Scientia Iranica D (2021) 28(6), 3273–3292

Sharif University of Technology
**Scientia Iranica**
*Transactions D: Computer Science & Engineering and Electrical Engineering*
http://scientiairanica.sharif.edu

# Domain adaptation via Bregman divergence minimization

**M. Zandifar, Sh. Noori Saray, and J. Tahmoresnezhad***

*Faculty of Information Technology & Computer Engineering, Urmia University of Technology, Urmia, Iran.*

Received 25 July 2018; received in revised form 13 November 2020; accepted 5 July 2021

**KEYWORDS**
Fisher linear discriminant analysis; Transfer learning; Bregman divergence; Dimensionality reduction.

**Abstract.** In recent years, the Fisher Linear Discriminant Analysis (FLDA)-based classification models are among the most successful approaches and have shown effective performance in different classification tasks. However, when the learning data (source domain) have a different distribution compared with the testing data (target domain), the FLDA-based models may not work well, and the performance degrades, dramatically. To face this issue, we offer an optimal domain adaptation via Bregman divergence minimization (DAB) approach, in which the discriminative features of source and target domains are simultaneously learned via domain invariant representation. DAB is designed based on the constraints of FLDA, with the aim of the coupled marginal and conditional distribution shifts adaptation through Bregman divergence minimization. Thus, the resulting representation can show well functionality like FLDA and simultaneously discriminate across various classes, as well. Moreover, our proposed approach can be easily kernelized to deal with nonlinear tasks. Different experiments on various benchmark datasets demonstrate that our DAB can constructively face with the cross domain divergence and outperforms other novel state-of-the-art domain adaptation approaches in cross-distribution domains.



## 1. Introduction

In general, the conventional Machine Learning (ML) approaches have high performance when the training and test samples have the similar distribution and feature space. Given the rejection of this hypothesis in the real world applications and the lack of sufficient labeled data in domains, the traditional supervised learning methods become ineffective. To compensate for this shortcoming, we can benefit from other available and relevant labeled samples. In this case, the distribution divergence between the training and test data may reduce the efficiency of the trained model on test data [1]. For an instance, suppose that for training an adaptive model to detect the pedestrians in night-time images, the day-time labeled images are employed where they are available with different distributions [2]. In this way, the trained model will not have the acceptable classification accuracy where the training and test images contain various expressions, postures, aging, and lighting conditions. In total, the distribution difference between the training and test samples is known as domain shift problem [3].

Domain Adaptation (DA) and Transfer Learning (TL) have been introduced as effective tools to address

*. *Corresponding author. Tel.: +98 44 31980236; Fax: +98 44 31980236*
*E-mail addresses: mozhdeh.zandifar@it.uut.ac.ir (M. Zandifar); shivanoorisaray@it.uut.ac.ir (Sh. Noori Saray); j.tahmores@it.uut.ac.ir (J. Tahmoresnezhad)*

doi: 10.24200/sci.2021.51486.2210

the domain shift problem where they utilize the gained knowledge from training instances to model test ones. According to the available information in input data, the TL perspectives are categorized into following two categories: (1) unsupervised TL, where no labeled data in target domain is accessible, and (2) semi-supervised TL where a small part of target domain is labeled but the amount of labeled data is not independently adequate to construct an efficient classifier [4].

In order to cope the shift problem between domains, the type of distribution shift should be realized. The distribution divergence between domains is derived from both marginal and conditional distribution divergences between domains. Given a specific domain, from the probability viewpoint, the marginal distribution mismatch is the probability distribution difference of domain features, and the conditional distribution mismatch is the prediction probability difference of similar instances of the source and target domains. Most previous TL methods reduce the distribution difference considering one of the marginal or conditional distributions. While, our proposed method covers both the marginal and the conditional distribution difference reductions of source and target domains to tackle the distribution mismatch across domains.

Therefore, we employ a benchmark to evaluate the divergency between domains. Bregman Divergences (BD) are the generalized distance measures that are defined between matrices, functions, and distributions. The BD are related to a strictly convex functions that evaluate the distribution divergencies of instances drawn from diverse domains. In principle, the BD work like the norm in a Reproducing Kernel Hilbert Space (RKHS) on Support Vector Machine (SVM) [5], where both methods reduce the solution space volumes. The BD can transfer the knowledge across training and test samples via decreasing the distribution discrepancies of source and target data.

However, the conventional dimensionality reduction algorithms find a shared feature space without considering the distribution diversity across the source and target instances where the domains are not identically and independent distributed (i.i.d.). To solve this issue, we use a transfer subspace learning framework using BD and Fisher Linear Discriminant Analysis (FLDA) to partially preserve the discriminant information. In this work, the integration of FLDA with BD is utilized to find an optimal subspace whereas the distribution discrepancy across domains is minimized and the data from various classes are well distinguished.

Therefore, we introduce DA via BD minimization (DAB) that is a novel unsupervised TL technique, which learns a new latent feature space to reconcile both the marginal and the conditional distributions across domains. The main aims of DAB are listed in the following:

1. DAB uses the BD to transfer the discriminant knowledge from training instances to test ones by dispelling or alleviating the difference across the distribution of training and test instances;
2. DAB uses an embedding matrix to map different domains data onto a shared embedded space based on the customized FLDA criterions in an unsupervised manner;
3. DAB extends the nonparametric BD to calculate the divergence across the source and target domains, and integrates the FLDA to form a robust feature representation;
4. DAB benefits from a base classifier (i.e., nearest neighbor (NN) [6] or SVM [7]) in projected subspace to estimate the pseudo labels of test data.

In fact, our proposed approach has the following major goals: 1) preserving the discriminant knowledge of source samples to effectively transfer the class knowledge, and 2) minimizing joint distribution disparities across source and target samples to decrease the domain variation statistically.

Extensive experiments are accomplished on 34 visual DA tasks over three benchmark datasets. The experiment results demonstrate that our DAB remarkably improves the classification accuracy against other newest DA methods.

The rest of the paper is formed as follows. In Section 2, we review the related work. The general requirements are presented in Section 3. We bring forward our unsupervised transfer subspace learning method in Section 4. Experimental settings and results on visual DA tasks are demonstrated in Section 5. In the end, Section 6 concludes the paper and the future works are discussed.

## 2. Related work

TL as an advanced variant of ML has attained great success in various fields, e.g., speech recognition [8,9], text mining [10], computer vision [11,12], and ubiquitous computing [13,14] over the last two decades. The existing TL approaches are categorized into following three main groups: (1) instance-based [15], (2) model-based [16,17] and (3) feature-based [18,19] approaches.

In instance-based approaches, the weights are assigned to rate the training instances in source domain. Instance-based methods focus on reweighting the source instances to lessen the distribution divergency across domains. A main idea of instance reweighting methods is to obtain an optimal classifier for unlabeled data by embedding the instance-dependent weights in the loss function. Lately, a landmark-based method is introduced that landmarks are a subset of labeled samples in training data that have the closest distribution

to target data [20]. Landmark selection is one of the instance-based methods, which profits from Maximum Mean Discrepancy (MMD) [21] to choose a part of source samples that obey the same distribution with target samples. In the other words, landmark selection is to link the source and target data by utilizing the selected landmarks. Landmarks Selection-based Subspace Alignment (LSSA) [22] is also instance-based method which opts some of the examples as landmarks and maps the training and test data into a new feature-space nonlinearly by considering the discovered landmarks. LSSA benefits the subspace alignment to match different domains through learning a linear mapping model. A Minimax Game for instance-based selective TL (MGTL) [23] is another approach that insists to choose the high-quality source samples to boost the transferring capability across diverse domains. In contrast to selecting samples one by one, in MGTL, the actions are sampled in groups to boost the training performance.

The model-based methods aim to create a robust classifier and reduce the distribution discrepancy across domains at the same time. They adapt the source classifier parameters to classify target samples. Adaptation Regularization based TL (ARTL) [24] pursues to find an adaptive classifier using the following three components: 1) minimizing the structural risk functional, 2) matching the dual distributions across domains, 3) utilizing the manifold consistency via marginal distribution. Adaptive Classifier learning with Transfer Component Analysis (ACTCA) [25] as a model-based method decreases the distribution shift across domains and then constructs a classifier on mapped data. However, it is obvious that creating a prediction function after obtaining a new feature representation may miss some discriminative knowledge and learning two steps at the same time can conduce a better performance. Joint Adaptive Classifier and Representation Learning (JACRL) [26] as a two-step model-based method learns an adaptive classifier via minimizing the functional structural risk and obtains a representation space through the distribution shift reduction across domains to maximize the manifold consistency of classifier.

The feature-based methods aim to learn new feature representation, which can capture common knowledge across domains. Most of recent feature-based methods contain the dimensionality reduction step to provide a common latent space among domains. The aim of dimensionality reduction methods is to embed the high-dimensional samples into a low-dimensional subspace where the intrinsic information contained in the data is preserved. Due to the literature survey [27], existing dimensionality reduction techniques can be divided into two categories: (1) PCA-based (principal component analysis) framework, which attempts to project instances along the direction of maximum variance, and (2) FLDA-based framework, which increases the mean value of Kullback-Leibler (KL) divergences [28] across diverse classes.

There are variety of PCA-based approaches [29,30]. Visual DA (VDA) [31] is a new framework which constructs a shared feature representation besides minimizing the joint distributions among the source and target data. In fact, VDA maintains the statistical and geometrical information of input samples using manifold assumptions. In addition, VDA employs domain invariant clustering in an embedded subspace to distinguish diverse classes of target data. Coupled Local-Global Adaptation (CLGA) [32] is another method, which reduces the distribution gap among domains by global and local matching. At the global step, CLGA minimizes dual distribution differences across the source and target domains. At the local step, CLGA uses both the class discrimination information and data geometric structures. Discriminative and Domain Invariant Subspace Alignment (DISA) for visual tasks [33] aims to embed the various domains data into the relevant feature spaces. DISA globally matches both domains through the distribution divergence minimization between domains. In addition, DISA separates various classes by adjusting the inter-class and the intra-class distances. Also, DISA conserves the manifold knowledge of data for local adaptation.

FLDA-based approaches consider class discrimination criteria besides DA criterions to match the distribution mismatch between domains. Scatter component Analysis (SCA) [34] is an FLDA-based dimensionality reduction approach, which seeks to learn a new representation to perform both DA and domain generalization effectively. Three main objectives of SCA are as follows: (1) maximizing the inter-class scatter, (2) minimizing the intra-class scatter, and (3) maximizing the general scatter.

In this paper, we introduce a novel FLDA-based dimensionality reduction method, which finds an optimal subspace and reduces the gap in marginal and conditional distributions of various domains. DAB embeds the source and target data into a common low dimensional feature space via BD minimization and FLDA criteria.

## 3. Proposed method

### *3.1. Motivation*

We are to propose a new TL algorithm based on the dimensionality reduction process to tackle domain shift problem. The main idea of our proposed approach is inspired from transfer subspace learning [2] on which the marginal distribution difference of source and target data is minimized. Nevertheless, our proposed method

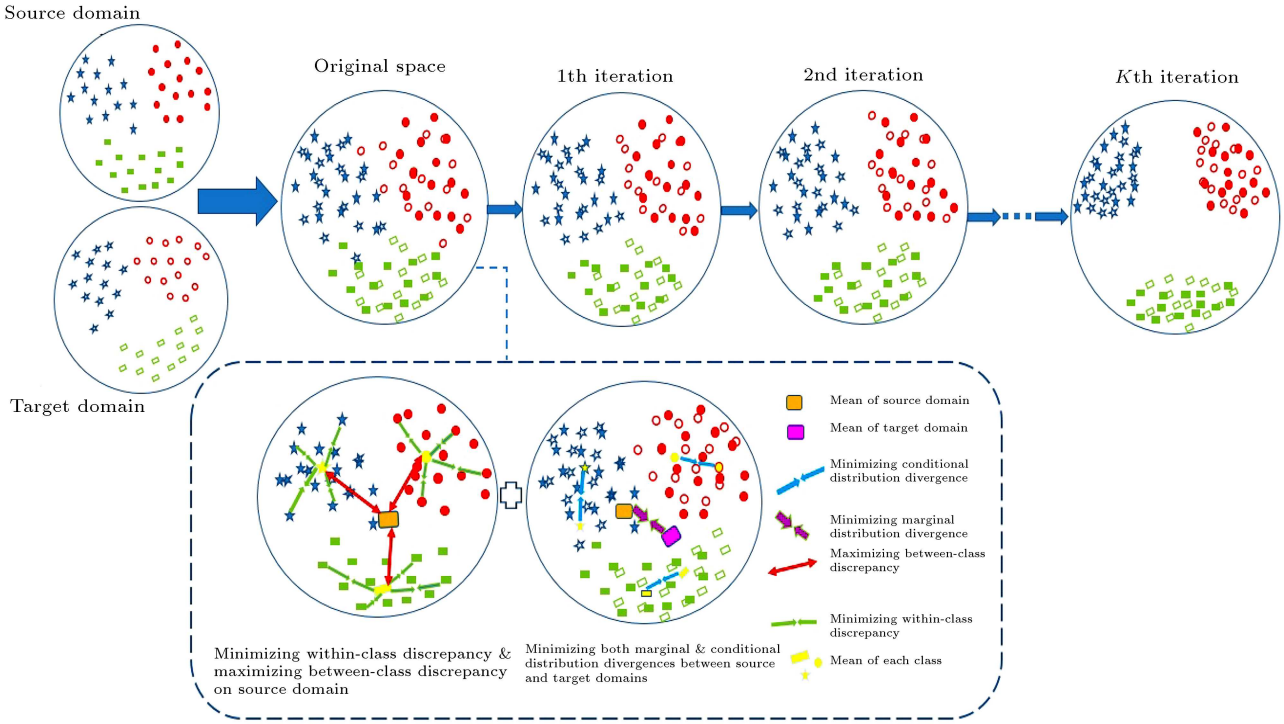



**Figure 1.** Overall scheme of Domain Adaptation via Bregman divergence minimization (DAB). DAB is a combination of Fisher Linear Discriminant Analysis (FLDA) and joint distribution adaptation via Bregman divergence. DAB uses FLDA to reduce the dimension and maintain the class structure where it increases the inter-class differences and reduces the intra-class differences. It also uses the joint distribution adaptation to lessen the gap across conditional and marginal distributions, simultaneously. In an iterative process, DAB seeks to obtain an optimal feature-space for matching the source and target domains.

minimizes joint marginal and conditional mismatches simultaneously, according to a modified version of BD to alleviate the conditional distribution discrepancy alongside the marginal distribution discrepancy. Also, DAB preserves the class structure using the FLDA-based criterions. Figure 1 illustrates the overall scheme of DAB. In the rest, we introduce the problem statement and then we represent our proposed approach.

### *3.2. Problem statement*

**Definition 1 (domain).** A domain $D$ is comprised of pairs $\mathcal{D} = \{\mathcal{X}, P(x)\}$, which is an $m$-dimensional feature space and $P(x)$ is a marginal probability distribution on $\mathcal{X}$ where $\mathcal{X} = \{x_1, \ldots, x_n\} \in \mathcal{X}$ . Input data includes the following two domains: source domain ($S$) and target domain ($T$). We mark the source domain as $\mathcal{D}_S = \{(x_1, y_1), \ldots, (x_{n_s}, y_{n_s})\}$ which is completely labeled. Similarly, we define the target domain as $\mathcal{D}_T = \{x_{n_s+1}, \ldots, x_{n_s+n_t}\}$ which is fully unlabeled. Also, $n_s$ and $n_t$ are the number of source and target samples, respectively.

**Definition 2 (task).** Given a specific domain $D$, a task for domain $D$ is denoted by $\mathcal{T} = \{\mathcal{Y}, f(x)\}$ and is composed of the following two components: $T$ is the set of labels of domain $D$ and $f(x)$ is a classifier, which can be employed to approximate the equivalent labels of the sample $x$. From a probabilistic standpoint, $f(x)$ can be expressed as the conditional probability distribution, i.e., $f(x) = Q(y \mid x)$ where $y \in \mathcal{Y}$. Our problem is to obtain a new feature subspace in which both the marginal and conditional probability differences among the source and target data are minimized, i.e., $P_s(x_s) \approx P_t(x_t)$ and $Q_s(y_s \mid x_s) \approx Q_t(y_t \mid x_t)$, respectively, where $\mathcal{X}_s = \mathcal{X}_t$ and $\mathcal{Y}_s = \mathcal{Y}_t$.

The aim of our DAB is to create an optimal subspace with the following characteristics: 1) The distances of both marginal and conditional probability of source and target data are minimized; 2) the separate information of source samples is preserved to improve the class information transfer.

### *3.3. Feature extraction using classical FLDA*

The main objective of FLDA is to perform dimensionality reduction besides preserving the class structure of information as much as possible. The class structure is considered using inter-class scatter ($S_B$) and intra-class scatter ($S_W$), which measures the difference across the various classes and the scatter of measurements around their relevant class centers. In this way, FLDA finds a feature space, which reduces the trace ratio of intra-class scatter and inter-class scatter matrices. Mathe-

matically, the inter-class scatter matrix is presented in Eq. (1):

$$S_B = \sum_{i=1}^{C} n_i(\vec{m}^{(i)} - m)(\vec{m}^{(i)} - m)^T, \quad (1)$$

and the intera-class scatter matrix is formulated as:

$$S_W = \sum_{i=1}^{C}\sum_{j=1}^{n_i}(x_j^{(i)} - \vec{m}^{(i)})(x_j^{(i)} - \vec{m}^{(i)})^T, \quad (2)$$

where $C$ demonstrates the number of classes, $n_i$ is the number of instances in $i$th class, $\vec{m}^{(i)}$is the mean of samples from the $i$th class, $x_j^{(i)}$ is $j$th sample from $i$th class and $m$ is the mean of all instances. Therefore, the FLDA subspace is calculated as follows:

$$F(W) = argmin\frac{tr(W^T S_W W)}{tr(W^T S_B W)}$$

$$= argmin \;\; tr^{-1}(W^T S_B W)tr(W^T S_W W), \quad (3)$$

where $tr^{-1}(X)$ is the inverse of matrix trace.

### 3.4. Bregman Divergence (BD)

To diminish the distribution shift across the source and target data, most of the previous studies have focused to design and optimize the objective functions via Euclidean distance to calculate the distribution discrepancy across domains. However, in some real-world applications, Euclidean distance may be unsuitable to measure the distribution gap between domains, since it cannot jointly increase the inter-class distance and decrease the intra-class distance. Thus, Bregman distance is an appropriate nonlinear distance function to calculate the distribution difference among the source and target domains. BD transfers the achieved knowledge from training instances to test instances by reducing a distance across the dual distributions of domains. We now introduce the BD with more details.

**Definition 3 (Bregman divergence).** Let $\Omega \to R$ be a strictly convex function on a convex set $\Omega \subseteq R^m$ supposed to be nonempty and differentiable. Then, for every $x, y \in R^m$ the BD corresponding to $\psi$ is formulated as:

$$BD(x,y)_\psi = \psi(x) - \psi(y) - < x - y, \bigtriangledown\psi(y) >, \quad (4)$$

where $\psi(.)$ is the mapping function and $\bigtriangledown\psi$ illustrates the gradient vector of $\psi$.

Therefore, the BD can be explained as the gap among the value of convex function at $x$ and its first order Taylor expansion at $y$, or correspondingly the remainder part of the first order Taylor expansion of $\psi$ at $y$. The geometric interpretation of BD is demonstrated in Figure 2. As illustrated in Figure 2, it is obvious that the BD calculates the ordinate distance among the value of convex function at $x$ and its tangent at $y$.

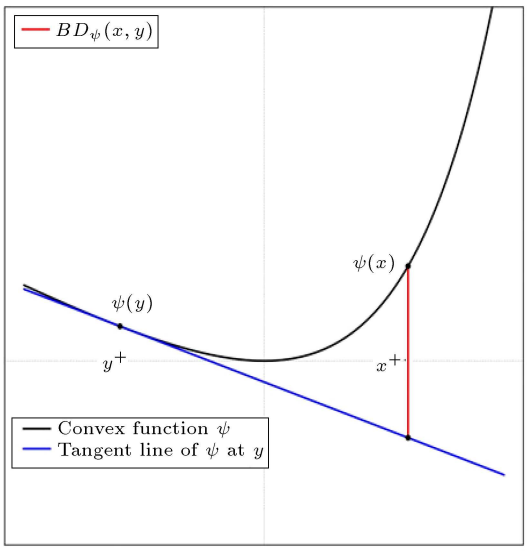



**Figure 2.** Geometrical illustration of Bregman divergence, $BD_\psi(x, y)$.

By setting $\psi(x) = x^2$ and $\psi(y) = y^2$, the BD is diminished to the form of squared error loss [35]. By employing the squared loss to calculate the distribution discrepancy, the following relation is achieved:

$$D(W, X_S, X_T) = \int (P_S(\vec{y}) - P_T(\vec{y}))^2 d\vec{y}, \quad (5)$$

where $X_S$ and $X_T$ are the source and target sample sets, $W$ is the projected subspace, $D(.)$ is the BD which calculates the gap among $P_S$ and $P_T$, $P_S(.)$ and $P_T(.)$ demonstrate the Probability Density Functions (PDFs) of the source and target samples in the embedded subspace, respectively. Therefore, the BD is a criterion of diversity in distributions of source and target instances in an embedded feature space.

The densities in the embedded feature spaces are calculated by utilizing the Kernel Density Estimation (KDE) method [36], which computes the density at each point $y \in R^d$ as a sum of kernels across $\vec{y}$ and other points $\vec{y}_i$, as follows:

$$p(\vec{y}) = (\frac{1}{n})\mathbf{G}_{\sum}(\vec{y} - \vec{y}_i), \quad (6)$$

where $n$ displays the number of instances and $\mathbf{G}_{\sum}(.)$ shows the $d$-dimensional Gaussian kernel with covariance matrix $\sum$. We extend the nonlinear BD to compute the discrepancy of both the marginal and the conditional distributions, and merge it with FLDA to create a new latent feature space, which is advantageous and robust for significant distribution divergence.

### 3.5. Marginal distribution adaptation

The existing dimensionality reduction methods obtain a linear combination of features that characterize or separate two or more classes of objects or events. However, they cannot guarantee to decrease the distribution mismatch across various domains in embedded subspace [37]. Thus, we are to extend an optimal dimensionality reduction paradigm to transfer knowledge across domains. We explicitly leverage the nonparametric distance measure, i.e., BD, in a projected subspace to assess the distance across expectations of source and target domains. Formally, the marginal distribution distance of domains is formulated in Eq. (7):

$$\begin{aligned} &Dist^{marginal}(D_S, D_T) \\ &= \left[\int (\frac{1}{n_s}\sum_{i=1}^{n} \mathbf{G}_{\sum_1}(\vec{y}-\vec{y}_i))^2 d\vec{y}\right] \\ &+ \left[\int (\frac{1}{n_t}\sum_{j=n_s+1}^{n} \mathbf{G}_{\sum_2}(\vec{y}-\vec{y}_i))^2 d\vec{y}\right] \\ &- \left[(\frac{1}{n_s n_t}\sum_{i=1}^{n}\sum_{j=n_s+1}^{n} \mathbf{G}_{\sum_{12}}(\vec{y}-\vec{y}_i))^2\right], \end{aligned} \tag{7}$$

where $Dist^{marginal}(D_S, D_T)$ measures the gap of marginal distributions across the source domain $D_S$ and target domain $D_T$. Also, $\mathbf{G}_{\sum_1}$, $\mathbf{G}_{\sum_2}$, and $\mathbf{G}_{\sum_{12}}$ are Gaussian kernel with $d$ dimensions for the covariance matrices of source instances, target instances and sum of source and target instances, respectively. The distance across the marginal distributions $P(X_S)$ and $P(X_T)$ is decreased by minimizing $Dist^{marginal}(D_S, D_T)$.

### 3.6. Conditional distribution adaptation

Most of the novel DA researches consider only the marginal probability distribution adaptation, supposing that the domains obey the same conditional probability distributions. While in factual examples, the conditional probability distribution discrepancies are remarkable. Moreover, alleviating the discrepancy of marginal distributions does not necessarily decrease the conditional distribution discrepancies. On the other hand, the computation of conditional distribution discrepancy is intractable, hence the class-conditional distribution is calculated instead [38]. Thus, we consider both the marginal and the class-conditional distributions in DA. Thus, the gap of class-conditional distributions is computed via the sum of empirical distances with respect to the class labels across the sub-domains with the same label between the source and target domains:

$$\begin{aligned} &Dist^{conditional}\sum_{c=1}^{C}(D_{S^c}, D_{T^c}) \\ &= \left[\int (\frac{1}{n_s^c}\sum_{i=1}^{n_s^c} \mathbf{G}_{\sum_1}(\vec{y}-\vec{y}_i))^2 d\vec{y}\right] \\ &+ \left[\int (\frac{1}{n_t^c}\sum_{j=n_s^c+1}^{n_s^c+n_t^c} \mathbf{G}_{\sum_2}(\vec{y}-\vec{y}_i))^2 d\vec{y}\right] \\ &- \left[(\frac{1}{n_s^c n_t^c}\sum_{i=1}^{n_s^c}\sum_{j=n_s^c+1}^{n_s^c+n_s^c} \mathbf{G}_{\sum_{12}}(\vec{y}-\vec{y}_i))^2\right], \end{aligned} \tag{8}$$

where $Dist^{conditional}\sum_{c=1}^{C}(D_{S^c}, D_{T^c})$ demonstrates the shift of class-conditional distributions among the source domain $D_S$ and the target domain $D_T$. Also, $n_s^c$ and $n_t^c$ indicate the samples number with the class $c$ in the source and target domains, respectively. Moreover, $D_{S^c}$ shows the sample set of source domain with class $c$, and $D_{T^c}$ is the sample set of target domain with class $c$. With minimizing $Dist^{conditional}\sum_{c=1}^{C}(D_{S^c}, D_{T^c})$, the discrepancy of conditional distributions across $D_{S^c}$ and $D_{T^c}$ is minimized.

### 3.7. Objective function

The intuition behind DAB is to learn a transfer matrix $W \in R^{D\times d}$, which maps the samples from $D$-dimensional space to $d$-dimensional space and persuades three following principal objectives: (1) the marginal distribution variation of source and target data is decreased, (2) the conditional distribution variation between the same class of source and target data is minimized, and (3) the discriminative information (i.e., data manifold structure) is preserved. Therefore, we embed Eqs. (7) and (8) into Eq. (3), to find the following optimization problem:

$$W = argmin F(W) + \lambda D(W, X_S, X_T), \quad \text{s.t.} \quad W^T W = I, \tag{9}$$

where $F(W)$ is the objective function of FLDA, $D(W, X_S, X_T)$ is the BD which computes the shift across $P_S$ and $P_T$ in an embedded subspace via $W$, and $\lambda \in [0, 1]$ demonstrates the regularization parameter which supervises the balance among $F(W)$ and $D(W, X_S, X_T)$. The solution of Eq. (9) can be found via the following gradient descent technique:

$$\begin{aligned} W_{k+1} &= W_k - \eta(k)\left(\frac{\partial F(W)}{\partial W}\right) \\ &+ \lambda\left(\sum_{i=1}^{n_s+n_t}\frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i}\frac{\partial \vec{y}_i}{\partial W}\right. \\ &\left. + \sum_{j=1}^{n_s^c+n_t^c}\frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i}\frac{\partial \vec{y}_i}{\partial W}\right), \end{aligned} \tag{10}$$

where $\partial$ is the gradient, $W_k$ is the primary subspace; $W_{k+1}$ is the update of $W_k$ to find the optimal linear subspace using the gradient descent method; and $\eta(k)$ illustrates the learning rate factor at the iteration $k$, which supervises the gradient step size for $k$th iteration.

The derivative of $F(W)$ due to $W$ is achieved via:

$$\frac{\partial F(W)}{\partial W} = 2tr^{-1}(W^T S_B W) S_W W - 2tr^{-2}(W^T S_B W) tr(W^T S_W W) S_B W, \quad (11)$$

where $tr^{-2}(X)$ displays the inverse of $tr(X)$ square root. We define the following equation inspired from [39] as:

$$\int \mathbf{G}_{\sum_1}(\vec{y} - \vec{y}_s) \sum_2 (\vec{y} - \vec{y}_t) dy = \mathbf{G}_{\sum_1 + \sum_2 (\vec{y}_s - \vec{y}_t)},$$

which belongs to two Gaussian kernels due to the quadratic form of Eqs. (7) and (8), where the derivative of $D(W, X_S, X_T)$ considering $W$ is:

$$\begin{aligned}
&\sum_{i=1}^{n_s+n_t} \frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i} \frac{\partial \vec{y}_i}{\partial W} = \sum_{i=1}^{n_s} \frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i} \vec{x}_i^T \\
&+ \sum_{i=1}^{n_s+n_t} \frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i} \vec{x}_i^T \\
&= \frac{1}{n_s^2} \sum_{s=1}^{n_s} \sum_{t=1}^{n_t} \mathbf{G}_{\sum_{11}}(\vec{y}_s - \vec{y}_t) \\
&+ \frac{1}{n_t^2} \sum_{s=1}^{n_s+n_t} \sum_{t=1}^{n_s+n_t} \mathbf{G}_{\sum_{22}}(\vec{y}_s - \vec{y}_t) \\
&- \frac{1}{n_s n_t} \sum_{s=1}^{n_s} \sum_{t=n_s+1}^{n_s+n_t} \mathbf{G}_{\sum_{12}}(\vec{y}_s - \vec{y}_t), \quad (12)
\end{aligned}$$

and:

$$\begin{aligned}
&\sum_{i=1}^{n_s^c+n_t^c} \frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i} \frac{\partial \vec{y}_i}{\partial W} = \sum_{i=1}^{n_s^c} \frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i} \vec{x}_i^T \\
&+ \sum_{i=n_s^c+1}^{n_s^c+n_t^c} \frac{\partial D(W, X_S, X_T)}{\partial \vec{y}_i} \vec{x}_i^T \\
&= \frac{1}{(n_s^c)^2} \sum_{s=1}^{n_s^c} \sum_{t=1}^{n_t^c} \mathbf{G}_{\sum_{11}}(\vec{y}_s - \vec{y}_t) \\
&+ \frac{1}{(n_t^c)^2} \sum_{s=n_s^c+1}^{n_s^c+n_t^c} \sum_{t=n_s^c+1}^{n_s^c+n_t^c} \mathbf{G}_{\sum_{22}}(\vec{y}_s - \vec{y}_t) \\
&- \frac{1}{n_s^c n_t^c} \sum_{s=1}^{n_s^c} \sum_{t=n_s^c+1}^{n_s^c+n_t^c} \mathbf{G}_{\sum_{12}}(\vec{y}_s - \vec{y}_t), \quad (13)
\end{aligned}$$

where:

$$\sum_{11} = \sum_1 + \sum_1, \qquad \sum_{22} = \sum_2 + \sum_2 \quad \text{and}$$

$$\sum_{12} = \sum_1 + \sum_2.$$

Algorithm 1 illustrates the whole procedure of DAB. Moreover, we utilize the pseudo-labels of target domain, which can be acquired by training a simple prediction function on the training samples to estimate the pseudo-labels of test samples. The prediction function could be a common ML classifier such as NN and SVM. It should be pointed out that, although most of the pseudo test labels may be false because of the discrepancies in distributions, we can still employ them to adapt the conditional distributions by modified BD measure formulated in Eq. (8). Thereby, we can apply the source classifier to boost the target classifier.

### *3.8. Computational complexity*

In this section, we look over the computational complexity of DAB by utilizing the big $O$ notation. We mark $n$ and $C$ as the instances number and the classes number, respectively. The computational complexity is calculated in the following: $O(n^2)$ for measuring the marginal distribution discrepancy across the source and target samples, $(Cn^2)$ for measuring the conditional distribution difference among the source and target samples, and $O(n^2)$ for computing matrix $W_{k+1}$. Consequently, the computational complexity of DAB is $O(Cn^2)$.

## 4. Experiments

In this section, at first we introduce the benchmark datasets. Then, the experimental setup and details of our DAB and other compared approaches are described. Finally, the results are discussed with details.

### *4.1. Datasets*

We employ the following three visual DA datasets as the popular benchmark to assess the visual DA algorithms: object (Office-Caltech-256), face (PIE) and digit (USPS, MNIST). We used the prepared datasets in [40] and provided 34 tasks. Table 1 shows the results of experiments.

#### *4.1.1. Office-Caltech-256 dataset*

The Office-Caltech-256 dataset is a benchmark dataset for visual DA tasks, which has 10 overlapping classes from following four domains: Amazon (A), Webcam (W), DSLR (D) and Caltech256 (C). Image-samples of Amazon and Caltech-256 domains are taken from amazon.com and office equipment, respectively. The Webcam and DSLR domains include images taken by

**Input:** training samples (source) $X_s$, test samples (target) $X_t$, labels of training samples $y_s$, regularization parameter $\lambda \in [0\ 1]$
**Output:** labels of test samples$y_t$
train 1-NN classifier $f$ on $(X_s, Y_s)$
estimate the pseudo labels of test data $(X_t)$, i.e., $Y_{t_0}$, by prediction function $f$
obtain $S_B$ via Equation 1
obtain $S_W$ via Equation 2
obtain $F(W_{DAB}, X_s)$ via Equation 3
**iterate to find accurate labels**
update matrix $W_{k+1}$ by Equation 10
solve Equation 10 and choose eigenvectors
train the classifier $f$ on $(W^T X_s, Y_s)$
update pseudo labels $Y_{t_0}$ on $(W^T X_t)$
**end iterate**
train final classifier on $(W^T X_s, Y_s)$
estimate the labels of $X_t$ by utilizing the ultimate prediction function
return the test data labels $y_t$ assigned by the ultimate prediction function $f$

**Algorithm 1.** Domain Adaptation via Bregman divergence minimization (DAB).

**Table 1.** The specifications of investigative datasets.

| Dataset | #Type | #Examples | #Features | #Classes | Subsets |
|---|---|---|---|---|---|
| *USPS* | Digit | 1,800 | 256 | 10 | U |
| *MNIST* | Digit | 2,000 | 256 | 10 | M |
| *PIE* | Face | 11,554 | 1,024 | 68 | P1, P2, P3, P4, P5 |
| *Office* | Object | 1,410 | 800 | 10 | A, W, D |
| *Caltech* | Object | 1,123 | 800 | 10 | C |

a webcam and a DSLR cameras, respectively. Each domain is collected under different conditions (i.e., pose, resolution, location, view angle, scene illumination, motion blur, and background clutter between scenes), which causes a distribution difference across disparate domains. Therefore, 12 cross DA tasks are provided, namely $A \rightarrow W, ..., C \rightarrow D$.

*4.1.2. USPS-MNIST dataset*

USPS (U) and MNIST (M) are common datasets which is utilized are pattern recognition and computer vision tasks. USPS dataset contains 9,298 labeled images, each of which with size of $16 \times 16$. The MNIST dataset contains 60,000 labeled images and 10,000 unlabeled images, each of which with size of $28 \times 28$. Notice that USPS and MNIST datasets are subject to disparate distributions and both have 10 classes. One USPS versus MNIST task is provided by randomly picking out 1800 images of USPS and 2000 images of MNIST as the training and test domains, respectively. As the same way, MNIST versus USPS task is created by switching the source and target domains. Therefore, we conduct the following two handwriting recognition tasks, i.e., USPS-MNIST and MNIST-USPS.

*4.1.3. PIE dataset*

PIE is another visual dataset, which has 41,368 face images with size of $32 \times 32$. The images were taken by 13 synchronized cameras and 21 flashes under varying illuminations, poses, and expressions. PIE dataset contains the following five domains based on the various poses: PIE1 (C05, left pose), PIE2 (C07, upward pose), PIE3 (C09, downward pose), PIE4 (C27, frontal pose), and PIE5 (C29, right pose). Accordingly, 20 cross domain tasks are constructed as follow: $P1 \rightarrow P2$, $P1 \rightarrow P3, ..., P5 \rightarrow P4$.

### *4.2. Comparison baselines*

We compare DAB with two baseline ML methods, i.e., NN, FLDA and eight novel DA methods according to the mentioned datasets (Joint Distribution Adaptation (JDA) [41], Transfer Joint Matching (TJM) [42], discriminative transfer subspace learning via Low-Rank and Sparse Representation (LRSR) [43], JACRL [26],

**Table 2.** Label prediction accuracy (%) on Office+Caltech-256 datasets. Domain Adaptation via Bregman divergence minimization (DAB) has superior performance compared with other Domain Adaptation (DA) methods (i.e., Fisher Linear Discriminant Analysis (FLDA), Joint Distribution Adaptation (JDA), Transfer Joint Matching (TJM), discriminative transfer subspace learning via Low-Rank and Sparse Representation (LRSR), Joint Adaptive Classifier and Representation Learning (JACRL), Visual Domain Adaptation (VDA), Coupled Local-Global Adaptation (CLGA), discriminative and Domain Invariant Subspace Alignment for visual tasks (DISA) and discriminative joint probability maximum mean discrepancy for domain adaptation (JPDA)) in 9 out of 12 tasks using Nearest Neighbor (NN) classifier, and 9 out of 12 tasks using Support Vector Machine (SVM) classifier. $DAB_M$ is the variant of DAB that performs marginal distribution matching without considering the conditional distribution mismatching and $DAB_C$ is the variant of DAB that performs conditional distribution matching without considering the marginal distribution mismatching.

| Dataset | NN classifier | | | | | | | | | | | | | SVM classifier | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NN | FLDA | JDA | TJM | LRSR | JACRL | VDA | CLGA | $DAB_M$ | $DAB_C$ | DISA | JPDA | DAB | LRSR | JACRL | DAB |
| $C \longrightarrow A$ | 23.70 | 40.22 | 44.78 | 46.76 | 51.25 | 56.26 | 46.14 | 48.2 | 52.3 | 53.4 | **57.93** | 47.6 | 51.43 | 53.34 | **55.53** | 51.07 |
| $C \longrightarrow W$ | 25.76 | 40.11 | 41.69 | 39.98 | 38.64 | 47.8 | 46.1 | 42.37 | 47.8 | 48.9 | 49.15 | 45.76 | **50.98** | 45.76 | 45.42 | **50.72** |
| $C \longrightarrow D$ | 25.48 | 39.99 | 45.22 | 44.59 | 47.13 | 43.95 | 51.59 | 49.04 | 45.86 | 47.6 | 49.04 | 46.5 | **53.43** | 50.96 | 45.86 | **53.18** |
| $A \longrightarrow C$ | 26 | 41.36 | 39.36 | 39.45 | 43.37 | 42.65 | 42.21 | 42.3 | 41.85 | 42.87 | 39.36 | 40.78 | **59.99** | 44.70 | 42.3 | **59.01** |
| $A \longrightarrow W$ | 29.83 | 41.65 | 37.97 | 42.03 | 36.61 | 41.69 | 51.19 | 41.36 | 39.39 | 40.43 | 50.51 | 40.68 | **54.11** | 38.31 | 43.73 | **54.22** |
| $A \longrightarrow D$ | 25.48 | 40.89 | 39.49 | 45.22 | 38.85 | 43.31 | 48.41 | 36.31 | 43.95 | 45.05 | 50.96 | 36.94 | **54.69** | 39.49 | 42.04 | **53.12** |
| $W \longrightarrow C$ | 19.86 | 40 | 31.17 | 30.19 | 29.83 | 34.64 | 27.6 | 32.95 | 33.39 | 35.48 | 34.02 | 34.55 | **51.78** | 30.28 | 35.26 | **52.09** |
| $W \longrightarrow A$ | 22.96 | 42.90 | 32.78 | 29.96 | 34.13 | 39.25 | 26.1 | 34.57 | 37.68 | 38.79 | 42.48 | 33.82 | **49.32** | 34.66 | 39.25 | **51.13** |
| $W \longrightarrow D$ | 59.24 | 41.52 | 89.17 | 89.17 | 82.80 | 85.99 | 89.18 | **92.36** | 69.43 | 73.72 | 90.45 | 88.54 | 49.63 | 82.80 | **86.62** | 49.74 |
| $D \longrightarrow C$ | 26.27 | 43.21 | 31.52 | 31.43 | 31.61 | 35.17 | 31.26 | 33.66 | 33.04 | 34.15 | 32.15 | 34.73 | **53.67** | 30.72 | 34.82 | **53.71** |
| $D \longrightarrow A$ | 28.5 | 42.56 | 33.09 | 32.78 | 33.19 | 37.89 | 37.68 | 35.99 | 36.43 | 37.55 | 39.35 | 34.66 | **46.35** | 33.19 | 37.47 | **48.47** |
| $D \longrightarrow W$ | 63.39 | 42.91 | 89.49 | 85.42 | 77.29 | 89.15 | 90.85 | 89.83 | 77.97 | 81.79 | **93.22** | 91.19 | 79.28 | 76.61 | **89.15** | 80.29 |
| Average | 31.37 | 41.38 | 46.31 | 46.45 | 45.39 | 49.81 | 49.02 | 48.33 | 46.59 | 48.31 | 52.38 | 48.29 | **54.55** | 46.73 | 49.78 | **54.72** |

VDA [31], CLGA [32], DISA [33] and discriminative joint probability MMD for DA (JPDA) [44]). Since these methods are considered as dimensionality reduction approaches, we train a prediction function on the labeled training data (i.e., NN and SVM classifier), and then apply it on test data to estimate the labels of target domain.

### *4.3. Implementation details*

In order to assess the efficiency of DAB against existing methods, the classification accuracy is utilized as the analysis criterion. The iterations number for DAB convergence is set to 20 and the regularization parameter is adjusted to $\lambda = 0.5$ (i.e., in Eq. (9)) for all datasets. In the rest, we will discuss about the parameter analysis.

## 5. Experimental results and discussion

In this section, we compare the efficiency of our proposed method with other ML and DA approaches on benchmark visual DA datasets.

### *5.1. Results evaluation*

Since the discussions of the experiment results with NN and SVM classifiers are similar, for the sake of page limitation, only the discussions of the experiment results with NN classifier are included.

The classification accuracies of the proposed DAB and other baseline methods are illustrated in Tables 2, 3 and 4. The best results for each cross-domain adaptation task is denoted.

**Experiments on Office+Caltech-256 datasets:** The classification accuracy of DAB and other methods on Office+Caltech datasets is reported in Table 2. In order to interpret better, the results are visualized in Figure 3. DAB gains the best performance in term of the average classification accuracy (54.55%). DAB obtains (17.24%) performance improvement compared to NN and (2.17%) improvement against the best compared method DISA on Office+Caltech dataset. The efficiency enhancement of DAB against FLDA is (13.17%) on Office+Caltech datasets.

**Experiments on USPS+MNIST datasets:** The classification accuracies of DAB and other basic methods on two cross-domain hand-written digit recognition tasks are demonstrated in Table 3. The results are figured in Figure 4 for better clarification. DAB gains best efficiency in terms of the average classification accuracy (76.54%) where it performs better than other novel DA methods in 1 out of 2 DA sub-problems. Moreover, due to the mismatched distribution among training and test datasets, the efficiency improvement of DAB over NN is (21.22%). In comparison to the best method DISA, DAB achieves (1.65%) performance improvement on USPS+MNIST dataset. The

**Table 3.** Label prediction accuracy (%) on USPS+ MNIST datasets. Domain Adaptation via Bregman divergence minimization (DAB) has superior performance compared with other Domain Adaptation (DA) methods (i.e., Fisher Linear Discriminant Analysis (FLDA), Joint Distribution Adaptation (JDA), Transfer Joint Matching (TJM), discriminative transfer subspace learning via Low-Rank and Sparse Representation (LRSR), Joint Adaptive Classifier and Representation Learning (JACRL), Visual Domain Adaptation (VDA), Coupled Local-Global Adaptation (CLGA), discriminative and Domain Invariant Subspace Alignment for visual tasks (DISA) and discriminative joint probability maximum mean discrepancy for domain adaptation (JPDA)) in 1 out of 2 tasks using Nearest Neighbor (NN) classifier, and 2 out of 2 tasks using Support Vector Machine (SVM) classifier. $DAB_M$ is the variant of DAB that performs marginal distribution matching without considering the conditional distribution mismatching and $DAB_C$ is the variant of DAB that performs conditional distribution matching without considering the marginal distribution mismatching.

| Dataset | NN classifier | | | | | | | | | | | | | SVM classifier | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NN | FLDA | JDA | TJM | LRSR | JACRL | VDA | CLGA | $DAB_M$ | $DAB_C$ | DISA | JPDA | DAB | LRSR | JACRL | DAB |
| $U \longrightarrow M$ | 44.7 | 73.51 | 59.65 | 52.25 | 54.51 | 42.18 | 62.95 | 58.35 | 60.54 | 66.05 | 69.90 | 59.2 | **73.75** | 53.83 | 41.8 | **72.05** |
| $M \longrightarrow U$ | 65.94 | 64.89 | 67.28 | 63.28 | 73.82 | 63.56 | 74.72 | 71.28 | 73.21 | 74.11 | **83.89** | 68.94 | 83.33 | 71.98 | 63.17 | **76.22** |
| Average | 55.32 | 58.31 | 63.46 | 57.76 | 64.16 | 52.85 | 68.83 | 64.81 | 66.87 | 70.08 | 76.89 | 64.26 | **78.54** | 62.90 | 52.48 | **74.13** |

**Table 4.** Label prediction accuracy (%) on PIE datasets. Domain Adaptation via Bregman divergence minimization (DAB) has superior performance compared with other Domain Adaptation (DA) methods (i.e., Fisher Linear Discriminant Analysis (FLDA), Joint Distribution Adaptation (JDA), Transfer Joint Matching (TJM), discriminative transfer subspace learning via Low-Rank and Sparse Representation (LRSR), Joint Adaptive Classifier and Representation Learning (JACRL), Visual Domain Adaptation (VDA), Coupled Local-Global Adaptation (CLGA), Discriminative and Domain Invariant Subspace Alignment for visual tasks (DISA) and discriminative joint probability maximum mean discrepancy for domain adaptation (JPDA)) in 11 out of 20 tasks using Nearest Neighbor (NN) classifier, and 18 out of 20 tasks using Support Vector Machine (SVM) classifier. $DAB_M$ is the variant of DAB that performs marginal distribution matching without considering the conditional distribution mismatching and $DAB_C$ is the variant of DAB that performs conditional distribution matching without considering the marginal distribution mismatching.

| Dataset | NN classifier | | | | | | | | | | | | | SVM classifier | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NN | FLDA | JDA | TJM | LRSR | JACRL | VDA | CLGA | $DAB_M$ | $DAB_C$ | DISA | JPDA | DAB | LRSR | JACRL | DAB |
| $P1 \longrightarrow P2$ | 26.09 | 33.89 | 58.81 | 23.87 | 65.87 | 51.2 | 73.48 | 67.83 | 37.63 | 69.61 | **77.29** | 59.36 | 73.15 | 65.44 | 51.75 | **71.02** |
| $P1 \longrightarrow P3$ | 26.59 | 33.56 | 54.23 | 28.86 | 64.09 | 57.6 | 62.92 | 63.85 | 47.55 | 61.71 | **74.69** | 66.67 | 71.28 | 62.87 | 57.05 | **66.95** |
| $P1 \longrightarrow P4$ | 30.67 | 32.93 | 84.5 | 43.37 | 82.03 | 86.66 | 90.51 | 88.95 | 69.18 | 47.32 | 91.35 | 83.99 | **95.32** | 81.29 | 85.94 | **91.56** |
| $P1 \longrightarrow P5$ | 16.67 | 38.79 | 49.75 | 19.3 | 54.90 | 52.39 | 57.29 | 61.76 | 36.09 | 61.42 | 63.30 | 49.51 | **72.17** | 54.23 | 51.96 | **69.38** |
| $P2 \longrightarrow P1$ | 24.49 | 35.29 | 57.62 | 26.14 | 45.54 | 65.55 | 70.02 | 71.4 | 38.6 | 76.36 | 80.25 | 63 | **80.56** | 45.59 | 65.1 | **75.65** |
| $P2 \longrightarrow P3$ | 46.63 | 34.78 | 62.93 | 37.93 | 53.49 | 68.5 | 73.04 | 72.98 | 44.91 | 68.67 | **81.31** | 60.85 | 74.37 | 52.70 | 67.83 | **71.48** |
| $P2 \longrightarrow P4$ | 54.07 | 35.17 | 75.82 | 50.53 | 71.43 | 81.71 | 84.29 | 86.24 | 69.84 | 82.41 | 90.90 | 77.05 | **95.2** | 72.24 | 81.47 | **89.96** |
| $P2 \longrightarrow P5$ | 26.53 | 32.41 | 39.89 | 21.63 | 47.97 | 54.53 | 54.66 | 51.23 | 34.01 | 64.56 | 69.91 | 47.67 | **74.36** | 48.41 | 54.66 | **67.76** |
| $P3 \longrightarrow P1$ | 21.37 | 37.36 | 50.96 | 28.66 | 52.49 | 69.39 | 67.35 | 70.17 | 44.48 | 74.61 | 81.12 | 59.78 | **86.05** | 53.30 | 68.61 | **81.49** |
| $P3 \longrightarrow P2$ | 41.01 | 37.03 | 57.95 | 35.97 | 55.56 | 61.76 | 70.41 | 73.48 | 40.7 | 72.93 | 81.52 | 63.35 | **82.42** | 56.97 | 60.96 | **77.78** |
| $P3 \longrightarrow P4$ | 46.53 | 38.45 | 68.45 | 51.97 | 77.50 | 89.74 | 84.47 | 69.81 | 69.42 | 81.23 | **93.45** | 74.47 | 92.09 | 75.94 | 89.4 | **92.47** |
| $P3 \longrightarrow P5$ | 26.23 | 32.59 | 39.95 | 25.31 | 54.11 | 60.54 | 52.39 | 55.51 | 46.81 | **77.54** | 75.37 | 52.7 | 73.39 | 53.43 | 59.62 | **63.41** |
| $P4 \longrightarrow P1$ | 32.95 | 34.53 | 80.58 | 45.71 | 81.54 | 89.08 | 91.6 | 89.56 | 72.09 | 89.41 | 94.8 | 84.87 | **95.06** | 79.71 | 88.63 | **95.69** |
| $P4 \longrightarrow P2$ | 62.68 | 35.21 | 82.63 | 57.58 | 58.39 | 85.64 | 91.47 | 92.94 | 65.56 | 83.11 | **95.89** | 83.24 | 93.58 | 87.23 | 84.78 | **93.78** |
| $P4 \longrightarrow P3$ | 73.22 | 34.96 | 87.25 | 71.63 | 82.23 | 86.34 | 90.93 | 93.08 | 72.79 | 78.56 | **95.04** | 87.44 | 92.67 | 81.13 | 85.29 | **92.88** |
| $P4 \longrightarrow P5$ | 37.19 | 34.80 | 54.66 | 30.94 | 72.61 | 76.04 | 63.36 | 71.63 | 59.25 | 65.28 | **82.60** | 65.38 | 73.17 | 71.02 | **75.43** | 73.67 |
| $P5 \longrightarrow P1$ | 18.49 | 31.81 | 46.46 | 27.13 | 52.19 | 71.94 | 55.7 | 57.68 | 39.83 | 62.48 | 63.90 | 53.63 | **78.11** | 51.80 | **72.09** | 71.28 |
| $P5 \longrightarrow P2$ | 24.1 | 29.28 | 42.05 | 22.65 | 49.41 | 47.45 | 61.57 | 55.43 | 34.62 | 71.32 | 73.36 | 51.32 | **76.3** | 50.09 | 46.35 | **72.43** |
| $P5 \longrightarrow P3$ | 28.31 | 34.06 | 53.31 | 28.86 | 58.45 | 65.5 | 55.58 | 58.03 | 50.67 | 68.26 | **76.78** | 55.76 | 75.08 | 58.09 | 64.28 | **66.57** |
| $P5 \longrightarrow P4$ | 31.24 | 29.12 | 57.01 | 32.59 | 64.31 | 79.9 | 68.82 | 71.85 | 62.06 | 78.38 | 76.74 | 58.49 | **86.34** | 66.09 | 79.03 | **80.89** |
| Average | 34.76 | 34.30 | 60.24 | 35.53 | 63.53 | 70.07 | 70.99 | 72.15 | 51.81 | 71.75 | 80.95 | 64.62 | **82.03** | 63.38 | 69.51 | **78.30** |

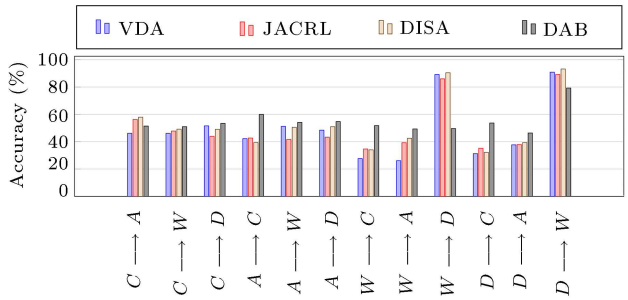



**Figure 3.** Label prediction accuracy (%) on Office+Caltech-256 datasets. Domain Adaptation via Bregman divergence minimization (DAB) has superior performance compared with other domain adaptation (DA) methods (i.e., Visual Domain Adaptation (VDA), Joint Adaptive Classifier and Representation Learning (JACRL), discriminative and Domain Invariant Subspace Alignment for visual tasks (DISA)) in 9 out of 12 tasks using Nearest Neighbor (NN) classifier.

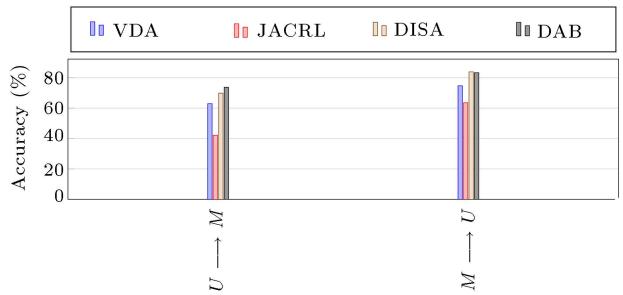



**Figure 4.** Label prediction accuracy (%) on USPS+ MNIST datasets. Domain adaptation via Bregman divergence minimization (DAB) has superior performance compared with other Domain Adaptation (DA) methods (i.e., Visual Domain Adaptation (VDA), Joint Adaptive Classifier and Representation Learning (JACRL), discriminative and Domain Invariant Subspace Alignment for visual tasks (DISA)) in 1 out of 2 tasks using nearest neighbor (NN) classifier.

performance improvements of DAB against FLDA is (18.23%) on USPS+MNIST dataset.

**Experiments on CMU PIE datasets:** To indicate the proficiency of our DAB in face recognition task, we compare DAB with nine state-of-the-art DA methods. Table 4 displays the accuracy of DAB and other methods on PIE datasets. In order to interpret better, the results are visualized in Figure 5. DAB has (43.22%) improvement over NN classifier. Also, the performance improvement of DAB in comparison to the best baseline method DISA is (1.08%) where DAB outperforms the modern DA methods in 11 out of 20 DA tasks. In the following, we compare our DAB with mentioned methods with details. The performance improvement of DAB against FLDA is (43.68%) on PIE dataset.

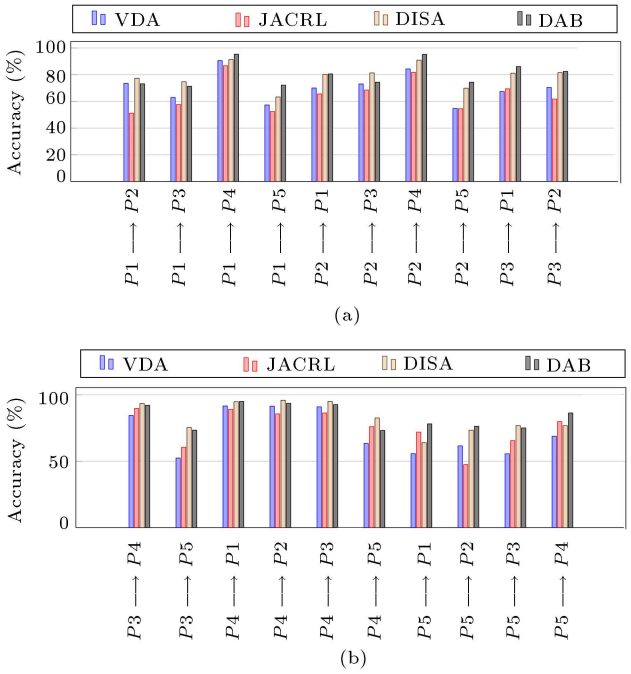



**Figure 5.** Label prediction accuracy (%) on PIE datasets. Domain Adaptation via Bregman divergence minimization (DAB) has superior performance compared with other Domain Adaptation (DA) methods (i.e., Visual Domain Adaptation (VDA), Joint Adaptive Classifier and Representation Learning (JACRL), discriminative and Domain Invariant Subspace Alignment for visual tasks (DISA)) in 11 out of 20 tasks using Nearest Neighbor (NN) classifier: (a) The first ten tasks and (b) the second ten tasks.

NN and FLDA are two conventional metric learning algorithms, which train the classifier on training data to be applied to test data. It is obvious that such algorithms cannot classify accurately because of differences between domains. Although, FLDA shows better efficiency in comparison with NN, it performs poorly versus other DA baseline methods.

To better analyze the contributions of our $DAB$, we evaluate its performance in three different settings: (1) $DAB_M$ which only focuses on the differences between the marginal distributions across domains, (2) $DAB_C$ which only focuses on the differences between the conditional distributions across domains, and (3) $DAB$ which concentrates on the differences between both distributions among domains.

According to Figure 6, $DAB_M$ has low efficiency compared to the other two variants, and $DAB_C$ works better than $DAB_M$, since $DAB_M$ only adapts the marginal distribution of samples and obviously does not reduce the conditional distribution discrepancy between domains. Furthermore, to create a classifier with high prediction accuracy, only the conditional distribution adaptation is not adequate when the domain diversity is significantly large. However, both primary variants cannot attain superior results than DAB. DAB

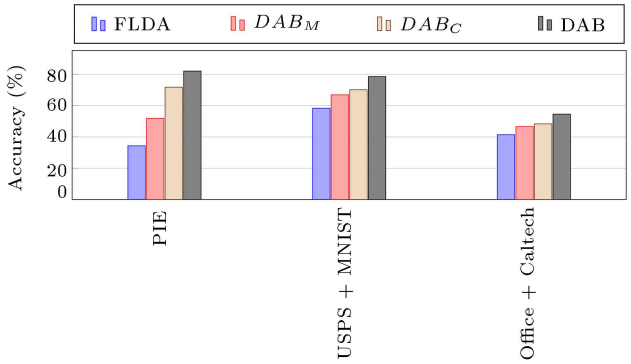



**Figure 6.** Average classification accuracy (%) of Fisher Linear Discriminant Analysis (FLDA), Domain Adaptation via Bregman divergence minimization (DAB), marginal variant of DAB ($DAB_M$) and conditional variant of DAB ($DAB_C$) on Office+Caltech-256, PIE and USPS+MNIST datasets.

minimizes the mismatch between joint marginal and conditional distributions among domains and benefits from the source data labels to create a new subspace. As a result, DAB can transfer knowledge across source and target data more effectively.

TJM is a novel DA technique which seeks to diminish the domain variation by simultaneously adapting the feature spaces and reweighting the samples between domains in a dimensionality reduction way, and constructs an optimal embedded representation, which is robust to both distribution diversity and unrelated samples. TJM only decreases the marginal distribution differences across domains. However, DAB focuses to jointly match both marginal and conditional distribution mismatches in a dimensionality reduction way, and provides an optimal representation which is efficient for considerable distribution diversity. The improvement accuracy of DAB is (18.78%) on digit datasets, (41.45%) on PIE dataset, and (8.14%) on Office+Caltech datasets in term of classification accuracy in comparison with TJM.

VDA is the next new framework that constructs a shared feature representation besides minimizing joint marginal and conditional distributions among domains. In fact, VDA preserves the statistical and geometrical pattern of the input data using manifold assumptions. In addition, VDA exploits domain invariant clustering in an embedded subspace to distinguish the variant classes of the target data. JDA is a novel TL framework that seeks to learn a shared feature subspace which simultaneously alleviates the marginal and conditional distribution discrepancies among the training and test data. JDA utilizes MMD to calculate the distance among the source and target distributions. VDA and JDA only aim to align the distributions of the source and target samples, while they ignore the discriminative properties between distinct classes in adapted domains. DAB tries to diminish the distances among the both domain distributions, while, the particular knowledge (e.g., domain manifold structure) is preserved. Moreover, DAB severally achieves (7.71%), (6.98%) and (5.53%) performance enhancement in average classification accuracy compared to VDA on digits, PIE and Office+Caltech datasets. DAB also severally gains (13.08%), (17.77%), and (9.23%) performance improvement compared to JDA on digits, PIE and Office+Caltech datasets.

LRSR is a new technique, which embeds instances of the source and the target domains into a shared feature-space, where each target instance can be shown by a composition of source instances such that differing domains instances can be well linked. The benefits of LRSR are noted below: (1) It can enlarge the distances among various classes, and (2) It minimizes the differences among the source and target distributions. In LRSR, the test samples are represented using the training samples. Therefore, LRSR could have poor performance on small-scale datasets. DAB performs well on both small and large datasets, and has considerable improvement against LRSR. DAB obtains (12.38%), (14.45%), and (9.16%) improvement compared to LRSR on digits, PIE and Office+Caltech datasets, respectively.

JACRL is another novel TL framework, which obtains an adaptive classifier and an embed representation space simultaneously via reducing the projected distribution difference among domains and the functional structural risk and also increasing the manifold consistency of the learned classifier. However, DAB outperforms JACRL in most cases since DAB considers both the distinctive knowledge included in the labeled instances and the distribution bias across the training and test instances. DAB gains (23.69%), (7.91%), and (4.74%) performance improvement compared to JACRL on digits, PIE and Office+Caltech datasets, respectively.

CLGA as a new unsupervised multi-source DA method has both the local and the global adaptations. CLGA uses multiple domains as the source domains while it mitigates the distribution gap between domains for maximizing the adaptation ability (i.e., global adaptation). Also, it adopts both class and domain manifold structures contained in domain instances for maximizing the discriminative ability (i.e., local adaptation). However, DAB jointly benefits from the representation and classification learning to adapt the source and target domains. Also, DAB uses class structure and distribution alignment to maximize the discriminative and adaptation abilities, respectively. DAB achieves (13.73%), (9.88%), and (6.22%) performance enhancement in average classification accuracy compared to CLGA on digits, PIE and Office+Caltech datasets, respectively.

DISA is a novel unsupervised DA technique,

which matches joint distributions of training and test samples. The focus of DISA is to construct distinct projection matrices to project different domains data to the discrete feature spaces with the following characteristics: 1) the discrepancy of marginal and conditional probability distributions across the source and target samples is calculated by MMD, 2) DISA discriminates various classes in the source domain utilizing the inter-class maximization and intra-class minimization, and 3) DISA preserves the local information of domains instances containing geometrical patterns of samples using sample labels. DISA learns a coupled subspaces to decrease the joint marginal and conditional distribution diversities among domains using MMD while DAB minimizes distribution distance across the marginal and conditional distribution of domains using BD. However, DAB outperforms DISA in most cases, because BD partially preserves the discriminative information. DAB obtains performance amelioration in terms of the average classification accuracy (1.65%), (1.08%), and (2.17%) compared to DISA on digits, PIE and Office+Caltech datasets, respectively.

JPDA is a noticeable method that introduces a new MMD for computing the distribution discrepancy, which is simpler and more accurate. Also, JPDA maximizes the discriminability and transferability, simultaneously. However, DAB exploits BD for computing the distribution discrepancy, which performs well in DA. In addition, DAB uses the source domain labels to construct a shared low dimensional subspace and also discriminate across various classes. Thus, DAB achieves (14.28%), (17.41%), and (6.26%) performance amelioration in average classification accuracy compared to JPDA on digits, PIE and Office+Caltech datasets, respectively.

### *5.2. Parameter analysis*

In this section, we discuss the sensitivity analysis on parameters of DAB. In general, the objective function of DAB, in Eq. (10), has two parameters $\lambda$ and $\eta(k)$. We empirically show the convergence property and parameter sensitivity of our DAB on three benchmark datasets.

1. Parameter $\eta(k)$ is the learning rate at the iteration $k$, where it supervises the gradient step size for $k$th iteration. The learning rate is reduced with the increase of the iterations' number. The large steps for several initial iterations indicates that $W$ is far away from its optimal solution and large size steps are efficient for a quick convergence. Moreover, small size steps for later iterations are beneficial to update $W$ in an acceptable scale such that the optimal solution is reached step by step.

   For a better understanding, Figures 10, 11, and 12 are included and related to the convergence evaluation of DAB on Office-Caltech, digits and PIE datasets. As can be seen from the figures, DAB has almost an identical process in all tasks regardless of the dataset being tested. In this way, with increasing the iteration number, DAB has an upward trend and reaches stability in the final steps. This shows that the $\eta(k)$ parameter has a significant effect on the accuracy of DA. Moreover, Figures 10 to 12 show that the learning rate in the early iterations has an upward trend with a steep slope, while with increasing the iterations number in the final steps, the learning rate decreases and experiences a steady trend. Therefore, the transfer matrix obtained in the final iterations is more robust and more efficient;
2. Parameter $\lambda \in [0, 1]$ is the tradeoff weight that is allocated to the regularization component, which supervises the trade-off between objective functions FLDA and BD. We consider $\lambda = 0.5$ for all datasets. Figures 7, 8 and 9 illustrate the experimental results for parameter $\lambda$ on Office+caltech, digits and PIE datasets, respectively.

Figure 7 shows the parameter sensitivity of $\lambda$ on Office-Caltech dataset. As can be seen, DAB has a significant performance improvement against other two methods. Also, in tasks such as C-W, A-C, W-C, W-A, D-C, and D-A, the prediction accuracy of DAB, VDA, and JDA methods increase by increasing the $\lambda$ value. This shows that the joint distribution discrepancies in mentioned tasks is low and with increasing $\lambda$ impact factor, the effect of this component on DA increases, which gives good results. In addition, in C-A, C-D, A-W and A-D tasks, the prediction accuracy has unstable but almost the uniform trend, while in W-D and D-W tasks trend of instability has a large interlude and downtrend is observed in the final steps. This behavior indicates that the conditional and marginal distribution divergences in the above tasks are large and with increasing the $\lambda$ value, the effect of this component on DA increases, which leads to performance degradation. As is clear from the plots, DAB achieves remarkable results with small values of $\lambda$. Therefore, we consider $\lambda = 0.5$ for Office+Caltech dataset.

Figure 8 shows the sensitivity of $\lambda$ parameter on digits dataset. As can be seen from the figure, the prediction accuracy is almost constant for low $\lambda$ values, and for high $\lambda$ values, the trend is unstable. This indicates that both USPS and MNIST domains have a large distribution discrepancy that increases the impact factor (i.e., parameter $\lambda$) of differences between both domains which causes a performance reduction. The presented results demonstrate that DAB achieves good accuracies with small values of $\lambda$ on digits dataset.

Figure 9 demonstrates the parameter assessment related to the classification accuracy and parameter $\lambda \in$ [0.00001 10] for PIE dataset. As the figure shows, DAB

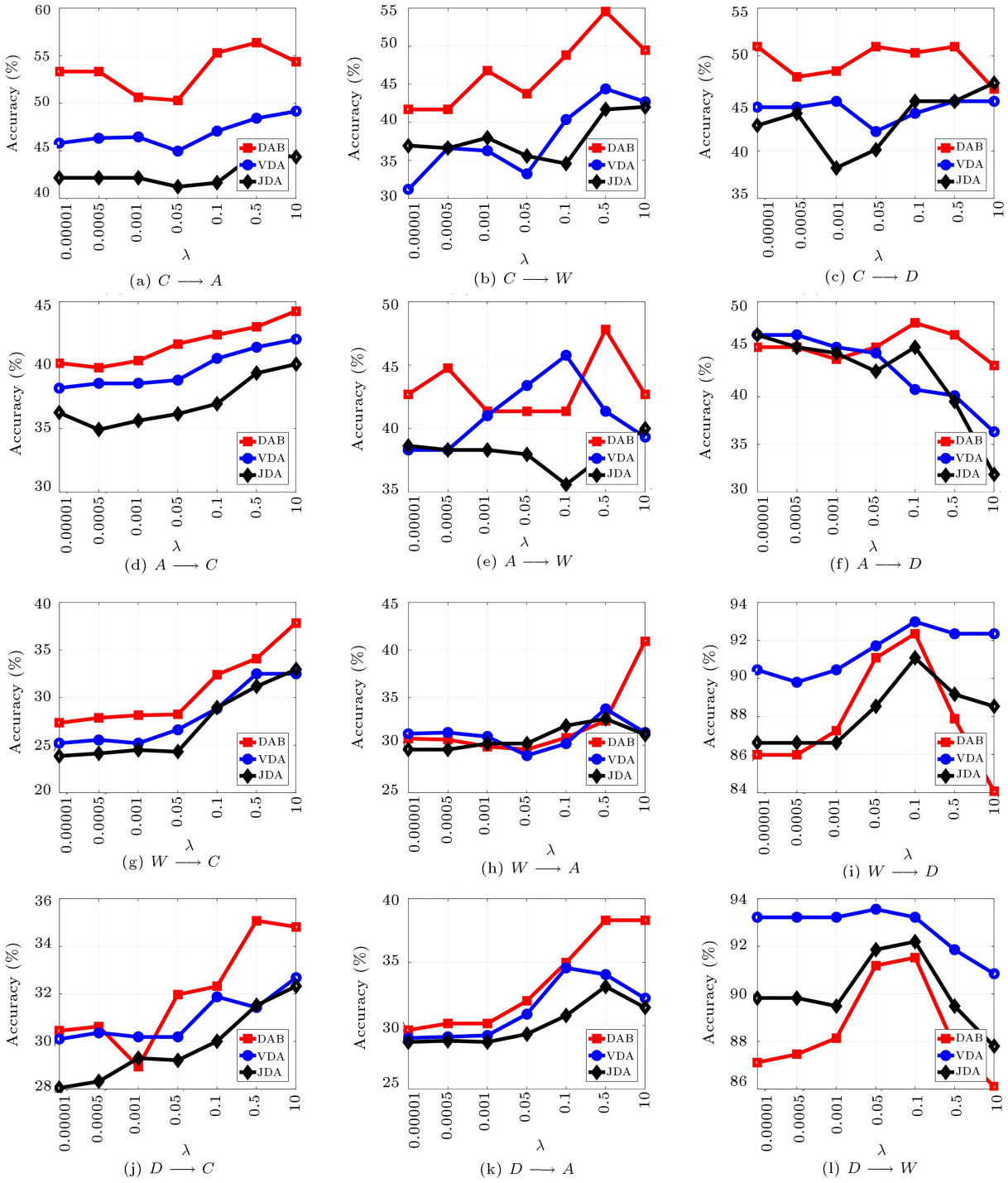
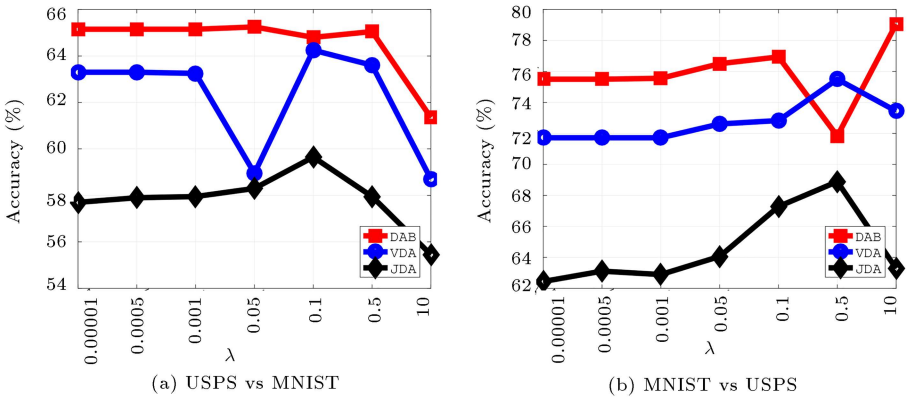



**Figure 7.** Parameter assessment according to label prediction accuracy (%) and parameter $\lambda$, for Visual Domain Adaptation (VDA), Joint Distribution Adaptation (JDA) and Domain Adaptation via Bregman divergence minimization (DAB) on Office+Caltech dataset. DAB achieves remarkable accuracies with large values of $\lambda$. We set $\lambda = 0.5$ for Office+Caltech dataset.

has a higher prediction accuracy than VDA and JDA methods. Also, DAB behavior is almost the same in all tasks. As is clear, for the $\lambda$ values that are less than 0.5, DAB prediction accuracy has an increasing and almost uniform trend, while for values more than 0.5, it has a decreasing trend. These behaviors indicate that the distribution difference in all tasks of PIE dataset is large and with increasing the effect of the distribution difference (i.e., the increase of $\lambda$) on the DA, the prediction accuracy decreases. As Figure 9 makes clear, in most cases, DAB achieves high performance with $\lambda \in [0.01\ 1]$. The best value of $\lambda$ is 0.5 for PIE dataset.




(a) USPS vs MNIST (b) MNIST vs USPS

**Figure 8.** Parameter assessment according to the label prediction accuracy (%) and the regularization parameter $\lambda$, for Visual Domain Adaptation (VDA), Joint Distribution Adaptation (JDA) and Domain Adaptation via Bregman divergence minimization (DAB) on digits dataset. DAB achieves remarkable accuracies with small values of $\lambda$ on digits dataset.

### *5.3. Effectiveness evaluation*

We conduct some experiments on all benchmark datasets to evaluate the efficiency of DAB and three baseline DA methods by considering their accuracies in 20 iterations. TJM, JDA, VDA, and DAB are iterated 20 times with their optimal parameters on Office+Caltech, digits and PIE datasets, and the results are indicated in Figures 10, 11, and 12, respectively.

Figure 10 demonstrates the average prediction accuracy of DAB and three baseline methods on Office+Caltech dataset. As evidenced from the figures, TJM alleviates the marginal distribution mismatch across domains via integrating feature matching and instance reweighting, but it has low performance compared with other baseline methods. JDA obtains desirable performance and outperforms TJM in 7 out of 12 experiments. VDA reduces the mismatch between joint distributions across the domains data and employs domain invariant clustering in an embedded subspace. VDA outperforms TJM and JDA in most cases. However, DAB incorporates TL and DA concurrently and reduces the distribution mismatch across domains. DAB outperforms VDA in 10 out of 12 experiments on Office+Caltech dataset. With the increase of iteration number, the prediction accuracy of DAB increases, which indicates that the learning rate increases with iteration number. In the first iterations, there is an increase with steep slope, while in the last steps, the slope of the prediction accuracy becomes mild, which indicates that the optimal transfer matrix is obtained with much iterations.

Figure 11 displays the efficiency of DAB and three baseline methods on digits dataset. As is apparent from the sub-figures, DAB possesses remarkable improvement on digits dataset in comparison with other DA methods. In the first iterations, DAB has less prediction accuracy against other compared methods, while in the last iterations, DAB prediction accuracy increases significantly. Also, DAB has an upward trend with a steep slope during the iteration increase.

Figure 12 shows the average prediction accuracy of DAB and three benchmark methods on PIE dataset. In general, DAB converges in 20 iterations in most cases. Although DAB is fluctuated in some cases, it has a limited oscillation range after 20 iterations and increasing the iteration number does not have significant impact on the efficiency enhancement of the proposed method. Although DAB method in the first iterations may have a lower prediction accuracy than the compared methods, in the last iterations, DAB prediction accuracy is significantly superior against other methods.

The ineffectiveness of increasing the iteration numbers in the performance of DAB indicates the convergence of DAB in the last iteration stages.

## 6. Conclusion and future work

In this paper, we introduced DAB framework for unsupervised DA tasks. DAB focuses to attain an optimal model with a best possible data representation, at the same time, and minimizes the joint marginal and conditional distribution mismatches. According to the proposed framework, the discriminative features are extracted to introduce a common predictive model. DAB maps the training and test samples onto a common low dimensional feature-space according to the FLDA criterions in an unsupervised manner. DAB efficiently preserves and utilizes the specific information among the samples from different domains. The obtained results indicate that DAB outperforms several state-of-the-art adaptation methods even if the distribution diversity is significantly large. Our upcoming work will focus to improve the efficiency of our proposed approach using other dimensionality reduction methods such as locality preserving projection.

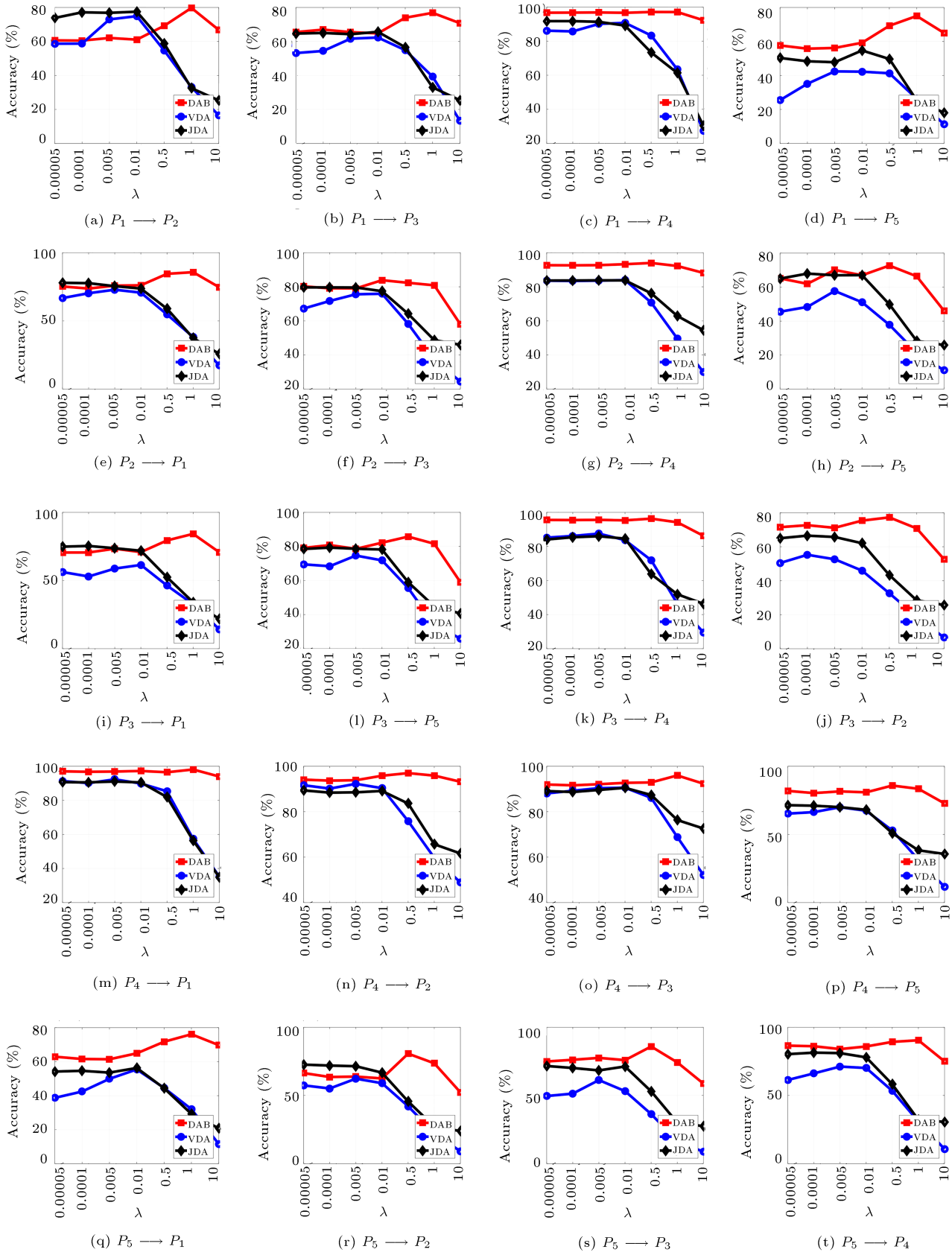



**Figure 9.** Parameter assessment according to the label prediction accuracy (%) and parameter $\lambda$ for Visual Domain Adaptation (VDA), Joint Distribution Adaptation (JDA) and Domain Adaptation via Bregman divergence minimization (DAB) on PIE dataset. In most cases, DAB has high efficiency with $\lambda \in [0\ 1]$. The best value of $\lambda$ is 0.5 for PIE dataset.

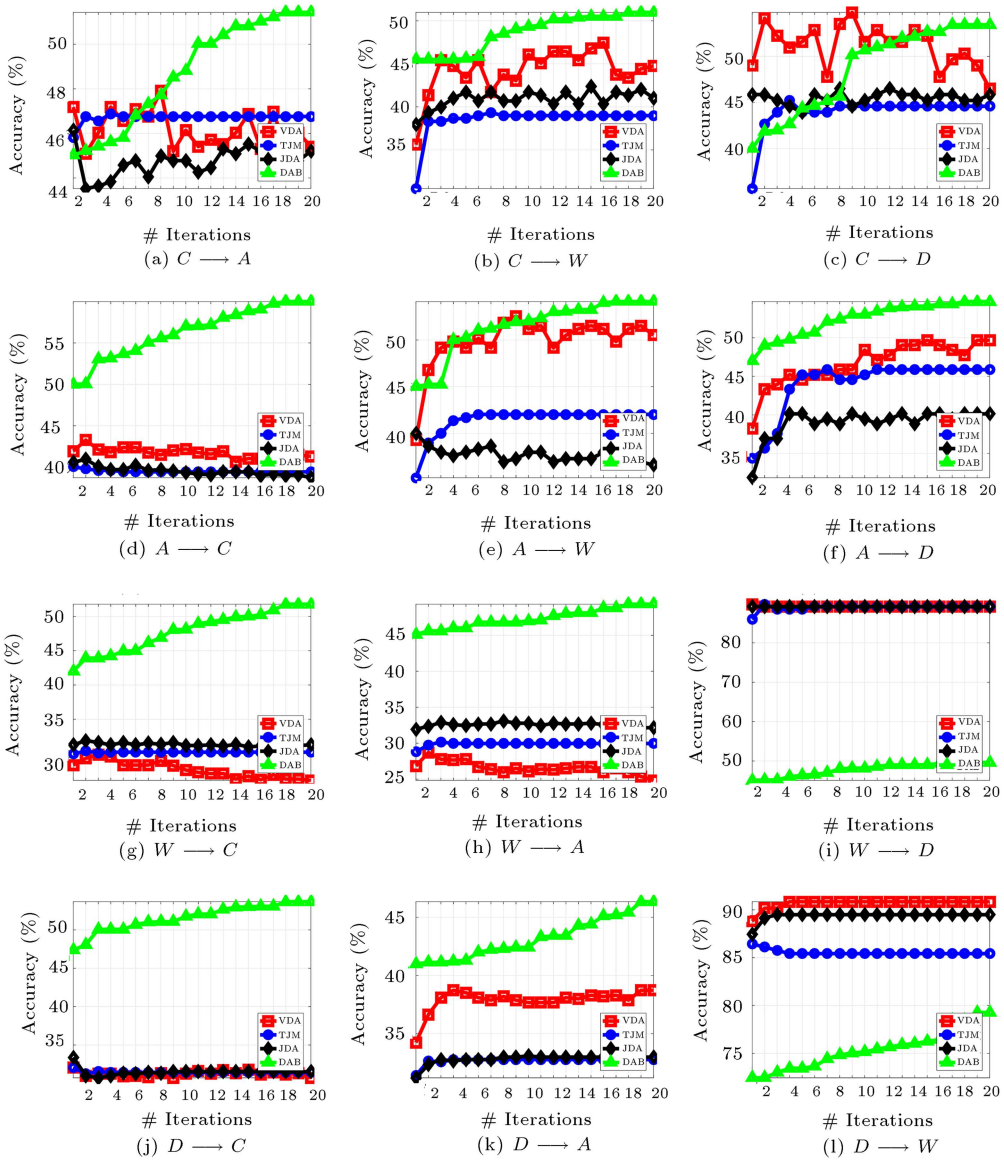



**Figure 10.** Average label prediction accuracy (%) according to the iterations number for Office+Caltech datasets against Domain Adaptation via Bregman divergence minimization (DAB), Visual Domain Adaptation (VDA), Transfer Joint Matching (TJM) and Joint Distribution Adaptation (JDA). DAB predicts the accurate labels for target samples in an iterative manner. Almost, the predicted labels of each step are better than the previous one.

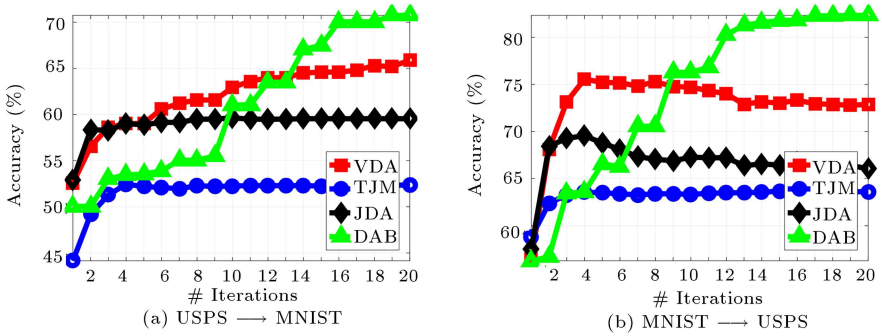



**Figure 11.** Average label prediction accuracy (%) according to the iterations number for digit datasets against Domain Adaptation via Bregman divergence minimization (DAB), Visual Domain Adaptation (VDA), Transfer Joint Matching (TJM) and Joint Distribution Adaptation (JDA). DAB predicts the accurate labels for target samples in an iterative manner. Almost, the predicted labels of each step are better than the previous one.

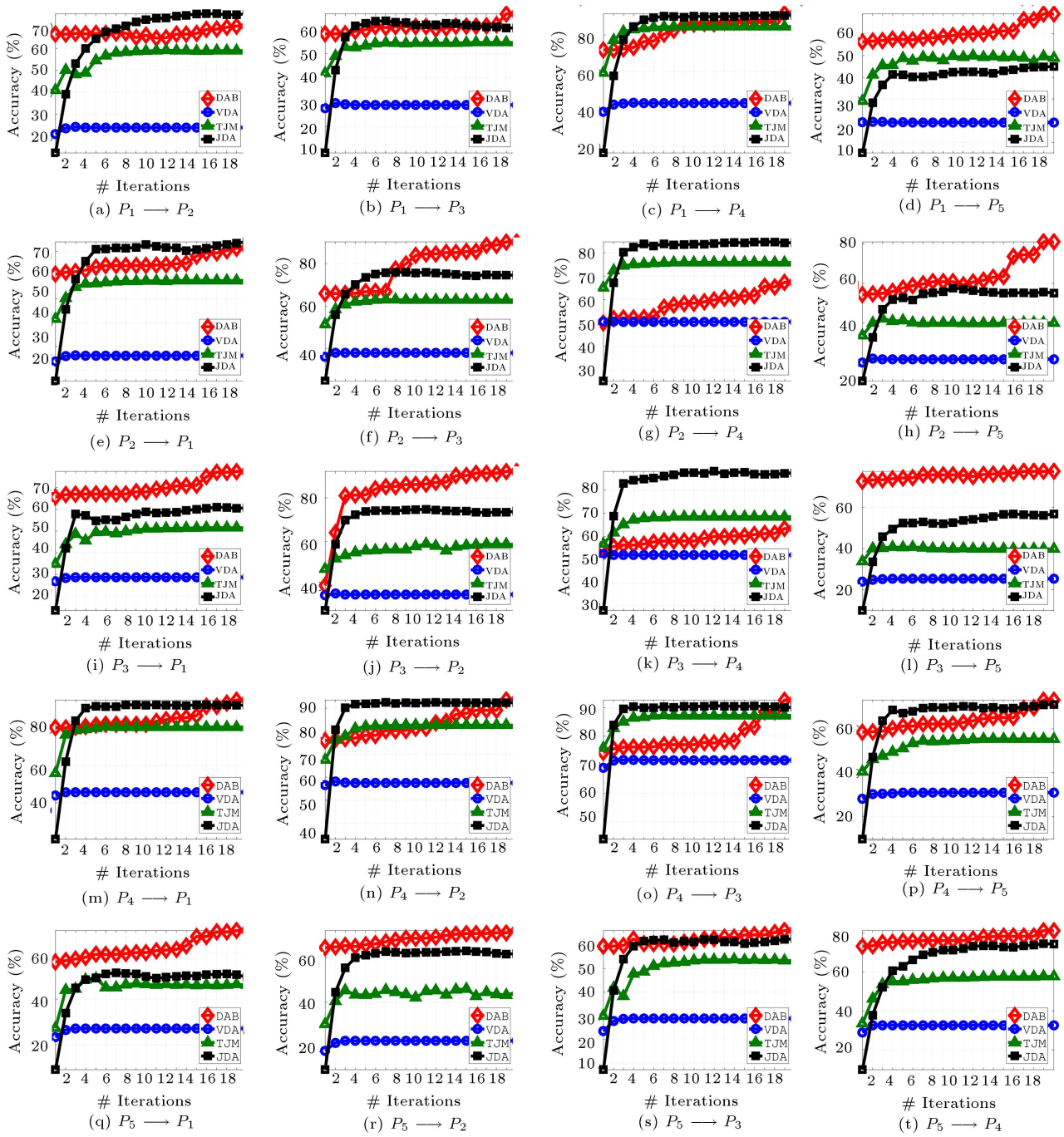



**Figure 12.** Average label prediction accuracy (%) according to the iterations number for PIE datasets against Domain Adaptation via Bregman divergence minimization (DAB), Visual Domain Adaptation (VDA), Transfer Joint Matching (TJM) and Joint Distribution Adaptation (JDA). DAB predicts the accurate labels for target samples in an iterative manner. Almost, the predicted labels of each step are better than the previous one.

## References

1. Zandifar, M. and Tahmoresnezhad, J. "Locality Fisher discriminant analysis for conditional domain adaptation", *Iran Journal of Computer Science*, **4**(1), pp. 17–34 (2020)

2. Si, S., Tao, D., and Geng, B. "Bregman divergence-based regularization for transfer subspace learning", *IEEE T KNOWL DATA EN*, **22**(7), pp. 929–942 (2010).

3. Karimpour, M., Saray, S.N., Tahmoresnezhad, J., et al. "Multi-source domain adaptation for image classification", *Machine Vision and Applications*, **31**(6), pp. 1–19 (2020).

4. Tahmoresnezhad, J. and Hashemi, S. "An efficient yet

effective random partitioning and feature weighting approach for transfer learning", *International Journal of Pattern Recognition and Artificial Intelligence*, **30**(02), p. 1651003 (2016).

5. Vapnik, V.N., *Lect Notes Math.*, Wiley (1998).
6. Cayton, L. "Fast nearest neighbor retrieval for bregman divergences", In *Proceedings of the 25th International Conference on Machine Learning*, pp. 112–119 (2008).
7. Noble, W.S. "What is a support vector machine", *Nature Biotechnology*, **24**(12), pp. 1565–1567, (2006).
8. Denisov, P. and Vu, N.T. "End-to-end multi-speaker speech recognition using speaker embeddings and transfer learning", *arXiv Preprint arXiv:1908.04737* (2019).
9. Shivakumar, P.G. and Georgiou, P. "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations", *Computer Speech and Language*, **63**, p. 101077 (2020).
10. Liu, R., Shi, Y., Ji, C. et al. "A survey of sentiment analysis based on transfer learning", *IEEE Access*, **7**, pp. 85401–85412 (2019).
11. Wei, W., Meng, D., Zhao, Q., Xu, Z., et al. "Semi-supervised transfer learning for image rain removal", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3877–3886 (2019).
12. Saray, S.N. and Tahmoresnezhad, J. "Joint distinct subspace learning and unsupervised transfer classification for visual domain adaptation", *Signal, Image and Video Processing*, pp. 1–9 (2020).
13. Wang, J., Zheng, V. W., Chen, Y., et al. "Deep transfer learning for cross-domain activity recognition", In *Proceedings of the 3rd International Conference on Crowd Science and Engineering*, pp. 1–8 (2018).
14. Zhu, H., Samtani, S., Chen, H., et al. "Human identification for activities of daily living: A deep transfer learning approach", *Journal of Management Information Systems*, **37**(2), pp. 457–483 (2020).
15. Cai, L., Gu, J., Ma, J., et al. "Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees", *Energies*, **12**(1), p. 159, (2019).
16. Sun, G. , Liang, L., Chen, T., et al. "Network traffic classification based on transfer learning", *Computers and Electrical Engineering*, **69**, pp. 920–927 (2018).
17. Hooshmand, A. and Sharma, R. "Energy predictive models with limited data using transfer learning", In *Proceedings of the Tenth ACM International Conference on Future Energy Systems*, pp. 12–16 (2019).
18. Zhong, X., Guo, S., Shan, H., et al. "Feature-based transfer learning based on distribution similarity", *IEEE Access*, **6**, pp. 35551–35557 (2018).
19. Gholenji, E. and Tahmoresnezhad, J. "Joint local and statistical discriminant learning via feature alignment", *Signal, Image and Video Processing*, **14**(3), pp. 1–8 (2019).
20. Gong, B., Grauman, K., and Sha, F. "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation", *ASTR SOC P*, pp. 222–230 (2013).
21. Xiao, T., Liu, P., Zhao W., et al. "Iterative landmark selection and subspace alignment for unsupervised domain adaptation", *Journal of Electronic Imaging*, **27**(3), p. 033037 (2018).
22. Aljundi, R., Emonet, R., Muselet, D., et al. "Landmarks-based kernelized subspace alignment for unsupervised domain adaptation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 56–63 (2015).
23. Wang, B., Qiu, M., Wang, X., et al. "A minimax game for instance based selective transfer learning", in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining ACM*, pp. 34–43 (2019).
24. Long, M., Wang, J., Ding, G., et al. "Adaptation regularization: A general framework for transfer learning", *IEEE Transactions on Knowledge and Data Engineering*, **26**(5), p. 1076–1089 (2014).
25. Gheisari, M. and Baghshah, M.S. "Unsupervised domain adaptation via representation learning and adaptive classifier learning", *Neurocomputing*, **165**, pp. 300–311 (2015).
26. Gheisari, M. and Baghshah, M.S. "Joint predictive model and representation learning for visual domain adaptation", *Eng Appl Artif Intel*, **58**, pp. 157–170 (2017).
27. Fodor, I.K. "A survey of dimension reduction techniques", *Cmr Worksh*, **9**, pp. 1–18 (2002).
28. Tao, D., Li, X., Wu, X., et al. "Geometric mean for subspace selection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(2), pp. 260–274 (2008).
29. Song, P., Zheng, W., Liu, J., et al. "A novel speech emotion recognition method via transfer PCA and sparse coding", In *Chinese Conference on Biometric Recognition*, pp. 393–400 (2015).
30. Song, S., Yu, H., Miao, Z., et al. "Domain adaptation for convolutional neural networks-based remote sensing scene classification", *IEEE Geoscience and Remote Sensing Letters*, **16**(8), pp. 1324–1328 (2019).
31. Tahmoresnezhad, J. and Hashemi, S. "Visual domain adaptation via transfer feature learning", *Knowl Inf Syst*, **50**(2), pp. 585–605 (2017).

32. Liu, J., Li, J., and Lu, K. "Coupled local-global adaptation for multi-source transfer learning", *Neurocomputing*, **275**, pp. 247–254 (2018).

33. Rezaei, S. and Tahmoresnezhad, J. "Discriminative and domain invariant subspace alignment for visual tasks", *Iran Journal of Computer Science*, **2**(4) pp. 219–230 (2019).

34. Ghifary, M., Balduzzi, D., Kleijn, W.B., et al. "Scatter component analysis: A unified framework for domain adaptation and domain generalization", *IEEE T Pattern Anal*, **39**(7), pp. 1414–1430, (2017).

35. Gneiting, T., Balabdaoui, F. and Raftery, A.E. "Probabilistic forecasts, calibration and sharpness", *J R Stat Soc*, **69**(2), pp. 243–268 (2007).

36. Wand M.P. and Jones, M.C., *Kernel Smoothing*, Crc Press (1994).

37. Pan, S.J., Kwok, J.T., and Yang, Q. "Transfer learning via dimensionality reduction", *AAAS R&D B*, **8**, pp. 677–682 (2008).

38. Wang, J., Chen, Y., Hao, S., et al. "Balanced distribution adaptation for transfer learning", In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 1129–1134 (2017).

39. Torkkola, K. "Feature extraction by non-parametric mutual information maximization", *J Mach Learn Res*, **3**(Mar), pp. 1415–1438 (2003).

40. Gong, B., Shi, Y., Sha, F., et al. "Geodesic flow kernel for unsupervised domain adaptation", *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073 (2012).

41. Long, M., Wang, J., Ding, G., et al. "Transfer feature learning with joint distribution adaptation", *Lect Notes Comput Sc*, pp. 2200–2207 (2013).

42. Long, M., Wang, J., Ding, G., et al. "Transfer joint matching for unsupervised domain adaptation", *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1410–1417 (2014).

43. Xu, Y., Fang, X., Wu, J., et al. "Discriminative transfer subspace learning via low-rank and sparse representation", *IEEE T Image Process*, **25**(2), pp. 850–863, (2016).

44. Zhang, W. and Wu, D. "Discriminative joint probability maximum mean discrepancy (DJP-MMD) for domain adaptation", In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE (2020).

## Biographies

**Mozhdeh Zandifar** was born in Hamedan, Iran, in 1992. She received her MSc degree from Department of Information Technology and Computer Engineering, Urmia University of Technology, Iran, in 2018. Her research interests include transfer learning, machine learning and image processing.

**Shiva Noori Saray** received her MSc degree in Information Technology Engineering from faculty of Information Technology and Computer Engineering, Urmia University of Technology, Urmia, Iran, in 2020. Her research interests include transfer learning, machine learning, sentiment analysis, data mining, deep learning.

**Jafar Tahmoresnezhad** received his PhD degree in Computer Science from Shiraz University, Shiraz, Iran, in 2015. Following academic appointments at Urmia University of Technology, he is currently an Associate Professor at Faculty of IT and Computer Engineering, Urmia University of Technology, Urmia, Iran. His research interests include pattern recognition, transfer learning, deep learning, data mining and computer security.