frontiers in Neuroscience

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fnins.2021.677106

Multi-Source Co-adaptation for EEG-Based Emotion Recognition by Mining Correlation Information

Jianwen Tao†* and *Yufang Dan*†

Institute of Artificial Intelligence Application, Ningbo Polytechnic, Zhejiang, China

Since each individual subject may present completely different encephalogram (EEG) patterns with respect to other subjects, existing subject-independent emotion classifiers trained on data sampled from cross-subjects or cross-dataset generally fail to achieve sound accuracy. In this scenario, the domain adaptation technique could be employed to address this problem, which has recently got extensive attention due to its effectiveness on cross-distribution learning. Focusing on cross-subject or cross-dataset automated emotion recognition with EEG features, we propose in this article a robust multi-source co-adaptation framework by mining diverse correlation information (MACI) among domains and features with $l_{2,1}-norm$ as well as correlation metric regularization. Specifically, by minimizing the statistical and semantic distribution differences between source and target domains, multiple subject-invariant classifiers can be learned together in a joint framework, which can make MACI use relevant knowledge from multiple sources by exploiting the developed correlation metric function. Comprehensive experimental evidence on DEAP and SEED datasets verifies the better performance of MACI in EEG-based emotion recognition.

Keywords: electroencephalogram, emotion recognition, multi-source adaptation, feature selection, maximum mean discrepancy

OPEN ACCESS

Edited by:
Yuanpeng Zhang,
Nantong University, China

Reviewed by:
Hongjiang Wei,
Shanghai Jiao Tong University, China
Cai Zhenhua,
Wuhan University of Technology, China

***Correspondence:**
Jianwen Tao
2019025@nbpt.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *07 March 2021*
Accepted: *22 March 2021*
Published: *13 May 2021*

Citation:
Tao J and Dan Y (2021) Multi-Source Co-adaptation for EEG-Based Emotion Recognition by Mining Correlation Information. Front. Neurosci. 15:677106. doi: 10.3389/fnins.2021.677106

INTRODUCTION

Although emotion can be easily captured by human beings due to its close relationship with human's cognition (Dolan, 2002), it cannot be readily recognized by instruments due to its complexity. Recently, as one of the most active research topics from the affective computing community, affection recognition had obtained a large amount of attention from pattern recognition and machine vision research fields (Kim et al., 2013). Generally, there are two categories on the responses of human emotion, i.e., external and internal responses. In this work, we focus on the latter. Basically, the representative internal responses include blood pressure, heart rate, respiration rate, electroencephalography (EEG), magneto encephalogram (Mühl et al., 2014), etc. Usually, the core components of a traditional emotion recognition system based on EEG are feature extraction and emotion classification (Lan et al., 2018). Practically, the time domain, frequency domain, and time–frequency domain are the main sources of the extracted EEG features (Jenke et al., 2014; Zhang et al., 2020b). The EEG feature extraction methods are more

comprehensively reviewed in Jenke et al. (2014). In the past few years, aiming at the problem of emotion classification, a large number of emotion recognition models based on EEG signals have been proposed (Musha et al., 1997; Kim et al., 2013). For instance, a new group sparse canonical correlation analysis method was proposed in Zheng (2017) for simultaneous EEG channel selection and emotion recognition; in Li et al. (2018c), a graph regularized sparse linear regression method was proposed to deal with EEG-based emotion recognition. In the recent years, the deep learning method based on EEG has shown better performance than those traditional methods in emotion recognition and widely exploited in feature extraction and emotion recognition at the same time. For example, Zheng et al. (2015) employed deep belief network for EEG-based emotion recognition; Song et al. (2018) modeled the multi-channel EEG features by utilizing a graph and then performed EEG-based emotion recognition on these features; the work by Li et al. (2018b) had proposed a novel neural network model for EEG-based emotion recognition task.

While many models and methods for emotion recognition based on EEG have been proposed, most of them worked well only in the scenario that the training and test data were from the same distribution or domain. Under this hypothesis, the classifier trained on the source domain can directly predict the labels of the target data. However, for the problem of cross-domain emotion recognition based on EEG, many EEG-based emotion recognition methods would fail because of the distribution mismatch of EEG features. To this end, domain adaptation (DA) emotion recognition algorithms have emerged to investigate and address the automated emotion recognition problem (Chu et al., 2017), in which one has target domain with few or even none of labeled images by leveraging other related but different source/auxiliary domain(s) (Bruzzone and Marconcini, 2010). A typical example is the cross-subject EEG emotion recognition problem, in which the training and testing EEG data are from different subjects. To deal with the challenging cross-subject EEG emotion recognition problem, Pandey and Seeja (2019) proposed a subject-independent approach for EEG emotion recognition. Li et al. (2018a) proposed another method for cross-subject EEG emotion recognition. In the past decade, deep neural networks (Ganin et al., 2016; Li et al., 2018b) have also driven rapid progress in DA. The DA issues can be solved by the domain adversarial neural network (Ganin et al., 2016). It remains unclear, however, whether the performance of deep DA methods is really contributed by their deep feature representation, the fine-tuned classifiers, or is rather an outcome of the adaptation regularization terms (Ghifary et al., 2017).

Although the existing DA method has obvious effectiveness and efficiency in the special use of emotion recognition (Chu et al., 2017) in order to make use of the correlation knowledge among domains and features, there is little work to use the joint feature selection method and then carry out the multi-source adaptive domain recognition of cross-dataset. Besides this, during DA, most of the multi-source domain adaptation (MDA) methods (Yang et al., 2007; Tommasi et al., 2014) generally cope with the sources independently without considering the correlation information among the source domains, which may destroy the discriminant structure (either intrinsic or extrinsic) of multi-source domains. Last but not the least, for a MDA system, it is crucial for source weight determination during learning based on the correlation and quality of source domains. To the best of our knowledge, these characters are not feasible enough in extant MDA methods.

In order to solve the above-mentioned problems in existing MDA, we exploit the relevant information among sources and features to learn a multi-source emotion recognition model. We mainly adopt the strategy of digging the relationship between multi-source domains and one target domain (including feature and distribution) for promoting multi-source adaptive emotion recognition. We aim to progress beyond existing works that have partially addressed those issues by exploring to solve all the above-mentioned issues in a unified framework. Specifically, we develop a robust multi-source co-adaptation method for EEG-based emotion recognition by employing the correlation information (MACI) among features and sources *via* $l_{2,1}-norm$ (Nie et al., 2010a) and correlation metric regularization. Under this framework, the correlation metric function is developed to mine the invariant knowledge among multi-source domains, the $l_{2,1}-norm$ loss function aims to reduce the influence of outliers or noise, and row sparsity is designed to obtain the solution of sparse feature selection (Zhang et al., 2020b). We match distributions between each domain pair (including both target and multi-source domains) by minimizing the nonparametric maximum mean discrepancy (MMD) (Gretton et al., 2009; Pan et al., 2011) in a reproducing kernel Hilbert space (RKHS). The contributions of this article are listed as follows:

1. We propose a unified multi-source adaptive emotion recognition framework with EEG features by combining $l_{2,1}-norm$ and correlation metric regularization.
2. Our framework selects features in a collaborative way and considers the correlated knowledge among features; the importance of each feature does not need to be evaluated separately. In addition, in our unified framework, we can learn multiple functions of feature selection for all source adaptation subjects synchronously so that our framework can use the correlated information of multiple sources as auxiliary information.
3. In this framework, the original geometric structure is retained by using the graph Laplacian regularization, and the $l_{2,1}-norm$ minimization sparse regression approach is used to suppress the influence of noise or outliers in the domains, which shows the robustness of the framework.
4. Through a large number of experiments on two EEG datasets, we prove the effectiveness and convergence of this framework.

The remainder of the paper is organized as follows. In section "Related Work," we discussed the related works with feature selection and MDA learning. In section "Proposed Framework," our framework MACI will be designed, while section "Algorithm" arranges the corresponding optimal algorithm of MACI. Section "Algorithm Analysis" gives algorithm analysis, including the convergence and generalization.

The experimental results and analysis on two real EEG datasets are presented in section "Experiments." Finally, we conclude in section "Conclusion."

RELATED WORK

In this section, we briefly review the prior emotion recognition with respect to EEG and multi-source adaptation techniques due to their relationships with our main ideas.

Multi-Source Domain Adaptation

In the past several years, the mismatch problem between source and target domains has been solved by many DA technologies, which are widely used in a large number of visual applications, such as image annotation/classification, video concept detection, target recognition, and so on (Yang et al., 2007; Duan et al., 2012b,c; Tao et al., 2015, 2016, 2017; Ghifary et al., 2017). In the existing works for conquering DA, discovering one or multiple domain-invariant classifier(s) is a widely research topic (Yang et al., 2007; Tommasi et al., 2014) by constructing certain common subspace to make different sources have the same (or similar) marginal distributions. Therefore, the source classifier would obtain well performance on the target domain. Several methods have been studied to measure the distribution similarities (Gretton et al., 2009), and the analysis from them (Mansour et al., 2009) show that the performance of the classifier on the target is in positive correlation with those similar sources.

Very recently, to overcome the so-called negative transfer issue (Rosenstein et al., 2005), MDA methodology has been put forward by leveraging knowledge from multiple sources (Zhang et al., 2015, 2019c). A common problem in MDA is how to reduce the distribution difference between domains (Ghifary et al., 2017). To solve this issue, existing MDA approaches can be simply grouped into two classes (Tao et al., 2019), i.e., classifier-centric learning and feature-centric learning. The former is mainly based on the learning of the source domain classifiers in the target domain to adjust for realizing the implicit adaptation in the target distribution (e.g., Yang et al., 2007; Duan et al., 2012c; Tao et al., 2012; Tommasi et al., 2014), while the latter tries to accomplish the distribution alignment by learning a new representation of the data through a certain transformation (e.g., Wang and Mahadevan, 2011; Ghifary et al., 2017). This article focuses on the research of unsupervised classifier approaches.

In real application scenarios, the classifier-centric MDA scheme usually aims to directly design multiple adaptive source classifiers by merging the multiple distributions' adaptation *via* feature representation or classifiers with model regularization. Lately, visual recognition works (Mansour et al., 2009; Nie et al., 2010a; Pan et al., 2011; Duan et al., 2012c; Tao et al., 2015, 2017) have proposed a great deal of classifier-centric MDA approaches. One part of classifier-centric MDA research assumes that there are enough number of unlabeled target instances and a large amount of labeled source instances in the training stage. Nevertheless, the remaining part of classifier-centric MDA research holds another hypothesis that only some labeled target instances are accessible in the training stage, which is also called model adaptation in the literature (Yang et al., 2007; Duan et al., 2012b,c; Tao et al., 2015, 2016, 2017; Ghifary et al., 2017). The model adaptation works effectively and efficiently just by exploiting the existing source models pre-trained on relevant but different source domains. Several representative state-of-the-arts include adaptive support vector machines (A-SVM) (Yang et al., 2007) *via* leveraging multiple source classifiers to suit a major target classifier, DA machine (or FastDAM) (Duan et al., 2012c) by employing sparsity regularizations and Laplacian manifold in least squares SVMs (Chai et al., 2016), etc. Recently, we also proposed some different model adaptation strategies (Tao et al., 2015, 2016, 2017) by leveraging the advantages of the low-rank and sparse representation.

Emotion Recognition

In recent research about affective computing, increasing attentions have been paid on emotion recognition in the community of brain–computer interfaces (BCIs) (Mühl et al., 2014; Chu et al., 2017). An ideal emotion-based BCI can detect the emotional state through spontaneous EEG signals without explicit input from the user (Zhang et al., 2019b) and make a corresponding response to different emotional states. This kind of BCI may enhance the consumer experience in the time of an interactive session. Therefore, different approaches in Zhang et al. (2016, 2017) have been designed to recognize various emotion signals from brain wave. The latest affective BCIs (aBCIs) took machine learning algorithms and depended on a few features with discriminative information (Jenke et al., 2014; Mühl et al., 2014). When recording EEG signals in order to generate a desired target emotion signal, it is necessary to provide users with affective stimulation of specific emotions. In the training/calibration session, the required features and corresponding emotion labels are extracted from EEG signals to train the classifier. In an ongoing BCI session, the feature extractor receives the real-time EEG data and then sends the extracted features to the classifier for real-time affection classification. In this paradigm (Mühl et al., 2014), many researchers have reported pleasing classification performance. However, even if the experimental results are encouraging, the performance of aBCI still could be impacted by some reason. Since the EEG-based emotion signals are different from subject to subject, it is indispensable to train a specific object classifier for the subject of interest. Even in the same subject, the EEG signals are unstable, and the earlier trained classifier may perform poorly in the same subject at a later time. Therefore, in order to maintain a satisfactory classification accuracy, it is necessary to recalibrate frequently.

Domain adaptation method (Judy et al., 2017; Tzeng et al., 2017; Ding et al., 2018) has nearly completely dominated the recent literature of BCI (Jayaram et al., 2016). In aBCI studies, Dolan (2002), Koelstra et al. (2012), Shi et al. (2013), Mühl et al. (2014), Zheng et al. (2015), Zheng and Lu (2015, 2016), Chai et al. (2016, 2017), Lan et al. (2018), Zhong et al. (2020), and search various DA approaches by exploiting the SEED dataset.

PROPOSED FRAMEWORK

Notations and Definitions

We describe the column vectors and matrices according to the small and capital letters, respectively, in this article. The often utilized symbols are listed in **Table 1**. The concatenation representation of k matrices according to row (horizontally) is like $[A_1, A_2, ..., A_k]$, and these matrices concatenation operations along a column (vertically) is denoted as $[A_1; A_2; ...; A_k]$. The $l_{2,1}-norm$ of A is defined as $||A||_{2,1} = \sum_{i=1}^{n} \left|\left|A_{i,:}\right|\right|_2 = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{d} A_{ij}^2}$, and the trace-norm of A is indicated as $||A||_* = tr((AA^T)^{\frac{1}{2}})$ (Lotfi and Akbarzadeh, 2014).

We mainly focus on the unsupervised MDA based on S various sources of c-class. Suppose there are n_a $(a = 1, 2, ..., S)$ instances in every source domain, respectively. In the $a-$th source domain, given $X^a = \{x_1^a, ..., x_{n_a}^a\} \in \mathrm{R}^{d \times n_a} \in \chi$, and it is a training instances matrix with c sub-classes, which are associated with their class labels $Y^a = [y_1^a, ..., y_{n_a}^a]^T \in \mathrm{R}^{n_a \times c} \in \Gamma = \{0, 1\}^{c \times 1}$, a target domain dataset is denoted as $X^t = \{x_1^t, x_2^t, ..., x_m^t\} \in \mathrm{R}^{d \times n_t} \in \chi$, with their pseudo-class labels $Y^t = [y_1^t, ..., y_{n_t}^t]^T \in \mathrm{R}^{n_t \times c} \in \Gamma$ obtained from some supervised models (e.g., SVMs) which are trained on the source domain with labeled data. Our ultimate goal is to recognize the ground-truth class of test data $x_k^t \in X^t$, under the conditions that each domain pair X^a and X^t is assumed to be of different marginal and conditional distributions. While we do not need to limit that the instances number in each source domain is identical with that assumed when shaped into the training matrix, for the sake of simplicity, we can extract the same number of training instances from each source domain.

We further denote by $X^{a(\bar{l})} \left(\bar{l} = 1, ..., c\right)$ the set of samples in X^a with the label $\bar{l}$. Similarly, the sample set in the target domain X^t with the label $\bar{l}$ is defined as $X^{t(\bar{l})}$. Note that the true labels of the set $X^{t(\bar{l})}$ are unknown. We therefore employ in this work a base classifier, e.g., SVM, to attribute pseudo-labels for the subset in the target domain. For easy expression, we further define the matrix $X_a = [X^a, X^t] \in \mathrm{R}^{d \times N}$ $(N = n_t + n_a)$ with its label matrix $Y_a = [Y^a, Y^t]$ in packing both source and target data with respect to the $a-$th source domain.

Definition 1 (MDA): Let $\Delta = \left\{\mathrm{P}^1, \ldots, \mathrm{P}^S\right\}$ be a set of S source domains and $\mathrm{P}^t \notin \Delta$ be a target domain. Denote by $_{X^a=\{x_i^a, y_i^a\}_{i=1}^{n_a} \sim \mathrm{P}^a (a=1,...,S)}$ the samples drawn from the $a-$th source domains and by $X^t = \left\{x_i^t\right\}_{i=1}^{n_t} \sim \mathrm{P}^t$ the samples drawn from the target domain. The task of MDA is to learn an ensemble function $f_{\mathrm{P}^t} : \mathcal{X} \to \Gamma$ by co-learning multiple classifiers given X^a $(a = 1, \ldots, S)$ and X^t as the training examples.

Definition 2 (Multi-source Domain Generalization): In this scenario, the target domain is inaccessible in the training stage. Given S source domains $\Delta = \left\{\mathrm{P}^1, \ldots, \mathrm{P}^S\right\}$ and denoted by $X^a = \left\{x_i^a, y_i^a\right\}_{i=1}^{n_a} \sim \mathrm{P}^a$ the samples drawn from the $a-$th source, the task of multi-source generalization is to co-learn multiple adaptive functions $f_{\mathrm{P}^a} : \mathcal{X} \to \Gamma$ only given $X^a, \forall a = 1, \ldots, S$ as the training examples, which could be well generalized to a certain unseen target domain.

TABLE 1 | Notations and descriptions.

Notations	Descriptions
n	Data size
d	Feature dimensionality of data
χ	Data space
Γ	Label space
$a = [a_1, a_2, ..., a_d]^T \in R^d$	Feature vector
$A \in R^{n \times d}$	Data matrix
$A_{i,j}$	The (i, j) entry of A
$A_{i,:}$ and $A_{:,j}$	The i-th row and j-th column of A
A^T and a^T	The transpose of matrix A and vector a
$tr(A)$	The trace of a matrix A
$\langle A_1, A_2 \rangle = tr(A_1^T A_2)$	The inner product of two matrices A_1 and A_2
$\|\|a\|\|_p := \left(\sum_{i=1}^{d} \|\|a_i\|\|^p\right)^{1/p}$	The p-norm of a vector a
$\|\|A\|\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} A_{i,j}^2}$	The Frobenius norm of A
I_r	Identity matrix of size $r \times r$
1_d	d-dimensional vector of ones
0_d	d-dimensional vector of zeroes

Problem Statement

In representative MDA, one can use the strategy of acquiring knowledge from multiple auxiliary sources to promote the target task of interest, which is better than learning each source task alone in emotion recognition. That is to say that common knowledge shared by multi-source domains is beneficial to emotion analysis. Moreover, some optimal recognition models have been developed in the latest works for the source domain and/or target domain separately. Furthermore, in these methods, joint multi-source adaption emotion recognition and feature selection has been largely unaddressed, and little or limited efforts have yet been devoted to the utilization of the correlated knowledge among sources.

To solve the above-mentioned issues, we propose in this work a robust multiple-source adaption emotion recognition method based on EEG features. The method utilizes the correlated knowledge among domains and features by joint $l_{2,1}-norm$ and correlation metric regularization and can process high-dimensional, sparse, outliers, and non-i.i.d EEG data at the same time. The designed method has three characteristics, which are integrated into a unified optimization formulation to find an effective emotion recognition model and align the feature distribution between source and target domains: (1) *via* employing the $l_{2,1}-norm$ minimization, a robust loss term is introduced to avoid the influence of noise or outliers in EEG signal, and a sparse regularization term is designed to eliminate over-fitting and a sparse feature subset is selected; (2) based on the designed regression model and the semantic distribution matching between each pair of domains, it not merely provides robustness on loss function but also retains the domain distribution (including local and global) structures and meanwhile maintains a high dependence on the (pseudo)-label knowledge of the source domains and the target domain (Zhang et al., 2020a) so as to obtain preferable generalization

performance; and (3) through our constructed metric function of correlation, we can make full use of the correlative information among multiple sources and transfer more discriminative knowledge to the target domain. To implement these properties, in the following part, we will detail the objective formulation of the proposed method.

General Formulation

In this section, we propose the general formulation of MACI framework underpinned by the robust regression principle and the regularization theory. In particular, our main purpose is to optimize a unified objective function by compromising the following three complementary objectives:

1. Robust multi-source co-regression with feature selection using $l_{2,1}-norm$ minimization, in which the domain label consistency is explicitly maximized through iterative linear label regression.
2. Aligning domain distributions including global statistical distributions and intra-domain semantic distributions or class conditional distributions.
3. Effectively utilizing correlation information among source domains *via* developing an effective correlation metric function.

For the multi-source adaptation emotion recognition of interest, we define the $a-$th $(a = 1, ..., S)$ classifier function as $f_a(X_a) = X_a^T W_a$, where W_a is the $a-$thclassifier model, and W_0 is certain reference model. Suppose there is a kernel feature map $\phi_a : \chi \rightarrow H_a$ [1] that projects the training data from the original feature space into certain RKHS (Nie et al., 2010b) H_a, the predictor weight W_a can be kernelized. We denote the kernel matrix as $(K_a)_{i,j} = \left\langle \phi(x_i^a), \phi(x_j^a) \right\rangle$, where $x_i^a, x_j^a \in X_a$. We present the empirical kernel map as discussed in Gretton et al. (2009):

$$\begin{aligned} &\phi_a : \quad \chi \rightarrow \mathrm{R}^N, \quad \text{for linear kernel mapping} \\ &x \rightarrow K_a(\cdot, x^a)|_{x_1^a, x_2^a, \ldots, x_N^a} = \left(K_a(x_1^a, x^a), \ldots, K_a(x_N^a, x^a)\right), \\ &\qquad \text{for nonlinear kernel mapping} \end{aligned}$$

We therefore have kernel matrices $K_a = \phi_a(X_a)$. Hence, the kernelized decision function on X_a $(a = 1, ..., S)$ becomes $f_a(X_a) = K_a^T W_a$. We further denote by $W = [W_1; ...; W_S]$ the concatenation matrix.

We then endeavor to find S cross-domain models parameterized by $\{W_a\}_{a=1}^S$ in some empirical RKHSs *via* jointly utilizing correlated knowledge among sources and features. In view of the above-cited objectives, we propose the following general formulation of MACI.

$$\Theta(W_a, F_a) = R(K_a^T W_a, Y_a) + \Omega_A(X^a, X^t) + Cor(W), \quad (1)$$

where $R(\cdot, \cdot)$ is the robust regression function with feature selection *via* $l_{2,1}-norm$ minimization, $\Omega_A(X^a, X^t)$ is certain distance metric function for aligning the domain distributions, and $Cor(\cdot)$ is a correlation metric function which is a global regularization term. In the subsequent sections, we focus on designing these components in the general formulation one by one to construct a unified framework.

[1] It is important to note that the feature mapping function ϕ_a $(1 \leq a \leq S)$ with respect to each source domain can be completely different from each other.

Design of Robust Multi-Source Co-regression With Feature Selection

To achieve the first objective mentioned above, one should jointly minimize each source regression loss and implement feature selection, in which the domain label consistency is explicitly maximized, and the data outliers are accounted for to avoid negative transfer. To this end, we first explain a predicted label matrix $F_a \in \mathrm{R}^{N \times c}$ $(a = 1, ..., S)$ into our predictive function (Nie et al., 2010b). The predicted values in this label matrix should satisfy local smoothness and global consistency, i.e., they should preserve the local geometry while fitting in with the true labels (Zhang et al., 2020a). To satisfy these requirements, we present a smooth regularization term on the label geometric structure between each source instance (Nie et al., 2010b; Yan et al., 2006), which is formulated as

$$g(F_a) = \left\{ \begin{array}{l} tr\left[\left(F_a - Y_a\right)^T \mathring{\mathrm{U}} \left(F_a - Y_a\right)\right] + \alpha tr(F_a L_a (F_a)^T) \\ \left|\left| F_a^T F_a = I_c, (F_a)_{i,j} \geq 0 \right.\right. \end{array} \right\},$$

where $\mathring{\mathrm{U}}$ is a diagonal matrix with $\mathring{\mathrm{U}}_{i,i} = \varsigma$ (ς is a large specified value) if $x_i^a \in X_a$ has a label, $\mathring{\mathrm{U}}_{i,i} = 0$ or else and a is a regularization parameter. L_a is the graph Laplacian matrix of the $a-$th source dataset, which is defined as $L_a = \Lambda_a - \prod_a$, where Λ_a is a diagonal matrix with $(\Lambda_a)_{i,i} = \sum_j (\prod_a)_{i,j}$, and $\prod_a$ is the weight matrix of the graph, which can be defined as:

$$(\prod_a)_{i,j} = \left\{ \begin{array}{c} \exp(-\gamma_a ||x_i^a - x_j^a||^2), \quad \text{if } x_i^a \in \delta_k(x_j^a) \text{ or } x_j^a \in \delta_k(x_i^a) \text{ and both have the same labels} \\ \exp(-\frac{\gamma_a}{||x_i^a - x_j^a||^2}), \quad \text{if } x_i^a \in \delta_k(x_j^a) \text{ or } x_j^a \in \delta_k(x_i^a) \text{ and both have different labels} \\ 0, \quad \text{otherwise} \end{array} \right.,$$

where the $_k$ nearest neighbors of x are assigned to $\delta_k(x)$, and γ_a is a hyper-parameter, which can be empirically selected as $\bar{\theta}_a \sqrt{c}$ by considering the impact of multi-class distribution on the affinity relationship among the domain data, where $\bar{\theta}_a$ is the square root of the mean norm of $_{X_a}$.

We therefore design the following multi-source sparse co-regression model for meeting the first objective.

$$\begin{aligned} &R(W_a^T \phi(X^a), F_a) = \sum_{a=1}^S \vartheta_a^{q_1} \left(||K_a^T W_a - F_a||_{2,1} + g(F_a) + \beta ||W_a||_{2,1} \right) \\ &s.t. \ \sum_{a=1}^S \vartheta_a = 1, \end{aligned} \quad (2)$$

where $\vartheta = [\vartheta_1, ..., \vartheta_a]^T$ is the weight vector to jointly combine all source regression loss, β is a regularization parameter, and $q_1 > 1$ is a tunable parameter for avoiding trivial solution. The model (2) is convex, and the $l_{2,1}-norm$ loss function $||K_a^T W_a - F_a||_{2,1}$ is robust to outliers (Li et al., 2015). In the meantime, the term $||W_a||_{2,1}$ assures that W_a can accomplish feature selection across different domains due to its sparsity. That is, by exploiting the correlation among different features, our approach can jointly evaluate all feature knowledge of source domains and target domain.

Design of Domain Distribution Alignment

As a nonparametric distribution discrepancy estimator, MMD (Gretton et al., 2009) was used to compare two distributions by transforming the distributions into a RKHS (Pan et al., 2011; Duan et al., 2012b; Tao et al., 2012; Chen et al., 2013; Long et al., 2014). Let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \to \mathrm{R}$. The MMD between two domains P and Q is defined as

$$MMD_{\mathcal{F}}[\mathrm{P}, \mathrm{Q}] := \sup_{f \in \mathcal{F}} \left(\mathop{\mathrm{E}}_{\mathrm{P}}[f(x)] - \mathop{\mathrm{E}}_{\mathrm{Q}}[f(x)] \right). \quad (3)$$

The MMD measures the similarity level between two domains from the side of function class $\mathcal{F}$. To make the MMD a proper regularization for the classifier model W_a, we adopt the following the empirical estimation of MMD between X^a and X^t, which is defined as

$$MMD_e(X^a, X^t) := \left\| \frac{1}{n_a} \sum_{i=1}^{n_a} f_a(x_i^a) - \frac{1}{n_t} \sum_{j=1}^{n_t} f_a(x_j^a) \right\|_H^2 = tr(W_a^T K_a M_a K_a W_a), \quad (4)$$

where $||\cdot||_H$ is the RKHS norm, $(K_a)_{i,j} = < \phi_a(x_i^a), \phi_a(x_j^a) >$ with $x_i^a, x_j^a \in X_a$, and

$$M_{i,j}^a = \begin{cases} \frac{1}{n_a^2}, \text{ when } x_i^a, x_j^a \in X^a \\ \frac{1}{n_t^2}, \text{ when } x_i^a, x_j^a \in X^t \\ \frac{-1}{n_a n_t}, \text{ otherwise} \end{cases}. \quad (5)$$

As for $MMD_e(X^a, X^t)$ in Eq. 4, whereas even if there is a perfect domain distribution match, it does not assure the instances from different domains, but the same class of labels will be mapped near the transform space. Lack of semantic consistency will be a major reason for performance degradation. Therefore, we use the following terms to develop a semantically matched MMD (Long et al., 2014):

$$\begin{aligned} &MMD_{CA}(X^a, X^t) = \sum_{l=1}^{c} MMD_{\varepsilon}(X^{a(\bar{l})}, X^{t(\bar{l})}) \\ &= \sum_{l=1}^{c} tr(W_a^T K_{a(\bar{l})} M_{a(\bar{l})} K_{a(\bar{l})} W_a), \end{aligned} \quad (6)$$

where $K_{a(l)} = \phi_a([X^{a(\bar{l})}, X^{t(\bar{l})}])$ with $(K_{a(l)})_{i,j} = < \phi_a(x_i^{a(\bar{l})}), \phi_a(x_j^{a(\bar{l})}) >$, $x_i^{a(\bar{l})} \in X^{a(\bar{l})}$ and $x_j^{t(\bar{l})} \in X^{t(\bar{l})}$, and

$$(M_{a(\bar{l})})_{i,j} = \begin{cases} \frac{1}{n_{\bar{al}}^2}, \text{ when } x_i^{a(\bar{l})}, x_j^{a(\bar{l})} \in X^{a(\bar{l})} \\ \frac{1}{m_{\bar{l}}^2}, \text{ when } x_i^{a(\bar{l})}, x_j^{a(\bar{l})} \in X^{t(\bar{l})} \\ \frac{-1}{n_{\bar{al}} m_{\bar{l}}}, \text{ otherwise} \end{cases}. \quad (7)$$

We call Eq. 7 conditional (or semantic) MMD, which explicitly encourages instances from various domains but with the same label to map to the nearest in multi-source subspace. Finally, we suggest that the domain distribution alignment could be approached by learning multiple optimal models such that

$$\begin{aligned} &\Omega_A(X^a, X^t) = MMD_e(X^a, X^t) + MMD_{CA}(X^a, X^t) \\ &= tr(W_a^T K_a M_a K_a W_a) + \sum_{l=1}^{c} tr(W_a^T K_{a(\bar{l})} M_{a(\bar{l})} K_{a(\bar{l})} W_a) \\ &= \sum_{\bar{l}=0}^{c} tr(W_a^T C_{a(\bar{l})} W_a) = tr(W_a^T \sum_{\bar{l}=0}^{c} C_{a(\bar{l})} W_a), \end{aligned} \quad (8)$$

where $C_{a(0)} = K_a M_a K_a$ and $C_{a(\bar{l})} = K_{a(\bar{l})} M_{a(\bar{l})} K_{a(\bar{l})}$ $\left(\bar{l} = 1, ..., c\right)$. Let $C_a = \sum_{\bar{l}=0}^{c} C_{a(\bar{l})}$, then we have $\Omega_A(X^a, X^t) = tr(W_a^T\ C_a\ W_a)$.

Design of Correlation Metric Function

As we know, a commonly used strategy in extant classifier-centric adaptation methods (Duan et al., 2012c; Tommasi et al., 2014) is to directly match the discriminant models between different domains, which is defined as:

Definition 3 (model discriminant discrepancy, MDD): Let $\mathcal{W}$ be a set of function parameters $\mathcal{W} : \mathcal{X} \to \mathrm{R}$. The model discriminant discrepancy between domains P and Q is defined as

$$MDD_{\mathcal{W}}[\mathrm{P}, \mathrm{Q}] := \sup_{W_{\mathrm{P}}, W_{\mathrm{Q}} \in \mathcal{W}} \left|\left| W_{\mathrm{P}} - W_{\mathrm{Q}} \right|\right|_F^2.$$

It may be difficult to push these two models respectively learnt from different domains when the distribution discrepancy between them is large. In our correlation metric function, we instead aim to guarantee each source model to be aligned with a global reference matrix W_0 so as to enable different source models to share the common knowledge for effectively utilizing correlation information among source domains. In essence, W_0 builds a transformation among source domains so that knowledge of one source can be used to another. They yield the following model alignment function (MAF):

Definition 4 (model alignment function): Given S domains $\{X^a\}_{a=1}^S$ on $\mathcal{X}$, we can think of the classification model set $\{W^a\}_{a=1}^S$ in some latent spaces. Their MAF is then defined as

$$\Psi\left(\{W_a\}_{a=1}^S\right) = \sum_{a=1}^{S} \eta_a ||W_a - W_0||_F^2$$

where $\eta = [\eta_1, ..., \eta_S]^T$ is a weight vector for discriminatively selecting different source knowledge with $\sum_{a=1}^S \eta_a = 1$, and W_0 is certain shared (common) discriminant model among these domains (Zhang et al., 2019a).

In essence, the MAF measures the similarity between two domain classifiers by the classification model. The next theorem is about from MAF to MDD between two domains.

Theorem 1 (MAF bounds MDD): The (squared) maximum discriminant discrepancy between domains P and Q on $\mathcal{X}$ is upper-bounded by their MAF with $\eta_{\mathrm{P}} = \eta_{\mathrm{Q}} = \eta_\Delta$:

$$\eta_\Delta MDD_{\mathcal{W}}^2[\mathrm{P}, \mathrm{Q}] \le \Psi\left(\left\{W_{\mathrm{P}}, W_{\mathrm{Q}}\right\}\right)$$

where $\mathcal{W} = \{W : \mathcal{X} \to \mathrm{R} | W \text{ is the classifier model}\}$ and $W_{\mathrm{P}}, W_{\mathrm{Q}} \in \mathcal{W}$. Specially if $\mathcal{W}$ is induced by a characteristic kernel on $\mathcal{X}$, then $\Psi\left(\left\{W_{\mathrm{P}}, W_{\mathrm{Q}}\right\}\right) = 0$ if and only if $\mathrm{P} = \mathrm{Q}$.

Proof. By definition 4 and triangle inequality theorem,

$$\begin{aligned} &\Psi\left(\left\{W_{\mathrm{P}}, W_{\mathrm{Q}}\right\}\right) \\ &= \eta_{\mathrm{P}} ||W_{\mathrm{P}} - W_0||_F^2 + \eta_{\mathrm{Q}} ||W_{\mathrm{P}} - W_0||_F^2 \\ &= \eta_\Delta \left(||W_{\mathrm{P}} - W_0||_F^2 + \left|\left| W_{\mathrm{Q}} - W_0 \right|\right|_F^2 \right) \\ &\ge \eta_\Delta \left|\left| (W_{\mathrm{P}} - W_0) - (W_{\mathrm{Q}} - W_0) \right|\right|_F^2 \\ &= \eta_\Delta \left|\left| W_{\mathrm{P}} - W_{\mathrm{Q}} \right|\right|_F^2 = \eta_\Delta MDD_{\mathcal{W}}^2[\mathrm{P}, \mathrm{Q}], \end{aligned}$$

that is, $\Psi\left(\{W_{\mathrm{P}}, W_{\mathrm{Q}}\}\right)$ bounds $MDD^2_{\mathcal{W}}[\mathrm{P}, \mathrm{Q}]$. If $\mathcal{W}$ is induced by some characteristic kernel on $\mathcal{X}$, then $\Psi\left(\{W_{\mathrm{P}}, W_{\mathrm{Q}}\}\right) = 0$ if and only if $\mathrm{P} = \mathrm{Q}$, which can be concluded from the result of Theorem 2.2 of Gretton et al. (2009).

Theorem 1 also indicates that the MAF is an effective metric method if the kernel on $\mathcal{X}$ is characteristic (Gretton et al., 2009).

Theorem 2: In particular, if $W_0 = \eta_{\mathrm{P}} W_{\mathrm{P}} + \eta_{\mathrm{Q}} W_{\mathrm{Q}}$ in MAF, we obtain

$$\Psi\left(\{W_{\mathrm{P}}, W_{\mathrm{Q}}\}\right) = \eta_{\mathrm{P}}\eta_{\mathrm{Q}} MDD^2_{\mathcal{W}}[\mathrm{P}, \mathrm{Q}] \leq \frac{1}{4} MDD^2_{\mathcal{W}}[\mathrm{P}, \mathrm{Q}].$$

Proof. By $\eta_P + \eta_Q = 1$, we have

$$\begin{aligned}
&\Psi\left(\{W_{\mathrm{P}}, W_{\mathrm{Q}}\}\right) \\
&\quad = \eta_P \left|\left| W_{\mathrm{P}} - \left(\eta_{\mathrm{P}} W_{\mathrm{P}} + \eta_{\mathrm{Q}} W_{\mathrm{Q}}\right)\right|\right|_F^2 \\
&\qquad + \eta_Q \left|\left| W_{\mathrm{Q}} - \left(\eta_{\mathrm{P}} W_{\mathrm{P}} + \eta_{\mathrm{Q}} W_{\mathrm{Q}}\right)\right|\right|_F^2 \\
&\quad = \eta_P \left|\left| \eta_{\mathrm{Q}} W_{\mathrm{P}} - \eta_{\mathrm{Q}} W_{\mathrm{Q}}\right|\right|_F^2 + \eta_Q \left|\left| \eta_{\mathrm{P}} W_{\mathrm{P}} - \eta_{\mathrm{P}} W_{\mathrm{Q}}\right|\right|_F^2 \\
&\quad = \left(\eta_{\mathrm{P}}\eta_{\mathrm{Q}}^2 + \eta_{\mathrm{Q}}\eta_{\mathrm{P}}^2\right) \left|\left| W_{\mathrm{P}} - W_{\mathrm{Q}}\right|\right|_F^2 = \eta_{\mathrm{P}}\eta_{\mathrm{Q}} MDD^2_{\mathcal{W}}[\mathrm{P}, \mathrm{Q}] \\
&\quad \leq \tfrac{1}{4} MDD^2_{\mathcal{W}}[\mathrm{P}, \mathrm{Q}],
\end{aligned}$$

where the last inequality follows after observing that $\eta_{\mathrm{P}}\eta_{\mathrm{Q}} \leq 1/4$ with the equality holding when $\eta_{\mathrm{P}} = \eta_{\mathrm{Q}} = 1/2$.

For achieving the third target of ours, the correlation metric function therefore was designed as follows:

$$Cor(W) = \sum_{a=1}^{S} \eta_a^{q_2} ||W_a - W_0||_F^2 + \frac{\lambda}{2} ||W||_*, \tag{9}$$

where $q_2 > 1$ is another tunable parameter to avoid a trivial solution. The regularization term $||W||_*$ in Eq. 9 enables different projection functions $\{W_a\}_{a=1}^S$ to share common information/parts through models of sources. Thus, the knowledge from multiple sources can be further shifted from one by one source domain.

Final Formulation

By using empirical kernel function and integrating Eqs 2, 8, and 9, we therefore propose the following unified framework to implement the combination of feature selection and domain adaptive learning by utilizing the correlated knowledge across multi-sources and features.

$$\begin{aligned}
&\min_{W_a, F_a, \vartheta_a, \eta_a} \sum_{a=1}^{S} \left(\vartheta_a^{q_1} \left(\left|\left|K_a^T W_a - F_a\right|\right|_{2,1} + tr(W_a^T C_a W_a)\right)\right. \\
&\qquad \left. + g(F_a) + \beta ||W_a||_{2,1}\right) + \sum_{a=1}^{S} \eta_a^{q_2} \left|\left| W_a - W_0\right|\right|_F^2 + \tfrac{\lambda}{2} ||W||_* \\
&s.t.\ \sum_{a=1}^{S} \vartheta_a = \sum_{a=1}^{S} \eta_a = 1.
\end{aligned} \tag{10}$$

Note that, in Eq. 10, the $l_{2,1} - norm$ loss function makes the model effectively robust to noises or outliers from domains. In addition, after $l_{2,1} - norm$ regularization is added to W_a, many rows in $W_a (a = 1, ..., S)$ become zero. Therefore, the characteristics corresponding to these rows with zeros are not significant for the target task learning. Therefore, for acquiring competitive performance, we can select features from the original domain data. Similarly (Nie et al., 2010a), we sort the rows in W_a by descending sequence in light of each row of values in $l_2 - norm$ and followed by selecting the top rows as the feature selection outcome.

Remark 1: Our proposed method (10) has some competitions in universalization and efficiency. Firstly, in order to deal with the change of feature dimensions and types in among source domains easily, we learn a separate classification model for each independent domain pair (i.e., one source domain and one target domain). Then, for dealing with more heterogeneous sources, the algorithm is easy to extend. Moreover, in order to improve the speed of the algorithm, the redundant and irrelevant knowledge in the original features is thrown away before classification through sparse feature selection method, and then the classification models are learned. Furthermore, joint $l_{2,1} - norm$ and trace-norm minimization can be used to express well the together learning of feature selection and classification models so as to assure that the common subspace in the multi-domain can be extracted.

ALGORITHM

In this section, we first give an iterative approach to optimize the objective function (10) followed by its complexity and classification function. Although matrix completion is realized by an alike optimization method from Yang et al. (2013), we focus on the other issue, that is, joint optimization of trace norm and $l_{2,1} - norm$. We then further present an effective and valuable extension to domain generalization when the target domain is inaccessible.

Algorithm Optimization

According to Nie et al. (2010a), the derivative of $tr\left(\mathring{\mathrm{T}}^T Q \mathring{\mathrm{T}}\right)$ is equal to the derivative of $\left|\left|\mathring{\mathrm{T}}\right|\right|_{2,1}$, i.e., $2tr\left(\mathring{\mathrm{T}}^T Q \mathring{\mathrm{T}}\right) = \left|\left|\mathring{\mathrm{T}}\right|\right|_{2,1}$, where Q is a diagonal matrix and its $i-$th diagonal value is $Q_{ii} = \frac{1}{2||\mathring{\mathrm{T}}_{i,:}||_2}$, and if $\mathring{\mathrm{T}}_{i,:} = 0$, we can let $Q_{ii} = \frac{1}{2||\mathring{\mathrm{T}}_{i,:}||_2 + \varepsilon}$, where ε is a very small given value. Hence, we can farther transform Eq. 10 into Eq. 11:

$$\begin{aligned}
&\min_{W_a, F_a, \vartheta_a, \eta_a} \sum_{a=1}^{S} (\vartheta_a^{q_1} tr(T_a^T Z_a T_a) + \vartheta_a^{q_1} tr(W_a^T C_a W_a) \\
&\qquad + g(F_a) + \beta tr(W_a^T G_a W_a)) \\
&\qquad + \sum_{a=1}^{S} \eta_a^{q_2} \left|\left| W_a - W_0\right|\right|_F^2 + \tfrac{\lambda}{2} tr\left(W^T \left(WW^T\right)^{-\frac{1}{2}} W\right) \\
&s.t.\ \sum_{a=1}^{S} \vartheta_a = \sum_{a=1}^{S} \eta_a = 1,
\end{aligned} \tag{11}$$

where $S_a = [\underbrace{0_r, ..., 0_r}_{a-1}, I_r, \underbrace{0_r, ..., 0_r}_{S-a}]^T$, the diagonal matrix G_a is based on W_a, where the $k-$th element is equal to $(G_a)_{kk} = \frac{1}{2||(W_a)_{k,:}||_2}$, the diagonal matrix Z_a is based on $T_a = \sqrt{\vartheta_a^{q_1}}(K_a^T W_a - F_a)$, and where the $k-$th entry is computed

by $(Z_a)_{kk} = \frac{1}{2||(T_a)_{k,:}||_2}$. By taking the derivative of Eq. (11) in reference to W_0 and equaling to zero, we obtain:

$$W_0 = W\eta_0, \tag{12}$$

where $\eta_0 = [\tilde{\eta}_1 I_r; ...; \tilde{\eta}_S I_r]$ with $\tilde{\eta}_a = \eta_a^{q_2} \Big/ \sum_{a=1}^{S} \eta_a^{q_2}$ $(a = 1, ..., S)$. Substituting Eq. 12 into Eq. 11, we have

$$\begin{aligned} \min_{W_a, F_a, \vartheta_a, \eta_a} & \sum_{a=1}^{S} \{tr\left(T_a^T Z_a T_a\right) + \vartheta_a^{q_1} tr(W_a^T C_a W_a) \\ & + g(F_a) + \beta tr\left(W_a^T G_a W_a\right)\} \\ & + \sum_{a=1}^{S} \eta_a^{q_2} ||WS_a - W\eta_0||_F^2 + \tfrac{\lambda}{2} tr\left(W^T \left(WW^T\right)^{-\frac{1}{2}} W\right). \end{aligned} \tag{13}$$

Note that the sub-gradient matrices Z_a and G_a in Eq. 13 are dependent on the matrices W_a and F_a, which are also unknown beforehand. Consequently, the objective function in Eq. 13 is a multi-variable optimization problem involving the variables W_a, F_a, ϑ_a, η_a. Since optimizing these variables simultaneously is difficult, we exploit the alternating iterative strategy to update one variable iteratively while the other variable(s) is(are) fixed. Therefore, the problem in Eq. 13 can be decomposed into four sets of convex sub-problems. We aim to find the optimal solution to each sub-problem alternatively and iteratively so that the objective function in Eq. 13 would converge to a local optimal solution. By initializing W_a and F_a, thus initializing Z_a and G_a, then we can start the iterations.

Optimize W_a and F_a by Fixing ϑ and η

By solving the derivative of Eq. 13 in reference to W_a and equaling to zero, we obtain:

$$W_a = \vartheta_a^{q_1} \left(\vartheta_a^{q_1} C_a + \vartheta_a^{q_1} K_a Z_a K_a + \beta G_a + \Omega + \lambda U\right)^{-1},$$

$$K_a Z_a F_a = E_a F_a \tag{14}$$

where

$$U = \frac{1}{2}\left(WW^T\right)^{-\frac{1}{2}}, \tag{15}$$

$$\Omega = \sum_{a=1}^{S} \left(\eta_a^{q_2} \left(S_a - \eta_0\right)^T \left(S_a - \eta_0\right)\right), \tag{16}$$

and

$$E_a = \vartheta_a^{q_1} \left(\vartheta_a^{q_1} C_a + \vartheta_a^{q_1} K_a Z_a K_a + \beta G_a + \Omega + \lambda U\right)^{-1} K_a Z_a. \tag{17}$$

Plugging Eq. 14 into Eq. 13, by mathematical calculating, we can get:

$$tr\left(F_a^T N_a F_a\right) + \alpha tr(F_a L_a (F_a)^T) + tr\left[\left(F_a - Y_a\right)^T \mathring{U} \left(F_a - Y_a\right)\right], \tag{18}$$

where

$$N_a = E_a^T \left(C_a + \beta G_a + \lambda U\right) E_a + \alpha L_a + \left(K_a^T E_a - I_n\right)^T Z_a \left(K_a^T E_a - I_n\right).$$

Lastly, substituting the optimal solution of the other variables into Eq. 18 to update F_a, the constraints $F_a^T F_a = I_c$ should be added additionally, and $(F_a)_{ij} \geq 0$. Then, we can get the objective function in reference to F_a :

$$\Theta(F_a) = \min_{F_a} tr\left(F_a^T \left(N_a + \alpha L_a\right) F_a\right) +$$

$$tr\left[\left(F_a - Y_a\right)^T \mathring{U} \left(F_a - Y_a\right)\right] + \frac{\zeta}{2}\left|\left|F_a^T F_a - I_c\right|\right|_F^2 + tr\left(\theta F_a^T\right), \tag{19}$$

where ζ [2] and θ are balance arguments. By solving the derivative of Eq. 19 in reference to $F_{i,j}$ and equaling to zero and exploiting the K.K.T. with constraint term $\theta_{ij} F_{i,j} = 0$, we can get:

$$\begin{aligned} \frac{\partial \Theta(F_a)}{\partial F_a} &= \left(N_a + \alpha L_a + \mathring{U}\right) F_a - \mathring{U} Y_a + \zeta F_a \left(F_a^T F_a - I_c\right) \\ &\quad + \theta/2 = 0 \\ &\Rightarrow (F_a)_{i,j} \leftarrow (F_a)_{i,j} \frac{\left(\mathring{U} Y_a + \zeta F_a\right)_{i,j}}{\left(\zeta F_a F_a^T F_a + \left(N_a + \alpha L_a + \mathring{U}\right) F_a\right)_{i,j}}. \end{aligned} \tag{20}$$

Optimize ϑ_a by Fixing W_a, F_a, and η_a

In this situation, the issue in Eq. 13 changes to a small problem as follows:

$$\min_{\vartheta_a \geq 0, \vartheta_a^T 1 = 1} \sum_{a=1}^{S} \left\{\vartheta_a^{q_1} tr\left(T_a^T Z_a T_a\right) + \vartheta_a^{q_1} tr(W_a^T C_a W_a)\right\}. \tag{21}$$

Let $g_a = tr\left(T_a^T Z_a T_a\right) + tr(W_a^T C_a W_a)$, the Lagrange function of Eq. 21 is

$$\Im(\vartheta_a, \phi) = \sum_{a=1}^{S} \vartheta_a^{q_1} g_a - \phi\left(\sum_{a=1}^{S} \vartheta_a - 1\right). \tag{22}$$

Setting the derivative of $\Im(\vartheta_a, \phi)$ with respect to ϑ_a is equivalent to 0, and we can obtain:

$$\vartheta_a = \left(\phi \big/ (q_1 g_a)\right)^{\frac{1}{q_1 - 1}}. \tag{23}$$

Substituting Eq. 23 into the constraint $\sum_{a=1}^{S} \vartheta_a = 1$, we obtain

$$\vartheta_a = (g_a)^{1/(1-q_1)} \Big/ \sum_{a=1}^{S} (g_a)^{1/(1-q_1)}. \tag{24}$$

Optimize η_a by Fixing W_a, F_a, and ϑ_a

By fixing W_a, F_a, and ϑ_a, the problem in Eq. 13 then becomes the following sub-problem:

$$\min_{\eta_a \geq 0, \eta_a^T 1 = 1} \sum_{a=1}^{S} \eta_a^{q_2} ||WS_a - W\eta_0||_F^2. \tag{25}$$

[2]For maximal consistency between $F^T F$ and I_c, the parameter ζ should be set as a relatively large value. In our experiments, we therefore empirically set $\zeta = 10^3$ without loss of performance to some extent.

Let $h_a = ||WS_a - W\eta_0||_F^2$, the Lagrange function of Eq. 25 is

$$\Im(\eta_a, \rho) = \sum_{a=1}^{S} \eta_a^{q_2} h_a - \rho\left(\sum_{a=1}^{S} \eta_a - 1\right). \tag{26}$$

Setting the derivative of $\Im(\eta_a, \rho)$ in reference to η_a is equivalent to 0, and we then get:

$$\eta_a = \left(\rho \big/ (q_2 h_a)\right)^{\frac{1}{q_2 - 1}}. \tag{27}$$

Substituting Eq. 27 into the constraint $\sum_{a=1}^{S} \eta_a = 1$, we obtain:

$$\eta_a = (h_a)^{1/(1-q_2)} \Big/ \sum_{a=1}^{S} (h_a)^{1/(1-q_2)}. \tag{28}$$

Overall Procedure

In this sub-section, we finally report the whole optimization process of MACI in Algorithm 1, where a window-based breaking criterion is employed to better obtain the convergence of the algorithm (Zhang et al., 2019a). Concretely speaking, defining a window size $\hbar$, we compute $\varsigma = ||Max\Theta_{itr} - Min\Theta_{itr}|| \big/ Max\Theta_{itr}$ in $itr-$th iteration, where $\Theta_{itr} = \{Obj_{itr-\hbar+1}, ..., Obj_{itr}\}$ represents the set of historical target values in the window. When ς is less than a given threshold ε, that is $\varsigma < \varepsilon$, the algorithm stops iterating. In our experiments, we set $\varepsilon = 10^{-5}$ and $\hbar = 6$ empirically without losing statistical performance. We will discuss in section "Convergence" why Algorithm 1 is convergent.

Algorithm 1. Multiple sources adaptation learning by utilizing correlation knowledge.

Input: Source datasets $\{X_i^s\}_{i=1}^S$, $\{L_i\}_{i=1}^S$, target dataset X^t, and parameters α, β, and λ, the maximal iteration number ℓ.
Output: Converged projection matrices $\{W_i\}_{i=1}^S$, and matrices $\{F_i\}_{i=1}^S$ and W_0.
Initialization: Set $itr = 0$, and initialize $\{W_i^{itr}\}_{i=0}^S$ randomly. Let $W^{itr} = [W_1^{itr}, ..., W_S^{itr}]$;
1: for $i = 1$ to S do
{
Compute matrix M_i^{itr} and $M_{i(l)}^{itr}$, and K_i^{itr} and $K_{i(l)}^{itr}$ with empirical kernel mapping, thus computing $C_i^{itr} = \sum_{l=0}^{c} C_{i(l)}^{itr}$ by $C_{i(0)}^{itr} = K_i^{itr} M_i^{itr} K_i^{itr}$ and $C_{i(l)}^{itr} = K_{i(l)}^{itr} M_{i(l)}^{itr} K_{i(l)}^{itr}$, $l = 1, ..., c$;
Compute η_i^{itr} by Eq. 28, and then construct matrix η^{itr} and $\eta_0^{itr} = [\eta_1^{itr} I_r; ...; \eta_S^{itr} I_r]$;
Compute $F_i^{itr} = K_i^T W_i^{itr}$;
}
2: repeat
{
Compute W_0^{itr} by (12);
Compute the diagonal matrix U^{itr} by (15);
Compute the matrix $\boldsymbol{\Omega}^{itr}$ by (16);
set $i = 1$;
repeat
{
Compute G_i^{itr} with respect to W_i^{itr} ;
Compute Z_i^{itr} with respect to $K_i^T W_i^{itr} - F_i^{itr}$;
Compute $g_i^{itr} = tr\left(\left(K_i^T W_i^{itr} - F_i^{itr}\right)^T Z_i \left(K_i^T W_i^{itr} - F_i^{itr}\right)\right)$
$+tr((W_i^{itr})^T C_i^{itr} W_i^{itr})$;
Compute ϑ_i^{itr} according to Eq. 24, and then construct ϑ^{itr} ;
Compute the matrix E_i^{itr} by Eq. 17, and then N_i^{itr} by Eq. 19;
Compute F_i^{itr} according to Eq. 20;
Compute W_i^{itr} according to Eq. 14;
$i = i + 1$;
} until $i > S$
Update $W_i^{itr+1} = W_i^{itr}$ s.t. $i = 1, .., S$;
Update $F_i^{itr+1} = F_i^{itr}$ according to (20) s.t. $i = 1, .., S$;
Update ϑ_i^{itr+1} according to (24) s.t. $i = 1, .., S$;
Update η_i^{itr+1} according to (27) s.t. $i = 1, .., S$;
Update W_0^{itr+1} according to (12);
Let $itr = itr + 1$;
}until $itr > \ell$ or $\varsigma < 10^{-5}$
3: return $\{W_a\}_{a=0}^S$ and $\{F_i\}_{i=1}^S$.

Computational Complexity

In this subsection, we give a formal analysis about the computational complexity of several main components in Algorithm 1 using the big O notation. Firstly, the construction of the $k-NN$ graph and computing of the kernel matrix $\{K_i\}_{i=1}^S$, respectively, need computational cost $O(Sdn_a^2)$ and $O(SdN^2)$. Then, the optimization proceeds according to step by step iteratively. The cost for computing F_a is $O\left(3n_a^3 + n_a^2 c\right)$. After F_a was updated, computing W_a would cost $O(n_a c^2 + dc^2)$. In a word, the whole calculating cost is $O\left(\ell S\left(3n_a^3 + n_a^2 c + n_a c^2 + dc^2\right) + Sdn_a^2 + SdN^2\right)$. We assume that all Laplacian matrixes $\{L_i\}_{i=1}^S$ can be pre-calculated before the iterative optimization of Algorithm 1, and multi-kernel can be pre-calculated and put into memory before training. Thus, Algorithm 1 is effective and efficient computationally.

Target Classification

The datasets of the unlabeled samples from the target domain are defined as $K_u^t = \{(k_j^t)\}_{j=1}^{n_t}$. In the unsupervised DA learning case, one can predict a target class using the classification model W_0. Specifically, one may use $\arg\max_{1 \le j \le c} (\Gamma_u^t)_j$ to classify a test sample $k_u^t \in K_u^t$ into one of the c target classes, where $\Gamma_u^t = (W_0)^T k_u^t$.

In our multi-source adaptation framework, nevertheless, another voting method defined as "sum" can be deduced, that is, once $\{W_a^s\}_{a=1}^S$ are obtained, for a test data $k_u^t \in K_u^t$, we can learn its label vector Γ_u^t by minimizing the residue between Γ_u^t and the projected vector of each source model:

$$\min_{\Gamma_u^t} \sum_{a=1}^{S} \vartheta_a \left|\left|\left(k_u^t\right)^T W_a^s - \Gamma_u^t\right|\right|_2^2. \tag{29}$$

The result of Eq. 29 can be acquired according to the constraint term $\sum_{a=1}^{S} \vartheta_a = 1$:

$$\Gamma_u^t = \sum_{a=1}^{S} \vartheta_a \left(k_u^t\right)^T W_a^s. \tag{30}$$

Once Γ_u^t is computed by using Eq. 30, we then use $\arg\max_{1 \le j \le c} (\Gamma_u^t)_j$ to determine the class for this test data.

ALGORITHM ANALYSIS

Convergence

We start with the next two lemmas and then demonstrate that the alternant optimization process, namely, step 2 in Algorithm 1, in the optimization issue of the Eq. 10, the optimal solution of $\{W_i\}_{i=1}^S$ converges.

Lemma 1. (Nie et al., 2010a) There are any two values $V_1, V_2 \in \mathrm{R}^d$, and they are not equal to zero; we can get the inequality as:

$$||V_1||_2 - \frac{||V_1||_2^2}{2||V_2||_2} \leq ||V_2||_2 - \frac{||V_2||_2^2}{2||V_2||_2}. \tag{31}$$

Lemma 2. (Nie et al., 2010a) For any invertible matrices $\mathring{\mathrm{P}}$ and $\mathring{\mathrm{Q}}$, the following inequality holds:

$$\frac{1}{2}tr\left(\mathring{\mathrm{P}}\mathring{\mathrm{Q}}^{-\frac{1}{2}}\right) - tr\left(\mathring{\mathrm{P}}^{\frac{1}{2}}\right) \geq \frac{1}{2}tr\left(\mathring{\mathrm{Q}}\mathring{\mathrm{Q}}^{-\frac{1}{2}}\right) - tr\left(\mathring{\mathrm{Q}}^{\frac{1}{2}}\right). \tag{32}$$

Then, the iterative method designed in Algorithm 1 can converge to the optimal solution, which will be proved in the next theorem.

Theorem 3: In each iteration of Algorithm 1, the objective function of issue in Eq. 10 will be monotonically decreasing and finally will converge to the optimal solution of the issue.

Proof. For easy description, we define the updated W_i and F_i in the iteration τ as W_i^τ and F_i^τ $(i = 1, ..., S)$ separately. The updating from step 2 of Algorithm 1 is equivalent to the optimum of the next problem:

$$\begin{aligned}\min_{W_i,F_i,\vartheta_i,\eta_i} &\sum_{i=1}^S \left\{\vartheta_i^{q_1}\left[tr\left(\left(K_i^T W_i - F_i\right)^T Z_i\left(K_i^T W_i - F_i\right)\right)\right.\right.\\ &\left.\left.+tr(W_i^T C_i W_i)\right] + \alpha g(F_i) + \beta tr\left(W_i^T G_i W_i\right)\right\}\\ &+tr\left(W^T \Omega W\right) + \lambda tr\left(W^T U W\right)\end{aligned}$$

Following the expressions of Z_i, G_i, and U, then we can get:

$$\begin{aligned}&\sum_{i=1}^S \left\{\begin{array}{l} tr\left((W_i^{\tau+1})^T C_i W_i^{\tau+1}\right) + \alpha g(F_i^{\tau+1})\\ +\vartheta_i^{q_1}\sum_{j=1}^n \frac{||Z_i(j,:)^{\tau+1}||_2^2}{||Z_i(j,:)^\tau||_2} + \beta\sum_{j=1}^n \frac{||W_i(j,:)^{\tau+1}||_2^2}{||W_i(j,:)^\tau||_2}\end{array}\right\}\\ &+tr\left((W^{\tau+1})^T \Omega^{\tau+1} W^{\tau+1}\right) + \lambda tr\left((W^{\tau+1})^T U^\tau W^{\tau+1}\right)\\ \leq &\sum_{i=1}^S \left\{\begin{array}{l} tr\left((W_i^\tau)^T C_i W_i^\tau\right) + \alpha g(F_i^\tau)\\ +\vartheta_i^{q_1}\sum_{j=1}^n \frac{||Z_i(j,:)^\tau||_2^2}{||Z_i(j,:)^\tau||_2} + \beta\sum_{j=1}^n \frac{||W_i(j,:)^\tau||_2^2}{||W_i(j,:)^\tau||_2}\end{array}\right\}\\ &+tr\left((W^\tau)^T \Omega^\tau W^\tau\right) + \lambda tr\left((W^\tau)^T U^\tau W^\tau\right).\end{aligned} \tag{33}$$

We can have the next inequality by Lemma 1:

$$\begin{aligned}&\sum_{j=1}^n \left(\left|\left|(W_i)_{j,:}^{\tau+1}\right|\right|_2 - \frac{\left|\left|(W_i)_{j,:}^{\tau+1}\right|\right|_2^2}{2\left|\left|(W_i)_{j,:}^\tau\right|\right|_2}\right)\\ &\leq \sum_{j=1}^n \left(\left|\left|(W_i)_{j,:}^\tau\right|\right|_2 - \frac{\left|\left|(W_i)_{j,:}^\tau\right|\right|_2^2}{2\left|\left|(W_i)_{j,:}^\tau\right|\right|_2}\right).\end{aligned} \tag{34}$$

Therefore, we have

$$\begin{aligned}&\sum_{i=1}^S \left\{\begin{array}{l} tr\left((W_i^{\tau+1})^T C_i W_i^{\tau+1}\right) + \alpha g(F_i^{\tau+1})\\ +\vartheta_i^{q_1}\sum_{j=1}^n \left|\left|(Z_i)_{j,:}^{\tau+1}\right|\right|_2 + \beta\sum_{j=1}^n \left|\left|(W_i)_{j,:}^{\tau+1}\right|\right|_2\end{array}\right\}\\ &+tr\left((W^{\tau+1})^T \Omega^{\tau+1} W^{\tau+1}\right) + \lambda tr\left((W^{\tau+1})^T U^{\tau+1} W^{\tau+1}\right)\\ \leq &\sum_{i=1}^S \left\{\begin{array}{l} tr\left((W_i^\tau)^T C_i W_i^\tau\right) + \alpha g(F_i^\tau)\\ +\vartheta_i^{q_1}\sum_{j=1}^n \left|\left|(Z_i)_{j,:}^\tau\right|\right|_2 + \beta\sum_{j=1}^n \left|\left|(W_i)_{j,:}^\tau\right|\right|_2\end{array}\right\}\\ &+tr\left((W^\tau)^T \Omega^\tau W^\tau\right) + \lambda tr\left((W^\tau)^T U^\tau W^\tau\right).\end{aligned} \tag{35}$$

Eq. 35 can be further rewritten as:

$$\begin{aligned}&\sum_{i=1}^S \{tr\left((W_i^{\tau+1})^T C_i W_i^{\tau+1}\right) + \alpha g(F_i^{\tau+1}) + \vartheta_i^{q_1}\\ &\sum_{j=1}^n \left|\left|(Z_i)_{j,:}^{\tau+1}\right|\right|_2 + \beta\sum_{j=1}^n \left|\left|(W_i)_{j,:}^{\tau+1}\right|\right|_2\}\\ &+tr\left((W^{\tau+1})^T \Omega^{\tau+1} W^{\tau+1}\right) + \lambda tr\left(W^{\tau+1}(W^{\tau+1})^T U^{\tau+1}\right)\\ &-\tfrac{\lambda}{2}tr\left(\left(W^{\tau+1}(W^{\tau+1})^T\right)^{\frac{1}{2}}\right) + \tfrac{\lambda}{2}tr\left(\left(W^{\tau+1}(W^{\tau+1})^T\right)^{\frac{1}{2}}\right)\\ &\leq \sum_{i=1}^S \{tr\left((W_i^\tau)^T C_i W_i^l\right) + \alpha g(F_i^\tau) + \vartheta_i^{q_1}\\ &\sum_{j=1}^n \left|\left|(Z_i)_{j,:}^\tau\right|\right|_2 + \beta\sum_{j=1}^n \left|\left|(W_i)_{j,:}^\tau\right|\right|_2\}\\ &+tr\left((W^\tau)^T \Omega^\tau W^\tau\right) + \lambda tr\left(W^\tau(W^\tau)^T U^\tau\right)\\ &-\tfrac{\lambda}{2}tr\left(\left(W^\tau(W^\tau)^T\right)^{\frac{1}{2}}\right) + \tfrac{\lambda}{2}tr\left(\left(W^\tau(W^\tau)^T\right)^{\frac{1}{2}}\right).\end{aligned} \tag{36}$$

Noting that $U^l = \frac{1}{2}\left(W^\tau(W^\tau)^T\right)^{-\frac{1}{2}}$ and according to Lemma 2, we get

$$\begin{aligned}&\lambda tr\left(W^{\tau+1}(W^{\tau+1})^T U^{\tau+1}\right) - \lambda tr\left(\left(W^{\tau+1}(W^{\tau+1})^T\right)^{\frac{1}{2}}\right)\\ &\geq \lambda tr\left(W^\tau(W^\tau)^T U^\tau\right) - \lambda tr\left(\left(W^\tau(W^\tau)^T\right)^{\frac{1}{2}}\right).\end{aligned} \tag{37}$$

Subtracting Eq. (37) from Eq. 36, we have

$$\begin{aligned}&\sum_{i=1}^S \{tr\left((W_i^{\tau+1})^T C_i W_i^{\tau+1}\right) + \alpha g(F_i^{\tau+1}) + \vartheta_i^{q_1}\\ &\sum_{j=1}^n \left|\left|(Z_i)_{j,:}^{\tau+1}\right|\right|_2 + \beta\sum_{j=1}^n \left|\left|(W_i)_{j,:}^{\tau+1}\right|\right|_2\}\\ &+tr\left((W^{\tau+1})^T \Omega^{\tau+1} W^{\tau+1}\right) + \lambda tr\left(\left(W^{\tau+1}(W^{\tau+1})^T\right)^{\frac{1}{2}}\right)\\ &\leq \sum_{i=1}^S \{tr\left((W_i^\tau)^T C_i W_i^\tau\right) + \alpha g(F_i^\tau) + \vartheta_i^{q_1}\\ &\sum_{j=1}^n \left|\left|(Z_i)_{j,:}^\tau\right|\right|_2 + \beta\sum_{j=1}^n \left|\left|(W_i)_{j,:}^\tau\right|\right|_2\}\\ &+tr\left((W^\tau)^T \Omega^\tau W^\tau\right) + \lambda tr\left(\left(W^\tau(W^\tau)^T\right)^{\frac{1}{2}}\right).\end{aligned} \tag{38}$$

That is to say

$$\begin{aligned}&\sum_{i=1}^S \{tr\left((W_i^{\tau+1})^T C_i W_i^{\tau+1}\right) + \alpha g(F_i^{\tau+1}) + \vartheta_i^{q_1}\\ &\left|\left|K_i^T W_i^{\tau+1} - F_i^{\tau+1}\right|\right|_{2,1} + \beta\left|\left|W_i^{\tau+1}\right|\right|_{2,1}\}\\ &+tr\left((W^{\tau+1})^T \Omega^{\tau+1} W^{\tau+1}\right) + \lambda\left|\left|W^{\tau+1}\right|\right|_*\\ &\leq \sum_{i=1}^S \{tr\left((W_i^\tau)^T C_i W_i^\tau\right) + \alpha g(F_i^\tau) + \vartheta_i^{q_1}\\ &\left|\left|K_i^T W_i^\tau - F_i^\tau\right|\right|_{2,1} + \beta\left|\left|W_i^\tau\right|\right|_{2,1}\}\\ &+tr\left((W^\tau)^T \Omega^\tau W^\tau\right) + \lambda\left|\left|W^\tau\right|\right|_*.\end{aligned} \tag{39}$$

Hence, the theorem has been verified.

According to the optimization strategy of Algorithm 1, the objective function is monotonically decreasing in problem Eq. 10, so it is easiest to observe that the algorithm is convergent.

Generalization

In this part, we derive an empirical bound for our method that shows how both MAF and PDS control the generalization performance under the situation of the squared loss $loss(a, b) = (a - b)^2$. The main idea is to merge the domain scatter into the proven adaptive range for the distance difference (Ghifary et al., 2017).

Denote by $H:= \{h : \mathcal{X} \to \mathcal{Y}\}$ a hypothesis class of functions in the RKHS $\mathcal{H}$, where $\mathcal{X}$ is a compact set and $\mathcal{Y}$ is a label space. Given a loss function $loss(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathrm{R}_+$ and a domain distribution $\mathcal{D}$ over $\mathcal{X}$, we denote by $\mathcal{L}_{\mathcal{D}}\left(h, \mathring{\mathrm{h}}\right) = E_{x\sim\mathcal{D}}\left[loss\left(h(x), \mathring{\mathrm{h}}(x)\right)\right]$ the expected loss for the given two functions $h, \mathring{\mathrm{h}} \in H$. Then, the distance of domain difference between two distributions P and Q is defined as:

$$disc(\mathrm{P}, \mathrm{Q}) = \sup_{h, \mathring{\mathrm{h}} \in H} \{\mathcal{L}_{\mathrm{P}}\left(h, \mathring{\mathrm{h}}\right) - \mathcal{L}_{\mathrm{Q}}\left(h, \mathring{\mathrm{h}}\right)\}, \tag{40}$$

By the notation in Eq. 40, we can obtain domain generalization bounds by domain scatter. Let f_{P} and f_{Q} be the true labeling functions for domain P and Q, respectively, and $h^*_{\mathrm{P}} := \underset{h\in H}{\operatorname{argmin}} \mathcal{L}_{\mathrm{P}}\left(h, f_{\mathrm{P}}\right)$ and $h^*_{\mathrm{Q}} := \underset{h\in H}{\operatorname{argmin}} \mathcal{L}_{\mathrm{Q}}\left(h, f_{\mathrm{Q}}\right)$ be the minimizers. The following theorem provides adaptation bounds with PDS (. . ..).

Theorem 4 (adaptation bounds with PDS) (Ghifary et al., 2017): Denote by $H:= \{f \in \mathcal{H} : \mathcal{X} \to \mathrm{R}, \left||f|\right|_{\mathcal{H}} \leq 1$ *and* $\left||f|\right|_{\infty} \leq r\}$ a class of functions in the RKHS $\mathcal{H}$ and by $X^{\mathrm{P}}_{\mathcal{X}} = \left(x^s_1, ..., x^s_{n_s}\right) \sim \mathrm{P}$ and $X^{\mathrm{Q}}_{\mathcal{X}} = \left(x^t_1, ..., x^t_{n_t}\right) \sim \mathrm{Q}$ the source and target dataset, respectively. Let $loss(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to [0, \Upsilon]$ be a $q-$Lipschitz loss function, i.e., for all $a, b \in \mathcal{Y} \times \mathcal{Y}$, $\left||loss(a) - loss(b)|\right| = q\left||a - b|\right|$. Then, for any hypothesis $h \in H$, with probability of at least $1 - \delta$, the following generalization bound holds with the Rademacher complexity $\Re_{X^{\mathrm{P}}_{\mathcal{X}}}(H)$ over $X^{\mathrm{P}}_{\mathcal{X}}$:

$$\begin{aligned}&\mathcal{L}_{\mathrm{Q}}\left(h, f_{\mathrm{Q}}\right) - \mathcal{L}_{\mathrm{Q}}\left(h^*_{\mathrm{Q}}, f_{\mathrm{Q}}\right) \leq \mathcal{L}_{\mathrm{P}}\left(h, h^*_{\mathrm{P}}\right) + 2q\Re_{X^{\mathrm{P}}_{\mathcal{X}}}(H)\\ &+3\Upsilon\sqrt{\frac{\log\frac{2}{\delta}}{2n_t}} + 8r\sqrt{\Xi_{\phi}(\{\mu_{\mathrm{Q}}, \mu_{\mathrm{P}}\})} + \mathcal{L}_{\mathrm{P}}\left(h^*_{\mathrm{Q}}, h^*_{\mathrm{P}}\right).\end{aligned} \tag{41}$$

Theorem 4 provides a generalization bound for DA by introducing PDS and Rademacher complexity that measures the level to which a class of functions can fit random noise. The Rademacher complexity measure is the basis of relating empirical loss with expected loss. From Theorem 4, for a successful DA, we shall make $\mathcal{L}_{\mathrm{P}}\left(h^*_{\mathrm{P}}, h^*_{\mathrm{Q}}\right)$ as small as possible. According to definition 3, the (squared) loss $\mathcal{L}_{\mathrm{P}}\left(h^*_{\mathrm{P}}, h^*_{\mathrm{Q}}\right)$ is essentially equivalent to MDD in some optimal RKHS. We then further provide the following adaptation bounds with PDS and MAF, which follows by Theorem 1 combined with Theorem 4.

Theorem 5 (adaptation bounds with PDS and MAF): Denote by $H:= \{f \in \mathcal{H} : \mathcal{X} \to \mathrm{R}, \left||f|\right|_{\mathcal{H}} \leq 1$ and $\left||f|\right|_{\infty} \leq r\}$ a class of functions in the RKHS $\mathcal{H}$ and by $X^{\mathrm{P}}_{\mathcal{X}} = \left(x^s_1, ..., x^s_{n_s}\right) \sim \mathrm{P}$ and $X^{\mathrm{Q}}_{\mathcal{X}} = \left(x^t_1, ..., x^t_{n_t}\right) \sim \mathrm{Q}$ the source and target dataset, respectively. Let $loss(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to [0, \Upsilon]$ be a $q-$Lipschitz loss function, i.e., for all $a, b \in \mathcal{Y} \times \mathcal{Y}$, $\left||loss(a) - loss(b)|\right| = q\left||a - b|\right|$. Denote by $W^*_{\mathrm{P}}, W^*_{\mathrm{Q}}$ the optimal functions learnt from domain P and domain Q, respectively; then, for any hypothesis $h \in H$, with probability of at least $1 - \delta$, the following generalization bound holds with the Rademacher complexity $\Re_{X^{\mathrm{P}}_{\mathcal{X}}}(H)$ over $X^{\mathrm{P}}_{\mathcal{X}}$:

$$\begin{aligned}&\mathcal{L}_{\mathrm{Q}}\left(h, f_{\mathrm{Q}}\right) - \mathcal{L}_{\mathrm{Q}}\left(h^*_{\mathrm{Q}}, f_{\mathrm{Q}}\right) \leq \mathcal{L}_{\mathrm{P}}\left(h, h^*_{\mathrm{P}}\right) + 2q\Re_{X^{\mathrm{P}}_{\mathcal{X}}}(H)\\ &+3B\sqrt{\frac{\log\frac{2}{\delta}}{2n_t}} + 8r\sqrt{\Xi_{\phi}(\{\mu_{\mathrm{Q}}, \mu_{\mathrm{P}}\})} + \Psi\left(\left\{W^*_{\mathrm{P}}, W^*_{\mathrm{Q}}\right\}\right).\end{aligned} \tag{42}$$

Theorem 5 clearly shows that the projected domain scatter $\Xi_{\phi}(\{\mu_{\mathrm{Q}}, \mu_{\mathrm{P}}\})$ and MAF $\Psi\left(\left\{W^*_{\mathrm{P}}, W^*_{\mathrm{Q}}\right\}\right)$ can control the generalization performance of MACI with its empirical measure, that is, to minimize the PDS (or, alternatively, the distributional scatter discrepancy) and MAF (or model discrimination discrepancy) in our methods can effectively improve the generalization bound in the setting of MDA or domain generalization, which is also supported by the following real-world experiments.

EXPERIMENTS

In this section, to evaluate the effectiveness of MACI for emotion recognition, we compare it with several state-of-the-art methods on two benchmark datasets, i.e., DEAP (Koelstra et al., 2012) and SEED (Zheng and Lu, 2015), which are also widely adopted as benchmark datasets for EEG-based emotion recognition (Mansour et al., 2009). Since existing deep DA models have demonstrated to be very effective, mainly applied to the EEG-based emotion recognition problems (Lotfi and Akbarzadeh, 2014), we divide our experiments into two parts, i.e., comparisons with shallow (traditional) DA methods on those emotion recognition tasks mentioned above and comparisons with the deep (CNN-based) DA methods for EEG-based emotion recognition on several cross-datasets.

DATA PREPARATION

At present, there are some EEG datasets for emotional state research. In this article, we used the following two public datasets: DEAP (Koelstra et al., 2012) and SEED (Zheng and Lu, 2015). As reported in Zhong et al. (2020) and Lan et al. (2018), there is a significant difference between these two databases due to some technical aspects. We also adopted the same feature extraction strategy with that in Lan et al. (2018). More details about these two databases can be found in Lan et al. (2018).

In our experiments, differential entropy (DE) (Zhong et al., 2020) is employed as the feature of emotion recognition. In the literature (Shi et al., 2013; Zheng et al., 2015; Chai et al., 2016; Zheng and Lu, 2016; Chai et al., 2017; Lan et al., 2018;

Zhong et al., 2020) about the DA emotion recognition based on EEG, DE features have been widely used. The details of DE are explained in Lan et al. (2018).

Baseline Setting

We compare our MACI method with the following state-of-the-art (related) baselines for multi-source emotion recognition tasks. Besides this, we also report the emotion recognition results of MACI using several deep features:

- No adaptation baseline FSSL (Yang et al., 2013)
- Multi-kernel adaptation method: FastDAM (Duan et al., 2012c)[3]
- Multi-KT (Tommasi et al., 2014)[4]: according to Tommasi et al. (2014), we here also use the l_2-norm constraint on p in Multi-KT algorithm
- Adaptive SVM: A-SVM (Yang et al., 2007)[5]
- Domain selection machine (DSM) (Duan et al., 2012a)
- Deep DA methods: DAN (Long et al., 2015) and ReverseGrad (Ganin and Lempitsky, 2015).

For the baseline FSSL without adaptation and the multi-source adaptation method A-SVM, we just equally fuse the decision values of all base classifiers with each classifier learned on one source domain[6].

In our MACI, only several vital parameters such as q_1, q_2, λ, α, and β in our model need to be predefined. Considering that parameter determination is a yet unaddressed open issue, we determine these parameters empirically as in our previous works (Tao et al., 2019). The parameters q_1 and q_2 play the same role in optimizing ϑ_a and η_a for preventing the trivial solution of these optimal variables. Since the larger q_1 (or q_2) would lead to the same weights with greater probability, we therefore empirically set $q_1 = q_2 = 2$ in our experiments in terms of the suggestion provided in Hou et al. (2017). Besides this, we discreetly choose the values of λ, α, and β by employing the grid search strategy in a heuristic way. Concretely, these regularization parameters are tuned from $\{10^{-4}, 10^{-3}, ..., 10^3, 10^4\}$. Finally, we search and fine-tune the number of nearest neighbors k in the set {3, 5, 10, 15, 17} for constructing the affinity graph in MACI (also in FSSL). For our algorithm, the maximum iteration number is set as $\tau = 100$.

For those nonlinear learning methods MACI, FastDAM, and Multi-KT, we borrow the Gaussian kernel [i.e., $K_{i,j} = \exp\left(-\sigma \left|\left|x_i - x_j\right|\right|^2\right)$] as the default kernel function, where σ is determined by setting it to be the reciprocal of feature dimension $1/d$. Following the same practice in Duan et al. (2012a), we predefine each source weight $\gamma_i = \frac{\exp(-\delta Dist(X_i^s, X))}{\sum_i \exp(-\delta Dist(X_i^s, X))}$ ($i = 1, ..., S$) in FastDAM, where $\delta = 100$.

Experiment I: Within-Dataset Emotion Recognition

It is worth noting that we may encounter difficulties with different subjects in EEG emotion recognition even if they belong to the same dataset because different subjects may have different EEG feature distributions due to personalized characteristics. Thereby, we may adopt the so-called leave-one-subject-out cross-validation strategy adopted also in Lan et al. (2018) to evaluate the performance of MACI on emotion recognition. Concretely, the left subject from the dataset of interest contributes to the target domain, and other subjects are constructed as the multi-source domains. We evaluate the multi-source adaptation performance of MACI compared with existing state-of-the-arts on SEED and DEAP, respectively.

There are totally 2,340 training data of 13 subjects with 60 data per class and 180 test data of each subject from three classes in DEAP. We extracted 2,775 samples consisting of 925 samples per class per class hour from each one of the 14 subjects in SEED, thus generating 38,850 training data from 14 subjects and 2,775 test samples from one target subject. Note that extant research (Chai et al., 2016, 2017; Zheng and Lu, 2016) have pointed out that it is almost impossible to train the DA methods by exploiting all training data from SEED due to the limitation of computational space. We thereby randomly sample 10% training data from SEED, i.e., 3,885 training data as the final multi-source domain data for all DA methods. We repeat each trial 10 times on SEED, and the final performance is the average of the results of 10 times.

Performance Comparison

We show in **Table 2** the emotion recognition performance of MACI and several baselines on within-DEAP and within-SEED, respectively.

As reported in Lan et al. (2018), the theoretical performance (or chance level) of random guessing is about 33.33%, which could be approached by real chance level when the number of training samples increase to infinity (Lan et al., 2018). As shown in **Table 1**, the baseline FSSL contributes 40.17% mean recognition accuracy on DEAP, which is very near to the random value. When there are finite samples, we obtain the empirical chance level by repeating the trials of the samples in question equipped with randomized class labels (Lan et al., 2018). The finally obtained chance levels with bound of 95% confidence interval are also recorded in **Table 2**. We can see from **Table 2** that the accuracy of FSSL significantly exceeds the upper bound of the real chance level at 5% significance level. However, the relatively lower performance of FSSL still indicates that emotion recognition with DA technique is imperative when there exists substantial divergence between the feature distributions of different subjects.

Almost all DA methods yield better recognition performance than FSSL for DEAP. Our MACI achieves the best performance (about 23.14% gains in performance over FSSL), closely followed by DSM. Note that though we acquired the relatively significant improvement measured by t-test with p-value > 0.05, the total recognition accuracy is still inferior. On SEED, FSSL achieves

[3]The code is available from http://vc.sce.ntu.edu.sg/transfer_learning_domain_adaptation_data/DAM-TNNLS2012.html.

[4]We use the code available from http://homes.esat.kuleuven.be/~ttommasi/source_code_CVPR10.html.

[5]The MATLAB code is available online http://www.robots.ox.ac.uk/~{}vgg/software/tabularasa/.

[6]For each source domain, we train one SVM by using the corresponding labeled samples. Then, for each test instance x, the decision values from p SVM classifiers are converted into the probability values by using the sigmoid function [i.e., $g(t) = 1/(1 + \exp(-t))$]. Finally, we average the p probability values as the final prediction of the test instance x.

TABLE 2 | Emotion recognition performance (mean % and SD %) of MACI and several baselines on within-datasets.

Method	DEAP		SEED							
			Session I		Session II		Session III		Average	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
FSSL	40.17	4.36	57.96	6.85	48.79	5.47	57.45	9.09	53.78	6.96
Multi-KT	55.83	5.59	73.56	4.37	68.89	3.43	72.57	7.38	70.68	5.09
A-SVM	49.49	7.92	65.82	7.86	64.00	7.09	69.08	10.77	65.25	8.53
FastDAM	57.37	5.50	72.31	6.86	69.45	7.18	75.64	7.37	71.52	7.04
DSM	60.22	6.50	72.76	6.86	70.10	5.18	76.35	7.37	72.27	6.47
MACI	63.31	4.50	73.42	6.86	70.81	6.18	77.43	7.37	73.23	6.66
Upp Bnd of Chn Lvl	38.85		34.58		34.65		34.60		34.61	

53.78% average recognition accuracy over three sessions, which obviously surpasses the upper bound of the chance level. Several multi-source adaptation methods, i.e., Multi-KT, FastDAM, and DSM, undoubtedly obtain more performance gains than FSSL on SEED. The proposed method MACI still effectively boost the mean recognition accuracy up to 73.23% under t-test with p-value > 0.05, which still demonstrates the best performance on SEED. It is worthy to note that all methods including FSSL work more effectively on SEED than on DEAP. This interesting observation is partially consistent with that in Lan et al. (2018). A possible explanation may be that the so-called negative transfer prevented the effective application of DA techniques in DEAP since larger discrepancies among different subjects may exist in DEAP than in SEED (Mansour et al., 2009; Lan et al., 2018).

Number of Source Samples

Figure 1 presents the effect of varying the number of source samples. The source dataset size varies from 100 to 2,300 on DEAP and 100 to 3,800 on SEED, respectively. It can be seen from the curves in **Figure 1** that all methods manifest the same trend of upgrade in the figure. This shows that larger source data is beneficial to improve the learning performance. It is worthy to note that the performance of MACI can be smoothly and steadily improved with the increase of the source samples, while other DA methods may only achieve satisfactory performance when the source samples are relatively large, i.e., larger than 500. In addition, A-SVM obtains the least performance on two datasets due to the so-called negative transfer issue in DA.

On DEAP, MACI, and DSM maintain a better accuracy than other methods with less than 500 source domain samples. DSM outperforms Multi-KT and FastDAM when the number of target training samples is relatively large due to properly choosing the weights to assign to each subject. Our method MACI obtains even more gains over DSM when the number of source samples is increasing asymptotically. The reason may be that, except for the use of correlation information among subjects, MACI can effectively select the most related sources with the optimally weighted multi-source adaptation regularization. Besides that, partial experimental results also show that FastDAM could occasion the "negative transfer" issue with the MMD-based weights assigned to all sources, which would deteriorate its performance. On SEED, the accuracy flattens at above 3,000 source samples. When the number of source samples increases to 3,500, MACI, FastDAM, and Multi-KT asymptotically approach a similar performance. From this point onwards, MACI, FastDAM, and Multi-KT perform similarly if we have sufficient source data.

Multi-Kernel Learning

We further evaluate the effectiveness of our method with different kernel functions (called MKMACI for short) for each source domain. Given the empirical kernel mapping set $\{\phi_a\}_{a=1}^{\upsilon}$, each mapping X_a into a different kernel space, we can integrate them orthogonally to the final space by concatenation, i.e., $\widetilde{\phi}(x_i) = [\phi_1(x_i)^T, \phi_2(x_i)^T, ..., \phi_\upsilon(x_i)^T]^T \in R^{\upsilon n_a}$, for $x_i \in X_a$. The final kernel matrix in this new space is defined as $K_{new} = [\mathring{\mathrm{K}}_1; \mathring{\mathrm{K}}_2; ...; \mathring{\mathrm{K}}_\upsilon]$, where $\mathring{\mathrm{K}}_i$ is the kernel matrix in the $i-$th feature space. Therefore, besides the above-mentioned Gaussian kernel, we additionally employ another three types of kernels in MKMACI: Laplacian kernel $K_{ij} = \exp\left(-\sqrt{\sigma}\left||x_i - x_j\right|\right)$, inverse square distance kernel $K_{ij} = 1\Big/\left(1 + \sigma\left||x_i - x_j\right|\right|^2\right)$, and inverse distance kernel $K_{ij} = 1/(1 + \sqrt{\sigma}\left||x_i - x_j\right|\right)$. It can be clearly seen in **Figure 2** that MKMACI is obviously better than MACI in terms of mean accuracies in all cases, which justifies that the multi-kernel trick can improve the quality of DA emotion recognition on within-datasets.

Experiment II: Cross-Dataset Emotion Recognition

Note that cross-dataset emotion recognition is more challenging in terms of the differences in acquisition and participant characteristics and behaviors. In the preceding experiments, we demonstrate the performance comparison of our method with other DA methods with the within-dataset (i.e., cross-subject) setting. We will, in this part, further evaluate the consistent effectiveness of MACI when performed on cross-dataset adaptation. In this scenario of experiment, we constructed multiple different schemes by sampling the training dataset and test dataset, respectively, with different EEG instruments and emotional stimuli. We therefore set up six trial settings, i.e., *DEAP* → *SEED I*, *DEAP* → *SEED II*, *DEAP* → *SEED III*, *SEED I* → *DEAP*, *SEED II* → *DEAP*, and *SEED III* → *DEAP*, to justify the effectiveness of MAC on cross-dataset emotion

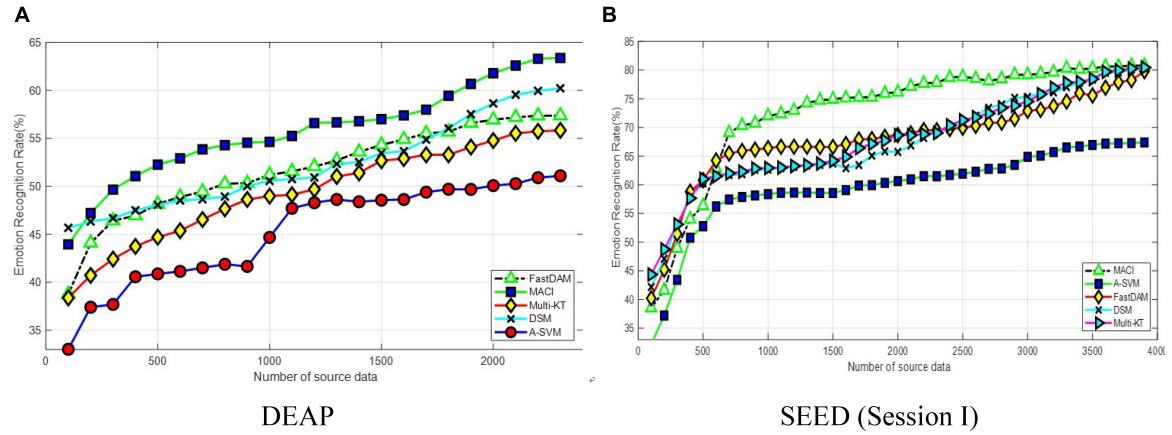

FIGURE 1 | Classification accuracy with varying numbers of source samples on **(A)** DEAP and **(B)** SEED (session I).

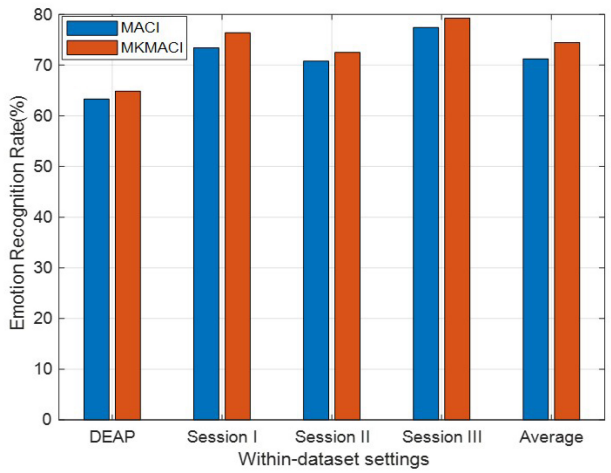

FIGURE 2 | Domain adaptation emotion recognition on within-dataset with multi-kernel learning.

recognition. In the context, we denote A→B by adaptation from dataset A to dataset B. For simplicity of expression, we respectively, coined SEED I, SEED II, and SEED III as the dataset of session I, session II, and session III in the database SEED.

In universe DA, a commonly used hypothesis is that the feature space of both source and target domains is the same. Consequently, only 32 channels between SEED and DEAP are employed to formulate a 160 dimensional feature space for both training and test datasets. In the first three experimental settings, there are 180 × 14 = 2,520 source samples from DEAP and 2,775 target samples from three different sessions in SEED. We evaluate the recognition accuracy for each subject in each session and report the final experimental results based on the mean over 15 subjects from SEED. In the other experimental settings, a total of 2,775 × 15 = 41,625 source samples from SEED are regarded as training datasets, and 180 samples contributed from DEAP are test dataset. We then evaluate the recognition accuracy of individual subjects in DEAP, and the results are recorded with the average over 14 subjects. We randomly sample 10% of the source data (4,162 samples) as the actual training data due to the limitation of memory (Shi et al., 2013; Zheng et al., 2015; Chai et al., 2016, 2017; Zheng and Lu, 2016; Lan et al., 2018; Zhong et al., 2020). Under each setting, we conduct the trial repeatedly 10 times and record the average performance over these 10 times.

Performance Comparison

We record the mean experimental results on six cross-dataset settings in **Table 3**, from which we can observe that the performance of the baseline FSSL is inferior to the upper bound of chance level with 95% confidence interval, that is, the baseline performance is almost close to the random guess with 5% significance level. This indicates that there exist larger distribution divergences between two datasets as well as the variance among different subjects than that in within-dataset. The importance of DA would be indispensable in this scenario. This is justified by the observation in **Table 3** that all DA methods outperform the baseline FSSL since DA could potentially reduce the technical discrepancies in cross-dataset applications. In most cases, MACI is found to be the best-performing method in the cross-dataset DA settings. In some scenario, Multi-KT and FastDAM occasionally obtain the best performance. A noticeable phenomenon can be observed in **Table 3**, such that the mean recognition accuracies of all methods are correspondingly worse than that in **Table 2** obtained on within-dataset due to the larger distribution discrepancy between different datasets.

Multi-Source Adaptation

In practical DA applications, one may expect that the number of prior sources grow in time, which would incur the so-called scalability issue. In this problem, it is necessary to explore the reliability of each prior source for the specific task (Tao et al., 2019). To this end, we additionally conduct multi-source adaptation trials on several cross-dataset settings. The average results of MACI, DSM, Multi-KT, FastDAM, and A-SVM with the average prior model are reported in **Figure 3**.

TABLE 3 | The recognition accuracy (mean%) with cross-dataset settings.

Method	DEAP→SEED I	DEAP→SEED II	DEAP→SEED III	SEED I→DEAP	SEED II→DEAP	SEED III→DEAP
FSSL	32.42	33.71	34.47	33.57	32.99	32.51
A-SVM	55.86	58.48	60.84	39.68	40.08	39.53
FastDAM	65.72	62.68	66.21	48.40	49.90	47.46
DSM	68.47	64.68	64.33	50.22	51.44	50.46
Multi-KT	67.74	65.51	64.65	48.73	52.16	51.27
MACI	69.36	67.60	65.43	54.37	51.88	51.76
Upp Bnd of Chn Lvl.	34.68	34.72	34.74	38.35	38.38	38.44

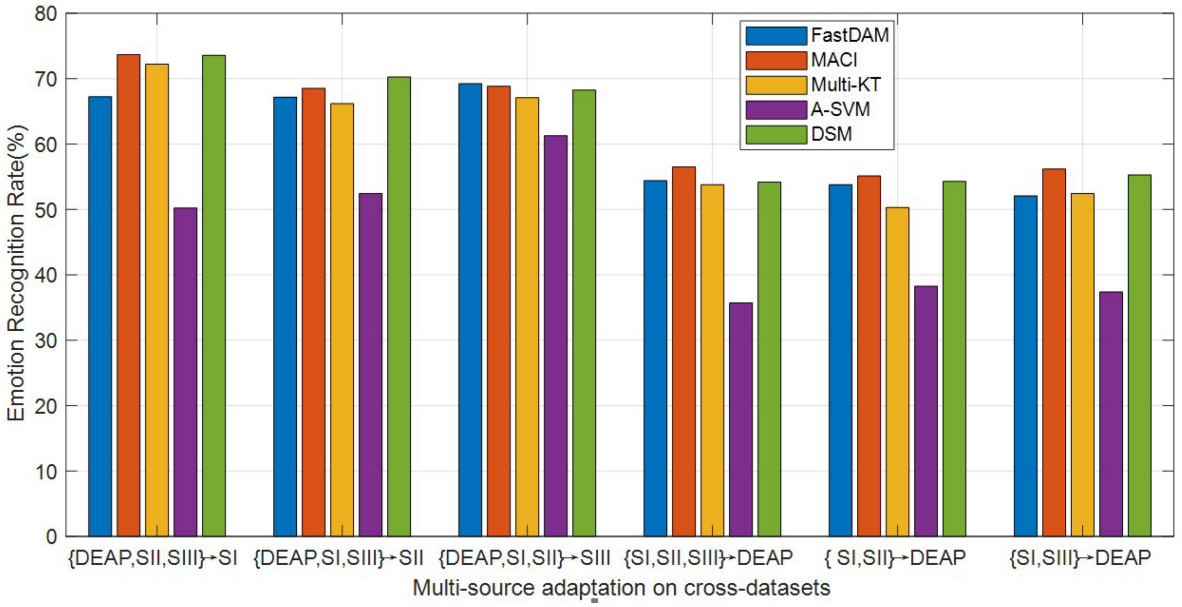

FIGURE 3 | Multi-source adaptation emotion recognition accuracy (SI, session I; SII, session II; SIII, session III).

TABLE 4 | Cross-dataset emotion recognition rates with different strategies of parameter settings.

Method	{DEAP, SII, SIII}→SI	{DEAP, SI, SIII}→SII	{DEAP, SI, SII}→SIII	{SI, SII, SIII}→DEAP	{SI, SII}→DEAP	{SI, SIII}→DEAP
MACI_NF	73.32	67.24	69.78	54.62	55.04	54.39
MACI_NS	69.88	65.31	66.07	52.92	53.44	52.81
MACI	73.69	68.52	68.85	56.52	55.12	56.17

SI, session I; SII, session II; SIII, session III.

It can be seen from the curves in **Figure 3** that A-SVM is still worse than the other DA methods in most cases in that it is difficult for A-SVM to minimize the between-domain distribution distance when the distribution varies greatly between domains. The accuracies of A-SVM tend to be downgraded when the number of sources is increasing in some cases, suggesting that negative transfer may have happened in A-SVM. MACI obtains a relatively much better performance in most cases, which demonstrates that our algorithm can improve the emotion recognition performance on cross-dataset. All methods except A-SVM manifest the same trend of upgrade with the increase of sources, and the accuracy improvements are significant with respect to that of within-dataset settings. This shows that utilizing the limited sources is beneficial to improve the learning performance. In addition, MACI and DSM usually outperform other DA methods due to properly choosing the weights to assign to each source. Our method MACI obtains even more gains over DSM, which may be attributed to the utilization of correlation information among sources in that MACI can effectively select the most related source domains with the optimally weighted multi-source adaptation regularization.

Adaptation With Deep Features

In the past decade, deep learning attracts more and more attention due to its powerful representation ability and dramatic improvement over the traditional shallow methods. We therefore additionally compare our MACI method with the recently proposed deep transfer learning models DAN and ReverseGrad

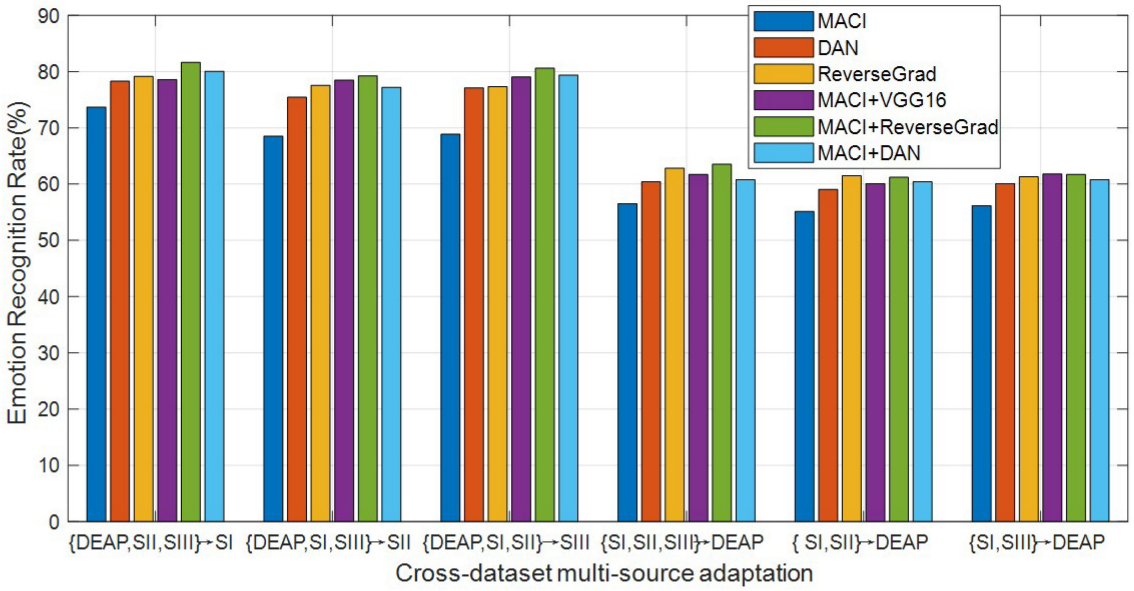

FIGURE 4 | Emotion recognition accuracies (%) of different methods using deeply extracted features (SI, session I; SII: session II; SIII, session III).

for cross-dataset emotion recognition using deeply extracted features in multi-source adaptation settings.

In our MACI, we can tackle the problem of deep DA with two steps: firstly, a higher-level feature extraction is learnt in an unsupervised fashion from all available domains using the popular deep architectures [e.g., VGG16 (Simonyan and Zisserman, 2014) or DAN]; secondly, our MACI is trained on the transformed data of all domains and then used to test the target domain. For fair comparison, however, we follow the experimental setup in Zhou et al. (2018) and Zhu et al. (2017). Specifically, we first fine-tune pretrained deep models (e.g., VGG16, DAN, and ReverseGrad) by using the labeled samples in the source domain and then use these fine-tuned CNN models to extract the features from EEG in both source and target domains. Finally, we perform emotion recognition using MACI on these deeply extracted features. In the context of our experiments, we denote our methods with different deep models as MACI + VGG16, MACI + DAN, and MACI + ReverseGrad, respectively. As for DAN and ReverseGrad, we use their released source codes and fine-tune the pre-trained deep models by using the suggested parameters in Long et al. (2015) and Ganin and Lempitsky (2015), respectively.

All experimental results are reported in **Figure 4**. As can be seen from this plot, the deep transfer learning methods are originally proposed to learn domain-invariant features, while our proposed method aims to improve the cross-domain generalization ability, namely, their methods focus on feature learning, while our work focuses on classification, so our proposed method can be used to further improve the recognition accuracies by co-learning the source classifiers with the features extracted by deep models, i.e., VGG16, DAN, and ReveseGrad. This indicates that the classification-level constraint can preserve all source discriminative structures for the guidance of target data classification, which demonstrates the effectiveness of MACI framework. From the plot bars of **Figure 4**, it can be observed that MACI + DAN consistently outperforms DAN, while MACI + ReverseGrad is consistently better than ReveseGrad, which demonstrates that our MACI method is complementary to the two deep transfer learning methods DAN and ReveseGrad by exploiting the correlation statistics to further enhance the generalization ability across domains.

Parameter Impact on MACI

There are mainly three model parameters to be tuned in our method, i.e., λ, β, and α. Note that larger α would make the predicted label matrix better meet the expected needs, thus with better results being achieved. Consequently, we empirically set $\alpha = 10^3$ in the following experiments. We firstly explore to set the extreme values of different parameters for validating the importance of each component in our framework. Specifically, we denote MACI without the feature selection (i.e., $\beta = 0$) by MACI_NF and MACI with $\lambda = \eta_a = 0$ by MACI_NS, which ignores correlation information among multiple sources. These settings are evaluated on cross-dataset settings for multi-source adaptation tasks. From **Table 4**, we can observe that MACI can be significantly improved from MACI_NS by exploiting the correlation information among multiple sources. Besides this, the performance of MACI_NF is slightly weaker than MACI, that is, MACI would degrade when the feature selection function is omitted. A possible reason may be that the features of EEG represented by DE introduced some noise/outlier data. In this case, the feature selection in MACI possesses indispensable importance for robust DA learning. In sum, the utilization of correlation knowledge among sources and features could make

MACI further boost its performance in cross-dataset emotion recognition applications. It is this argument that constitutes the basic principle of our MACI framework.

CONCLUSION

In this work, we explore to cope with the cross-dataset emotion recognition where existing BCI methods cannot work well. To this end, we proposed an effective multi-source co-adaptation framework (MACI) for EEG-based emotion recognition mainly by leveraging correlation knowledge among sources and features in the objective function, which dampens unimportant evidence (within features and between sources) and amplifies useful knowledge. In MACI, multiple domain-invariant classification functions corresponding to different sources are co-learned by bridging both statistical and semantic distribution discrepancy between source and target domains, thus making MACI utilize the correlated knowledge among multiple sources by exploiting the developed correlation metric function. A large number of experimental results conducted on two publicly available EEG datasets show that MACI are much better than several representative baseline methods and provide the state-of-the-art performance on within/cross-dataset emotion recognition in most cases. This demonstrates the effectiveness of MACI in addressing feature distribution discrepancy between individual subjects as well as different datasets due to technical discrepancies.

To boost the efficiency of our method, however, a more efficient iterative algorithm would be developed or further elaborated in our future works. Besides this, the pseudo-labels strategy (i.e., iteratively updating target label matrix in the training stage) for bridging semantic distribution discrepancy between different domains would be unreliable or even misleading in training. This therefore arouses another challenge, i.e., how to effectively infer and incorporate target labels in unsupervised DA, which would be an urgent and valuable work in our future research.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported partly by High level talents introduction research project of Ningbo Polytechnic (under grant no. RC201902), partly by Zhejiang Provincial Natural Science Foundation of China (under grant no. LY19F020012), and partly by Foundation of Zhejiang Educational Committee (under grant no. Y201941140).

REFERENCES

Bruzzone, L., and Marconcini, M. (2010). Domain adaptation problems: a DASVM classification technique and a circular validation strategy. *IEEE Trans. PAMI* 32, 770–787. doi: 10.1109/tpami.2009.57

Chai, X., Wang, Q., Zhao, Y., Li, Y., Liu, D., Liu, X., et al. (2017). A fast, efficient domain adaptation technique for cross-domain electroencephalography (EEG)-based emotion recognition. *Sensors* 17:1014. doi: 10.3390/s17051014

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Comput. Biol. Med.* 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chen, B., Lam, W., Tsang, I. W., and Wong, T.-L. (2013). Discovering low-rank shared concept space for adapting text mining models. *IEEE Transact. Patt. Anal. Mach. Intell.* 35, 1284–1297. doi: 10.1109/tpami.2012.243

Chu, W. S., Del, T. F., and Cohn, J. F. (2017). Selective transfer machine for personalized facial action unit detection. *IEEE Transact. Patt. Anal. Mach. Intell.* 39, 529–545.

Ding, Z., Sheng, L., Ming, S., and Fu, Y. (2018). "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proceedings of the 15th European Conference (ECCV2018), Munich, Germany, September 8-14, 2018*, (Cham: Springer).

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194.

Duan, L., Dong, X., and Shih-Fu, C. (2012a). "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *Proceedings of the CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, (New York, NY: IEEE), 1338–1345.

Duan, L., Tsang, I. W., and Xu, D. (2012b). Domain transfer multiple kernel learning. *IEEE Transact. Patt. Anal. Mach. Intell.* 34, 465–479. doi: 10.1109/tpami.2011.114

Duan, L., Xu, D., and Tsang, I. W. (2012c). Domain adaptation from multiple sources: a domain-dependent regularization approach. *IEEE Transact. Neur. Netw. Learn. Syst.* 23, 504–518. doi: 10.1109/tnnls.2011.2178556

Ganin, Y., and Lempitsky, V. (2015). "Unsupervised domain adaptation by back-propagation," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Lavi-olette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 1–35. doi: 10.1109/tnnls.2020.3025954

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. (2017). Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Transact. Patt. Anal. Mach. Intell.* 99, 1–1.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). "A fast, consistent kernel two-sample test," in *Proceedings of the Conference on Neural Information Processing Systems 22*, (Vancouver, BC: MIT Press), 673–681.

Hou, C., Feiping, N., Hong, T., and Yi, D. (2017). Multi-view unsupervised feature selection with adaptive similarity and view weight. *IEEE Trans. Knowl. Data Eng.* 29, 1998–2011. doi: 10.1109/tkde.2017.2681670

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Comput. Intell. Magaz.* 11, 20–31.

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Transact. Aff. Comput.* 5, 327–339. doi: 10.1109/taffc.2014.2339834

Judy, H., Tzeng, E., Darrell, T., and Saenko, K. (2017). Simultaneous deep transfer across domains and tasks. *Dom. Adaptat. Comput. Vis. Appl.* 17, 173–187. doi: 10.1007/978-3-319-58347-1_9

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human EEG. *Comput. Mathematic. Methods Med.* 2013, 1–3. doi: 10.1155/2013/573734

Koelstra, S., Mühl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: a database for emotion analysis using physiological signals. *IEEE Transact. Affect. Comput.* 3, 18–31. doi: 10.1109/t-affc.2011.15

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Müller-Putz, G. R. (2018). Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets. *IEEE Transact. Cogn. Dev. Syst.* 11, 85–94. doi: 10.1109/tcds.2018.2826840

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018a). Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci.* 12:162.

Li, Y., Zheng, W., Cui, Z., Zhang, T., and Zong, Y. (2018b). "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, 1561–1567.

Li, Y., Zheng, W., Cui, Z., Zong, Y., and Ge, S. (2018c). EEG emotion recognition based on graph regularized sparse linear regression. *Neur. Proces. Lett.* 8, 1–17. doi: 10.1109/taffc.2020.2994159

Li, Z., Liu, J., Tang, J., and Lu, H. (2015). Robust structured subspace learning for data representation. *IEEE Transact. Patt. Anal. Mach. Intell.* 37, 2085–2098. doi: 10.1109/tpami.2015.2400461

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, 97–105.

Long, M., Wang, J., Ding, G., and Pan, S. J. (2014). Adaptation regularization: a general framework for transfer learning. *IEEE Transact. Knowl. Data Eng.* 26, 1076–1089. doi: 10.1109/tkde.2013.111

Lotfi, E., and Akbarzadeh, M.-R. (2014). Practical emotional neural networks. *Neur. Netw.* 59, 61–72. doi: 10.1016/j.neunet.2014.06.012

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). "Domain adaptation with multiple sources," in *Proceedings of the Conference on Neural Information Processing Systems*, (Vancouver, BC: MIT Press), 1041–1048.

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain Comput. Interfac.* 1, 66–84. doi: 10.1080/2326263x.2014.912881

Musha, T., Terasaki, Y., Haque, H. A., and Ivamitsky, G. A. (1997). Feature extraction from EEGs associated with emotions. *Artif. Life Robot.* 1, 15–19. doi: 10.1007/bf02471106

Nie, F., Huang, H., Cai, X., and Ding, C. (2010a). "Efficient and robust feature selection via joint -norms minimization," in *Proceedings of the International Conference on Neural Information Processing Systems*, (Vancouver, BC: Curran Associates Inc), 1813–1821.

Nie, F., Xu, D., Tsang, I. V.-H., and Zhang, C. (2010b). Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction. *IEEE Transact. Image Proces.* 19, 1921–1932. doi: 10.1109/tip.2010.2044958

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Transact. Neural Netw.* 22, 199–210. doi: 10.1109/tnn.2010.2091281

Pandey, P., and Seeja, K. (2019). "Emotional state recognition with EEG signals using subject independent approach," in *Data Science and Big Data Analytics. Lecture Notes on Data Engineering and Communications Technologies*, Vol. 16, eds D. Mishra, X. S. Yang, and A. Unal (Singapore: Springer), 117–124. doi: 10.1007/978-981-10-7641-1_10

Rosenstein, M. T., Marx, Z., and Kaelbling, L. P. (2005). "To Transfer or not to transfer," in *Proceedings of the Conference on Neural Information Processing Systems*, (Cambridge, MA: MIT Press).

Shi, L. C., Jiao, Y. Y., and Lu, B. L. (2013). "Differential entropy feature for EEG-based vigilance estimation," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, 6627–6630.

Simonyan, K., and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference Learning Representations (ICLR)*, Banff, 1–14.

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transact. Affect. Comput.* 18, 1–1.

Tao, J. W., Song, D., Wen, S., and Hu, W. (2017). Robust multi-source adaptation visual classification using supervised low-rank representation. *Patt. Recogn.* 61, 47–65. doi: 10.1016/j.patcog.2016.07.006

Tao, J., Chung, F. L., and Wang, S. (2012). On minimum distribution discrepancy support vector machine for domain adaptation. *Patt. Recogn.* 45, 3962–3984. doi: 10.1016/j.patcog.2012.04.014

Tao, J., Di, Z., Fangyu, L., and Bin, Z. (2019). Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models. *Patt. Recogn.* 87, 296–316. doi: 10.1016/j.patcog.2018.10.023

Tao, J., Wen, S., and Hu, W. (2015). L1-norm locally linear representation regularization multi-source adaptation learning. *Neural Netw.* 69, 80–98. doi: 10.1016/j.neunet.2015.01.009

Tao, J., Wen, S., and Hu, W. (2016). Multi-source adaptation learning with global and local regularization by exploiting joint kernel sparse representation. *Knowl. Based Syst.* 98, 76–94. doi: 10.1016/j.knosys.2016.01.021

Tommasi, T., Orabona, F., and Caputo, B. (2014). Learning categories from few examples with multi model knowledge transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 928–941. doi: 10.1109/tpami.2013.197

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial discriminative domain adaptation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2962–2971.

Wang, C., and Mahadevan, S. (2011). "Heterogeneous domain adaptation using manifold alignment," in *Proceedings of the International Joint Conference on Artificial Intelligence*, Barcelona, 1541–1546.

Yan, S., Xu, D., Zhang, B., Yang, Q., and Lin, S. (2006). Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Transact. Patt. Anal. Mach. Intell.* 29:40. doi: 10.1109/tpami.2007.250598

Yang, J., Yan, R., and Hauptmann, A. G. (2007). "Cross-domain video concept detection using adaptive svms," in *Proceedings of the ACM International Conference on Multimedia*, (New York, NY: ACM), 188–197.

Yang, Y., Ma, Z., Hauptmann, A. G., and Sebe, N. (2013). Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transact. Multim.* 15, 661–669. doi: 10.1109/tmm.2012.2237023

Zhang, K., Gong, M., and Schölkopf, B. (2015). "Multi-source domain adaptation: a causal view," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX, 3150–3157.

Zhang, Y., Chung, F. L., and Wang, S. (2019a). Takagi-sugeno-kang fuzzy systems with dynamic rule weights. *J. Intell. Fuzzy Syst.* 37, 8535–8550. doi: 10.3233/jifs-182561

Zhang, Y., Chung, F., and Wang, S. (2020a). Clustering by transmission learning from data density to label manifold with statistical diffusion. *Knowl Based Syst.* 193:105330. doi: 10.1016/j.knosys.2019.105330

Zhang, Y., Dong, J., Zhu, J., and Wu, C. (2019b). Common and special knowledge-driven TSK fuzzy system and its modeling and application for epileptic EEG signals recognition. *IEEE Access* 7, 127600–127614. doi: 10.1109/access.2019.2937657

Zhang, Y., Li, J., Zhou, X., Zhang, M., Ren, J., and Yang, J. (2019c). A view-reduction based multi-view TSK fuzzy system and its application for textile color classification. *J. Amb. Intell. Human. Comput.* 19, 1–11.

Zhang, Y., Tian, F., Wu, H., Xingyun, G., Danmin, Q., Jiancheng, D., et al. (2017). Brain MRI tissue classification based fuzzy clustering with competitive learning. *J. Med. Imag. Health Inform.* 7, 1654–1659. doi: 10.1166/jmihi.2017.2181

Zhang, Y., Wang, L., Wu, H., Geng, X., Yao, D., and Dong, J. (2016). A clustering method based on fast exemplar finding and its application on brain magnetic resonance images segmentation. *J. Med. Imag. Health Inform.* 6, 1337–1344. doi: 10.1166/jmihi.2016.1923

Zhang, Y., Wang, S., Xia, K., Jiang, Y., and Qian, P. (2020b). Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. *Inform. Fus.* 66, 170–183. doi: 10.1016/j.inffus.2020.09.002

Zheng, W. (2017). Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis. *IEEE Transact. Cogn. Dev. Syst.* 9, 281–290. doi: 10.1109/tcds.2016.2587290

Zheng, W. L., and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transact. Autonom. Mental Dev.* 7, 162–175. doi: 10.1109/tamd.2015.2431497

Zheng, W. L., and Lu, B. L. (2016). "Personalizing EEG-based affective models with transfer learning," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, Palo Alto, CA, 2732–2738.

Zheng, W. L., Zhang, Y. Q., Zhu, J. Y., and Lu, B. L. (2015). "Transfer components between subjects for EEG-based emotion recognition," in *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, 917–922.

Zhong, P., Wang, D., and Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Transact. Affect. Comput.* 99, 1–1.

Zhou, X., Jin, K., Shang, Y., and Guo, G. (2018). Visually interpretable representation learning for depression recognition from facial Im-ages. *IEEE Transact. Affect. Comput.* 11, 542–552. doi: 10.1109/TAFFC.2018.2828819

Zhu, Y., Shang, Y., Shao, Z., and Guo, G. (2017). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transact. Affect. Comput.* 9, 578–584. doi: 10.1109/TAFFC.2017.2650899 doi: 10.1109/taffc.2017.2650899

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.