Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(11): 6338–6363.
DOI: 10.3934/era.2024295
Received: 02 August 2024
Revised: 04 November 2024
Accepted: 15 November 2024
Published: 22 November 2024

Research article

Structure preserved ordinal unsupervised domain adaptation

Qing Tian[1,2,3,*] **and Canyu Sun**[1]

[1] School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China

[2] Wuxi Institute of Technology, Nanjing University of Information Science and Technology, Wuxi 214000, China

[3] MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

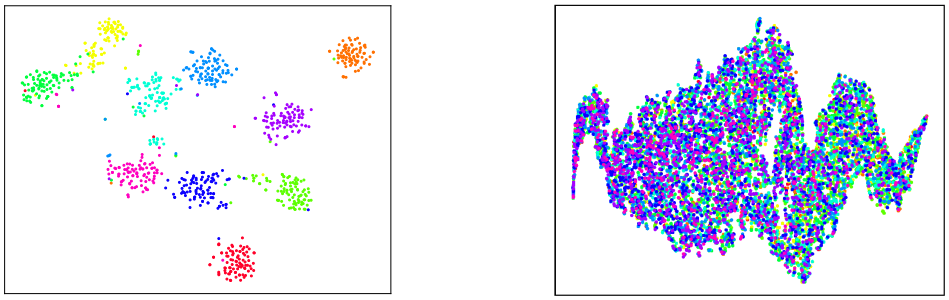
* **Correspondence:** Email: tianqing@nuist.edu.cn.

Abstract: Unsupervised domain adaptation (UDA) aims to transfer the knowledge from labeled source domain to unlabeled target domain. The main challenge of UDA stems from the domain shift between the source and target domains. Currently, in the discrete classification problems, most existing UDA methods usually adopt the distribution alignment strategy while enforcing unstable instances to pass through the low-density areas. However, the scenario of ordinal regression (OR) is rarely researched in UDA, and the traditional UDA methods cannot preferably handle OR since they do not preserve the order relationships in data labels, like in human age estimation. To address this issue, we proposed a structure-oriented adaptation strategy, namely, structure preserved ordinal unsupervised domain adaptation (SPODA). More specifically, on one hand, the global structure information was modeled and embedded into an auto-encoder framework via a low-rank transferred structure matrix. On the other hand, the local structure information was preserved through a weighted pair-wise strategy in the latent space. Guided by both the local and global structure information, a well-performance latent space was generated, whose geometric structure was adopted to further obtain a more discriminant ordinal regressor. To further enhance its generalization, a counterpart of SPODA with deep architecture was developed. Finally, extensive experiments indicated that in addressing the OR problem, SPODA was more effective and advanced than existing related domain adaptation methods.

Keywords: unsupervised domain adaptation; ordinal domain adaptation; structure-oriented adaptation; ordinal regression

1. Introduction

Nowadays, in-depth research on neural networks has promoted the rapid development of machine learning, such as convolutional neural networks (CNN) for computer vision [1–3], recurrent neural networks (RNN) for natural language processing [4–6], and graph neural networks (GNN) for recommendation systems [7–9]. Although the above methods have received success in various tasks, a strong hypothesis should be guaranteed, i.e., the training and test data must comply with the independent and identically distribution. However, the hypothesis is too strict and even impracticable for real-world applications. For instance, a human facial age predictor, trained on training images with ideal lighting, tends to make mistakes when deployed on wild environments [10]. The reason is that facial appearance and recognizability are highly susceptible to environmental factors, such as scene illumination. This phenomenon induces a robust model to be constructed for out-of-distribution data to handle the issue of distribution shift.



(a) Amazon (b) AgeDB

Figure 1. Visualization of the latent representation space of the Amazon dataset and the AgeDB dataset. The representations of the former are extracted from DeCaf architecture with Cross-Entropy loss, while the representations of the latter are extracted from ResNet-50 architecture with ordinal margin loss, namely, ODFL [11]. To visualize more clearly, on the AgeDB dataset, each neighboring five ages are set as one class, and each age contains about 40 facial instances. The circles with different colors represent different classes.

Over the past decades, domain adaptation community has emerged and became more and more active. The aim of unsupervised domain adaptation (UDA) is to transfer the knowledge from the labeled data to the unlabeled data [1, 12, 13]. In this field, the labeled data sampled from a certain distribution is named source domain, and the unlabeled data sampled from another distribution is distinguished as target domain. As a result, the main challenge of UDA stems from the domain shift between the source and target domains.

To quantitatively measure the difference between two domains, the $\mathcal{H}$-divergence [14, 15] is introduced to derive the generalization bounds based on the Vapnik-Charvonenkis (VC) dimension [16]. Along this line, the UDA has theoretical guarantee and a variety of valid methods have been successively derived, which follows two major methodologies, i.e., feature matching and instance reweighing. Specifically, the feature matching aims to seek a latent representation space in which either the marginal distributions [17], the conditional distributions [18], or both of them [19, 20] across the domains are aligned. For the instance reweighing methodology, it mainly assumes that the target domain

sample space can be constructed by the reweighed source samples [21]. Hence, its goal is to estimate the weights of the source domain data similar to the target domain [22, 23]. To further enhance the discrimination performance of UDA, in recent years, entropy regularization [24] has been introduced into UDA [25, 26] and other DA settings, such as source-free UDA [27,28], open-set UDA [29,30], universal UDA [31, 32] and so on. The reason is that the entropy regularization can enforce those unstable instances across the domains to pass the low-density area, enlarging the interclass margins and consequently improving the discrimination.

The aforementioned methods focus on handling the discrete classification problems, and unfortunately cannot better handle the continuous regression problems since they do not preserve the "orderness" characteristics of the distributed ordinal data in labels. Figure 1 demonstrates this phenomenon visually on the Amazon dataset [33] and Morph Album II dataset [34]. Herein, the former denotes the discrete case, while the latter denotes the continuous case. Beyond the expectation, surprisingly, the performance of the regression case is poor despite more training data with deeper architecture being adopted. We can observe that compared with the clearly separated data clusters in the classification task on Amazon, there is no clear data class boundaries but continuous manifold in the ordinal regression task on AgeDB. Therefore, intuitively, such strategies that enforce unstable instances across the low-density area or enlarge the margin in the target domain, like large-margin learning or entropy regularization, are no longer applicable for the ordinal UDA scenario. To perform ordinal DA, the class-related latent factors are modeled through recursive conditional Gaussian (RCG) to capture sequential structures. Moreover, the separation of class-related and class-unrelated factors and self-training conducted on a shared ordinal latent space are introduced to realize the alignment of source and target domains [35,36]. Although the above methods have achieved good adaptation performance, the global and local structural information has not been well preserved, which may lead to suboptimal decision boundary.

In the ordinal regression scenarios, the data instances like human facial images are typically distributed within a low-dimensional manifold space [37], as illustrated in Figure 1(b). Given that the domains share the same task in ordinal UDA, their underlying manifold structures should be similar. In addition, target instances from the same class tend to exhibit higher affinity than those from different classes. Motivated by these observations, we investigate a structure-oriented adaptation strategy, coined as structure preserved ordinal unsupervised domain adaptation (SPODA). More specifically, inspired by the self-representation learning in spatial clustering [38], the global structure information is embedded into an auto-encoder framework by a transferred structure matrix in SPODA. In this way, the cross-domain structure knowledge is captured through the structure matrix. Meanwhile, the style transformation can be represented to further guide a more well-performance latent space through the autoencoder framework. In addition, not only intra-domain but also inter-domain local structure information can be explored and exploited simultaneously. Besides, the above local information can also assist in the reconstruction of input space. As a result, a common latent space is consequently generated and guided by both the local and global structures, which will be beneficial in improving the generalization of the ordinal UDA regressor to learn. Further, in the generated latent space, a more discriminative regressor guided by the geometric structure automatically [39] is designed to generate pseudo labels in the target domain. Then, with the confident pseudo label annotation, the previous steps in SPODA are updated in turn until the whole procedure converges, and, consequently, the ordinal UDA can be fulfilled. To sum up, the main contributions of this article are highlighted as follows:

1. A novel kind of SPODA is proposed by exploiting both the local and global structure information from the source and target domains, in which the global structure knowledge is captured by low-rank learning while the intra/inter-domain local structure information is preserved by manifold learning in designed latent space.
2. SPODA adopts an autoencoder reconstruction idea and a novel regularization term to obtain domain-consistent feature representations and protect cross-domain intra-class structural information in the hidden space. Moreover, the MLG-LSC model and cumulative attribute coding are introduced to derive more discriminative regression boundaries.
3. An alternating optimization algorithm is derived to efficiently solve the SPODA model. Besides, SPODA is extended with deep neural network architecture to further boost its generalization performance.
4. Extensive experiments on different benchmark datasets are conducted to evaluate the effectiveness of the SPODA method.

The remainder of this article is organized as follows. In Section 2, the existing UDA methods are reviewed briefly. In Section 3, our proposed method, SPODA, and the optimization algorithm are presented in detail. In Section 4, comprehensive experiments are reported and the corresponding results are analyzed. Finally, in Section 6, conclusions are drawn.

2. Related work

In this section, we review the related domain adaptation methods briefly, which can be classified into two taxonomies: domain adaptation classification and DA regression.

2.1. DA classification

In terms of DA classification, Ben-David et al. [14] proposed a generalization bound based on $\mathcal{H}$-divergence [15] to theoretically guarantee the feasibility of the learning paradigm. Inspired by this theorem, a variety of methods have been developed, which can be classified into two categories, i.e., feature matching and instance reweighing. For feature matching, the maximum mean discrepancy (MMD) [17] and adversarial model [40] are widely adopted strategies. Along this line, TCA [17] alleviates the marginal distribution between the source and target domains via the MMD metric. Moreover, both the JDA [20] and BDA [19] leverage the marginal and conditional distributions using the pseudo label strategy. Subsequently, BDA is further introduced into heterogeneous domain adaptation scenarios, which is named DDA [41]. To reduce the computational complexity of kernel function in MMD, the CMD measures [42] with lower moments alignment has been proposed. Further, instead of central mean alignment, CORAL [43] matches the cross-domain covariance, i.e, the second-order moment. In addition, the MMD is incorporated with multilayer representation learning in DAN [44] to yield unbiased deep features. Moreover, to improve the discriminability in DA procedure, the margin dragging strategy is employed in JDDA [45]. Following the methodology of the generative adversarial network (GAN) [46], the DANN model [40] is built with the gradient reversal layer to obtain better domain invariant representations. The discriminative modeling, untied weight sharing, and adversarial learning are integrated together in ADDA [47] to perform more powerful DA. Furthermore,

in DADA [48], the domains are aligned in terms of their joint distributions through a decoupling category and entropy regularization. Moreover, a novel bi-classifier adversarial paradigm [49, 50] by maximizing classifier discrepancy is introduced to realize cross-domain knowledge transfer. Recently, to enhance the discrimination and confidence of the output predictions, entropy regularization [25,27] is employed to handle domain adaptation [29,32,51]. Along the direction of instance reweighing, the nonparametric feature learning [52] is proposed to perform DA. Moreover, in SPL [53], the progressive sample selection strategy is adopted to weigh the adjacent pair-wise instances between the source and target domains. Besides, the block-wise generative transfer methodologies are also employed to select similar source data for taregt adaptation [23,54]. Further, the manifold regularization is placed on the weighted data samples to preserve their structure in the process of DA [12]. Impressed by the success of MoCo [55] and SimCLR [56], some works [57–59] attempt to introduce the contrastive learning into DA. For example, CDCL [57] generates the target pseudo-labels based on the prototype-initialized K-Means clustering and realizes the cross-domain contrastive learning by attracting samples from the same category in two domains. Furthermore, EIDCo [59] explores the limitations of directly transferring the instance discrimination contrastive loss to DA, and introduces the class relationship embedded features and target-dominant mixup to overcome the above restrictions. In recent works, Wang et al. [2] introduced a new property termed equity to reveal the effectiveness of nuclear norm maximization in UDA. In addition, two novel loss functions that incorporate equity constraints into the squares loss are designed to encourage predictive discriminability and equity. Furthermore, CDSA [3] generates more diverse augmented data through the proposed CrossSmooth technique, and CrossVariance is developed to enable each domain to capture the styles of multiple domains.

For structure preserved UDA methods, Liu et al. [60] implemented multisource UDA tasks by mixing data from the source and target domains for clustering and simultaneously exploring the structure of both domains. On this basis, SP-UDA [61] redefines the domain adaptation problem as a semi-supervised clustering problem, which guides the learning of the target domain structure by preserving the inherent structure of the source domain. Specially, an augmented matrix and a nontrivial solution are developed to transform the UDA problem into a K-means optimization problem. Further, SPTR [62] preserves local semantic structure during the knowledge transfer by enforcing structural consistency in the feature space and label space of the source and target domains. Besides, a novel sample reweighing strategy is introduced to reduce the harm of inaccurate pseudo labels to the target model. Moreover, HCSA [63] applies the concept of spatial structure preservation to heterogeneous domain adaptation to achieve more accurate classification decisions.

However, as shown in Figure 1(a), the aforementioned methods focus on the discrete classification tasks. In this way, there are usually large low-density areas in the representation space or label space. Therefore, strategies such as entropy regularization are taken into consideration to force the marginal instances to be more confident.

2.2. DA regression

Different from the DA classification, the challenge of DA regression stems not only from the domain shift but also the continuous data distributions. Unfortunately, few of the existing DA methods pay attention on such scenarios. Due to the sparsity of regression output space in keypoint detection, the authors of [64] bridge the gap between regression and classification by optimizing an adversarial regressor complying with a spatial probability distribution. In reference [65], a transfer ordinal label

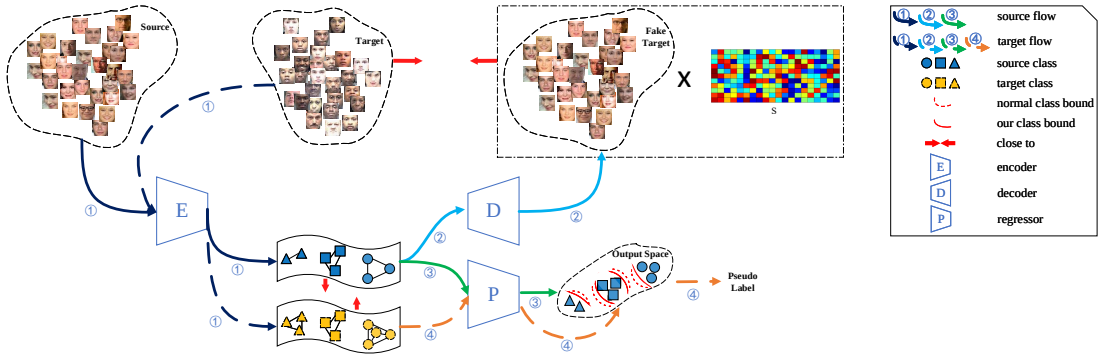


Figure 2. The framework of our proposed method SPODA. It contains four modules: i) the encoder $\boldsymbol{E}$, ii) the decoder $\boldsymbol{D}$, iii) the transferred structure matrix $\boldsymbol{S}$, and iv) the regressor $\boldsymbol{P}$. Flow ② denotes a procedure of the reconstruction, where a fake target domain would be transformed from the source domain through $\boldsymbol{E}$ and $\boldsymbol{D}$ and combined with $\boldsymbol{S}$ to reconstruct the target domain. Flow ① denotes a procedure of the explioting structure of the latent space through $\boldsymbol{E}$. Flow ③ indicates the procedure of developing regressor with the underlying structure under the source domain through $\boldsymbol{E}$ and $\boldsymbol{P}$. Flow ④ denotes a procedure of the prediction through $\boldsymbol{E}$ and $\boldsymbol{P}$.

learning model is proposed by expanding the solution space with an ensemble of ordinal classifiers from multiple relevant source domains. The research [66] reveals that aligning the distribution of deep representations would alter feature scale and hamper DA regression. To alleviate such issues, they attempt to reduce the domain shift via a set of orthogonal bases of the representation space instead. DINO [67] studies the UDA regression problem by integrating the ideas of feature matching, sample weight, and adaptive Gaussian process into the proposed distribution-informed neural networks. Moreover, DARE-GRAM [68] addresses the cross-domain regression challenge by matching the scale and angle within a subspace formed by the pseudo-inverse gram matrices of the two domains. Unlike the above works, we concentrate on a more challenging scenario, that is ordinal UDA, in which all the data patterns have an orderly trend and neighboring classes even overlap, as shown in Figure 1(b). It is worth noting that although the setting of CIDA [69] is quite similar to ours, their sampling domains are continuous and sampling classes discrete, so this method can be regarded as a classification model essentially while ours is definitely ordinal regression.

3. Proposed method

In this section, we first propose some preliminaries. Along this line, we present our proposed method, SPODA, and the corresponding optimization algorithm in details. Subsequently, the counterpart with deep architecture is developed to further boost its generalization performance.

3.1. Preliminaries

In our proposed ordinal UDA scenario, we are given the labeled source domain $\mathcal{D}_s = \left\{\left(x_i^s, y_i^s\right)\right\}_{i=1}^{n_s}$, where $x_i^s \in \mathbb{R}^d$ represents the representation of the i-th instance with d dimensions, $y_i^s \in \mathbb{R}$ represents the ground truth of the i-th instance, and n_s represents the number of instances sampling from the

source domain. C_s is the label set of the source domain, and $y_i \in \{1, 2, \ldots, K\}$, where $K = |C_s|$ is the number of classes. In the target domain, we are also given the unlabeled instances $\mathcal{D}_t = \left\{x_i^t\right\}_{i=1}^{n_t}$, where $x_i^t \in \mathbb{R}^d$ represents the representation of the i-th target domain instance, and n_t represents the number of instances sampling from the target domain. C_t is the label set of the target domain as the same as C_s. The marginal distributions of the source and target domains are denoted as $\mathcal{P}_s$ and $\mathcal{P}_t$, while their conditional distributions are denoted as $\mathcal{Q}_s$ and $\mathcal{Q}_t$, respectively. In our settings, we tackle a close-set UDA task, where $|C_s| = |C_t|$, $\mathcal{P}_s \neq \mathcal{P}_t$, and $\mathcal{Q}_s \neq \mathcal{Q}_t$. It is worth noting that while our regression settings are consistent with the domain adaptation classification task, they, however, are essentially different. On the one hand, at the label level, our classes are mutually dependent while they are independent in domain adaptation classification. On the other hand, except for the domain shift, at the representation level we have a relatively dense representation space and the neighbor classes may overlap and differ from classification task where the representations are clustering. These observations are demonstrated in Figure 1 and bring more challenges.

3.2. SPODA

To tackle the issues above, we propose a structure-oriented adaptation strategy, namely, SPODA, whose framework is exhibited in Figure 2. Specifically, inspired by the self-representation learning [38] in the spatial clustering, we adapt it for the target domain and obtain the objective function as follows:

$$\min_{\boldsymbol{S}} \quad \|\boldsymbol{X}_t - \boldsymbol{X}_t\boldsymbol{S}\|_F^2 + \lambda f(\boldsymbol{S}) \tag{3.1}$$

where $\boldsymbol{S}$ represents the self-representation matrix in clustering, which aims to explore and exploit the relationships of the global structure among instances, and $f(\boldsymbol{S})$ represents a regularization term to further capture more precise relationships. λ represents a hyper-parameter to leverage the balance between the self-representation loss and the regularization term.

To transfer the knowledge maintained in the source domain, we plan to adopt a subspace alignment learning paradigm, assuming that after the spatial transformation of the projection matrix $\boldsymbol{W} \in \mathbb{R}^{d\times p}$, $P_s\left(W^T X_s\right) = P_t\left(W^T X_t\right)$. However, unlike existing works that align directly in the hidden space, in order to obtain stable feature representations between domains, we extend the above self-representation learning to a cross-domain structure-exploiting strategy analogous to PCA or auto-encoder [33]. Therefore, Eq (3.1) can be reformulated:

$$\begin{array}{ll} \min_{\boldsymbol{W},\boldsymbol{S}} & \left\|\boldsymbol{X}_t - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s\boldsymbol{S}\right\|_F^2 + \lambda f(\boldsymbol{S}) \\ s.t. & \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_p \end{array} \tag{3.2}$$

where $\boldsymbol{S} \in \mathbb{R}^{n_s\times n_t}$ is used to capture the global cross-domain structure information and is slightly different from the one in Eq (3.1), and $\boldsymbol{I}_p$ represents a p-order identity matrix. The orthogonal constraint prevents trivial solutions w.r.t. $\boldsymbol{W}$. For such a modeling methodology, the following advantages can be drawn:

1. Aligning inter-domain covariance: Compared to the modeling strategy of subspace alignment inter-domain expectations, using a PCA-like modeling strategy can implicitly align inter-domain covariance.

2. Removing essential differences: The MMD metric treats each dimensional feature representation equally, but not all feature representations are effective. We use auto-encoder to remove inconsistent features between domains and improve the robustness of the hidden space.

3. Mining spatial structure: While aligning covariance, use the representation matrix $\boldsymbol{S}$ to synchronously mine potential spatial structure relationships between domains.

Nevertheless, the structure information at the class level has not been explored and exploited. Moreover, the local neighbor structure may be neglected in the latent space so that the knowledge transfer is quite an effort and the global structure, captured by $\boldsymbol{S}$, may lose its effectiveness. To alleviate this issue, while preserving the manifold structure inherent in data, the linear approximation of the nonlinear Laplacian eigen-map is constructed to explore and exploit not only intra-domain but also inter-domain local structure information and to further assist in the reconstruction of input space. To this end, Eq (3.2) can be derived as follows:

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{S}} \quad & \left\|\boldsymbol{X}_t - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s\boldsymbol{S}\right\|_F^2 + \lambda f(\boldsymbol{S}) \\ & +\alpha \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\right\|_2^2 m_{ij} + \beta \left\|\boldsymbol{W}^T\boldsymbol{X}\right\|_F^2 \\ s.t. \quad & \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_p \end{aligned} \tag{3.3}$$

where $n = n_s + n_t$ represents the total number of the instances both in the source domain and the target domain, $\boldsymbol{X} = [\boldsymbol{X}_s, \boldsymbol{X}_t] \in \mathbb{R}^{d\times n}$ represents the instances matrix consisted by all instances both in two domains, and $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent the i-th instance and the j-th instance in $\boldsymbol{X}$, respectively. The last term controls the complexity of latent space. The third term indicates the cross-domain manifold structure information with neighbor instances in the latent space and m_{ij} represents the weights between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which is defined as follows:

$$m_{ij} = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2}\right), & y_i = y_j \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

where y_i and y_j represent the ground truth or the pseudo label in the source domain or the target domain, respectively.

To explore the global structure more effectively and eliminate redundant information, we adopt the low-rank regularizor to the transfer structure matrix $\boldsymbol{S}$, i.e., $rank(\boldsymbol{S})$. However, it belongs to an NP-hard problem [70] for solving the $rank(\cdot)$ regularizer directly. Instead, we utilize the nuclear-norm regularization $\|\cdot\|_*$ [70] to calculate its approximate lower-bound, and we can reformulate Eq (3.3) to obtain the final objective function as follows:

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{S}} \quad & \tfrac{1}{n_t}\left\|\boldsymbol{X}_t - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s\boldsymbol{S}\right\|_F^2 \\ & +\tfrac{\lambda_1}{N_{m_{ij}\neq 0}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\right\|_2^2 m_{ij} \\ & +\lambda_2 \left\|\boldsymbol{W}^T\boldsymbol{X}\right\|_F^2 + \lambda_3 \|\boldsymbol{S}\|_* \\ s.t. \quad & \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_p \end{aligned} \tag{3.5}$$

where $N_{m_{ij}\neq 0}$ represents the number of nonzero weight m_{ij}.

To fit the trend in regression, the least square regression (LSR) [39] as a basic cost is adopted in the procedure of prediction. Differing from classification task, there are certain correlations between

classes in regression task, especially for the tasks such as age estimation [71]. Along this line, we herein encode the class label from the source domain as cumulative attribute (CA) coding [71]:

$$Y_{ij} = \begin{cases} 1, & j \le y_i \\ 0, & j > y_i \end{cases}$$

where Y_{ij} represents the j-th coding of the i-th instance. Reflected from the above definition, this coding can depict the inherent characteristics of the neighbor-similarity and ordinality.

Moreover, due to the tuning of Eq (3.5), the well-performance representations in the latent space can be obtained while retaining the structure information. To this end, we exploit this structure information from the source domain as the transferred knowledge to predict the labels of the instances from the target domain. Herein, we apply MLG-LSC [39], which is a variation of LSR with geometric mean learning (GMML) [72], and combine it with CA coding. MLG-LSC contains a two-stage modeling strategy, and its objective function is defined as follows:

$$\min_{\boldsymbol{P},\boldsymbol{b}} \quad \|\boldsymbol{P}^T\widetilde{\boldsymbol{X}}_s + \boldsymbol{b}\mathbf{1}_N^T - \boldsymbol{Y}_s\|_F^2 + \lambda\|\boldsymbol{P}\|_F^2 \tag{3.6}$$

$$\min_{\boldsymbol{A}>0} \quad \sum_{k=1}^{K}\left(\|\boldsymbol{C}_k\|_{\boldsymbol{A}}^2 + \|\boldsymbol{D}_{-k}\|_{\boldsymbol{A}^{-1}}^2\right) \tag{3.7}$$

where $\widetilde{\boldsymbol{X}}_s = \boldsymbol{W}^T\boldsymbol{X}_s \in \mathbb{R}^{p\times n_s}$ represents the representations of the instances from the source domain in the latent space, and $\boldsymbol{Y}_s \in \mathbb{R}^{K\times n_s}$ represents the CA coding matrix generalized through the ground truth in the source domain. In regression, $\boldsymbol{P} \in \mathbb{R}^{p\times K}$ represents the projection matrix, $\boldsymbol{b} \in \mathbb{R}^K$ represents the bias vector, and $\mathbf{1}_n \in \mathbb{R}^n$ represents a vector, whose each element is one. $\boldsymbol{A}$ represents the metric matrix guided by the data, and $\|\boldsymbol{C}_k\|_{\boldsymbol{A}}^2$ measures the distance of the instances from the k-th class while $\|\boldsymbol{D}_{-k}\|_{\boldsymbol{A}^{-1}}^2$ measures the distance of the instances besides the k-th class. For more details, please refer to reference [39]. In essence, Eq (3.6) is the objective function of LSR and Eq (3.7) is its GMML extension. Optimizing the above two objective functions jointly, we can obtain the optimized parameters $\boldsymbol{P}$, $\boldsymbol{b}$ and $\boldsymbol{A}$ based on the data from the source domain. Subsequently, the predicted labels of instances from the target domain can be obtained through the following rule:

$$\ell(\boldsymbol{x}_{ti}) = \arg\min_k \left\{\left(\boldsymbol{P}^T\widetilde{\boldsymbol{x}}_{ti} + \boldsymbol{b} - \boldsymbol{Y}^k\right)^T \boldsymbol{A}\left(\boldsymbol{P}^T\widetilde{\boldsymbol{x}}_{ti} + \boldsymbol{b} - \boldsymbol{Y}^k\right)\right\} \tag{3.8}$$

where $\widetilde{\boldsymbol{x}}_{ti} = \boldsymbol{W}^T\boldsymbol{x}_{ti}$ represents the representation of i-th instance from the target domain in the latent space, and $\boldsymbol{Y}^k$ represents the CA coding of the k-th class.

3.3. Optimization algorithm

In this subsection, we provide the optimization algorithm in detail. Specifically, due to the non-convexity of Eq (3.5), we adopt the alternating optimization strategy to solve each of the optimization subproblems. Meanwhile, to optimize the nuclear norm regularization $\|\boldsymbol{S}\|_*$, we adopt the SVT strategy [70] to obtain the analytic solution of S.

To this end, for convenience, we reformulate Eq (3.5) through an auxiliary variable $\boldsymbol{Z}$.

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{S},\boldsymbol{Z}} \quad & \frac{1}{n_t}\left\|\boldsymbol{X}_t - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s\boldsymbol{S}\right\|_F^2 \\ & + \frac{\lambda_1}{N_{c_{ij}\neq 0}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\right\|_2^2 m_{ij} \\ & + \lambda_2\left\|\boldsymbol{W}^T\boldsymbol{X}\right\|_F^2 + \lambda_3\left\|\boldsymbol{Z}\right\|_* \\ s.t. \quad & \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_p \\ & \boldsymbol{S} = \boldsymbol{Z} \end{aligned} \tag{3.9}$$

Then, we adopt the inexact augmented Lagrange multiplier (ALM) optimization algorithm [70] to solve the Eq (3.9) and the eventual optimization objective can be formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{S},\boldsymbol{Z}} \quad & \frac{1}{n_t}\left\|\boldsymbol{X}_t - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s\boldsymbol{S}\right\|_F^2 \\ & + \frac{\lambda_1}{N_{c_{ij}\neq 0}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\right\|_2^2 m_{ij} \\ & + \lambda_2\left\|\boldsymbol{W}^T\boldsymbol{X}\right\|_F^2 + \lambda_3\left\|\boldsymbol{Z}\right\|_* \\ & + tr\left(\Phi\left(\boldsymbol{S} - \boldsymbol{Z}\right)\right) + \frac{\mu}{2}\left\|\boldsymbol{S} - \boldsymbol{Z}\right\|_F^2 \\ s.t. \quad & \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_p \end{aligned} \tag{3.10}$$

Optimize $\boldsymbol{W}$ when fixing $\boldsymbol{S}$ and $\boldsymbol{Z}$. According to reference [73], we can obtain the following objective function w.r.t. $\boldsymbol{W}$:

$$\begin{aligned} \min_{\boldsymbol{W}} \quad & tr\left(\boldsymbol{W}^T\boldsymbol{X}\left(\frac{N_{c_{ij}\neq 0}}{\lambda_1 n_t}\begin{pmatrix} \boldsymbol{S}\boldsymbol{S}^T & -\boldsymbol{S} \\ -\boldsymbol{S}^T & \boldsymbol{0}_{n_t} \end{pmatrix}\right.\right. \\ & \left.\left. + \boldsymbol{L} + \frac{\lambda_2 N_{c_{ij}\neq 0}}{\lambda_1}\boldsymbol{I}_n\right)\boldsymbol{X}^T\boldsymbol{W}\right) \\ s.t. \quad & \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W} = \boldsymbol{I}_p \end{aligned} \tag{3.11}$$

where $\boldsymbol{L} = \boldsymbol{H} - \boldsymbol{M} \in \mathbb{R}^{n\times n}$ represents a Laplacian matrix, $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ represents the adjacency matrix composed of m_{ij}, and $\boldsymbol{H} = diag\left\{\sum_j^n m_{0j}, \cdots, \sum_j^n m_{nj}\right\}$. $\boldsymbol{I}_n$ represents the n-order indentity matrix and $\boldsymbol{0}_{n_s}$ represents the n_s-order zero matrix. Herein, the optimization of Eq (3.11) can be derived as a generalized eigen-decomposition problem:

$$\begin{aligned} \boldsymbol{X}\left(\frac{N_{c_{ij}\neq 0}}{\lambda_1 n_t}\begin{pmatrix} \boldsymbol{S}\boldsymbol{S}^T & -\boldsymbol{S} \\ -\boldsymbol{S}^T & \boldsymbol{0}_{n_t} \end{pmatrix} + \boldsymbol{L} + \frac{\lambda_2 N_{c_{ij}\neq 0}}{\lambda_1}\boldsymbol{I}_n\right)\boldsymbol{X}^T\boldsymbol{W} \\ = \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Lambda} \end{aligned} \tag{3.12}$$

where $\boldsymbol{\Lambda}$ represents the diagonal matrix composed of all Lagrange multipliers, that is, eigenvalues, and $\boldsymbol{W}$ consists of corresponding p smallest eigenvectors.

Optimize $\boldsymbol{S}$ when fixing $\boldsymbol{W}$ and $\boldsymbol{Z}$. Along this line, we reformulate Eq (3.9) w.r.t. $\boldsymbol{S}$ as follows:

$$\begin{aligned} \mathcal{J}\left(\boldsymbol{S}\right) = \min_{\boldsymbol{S}} \quad & \frac{1}{n_t}\left\|\boldsymbol{X}_t - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s\boldsymbol{S}\right\|_F^2 \\ & + tr\left(\Phi\left(\boldsymbol{S} - \boldsymbol{Z}\right)\right) + \frac{\mu}{2}\left\|\boldsymbol{S} - \boldsymbol{Z}\right\|_F^2 \end{aligned} \tag{3.13}$$

Set derivative $\frac{\partial\mathcal{J}(\boldsymbol{S})}{\partial\boldsymbol{S}} = 0$, and we can obtain the closed form of $\boldsymbol{S}$:

$$\begin{aligned} \frac{\partial\mathcal{J}(\boldsymbol{S})}{\partial\boldsymbol{S}} = & \left(2\boldsymbol{X}_s^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s + \mu\boldsymbol{I}_{n_s}\right)\boldsymbol{S} \\ & -2\boldsymbol{X}_s^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_t - \mu\boldsymbol{Z} + \Phi^T \\ = & \ 0 \end{aligned} \tag{3.14}$$

$$\Rightarrow \boldsymbol{S} = \begin{array}{l} \left(2\boldsymbol{X}_s^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_s + \mu\boldsymbol{I}_{n_s}\right)^{-1} \\ \cdot\left(2\boldsymbol{X}_s^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}_t + \mu\boldsymbol{Z} - \Phi^T\right) \end{array} \tag{3.15}$$

Optimize $\boldsymbol{Z}$ when fixing $\boldsymbol{W}$ and $\boldsymbol{S}$. Analogous to optimizing $\boldsymbol{S}$, we can reformulate Eq (3.9) w.r.t. $\boldsymbol{Z}$ as follows:

$$\mathcal{J}(\boldsymbol{Z}) = \min_{\boldsymbol{Z}} \quad \tfrac{\lambda_3}{\mu}\|\boldsymbol{Z}\|_* + \tfrac{1}{2}\|\boldsymbol{Z} - \boldsymbol{S} - \boldsymbol{\Phi}/\mu\|_F^2 \tag{3.16}$$

Set SVD $(\boldsymbol{Z} + \boldsymbol{\Phi}/\boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where σ_i represents the i-th singular value. Therefore, according to SVT, we can obtain the analytic solution of $\boldsymbol{Z}$.

$$\boldsymbol{Z}_{t+1} = \quad \boldsymbol{U} \times diag\left(\left\{\max\left(0, \sigma_i - \tfrac{\lambda_3}{\mu}\right)_{1\le i\le r}\right\}\right) \times \boldsymbol{V}^T \tag{3.17}$$

Optimize $\boldsymbol{P}$, $\boldsymbol{b}$, and $\boldsymbol{A}$. According to reference [39], we give their solutions directly.

$$\boldsymbol{P} = (\widetilde{\boldsymbol{X}}_s\boldsymbol{G}\widetilde{\boldsymbol{X}}_s^T + \lambda\boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{X}}_s\boldsymbol{G}\boldsymbol{Y}_s^T \tag{3.18}$$

$$\boldsymbol{b} = \frac{1}{n_s}\left(\boldsymbol{Y}_s\boldsymbol{1}_{n_s} - \boldsymbol{P}^T\boldsymbol{X}_s\boldsymbol{1}_{n_s}\right) \tag{3.19}$$

$$\boldsymbol{A} = (\boldsymbol{C} + \beta\boldsymbol{I}_k)^{-1}\sharp_\alpha(\boldsymbol{D} + \beta\boldsymbol{I}_k) \tag{3.20}$$

where $\sharp$ represents the geometric mean operator and $\boldsymbol{C} = \sum_{k=1}^{K}\boldsymbol{C}_k^T\boldsymbol{C}_k$, $\boldsymbol{D} = \sum_{k=1}^{K}\boldsymbol{D}_{-k}^T\boldsymbol{D}_{-k}$. $\alpha \in [0, 1]$, λ, and β are three nonnegative hyper-parameters, and $\boldsymbol{G} = \boldsymbol{I}_{n_s} - (1/n_s)\,\boldsymbol{1}_{n_s}\boldsymbol{1}_{n_s}^T$. For more details of solutions, please refer to reference [39].

By repeating the above steps iteratively until convergence, we can predict the label y_{tp} of the instances from the target domain without access to the ground truth. It's worth noting that λ and α are chosen through the gird search with the data from the source domain in each iteration. We summarize the complete optimization algorithm in Algorithm 1.

3.4. Time and space complexity analysis

First, we provide the time complexity with optimization variables for single step updates. When $\boldsymbol{S}$ and $\boldsymbol{Z}$ are fixed, the time complexity of updating $\boldsymbol{W}$ is $O\left(nd^2 + dn^2 + nd^2 + d^3\right)$. When $\boldsymbol{W}$ and $\boldsymbol{Z}$ are fixed, the time complexity of updating $\boldsymbol{S}$ is $O\left(n_s^3 + pn_s^2 + pdn_s + pn_sn_t + pdn_t + n_tn_s^2\right)$. When $\boldsymbol{W}$ and $\boldsymbol{S}$ are fixed, the time complexity of updating $\boldsymbol{Z}$ is $O\left(n_t^3\right)$. The time complexity of updating $\boldsymbol{P}$ is $O\left(pn_s^2 + n_s^3 + Kpn_s\right)$. The time complexity of updating $\boldsymbol{b}$ is $O(Kn_s + Kpn_s)$. The time complexity of updating $\boldsymbol{A}$ is $O\left(K^3 + K^2n_s\right)$. The time complexity of updating y_{tp} is $O\left(K^2n_t + Kn_t\right)$. Second, for the sake of clarity, we assume that the number of iterations for convergence is L, and the combination of λ and α is u. Generally, we can have $K < p < d$ and $d < \min\{n_s, n_t\}$. The complete time complexity of Algorithm 1 is $O\left(Ldn^2 + L\max\left\{n_s^3, n_tn_s^2\right\} + Ln_t^3 + Lun_s^3 + LuK^2n_s + LuK^2n_t\right)$. Obviously, the space complexity is $O(n_s \cdot n_t)$, which is decided by the transfer structure matrix $\boldsymbol{S}$.

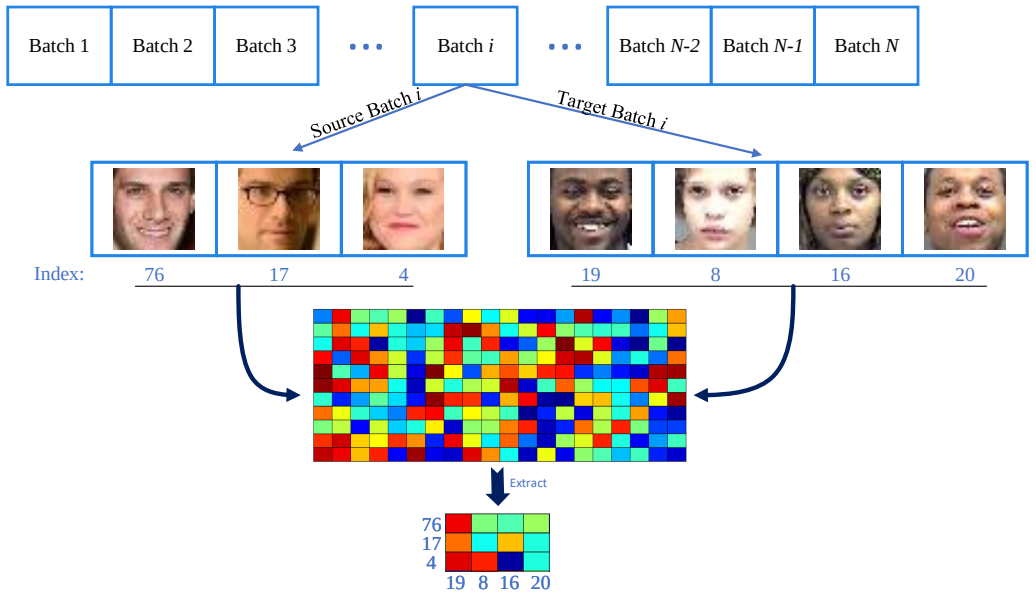


Figure 3. Illustration of gerenating $\boldsymbol{S}$. Take the i-th batch as an example, which consists of several instances from the source domain and the target domain, respectively. We can also obtain the indexes of corresponding instances. As shown in the figure, the '4', '17', and '76' represent the 4th, 17th, 76th instances from the source domain, respectively, and the '8', '16', '19', '20' from the target domain have the same meaning. Thus, in the i-th batch during training, the corresponding FC layer parameters are extracted from the whole transferred structure matrix $\boldsymbol{S}$ via the current indexes.

3.5. SPODA with deep architecture

Motivated by the huge success of deep learning across various tasks, we extend SPODA to a deep network architecture, which we refer to as D-SPODA. Specifically, we herein regard the components $\boldsymbol{E}$ and $\boldsymbol{D}$ as an auto-encoder network. In this work, the encoder and the decoder are constructed based on AlexNet [74]. The specific architecture of the encoder is shown in Figure 4, and the decoder adopts the opposite transmission structure. Moreover, the component $\boldsymbol{P}$ consists of multiple FC layers, functioning as a classifier following the encoder network. To facilitate the optimization process, we treat the transferred structure matrix $\boldsymbol{S}$ as an FC layer positioned after the decoder network. Given the batch training mechanism, we adopt a dynamic FC layer generated from the index of each training batch, as depicted in Figure 3. Intuitively, the entire transferred structure matrix can become low-rank to some extent after training, since the partial matrices generated by the indices of randomly selected training data aim to achieve low-rank characteristics during each epoch. At the input layer, aligned face images of size $227 \times 227 \times 3$ are fed to the network. At the output layer, we maintain pretrained auto-encoder deep architecture and a classifier network by replacing its loss function with our SPODA objective function in Eq (3.5) instead of the conventional cross-entropy loss with the softmax function. Furthermore, cumulative attribute coding replaces one-hot coding in our output targets, optimized using an SGD solver. In this way, the number of last full connection layer output is adjusted to the total number of data classes K. Additionally, we employ the MLG-LSC strategy to update the parameters involved in predicting pseudo-labels at decreasing intervals.

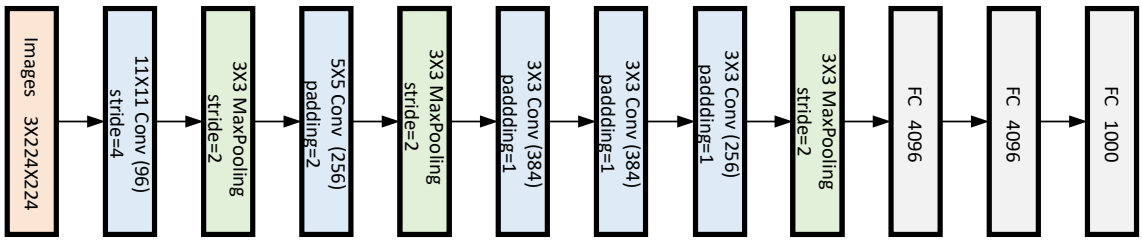


Figure 4. The network architecture of encoder $\boldsymbol{E}$.

4. Experiment

In this section, we construct extensive experiments to validate the effectiveness of our proposed SPODA methods digit datasets and human facial datasets. Besides, we also analyze other properties of SPODA with charts, e.g., the hyper-parameters sensitivity, and the time cost.

4.1. Datasets and settings

In this article, three widely used digit datasets and three real-world human facial datasets are used to validate the effectiveness and superiority of our proposed methods. The digit datasets [33] contain MNIST (M), SVHN (S), and USPS (U), and the human facial benchmark datasets contain AgeDB (A) [75], Morph Ablum II (II) [34], and CACD (C) [71]. All digit datasets contain ten digit images, which belong to $\{0, \ldots, 9\}$ sampling from different scenarios. Meanwhile, the AgeDB dataset contains 16,516 face images aged from 0 to 101 years. The Morph Ablum II dataset contains more than 55, 000 face images aged from 15 to 85 years. The CACD dataset is the largest dataset widely used for human facial age estimation, which contains more than 160,000 face images aged from 14 to 62 years. Different from the digit datasets, where only semantic labels are ordinal, the ordinarily is both in representation level and semantic level on the human facial datasets.

In the procedure of feature extraction from images in conventional machine learning, to effectively evaluate the generalization ability of feature representation variances, we herein extracted the DeCAF representations [23] for the digit datasets and the ResNet-50 representations pretrained by ODL strategy [11] with cross-entropy loss for the human facial datasets. Since the label space of the target domain coincides with the source domain, we adopt the overlapped classes among the three above datasets, that is all instances aged from 16 to 62 years. In deep architecture, described as subsection 3.5, the AlexNet architectures [74] are adapted to build the encoder and the decoder network, and a classifier network is built by two FC layers, where a ReLU layer followed by a dropout layer is sandwiched between them. The transferred structure matrix is implemented by a dynamic FC layer according to each batch. The input size of both digit datasets and human facial datasets are $227 \times 227 \times 3$ randomly cropped from $256 \times 256 \times 3$ resized images.

For the hyper-parameters involved in the proposed method, in conventional machine learning, we set $\lambda_1 \in \{1e\text{-}1, 1e0, \cdots, 1e4\}$, $\lambda_2 \in \{1e\text{-}4, 1e\text{-}3, \cdots, 1e1\}$, $\lambda_3 \in \{1e\text{-}4, 1e\text{-}3, \cdots, 1e3\}$, and the dimension of latent space $p \in \{50, 60, 70, \cdots, 200\}$. Then, five-fold cross validation with grid search is applied to choose the suitable hyper-parameters. Meanwhile, the hyper-parameters built in Eq (3.8), i.e., λ, α, and β, are set in $\{1e\text{-}2, 1e\text{-}1, 1e0\}$, $\{0, 0.1, \cdots, 1\}$, and $\{1e\text{-}6\}$, respectively. They are also chosen

Algorithm 1 Optimization Algorithm for SPODA

Input:

The data matrix $\boldsymbol{X}_s$ from the source domain, their ground truth $\boldsymbol{y}_s$, the corresponding CA coding matrix $\boldsymbol{Y}_s$, the data matrix $\boldsymbol{X}_t$ from the source domain.

Hyper-parameters λ_1, λ_2, λ_3, β.

Output:

The predicted label y_{tp} of the instances from the target domain.

Initialize $\boldsymbol{A} = \boldsymbol{I}_K$, $\boldsymbol{Z} = \boldsymbol{I}_{n_s \times n_t}$, $\mu = 1e\text{-}6$, $\boldsymbol{\Phi} = \boldsymbol{0}_K$, $\mu_{\max} = 1e2$, $\rho = 1.1$.
Initialize y_{tp} via k-NN (k = 15).
repeat
  Update m_{ij} Eq (3.4).
  Update $\boldsymbol{W}$ Eq (3.12);
  Update $\boldsymbol{S}$ Eq (3.15);
  Update $\boldsymbol{Z}$ Eq (3.17);
  for the combinations of λ and α **do**
    Update $\boldsymbol{P}$ Eq (3.18);
    Update $\boldsymbol{b}$ Eq (3.19);
    Update $\boldsymbol{A}$ Eq (3.20);
  end for
  Update y_{tp} Eq (3.8);
  Update $\boldsymbol{\Phi} = \boldsymbol{\Phi} + \mu(\boldsymbol{S} - \boldsymbol{Z})$;
  Update $\mu = \min(\rho\mu, \mu_{\max})$;
until convergence of Eq (3.10)

by grid search during alternately optimizing, but not be chosen coupled with λ_1, λ_2, and λ_3. In deep architecture, they are set by different constant values according to each task.

To evaluate the generalization ability, mean absolute error (MAE) and cumulative score (CS) are adopted, in which they are respectively defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_{ti} - y_{ti}|$$

$$CS = \frac{N_{\varepsilon \le \theta}}{N} \times 100\%$$

where $\hat{y}_{ti}$ and y_{ti} denote, respectively, the predicted and ground-truth labels of i-th instances from target domain, and $N_{\varepsilon \le \theta}$ denotes the number of the instances, whose error ε between the predicted and ground-truth label is not greater than the error level θ.

To validate the effectiveness of our proposed SPODA, several related methods are introduced for comparison, which are i) TCA [17], ii) JDA [20], iii) BDA [19], iv) CORAL [43], v) MEDA [12], vi) Easy-TL [52], vii) MCTL [23], viii) GSL [54] for conventional methods, and i) Deep-CORAL [76], ii) DANN [40], iii) DAAN [77], and iv) RSD [66] for deep architecture, respectively. Additionally, their hyper-parameters are assigned by five-fold cross-validation according to the corresponding literatures. It's worth noting that except for Easy-TL and SPODA (ours), in the conventional setting, all the others adopt kernel trick with radial basis function (RBF) kernel [71].

Finally, all experiments are run on Python 3.8 and PyTorch 1.5 platform of Ubuntu 20.04 LTS with Intel Xeon Gold 5217 CPU@3.7Ghz, 128GB RAM, and Nvidia RTX Titan GPU.

Table 1. MAE results comparison on three digit datasets.

Methods	M → S	M → U	S → M	S → U	U → M	U → S	Avg
TCA	3.244	3.393	1.394	3.441	1.479	3.198	2.692
JDA	3.221	3.382	1.424	3.466	1.376	3.006	2.646
BDA	3.219	3.347	1.425	3.463	1.372	3.008	2.639
CORAL	3.370	3.900	1.558	3.246	1.229	3.061	2.727
MEDA	2.901	3.671	1.369	3.261	1.220	3.025	2.575
Easy-TL	2.369	3.275	1.372	3.008	1.144	3.105	2.379
MCTL	1.815	2.974	1.301	3.190	1.620	3.099	2.333
GSL	1.842	2.477	1.286	2.586	1.483	2.862	2.089
SPODA	2.236	2.381	1.352	2.375	1.363	2.493	2.033
DANN	*1.467*	1.334	*1.169*	*2.070*	1.074	***1.829***	1.491
DAAN	***1.454***	1.111	***1.135***	***2.036***	0.947	1.963	*1.441*
RSD	1.484	*1.092*	1.351	2.192	*0.936*	1.851	1.484
D-SPODA	1.503	***1.064***	1.200	2.071	***0.891***	*1.834*	***1.427***

4.2. Results and analysis

With the setup in the Section 4.1, the experiments are run on all instances on the digit datasets. Since the human facial datasets are too large, to avoid high memory usage and computational complexity caused by the RBF kernel, we sample 50 instances from each class of the source domain and the target domain randomly. Along this line, the experiments are run on 10 trials on the human facial datasets in conventional cases. Therefore, the results are reported in both Tables 1 and 3, and the averaged results with standard deviations are reported in Table 2. In each table, the best results are in bold and italic face and the second are in underline and italic face.

From Tables 1–3, we can obverse the following findings. First, in most cases, our proposed methods SPODA or D-SPODA achieved the best MAE results, especially in conventional case. It is worth noting that SPODA adopts linear trick while kernel trick are adopted into the others in conventional case, and D-SPODA utilizes AlexNet, whose architecture is more shallow than the others in deep case. These facts directly testify its generalization, effectiveness, and superiority. Second, the MAE results of SPODA are unstable on the digit datasets. We argue that the ordinality of the digit datasets is in the semantic level, but the representation level is relatively discrete according to the human facial datasets. The characteristic promotes the performances of the methods about domain adaptation classification. Third, in the almost case, the MAE results of TCA, JDA, and BDA are inferior to the other models. This indicates that on the regression problem with fuzzy category boundary, although the alignment strategy, i.e., MMD, can extract the most relevant features between domains, it is easy to destroy the original structure relationships of representations, leading to the loss of discrimination ability of the model. Fourth, the MAE results of MEDA are better than the above methods with MMD, and have certain advantages over CORAL. It indicates that in regression, even in human facial age estimation, the scheme is effective to a certain extent, where the first step is to align the representation through MMD and the second step is to take the manifold regularization term into the objective to preserve the discrimination ability. Fifth, in the conventional case, the MAE results of the last five methods are

Table 2. MAE results comparison on three human facial datasets in conventional case.

Methods	A → II	A → C	II → A	II → C	C → A	C → II	Avg
TCA	14.760 ± 0.473	15.592 ± 0.326	15.230 ± 0.471	14.454 ± 0.425	14.431 ± 0.322	14.775 ± 0.434	14.874 ± 0.409
JDA	14.339 ± 0.413	15.880 ± 0.324	14.623 ± 0.292	14.741 ± 0.342	14.748 ± 0.248	14.736 ± 0.452	14.859 ± 0.377
BDA	14.213 ± 0.409	15.560 ± 0.305	14.450 ± 0.226	14.643 ± 0.335	14.563 ± 0.305	14.614 ± 0.411	14.797 ± 0.362
CORAL	14.084 ± 0.556	15.657 ± 0.332	13.900 ± 0.275	14.000 ± 0.325	14.772 ± 0.517	14.624 ± 0.455	14.725 ± 0.374
MEDA	13.778 ± 0.549	15.088 ± 0.308	13.901 ± 0.316	13.710 ± 0.227	14.269 ± 0.349	14.136 ± 0.358	14.609 ± 0.369
Easy-TL	13.792 ± 0.539	14.399 ± 0.360	14.201 ± 0.387	13.646 ± 0.263	14.800 ± 0.324	13.636 ± 0.430	14.521 ± 0.372
MCTL	13.529 ± 0.410	14.523 ± 0.209	*12.963 ± 0.527*	*13.006 ± 0.196*	13.635 ± 0.310	13.675 ± 0.423	14.383 ± 0.368
GSL	*13.194 ± 0.243*	*13.317 ± 0.442*	13.069 ± 0.374	13.384 ± 0.232	*12.492 ± 0.457*	*12.922 ± 0.448*	*14.218 ± 0.368*
SPODA	***10.977 ± 0.388***	***12.046 ± 0.389***	***11.053 ± 0.219***	***11.405 ± 0.443***	***11.257 ± 0.298***	***12.078 ± 0.397***	***13.912 ± 0.366***

Table 3. MAE results comparison on three human facial datasets with deep architecture.

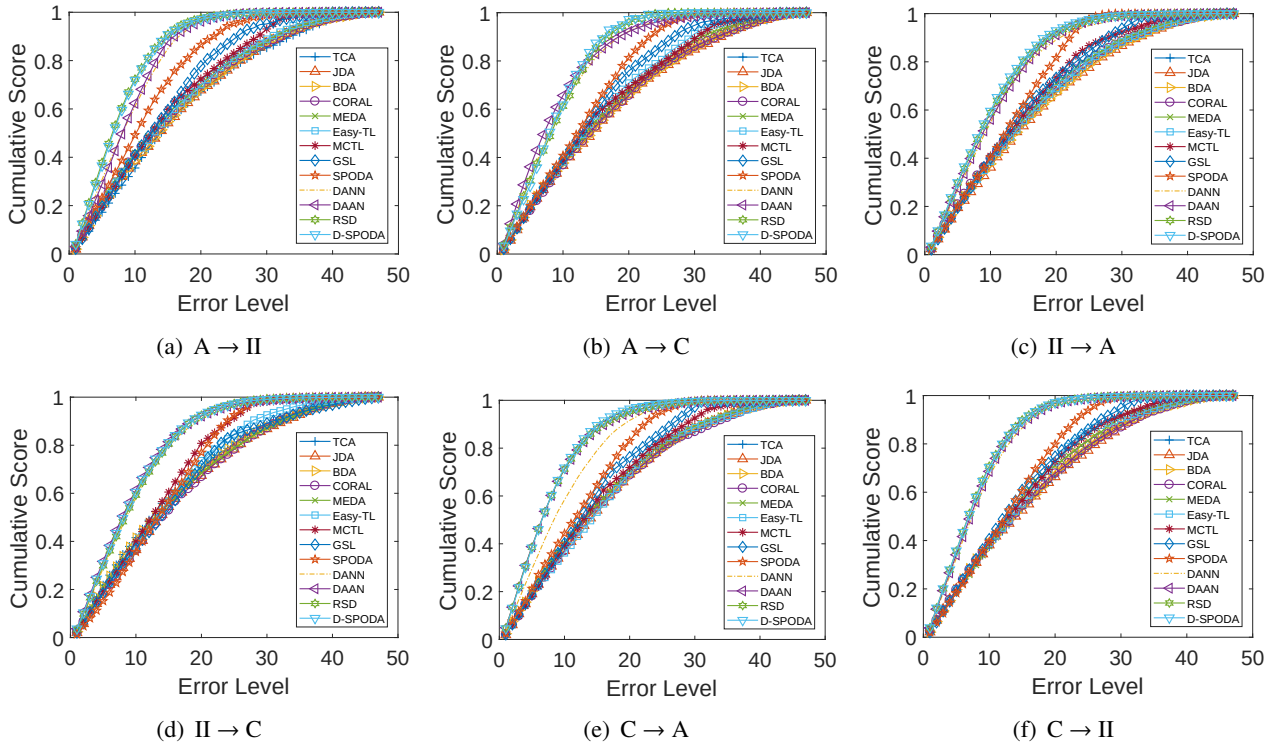
Methods	A → II	A → C	II → A	II → C	C → A	C → II	Avg
DANN	7.362	8.939	9.472	9.451	9.235	8.056	8.752
DAAN	6.804	*8.743*	8.971	9.465	7.401	7.594	8.163
RSD	***6.535***	8.863	*8.362*	*9.011*	*7.365*	***7.364***	*7.917*
D-SPODA	*6.638*	***8.676***	***8.215***	***8.861***	***7.249***	*7.499*	***7.856***

usually better than the others. It indicates that in regression, especially in human facial age estimation, it is necessary to preserve the structure information between inter-/intra- domains. Sixth, the MAE results of GSL are a bit better than MCTL. It indicates that although MCTL generates a fake target domain through the data from the source domain while preserving the structure information, the supervised information from the source domain is adopted insufficiently during training, learning to reduce discriminability. Seventh, the MAE results of GSL are worse than ours. We argue that there are two reasons. On the one hand, GSL loses the preservation of the interclass structure. On the other hand, the ordinality of the task is grossly underused in GSL. Eighth, the MAE results with deep architecture are much better than conventional case, due to the strong feature description ability of deep architecture. Ninth, analogous to MCTL and GSL, the MAE results of DANN and DAAN as classification methods are worse than regression methods such as RSD and ours. Finally, in most cases, the MAE results of RSD are worse than ours, since it could not adopt the ordinality of data. In summary, the generalization of our proposed methods are better than the others.

We also illustrate the results of CS criterion compared with all the above methods. Considering that the CS rules on digit datasets are quite similar to these on human facial datasets, we herein only show the results on human facial datasets in Figure 5. We can observe that both in the conventional case, deep case, and almost case, the curve of our proposed method is closest to the upper left corner. It validates the effectiveness and superiority of our proposed methods for DA regression once again.

Table 4. Ablation study on three human facial datasets.

Settings	A → II	A → C	II → A	II → C	C → A	C → II	Avg
$\lambda_1 = 0$	11.154	12.684	11.734	12.354	11.824	14.423	12.362
$\lambda_2 = 0$	11.128	12.634	11.685	12.272	12.884	11.624	12.038
$\lambda_3 = 0$	11.452	12.624	13.584	11.984	11.654	12.745	12.341
Full	***10.977***	***12.046***	***11.053***	***11.405***	***11.257***	***12.078***	***11.469***



(a) A → II (b) A → C (c) II → A

(d) II → C (e) C → A (f) C → II

Figure 5. Cumulative scores comparison on human facial datasets.

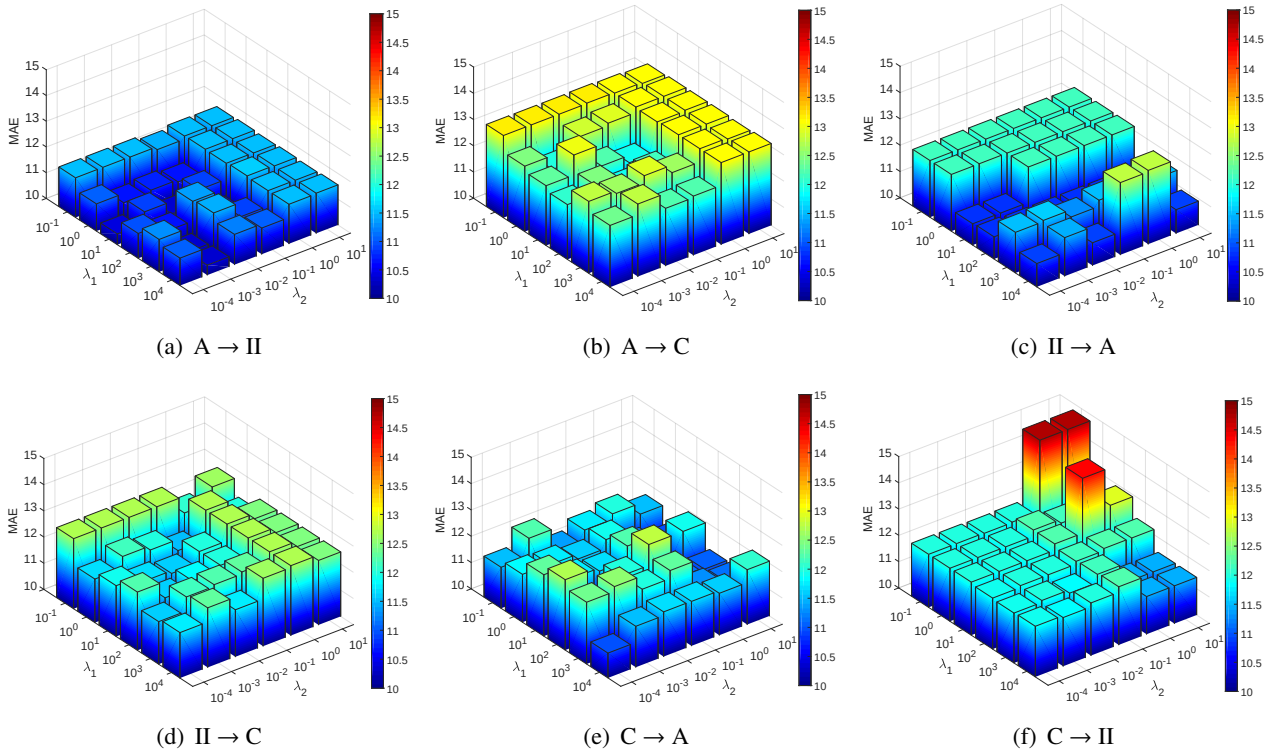
4.3. Ablation study

To demonstrate the effectiveness of each proposed module, we conduct the ablation study on three human facial datasets by setting the corresponding hyper-parameters before losses to zero. The specific experimental results are shown in Table 4. We can find that removing the cross-domain structure regularization (the LPP manifold regularization term) will result in the maximum performance degradation of the target model. Besides, the F-norm term of the latent space and low-rank regularization play a certain role in improving the adaptation performance of the model by exploring the cross-domain structure and preserving the structure information.

4.4. Parameter analysis

In this section, to analyze the parameters sensitivity, we conduct the experiments for the hyper-parameters λ_1, λ_2, λ_3 and the dimension of latent space p. Without loss of generality, we take the human facial datasets as the sample in the conventional case, uniformly. More specifically, we analyze the relationship between cross-domain structure regularization and the complexity of the latent space, and the influence of low-rank regularization and the dimension of latent space. The experimental setup is the same as that in Subsection 4.1, where the parameters are adjusted through the grid search while the others are fixed.

Relationship between cross-domain structure regularization and the complexity of the latent space We first analyze the leverage between the LPP manifold regularization term and the F-norm term of the latent space, as shown in Figure 6 for details. We can observe that the best MAE results are generally achieved when $\lambda_1 \in \{1e1, 1e2, 1e3\}$. Unlike relatively stable λ_1, λ_2 is sensitive on different



(a) A → II (b) A → C (c) II → A

(d) II → C (e) C → A (f) C → II

Figure 6. Estimation performance with varying combinations of cross-domain structure regularization coefficient λ_2 and the complexity coefficient of the latent space λ_2 on human facial datasets.

dataset pairs. Nevertheless, we can take an attempt to achieve the best performance where $\lambda_2 \in \{1e\text{-}2, 1e\text{-}1\}$.

Influence of low-rank regularization From Figure 6(a)–(f), we can observe that in most cases, the MAE result decreases as the weight of the low-rank regularization becomes larger. This fact reveals that the penalty of low rank is positive to explore and exploit the cross-domain structure, and to preserve the structure information.

Influence of the dimension of latent space From Figure 6(g)–(l), we can observe that as a whole, the MAE result has an increasing trend as the dimension increases. Also, we find that the best performance can be achieved when $p \in \{50, \cdots, 70\}$. Meanwhile, this fact reveals that the reconstruction strategy can drop the style information of the source domain and extract more robust representations to some extent.

4.5. Time efficiency comparsion

To evalute the time cost of SPODA during training, as the same as Section 4.4, we herein provide the results about time efficiency with other methods in the conventional case, reported in Table 5.

We can observe the following findings. First, the time cost of CORAL and Easy-TL is less than the others evidently. The reason is that CORAL is convex directly without iterative optimization, and Easy-TL based on CORAL owns one more step to calculate the centers of the source domain classes. Second, the time cost of JDA, BDA, and MCTL is more than the others evidently. The reason is that

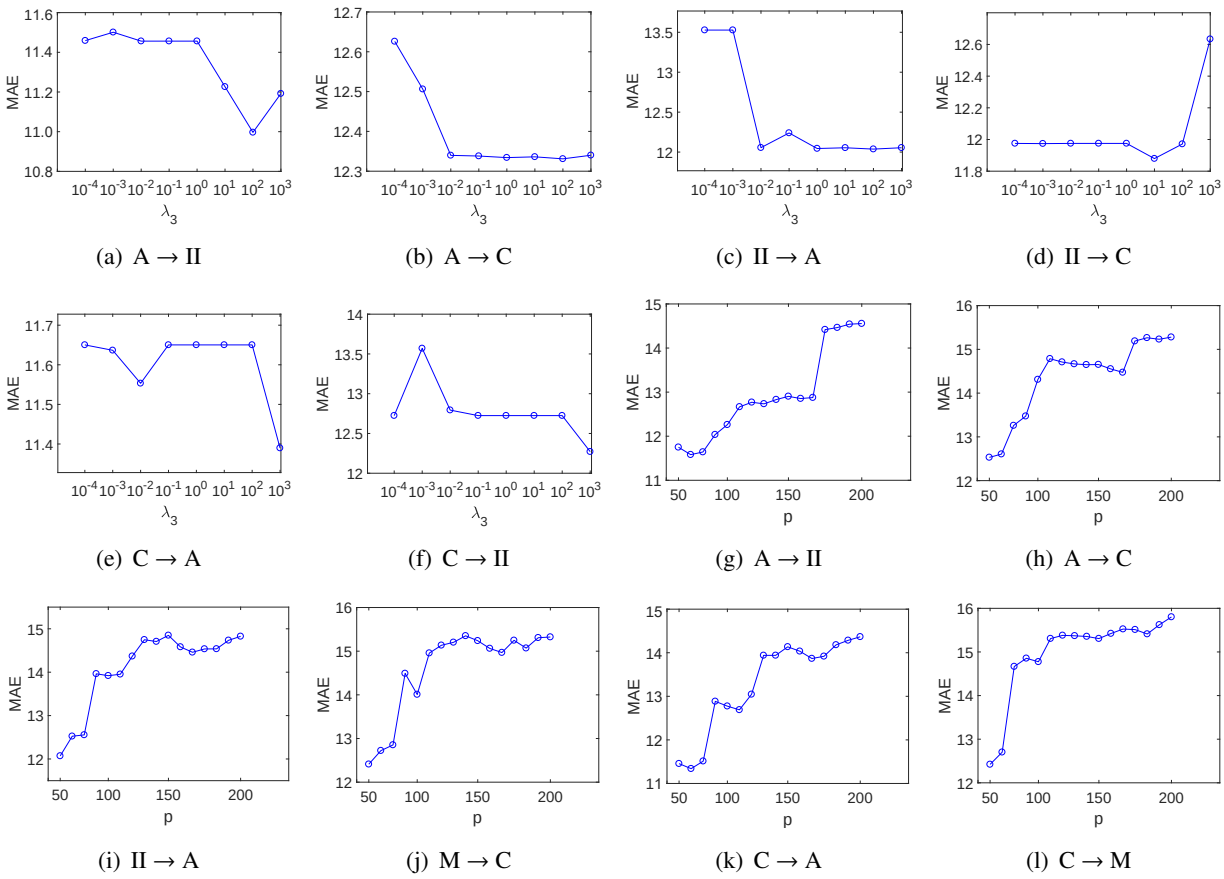


Figure 7. (a)–(f): Estimation performance with varying combinations of low-rank regularization coefficient λ_3 on human facial datasets. (g)–(l): Estimation performance with varying combinations of the dimension of latent space p on human facial datasets.

they adopt the kernel trick leading to dimensionality upgrading while their solution must be optimized iteratively. Third, the time cost of TCA and MEDA is the second highest. Although TCA is convex directly, the kernel trick is adopted and leads to slow running efficiency. Analogous to TCA, MEDA adopts matrix inverse operation instead. Fourth, the time cost of our proposed method SPODA is at a medium level and faster than GSL.

5. Limitations and potential drawbacks

While this work provides some valuable insights into addressing the problem of ordinal UDA, we must acknowledge the limitations and potential shortcomings of our approach. Based on the time complexity analysis and time efficiency comparison, we can find that SPODA is computationally intensive and requires higher computational costs. A more concise and efficient framework needs to be developed to accelerate model convergence and improve model performance. Besides, the existing datasets may suffer from class imbalance, where each dataset contains facial images of different ages for age estimation. How to overcome this inherent problem at the data level to avoid information masking is a direction that needs to be studied.

Table 5. Time efficiency comparison (in seconds) on three human facial datasets in conventional case.

Methods	A → II	A → C	II → A	II → C	C → A	C → II
TCA	294.507±0.323	298.788±3.268	300.365±2.326	286.415±0.964	296.663±2.224	310.216±0.857
JDA	1496.324±26.319	1526.914±20.844	1483.640±32.618	1500.306±33.325	14863.351±25.639	1524.446±23.633
BDA	1511.169±24.812	1501.681±36.153	1492.325±36.954	1563.360±32.324	1493.659±40.327	1596.326±26.350
CORAL	37.348±0.303	38.440±1.030	37.178±0.506	38.006±2.045	41.007±0.903	38.957±1.083
MEDA	234.510±1.535	239.628±1.110	239.062±1.093	242.656±2.123	227.628±8.341	218.285±2.100
Easy-TL	44.157±0.044	45.569±0.177	44.350±0.152	45.511±0.099	49.103±0.288	46.080±0.431
MCTL	1591.829±24.320	1630.797±29.654	1551.575±26.362	1600.721±20.365	1617.210±30.652	1640.070±35.120
GSL	931.768±30.397	890.497±20.862	957.199±17.662	963.023±10.247	987.655±29.884	984.602±38.727
SPODA	651.656±28.316	612.778±39.513	686.732±22.396	627.042±37.828	609.924±26.639	626.326±40.792

6. Conclusions

In this article, we proposed a structure-oriented adaptation model, namely, SPODA. Specifically, we achieved exploring the cross-domain structure knowledge through cross-domain structure transfer learning via an auto-encoder. In addition, the manifold prior was incorporated to preserve the cross-domain local structure. Considering the neighbor similarity and ordinality of the sample in order to depict these inherent characteristics more precisely, CA coding was introduced to encode the label of the source domain sample. The MLG-LSC method was embedded and effectively combined with CA coding to construct the proposed model. In this way, the discriminative boundary was obtained so that the performance for the prediction of the instances from the target domain was improved. Then, inspired by the inexact ALM optimization algorithm, we derived an alternating optimization algorithm to efficiently solve the SPODA model. To further boost the performance and generalization, SPODA was extended with deep neural network architecture and achieved better results than before. Finally, through extensive experiments, we have verified the effectiveness and superiority of the proposed methods.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 62176128, the Natural Science Foundation of Jiangsu Province under Grant BK20231143, the Fundamental Research Funds for the Central Universities No. NJ2023032, the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund, the Qing Lan Project of Jiangsu Province, as well as Postgraduate Research & Practice Innovation Program of Jiangsu Province SJCX24_0455.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. Y. Liu, Z. Zhou, B. Sun, Cot: Unsupervised domain adaptation with clustering and optimal transport, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 19998–20007. https://doi.org/10.1109/CVPR52729.2023.01915
2. M. Wang, S. Wang, X. Yang, J. Yuan, W. Zhang, Equity in unsupervised domain adaptation by nuclear norm maximization, *IEEE Trans. Circuits Syst. Video Technol.*, **34** (2024), 5533–5545. https://doi.org/10.1109/TCSVT.2023.3346444
3. M. Wang, Y. Liu, J. Yuan, S. Wang, Z. Wang, W. Wang, Inter-class and inter-domain semantic augmentation for domain generalization, *IEEE Trans. Image Process.*, **33** (2024), 1338–1347. https://doi.org/10.1109/TIP.2024.3354420
4. T. Mikolov, M. Karafiát, L. Burget, J.Cernockỳ, S. Khudanpur, Recurrent neural network based language model, in *11th Annual Conference of the International Speech Communication Association*, (2010), 1045–1048. https://doi.org/10.21437/INTERSPEECH.2010-343
5. I. Sutskever, J. Martens, G. Hinton, Generating text with recurrent neural networks, in *Proceedings of the 28th international conference on machine learning*, (2011), 1017–1024.
6. T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, S. Khudanpur, Extensions of recurrent neural network language model, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, (2019), 5528–5531. https://doi.org/10.1109/ICASSP.2011.5947611
7. Y. Wang, S. Tang, Y. Lei, W. Song, S. Wang, M. Zhang, Disenhan: Disentangled heterogeneous graph attention network for recommendation, in *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, (2020), 1605–1614. https://doi.org/10.1145/3340531.3411996
8. Y. Wang, Y. Li, S. Li, W. Song, J. Fan, S. Gao, et al., Deep graph mutual learning for cross-domain recommendation, in *International Conference on Database Systems for Advanced Applications*, (2022), 298–305. https://doi.org/10.1007/978-3-031-00126-0_22
9. Y. Wang, X. Luo, C. Chen, X. Hua, M. Zhang, W. Ju, Disensemi: Semi-supervised graph classification via disentangled representation learning, *IEEE Trans. Neural Networks Learn. Syst.*, (2024). https://doi.org/10.1109/TNNLS.2024.3431871
10. I. Shlizerman, S. Suwajanakorn, S. Seitz. Illumination-aware age progression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2014), 3334–3341. https://doi.org/10.1109/CVPR.2014.426
11. H. Liu, J. Lu, J. Feng, J. Zhou, Ordinal deep learning for facial age estimation, *IEEE Trans. Circuits Syst. Video Technol.*, **29** (2019). https://doi.org/10.1109/TCSVT.2017.2782709
12. J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, P. Yu, Visual domain adaptation with manifold embedded distribution alignment, in *Proceedings of the 26th ACM international Conference on Multimedia*, (2018), 402–410. https://doi.org/10.1145/3240508.3240512
13. C. Ren, Y. Zhai, Y. Luo, H. Yan, Towards unsupervised domain adaptation via domain-transformer, *Int. J. Comput. Vis.*, **132** (2024), 6163–6183. https://link.springer.com/article/10.1007/s11263-024-02174-9

14. S. David, J. Blitzer, K. Crammer, F. Pereira, Analysis of representations for domain adaptation, *Adv. Neural Inf. Process. Syst.*, **19** (2007), 137–144.

15. S. David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, A theory of learning from different domains, *Mach. learn.*, **79** (2010), 151–175. https://doi.org/10.1007/s10994-009-5152-4

16. V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer Science & Business Media, 2006. https://doi.org/10.1007/0-387-34239-7

17. S. Pan, I. Tsang, J. Kwok, Q. Yang, Domain adaptation via transfer component analysis, *IEEE Trans. Neural Networks Learn. Syst.*, **22** (2010), 199–210. https://doi.org/10.1109/TNN.2010.2091281

18. R. Combes, H. Zhao, Y. Wang, G. Gordon, Domain adaptation with conditional distribution matching and generalized label shift, *Adv. Neural Inf. Process. Syst.*, **33** (2020).

19. J. Wang, Y. Chen, S. Hao, W. Feng, Z. Shen, Balanced distribution adaptation for transfer learning, in *2017 IEEE International Conference on Data Mining*, (2017), 1129–1134. https://doi.org/10.1109/ICDM.2017.150

20. M. Long, J. Wang, G. Ding, J. Sun, P. Yu, Transfer feature learning with joint distribution adaptation, in *Proceedings of the IEEE International Conference on Computer Vision*, (2013), 2200–2207. https://doi.org/10.1109/ICCV.2013.274

21. S. Li, S. Song, G. Huang, Prediction reweighting for domain adaptation, *IEEE Trans. Neural Networks Learn. Syst.*, **28** (2016), 1682–1695. https://doi.org/10.1109/TNNLS.2016.2538282

22. S. Chen, F. Zhou, Q. Liao, Visual domain adaptation using weighted subspace alignment, in *2016 Visual Communications and Image Processing (VCIP)*, (2016), 1–4. https://doi.org/10.1109/VCIP.2016.7805516

23. L. Zhang, S. Wang, G. Huang, W. Zuo, J. Yang, D. Zhang, Manifold criterion guided transfer learning via intermediate domain generation, *IEEE Trans. Neural Networks Learn. Syst.*, **30** (2019), 3759–3773. https://doi.org/10.1109/TNNLS.2019.2899037

24. Y. Grandvalet, Y. Bengio, Semi-supervised learning by entropy minimization, *Neural Inf. Process. Syst.*, (2004), 529–536.

25. S. Ahmed, D. Raychaudhuri, S. Paul, S. Oymak, A. RoyChowdhury, Unsupervised multi-source domain adaptation without access to source data, in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, (2021), 10103–10112. https://doi.org/10.1109/CVPR46437.2021.00997

26. Q. Tian, Y. Zhu, H. Sun, S. Chen, H. Yin, Unsupervised domain adaptation through dynamically aligning both the feature and label spaces, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 8562–8573. https://doi.org/10.1109/TCSVT.2022.3192135

27. S. Roy, M. Trapp, A. Pilzer, J. Kannala, N. Sebe, E. Ricci, et al., Uncertainty-guided source-free domain adaptation, in *European Conference on Computer Vision*, (2022), 537–555. https://doi.org/10.1007/978-3-031-19806-9_31

28. H. Mao, L. Du, Y. Zheng, Q. Fu, Z. Li, X. Chen, et al., Source free graph unsupervised domain adaptation, in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, (2024), 520–528. https://doi.org/10.1145/3616855.3635802

29. X. Wu, L. Cheng, S. Zhang, Open set domain adaptation with entropy minimization, in *Pattern Recognition and Computer Vision: Third Chinese Conference*, (2020), 29–41.

30. J. Kundu, N. Venkat, A. Revanur, R. Babu, Towards inheritable models for open-set domain adaptation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 12376–12385. https://doi.org/10.1109/CVPR42600.2020.01239

31. K. Saito, K. Saenko, Ovanet: One-vs-all network for universal domain adaptation, in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, (2021), 9000–9009. https://doi.org/10.1109/ICCV48922.2021.00887

32. Y. Lu, M. Shen, A. Ma, X. Xie, J. Lai, Mlnet: Mutual learning network with neighborhood invariance for universal domain adaptation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **38** (2024), 3900–3908. https://doi.org/10.1609/aaai.v38i4.28182

33. F. Qiao, L. Zhao, X. Peng, Learning to learn single domain generalization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 12556–12565. https://doi.org/10.1109/CVPR42600.2020.01257

34. K. Ricanek, T Tesafaye, Morph: A longitudinal image database of normal adult age-progression, in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, (2006), 341–345. https://doi.org/10.1109/FGR.2006.78

35. X. Liu, S. Li, Y. Ge, P. Ye, J. You, J. Lu, Recursively conditional gaussian for ordinal unsupervised domain adaptation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 764–773. https://doi.org/10.1109/ICCV48922.2021.00080

36. X. Liu, S. Li, Y. Ge, P. Ye, J. You, J. Lu, Ordinal unsupervised domain adaptation with recursively conditional gaussian imposed variational disentanglement, *IEEE Trans. Pattern Anal. Mach. Intell.*, (2022), 1–14. https://doi.org/10.1109/TPAMI.2022.3183115

37. Q. Tian, W. Zhang, M. Cao, L. Wang, S. Chen, H. Yin, Moment-guided discriminative manifold correlation learning on ordinal data, *ACM Trans. Intell. Syst. Technol. (TIST)*, **11** (2020), 1–18. https://doi.org/10.1145/3402445

38. Z. Kang, Y. Lu, Y. Su, C. Li, Z. Xu, Similarity learning via kernel preserving embedding, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 4057–4064. https://doi.org/10.1609/aaai.v33i01.33014057

39. C. Geng, S. Chen, Metric learning-guided least squares classifier learning, *IEEE Trans. Neural Networks Learn. Syst.*, **29** (2018), 6409–6414. https://doi.org/10.1109/TNNLS.2018.2830802

40. Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in *International Conference on Machine Learning*, (2015), 1180–1189. http://proceedings.mlr.press/v37/ganin15.html

41. Y. Yao, Y. Zhang, X. Li, Y. Ye, Discriminative distribution alignment: A unified framework for heterogeneous domain adaptation, *Pattern Recognit.*, **101** (2020), 107165. https://doi.org/10.1016/j.patcog.2019.107165

42. W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, S. Platz, Central moment discrepancy (cmd) for domain-invariant representation learning, preprint, arXiv:1702.08811.

43. B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **30** (2016), 2058–2065. https://doi.org/10.1609/aaai.v30i1.10306

44. M Long, Y Cao, J Wang, M Jordan, Learning transferable features with deep adaptation networks, in *International Conference on Machine Learning*, **37** (2015), 97–105.

45. C. Chen, Z. Chen, B. Jiang, X. Jin, Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 3296–3303. https://doi.org/10.1609/aaai.v33i01.33013296

46. I. Goodfellow, J. Abadie, M. Mirza, B. Xu, D. Farley, S. Ozair, et al., Generative adversarial nets, *Adv. Neural Inf. Process. Syst.*, (2014), 2672–2680.

47. E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 7167–7176. https://doi.org/10.1109/CVPR.2017.316

48. H. Tang, K. Jia, Discriminative adversarial domain adaptation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 5940–5947. https://doi.org/10.1609/aaai.v34i04.6054

49. K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 3723–3732. https://doi.org/10.1109/CVPR.2018.00392

50. L. Zhou, M. Ye, X. Zhu, S. Li, Y. Liu, Class discriminative adversarial learning for unsupervised domain adaptation, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), https://doi.org/10.1145/3503161.3548143

51. K. Saito, D. Kim, S. Sclaroff, K. Saenko, Universal domain adaptation through self supervision, *Adv. Neural Inf. Process. Syst.*, **33** (2020), 16282–16292.

52. J. Wang, Y. Chen, H. Yu, M. Huang, Q. Yang, Easy transfer learning by exploiting intra-domain structures, in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, (2019), 1210–1215. https://doi.org/10.1109/ICME.2019.00211

53. Q. Wang, T. Breckon, Unsupervised domain adaptation via structured prediction based selective pseudo-labeling, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 6243–6250. https://doi.org/10.1609/aaai.v34i04.6091

54. L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, C. Chen, Guide subspace learning for unsupervised domain adaptation, *IEEE Trans. Neural Networks Learn. Syst.*, **31** (2020), 3374–3388. https://doi.org/10.1109/TNNLS.2019.2944455

55. K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 9729–9738. https://doi.org/10.1109/CVPR42600.2020.00975

56. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in *International Conference on Machine Learning*, (2020), 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

57. R. Wang, Z. Wu, Z. Weng, J. Chen, G. Qi, Y. Jiang, Cross-domain contrastive learning for unsupervised domain adaptation, *IEEE Trans. Multim.*, **25** (2023), 1665–1673. https://doi.org/10.1109/TMM.2022.3146744

58. W. Ma, J. Zhang, S. Li, C. Liu, Y. Wang, W. Li, Making the best of both worlds: A domain-oriented transformer for unsupervised domain adaptation, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 5620–5629. https://doi.org/10.1145/3503161.3548229

59. Y. Zhang, Z. Wang, J. Li, J. Zhuang, Z. Lin, Towards effective instance discrimination contrastive loss for unsupervised domain adaptation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), 11388–11399. https://doi.org/10.1109/ICCV51070.2023.01046

60. H. Liu, M. Shao, Y. Fu, Structure-preserved multi-source domain adaptation, in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, (2016), 1059–1064. https://doi.org/10.1109/ICDM.2016.0136

61. H. Liu, M. Shao, Z. Ding, Y. Fu, Structure-preserved unsupervised domain adaptation, *IEEE Trans. Knowl. Data Eng.*, **31** (2018), 799–812. https://doi.org/10.1109/TKDE.2018.2843342

62. M. Meng, Q. Chen, J. Wu, Structure preservation adversarial network for visual domain adaptation, *Inf. Sci.*, **579** (2021), 266–280. https://doi.org/10.1016/j.ins.2021.07.085

63. Q. Tian, H. Sun, C. Ma, M. Cao, Y. Chu, S. Chen, Heterogeneous domain adaptation with structure and classiffcation space alignment, *IEEE Trans. Cybern.*, **52** (2022), 10328–10338. https://doi.org/10.1109/TCYB.2021.3070545

64. J. Jiang, Y. Ji, X. Wang, Y. Liu, J. Wang, M. Long, Regressive domain adaptation for unsupervised keypoint detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 6780–6789. https://doi.org/10.1109/CVPR46437.2021.00671

65. C. Seah, I. Tsang, Y. Ong, Transfer ordinal label learning, *IEEE Trans. Neural Networks Learn. Syst.*, **24** (2013), 1863–1876. https://doi.org/10.1109/TNNLS.2013.2268541

66. X. Chen, S. Wang, J. Wang, M. Long, Representation subspace distance for domain adaptation regression, in *International Conference on Machine Learning*, (2021), 1749–1759. http://proceedings.mlr.press/v139/chen21u.html

67. W. Wu, J. He, S. Wang, K. Guan, E. Ainsworth, Distribution-informed neural networks for domain adaptation regression, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 10040–10054.

68. I. Nejjar, Q. Wang, O. Fink, Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 11744–11754. https://doi.org/10.1109/CVPR52729.2023.01130

69. H. Wang, H. He, D. Katabi, Continuously indexed domain adaptation, preprint, arXiv:2007.01807

70. X. Zhong, L. Xu, Y. Li, Z. Liu, E. Chen, A nonconvex relaxation approach for rank minimization problems, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **29** (2015), 266–280. https://doi.org/10.1609/aaai.v29i1.9482

71. Q. Tian, M. Cao, S. Chen, H. Yin, Structure-exploiting discriminative ordinal multioutput regression, *IEEE Trans. Neural Networks Learn. Syst.*, **32** (2020), 266–280. https://doi.org/10.1109/TNNLS.2020.2978508

72. P. Zadeh, R. Hosseini, S. Sra. Geometric mean metric learning, in *International Conference on Machine Learning*, (2016), 2464–2471. http://proceedings.mlr.press/v48/zadeh16.html

73. X. He, P. Niyogi, Locality preserving projections, *Adv. Neural Inf. Process. Syst.*, (2003), 153–160.

74. Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 4920–4928. https://doi.org/10.1109/CVPR.2016.532

75. S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, S. Zafeiriou Agedb: the first manually collected, in-the-wild age database, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2017), 51–59. https://doi.org/10.1109/CVPRW.2017.250

76. B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation, in *European Conference on Computer Vision*, (2016), 443–450. https://doi.org/10.1007/978-3-319-49409-8_35

77. C. Yu, J. Wang, Y. Chen, M Huang, Transfer learning with dynamic adversarial adaptation network, in *2019 IEEE International Conference on Data Mining (ICDM)*, (2019), 778–786. https://doi.org/10.1109/ICDM.2019.00088

AIMS Press