Unsupervised Domain Adaptation with Label and Structural Consistency

Technical Report

Abstract

Unsupervised domain adaptation deals with scenarios in which labeled data are available in the source domain, but only unlabeled data can be observed in the target domain of interest. Since using the classifiers trained by source-domain data would not be expected to generalize well in the target domain, how to transfer the label information from source to target-domain data becomes a challenging task. In this paper, we aim at adapting features of cross-domain data so that the differences between the associated marginal and conditional distributions can be suppressed. In particular, for unlabeled target-domain data, we propose to utilize the label information inferred from the source domain, while the observed structural information in the target domain will be exploited for adaptation purposes. With such cross-domain knowledge, our proposed model not only reduces the mismatch between domains for adaptation purposes, improved recognition of target-domain data can be achieved simultaneously. Experiments on benchmark datasets will verify the effectiveness of our method, which is shown to outperform several state-of-the-art domain adaptation approaches.

1. Introduction

In many real-world classification tasks, one cannot expect that the data to be recognized always exhibit the same or similar distribution as the training data does. The distribution mismatch between training and test data typically comes from the fact that such data are collected from different domains (e.g., videos captured by cameras at different views, images taken by cameras with different resolutions, etc.) [23, 20]. For the above scenarios, classifiers learned from training data cannot be expected to generalize well when recognizing test data.

To address the aforementioned problems, researchers advance the idea of *domain adaptation* and aim at associating cross-domain data for recognition purposes. If the difference between source and target domains can be eliminated, test data observed in the target domain can be recognized by source-domain training data accordingly. Thus, domain adaptation and its applications has been widely exploited in computer vision [20, 16, 9] and machine learning [22, 7, 18] communities.

Depending on the availability of labeled data in the target domain, domain adaptation approaches can be generally divided into two different categories. For *semi-supervised domain adaptation* [20, 4], one can collect source-domain labeled data in advance, but only a small amount of labeled data can be observed in the target domain. Given such cross-domain data and label information, the task is to recognize the remaining target-domain data. On the other hand, *unsupervised domain adaptation* [9, 16] deals with totally unlabeled target-domain data, with only labeled data available in the source domain. In this paper, we focus on unsupervised domain adaptation.

Among existing domain adaptation methods, the most common strategy is to derive feature representations for reducing the domain differences [16, 22, 18, 9], so that recognition can be performed in the resulting feature spaces. While some advocated the adaptation of marginal distributions across data domains [18, 22], several works have been proposed to further adapt both marginal and conditional distributions for improved performance [26, 16]. It is worth noting that, however, adaptation of conditional distributions is not trivial for unsupervised domain adaptation problems. This is because that, only unlabeled data can be observed in the target domain. Therefore, how to properly transfer the source-domain label information to the target domain for associating cross-domain data becomes a challenging task.

As noted above, we particularly address the unsupervised domain adaptation problem in this paper. The overview of our proposed method is shown in Figure 1. Motivated by existing *Maximum Mean Discrepancy* (MMD) [11] based feature adaption approaches, we propose to exploit the *structural* information of target-domain data, together with the *label* information transferred from the source domain. By utilizing such extensive cross-domain knowledge, we approach domain adaptation by solving a *label-propagation* based optimization task, which improves the matching of cross-domain marginal and conditional feature distributions. As verified in our experiments, the proposed method not only exhibits improved domain adaptation ability, it also outperforms several state-of-the-art unsupervised domain adaptation approaches in terms of cross-domain visual classification performance.

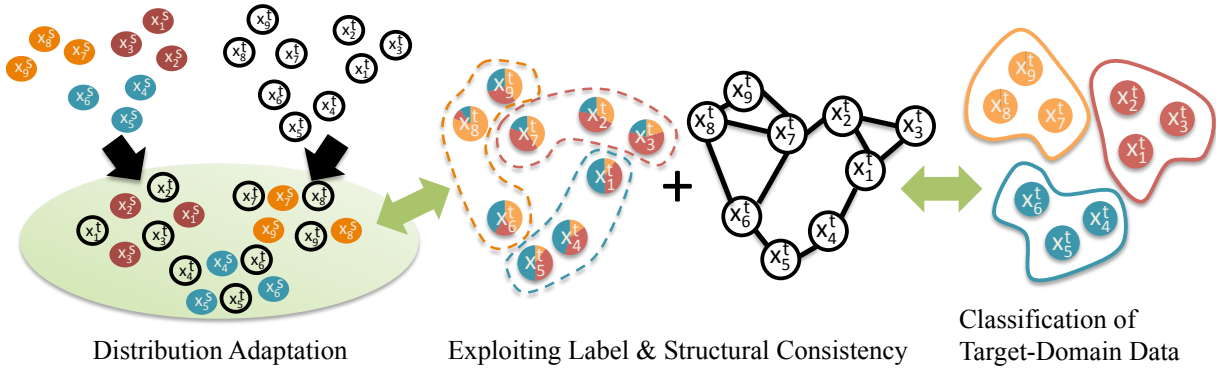

Figure 1. Overview of our proposed method for unsupervised domain adaptation. Note that different colors indicate the class labels, while $\mathbf{x}^s$ and $\mathbf{x}^t$ denote data in source and target domains, respectively.

We now summarize the contributions of this paper:

- When adapting of cross-domain feature distributions for domain adaptation, we exploit the local structural information of the target-domain data. With the label information inferred from the source domain, additional data discriminating capabilities can be introduced. (Section 2)
- We verify the capability of our method in adapting cross-domain distributions. Compared to [16, 17, 18], improved performance can be observed by our proposed method. Our approach also performs favorably against the scenarios in which a substantial amount of ground truth labels from the target domain are available. (Section 4)
- In our experiments, we conduct experiments on several cross-domain image classification datasets, and verify the effectiveness of our method. We show that our method is able to achieve promising recognition performance, and it outperforms several state-of-the-art unsupervised domain adaptation methods. (Section 4)

2. Our Proposed Method

2.1. Motivation

We first start from the problem definition, and introduce the notations which will be used in the following of this paper. Let $\mathcal{D}_S = \{(\mathbf{x}_1^s, y_1^s), \ldots, (\mathbf{x}_M^s, y_M^s)\} = \{\mathbf{X}_S, \mathbf{y}_S\}$, where $\mathbf{X}_S \in \mathbb{R}^{d \times M}$ represents M d-dimensional data in the source domain, and each entry in $\mathbf{y}_S \in \mathbb{R}^{M \times 1}$ indicates the corresponding label (from 1 to C). On the other hand, we have N unlabeled instances observed in the target domain (with the same feature dimension), i.e., $\mathcal{D}_T = \{\mathbf{x}_n^t\}_{n=1}^N = \mathbf{X}_T \in \mathbb{R}^{d \times N}$. Thus, we determine the cross-domain data matrix as $\mathbf{X} = [\mathbf{X}_S, \mathbf{X}_T] \in \mathbb{R}^{d \times (M+N)}$. With the assumption that both source and target domains contain data of the same C classes of interest, the goal of our work is to predict the label vector $\mathbf{y}_T \in \mathbb{R}^{N \times 1}$ for classification purposes, while each element in $\mathbf{y}_T$ is the assigned class label for the corresponding instance in the target domain.

In this paper, we perform transfer feature learning for unsupervised domain adaptation (i.e., only labeled and unlabeled data are available in source and target domains, respectively). We not only eliminate domain differences for associating cross-domain data, we also need to leverage label information from source to target domains for recognition purposes. To address the above issues, we propose and integrate two components highlighted below, which will be detailed in Sections 2.2 and 2.3, respectively:

i) Adaptation of joint feature distributions. Let $P_S(\mathbf{X}_S)$ and $P_T(\mathbf{X}_T)$ as the marginal distributions of data in source and target domains, respectively, and we have $P_S(\mathbf{y}_S|\mathbf{X}_S)$ and $P_T(\mathbf{y}_T|\mathbf{X}_T)$ as the corresponding conditional distributions. As noted in [18, 16], we typically observe $P_S(\mathbf{X}_S) \neq P_T(\mathbf{X}_T)$ and $P_S(\mathbf{y}_S|\mathbf{X}_S) \neq P_T(\mathbf{y}_T|\mathbf{X}_T)$ for cross-domain data. Thus, the goal of domain adaptation is to eliminate the domain bias, so that both marginal and conditional distributions can be matched, and recognition of target-domain data can be performed accordingly. A major contribution of our work is the ability to match cross-domain conditional distributions, given only unlabeled data $\mathbf{X}_T$ in the target domain.

ii) Exploitation of cross-domain data with label and structural consistency. We advance the technique of label propagation [27] for domain adaptation. More specifically, we utilize label information inferred from the source domain and observe the target-domain data structure for performing adaptation. This allows us to tackle the unsupervised domain adaptation problem with improved recognition of target-domain data.

2.2. Distribution Adaptation

As highlighted in Section 2.1, the primary goal of this work is to match both marginal and conditional feature dis-

tributions of cross-domain data, so that data in the target domain can be classified accordingly. However, as noted in [16], since the modeling of conditional distributions $P(\mathbf{y}_S|\mathbf{X}_S)$ and $P(\mathbf{y}_T|\mathbf{X}_T)$ is not explicitly applicable, an alternative way is to observe and adapt class-conditional distributions $P(\mathbf{X}_S|\mathbf{y}_S)$ and $P(\mathbf{X}_T|\mathbf{y}_T)$ based on their sufficient statistics.

In our work, we aim at determining a feature transformation Φ for cross-domain data, so that both $P_S(\Phi(\mathbf{X}_S)) \approx P_T(\Phi(\mathbf{X}_T))$ and $P_S(\Phi(\mathbf{X}_S)|\mathbf{y}_S) \approx P_T(\Phi(\mathbf{X}_T)|\mathbf{y}_T)$ can be satisfied. For simplicity, we apply empirical criteria of MMD [11] for adapting the above distribution. To be more precise, we need to minimize the difference between feature distributions is calculated by the distance between data means in a reproducing kernel Hilbert space (RKHS):

$$Dist(P_S(\mathbf{X}_S), P_T(\mathbf{X}_T)) + Dist(P_S(\mathbf{X}_S|\mathbf{y}_S), P_T(\mathbf{X}_T|\mathbf{y}_T)) = \left\| \frac{1}{M}\sum_{i=1}^{M}\phi(\mathbf{x}_i^s) - \frac{1}{N}\sum_{j=1}^{N}\phi(x_j^t) \right\|_{\mathcal{H}}^2 + \sum_{c=1}^{C}\left\| \frac{1}{|\mathcal{D}_S^{(c)}|}\sum_{\mathbf{x}_i^s\in\mathcal{D}_S^{(c)}}\phi(\mathbf{x}_i^s) - \frac{1}{|\hat{\mathcal{D}}_T^{(c)}|}\sum_{\mathbf{x}_i^t\in\hat{\mathcal{D}}_T^{(c)}}\phi(\mathbf{x}_i^t) \right\|_{\mathcal{H}}^2, \tag{1}$$

where $Dist$ measures the distance between feature distributions, and ϕ is the feature transformation induced by universal kernels. In (1), we have $\mathcal{D}_S^{(c)} = \{\mathbf{x}_i^s : y_i^s = c\}$ indicate source-domain data of class c, and $\hat{\mathcal{D}}_T^{(c)} = \{\mathbf{x}_i^t : \hat{y}_i^t = c\}$ as those in the target domain with the same predicted label. It is worth repeating that, for unsupervised domain adaptation, a major challenge is to transfer the label information from source to target domains when reducing domain biases. Later in Section 2.3, we will explain how we determine the label information for target-domain data, so that the adaptation of the above conditional distributions can be achieved.

By kernel tricks, we rewrite (1) as $tr(\mathbf{KL}) + \sum_{c=1}^{C} tr(\mathbf{KL}_c)$, where $\mathbf{K} \in \mathbb{R}^{(M+N)\times(M+N)}$ indicates the kernel matrix of data matrix $\mathbf{X}$, and

$$L_{ij} = \begin{cases} \frac{1}{M^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_S \\ \frac{1}{N^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_T \\ \frac{-1}{MN}, & \text{otherwise.} \end{cases}$$

$$(L_c)_{ij} = \begin{cases} \frac{1}{|D_S^{(c)}|^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_S^{(c)} \\ \frac{1}{|\hat{D}_T^{(c)}|^2}, & \mathbf{x}_i, \mathbf{x}_j \in \hat{\mathcal{D}}_T^{(c)} \\ \frac{-1}{|D_S^{(c)}||\hat{D}_T^{(c)}|}, & \begin{cases} \mathbf{x}_i \in \mathcal{D}_S^{(c)}, \mathbf{x}_j \in \hat{\mathcal{D}}_T^{(c)} \\ \mathbf{x}_i \in \hat{\mathcal{D}}_T^{(c)}, \mathbf{x}_j \in \mathcal{D}_S^{(c)} \end{cases} \\ 0, & \text{otherwise.} \end{cases}$$

As noted in [18], the above optimization problem with respect to $\mathbf{K}$ would require high computational costs. In our work, we utilize the empirical kernel map [21] as suggested in [18, 16], and derive a lower-dimensional space of fixed $\mathbf{K}$ by determining the projection matrix $\mathbf{W}\in\mathbb{R}^{(M+N)\times k}$ instead ($k << d$). As a result, we have the objective function as:

$$\min_{\mathbf{W}} tr(\mathbf{W}^\top\mathbf{KLK}^\top\mathbf{W}) + \sum_{c=1}^{C} tr(\mathbf{W}^\top\mathbf{KL_cK}^\top\mathbf{W}) + \lambda\|\mathbf{W}\|_F^2 \quad \text{s.t. } \mathbf{W}^\top\mathbf{KHK}^\top\mathbf{W} = \mathbf{I}. \tag{2}$$

It can be seen that, the first two terms in (2) are associated with the adaptation of marginal and conditional distributions, respectively. The sum of the these two terms corresponds to the MMD distance between the cross-domain data. The third term in (2) regularizes the projection $\mathbf{W}$, weighted by parameter λ. The centering matrix $\mathbf{H}$ in the constraint of (2) is defined as $\mathbf{H} = \mathbf{I} - \frac{1}{M+N}\mathbf{1}$, in which $\mathbf{1}$ is the matrix of ones. As noted in [18, 16], adding this constraint would preserve the data variance after adaptation, which implies and introduces additional data discriminating ability into the learned model $\mathbf{W}$. By applying Lagrange techniques, we can rewrite the objective function of (2) into the following Lagrangian function:

$$\mathcal{L}(\mathbf{W}, \mathbf{\Psi}) \equiv tr(\mathbf{W}^\top(\mathbf{KLK}^\top + \mathbf{K}\sum_{c=1}^{C}\mathbf{L}_c\mathbf{K}^\top + \lambda\mathbf{I})\mathbf{W}) + tr((\mathbf{I} - \mathbf{W}^\top\mathbf{KHK}^\top\mathbf{W})\mathbf{\Psi}), \tag{3}$$

where $\mathbf{\Psi}$ is a diagonal matrix with Lagrange Multipliers (i.e., $\mathbf{\Psi} = diag(\psi_1, ..., \psi_k)\in\mathbb{R}^{k\times k}$). By setting the derivative of (3) with respective to $\mathbf{W}$ equal to zero, we approach the original optimization problem by solving the following generalized eigen-decomposition problem:

$$(\mathbf{KLK}^\top + \mathbf{K}\sum_{c=1}^{C}\mathbf{L}_c\mathbf{K}^\top + \lambda\mathbf{I})\mathbf{W} = \mathbf{KHK}^\top\mathbf{W}\mathbf{\Psi}. \tag{4}$$

Taking the $k-$smallest eigenvectors from (4) would satisfy (2), which determines the optimal solution of $\mathbf{W}$ (recall that $k << d$). Once $\mathbf{W}$ is obtained, we project cross-domain data into the resulting k-dimensional latent space i.e., $\mathbf{Z} = \mathbf{W}^\top\mathbf{K} = [\mathbf{Z}_S, \mathbf{Z}_T]\in\mathbb{R}^{k\times(M+N)}$, where $\mathbf{Z}_S$ and $\mathbf{Z}_T$ represent the transformed data projected from source and target domains, respectively. In other words, the data matrix $\mathbf{Z}$ can be viewed as adapted cross-domain data with matched marginal and conditional distributions.

2.3. Exploiting Label and Structural Consistency for Unsupervised Domain Adaptation

Due to the lack of label information in the target domain, matching of cross-domain marginal and conditional distributions would be a challenging task. In particular, without proper prediction of class labels for the target-domain data $\hat{\mathcal{D}}_T$, adapting the conditional distributions for cross-domain data cannot be achieved.

To solve the above problem, we propose to take the knowledge which is exploited across domains into the adaptation process, with the goal of suppressing domain biases with cross-domain recognition guarantees. To begin with, we apply SVM-based classifiers [2, 24] for estimating the class posterior probability of the transformed target-domain data $\mathbf{Z}_T$. For the setting of unsupervised domain adaptation, these SVM classifiers are trained by labeled source-domain data in the transformed feature space (i.e., $\mathbf{Z}_S$). Based on the estimated posterior probabilities, we construct an *uncertain label matrix* $\mathbf{Y}\in\mathbb{R}^{N\times C}$ for target-domain instances, in which each entry is defined as:

$$Y_{ij} = \begin{cases} 2p(y_i^t = j|z_i^t) - 1, & \begin{cases} p(y_i^t = j|z_i^t) > \delta \\ p(y_i^t = j|z_i^t) = \max\limits_c p(y_i^t = c|z_i^t) \end{cases} \\ -1, & \text{otherwise.} \end{cases} \tag{5}$$

It is worth noting that, the posterior probability $p(y_i^t = j|z_i^t)$ indicates how likely the projected target-domain instance z_i^t belongs to class j. Obviously, we have $-1 \leqslant Y_{ij} \leqslant 1$, and a larger Y_{ij} value implies that the instance of interest is of the corresponding class. We note that, the parameter δ controls the number of uncertain labels to be transferred from the source domain (i.e., the aggressiveness of label propagation). For simplicity, we simply set δ equal to the lower quantile (i.e., 25%) of the maximum posterior probabilities observed from each target-domain instance. In other words, 75% of target-domain data will be assigned the predicted uncertain labels for adaptation purposes. Later in the experiments, we will provide additional remarks on our choice of δ.

Once the above uncertain label matrix is constructed, we effectively set a semi-supervised setting for the target-domain data. However, unlike standard semi-supervised learning problems in which a portion of the data are given specific class labels, we do not directly take the labels predicted by source-domain data due to possible domain mismatch. In other words, we cannot directly apply existing semi-supervised techniques, since they cannot deal with data collected from different domains. This is the reason why the use of our uncertain label matrix together with feature adaptation is preferable, which offers additional robustness in adapting and assigning class labels.

In addition to the use of uncertain labels predicted from the source domain, we further take the data structure observed in the target domain into our adaptation process. This allows us to better determine the target-domain labels for improved recognition. To observe target-domain structural information, we advance graph-based semi-supervised learning by constructing a k-nearest neighbors (k-NN) graph over target-domain data [27, 19, 6]. We note that, we choose to construct this k-NN based graph in the transformed space (i.e., $\mathbf{Z}_T$). The use of the transformed space not only allows us to better observe data structural information due to reduced feature dimensions, the learned transformation model $\mathbf{W}$ also exhibits capabilities in eliminating biases across source and target domains. This is why improved recognition of target-domain data can be expected.

Algorithm 1 Our Proposed Model

Input: Kernel matrix $\mathbf{K}$ of cross-domain data, labels $\mathbf{y}_S$ of source-domain data

1. Initialize $\hat{D}_T^{(c)}$ as $\emptyset$

while not converged **do**

2. $\mathbf{W} \leftarrow$ *Distribution adaptation* $(\hat{D}_T^{(c)}, \hat{\mathbf{y}}_T)$ in (4) and let $[\mathbf{Z}_S, \mathbf{Z}_T] = \mathbf{W}^\top \mathbf{K}$
3. Assign $\mathbf{Y}^{(0)}$ by classifiers trained by $\mathbf{Z}_S$ and (5)
4. Construct the k-NN graph matrix $\mathbf{E}$ and $\mathbf{S}$ within target domain $\mathbf{Z}_T$
5. $(\hat{D}_T^{(c)}, \hat{\mathbf{y}}_\mathbf{T}) \leftarrow$ *label propagation* $(\mathbf{Y}^{(0)}, \mathbf{S})$ in (7)

end while

6. $\mathbf{y}_T \leftarrow \hat{\mathbf{y}}_T$

Output: $\mathbf{y}_T$ as labels of target-domain data

Based on the above observations, we calculate the distance between target-domain data pairs as $d(\mathbf{z}_i^t, \mathbf{z}_j^t) = \|\mathbf{z}_i^t - \mathbf{z}_j^t\|$, and apply Gaussian kernels for converting such distances into similarity scores: $s(\mathbf{z}_i^t, \mathbf{z}_j^t) = exp(-d(\mathbf{z}_i^t, \mathbf{z}_j^t)/2\sigma^2)$. With this structural information determined, the k-NN based similarity matrix $\mathbf{E}\in\mathbb{R}^{N\times N}$ can be formulated, in which each entry is:

$$E_{ij} = \begin{cases} s(\mathbf{z}_i^t, \mathbf{z}_j^t), & \text{if } \mathbf{z}_j^t \text{ is one of k-NN of } \mathbf{z}_i^t \text{ and } i \neq j \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

With the uncertain label matrix $\mathbf{Y}$ and the structural similarity matrix $\mathbf{E}$ of target-domain data obtained, we advance the technique of label propagation [27] for updating the label information for each instance in the target domain. We note that, label propagation exhibits capabilities in incorporating (partial) label and structural information for determining the final labels for each instance of interest. In our work, we construct the matrix $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{E}\mathbf{D}^{-1/2}$, in which $\mathbf{D}$ is a diagonal matrix that $d_{ii} = \sum_j E_{ij}$. We perform label propagation which iteratively propagates and updates the labels of each instance via the constructed similarity graph until convergence: $\mathbf{Y}^{(t+1)} = \alpha\mathbf{S}\mathbf{Y}^{(t)} + (1-\alpha)\mathbf{Y}^{(0)}$ with the regularization parameter $\alpha\in(0,1]$ and $\mathbf{Y}^{(0)} = \mathbf{Y}$. According to [27], the optimal label matrix $\mathbf{Y}^*$ has the following closed-form solution:

$$\mathbf{Y}^* = (\mathbf{I} - \alpha\mathbf{S})^{-1}\mathbf{Y}^{(0)}. \tag{7}$$

Once $\mathbf{Y}^*$ is derived, the final label of each target-domain instance is determined by $\hat{y}_i^t = arg\max_{j\leqslant C} Y_{ij}^*$ based on the winner-take-all strategy.

2.4. Adaptation and Recognition via Iterative Optimization

Finally, we integrate the techniques and learning models presented in Sections 2.2 and 2.3 for performing joint unsupervised domain adaptation and cross-domain recognition. The proposed method is summarized in Algorithm 1. It can be seen that, except for the initialization stage which only adapts marginal distributions of cross-domain data, the optimization process take both marginal and conditional distributions into consideration, while the class labels are updated from the previous adaptation iterations. Later in the experiments, we will show that the proposed method converges to the optimal solution in terms of both MMD and accuracy in few iterations, which verify the effectiveness of the proposed model for domain adaptation and recognition.

3. Related Works

Recently, domain adaptation has attracted the attention from researchers in the fields of machine learning [13, 25, 4], speech and language processing [3, 26, 1], and computer vision [16, 9, 7, 20, 10, 8]. Among different adaptation approaches, MMD-based feature adaptation approaches have been studied in [18, 16, 7, 8, 13], which aim at matching cross-domain information (e.g., data or class-conditional means) for adaptation and recognition purposes.

For unsupervised domain adaptation, since only unlabeled data is available in the target domain, modeling of the associated class-conditional feature distributions is not a trivial task. Previously, researchers chose to model and match cross-domain marginal distributions only. For example, Huang et al. [13] proposed kernel mean matching by weighting source-domain data, so that the mean difference between cross-domain data in a predetermined kernel space can be minimized. To provide additional robustness, Pan et al. [18] proposed Transfer Component Analysis (TCA) to determine low-dimensional embeddings for cross-domain data, and they performed matching of cross-domain marginal distributions in the derived lower-dimensional feature space. Based on TCA, Long et al. [17] further proposed Transfer Feature Matching (TJM) combining instance reweighting and distribution adaptation techniques for further improving the performance.

However, adapting marginal distributions only would not be expected to be sufficient for relating cross-domain data, especially if recognition of target-domain data is desirable. To address this issue, Long et al. [16] proposed Joint Distribution Adaptation (JDA) to match both marginal and conditional feature distributions. To deal with unsupervised domain adaptation settings, they applied the outputs of source-domain classifiers as the pseudo labels of target-domain data, and thus the conditional distributions can be observed accordingly. Since the direct use of such pseudo labels might not be preferable due to possible domain mismatch, we not only transfer label information but further exploit target-domain structural information during the adaptation process. Inspired by recent semi-supervised learning works of [6, 19, 27], we approach the problem of unsupervised domain adaptation by solving a label-propagation based optimization task, which allows us to better associate cross-domain marginal and conditional feature distributions. By jointly solving the adaptation and recognition tasks in a unified framework, improved recognition performance can be expected in the target domain.

4. Experiments

4.1. Datasets and Settings

MNIST and USPS: We first consider cross-domain digit recognition using MNIST [15] and USPS [14] datasets. MNIST contains a training set of 60,000 images and a test set of 10,000 images, while each image is of size 28×28 pixels. On the other hand, each image in USPS is of size 16×16 pixels, and a total of 7291 and 2007 images are available for training and testing, respectively. Figure 2(a) shows example images of these two datasets.

In our experiments, we follow the same setting as [16] did. We randomly sample 2000 and 1800 images from MNIST and USPS (scaled to the same 16×16 pixels), respectively, and take pixel intensities as the features. Two cross-domain pairs are considered: MNIST → USPS and USPS → MNIST. Take MNIST → USPS for example, we have MNIST as the source domain with 2000 labeled training data, and USPS as the target domain with 1800 instances to be recognized. Similar remarks can be applied to USPS → MNIST.

Caltech-256 and Office: For experiments on cross-domain object recognition, we consider the Caltech-256 [12] and Office [20, 9] datasets. The former consists of object images of 256 categories (with at least 80 instances per category), while the latter contains 31 objects categories collected from three different sub-datasets: Amazon, DSLR, and webcam. Following the same settings applied in [16, 7, 8], we select the 10 over-lapping object categories of Caltech-256 and Office for experiments, and produce four different domains of interest: Caltech (C), Amazon (A), DSLR (D), and webcam (W). As a result, a total of 12 different cross-domain pairs will be available (e.g., C → A, C → W, etc.).

To describe each object image in the Caltech-256 and Office datasets, we apply the $DeCAF_6$ features [5]. As shown in [5], these features are able to achieve very promising results for image classification. With the use of $DeCAF_6$ features, each image will be converted into a 4096-dimensional representation for training and testing.

It is worth noting that, since only unlabeled (test) data

Figure 2. Example images of (a) *MNIST + USPS* datasets and (b) *Caltech-256 + Office* datasets.

Table 1. Comparisons of recognition rates (%) for cross-domain digit recognition. Note that S and T denote source and target domains, respectively.

S → T	MNIST → USPS	USPS → MNIST
Direct	50.1	33.2
TCA [18]	52.7	45.7
JDA [16]	68.5	56.0
TJM [17]	63.5	52.7
Ours*	70.6	62.7
Ours	**72.3**	**65.5**

are available in the target domain, one cannot apply cross-validation to select the parameters for the learning models. For fair comparisons, we follow the same parameter settings as [16] did, and set $\lambda = 0.1$ and 1 for digit and object datasets, respectively. When performing data embedding, we choose $k = 100$ as the reduced feature dimension. In addition, we follow the recent works of [16, 18] and apply the linear kernels for constructing the kernel matrix $\mathbf{K}$. For simplicity, we fix the parameter $\alpha = 0.5$ for label propagation, and set the number of neighbors (for the graph-based similarity matrix $\mathbf{E}$) as 15 for all our experiments.

4.2. Evaluation

For cross-domain digit recognition, we consider the approaches of TCA [18], JDA [16] and TJM [17]. It is worth repeating that, JDA also adapts both marginal and conditional distributions for unsupervised domain adaptation as we do. For baseline approaches, we consider the direct use of SVMs trained by source-domain data in the original feature space (i.e., no domain adaptation). Table 1 lists the recognition results of cross-domain digit recognition.

Recall that, when using our proposed method, recognition is achieved when the domain adaptation process is complete (i.e., via label propagation). To show that we can also train the SVM classifiers in the derived transformed feature space using projected labeled source-domain data, and apply such classifiers to recognize the projected target-domain data as other recent methods do (e.g., TCA and JDA), we provide additional results of ours in Table 1 (denoted as Ours*). Nevertheless, as shown in this table, our methods clearly outperformed baseline and state-of-the-art methods for the task of cross-domain digit recognition.

As for cross-domain object recognition, we further consider another state-of-the-art method of Landmarks (LM) [7, 8]. We present and compare the recognition performance of different methods in Table 2. It is worth noting that LM can only be applied to 9 out of 12 cross-domain pairs. This is because that, as noted in [8], LM requires a sufficient amount of source-domain data for adaptation, and it cannot be applied to the cases when DSLR is applied as the source domain. Our proposed method does not have this limitation. More importantly, from the results presented in this table, we see that our method significantly outperformed all others in all cases. We also visualize the table in Figure 3 for more clear interpretation. This supports the use of our proposed model for unsupervised domain adaptation.

4.3. Remarks on Adapting Cross-Domain Feature Distributions

As discussed in Section 2.3, a major contribution of our approach is its ability in exploiting label and structure consistency, which allows us to better match cross-domain conditional distributions for improved domain adaptation. For verification purposes, we consider different amounts of ground-truth labeled data available in the target domain for solving (2). To be more precise, we utilize different amounts of target-domain instances and their labels to construct $\mathbf{L_c}$ in (2). With more labeled target-domain data observed for adaptation, the improvements of cross-domain recognition can be expected. Figure 4 compares our results with those with varying amounts of labeled target-domain instances on selected cross-domain datasets. It can be seen that our method with the challenging unsupervised setting actually achieved comparable results with those utilizing a large amount of ground truth target domain labels for adaptation. As a result, the capability of our method in associating cross-domain data distributions for unsupervised domain adaptation can be successfully verified.

Table 2. Comparisons of recognition rates (%) on *Caltech256 + Office* datasets.

S → T	C → A	D → A	W → A	A → C	D → C	W → C	A → D	C → D	W → D	A → W	C → W	D → W	*Average*
Direct	91.86	72.13	74.63	82.64	60.20	64.56	81.53	86.62	99.36	74.58	79.66	96.61	80.36
TCA [18]	90.21	87.68	82.67	85.04	79.70	77.38	82.16	87.26	98.22	76.94	81.02	97.02	85.44
JDA [16]	92.02	90.28	87.02	86.33	83.88	83.64	88.54	90.36	100	83.78	85.08	97.98	88.91
LM [8]	92.28	-	86.01	84.42	-	70.53	84.71	89.17	99.36	84.07	85.42	-	-
TJM [17]	92.17	88.73	83.51	85.04	81.03	80.14	83.44	87.26	99.36	80.34	81.36	97.02	86.62
Ours*	92.80	92.17	92.07	87.44	86.02	86.02	91.08	93.63	100	87.12	89.67	98.63	91.16
Ours	**94.26**	**92.37**	**93.31**	**87.88**	**86.19**	**87.97**	**94.9**	**95.26**	**100**	**88.81**	**91.18**	**99.32**	**92.62**

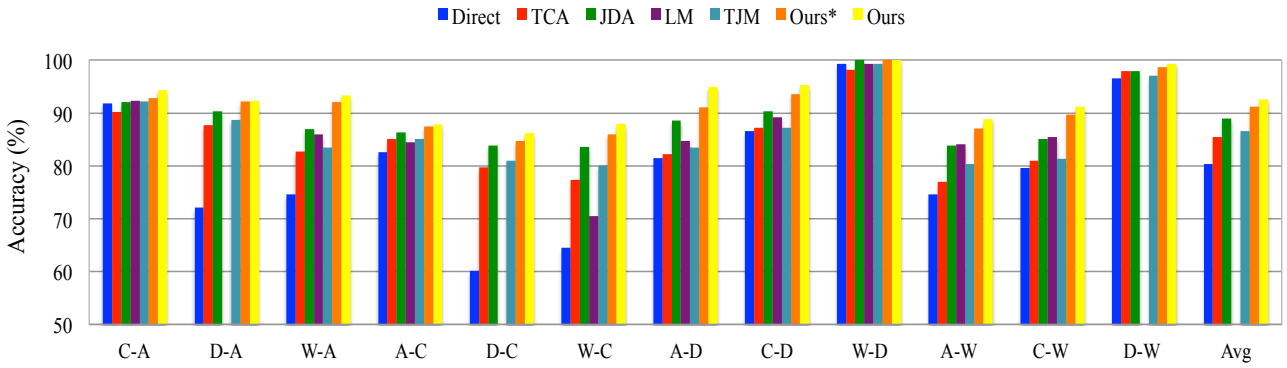

Figure 3. Classification Accuracy (%) on Office and Caltech datasets.

4.4. Remarks on Convergence and δ

As noted in Section 2.4, we iteratively solve the proposed model for adapting cross-domain data. To assess its convergence property, Figures 5(a) and (b) show the recognition accuracy and MMD distance with increasing iteration numbers, respectively. Recall that, the MMD distance is calculated by summing up the first two terms in (2) using target-domain data with ground truth labels. Due to space limit, only selected cross-domain data pairs are presented. From the above figures, we see that both accuracy and distance converged within 5-10 iterations during optimization.

We comment on the parameter δ in (5). Recall that, δ determines the aggressiveness of label propagation by controlling the number of uncertain labels to be transferred from the source domain. In our experiments, we simply set δ as the lower quantile (i.e., 25%) of the maximum posterior probabilities observed from each target-domain instance, which allows 75% of target-domain data to be assigned uncertain labels during the adaptation process. Although a larger δ value would imply fewer (and thus less noisy) target-domain instances with uncertain labels for propagation, the adaptation capability would be limited due to less information adapted from the source domain. In other words, the choice of δ is a tradeoff between adaptation and propagation.

Table 3. Comparisons of runtime estimates (in seconds) of different methods using A → W domain pair. Note that JDA, TJM, and ours all converged at 10 iterations.

TCA [18]	JDA [16]	LM [8]	TJM [17]	Ours
4.15 (s)	34.12 (s)	1204 (s)	36.17 (s)	45.41 (s)

Figures 5(c) and (d) compare the recognition performance of two example domain pairs over different δ values. It can be seen that, while extreme δ (i.e., close to 0 or 1) cannot achieve satisfactory performance, our δ choice (based on the above guideline) was able to achieve improved results when comparing to state-of-the-art methods. Intuitively, the choice of δ should be domain dependent. For example, if the mismatch between source and target domains is marginal, one would expect that a small δ would be sufficient for performing adaptation. The study of domain biases and its effect on adaptatoin/propagation aggressiveness would be among our future research directions.

Finally, we compare the computation timE of different methods on the domain pair of A → W in Table 3. The runtime estimates were performed on an Intel Core i5 PC with 2.6 GHz CPU and 8G RAM. It can be seen that the

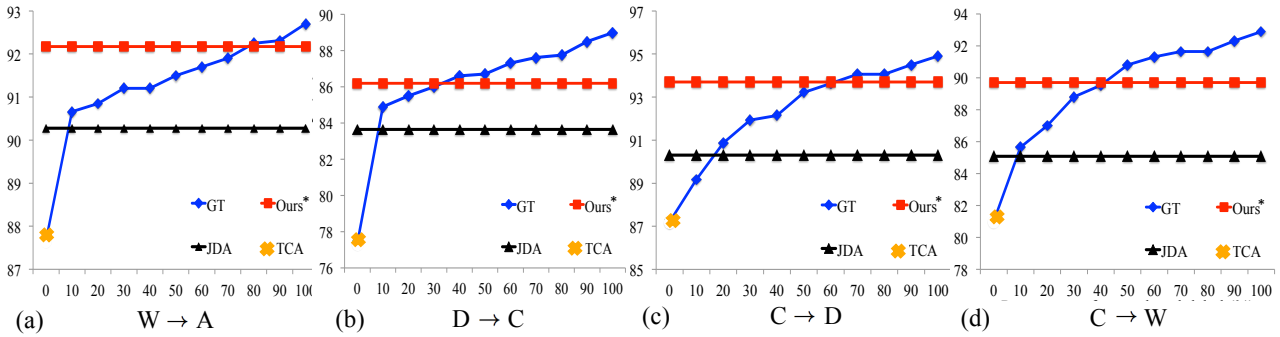

Figure 4. Verification of distribution adaptation on (a) W → A, (b) D → C, (c) C → D and (d) C → W. The vertical axis denotes the accuracy (%), and the horizontal axis indicates the percentage of ground-truth labels used the target domain. Note that only selected cross-domain data pairs are presented due to space limit.

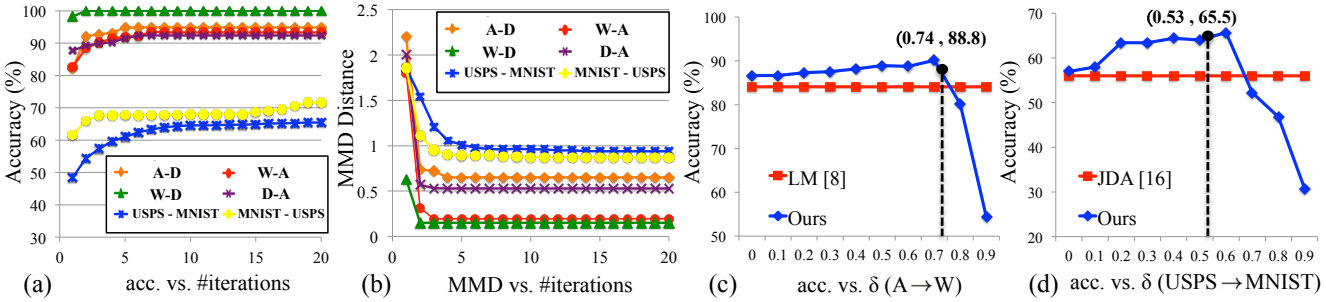

Figure 5. Convergence analysis and parameter sensitivity. The former is verified by reporting (a) recognition accuracy and (b) MMD distance. The latter presents the recognition results over different δ values on (c) A→W and (d) USPS→MNIST. Note that the vertical dash line indicate the δ value which we determine (see Section 2.3).

computation time of our proposed approach (including iterative optimization and label propagation) was comparable to those of state-of-the-art methods, while the recognition performance was greatly improved.

5. Conclusion

We proposed an unsupervised domain adaptation based on transfer feature learning. In addition to matching both marginal and conditional distributions of cross-domain data, our proposed model further leverages rich label and structural information across domains. This allows us to achieve improved adaptation and recognition of cross-domain data. Our experiments on cross-domain digit and object recognition confirmed that our proposed model performed favorably against state-of-the-art domain adaptation methods. Future research directions include landmark (i.e., instance) selection for cross-domain data and domain-adaptive label propagation, which could further improve the domain adaptation and recognition performance.

References

[1] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006. 5

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 4

[3] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007. 5

[4] H. Daumé III, A. Kumar, and A. Saha. Co-regularization based semi-supervised domain adaptation. In *NIPS*, 2010. 1, 5

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *ICML*, 2014. 6

[6] S. Ebert, D. Larlus, and B. Schiele. Extracting structures in image collections for object recognition. In *ECCV*. 2010. 4, 5

[7] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013. 1, 5, 6

[8] B. Gong, K. Grauman, and F. Sha. Learning kernels for unsupervised domain adaptation with applications to visual object recognition. *IJCV*, 2014. 5, 6, 7

[9] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE CVPR*, 2012. 1, 5

[10] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *IEEE ICCV*, 2011. 5

[11] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two sample problem. In *NIPS*, 2007. 1, 3

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 5

[13] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007. 5

[14] J. J. Hull. A database for handwritten text recognition research. *PAMI*, 16(5):550–554, 1994. 5

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[16] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *IEEE ICCV*, 2013. 1, 2, 3, 5, 6, 7

[17] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer joint matching for unsupervised domain adaptation. In *IEEE CVPR*, 2014. 2, 5, 6, 7

[18] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans Neural Networks*, 2011. 1, 2, 3, 5, 6, 7

[19] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013. 4, 5

[20] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*. 2010. 1, 5

[21] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998. 3

[22] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE TKDE*, 2010. 1

[23] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE CVPR*, 2011. 1

[24] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *JMLR*, 5(975-1005):4, 2004. 4

[25] K. Zhang, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *ICML*, 2013. 5

[26] E. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. Turaga, and O. Verscheure. Cross domain distribution adaptation via kernel mapping. In *ACM KDD*, 2009. 1, 5

[27] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2003. 2, 4, 5