Iterative joint classifier and domain adaptation for visual transfer learning

Shiva Noori Saray¹ · Jafar Tahmoresnezhad¹

Received: 20 November 2020 / Accepted: 12 September 2021 / Published online: 29 September 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Current available supervised classifiers cannot generalize across various domains due to distribution mismatch among them. Domain adaptation and transfer learning algorithms are proposed to tackle domain shift problem that originates from different data collection conditions. In this paper, we propose a transfer learning framework called iterative joint classifier and domain adaptation for visual transfer learning (ICDAV), which utilizes the balanced maximum mean discrepancy to better transfer knowledge across domains. Also, for learning a robust classifier against domain shift, a set of graph manifold regularizer and modified joint probability maximum mean discrepancy are simultaneously exploited to capture the domain structures and adapt the distribution of projected samples during the model learning process. Variety of experiments on several public datasets indicates that our approach achieves remarkable performance on visual domain adaptation and transfer learning tasks.

Keywords Transfer learning · Domain adaptation · Visual classification · Manifold regularization

1 Introduction

In today's world where the information is derived from classified data, and given that in any specialized fields, the data (text, audio and video) classification is one of the most important issues. Artificial intelligence is one of the most important solutions to this issue. Data classification is widely used in many specialized fields including medicine, military, traffic guidance and weather forecasting [4, 32, 45, 48]. One example of medical application is the diagnosis of disease. Now that the world is facing the Covid-19 problem, the robust diagnostic models are needed to identify the symptoms of Covid-19 [13].

Supervised learning as a well-studied machine learning method needs massive amounts of labeled data to create the robust models [1]. Since the manual labeling of training data to learn a model is prohibitive and the input dataset may be bias, the learned model most probably will perform poorly on new data. For example, the image properties of people

can be vary in different countries, which causes the distribution shift across domains. Transfer learning is one of the promising solutions to solve the label famine and dataset bias problems [28, 39, 51].

Transfer learning benefits from abundant labeled samples of source domain to reduce the dataset bias in target data labeling. Auxiliary source domain often has different distribution from target domain that makes traditional machine learning algorithms useless, since they assume that the source and target data are under the same distributions. In such a circumstance, transfer learning methods are used to reduce the distribution divergence where they can be divided into following three categories: (a) instance-based methods [47], which reuse the samples from source domain according to re-weighting techniques, (b) feature-based methods [46, 54], which learn the subspace with shared features to represent the source and target data under common conditions or perform distribution alignment to minimize the marginal or conditional distribution divergences between domains, and (c) model-based methods [31], which transfer parameters of source model to improve the performance of target model.

One of the effective factors in domain discrepancy is the marginal and conditional distribution discrepancies. Joint marginal and conditional distribution adaptation is a useful method in transfer learning, which measures the distribution shift between domains by a metric such as maximum

---

✉ Jafar Tahmoresnezhad
j.tahmores@it.uut.ac.ir

Shiva Noori Saray
shivanoorisaray@it.uut.ac.ir

¹ Faculty of Information Technology and Computer Engineering, Urmia University of Technology, Urmia, Iran

mean discrepancy (MMD) [42]. In many joint distribution adaptation based approaches, the marginal and conditional distributions are often treated equally, which may not be optimal [9, 34, 43]. In the other words, if two domains are very dissimilar, it means that the marginal distributions have more discrepancy and need more notice to align; while if two domains are similar (i.e., the marginal distributions are close), it means that the conditional distributions needs more attention and should be given more weight. Hence, considering the distributions with equal importance will degrade the efficiency of methods. To this end, in the first step, we use the dynamic balanced distribution adaptation to specify the relative importance of marginal and conditional distributions by $\mathcal{A}$ -distance and learning a domain-invariant representation. Also, in the second step to have a domain-invariant transfer classifier, the joint probability distribution discrepancy is utilized. Most of the available joint adaptation methods compute the sum of marginal and conditional distribution discrepancies, while the joint probability distribution discrepancy uses the product of class-conditional probability and the class prior probability to compute the discrepancy across domains. In fact, the calculation of joint probability distribution discrepancy is simpler and more accurate. Also, joint probability distribution discrepancy increases both the consistency of the same class samples from various domains (i.e., inter-domain transferability maximization) and the discrepancy of various classes (i.e., inter-class discriminability maximization) to facilitate the classification.

In addition, for obtaining a robust transfer classifier with superior performance, the principle of structural risk minimization (SRM) [22] and a group of graph manifold regularizer are exploited. SRM consists of loss function and regularizer term to minimize the loss on predicted labels and preserve the structure of domains in transfer learning process.

In summary, our proposed method brings the following contributions.

(1) We introduce a modified visual domain adaptation (MVDA) to learn a new representation to avoid feature distortion in original space. MVDA, in addition exploits the domain-invariant clusters to discriminate between different classes, and considers the relative importance of marginal and conditional distributions, quantitatively, for domain adaptation.

(2) We use two forms of MMD criteria to benefit the advantages of it in learning the new subspace and improving the classifier efficiency. We utilize the dynamic balanced MMD form in a novel representation learning step via proposed MVDA to perform an optimal distribution adaptation and obtain a new subspace. Dynamic balanced MMD evaluates the importance of marginal and conditional distributions to perform the optimal domain adaptation. Also, we benefit the joint probability MMD to boost the transferability and

discriminability of classifier by reducing the discrepancy of the same class samples of various domains and increasing the discrepancy between different classes' samples of various domains.

(3) We introduce the projected joint probability distribution discrepancy, which consists of embedded classifier to increase the between-class discriminability and between-domain transferability by participating the prediction function to improve the transfer learning performance.

(4) For increasing the compatibility between the intrinsic manifold structure of data and classifier structure, two graph-based regularizers (i.e., Hessian and Laplacian graphs) are utilized by embedded classifier.

(5) We conduct extensive experiments on real-world benchmark datasets to demonstrate the advantage of our proposed ICDAV over state-of-the-art traditional transfer learning and domain adaptation methods. The results illustrate a noteworthy improvement of ICDAV in term of average classification accuracy compared to the other transfer learning methods on most datasets.

The remainder of this paper is organized as follows. We review the related work in Sect. 2. In Sect. 3, we describe our proposed ICDAV method and its components. Extensive experiments are shown in Sect. 4, where it consists of dataset definition and performance results. Moreover, we extensively analyze the performance of ICDAV in Sects. 5 and 6. At the end, Section 7 concludes the paper.

2 Related work

The scope of transfer learning is very wide, including homogeneous and heterogeneous transfer learning [19], visual domain adaptation [15, 36, 40] and cross-dataset recognition [50], which is applicable in various areas such as object recognition [8], text classification [16], speech recognition [12] and face recognition [49].

Transfer learning methods are divided into the following three general categories based on the type of transferred information between source and target data [29]: instance-based transfer learning, feature-based transfer learning and model-based transfer learning.

Instance-based transfer learning: In this perspective, some of the labeled source data that are more similar to target data achieve more weight to be used in training of a target classifier. Hasse et al. [11] introduced “Instance-weighted transfer learning of active appearance models” approach, which integrates the training data from an auxiliary source domain via predicting the sample-specific weights and selects similar but still informative samples that improves the degree of transfer. Chen et al. [5] introduced “Instance weighting framework for genetic programming (GP) based symbolic regression for transfer learning”,

which utilizes the differential evolution to achieve the optimal weights during the evolutionary process of GP. This framework helps GP to recognize and remove the less useful source domain instances and learn from more useful source domain instances.

Feature-based transfer learning: This consists of subspace and manifold learning methods that map data from source and target domains into a new subspace with a better representation. Also, the distribution alignment methods, which they reduce the marginal, conditional or both distribution discrepancies across domains falls into this perspective. Principal component analysis (PCA) [14] is one of the most widely used methods, which reduces the dimensionality of data and finds a new subspace, while preserving the variability as much as possible. Kute et al. [17] proposed “Cross domain association using transfer subspace learning” that uses Bregman divergence regularization to reduce the distribution divergence across training and test domains to map into a common subspace. Also, it utilizes Fisher linear discriminant analysis subspace learning algorithm [37] to obtain the projection matrix to discriminate samples within the individual domains. Li et al. [18] proposed “Manifold alignment and balanced distribution adaptation (MBDA)” approach, which uses the balance factor to weigh the marginal and conditional distribution adaptation. Also, MBDA uses the marginal distribution to preserve the neighboring structures of domains, and reduce the dimension as much as possible. Manifold transfer learning via discriminant regression analysis (MTL-DRA) [24] is also one of the newest feature-based methods. MTL-DRA utilizes within- and between-class graphs to transfer the local geometry structure information across the source and target domains. Moreover, it provides a transform matrix, which is robust and sparse to deal with the noise and the negative transfer learning.

Model-based transfer learning: In this perspective, the goal is to adapt the trained classifier or regression model via the source domain to use in target domain. A novel shape deviation modeling scheme is proposed in [6], which models the dimensional error of the product in a parameter-based transfer learning approach. In particular, the shape deviation is decomposed into the following two components: i) the shape-independent error, which is modeled by multiple linear regression models in a Cartesian coordinate system, and ii) the shape-specific error, which is separated from the shape-independent error by analyzing the error generating mechanism with a particularly designed experiment. Thus, for prediction of the shape independent error of any new shape and minimization of the need in creating more test parts, the earned knowledge is transferred.

In this paper, our proposed method is a combination of both feature- and model-based transfer learning perspectives. In this approach, the feature-based learning perspective considers the adaptation of source and target domains

using dynamic balanced distribution adaptation and dense clusters. Whereas, in the perspective of model-based transfer learning, the domain-invariant optimal classifier learning is desired in a new subspace. In the classifier learning phase, in addition to minimizing the loss function, the joint probability distribution discrepancy adaptation is used to reduce the difference between conditional and marginal distributions, jointly. Also, a set of graph-based regularizer is utilized in classifier learning process, which benefits from the manifold structures.

3 Proposed method

Let domain D is composed of two components, the feature space X and the marginal probability distribution $P(X)$. The source domain and target domain are defined as $D_s = \{X_s, Y_s\}$ and $D_t = \{X_t\}$, respectively, which X_s and X_t illustrate the data from source and target datasets with the same class space, and Y_s induces the source labels. The data in domain D_t is unlabeled. The number of samples in D_s and D_t are n_s and n_t , in turn. Also, each domain has another distribution called conditional distribution, i.e., $P(Y|X)$, which is equivalent to the classifier $f(X)$. In fact, the conditional distribution indicates the probability of occurring label $y \in Y$ for input $x \in X$.

Accordingly, ICDAV first learns the transferable features to represent data with the best properties. Then, ICDAV performs the adaptive classification to classify target data. These two steps are iteratively optimized via several iterations. Therefore, we consider $P(X_s) \neq P(X_t)$ and $P(Y_s|X_s) \neq P(Y_t|X_t)$, separately, to distribution adaptation. After that, we assume $P(X_s, Y_s) \neq P(X_t, Y_t)$, directly, to find the better domain-invariant classifier.

3.1 Modified visual domain adaptation

In this step, we seek to introduce the modified visual domain adaptation to eliminate domain shifts. For this, we first briefly define VDA, which is the basis of MVDA.

The goal of VDA is to obtain an adaptation matrix $A \in R^{m \times k}$, which projects the source and target domains from m -dimensional feature subspace into a k -dimensional invariant feature subspace where the sample properties in the new representation is preserved. For finding the projection matrix A , Eq. 1 is utilized:

$$D = \underset{A}{\operatorname{argmin}} \operatorname{tr}(A^T X W_0 X^T A) + \operatorname{tr}(A^T X W_c X^T A) + \operatorname{tr}(A^T S_w A) \quad (1)$$

where the first and second components introduce the marginal and class-conditional distribution distances across

domains, respectively, where it uses the maximum mean discrepancy as a distance metric.

The third component $S_w = \sum_{c \in C} \sum_{x_i \in X_s^c} (x_i - \mu_c)^T (x_i - \mu_c)$ also shows the distance of each sample from its mean (μ_c) in class c . Moreover, $A, X = [X_s, X_t]$ and $tr(\cdot)$ denote the adaptation matrix, the set of source and target data and the trace of matrix, in turn. Also, $W_0 = \begin{bmatrix} (W_0)_{s,s} & (W_0)_{s,t} \\ (W_0)_{t,s} & (W_0)_{t,t} \end{bmatrix} \in R^{(n_s+n_t) \times (n_s+n_t)}$ is MMD coefficient matrix where $(W_0)_{s,s}$, $(W_0)_{t,t}$ and $(W_0)_{s,t} = (W_0)_{t,s}$ are computed through $\frac{1}{n_s^2}$, $\frac{1}{n_t^2}$ and $-\frac{1}{n_s n_t}$, respectively. As the same way, matrix $W_c = \begin{bmatrix} (W_c)_{s,s} & (W_c)_{s,t} \\ (W_c)_{t,s} & (W_c)_{t,t} \end{bmatrix} \in R^{(n_s+n_t) \times (n_s+n_t)}$ are calculated by $(W_c)_{s,s} = \frac{1}{(n_s^c)^2}$, $(W_c)_{t,t} = \frac{1}{(n_t^c)^2}$ and $(W_c)_{s,t} = (W_c)_{t,s} = -\frac{1}{n_s^c n_t^c}$, respectively, where n_s^c and n_t^c represent in order the number of samples in source and target domains related to class c . In VDA, PCA is used to reconstruct data such that $A^T X H X^T A$ is maximized subject to $A^T A = I$ where $X H X^T$ induces the covariance matrix of input data. Also, $H = I - \frac{1}{n} 11^T$ is the centering matrix where I and 1 are the identity matrix and the ones vector, in turn. Thus, the optimization problem is achieved as follow:

$$\min_{A^T X H X^T A = I} \sum_{c=1}^C tr(A^T (X W_c X^T + S_w) A) + \lambda \|A\|_F^2 \quad (2)$$

where $\|\cdot\|_F$ and λ are the Frobenius norm and the regularization parameter, respectively. We propose the modified VDA in which the importance of aligning marginal (i.e., $P(X)$) and conditional (i.e., $P(Y|X)$) distributions in domain adaptation is considered differently due to domain shift scale. To this end, we refine Eq. 2 as follow:

$$\min_{A^T X H X^T A = I} \sum_{c=1}^C tr(A^T (X (\eta W_c + (1 - \eta) W_0) X^T + S_w) A) + \lambda \|A\|_F^2 \quad (3)$$

where $\eta \in [0, 1]$ is the adaptive factor that is estimated through Eq. 4:

$$\hat{\eta} = 1 - \frac{d_M}{d_M + \sum_{c=1}^C d_c} \quad (4)$$

where d_M and d_c are the marginal and conditional A -distances and are introduced as the errors of a linear classifier ($\epsilon(h)$) in discrimination of source and target domains as

$$d_A(D_s, D_t) = 2(1 - 2\epsilon(h)). \quad (5)$$

Accordingly, the Lagrange function for Eq. 3 can be written as

$$\mathcal{J} = tr \left(A^T \left(X \sum_{c=1}^C (\eta W_c + (1 - \eta) W_0) X^T + S_w + \lambda I \right) A \right) + tr((I - A^T X H X^T A) \phi) \quad (6)$$

where $\phi = \text{diag}(\phi_1, \dots, \phi_k) \in R^{k \times k}$ illustrates the Lagrange multipliers. By setting $\frac{\partial \mathcal{J}}{\partial A} = 0$, the generalized eigen-decomposition is obtained as

$$\left(X \sum_{c=1}^C (\eta W_c + (1 - \eta) W_0) X^T + S_w + \lambda I \right) A = X H X^T A \phi. \quad (7)$$

The adaptation matrix A is then formed by the k trailing eigen-vectors. At the end, the new representation of source and target data can be achieved as $A^T X_s$ and $A^T X_t$.

Note that due to the lack of label in target domain in the first iteration, the marginal adaptation only is participated in distribution adaptation. After that, both the marginal and conditional distributions are utilized together in distribution adaptation and the labels of target domain is refined in later iterations.

3.2 Adaptive classifier

The aim of this step is to train a transferable classifier f , which has the least error on target domain.

3.2.1 Structural risk minimization

For obtaining the prediction function f , we use the structural risk minimization principle [41], which is formulated as

$$f = \arg \min_{f \in \mathcal{H}_K} \mathcal{J}(f(z), y) + E(f) \quad (8)$$

where $\mathcal{J}(\cdot, \cdot)$ represents the loss function and $f(z)$ and y are the predicted label and true label of sample z , in turn. $\mathcal{H}_K$ is the reproduced kernel Hilbert space (RKHS) [3] induced by Mercer kernel [26] function $K(\cdot, \cdot)$. Also, $E(f)$ represents the regularization term, which can be composed of several components. We describe the regularization $E(f)$ as follows:

$$E(f) = \gamma d_f(D_s, D_t) + \omega \hat{S}_{Hess_f}(D_s, D_t) + \rho \Psi_f(D_s, D_t) \quad (9)$$

where the definition of different components of Eq. 9 are explained in the rest. The first term $d_f(D_s, D_t)$ induces the modified joint probability MMD that performs the joint distribution alignment, which uses the maximum mean discrepancy measurement to compute the joint distribution shift across domains. This term performs the discriminability concurrently with transferability. Also, $\hat{S}_{Hess_f}(D_s, D_t)$ and $\Psi_f(D_s, D_t)$ are graph manifold regularizer based on Hessian and Laplacian regularizations, respectively. Moreover, γ, ω ,

ρ are the regularization coefficients. Thus, the embedding of Eq. 9 in Eq. 8 results the Eq. 10 as follows:

$$f = \operatorname{argmin}_{f \in \mathcal{H}_K} \mathcal{J}(f(z), y) + \gamma d_f(D_s, D_t) + \omega \hat{S}_{Hess_f}(D_s, D_t) + \rho \Psi_f(D_s, D_t). \quad (10)$$

Classifier f can be defined as Eq. 11 due to the representer theorem [2]:

$$f(z) = \sum_{i=1}^{n_s+n_t} \theta_i K(z_i, z) \quad (11)$$

where $\Theta = (\theta_1, \theta_2, \dots)^T \in \mathbb{R}^{(n_s+n_t) \times 1}$ denotes the coefficients vector.

3.2.2 Loss function

For minimizing the error of classifier $\mathcal{J}(f(z), y)$, the square loss function $\sum_{i=1}^n (y_i - f(z_i))^2$ is adopted where z is the projected samples of source and target domains. By considering Eq. 11, the first term of Eq. 10 becomes

$$\begin{aligned} \mathcal{J}(f(z), y) &= \sum_{i=1}^n (y_i - f(z_i))^2 + \zeta \|f\|_K^2 \\ &= \sum_{i=1}^{n_s+n_t} \mathcal{V}_{ii} (y_i - f(z_i))^2 + \zeta \|f\|_K^2 = \\ &\quad \left\| (Y - \Theta^T K) \mathcal{V} \right\|_F^2 + \zeta \operatorname{tr}(\Theta^T K \Theta) \end{aligned} \quad (12)$$

where $\|f\|_K^2$ is the squared norm of classifier f in RKHS while $\|\cdot\|_F$ indicates the Frobenious norm. Coefficient ζ denotes the regularization parameter. Also, $K \in \mathbb{R}^{(n_s+n_t) \times (n_s+n_t)}$ induces the kernel matrix whose entries are defined as $K_{ij} = K(z_i, z_j)$, and $\mathcal{V}$ is a diagonal domain indicator that $\mathcal{V}_{ii}$ has value 1 if $i \in D_s$ and value 0 if $i \in D_t$. Moreover, $Y = [y_1, \dots, y_{n_s+n_t}]$ denotes the source and target label matrix. In the rest, we introduce the components of regularization term in details.

3.2.3 Modified joint probability MMD adaptation

In this step, instead of considering $P(Z_s) \neq P(Z_t)$ (i.e., the marginal distribution discrepancy) and $P(Y_s|Z_s) \neq P(Y_t|Z_t)$ (i.e., the conditional distribution discrepancy), separately, $P(Z_s, Y_s) \neq P(Z_t, Y_t)$ (i.e., joint marginal and conditional distribution discrepancy) is jointly considered to distribution adaptation.

Joint probability MMD [52] is defined as Eq. 13, which calculates the joint probability discrepancy directly, and increases both the class discriminability and the domain transferability, simultaneously, as follows:

$$\begin{aligned} JPMMD &= \left\| A^T K_s N_s - A^T K_t N_t \right\|_F^2 \\ &\quad - \mu \left\| A^T K_s M_s - A^T K_t M_t \right\|_F^2 = \\ &\quad \operatorname{tr}(A^T K (R_{min} - \mu R_{max}) K^T A) \end{aligned} \quad (13)$$

where μ is a trade-off coefficient, while $K_s = \Phi(Z)^T \Phi(Z_s)$, $K_t = \Phi(Z)^T \Phi(Z_t)$ and $K = [K_s, K_t]$ are kernel matrices of projected samples of source, target and total matrices, respectively. Also, $R_{min} \in \mathbb{R}^{n \times n}$ and $R_{max} \in \mathbb{R}^{n \times n}$ are defined

$$\begin{aligned} R_{min} &= \begin{bmatrix} N_s N_s^T & -N_s N_t^T \\ -N_t N_s^T & N_t N_t^T \end{bmatrix} \quad \text{a n d} \\ R_{max} &= \begin{bmatrix} M_s M_s^T & -M_s M_t^T \\ -M_t M_s^T & M_t M_t^T \end{bmatrix} \quad \text{where } N_s = \frac{Y_s}{n_s}, N_t = \frac{\hat{Y}_t}{n_t}, \\ M_s &= \frac{F_s}{n_s} \text{ and } M_t = \frac{\hat{F}_t}{n_t}. \end{aligned}$$

such that $Y_s = [y_{s,1}, \dots, y_{s,n_s}]$ and $\hat{Y}_t = [\hat{y}_{t,1}, \dots, \hat{y}_{t,n_t}]$ are one-hot coding label matrices of source and target domains. Also, $F_s = [Y_s(:, 1) * (C-1), Y_s(:, 2) * (C-2), \dots, Y_s(:, C-1)]$ and $\hat{F}_t = [\hat{Y}_t(:, 2 : C), \hat{Y}_t(:, 3 : C), \dots, \hat{Y}_t(:, C)]$ are defined to calculate R_{max} where $Y_s(:, c)$ and $(C-1)$ demonstrate the c^{th} column of Y_s and repeat time of Y_s , respectively. In the same way, $\hat{Y}_t(:, 2 : C) \in \mathbb{R}^{n_t \times (C-1)}$ illustrates the 2^{nd} to C^{th} columns of matrix $\hat{Y}_t$.

Therefore, the modified joint probability MMD $d_f(D_s, D_t)$ is achieved in Eq. 14 by embedded classifier f (using Eq. 11) as the following feature map function:

$$d_f(D_s, D_t) = \operatorname{tr}(\Theta^T K (R_{min} - \mu R_{max}) K^T \Theta). \quad (14)$$

3.2.4 Set of graph manifold regularizer

The fourth $\hat{S}_{Hess_f}(D_s, D_t)$ and fifth $\Psi_f(D_s, D_t)$ terms of Eq. 10 induce the second-order Hessian energy regularizer and the Laplacian regularizer, respectively. The set of both graph regularizer based on the Hessian and Laplacian graphs are utilized to preserve the manifold structure for global label consistency.

The second-order Hessian energy regularizer performs extrapolation better than the Laplacian regularizer if the number of labeled data is low. The definition of Hessian energy is as

$$\begin{aligned} \hat{S}_{Hess_f}(D_s, D_t) &= \sum_{i=1}^n \left\| \nabla_a \nabla_b f \right\|_{Z_i}^2 = \sum_{i=1}^n \sum_{r,s=1}^d \left(\frac{\partial^2 f}{\partial x_r \partial x_s} \Big|_{Z_i} \right)^2 \\ &= \sum_{i=1}^n \sum_{\alpha \in N_p(z_i)} \sum_{\beta \in N_p(z_i)} f_{\alpha\beta} \Omega_{\alpha\beta}^{(i)} = \operatorname{tr}(\Theta^T K \Omega K \Theta) \end{aligned} \quad (15)$$

where $\left\| \nabla_a \nabla_b f \right\|_{Z_i}^2$ is the squared norm of second covariant derivative of f in data point Z_i , which corresponds to the Frobenious norm of Hessian of f at normal coordinates. Parameters $n = n_s + n_t$ denotes the set of source and target

data and d illustrates the PCA eigenvectors of p nearest neighbors space $N_p(z_i)$, which forms the orthogonal basis of local tangent space $T_{z_i}M$. Also, x_r and x_s are variables of function f . Moreover, α and $f_\alpha = f(z_\alpha)$ induce the neighboring point z_i and function value at point z_α , in turn. Moreover, the symbol Ω denotes the Hessian energy matrix, which is the sum up of all $\Omega^{(i)}$ matrices. The definition of $\Omega_{\alpha\beta}^{(i)}$ is given as Eq. 16 where H is an operator to estimate the Hessian of f at Z_i . Due to each sample is only associated with its neighbors, the Hessian energy matrix is sparse as follows:

$$\Omega_{\alpha\beta}^{(i)} = \sum_{r,s=1}^d H_{rs\alpha}^{(i)} H_{rs\beta}^{(i)}. \quad (16)$$

In fact, the Laplacian graph is utilized to preserve the label information in built manifold on training data, which is defined as

$$\Psi_f(D_s, D_t) = \sum_i \sum_j \mathcal{A}_{ij} \|f(z_i) - f(z_j)\|_2^2 = \text{tr}(\Theta^T K \mathcal{L} K \Theta) \quad (17)$$

where $\mathcal{L} = \mathcal{D} - \mathcal{A}$ represents the Laplacian graph matrix, in which $\mathcal{D}$ shows a diagonal matrix that is calculated through $\mathcal{D}_{ii} = \sum_j \mathcal{A}_{ij}$. Also, $\mathcal{A}$ illustrates the affinity matrix whose entries are described as

$$\mathcal{A}_{ij} = \begin{cases} 1, & \text{if } z_i \in \mathcal{N}_p(z_j) \text{ or } z_j \in \mathcal{N}_p(z_i) \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

where $\mathcal{N}_p(z_j)$ is the local set consisting of the p -nearest neighbors of z_j .

3.2.5 Overall reformulation

Thus, the combination of Eqs. (12), (14), (15) and (17) results the Eq. (19) as follow:

$$f = \underset{f \in \mathcal{H}_K}{\text{argmin}} \left\| (Y - \Theta^T K) \mathcal{V} \right\|_F^2 + \zeta \text{tr}(\Theta^T K \Theta) + \gamma \text{tr}(\Theta^T K (R_{\min} - \mu R_{\max}) K^T \Theta) + \omega \text{tr}(\Theta^T K \Omega K \Theta) + \rho \text{tr}(\Theta^T K \mathcal{L} K \Theta). \quad (19)$$

By setting the derivative $\frac{\partial f}{\partial \Theta} = 0$, the solution for Eq. (19) is obtained as follow:

$$\Theta = ((\mathcal{V} + \gamma(R_{\min} - \mu R_{\max}) + \omega \Omega + \rho \mathcal{L}) K + \zeta I)^{-1} \mathcal{V} Y^T. \quad (20)$$

In this way, we introduce a two-step framework for image classification that has considerable performance compared to other state-of-the-art transfer learning and domain adaptation methods. In the next section, to evaluate the performance of the proposed method, we provide the classification results on benchmark visual datasets.

4 Experiments

Here, we describe the used benchmark visual datasets in our experiments. We utilize following five datasets to evaluate the prediction accuracy of the proposed method compared to the other methods.

I. Office + Caltech-10 dataset [10, 35]

Office + Caltech-10 dataset contains 4 domains, in which 3 domains (i.e., Amazon (A), DSLR (D), Webcam (W)) are from the Office-31 and the another one (i.e., Caltech10 (C)) is from Caltech-256. In experiments, the samples number of Amazon, DSLR, Webcam and Caltech10 domains are 958, 157, 295 and 1123, respectively. The common 10 classes (i.e., projector, calculator, head-phones, keyboard, bike, mouse, monitor, backpack, mug, and laptop) among the Office-31 and Caltech-256 are exploited to perform experiments across 4 domains. In evaluation, 12 tasks for 4 domains are considered as follows: A→D, A→C, A→W, D→A, D→C, D→W, C→A, C→D, C→W, W→A, W→C and W→D.

II. Digits dataset [7]

MNIST (M) and USPS (U) are two datasets containing the images of 0 to 9 digits with different distributions. The MNIST consists of 60,000 training images and 10,000 test images. The USPS dataset consists of 7291 training images and 2007 test images. In experiments, 2000 and 1800 images are randomly sampled from MNIST and USPS, respectively. Also, each image are described with a 256-dimensional feature vector. In evaluation, two cross-domain tasks, USPS→MNIST and MNIST→USPS are examined.

III. Multi-PIE dataset [38]

Multi-PIE consists of 41,368 face images of 68 different identities with different poses, illuminations and expressions. In experiments, cross pose face recognition is performed on five different face orientations, including PIE05 (P1): left pose, PIE07 (P2): upward pose, PIE09 (P3): downward pose,

PIE27 (P4): front pose and PIE29 (P5): right pose. Totally, 3332, 1629, 1632, 3329, and 1632 facial images are contained in P1, P2, P3, P4 and P5, respectively. Therefore, 20 tasks were evaluated, like P1→P2, P1→P3, P1→P4, P1→P5.

IV. ImageNet-VOC 2007 [44]

ImageNet (I) and VOC2007 (V) are large-scale image classification datasets. ImageNet-VOC 2007 couple is much larger in diversity and much more accurate compared to the current image datasets. However, five shared classes (i.e., chair, cat, bird, person and dog) are selected from both datasets. To verify the effectiveness of proposed approach,

we use two tasks ImageNet-VOC 2007 (I-V) and VOC 2007-ImageNet (V-I), which each dataset is considered as one domain.

V. COIL20 [33]

The Columbia Object Image Library (COIL-20) dataset consists of 1440 gray scale images with 20 classes (objects such as duck) and 72 images per class (object). Different images of the same class are captured in different angles (i.e., at the interval of five degrees), which makes the images of the same class to have different distribution. We use two subsets C1 (all images taken in the directions of $[0^\circ, 85^\circ] \cup [180^\circ, 265^\circ]$) and C2 (all images taken in the directions of $[90^\circ, 175^\circ] \cup [270^\circ, 355^\circ]$) to construct two tasks C1-C2 and C2-C1.

4.1 Method evaluation

In this section, we present the performance of our ICDAV algorithm on five widely used object recognition datasets (Table 1). A comparison of the proposed method with other DA methods including TCA [30], JDA [23], CDDA [25], BDA (Wang et al. 2018), CLGA [20], GSL [52], JPDA [53], UCGS [21] and CIDA [27] is also provided.

All of the reported results are the classification accuracy on target domain, which is defined as follows:

$$accuracy = \frac{|x : x \in D_t \wedge \hat{y}(x) = y(x)|}{|x : x \in D_t|} \quad (21)$$

where $\hat{y}(x)$ and $y(x)$ are the predicted label of target dataset and the real label vector. In the rest, we compare our ICDAV with other state-of-the-art methods.

Transfer component analysis (TCA) seeks to obtain transfer components across domains and find a new latent subspace for marginal distribution adaptation by maximum mean discrepancy (MMD). While ICDAV uses the balanced marginal and conditional distribution adaptation to improve the transfer learning between the source and target domains. Our results show that ICDAV gets (10.94%), (19.94%), (31.07%), (12.28%) and (12.5%) significant classification accuracy improvement compared to TCA on Office-Caltech-10, Digits, PIE, ImageNet-VOC and COIL20 datasets, respectively.

Joint distribution adaptation (JDA) treats with both of the marginal and conditional distributions equally to adapt domains. Also, JDA measures the distribution shift across domains by maximum mean discrepancy. However, ICDAV adapts both distributions by considering different significance for each distribution to optimally adapt domains. As is clear from the results, ICDAV achieves (8.24%), (10.14%), (15.58%), (9.78%) and (10.76%) improvement in classification accuracy compared to JDA on Office-Caltech-10, Digits, PIE, ImageNet-VOC and COIL20 datasets, respectively.

Close yet discriminative domain adaptation (CDDA) obtains a latent feature representation by minimizing both marginal and conditional probability distribution discrepancies of source and target data via MMD and maximizing the distances between classes using repulsive force term. However, ICDAV increases the dependence between the labels and projected features alongside reducing the difference between the marginal and conditional distributions. ICDAV obtains (6.33%), (4.47%), (12.72%) and (6.94%) improvement against CDDA in prediction accuracy on Office-Caltech-10, Digits, PIE and COIL20 datasets, respectively.

Balanced domain adaptation (BDA) is an extended version of JDA, which allocates different weights to marginal and conditional distributions by several methods. BDA considers domain adaptation as the sum of the marginal and conditional distribution discrepancies, while ICDAV utilizes both the sum and joint probability forms of distribution discrepancies for domain alignment. ICDAV on Office-Caltech-10 datasets obtains (8.3%) improvement, and in Digits, PIE, ImageNet-VOC and COIL20 datasets obtains (9.96%), (15.9%), (7.04%) and (10.76%) performance improvement against BDA, respectively.

Coupled local-global adaptation (CLGA) is an unsupervised multi-source domain adaptation approach, which jointly reduces both the marginal and conditional distribution shifts to maximize the adaptation ability between source and target dataset (i.e., global level). Moreover, CLGA maximizes the discriminative ability (i.e., local level) to obtain a shared subspace by exploiting both class and domain manifold structures of data samples. However, ICDAV finds domain-invariant subspace for source and target domains and minimizes the distributional and geometrical discrepancies across domains. According to the results, ICDAV achieves (6.32%), (8.79%), (3.67%), (2.85%) and (5.69%) improvement in average accuracy compared to CLGA on Office-Caltech-10, Digits, PIE, ImageNet-VOC and COIL20 datasets, respectively.

Guide subspace learning for unsupervised domain adaptation (GSL) integrates three guidance losses including subspace guidance for distribution adaptation, data guidance for domain adaptation, and label guidance for label prediction to obtain target subspace. While our ICDAV benefits from joint MMD-based distribution adaptation and clustering to adapt domains. Also, unlike GSL which uses the basic classifiers for data classification, ICDAV proposes a novel classifier, which is more accurate and robust against noise. Moreover, ICDAV illustrates further improvement and achieves the state-of-the-art performance (1.65%), (33.81%), (15.31%), (9.72%) and (8.55%) on Office-Caltech-10, Digits, PIE, ImageNet-VOC and COIL20 datasets, respectively, against GSL.

Joint probability domain adaptation (JPDA) reduces the joint probability distribution disparity of the same class across domains for transferability, and increases the joint

Table 1 Classification accuracy (%) on Office + Caltech-10, Multi-PIE, MNIST + USPS, ImageNet-VOC and COIL20 datasets

Dataset	source	target	TCA (2011)	JDA (2013)	CDDA (2017)	BDA (2018)	CLGA (2018)	GSL (2019)	JPDA (2020)	UCGS (2021)	CIDA (2021)	ICDAV
Office-Caltech-10	C	A	38.20	44.78	48.33	44.57	48.02	58.7	48.54	–	41.76	59.19
		W	38.64	41.69	44.75	40.34	42.37	59.7	45.76	–	33.48	57.97
		D	41.4	45.22	48.41	45.22	49.04	49.7	45.86	–	33.04	52.23
	A	C	37.76	39.36	42.12	39.27	42.3	46	42.21	–	52.30	48.98
		W	37.63	37.97	41.69	37.97	41.36	44.1	42.03	–	37.68	51.86
		D	33.12	39.49	37.58	40.76	36.31	47.1	36.94	–	36.33	48.41
	W	C	29.30	31.17	31.97	31.43	32.95	37.9	35.17	–	47.8	31.61
		A	30.06	32.78	37.27	32.46	34.57	41.8	33.82	–	39.32	42.28
		D	87.26	89.17	87.9	89.17	92.36	88.5	89.17	–	80.68	93.63
	D	C	31.70	31.52	34.64	31.17	33.66	35.3	34.46	–	46.50	36.15
Multi-PIE		A	32.15	33.09	33.51	33.19	35.99	40.6	34.34	–	43.95	41.75
		W	86.10	89.49	90.51	89.49	89.83	85.8	91.19	–	72.61	90.51
	P1	P2	40.76	58.81	60.22	58.20	67.83	50.15	58.20	65.12	38.60	71.21
		P3	41.79	54.23	58.7	52.82	63.85	59.68	66.54	62.81	44.48	65.69
		P4	59.63	84.50	83.48	83.03	88.95	84.81	82.88	79.69	72.09	90.18
		P5	29.35	49.75	54.17	49.14	61.76	54.72	49.75	51.29	39.83	61.64
	P2	P1	41.81	57.62	62.33	57.35	71.4	48.02	63.36	62.61	37.63	73.74
		P3	51.47	62.93	64.64	62.75	72.98	42.16	60.48	63.91	40.7	74.75
		P4	64.73	75.82	79.9	75.76	86.24	73.36	77.53	80.84	65.56	91.44
		P5	33.70	39.89	44	39.71	51.23	37.50	47.79	55.27	34.62	60.72
Digits	P3	P1	34.69	50.96	58.46	51.35	70.17	55.34	59.03	60.02	47.55	75.93
		P2	47.70	57.95	59.73	56.41	73.48	53.10	61.51	62.49	44.91	70.60
		P4	56.23	68.46	77.2	67.86	89.31	73.27	74.8	78.13	72.79	91.98
		P5	33.15	39.95	47.24	42.40	55.51	55.82	51.16	58.27	50.67	61.03
	P4	P1	55.64	80.58	83.1	80.52	89.56	86.46	84.21	82.62	69.18	93.52
		P2	67.83	82.63	82.26	83.06	92.94	78.94	83.18	84.84	69.84	93.5
		P3	75.86	87.25	86.64	87.25	93.08	81.07	86.76	83.09	69.42	90.07
		P5	40.26	54.66	58.33	54.53	71.63	72.67	64.71	68.92	62.06	77.45
	P5	P1	26.98	46.46	48.02	47.99	57.68	45.62	53.39	58.07	36.09	64.98
		P2	29.90	42.05	45.61	43.22	55.43	39.47	49.85	54.92	34.01	61.76
ImageNet-Voc 2007	P3	P3	29.90	53.31	52.02	47.92	58.03	50.98	57.35	61.69	46.81	69.79
		P4	33.64	57.01	55.99	57.10	71.85	66.96	59.84	70.71	59.25	76.48
	U	M	51.05	59.65	62.05	59.90	58.35	28.36	59.35	–	–	66.10
	M	U	56.28	67.28	76.22	67.39	71.28	51.23	69.17	–	–	81.11
Coil20	I	V	63.7	63.4	–	65.02	68.16	61.26	62.71	68.72	–	70.41
	V	I	64.9	70.2	–	74.06	79.29	72.46	72.35	78.89	–	82.75
	C1	C2	88.47	89.31	91.53	89.44	96.81	92.9	92.08	–	–	99.72

Table 1 (continued)

Dataset	source	target	TCA (2011)	JDA (2013)	CDDA (2017)	BDA (2018)	CLGA (2018)	GSL (2019)	JPDA (2020)	UCGS (2021)	CIDA (2021)	ICDAV
	C2	C1	85.83	88.47	93.89	88.33	91.11	89.3	89.86	–	–	99.58
Avg. Office-Caltech-10			43.61	46.31	48.22	46.25	48.23	52.9	48.29	–	47.12	54.55
Avg. Multi-PIE			44.75	60.24	63.10	59.92	72.15	60.51	64.62	67.27	51.81	75.82
Avg. Digits			53.67	63.47	69.14	63.65	64.82	39.8	64.26	–	–	73.61
Avg. Image-Voc			64.3	66.8	–	69.54	73.73	66.86	67.53	73.81	–	76.58
Avg. Coil20			87.15	88.89	92.71	88.89	93.96	91.1	90.97	–	–	99.65
Overall Avg.			58.7	65.14	–	65.65	70.58	62.23	67.13	–	–	76.04

probability distribution disparity between different classes of different domains for discriminability, simultaneously. However, ICDAV uses the balanced and joint probability of distribution disparities, jointly for effective transfer across domains. Compared to JPDA, the average performance improvement of ICDAV is (6.26%), (9.35%), (11.2%), (9.05%) and (8.68%) on Office-Caltech-10, Digits, PIE, ImageNet-VOC and COIL20 datasets, respectively.

Unified cross-domain classification method via geometric and statistical adaptations (UCGS) unifies the structural risk minimization, frequently-used joint MMD adaptation, and the Nyström method for learning an adaptive classifier by considering the statistical and geometric adaptations. While ICDAV uses discriminative JPMMD, which is more accurate and enhances the discriminability and transferability at the same time. Also, ICDAV utilizes a set of graph manifold regularizer (i.e., Hessian and Laplacian) that explores better geometric structure across domains compared to the Nyström method. The performance improvement of ICDAV is (8.55%) and (2.77%) on PIE and ImageNet-VOC datasets, respectively, compared to the method UCGS.

Cross- and multiple-domains visual transfer learning via iterative Fisher linear discriminant analysis (CIDA) uses modified Fisher linear discriminant analysis to adapt domains. In the other words, CIDA maximizes the marginal distribution discrepancy of different classes of source and target domains and minimizes the difference between marginal distribution of the same class samples and the amount of variance between the various classes. Our ICDAV utilizes the clustering and balanced distribution adaptation for domain adaptation. Also, CIDA uses the empirical risk minimization and Laplacian matrix to construct a classifier, which does not consider the distribution adaptation. While ICDAV exploits the joint probability distribution adaptation and coupled Laplacian and Hessian energy regularizers to construct a robust adaptive classifier. ICDAV gains (7.43%) and (24.01%) on Office-Caltech-10 and PIE datasets, respectively, compared to CIDA in terms of the average classification accuracy.

5 Parameter evaluation

In this section, for performance evaluation of our proposed method in different conditions, we provide the parameter evaluation on tasks of following five benchmark datasets Office-Caltech-10, Multi-PIE, Digits, ImageNet-VOC 2007 and COIL20. Therefore, the classification accuracy changes according to the different values of parameters γ , k , λ , μ , ω , p , ρ and ζ on five selected tasks C-A (from Office-Caltech-10), P1-P2 (from Multi-PIE), U-M (from Digits), I-V (from ImageNet-VOC 2007) and C2-C1 (from COIL20) as shown in Fig. 1.

Figure 1. (a) shows the classification accuracy of parameter $\gamma \in [0.0 \ 1.0]$ on five tasks C-A, P1-P2, U-M, I-V and C2-C1. The parameter γ is the regularization parameter of modified joint probability MMD. As is clear from the plot, ICDAV has smooth flow performance in different values of parameter γ on two tasks (i.e., I-V and C2-C1), which it means that the joint probability MMD has no effect on I-V and C2-C1 tasks because there is not much distribution difference between domains, therefore, this component does not help in transferability and

discriminability between domain samples. As the same way, the prediction accuracy in two other tasks (i.e., P1-P2 and C-A) increases with a low slope, which means that the domain data in two tasks are very similar, and the component has a constructive effect on transferability and separability of the classifier. By increasing the value of γ , the performance of our method on U-M task increases at first and then decreases, which indicates a high distribution difference between domains. In the other words, when the difference of domain distributions is large, the joint

Fig. 1 The performance analysis of ICDAV on three tasks C-A (from Office-Caltech-10), P1-P2 (from Multi-PIE), U-M (from Digits), I-V (from ImageNet-VOC) and C2-C1 (from COIL20) with respect to the parameters γ , k , λ , μ , ω , p , ρ and ζ

probability MMD component can reasonably improve the transferability of the same class between different domains and discriminability between different classes of different domains, but increasing the impact of this component decreases the classifier performance.

Figure 1. (b) induces the experiments for evaluating the parameter $k \in \{5, 10, \dots, 60\}$. The parameter k illustrates the dimensionality of latent subspace. The plot indicates that in all five cases, increasing the value of k increases the ICDAV performance. In fact, ICDAV determines better performance in subspaces with high dimension because in this case the novel latent subspace is closer to the original subspace and has less information loss, so we will have a higher classification accuracy with high values of k .

Figure 1. (c) indicates the experiments of parameter λ on tasks of Office-Caltech-10, Multi-PIE, Digits, ImageNet-VOC 2007 and COIL20 datasets. The evaluation range of parameter λ is 0.0 to 1.0. The classification accuracy on task P1-P2 has negative slope after $\lambda = 0.1$ while on tasks U-M, I-V and C2-C1, it almost behaves smoothly. Also, the classification accuracy has partially uniform manner on C-A task. The descending manner on P1-P2 shows that ICDAV cannot learn a robust representation for cross-domain classification in Multi-PIE dataset at this interval. While, uniform slope of prediction accuracy on other tasks for the λ values more than 0.1 indicates that the robust representation is constructed in $\lambda = 0.1$ and more without considering the value of λ .

Figure 1. (d) illustrates the classification accuracy of ICDAV with respect to $\mu \in [0.0 \text{ } 1.0]$ where μ is the trade-off parameter between transferability and discriminability of different classes and different domains. ICDAV on tasks U-M and I-V has partially uniform trend, which induces the minor role of discriminability between different classes of different domains. In the other words, in these tasks, various classes' samples in various domains have enough distribution differences, and discriminability component cannot further increase the distribution difference. ICDAV behaves differently on Office-Caltech-10 (C-A), Multi-PIE (P1-P2) and COIL20 (C2-C1) datasets and has a completely downward trend. The downward trend on both tasks C-A and P1-P2 induces that the cross-domain transferability component contribute most to the discrepancy and increasing the coefficient of cross-class discriminability component reduces the performance of ICDAV. However, there is a sudden decrease on task C2-C1, which it means that the distributions of the same class samples between different domains is similar and the transferability should be given more weight than the discriminability. Therefore, allocating more weight in the discriminability component decreases the efficiency of the method, suddenly.

Figure 1. (e) presents the results for parameter ω on five tasks (i.e., C-A, P1-P2, U-M, I-V and C2-C1) where ω is the regularization parameter of second-order Hessian

energy regularizer, which preserves the manifold structure in domain adaptation. ICDAV behaves unbalanced but ascending on both tasks P1-P2 and U-M while it has mild descending manner on tasks C-A, I-V and C2-C1. This behavior shows that the similarity of domain pairs structures in P1-P2 and U-M tasks is high, and with increasing the value of parameter ω , the effect of the Hessian component on the domain adaptation increases, and has a positive effect on the transfer of inter-domain knowledge. While in other three tasks (i.e., C-A, I-V and C2-C1), domain structures are not very similar and for high values of ω parameter, the effect of the mentioned component in domain adaptation is negative.

Figure 1. (f) shows the effect of varying values of p on ICDAV performance. Therefore, $p \in \{2, 4, \dots, 20\}$ induces p -nearest neighbors in the second-order Hessian energy regularizer and Laplacian regularizer. ICDAV has stable behavior during the value changes of parameter p on four tasks (i.e., C-A, U-M, I-V and C2-C1) while the inconstant trend is observed in the middle of the chart on P1-P2. In four tasks C-A, U-M, I-V and C2-C1, the best geometric structure of the domain is obtained in p with values much more than 60, so these tasks experience a constant slope at low values p . Whereas in task P1-P2, the best geometric structure of domains is obtained at several points, even at low values of p .

Figure 1. (g) demonstrates the performance of ICDAV relation to the change in values of $\rho \in [0.0 \text{ } 1.0]$. Parameter ρ indicates the importance of Laplacian regularizer in optimization problem. As seen in the plot, by increasing the value of the parameter ρ , the classification accuracy fluctuates slightly on tasks P1-P2 and U-M. Also, the performance is declining by increasing the parameter ρ on C-A, I-V and C2-C1 tasks. In fact, the Laplacian is the neighborhood graph of samples and the best Laplacian graph gives the best data structure when the similar data place in the same neighborhood graph as much as possible. The best Laplacian structure of domains on tasks C-A, I-V and C2-C1 is achieved in low values of ρ because the dissimilarity of domain samples in these tasks causes that the good graphs are not formed. Therefore, by increasing the ρ value, the accuracy of classification decreases. While on tasks P1-P2 and U-M, the optimal Laplacian structures of domains are obtained in different points.

Figure 1. (h) displays the performance sensitivity of ICDAV to parameter $\zeta \in [0.0 \text{ } 1.0]$, which is the regularization parameter for squared norm of classifier f in RKHS. In fact, ζ supervises the complexity of the prediction function. Therefore, the accuracy has stable manner in four tasks C-A, U-M, I-V and C2-C1 for values more than 0.1 while it is a bit non-stable on task P1-P2. In fact, 0.1 is a critical value for parameter ζ on all tasks. When $\zeta \rightarrow 0$, the classifier performance decreases and the classifier overfits. While, by $\zeta \rightarrow \infty$,

the classifier is not fitted on input samples and underfitting occurs where the diagram displays the constant slope.

6 Ablation study

In this section, we evaluate the impact of each component on prediction accuracy of proposed method. Therefore, we provide Table 2, which shows the effect of each component by setting its coefficient to zero.

Parameter η is the balance parameter of the marginal and conditional distributions in Eq. 7. Our ICDAV by considering the η parameter, has a high performance in 9 out of 12 tasks in Office-Caltech dataset, 12 out of 20 tasks in PIE dataset, and 1 out of 2 tasks in Digits dataset. This result indicates that the balance parameter η improves the prediction accuracy in more than half of tasks. Also, the average accuracy is improved 0.1% on Office-Caltech dataset, 1.44% on PIE dataset and 0.15% on Digits dataset, by using the parameter η . While the parameter η has no effect on ImageNet-VOC and COIL20 datasets because the marginal and conditional distributions are with equal importance on these datasets.

The parameter λ is the regularizer parameter in Eq. 7. As can be seen from the results, the average accuracy of the label prediction is significantly improved by considering parameter λ . In the other words, the effect of this parameter on improving the prediction accuracy is 4.41% on Office-Caltech dataset, 40.21% on PIE dataset, 35.27% on Digits dataset, 2.64% on ImageNet-VOC dataset and 94.65% on COIL20 dataset.

Parameter γ is the coefficient of influence of joint probability distribution in Eq. 20. According to Table 2, the mentioned component in ICDAV results 1.57%, 1.07%, 0.98%, 4.31% and 0.21% classification accuracy improvement on the Office-Caltech, PIE, Digits, ImageNet-VOC and COIL20 datasets, respectively. Also, the ICDAV method has a superior performance in 26 out of 38 tasks by considering parameter γ .

Parameter μ is the trade-off parameter of joint probability distribution in Eq. 20. It establishes the balance between the R_{min} and R_{max} matrices. According to the results, parameter μ helps ICDAV to perform the same or better in 33 out of 38 tasks. In addition, the effect of parameter μ on prediction accuracy is 0.22% improvement on Office-Caltech dataset, 0.88% improvement on PIE dataset, 0.77% improvement on Digits dataset, 2.28% improvement on ImageNet-VOC dataset and 2.84% improvement on COIL20 dataset.

Parameter ω is the regularization coefficient of the second-order Hessian energy regularizer in Eq. 20. ICDAV obtains 3.58% prediction accuracy improvement and

achieves high efficiency in 30 out of 38 tasks by considering the second-order Hessian energy regularizer.

Parameter ρ is the regularization parameter of Laplacian regularizer in Eq. 20. Parameter ρ has the greatest impact on ImageNet-VOC dataset with 10.17% prediction accuracy improvement. Also, the effect of Laplacian regularizer is about 0.05%, 0.82% and 2.84% improvement on the Office-Caltech, PIE and COIL20 datasets, while in the Digits dataset is not improved. This indicates that the effect of parameter ρ depends on the type of dataset.

Parameter ζ is the adjustment parameter of squared norm of function f in Hilbert space, which has a significant effect on improving the classification prediction accuracy. The mentioned component ζ increases the efficiency of ICDAV in 36 out of 38 tasks and improves (29.52%) the average prediction accuracy of 38 tasks. In the other words, the presence of ζ in objective function results 23%, 9.98%, 25.18%, 60.5% and 28.96% improvement on Office-Caltech, PIE, Digits, ImageNet-VOC and COIL20 datasets, respectively.

As a result, ICDAV can have the best performance in domain adaptation by setting the optimal values of parameters. Also, the all parameters together have significant improvement on prediction accuracy of ICDAV.

7 Conclusion

As is from the results, our proposed ICDAV can cope with the distribution disparities across domains, with a successful approach. ICDAV minimizes data discrepancies through the modified visual domain adaptation method and obtains an adaptive classifier by maintaining the data structure and reducing the domain distribution discrepancies. To maintain the data structures, a set of graph manifold regularizer is exploited while the discrepancy reduction between domains is achieved using modified joint probability MMD. ICDAV uses two different forms to calculate MMD between data distribution, which has a tremendous impact on the quality of transfer learning across domains. Also our experiments induce the competitive results on all considered benchmark datasets. In future, we aim to perform some developments on proposed method to achieve more acceptable results in transfer learning. The developments can include multi-source experiments and embedded deep learning.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article

Table 2 Impacts of ICDAV's parameters on recognition accuracies on Office-Caltech-10, Multi-PIE, Digits, ImageNet-VOC and COIL20 data-sets

Dataset	Source	Target	Without η	Without λ	Without γ	Without μ	Without ω	Without ρ	Without ζ	ICDAV
Office-Caltech-10	C	A	59.08	54.91	56.05	59.19	59.92	59.60	9.60	59.19
		W	58.31	53.90	57.63	57.97	52.88	58.31	49.49	57.97
		D	52.23	51.59	50.32	52.23	45.86	53.50	45.22	52.23
	A	C	48.17	47.82	46.48	48.17	49.78	48.44	6.59	48.98
		W	50.17	46.10	50.17	50.51	46.78	52.54	12.54	51.86
		D	49.04	44.59	45.86	48.41	40.13	47.13	14.01	48.41
	W	C	32.77	28.50	34.28	31.43	32.77	31.26	26.80	31.61
		A	42.07	28.71	33.61	42.07	39.87	42.38	9.60	42.28
		D	93.63	88.54	92.99	93.63	88.54	94.90	94.90	93.63
	D	C	35.80	34.11	36.95	36.15	35.98	34.55	10.42	36.15
		A	41.65	33.82	42.90	41.75	43.95	40.92	9.60	41.75
		W	90.51	89.15	88.47	90.51	90.51	90.51	89.83	90.51
Multi-PIE	P1	P2	69.31	20.14	69.67	70.10	62.86	71.33	54.21	71.21
		P3	65.63	5.70	65.75	65.56	59.99	69.30	58.52	65.69
		P4	88.77	14.60	89.37	89.82	89.64	88.98	94.47	90.18
		P5	62.87	21.32	57.48	61.40	63.60	63.97	32.90	61.64
	P2	P1	70.44	32.35	78.36	70.02	77.16	65.52	65.40	73.74
		P3	61.46	32.05	59.25	73.47	62.81	57.60	67.95	74.75
		P4	91.50	1.47	91.47	91.50	90.78	91.50	90.06	91.44
		P5	62.87	37.38	62.68	62.75	50.55	62.87	46.94	60.72
	P3	P1	74.43	23.62	74.58	73.89	67.89	73.56	56.99	75.93
		P2	69.31	56.23	63.72	70.60	56.72	71.21	68.02	70.60
		P4	92.04	78.82	91.86	91.92	89.85	93.42	90.09	91.98
		P5	61.52	54.17	60.11	60.36	57.17	63.97	56.19	61.03
	P4	P1	92.65	1.47	92.44	92.02	91.27	93.19	90.19	93.52
		P2	93.74	82.44	91.96	93.68	91.22	94.05	91.10	93.49
		P3	90.38	85.60	90.63	90.44	90.32	89.95	89.09	90.07
		P5	77.21	66.61	79.17	78.31	76.23	79.90	65.50	77.45
	P5	P1	63.90	1.47	65.97	63.84	67.14	65.94	29.02	64.98
		P2	56.05	9.82	66.85	56.29	51.81	60.10	48.31	61.76
		P3	70.83	27.70	70.53	68.69	65.32	70.53	50.06	69.79
		P4	72.78	59.27	73.18	74.11	72.06	73.09	71.79	76.48
Digits	U	M	65.80	48.30	65.70	65.85	64.55	65.60	35.20	66.10
	M	U	81.11	28.39	79.56	79.83	73.44	81.61	61.67	81.11
ImageNet-voc 2007	I	V	70.41	68.39	67.65	68.99	68.93	62.50	13.63	70.41
	V	I	82.75	79.49	76.88	79.61	78.93	70.32	18.53	82.75
Coil20	C1	C2	99.72	5	99.72	96.11	93.75	96.11	66.53	99.72
	C2	C1	99.58	5	99.17	97.5	97.08	97.5	74.86	99.58
Avg. Office-Caltech-10			54.45	50.14	52.98	54.33	52.25	54.50	31.55	54.55
Avg. Multi-PIE			74.38	35.61	74.75	74.94	71.72	75	65.84	75.82
Avg. Digits			73.46	38.34	72.63	72.84	69	73.61	48.43	73.61
Avg. Image-Voc			76.58	73.94	72.27	74.30	73.93	66.41	16.08	76.58
Avg. Coil20			99.65	5	99.44	96.81	95.42	96.81	70.69	99.65
Overall Avg.			75.70	40.61	74.41	74.64	72.46	73.27	46.52	76.04

References

- Arnold A, Nallapati R, Cohen WW (2007) A comparative study of methods for transductive transfer learning. In: Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), IEEE, pp 77–82. <https://doi.org/10.1109/ICDMW.2007.109>
- Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J Mach Learn Res* 7(11):2399–2434
- Berlinet A, Thomas-Agnan C (2011) Reproducing kernel Hilbert spaces in probability and statistics. Springer, New York
- Cai L, Gu J, Ma J, Jin Z (2019) Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees. *Energies* 12(1):159. <https://doi.org/10.3390/en12010159>
- Chen Q, Xue B, Zhang M (2019) Differential evolution for instance based transfer learning in genetic programming for symbolic regression. In: Proceedings of the genetic and evolutionary computation conference companion, pp 161–162. <https://doi.org/10.1145/3319619.3321941>
- Cheng L, Tsung F, Wang A (2017) A statistical transfer learning perspective for modeling shape deviations in additive manufacturing. *IEEE Robot Autom Lett* 2(4):1988–1993. <https://doi.org/10.1109/LRA.2017.2713238>
- Fernando B, Tommasi T, Tuytelaars T (2015) Joint cross-domain classification and subspace learning for unsupervised adaptation. *Pattern Recogn Lett* 65:60–66. <https://doi.org/10.1016/j.patrec.2015.07.009>
- Ghifary M, Balduzzi D, Kleijn WB, Zhang M (2016) Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Trans Pattern Anal Mach Intell* 39(7):1414–1430
- Gholenji E, Tahmoresnezhad J (2020) Joint local and statistical discriminant learning via feature alignment. *SIViP* 14(3):609–616. <https://doi.org/10.1007/s11760-019-01587-1>
- Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2066–2073. <https://doi.org/10.1109/CVPR.2012.6247911>
- Haase D, Rodner E, Denzler J (2014) Instance-weighted transfer learning of active appearance models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1426–1433
- Hsu W-N, Harwath D, Glass J (2019) Transfer learning from audio-visual grounding to speech recognition. *arXiv:1907.04355*
- Jaiswal A, Gianchandani N, Singh D, Kumar V, Kaur M (2020) Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *J Biomol Struct Dyn*. <https://doi.org/10.1080/07391102.2020.1788642>
- Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. *Philos Trans R Soc A: Math Phys Eng Sci* 374(2065):20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Karimpour M, Saray SN, Tahmoresnezhad J, Aghababa MP (2020) Multi-source domain adaptation for image classification. *Mach Vis Appl* 31(6):1–19. <https://doi.org/10.1007/s00138-020-01093-2>
- Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: a survey. *Information* 10(4):150. <https://doi.org/10.3390/info10040150>
- Kute RS, Vyas V, Anuse A (2019) Cross domain association using transfer subspace learning. *Evol Intel* 12(2):201–209. <https://doi.org/10.1007/s12065-019-00211-y>
- Li Y, Cheng L, Peng Y, Wen Z, Ying S (2019) Manifold alignment and distribution adaptation for unsupervised domain adaptation. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 688–693. <https://doi.org/10.1109/ICME.2019.00124>
- Liu F, Zhang G, Lu J (2020) Heterogeneous domain adaptation: an unsupervised approach. *IEEE Trans Neural Netw Learn Syst* 31(12):5588–5602. <https://doi.org/10.1109/TNNLS.2020.2973293>
- Liu J, Li J, Lu K (2018) Coupled local-global adaptation for multi-source transfer learning. *Neurocomputing* 275:247–254. <https://doi.org/10.1016/j.neucom.2017.06.051>
- Liu W, Li J, Liu B, Guan W, Zhou Y, Xu C (2021) Unified cross-domain classification via geometric and statistical adaptations. *Pattern Recognit* 110:107658. <https://doi.org/10.1016/j.patcog.2020.107658>
- Long M, Wang J, Ding G, Pan SJ, Philip SY (2013a) Adaptation regularization: a general framework for transfer learning. *IEEE Trans Knowl Data Eng* 26(5):1076–1089. <https://doi.org/10.1109/TKDE.2013.111>
- Long M, Wang J, Ding G, Sun J, Yu PS (2013b) Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE international conference on computer vision, IEEE, pp 2200–2207
- Lu Y, Wang W, Yuan C, Li X, Lai Z (2020) Manifold transfer learning via discriminant regression analysis. *IEEE Trans Multimed* 23:2056–2070. <https://doi.org/10.1109/TMM.2020.3007340>
- Luo L, Wang X, Hu S, Wang C, Tang Y, Chen L (2017) Close yet distinctive domain adaptation. *arxiv:1704.04235*
- Lyu S (2005) Mercer kernels for object recognition with local features. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, pp 223–229. <https://doi.org/10.1109/CVPR.2005.223>
- Mardani M, Tahmoresnezhad J (2021) Cross-and multiple-domains visual transfer learning via iterative Fischer linear discriminant analysis. *Knowl Inf Syst* 1–32. <https://doi.org/10.1007/s10115-021-01586-0>
- Noori Saray S, Tahmoresnezhad J (2020) Kernelized domain adaptation and balanced distribution alignment for image classification. *J Soft Comput Inf Technol* 9(2):48–60
- Pan S, Yang Q (2009) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
- Pan SJ, Tsang IW, Kwok JT, Yang Q (2010) Domain adaptation via transfer component analysis. *IEEE Trans Neural Netw* 22(2):199–210. <https://doi.org/10.1109/TNN.2010.2091281>
- Perrone V, Jenatton R, Seeger MW, Archambeau C (2018) Scalable hyperparameter transfer learning. In: Proceedings of the 32nd international conference on neural information processing systems, pp 6846–6856
- Raghu M, Zhang C, Kleinberg J, Bengio S (2019) Transfusion: understanding transfer learning for medical imaging. *arXiv:1902.07208*
- Rate C, Retrieval C (2011) Columbia object image library (coil-20). Computer
- Rezaei S, Tahmoresnezhad J, Solouk V (2021) A transductive transfer learning approach for image classification. *Int J Mach Learn Cybern* 12(3):747–762. <https://doi.org/10.1007/s13042-020-01200-9>
- Saenko K, Kulis B, Fritz M, Darrell T (2010) Adapting visual category models to new domains. European conference on computer vision, Springer, Berlin 6314:213–226. <https://doi.org/10.1007/978-3-642-15561-1-16>
- Saray SN, Tahmoresnezhad J (2021) Joint distinct subspace learning and unsupervised transfer classification for visual domain

- adaptation. *SIViP* 15(2):279–287. <https://doi.org/10.1007/s11760-020-01745-w>
37. Si S, Tao D, Geng B (2009) Bregman divergence-based regularization for transfer subspace learning. *IEEE Trans Knowl Data Eng* 22(7):929–942. <https://doi.org/10.1109/TKDE.2009.126>
  38. Sim T, Baker S, Bsat M (2002) The CMU pose, illumination, and expression (PIE) database. In: *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, IEEE, pp 53–58. <https://doi.org/10.1109/AFGR.2002.1004130>
  39. Tahmoresnezhad J, Hashemi S (2017a) Exploiting kernel-based feature weighting and instance clustering to transfer knowledge across domains. *Turk J Electr Eng Comput Sci* 25(1):292–307
  40. Tahmoresnezhad J, Hashemi S (2017b) Visual domain adaptation via transfer feature learning. *Knowl Inf Syst* 50(2):585–605. <https://doi.org/10.1007/s10115-016-0944-x>
  41. Vapnik VN, Vapnik V (1998) *Statistical learning theory*, vol 1. Wiley, New York
  42. Wang J, Chen Y, Feng W, Yu H, Huang M, Yang Q (2020) Transfer learning with dynamic distribution adaptation. *ACM Trans Intell Syst Technol (TIST)* 11(1):1–25. <https://doi.org/10.1145/3360309>
  43. Wang J, Chen Y, Hao S, Feng W, Shen Z (2017) Balanced distribution adaptation for transfer learning. In: *2017 IEEE international conference on data mining (ICDM)*, pp 1129–1134. <https://doi.org/10.1109/ICDM.2017.150>
  44. Wang J, Feng W, Chen Y, Yu H, Huang M, Yu PS (2018a) Visual domain adaptation with manifold embedded distribution alignment. In: *Proceedings of the 26th ACM international conference on Multimedia*, pp 402–410. <https://doi.org/10.1145/3240508.3240512>
  45. Wang P, Li L, Jin Y, Wang G (2018b) Detection of unwanted traffic congestion based on existing surveillance system using in freeway via a CNN-architecture trafficnet. In: *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, pp 1134–1139. <https://doi.org/10.1109/ICIEA.2018.8397881>
  46. Wang P, Lu L, Li J, Gan W (2019) Transfer learning with joint distribution adaptation and maximum margin criterion. In: *Journal of Physics: Conference Series*, 1169(1). IOP Publishing, p 012028
  47. Xu F, Yu J, Xia R (2018) Instance-based domain adaptation via multiclustering logistic approximation. *IEEE Intell Syst* 33(1):78–88. <https://doi.org/10.1109/MIS.2018.012001555>
  48. Yang Z et al (2019) Deep transfer learning for military object recognition under small training set condition. *Neural Comput Appl* 31(10):6469–6478. <https://doi.org/10.1007/s00521-018-3468-3>
  49. Yin X, Yu X, Sohn K, Liu X, Chandraker M (2019) Feature transfer learning for face recognition with under-represented data. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 5704–5713
  50. Zhang J, Li W, Ogunbona P (2017) Transfer learning for cross-dataset recognition: a survey. [arXiv:1705.04396](https://arxiv.org/abs/1705.04396)
  51. Zandifar M, Noori Saray S, Tahmoresnezhad J (2021) Domain adaptation via Bregman divergence minimization. *Sci Iran*. <https://doi.org/10.24200/sci.2021.51486.2210>
  52. Zhang L, Fu J, Wang S, Zhang D, Dong Z, Chen CP (2019) Guide subspace learning for unsupervised domain adaptation. *IEEE Trans Neural Netw Learn Syst* 31(9):3374–3388. <https://doi.org/10.1109/TNNLS.2019.2944455>
  53. Zhang W, Wu D (2020) Discriminative Joint Probability Maximum Mean Discrepancy (DJP-MMD) for domain adaptation. In: *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp 1–8. <https://doi.org/10.1109/IJCNN48605.2020.9207365>
  54. Zhong X, Guo S, Shan H, Gao L, Xue D, Zhao N (2018) Feature-based transfer learning based on distribution similarity. *IEEE Access* 6:35551–35557. <https://doi.org/10.1109/ACCESS.2018.2843773>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.