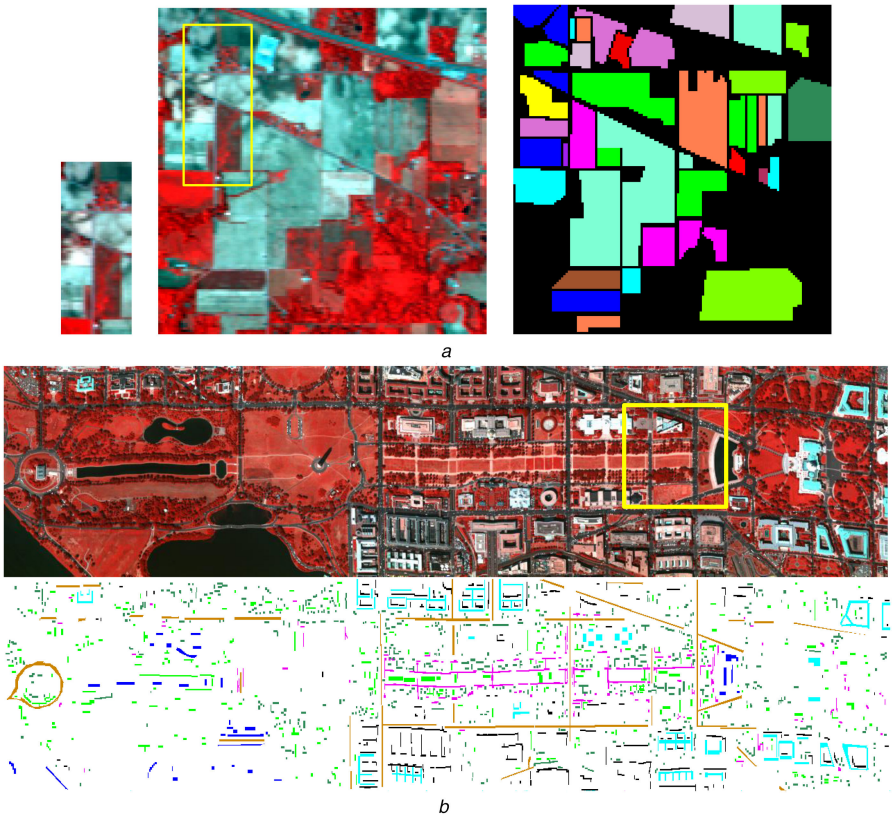

Fig. 1 *Pseudocolour image and ground truth map for the hyperspectral data. The yellow rectangles represent the source domain area (best viewed online in colour)*
(a) Indian Pines dataset, ***(b)*** WDC dataset

Table 1 OA and kappa values with the two different experimental settings for the Indian Pines dataset

	Biased sampling		Non-biased sampling	
	OA	Kappa	OA	Kappa
LDA	0.4617 ± 0.0130	0.4369	0.7659 ± 0.0128	0.7210
MFA	0.4763 ± 0.0161	0.4593	0.7963 ± 0.0134	0.7824
DLA	0.4978 ± 0.0142	0.4867	0.7836 ± 0.0136	0.7663
MMDE	0.5046 ± 0.0158	0.5349	0.7319 ± 0.0118	0.7091
TLDA	0.5234 ± 0.0147	0.5993	0.7721 ± 0.0152	0.7564
TMFA	0.5335 ± 0.0134	0.6027	0.8024 ± 0.0163	0.7876
TDLA	0.5431 ± 0.0128	0.6134	0.7967 ± 0.0169	0.7739
TCA	0.5124 ± 0.0144	0.5087	0.7832 ± 0.0157	0.7723

Table 2 OA and kappa values with the two different experimental settings for the WDC dataset

	TF2WF		WF2WF	
	OA	Kappa	OA	Kappa
LDA	0.8434 ± 0.0130	0.8319	0.9434 ± 0.006	0.9364
MFA	0.8561 ± 0.0161	0.8436	0.9567 ± 0.007	0.9315
DLA	0.8610 ± 0.0142	0.8394	0.9624 ± 0.006	0.9501
MMDE	0.8534 ± 0.0103	0.8335	0.9468 ± 0.005	0.9296
TLDA	0.8684 ± 0.0157	0.8537	0.9539 ± 0.008	0.9316
TMFA	0.8845 ± 0.0143	0.8581	0.9542 ± 0.010	0.9342
TDLA	0.8831 ± 0.0161	0.8628	0.9327 ± 0.008	0.9185
TCA	0.8641 ± 0.0124	0.8514	0.9254 ± 0.004	0.9157

There are two different settings: biased sampling and non-biased sampling. The first setting is selecting the labelled samples in the source domain; meanwhile, the test samples come from the whole image. The second setting is that both the training and test samples come from the whole image. Tables 1 and 2 show the overall accuracy (OA) and kappa values with two different experimental settings for the Indian Pines dataset and the WDC dataset. Fig. 2 presents the overall classification accuracy versus the subspace dimension by utilising different dimension reduction method.

The TSL-based versions outperform MMDE and TCA. Tables 1 and 2 show the classification accuracies of different subspace learning algorithm with regard to the optimal number of subspace dimension. Specifically, the conventional subspace learning methods represented poor performance because they assume that the training set and test sample set share exactly same distribution, and this assumption is unsuitable for the practical condition.

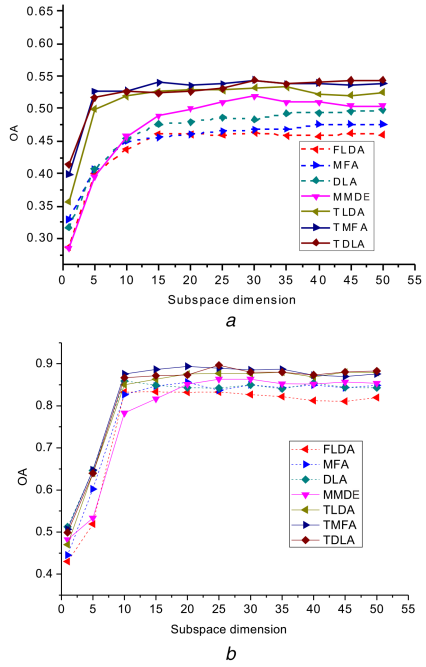

Fig. 2 *Recognition rates versus subspace dimensions under the biased sampling strategies for*
(a) Indian Pines dataset, ***(b)*** WDC dataset

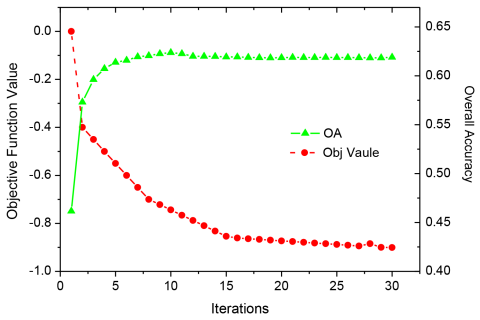

Fig. 3 *Objective function value and overall classification accuracy versus iterations*

Although MMDE also takes the distribution bias between the source domain and the target domain into consideration, it discards the discriminative information in the reduced space. The TSL-based algorithms outperform other counterpart methods because discriminative training information is preserved in the lower subspace, at the same time, distribution difference between the two domains can be minimised. Furthermore, under the non-biased setting, the performances of the TSL-based algorithms are still slightly better than the traditional subspace methods, which proves that even under the random sampling strategy, the sample set cannot guarantee to be strict non-bias.

Fig. 2 presents the overall accuracy versus the dimension with the different subspace learning algorithms. It can be seen that the TSL examples significantly outperform the others. Fig. 3 presents the trends of the classification accuracy and transfer linear discriminant analysis' (TFLDA's) objective function values $F(W) + \lambda D_W$ with regard to the training iterations. As the other TSL-based algorithms also show similar running characteristic in the stage of finding subspace, we just show the result and analysis of TFLDA with the Indian Pines dataset. For the initial projection matrix $\boldsymbol{W}_0$, we use ten random initialisations to find the most suitable $\boldsymbol{W}_0$. Fig. 3 shows that, with the increase of training iteration, the value of objective function of TFLDA and corresponding variance consistently decrease.

5 Conclusion

In this paper, we introduced a regularisation term based on Bregman divergence into traditional subspace learning framework, which is used to minimise the divergence of the probability distribution between the source domain and target domain. The regularisation item could transfer the discriminative knowledge learned from the source domain, meanwhile, model and reduce the distribution difference between the target domain and source domain. Furthermore, by adding the regularisation term, the discriminative information in training samples from source domain could be preserved. The experimental results show that the proposed method could effectively improve the predictability of classifier which is learned from the source domain. However, the performances of all the methods under the non-biased setting are higher than the performances of the TSL-based algorithms under the biased setting, which proves the limitation of the unsupervised transfer learning methods; thus, the supervised transfer learning method will be further studied to narrow the difference between the two settings.

6 References

[1] Li, W., Chen, C., Su, H., *et al.*: 'Local binary patterns and extreme learning machine for hyperspectral imagery classification', *IEEE Trans. Geosci. Remote Sens.*, 2012, **53**, (7), pp. 3681–3693

[2] Du, B., Zhang, L.: 'A discriminative metric learning based anomaly detection method', *IEEE Trans. Geosci. Remote Sens.*, 2014, **52**, (11), pp. 6844–6857

[3] Wang, Q., Gao, J., Yuan, Y.: 'Embedding structured contour and location prior in siamesed fully convolutional networks for road detection', *IEEE Trans. Intell. Transp. Syst.*, 2018, **19**, (1), pp. 230–241

[4] Wu, J., Pan, S., Zhang, P., *et al.*: 'Direct discriminative bag mapping for multi-instance learning'. Proc. of the 30th AAAI Conf. on Artificial Intelligence (AAAI'16), Phoenix, Arizona, USA, 12–17 February 2016

[5] Du, B., Zhang, L.: 'Target detection based on a dynamic subspace', *Pattern Recognit.*, 2014, **47**, (1), pp. 344–358

[6] Wu, J., Zhu, X., Zhang, C., *et al.*: 'Bag constrained structure pattern mining for multi-graph classification', *IEEE Trans. Knowl. Data Eng.*, 2014, **26**, (10), pp. 2382–2396

[7] Du, B., Zhang, L.: 'Random-selection-based anomaly detector for hyperspectral imagery', *IEEE Trans. Geosci. Remote Sens.*, 2011, **49**, (5), pp. 1578–1589

[8] Wang, Q., Wan, J., Yuan, Y.: 'Locality Constraint Distance Metric Learning for Traffic Congestion Detection', *Pattern Recognit.*, 2018, **75**, pp. 272–281

[9] Wu, J., Pan, S., Zhu, X., *et al.*: 'Multiple structure-view learning for graph classification', *IEEE Trans. Neural Netw. Learn. Syst.*, 2018, **29**, (7), pp. 3236–3251

[10] Du, B., Zhang, L., Tao, D., *et al.*: 'Unsupervised transfer learning for target detection from hyperspectral images', *Neurocomputing*, 2013, **120**, pp. 72–82

[11] Wang, Q., Gao, J., Yuan, Y.: 'A joint convolutional neural networks and context transfer for street scenes labeling', *IEEE Trans. Intell. Transp. Syst.*, 2017, **19**, (5), pp. 1457–1470

[12] Bahirat, K., Bovolo, F., Bruzzone, L., *et al.*: 'A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains', *IEEE Trans. Geosci. Remote Sens.*, 2012, **50**, (7), pp. 2810–2826

[13] Du, B., Tang, X., Zhang, L.: 'Robust graph-based semi-supervised learning for noisy labeled data via maximum correntropy criterion', *IEEE Trans. Cybern.*, 2018, pp. 1–14, doi: 10.1109/TCYB.2018.2804326

[14] Shi, Q., Du, B., Zhang, L.: 'Domain adaptation for remote sensing image classification: a low-rank reconstruction and instance weighting label propagation inspired algorithm', *IEEE Trans. Geosci. Remote Sens.*, 2015, **53**, (10), pp. 5677–5689

[15] Wu, J., Pan, S., Zhu, X., *et al.*: 'Positive and unlabeled multi-graph learning', *IEEE Trans. Cybern.*, 2016, **PP**, (99), pp. 1–12

[16] Du, B., Huang, Z., Wang, N., *et al.*: 'A band-wise noise model combined with low-rank matrix factorization for hyperspectral image denoising', *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2018, **11**, (4), pp. 1070–1081

[17] Zhang, L., Zhu, X., Zhang, L., *et al.*: 'Multidomain subspace classification for hyperspectral images', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, (10), pp. 6138–6150

[18] Pan, S.J., Yang, Q.: 'A survey on transfer learning', *IEEE Trans. Knowl. Data Eng.*, 2010, **22**, (10), pp. 1345–1359

[19] Li, X., Zhang, L., Du, B., *et al.*: 'An iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification', *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2017, **10**, (5), pp. 2022–2035

[20] Ben-David, S., Blitzer, J., Crammer, K., *et al.*: 'Analysis of representations for domain adaptation', in Bernhard, S., John, P., Thomas, H. (Eds.): *Advances in neural information processing systems*, vol. **19** (MIT Press, Cambridge, MA, 2007), pp. 137–144

[21] Zhang, L., Zhu, X., Zhang, L., *et al.*: 'Multi-domain subspace classification for hyperspectral images', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, (10), pp. 6138–6150

[22] Du, B., Zhang, M.: 'PLTD: patch-based Low-rank tensor decomposition for hyperspectral images', *IEEE Trans. Multimed.*, 2017, **19**, (1), pp. 67–79

[23] Wu, J., Pan, S., Zhu, X., *et al.*: 'Boosting for multi-graph classification', *IEEE Trans. Cybern.*, 2015, **45**, (3), pp. 416–429

[24] Du, B., Wang, Z., Zhang, L., *et al.*: 'Exploring representativeness and informativeness for active learning', *IEEE Trans. Cybern.*, 2017, **47**, (1), pp. 14–26
[25] Du, B., Xiong, X., Zhang, L., *et al.*: 'Stacked convolutional denoising auto-encoders for feature representation', *IEEE Trans. Cybern.*, 2017, **47**, (4), pp. 1017–1027
[26] Bünau, P.V., Meinecke, F.C., Müller, K.R., *et al.*: 'Stationary subspace analysis', *Lect. Notes Comput. Sci.*, 2012, **5441**, (12), pp. 1–8
[27] Du, B., Wang, Z.: 'Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion', *IEEE Trans. Image Process.*, 2017, **26**, (4), pp. 1694–1707
[28] Pan, S.J., Kwok, J.T., Yang, Q.: 'Transfer learning via dimensionality reduction'. Proc. 23rd AAAI Conf. Artificial Intelligence, Chicago, IL, July 2008, pp. 677–682
[29] Pan, S.J., Tsang, I.W., Kwok, J.T., *et al.*: 'Domain adaptation via transfer component analysis', *IEEE Trans. Neural Netw.*, 2011, **22**, pp. 199–210
[30] Du, B., Zhang, Y., Zhang, L., *et al.*: 'Beyond the sparsity-based target detector: a hybrid sparsity and statistics based detector for hyperspectral images', *IEEE Trans. Image Process.*, 2016, **25**, (11), pp. 5345–5357
[31] Sun, Z., Wang, C., Wang, H., *et al.*: 'Learn multiple-kernel SVMs for domain adaptation in hyperspectral data', *IEEE Geosci. Remote Sens. Lett.*, 2013, **10**, pp. 1224–1228
[32] Si, S., Tao, D., Geng, B.: 'Bregman divergence-based regularization for transfer subspace learning', *IEEE Trans. Knowl. Data Eng.*, 2010, **22**, pp. 929–942
[33] Fisher, R.A.: 'The use of multiple measurements in taxonomic problems', *Ann. Eugenics*, 1936, **7**, pp. 179–188
[34] Yan, S., Xu, D., Zhang, B.Y., *et al.*: 'Graph embedding: a general framework for dimensionality reduction'. IEEE Computer Society Conf. Computer Vision and Pattern Recognition 2005, San Diego, CA, USA, 2005, pp. 830–837
[35] Zhang, T., Tao, D., Yang, J.: 'Discriminative locality alignment', in 10th European Conf. on Computer Vision, Marseille, France, October 12-18, 2008
[36] Shi, Q., Zhang, L., Du, B.: 'Semisupervised discriminative locally enhanced alignment for hyperspectral image classification', *IEEE Trans. Geosci. Remote Sens.*, 2013, **51**, pp. 4800–4815
[37] Landgrebe, D.A.: '*Signal theory methods in multispectral remote sensing*', vol. **29** (Wiley-Interscience, Hoboken, NJ, USA, 2005)