Knowledge-Based Systems 156 (2018) 43–61

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Sparse feature space representation: A unified framework for semi-supervised and domain adaptation learning

Liu Long[a], Yang Lechao[a], Zhu Bin[b,*]

[a] School of Automation and Information Engineering, Xi'an University of Technology Xi'an, 710048, China
[b] College of Communication Engineering, Chongqing University, Chongqing 400040, China

ARTICLE INFO

Keywords:
Domain adaptation
Sparse representation
Laplacian regularization
Feature space embedding

ABSTRACT

In a semi-supervised domain adaptation (DA) task, one has access to only few labeled target examples. In this case, the success of DA needs the effective utilization of a large number of unlabeled target data to extract more discriminative information that is useful for generalization. To this end, we exploit in this paper the feature space embeddings of the target data as well as multi-source prior models to augment the discrimination space for the target function learning. Therefore, we propose a novel multi-source adaptation learning framework based on Sparse Feature Space Representation (SFSR), or called SFSR-MSAL for short. Specifically, the SFSR algorithm is first presented for the further construction of robust graph, on which the discriminative information can be smoothly propagated into the unlabeled target data by additionally incorporating the geometric structure of the target data. Considering the robustness in the semi-supervised DA, we replace the traditional l_2-norm based least squares regression with the $l_{2,\ 1}$-norm sparse regression, and then construct the SFSR-graph based semi-supervised DA framework with multi-source adaptation constraints. Our framework is universal and can be easily degraded into semi-supervised learning by just tuning the regularization parameter. Moreover, to select the discriminative SFSR-graph Laplacians, we also introduce the ensemble SFSR-graph Laplacians regularization into SFSR-MSAL, thus further improving the performance of SFSR-MSAL. The validity of our methods including semi-supervised and DA learning are examined by several visual recognition tasks on some benchmark datasets, which demonstrate the superiority of our methods in comparison with other related state-of-the-art algorithms.

1. Introduction

Traditional learning tasks usually require large amounts of training data to establish a classifier with satisfactory generalization capability. However, the acquisition of labeled training data is usually nontrivial for one learning task in that it needs to manually annotate input data with ground-truth labels by experts, which is often difficult, expensive, and time-consuming [1]. In the past decades, semi-supervised learning (SSL) has become a classical paradigm by exploiting a large number of unlabeled data [2]. The success of SSL is mainly attributed to certain assumptions that holds for the data distribution (also referred to as a cluster or a manifold assumption [3]), which considers both local and structure smoothness of the data distribution [4]. In particular, the manifold assumption has been applied for regularization where the geometric structure behind labeled and unlabeled data is explored with a graph-based representation [3]. In such a representation, examples are expressed as the vertices and the pairwise similarity between data is described as a weighted edge. Thus, graph-based SSL (GSSL) algorithms make good use of the manifold structure to propagate the known label information over the graph into the unlabeled data [5,8,11,12].

Although most of existing GSSL algorithms have shown promising achievements in their specific applications, there still exist several limitations in them, e.g., performance sensitivities to model parameters and noise data [8]. Besides, while exploiting the vast amount of unlabeled data directly in the GSSL paradigm is valuable in its own right, it is still beneficial to leverage a plenty of labeled data of relevant categories across domains. For example, it is increasingly popular to enrich our limited collection of training data with those from the Internet. One problem with this strategy, however, arises from the possible misalignment between the target domain of interest and the auxiliary (or source) domain that provides some prior information. This makes it harmful to directly incorporate data from the source domain into the target domain. Addressing this issue has inspired recent research efforts into the domain adaptation (DA) problems [14,15] in computer vision and machine learning [16]. Recently, many cross-domain learning techniques have been proposed to solve the problem of distribution mismatch in the field of image or video concept detection and classification [17–29]. Depending on how the source information is

* Corresponding author.
E-mail addresses: zhubin@cqu.edu.cn, bin_zhu@aliyun.com (B. Zhu).

https://doi.org/10.1016/j.knosys.2018.05.011
Received 5 January 2018; Received in revised form 8 May 2018; Accepted 9 May 2018
Available online 22 May 2018

exploited, the division of DA methods is between model/classifier-centric [13,14,17–23,43,45–47] and representation-centric adaptation [24–29,42]. The former advocates implicit adaptation to the target distribution by adjusting a classifier from the source domain; and the latter attempts to achieve alignment by adjusting the representation of the source data via learning some transformation. Orthogonal to this, the extant methods can also be classified into semi-supervised (e.g., [13], and [18–20]) and unsupervised (e.g., [17] and [21–27]) DA, based on whether target labels have been exploited during the adaptation. In this paper, we mainly focus on the classifier-centric semi-supervised DA methodology.

In some specific visual recognition tasks, one has access only to a small number of training data with few labeled samples. In this scenario, there still exists an unsolved challenge in DA and SSL, i.e., how to effectively utilize the auxiliary prior knowledge as well as a large number of unlabeled data to extract more discriminative evidence for generalization. Towards this end, we explore to augment the discriminative space of the target domain by sparsely reconstructing/representing each target data with their feature space projections [7], respectively. We would like to ensure that the solution is robust and smooth with respect to both the ambient feature space and the target label space by considering sparse reconstruction measure [10]. We therefore propose a novel Sparse Feature Space Representation (SFSR) regularization multi-source adaptation framework, called SFSR-MSAL for short. To our best knowledge, there are no GSSL algorithms with SFSR regularization working on multi-source DA scenarios in literature. The main contributions of this paper are highlighted as follows.

(1) To augment the discrimination space of the target data, the SFSR algorithm is originally presented for robust graph construction using the sparse reconstruction measure. SFSR could have more robust reconstruction capacity when data is contaminated by some noises or outliers.
(2) By introducing the SFSR-graph Laplacian regularization, we construct a new semi-supervised DA framework with multi-source adaptation constraints. Moreover, to select the discriminative SFSR-graph Laplacians, we also introduce the ensemble SFSR-graph Laplacians regularization into the framework of SFSR-MSAL.
(3) Our proposed framework provides a unified view to explain and understand robust GSSL with multi-source adaptation, including semi-supervised learning (or SFSR-GSSL for short), and multi-source adaptation techniques.
(4) Based on the tool of transductive Rademacher complexity [30], a generalization error bound for our framework is derived. A serial of experiments on several real-world datasets demonstrate promising performance by using the proposed models.

The remainder of this paper is organized as follows. In Section 2, we briefly review some related works. In Section 3 we propose the SFSR algorithm. Section 4 details our SFSR-MSAL framework and its optimization algorithm. In Section 5, we further discuss about ensemble SFSR-graph Laplacians, generalization error bound of SFSR-MSAL, and selection of source weights. Experimental results on several real-world datasets with SFSR-GSSL and SFSR-MSAL are respectively reported in Section 6 involving parameter selection and results analysis. Finally, Section 7 concludes the whole paper.

2. Brief review of prior works

2.1. Notations

In the following, we denote by $A \in \mathbb{R}^{d\times n}$ a matrix of size $d \times n$ with A_{ij} corresponding to the (i, j) element in$A$. $(\cdot)^T$ denotes the transpose of a vector or matrix $(\cdot)$. When only one subscripted index is present (e.g.,A_i), it denotes the column index of a matrix. Moreover, we indicate with $\|A\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{d} A_{ij}^2}$ the $l_{2,\ 1}$-norm, and with $\|A\|_F$ the Euclidean norm of the matrix A, respectively. The trace of a matrix A is represented as $tr(A)$. Let us also define I_n as the $n \times n$ identity matrix and $1_n \in \mathbb{R}^n$ as the column vectors of all ones, respectively. Given a vector $u = [u_1, ...,u_n]^T$, the inequality $u \geq 0$ means $u_i \geq 0$ for $i = 1, 2, ...,n$.

In DA tasks, the target domain of interest is denoted by a dataset $X = [x_1, ...,x_n] \in \mathbb{R}^{d\times n}$, and $Y = [y_1, y_2, ...,y_n]^T \in \{0, 1\}^{n\times c}$ is the corresponding label matrix with $y_{ij} = 1$ if x_i is labeled as $y_i = j$ and $y_{ij} = 0$ otherwise, where d is the feature dimension, n is the total number of target samples, and c is the number of classes. Let us further respectively represent the labeled and unlabeled instances from the target domain as $X_l = [x_1, ...,x_{n_l}] \in \mathbb{R}^{d\times n_l}$ and $X_u = [x_{n_l+1}, ...,x_n] \in \mathbb{R}^{d\times n_u}$, where n_l (resp. n_u) is the number of labeled (resp. unlabeled) samples and $n_l \ll n_u$, i.e., $n = n_l + n_u$. Correspondingly, we respectively define their class label matrix as $Y_l = [y_1, y_2, ...,y_{n_l}]^T$ and $Y_u = [y_{n_l+1}, ...,y_n]^T$which is a matrix with all zeros, i.e., $Y = [Y_l, Y_u]^T$. We also denote by $X^s = \{x_i^s\}_{i=1}^{n_s} \in \mathbb{R}^{d\times n_s}$ the dataset of size n_s from the sth $(s = 1, 2, ...,a)$ source domain and by $Y^s = [y_1^s, ...,y_{n_s}^s]^T \in \{0, 1\}^{n_s\times c}$ the corresponding label matrix. We assume that both the feature spaces and label spaces between domains are the same, and all source domains are independent of each other.

2.2. Graph-based semi-supervised learning

In the past several decades, one has witnessed a prominent development of the graph-based semi-supervised learning (GSSL) strategy [5,40]. Given a target dataset X, GSSL models the whole dataset in the form of an undirected graph $G = (X, S)$ with the vertex set X and the edge weight set $S \in \mathbb{R}^{n\times n}$, in which each element S_{ij} of the symmetric matrix S reflects the similarity between x_i and x_j on the graph as $S_{ij} = \exp(-\|x_i - x_j\|/\sigma^2)$ if x_i and x_j are k-nearest neighbors, $S_{ij} = 0$ otherwise. The graph Laplacian matrix $L \in \mathbb{R}^{n\times n}$ is denoted as $L = D - S$, where D is a diagonal matrix with the diagonal elements as $D_{ii} = \sum_j S_{ij}, i = 1, ...,n$. Generally, there are two types of GSSL tasks, i.e., transductive GSSL [4,8,11,12] and inductive GSSL [3,32].

The transductive GSSL aims at predicting the labels of the unlabeled vertices only. The representative transductive GSSL methods include [2,4,8]. In [8], the linear neighborhood propagation (LNP) is proposed based on the basic assumption of local linear embedding (LLE) [9]. Besides, Hong et al. [11] and Fan et al. [12] respectively put forward sparse representation based graph regularization SSL methods, which are motivated by the idea that the sparse representation of visual data has natural discriminative ability [10]. In nature, most of graph-based regularization algorithms are transductive, although they can be converted into inductive algorithms with the out-of-sample extension [31].

The inductive GSSL tries to induce a decision function that has a low error rate on the whole sample space. For example, Belkin et al. [3] propose a general Manifold Regularization (MR) framework based on the assumption that data lies on an intrinsic low-dimensional manifold. It is shown in [32] that MR is a varied out-of-sample extension of LGC [4] when a graph Laplacian matrix L satisfying $L1_n = 0$ and $1_n^T L = 0^T$ is used. Recently, Nie et al. [32] propose a novel manifold learning framework termed Flexible Manifold Embedding (FME) for multi-class setting by modeling a regression residue, which naturally unifies many existing GSSL methods. Specifically, FME sets the prediction labels as $F = H(X) + F_0$, where $H(X)$ is a regression function for mapping new data and F_0 is the regression residue modeling the mismatch between F and $H(X)$. Therefore, FME aims to find the optimal prediction labels F, the linear regression function $H(X)$ and the regression residue F_0 simultaneously.

2.3. Classifier-centric domain adaptation

In practical applications, we often have a lot of annotated data

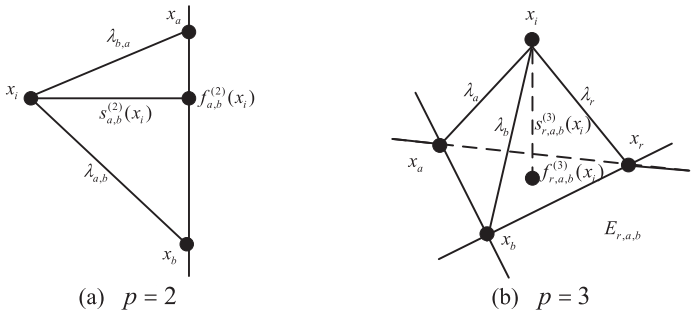


Fig. 1. Feature space projections and the weight setting for (a) $p = 2$and (b) $p = 3$.

supplied from other related auxiliary/source domains while need to solve a different target problem with few or even none of labeled data, where both auxiliary and target domains present a distribution disparity. In this scenario, the classifier-centric DA scheme [1,15,22] is proposed to decrease the effort of collecting new training samples, and at the same time reduce the risk of over-fitting by leveraging some existing prior knowledge to solve a target task. This scheme usually aims to directly design an adaptive source and/or target classifier(s) by incorporating the adaptation of different distributions or classifiers through model regularization. Recently, a large number of classifier-centric DA methods have been proposed in visual recognition tasks based on different assumptions of the source and target domains [18–20,42–44].

One line of classifier-centric DA researches focuses on the assumption that one has access to the enough number of labeled source samples as well as a large number of unlabeled target samples. This line can be subdivided into two categories, i.e., Classifier Induction with Subspace Learning (CISL), and Classifier Induction with Distribution Adaptation (CIDA). CISL simultaneously extracts a shared subspace and trains a supervised [47,48] or semi-supervised classifier [24,28,46] in this subspace, while CIDA directly integrates the minimization of distribution discrepancy as a regularizer into the classic SVM model [21–23,41]. Since the adaptive classifiers in these methods are found by exploiting the knowledge from both source(s) and target domains based on the accessibility of two or multiple different domains, their adaptation and discriminability could be captured more precisely, and therefore they are usually more accurate than other state-of-the-art methods in certain specific DA applications [45]. However, the computational complexity of CISL or CIDA is relatively high which could make these methods inapplicable for some visual applications with large-scale source samples. To this end, there exist another line of classifier-centric DA works based on a different assumption that one has access only to the target domain with a few labeled target samples, i.e., Classifier Induction with Model Adaptation (CIMA) or model adaptation for simplicity, which just leverages the prior source models previously trained on related sources for adaptation [18–20]. Our proposed framework belongs to this subcategory with substantial extensions.

Recently, along the line of the researches in [18] and [20], we have also proposed several diverse model adaptation strategies [42–44] by exploiting the superiorities of the sparse and low-rank representation. While the variations of model adaptation methods are limited and most of them apply adaptation in the original space of the source and target domains [45], they are very important because of their low time and space complexities [45] and they are well suited for multi-source adaptation tasks [18–20,42–44] where multiple prior models can be easily accessed.

As can be seen from our work, the works of Tomassi et al. [20] and Duan et al. [19] are somewhat close to our method since they also use a discriminative SVM framework and include the previously learnt models as a regularizer. However, the distinction between ours and those related is that we explicitly consider to augment the discrimination space of target domain and exploit the intrinsic discriminative structure of the target distribution by modeling it as a feature space representation based sparse reconstruction problem, thus further learning with both labeled and unlabeled dataset a robust classifier in the target domain. Besides, those related methods would place heavy reliance on the labeled dataset from the target domain [20], and lack of robustness on real datasets in which noise/outliers may abound. Instead we particularly introduce a predicted label matrix and an $l_{2,\,1}$-norm least squares regression term [31] in the proposed framework to avoid these issues. Therefore, our aim is to extend the mentioned above related works by developing a new model and also by considering the more challenging robust GSSL problem.

3. Sparse feature space representation

Given the target dataset$X \in \mathbb{R}^{d\times n}$ that is sampled from some fixed but unknown distribution$p(x)$, the basic procedure of the SFSR method includes: (1) calculation of feature space projections w.r.t the original target data; (2) sparsely reconstructing the original target data by their corresponding feature space projections, thus obtaining a sparse reconstruction matrix S for constructing a weighted graph $G(X, S)$.

3.1. Calculation of feature space projection

For the target dataset$X = \{x_1, x_2, ...,x_n\}$, the distance between any object $x_i \in X$ and its feature space projection is:

$$\|x_i - f^{(p)}(x_i)\|^2, \tag{1}$$

where $f^{(p)}$ is the feature space spanned by p ($p = 1, 2, ...,n - 1$) objects and $f^{(p)}(x_i)$ is the projection (or embedding) of object x_i in feature space$f^{(p)}$. Fig. 1 shows the feature space projections of the object x_i and the corresponding weights in the case of $p = 2$ and $p = 3$, respectively. When $p = 1$, formula (1) shows the distance measure of point pairs in the original feature space. Next, we will mainly focus in this paper on the calculation of $f^{(p)}(x_i)$ w.r.t the object x_i when $p \geq 2$.

(1) When $p = 2$, formula (1) changes into the distance measure between x_i and the feature line $\overline{x_a x_b}$, which is represented as $\|x_i - f^{(2)}_{a,b}(x_i)\|$, and the projection point can be calculated as:

$$f^{(2)}_{a,b}(x_i) = x_a + \pi_{a,b}(x_a - x_b), \tag{2}$$

where, the weights $\pi_{a,b} = (x_i - x_a)^T(x_a - x_b)/(x_a - x_b)^T(x_a - x_b)$ with $\pi_{a,b} + \pi_{b,a} = 1$ ($i \neq a \neq b$). We then have

$$\begin{aligned} x_i - f^{(2)}_{a,b}(x_i) &= x_i - x_a + \pi_{a,b}(x_a - x_b) \\ &= x_i - (1 - \pi_{a,b})x_a - \pi_{a,b}x_b = x_i - \pi_{a,b}x_a - \pi_{b,a}x_b \end{aligned}. \tag{3}$$

(2) When $p = 3$, formula (1) changes into the distance measure between the point x_i and the feature face $E_{r,\,a,\,b}$, which is represented as$\|x_i - f^{(3)}_{r,a,b}(x_i)\|$. The feature face $E_{r,\,a,\,b}$ consists of objectsx_r,x_a and

x_b(shown as Fig. 1(b)). The projection $f^{(3)}_{r,a,b}(x_i)$ of the feature face$E_{r,a,b}$, is calculated by formula (4):

$$\begin{aligned} f^{(3)}_{r,a,b}(x_i) &= \Gamma_{r,a,b}(\Gamma^T_{r,a,b}\Gamma_{r,a,b})^{-1}\Gamma^T_{r,a,b}(x_i - x_r) + x_r \\ &= [(x_a - x_r) \quad (x_b - x_r)][\pi_a \quad \pi_b]^T + x_r \\ &= (1 - \pi_a - \pi_b)x_r + \pi_a x_a + \pi_b x_b \\ &= \lambda_r x_r + \pi_a x_a + \pi_b x_b \end{aligned}, \tag{4}$$

where $\Gamma_{r,a,b} = [(x_a - x_r) \quad (x_b - x_r)]$ is a matrix of $r \times 2$. The matrix $(\Gamma^T_{r,a,b}\Gamma_{r,a,b})^{-1}\Gamma^T_{r,a,b}(x_i - x_r)$ is represented as $[\pi_a \quad \pi_b]^T$, and$\pi_r + \pi_a + \pi_b = 1$. From Eq. (4), we can see that the projection $f^{(3)}_{r,a,b}(x_i)$ is represented as the linear combination of objectsx_r,x_a andx_b.

(3) Generally speaking, formula (1) changes into the distance measure between the point x_i and the feature space for$p > 3$, which is represented as$x_i - f^{(p)}(x_i)$. The projection$f^{(p)}(x_i)$ in the feature space $f^{(p)}$ is calculated through formula (5).

$$f^{(p)}(x_i) = \Gamma_{1:p}(\Gamma^T_{1:p}\Gamma_{1:p})^{-1}\Gamma^T_{1:p}(x_i - x_1) + x_1 = \sum_{j=1}^{p} \pi_j x_j, \tag{5}$$

where $\Gamma_{1:p} = [x_2 - x_1 \quad x_3 - x_1 \quad \dots \quad x_p - x_1]$ and $\sum_{j=1}^{p} \pi_j = 1$.

3.2. Construction of SFSR graph

In GSSL, to construct an informative graph for effective label propagation is crucially important. Unfortunately, for the original high-dimensional data (such as facial image data), the nearest neighbor principles, commonly employed in the traditional GSSL methods, cannot attain good performance [7], since it is very difficult to choose the optimal model parameters when just a few labeled data are accessible [11]. Therefore, we need to find a more reliable and stable method to construct the graph model. The recent researches have proven that sparse representation (SR) [10] has natural discriminative ability and can get better performance under high-dimensional data space [10]. Furthermore, the discriminative ability only relates to the number of object class tightly but not relates to the number of samples. Therefore, we construct the graph model based on the theory of SR so as to contain more discrimination information in the case of learning with limited labeled data.

Denote by $T^{(p)}$ a feature space set of size C_p^{n-1} generated by the dataset X, and by $T^{(p)}(x_i)$ the project set of $x_i \in X$ $(i = 1, \dots, n)$ w.r.t $T^{(p)}$. We try to avoid the original point pair measure in the traditional GSSL and use $T^{(p)}(x_i)$ to sparsely reconstruct x_i, thus a sparse reconstruction matrix S being obtained. We then construct the weighted graph $G = \{T^{(p)}, S\}$, where $s^{(p)}(x_i) \in S$ is the vector of sparse reconstruction coefficients between x_i and its feature space project set $f^{(p)}(x_i) \in T^{(p)}(x_i)$. In particular, the sparse reconstruction coefficients vector w.r.t x_i can be computed by the following minimization problem [10]:

$$\min \ \|x_i - s^{(p)}(x_i)f^{(p)}(x_i)\|^2 + \widetilde{C}\,\|s_i^{(p)}\|_1, \ i = 1, \dots, n, \tag{6}$$

where $\widetilde{C}$ is the regularization parameter to control the balance between the sparsity and reconstruction error. $s_i^{(p)} = \left[s_{i1}^{(p)}, s_{i2}^{(p)}, \dots, s_{i(j-1)}^{(p)}, 0, s_{i(j+1)}^{(p)}, \dots, s_{iC_p^{n-1}}^{(p)}\right]^T$ is a C_p^{n-1} dimensional column vector, in which the ith element is equal to zero, implying that the feature space project point $\widetilde{x}_j \in T^{(p)}(x_i)$ is removed from$T^{(p)}(x_i)$, and $s_{ij}(j \neq i)$ denotes the contribution of $\widetilde{x}_j$ for reconstructing x_i. We further constrain $\sum_j s_{ij}^{(p)} = 1$ and $s_{ij}^{(p)} \geq 0$. When $p = 2$ and$p = 3$, formula (6) changes respectively into:

$$\min \ \left\|x_i - \sum_{i \neq a \neq b} s^{(2)}_{a,b}(x_i) f^{(2)}_{a,b}(x_i)\right\|^2 + \widetilde{C}\,\|s_i^{(2)}\|_1, \tag{7}$$

$$\min \ \left\|x_i - \sum_{i \neq r \neq a \neq b} s^{(3)}_{r,a,b}(x_i) f^{(3)}_{r,a,b}(x_i)\right\|^2 + \widetilde{C}\,\|s_i^{(3)}\|_1, \tag{8}$$

where, $s_i^{(2)}$ and $s_i^{(3)}$ are the $C_2^{n-1}$and C_3^{n-1}dimensional column vectors of the sparse weights $s^{(2)}_{a,b}(x_i)$ and $s^{(3)}_{r,a,b}(x_i)$, respectively.

We can easily extend the feature-sign search algorithm in [33] to solve the sparse coding problem in (6). After obtaining the entire optimal reconstruction coefficient vectors $\widehat{\mathbf{s}_i^{(p)}}(1 \leq i \leq n)$ for each datax_i, a sparse weight matrix $S = \left[\widehat{\mathbf{s}_1^{(p)}}, \widehat{\mathbf{s}_2^{(p)}}, \dots, \widehat{\mathbf{s}_n^{(p)}}\right]$ can be constructed, and based upon which we can obtain the newly constructed graph $G = \{X, S\}$. Note that the element $s_{ij}^{(P)}$ in S isn't the simple similarity measure between x_i and $\widetilde{x}_j$, which is essentially different from the weight in the traditional graph regularization algorithms. In the matrix S, each weight vector $\widehat{\mathbf{s}_i^{(p)}}$ obeys an important symmetry, i.e., it is invariant to rotations and invariant to translations due to the constraint $1^T_{C_p^{n-1}}\widehat{\mathbf{s}_i^{(p)}} = 1$.

According to the manifold assumption [9], if two samples in the original feature space are close to each other, then their embeddings in the low-dimensional manifold space should also be “close” to each other. We thereby expect that the desirable characteristics in the class label space can be preserved by sparse representation as that in the feature space [12]. Denote by $F = (f_1, f_2, \dots f_c) \in \mathbb{R}^{n \times c}$ the class assignment matrix, where $f_j = [f_j(x_1), f_j(x_2), \dots f_j(x_n)]^T$ $(1 \leq j \leq c)$ with $f_j(\cdot)$ being the jth classifier in the case of multi-class classification. We thus can construct the sparse representation regularization in the label space by minimizing the following object functions, which preserve the sparse reconstruction vector of the feature space representation w.r.t $x_i(1 \leq i \leq n)$:

$$\Omega_1(F) = \sum_i \left\| \sum_{j:L^{(p)} \in F^{(P)}} (f_i - L^{(p)}(f_i))s_j^{(p)}(x_i) \right\|^2, \tag{9}$$

$$\Omega_2(F) = \sum_i \sum_{j:L^{(p)} \in F^{(p)}} \|f_i - L^{(p)}(f_i)\|^2 s_j^{(p)}(x_i), \tag{10}$$

where $s_j^{(p)}(x_i)$ is the sparse representation weight for the jth feature space representation, $L^{(p)}$ is the label space spanned by p label vectors, $L^{(p)}(f_i)$ is the projection of f_i in the label space $L^{(p)}$, and $F^{(p)}$ is C_p^{n-1} possible label space sets generated by the original label sets. Formula (9) and (10) are the sparse label space representation functions. The difference between $\Omega_1(F)$ and $\Omega_2(F)$ is the sequence of “add” operations.

Theorem 1. *The object functions $\Omega_1(F)$ and $\Omega_2(F)$ can be unified as the following Laplacian regularization formulation.*

$$\Omega(F) = tr(FLF^T), \tag{11}$$

where $L = L_1$ or $L = L_2$.

Proof. The detailed procedure of proof for Theorem 1 is shown in Appendix A.

4. Proposed framework

4.1. Problem statement

For a multi-class learning task, we may learn c prediction functions (classifiers) $\{f_i(\cdot)\}_{i=1}^c$ based on the SFSR-graph $G = \{S, X\}$ constructed in Section 3. This learning process can be formalized as an optimization problem for finding an optimal classifier f in the hypothesis space of function H, which minimizes the following structure risk:

$$\Omega(\|f\|_H) + C\sum_{i=1}^{n} loss(f(x_i), y_i). \tag{12}$$

Here $\Omega(\|f\|_H)$ is a regularizer, which encodes some notion of smoothness for f, and guarantees good generalization performance avoiding overfitting. In the second term, $loss(\cdot)$ is some convex non-negative loss function which assesses the quality of the function f on the

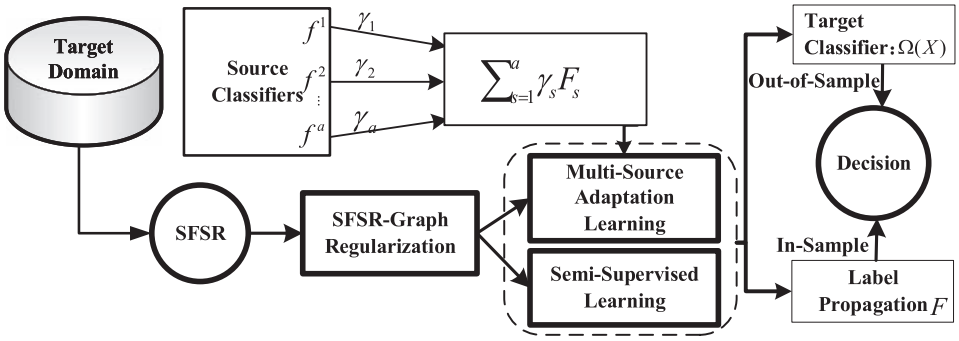


Fig. 2. Flowchart of the unified framework of SFSR-MSAL.

stance and label pair $\{x_i, y_i\}$. The predictivity is a trade-off between fitting the training data and keeping the complexity of the solution low, controlled by the parameter $C > 0$. For simplicity, we set H equal to space of all linear models of the form $f(x_i) = Q^T x_i$, where Q is the classification model.

However, this classical learning scheme in (12) fails to take other source knowledge gained from the prior source domain X^s ($s = 1, 2, ...,a$) extracted from a distribution different with that of target domain into consideration. We therefore propose a novel SFSR-graph regularization Multi-Source Adaptation Learning framework (SFSR-MSAL), where we integrate multi-source DA and GSSL into a unified framework for target task. The flowchart of the overall framework is illustrated in Fig. 2. Note that our framework is inductive, making it applicable to the large-scale dataset which may grow dynamically. Next, we will detail the formulation of the proposed framework.

4.2. Formulation of proposed framework

In DA tasks, one of main challenges is to minimize the domain discrepancy. To this end, a common-used strategy is to re-represent the domain features in some optimal subspace, thus reducing the domain distance. Therefore, we present in this paper to simultaneously learn an appropriate target model and an optimal latent space. It is reasonable to assume that the latent space can be presented by a linear transformation matrix $P \in \mathbb{R}^{d\times r}$, where r is the dimensionality of the latent space. Thus, the learning problem (12) becomes:

$$\Omega(f) + C\sum_{i=1}^{n} loss(Q^T(P^T x), y_i). \tag{13}$$

Moreover, we additionally introduce a predicted label matrix $F \in \mathbb{R}^{n\times c}$ into the predictive functions, where each row corresponds to the predicted label vector of each target data. This predicted label matrix should satisfy: (1) The predicted labels are supposed to preserve the sparse representation relationship on the SFSR-graph; (2) The predicted labels should be globally consistent, i.e., they should be consistent with the ground-truth labels; and (3) The predictive functions should be robust to outliers and noise. To satisfy the first and second attributes, we introduce a smooth regularization on the latent geometric structure between target samples with the proposed sparse label space representation, which is formulated as

$$tr(FLF^T) + tr((F - Y)^T U(F - Y)), \tag{14}$$

where $L = E^{-1/2}(E - S)E^{-1/2}$ is the normalized SFSR-graph Laplacian, E is a diagonal matrix with $E_{ii} = \sum_j^n S_{ij}$, and $U \in \mathbb{R}^{n\times n}$ is a diagonal matrix, whose diagonal element $U_{ii} = \tau$ (τ is a large constant) if x_i is a labeled sample and $U_{ii} = 0$ otherwise. Usually we can employ the classical least squares loss function in (13), which is always used to learn the predictive functions [4,5,8,12]. However, it is sensitive to outliers and noise. As indicated in [31], the $l_{2,1}$-norm loss function can achieve more robust performance when the target data are contaminated by noise/outliers. Hence, while other loss functions, e.g., hinge loss or logistic loss, may be used as well, we use the $l_{2,1}$-norm loss function in our framework due to its robustness. Therefore, we further propose the following objective function to satisfy the third attribute.

$$\|X^T PQ - F\|_{2,1} + \alpha \|PQ\|_{2,1},\ s.\ t.\ P^T P = I_r, \tag{15}$$

where α is a nonnegative regularization parameter. By jointly optimize the above objectives (14) and (15), we obtain the following objective function.

$$\begin{aligned} &\arg\min_{P,Q,F} \|X^T PQ - F\|_{2,1} + \alpha \|PQ\|_{2,1} + tr(FLF^T) \\ &\quad + tr((F - Y)^T U(F - Y)) \\ &s.\ t.\ P^T P = I_r \end{aligned} \tag{16}$$

Given the ith sample x_i in the target domain, we denote its labels predicted respectively by the target and the sth source classifiers as $f_i = f(x_i)$ and $f_i^s = f^s(x_i)$, where f^s represents the sth source classifier function. We thus have $F = [f_1, ...,f_n]^T$, and $F_s = [f_1^s, ...,f_n^s]^T$ which denotes the target label matrix predicted by the sth source classifier. To exploit prior source knowledge for target tasks, it is reasonable to assume that the prediction label matrix from the target domain can be represented by those from some related source domains. In other words, we assume that the target classifier $f(x)$ should have similar decision values on the unlabeled samples with that predicted by the pre-computed source classifiers. Hence, the final label of each target sample can be predicted by minimizing the objective $\Theta_s(F) = \sum_i^n \|f_i - \sum_{s=1}^a \gamma_s f_i^s\|^2$. The parameter γ_s is defined as the weight for measuring the distribution relevance between the sth source domain and the target domain.[1] If the sth source domain and the target domain are relevant, i.e., γ_s is large, we enforce that f_i^s should be close to f_i on the unlabeled instances in the target domain.

To exploit certain prior knowledge in (16), we propose to simultaneously learn the SFSR-graph regularization prediction label matrix and the target classifier in a unified framework by leveraging multi-source prior models, which is further formulated as:

$$\begin{aligned} &\arg\min_{P,Q,F} \|X^T PQ - F\|_{2,1} + \alpha \|PQ\|_{2,1} + tr(FLF^T) \\ &\quad + tr((F - Y)^T U(F - Y)) + \lambda\Theta_s(F) \\ &s.\ t.\ P^T P = I_r \end{aligned} \tag{17}$$

where $\lambda \geq 0$ is a regularization parameter. Note that $\sum_i^n \|f_i - \sum_{s=1}^a \gamma_s f_i^s\|^2 = \|F - \sum_{s=1}^a \gamma_s F_s\|^2$. We further let $W = PQ$ be the classifier model for the target domain, then our objective can be rewritten as

$$\begin{aligned} J = \arg\min_{W,F,Q,P^T P==I_r} &\|X^T W - F\|_{2,1} + \alpha \|W\|_{2,1} + \beta \|W - PQ\|_F^2 \\ &+ tr(FLF^T) + tr((F - Y)^T U(F - Y)) + \lambda \left\| F - \sum_{p=1}^{a} \gamma_s F_s \right\|^2, \end{aligned} \tag{18}$$

[1] More discussions on γ_s can be found later.

where β is a regularization parameter, which can tune the closeness between W and PQ.

As we analyzed above, the new SFSR contains much useful discrimination information of domain data. In (18), the objective function J is convex, and the loss function $\|X^TW - F\|_{2,1}$ is robust to outliers and noise [31]. The graph regularization term makes learning procedure preserve the discriminative structures of domains data. The $l_{2,1}$-norm regularization of matrix $\|W\|_{2,1}$ added on projection matrix W forces most of the rows in W shrink to zero, which implies the corresponding features of these zero rows are not important to the target learning. Therefore, we can perform feature selection technique [54] on original domains data to get the efficient discriminative model. We rank the rows in the W in descending order according to l_2-norm values of each row and then select the top-ranked rows as the results of feature selection.

In (18), by simply setting the regularization parameter λ equal to 0, the semi-supervised version of the proposed framework can be easily recovered as follows.

$$J(F, W, P, Q) = \min_{P^TP==I} \|X^TW - F\|_{2,1} + \alpha \|W\|_{2,1} + \beta \|W - PQ\|_F^2 + tr(FLF^T) + tr((F - Y)^TU(F - Y)), \tag{19}$$

which is termed as SFSR-GSSL in the sequence. It is worth mentioning that the framework shown in (19) is intrinsically different from the traditional graph regularization framework proposed in [8] and [32] which has been frequently used in previous transductive graph-based algorithms. Besides, facing robust multi-source adaptation problems, our proposed framework (18) has completely different performance compared with the art in [18]. These are attributed to the SFSR-graph regularization as well as $l_{2,1}$-norm least squares regression in our framework.

4.3. Optimization algorithm

In this section, we give an iterative approach to optimize the objective function in (18) (the same procedure for the objective function in (19) by just setting $\lambda = 0$), i.e., a block coordinate descent method is adopted to iteratively update each optimal variable in (18). While a similar optimization approach has been proposed in [31] for matrix completion, we are focusing on a different problem which jointly optimizes trace norm and $l_{2,1}$-norm. Specifically, by denoting $X^TW - F = [z_1, ..., z_n]^T$ and $W = [w_1, ..., w_d]^T$, the optimization problem in (18) is equivalent to

$$\begin{aligned} J(F, W, P, Q) &= \min tr[(X^TW - F)^T\overline{D}(X^TW - F)] + \alpha tr(W^T\widetilde{D}W) \\ &\quad + \beta \|W - PQ\|_F^2 + \Omega(F) \\ s.t. \quad & P^TP = I_r \end{aligned}, \tag{20}$$

where $\Omega(F) = tr(FLF^T) + tr((F - Y)^TU(F - Y)) + \lambda\left\|F - \sum_{p=1}^{a}\gamma_s F_s\right\|^2$, $\overline{D}$ and $\widetilde{D}$ are two diagonal matrices with their diagonal elements being $\overline{D}_{ii} = 1/2\|z_i\|_2$ and $\widetilde{D}_{ii} = 1/2\|w_i\|_2$, respectively. Note that for an arbitrary matrix A, $\|A\|_F^2 = tr(A^TA)$. Thus, (20) becomes

$$\begin{aligned} J(F, W, P, Q) &= \min tr[(X^TW - F)^T\overline{D}(X^TW - F)] + \alpha tr(W^T\widetilde{D}W) \\ &\quad + \beta tr[(W - PQ)^T(W - PQ)] + \Omega(F) \\ s.t. \quad & P^TP = I_r \end{aligned}, \tag{21}$$

The problem (21) therefore can be solved by performing gradient descent on the optimal variables, which involves the $l_{2,1}$-norm that is non-smooth and cannot have a close form solution. Consequently, we propose an iterative algorithm and introduce the following update rules in brief.

4.3.1. Update Q as given W, F, and P

In (21), by setting the derivative $\partial J/\partial Q = 0$ and using the property $P^TP = I_r$, we obtain

$$\beta(P^TPQ - P^TW) = 0 \Rightarrow Q = P^TW, \tag{22}$$

Since $(I - PP^T)(I - PP^T) = (I - PP^T)$, substituting Q in (21) with (22) we have

$$\begin{aligned} & arg\min_{W,P^TP=I_r} tr[(X^TW - F)^T\overline{D}(X^TW - F)] + \alpha tr(W^T\widetilde{D}W) \\ & \quad + \beta tr[(W - PP^TW)^T(W - PP^TW)] + \Omega(F) \\ \Rightarrow & arg\min_{W,P^TP=I_r} tr[(X^TW - F)^T\overline{D}(X^TW - F)] + \alpha tr(W^T\widetilde{D}W) \\ & \quad + \beta tr[W^T(I - PP^T)(I - PP^T)W] + \Omega(F) \\ \Rightarrow & arg\min_{W,P^TP=I_r} tr[(X^TW - F)^T\overline{D}(X^TW - F)] + \alpha tr(W^T\widetilde{D}W) \\ & \quad + tr[W^T(\alpha\widetilde{D} + \beta I - \beta PP^T)W] + \Omega(F) \end{aligned}, \tag{23}$$

4.3.2. Update W as given Q, F, and P

By setting the derivative of (23) w.r.t $W$to zero, we obtain

$$\begin{aligned} & (X\overline{D}X^T + \beta I_d - \beta PP^T + \alpha\widetilde{D})W - X\overline{D}F = 0 \\ \Rightarrow & W = (X\overline{D}X^T + \beta I_d - \beta PP^T + \alpha\widetilde{D})^{-1}X\overline{D}F \\ & W = (M - \beta PP^T)^{-1}X\overline{D}F \\ \Rightarrow & W = (N)^{-1}X\overline{D}F \end{aligned}, \tag{24}$$

Where $M = X\overline{D}X^T + \beta I_d + \alpha\widetilde{D}$, $N = M - \beta PP^T$, thus $N = N^T$.

4.3.3. Update P and F as given Q, and W

First, we rewrite (23) as follows.

$$\begin{aligned} \arg\min_{P,F,W} &= tr[(X^TW - F)^T\overline{D}(X^TW - F)] + \alpha tr(W^T\widetilde{D}W) \\ &\quad + tr[W^T(\alpha\widetilde{D} + \beta I - \beta PP^T)W] + \Omega(F) \\ &= \Omega(F) + tr[W^TX\overline{D}X^TW] - 2tr[W^TX\overline{D}F] + tr[F^T\overline{D}F] \\ &\quad + tr[W^T(\beta I_d - \beta PP^T + \alpha\widetilde{D})W] \\ &= \Omega(F) - 2tr[W^TX\overline{D}F] + tr[F^T\overline{D}F] + tr[W^TNW] \end{aligned}, \tag{25}$$

By substituting the expression for W in (24) into (25), since $N = N^T$, we have the following equation.

$$J = \arg\min_{P,F} - tr[F^T\overline{D}X^TN^{-1}X\overline{D}F] + tr[F^T\overline{D}F] + \Omega(F), \tag{26}$$

By setting the derivative $\partial J/\partial F = 0$ in (26), we obtain

$$\begin{aligned} & LF + U(F - Y) + \lambda(F - F_s) + \overline{D}F - \overline{D}X^TN^{-1}X\overline{D}F = 0 \\ \Rightarrow & F = (L + U + \lambda I_d + \overline{D} - \overline{D}X^TN^{-1}X\overline{D})^{-1}(UY + \lambda F^s) \\ \Rightarrow & F = (B - \overline{D}X^TN^{-1}X\overline{D})^{-1}(UY + \lambda F^s) \end{aligned}, \tag{27}$$

where $F^s = \sum_{p=1}^{a}\gamma_s F_s$, $B = L + U + \lambda I_d + \overline{D}$. Finally, we obtain the following objective function by using (27).

$$J = \max_{P^TP=I} tr((UY + \lambda F^s)^T(B - \overline{D}X^TN^{-1}X\overline{D})^{-1}(UY + \lambda F^s)). \tag{28}$$

To compute the matrix inverse, using the Sherman-Morrison-Woodbury formula [34]:

$$(A - BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1},$$

then we have

$$(B - \overline{D}X^TN^{-1}X\overline{D})^{-1} = B^{-1} + B^{-1}\overline{D}X^T(N - \overline{D}X^TB^{-1}X\overline{D})^{-1}X\overline{D}B^{-1}. \tag{29}$$

Thus, by using the matrix property $tr(AB) = tr(BA)$, the optimization problem (28) is equivalent to

$$J = \max_{P^TP=I} tr(\widetilde{Y}^TB^{-1}\overline{D}X^T\Sigma^{-1}X\overline{D}B^{-1}\widetilde{Y}), \tag{30}$$

where

$$\widetilde{Y}^v = UY + \lambda F^s,\ \Sigma = N - \overline{D}X^TB^{-1}X\overline{D}. \tag{31}$$

Next, we show in Theorem 2 that the optimization problem in (30) can be transformed into a generalized eigenvalue problem.

Theorem 2. *The global optimal P of (30) can be obtained by solving the following ratio trace maximization problem:*

$$\max_{(P^v)^TP^v=I} tr(P^T\widetilde{S}P)^{-1}(P^TRP), \tag{32}$$

where

$$\begin{aligned}&\widetilde{S} = I_r - \beta C^{-1},\\ &R = C^{-1}X\overline{D}B^{-1}\widetilde{Y}\widetilde{Y}^TB^{-1}\overline{D}X^TC^{-1},\\ &C = M - \overline{D}X^TB^{-1}X\overline{D}.\end{aligned} \tag{33}$$

Proof. According to the Woodbury theorem, we have:

$$\begin{aligned}&\Sigma^{-1} = (N - \overline{D}X^TB^{-1}X\overline{D})^{-1}\\ &= (M - \beta PP^T - \overline{D}X^TB^{-1}X\overline{D})^{-1} = (C - \beta PP^T)^{-1}\\ &= C^{-1} + \beta C^{-1}P(I_r - \beta P^TC^{-1}P)^{-1}P^TC^{-1}\\ &= C^{-1} + \beta C^{-1}P(P^T(I_r - \beta C^{-1})P)^{-1}P^TC^{-1}\end{aligned} \tag{34}$$

where $C = M - \overline{D}X^TB^{-1}X\overline{D}$. It is obvious that C is independent to P. By using Eq. (34), the objective function (30) is equivalent to

$$\max_{P^TP=I} tr[(P^T(I_r - \alpha C^{-1})P)^{-1}(P^TC^{-1}X\overline{D}B^{-1}\widetilde{Y}\widetilde{Y}^TB^{-1}\overline{D}X^TC^{-1}P)].$$

Let

$$\widetilde{S} = I_r - \beta C^{-1},$$

$$R = C^{-1}X\overline{D}B^{-1}\widetilde{Y}\widetilde{Y}^TB^{-1}\overline{D}X^TC^{-1},$$

we then obtain (32). ■

Theorem 3. *Given the semi-definite matrices L and U, The matrix $\widetilde{S}$ is a nonsingular matrix.*

Proof. Because L and U are positive semi-definite, $B = L + U + \lambda I_d + \overline{D}$ is positive definite. Note that $P^TP = I_r$. Then, the largest eigenvalue of PP^T is 1. Thus, the smallest eigenvalue of $\beta I_d + \alpha\widetilde{D} - \beta PP^T$ is larger than 0. Therefore, $N = X\overline{D}X^T + \beta I_d + \alpha\widetilde{D} - \beta PP^T$ is positive definite. Because all of the eigenvalues of $I_d - (L + U + \lambda I_d + \overline{D})^{-1}\overline{D}$ is larger than 0, $\overline{D}(I_d - (L + U + \lambda I_d + \overline{D})^{-1}\overline{D})$ is positive definite. Recall that the largest eigenvalue of $PP^T$is 1. We have

$$\begin{aligned}&\Sigma = N - \overline{D}X^TB^{-1}X\overline{D}\\ &= X\overline{D}X^T + \beta I_d + \alpha\widetilde{D} - \beta PP^T - \overline{D}X^TB^{-1}X\overline{D}\\ &= \beta I_d + \alpha\widetilde{D} - \beta PP^T + X(\overline{D} - \overline{D}B^{-1}\overline{D})X^T\\ &= \beta I_d + \alpha\widetilde{D} - \beta PP^T + X(\overline{D}(I_d - (L + U + \lambda I_d + \overline{D})^{-1}\overline{D}^v))X^T\end{aligned}$$

Therefore, Σ is positive definite. Moreover, we have

$$\begin{aligned}&C = M - \overline{D}X^TB^{-1}X\overline{D}\\ &= X\overline{D}X^T + \beta I_d + \alpha\widetilde{D} - \overline{D}X^TB^{-1}X\overline{D}\\ &= \beta I_d + \alpha\widetilde{D} + X(\overline{D}(I_d - (L + U + \lambda I_d + \overline{D})^{-1}\overline{D}^v))X^T\end{aligned}$$

Therefore, C is positive definite. Then $\widetilde{S} = I_r - \beta C^{-1}$ is positive definite.

By the Theorem 3, the optimal P can be effectively obtained by eigen-decomposition of $\widetilde{S}^{-1}R$ since $\widetilde{S}$ is positive definite.

4.4. Overall algorithm

Solving (18) is a challenging task and it appears that such a global solution might not be analytically available. What we propose in the following is an iterative algorithm that alternates between variables such that the updates are tractable. From the above analysis, we can see that $\overline{D}$, $\widetilde{D}$ and related to W is required to solve $P$and F, and it is still not straightforward to obtain W, P and F. To this end, we design an iterative algorithm to solve the proposed formulation, which is summarized in Algorithm 1. In Algorithm 1, we adopt a window based stopping criterion: for a given window size h, at every iteration t, we calculate $\Delta = |MaxJ_t - MinJ_t|/MaxJ_t$, where the set $J_t = \{Obj_{t-h+1}, ...,Obj_t\}$ consists of history objective values in a window h. If $\Delta < 10^{-4}$, the algorithm stops iterating. Specifically, we set $h = 6$ in our experiments.

Algorithm 1
Solving the optimization problem (18).

Input: Target training dataset $X = \{(x_i, y_i),\ i = 1, ...,n\}$; a source predicted label matrices $\{F_s\}_{s=1}^a$, and parameters α, β, λ.
Output: Converged matrix F,W.
Initialization: Set $t = 0$, and initialize W_0, F_0, P_0randomly, and set Ω_0, $\overline{D}_0$ and $\widetilde{D}_0$ as identity matrix.
1: Construct the SFSR-graphs and calculate the SFSR-graph Laplacian matrix L.
2: Compute the selection matrix U.
3: **repeat**
  Compute $M = X\overline{D}X^T + \beta I_d + \alpha\widetilde{D}$;
  Compute $B = L + U + \lambda I_d + \overline{D}$
  Compute $C = M - \overline{D}X^TB^{-1}X\overline{D}$
  Compute $F^s = \sum_{p=1}^a \gamma_s F_s$, then $\widetilde{Y}_t = UY + \lambda F^s$;
  Compute $\widetilde{S} = I_r - \beta C^{-1}$;
  Compute $R = C^{-1}X\overline{D}B^{-1}\widetilde{Y}\widetilde{Y}^TB^{-1}\overline{D}X^TC^{-1}$;
  Obtain P_{t+1} by the ratio trace maximization problem in (32)
  Compute $N = M - \beta PP^T$;
  Update $F_{t+1} = (B - \overline{D}X^TN^{-1}X\overline{D})^{-1}\widetilde{Y}_t$;
  Update $W_{t+1} = N_t^{-1}X\overline{D}_tF_t$;
  Compute $Z_{t+1} = X^TW_{t+1} - F_{t+1}$;
  Update $Q_{t+1} = P_{t+1}^TW_{t+1}$;
  Update $\overline{D}_{t+1}$ and $\widetilde{D}_{t+1}$;
  Set $t = t + 1$
4: **until** $|MaxJ_t - MinJ_t|/MaxJ_t < 10^{-4}$
5: **return** F and W.

Once P and W are obtained, given a testing data $x \in X$, its latent representation and label vector are computed by $b = x^TP$, $y_x = W^Tx$, respectively. Besides, W can be deemed as a feature selection matrix in the original target space.

After the latent representation matrix F and the optimal model parameter W is learned, the predicted label of $x_u \in X_u (n_l + 1 \le u \le n)$ is determined by

$$f_u = \arg\max_{1\le j\le c} f_{uj} \tag{35}$$

For the out-of-sample data, we also can use the linear classifier $f(X_u) = (X_u)^TW$ for target classification. In other words, the label for some test data $x_i \in X_u$is given by:

$$j = \arg\max_j (y_i = x_i^TW)_j. \tag{36}$$

4.5. Computational complexity

In the block coordinate descent procedure, the problem (18) is solved by iteratively using Algorithm 1. We focus on discussing the computational complexity of the main components involved in each iteration of these two subproblems. For the Algorithm 1, the main computational cost comes from computing the gradient of P, Q, F, and W. Specifically, the computational cost for these terms in (18) are $O(d^3 + d^2n + dn^2 + dnc)$, $O(r + dr^2 + drc)$, $O(n^2c)$ and $O(d^2n + dn + dnc)$, respectively. Therefore, the overall computational complexity of the problem (18) is $O(maxT(d^3 + d^2n + dn^2 + dr^2 + n^2c + dnc + drc))$with $maxT$ denotes the maximum number of iterations in our algorithm. Since $r \ll n$ and $r \ll d$, the whole cost of the problem (18) is $O(maxT(d^3 + d^2n + (d + c)n^2 + dnc))$, which grows cubically with the feature dimensionality, quadratically with the number of target data, and linearly with the number of classes. When the number of

the target samples is relatively small, the whole cost of the problem (18) can be further simplified as $O(d^3)$, which only grows cubically with the feature dimensionality.

5. Discussion

5.1. Ensemble SFSR-graph Laplacian

Since there is no evidence to show that the performance of SFSR with $p \geq 3$ is better than that with $p = 2$ [7], we further propose an ensemble strategy of multiple SFSR-graphs. Specifically, we give the following ensemble SFSR-graph Laplacian matrix.

$$L = \sum_{p=1}^{g} \mu_p L_p,\ g \in \{1, 2, ...,n\},\ s.t.\ \sum_{p=1}^{g} \mu_p = 1, \mu_p \geq 0, for p = 1, ...,g \tag{37}$$

where L_p corresponds to the afore-mentioned SFSR-graph Laplacian matrix with a given p. In fact, we define in (37) a set of SFSR-graph Laplacians $\Theta = \{L_1, ...,L_g\}$ and denote a convex hull of set Ψ as $conv(\Psi) = \{\sum_{i=1}^{g} \theta_i x_i | \sum_{i=1}^{g} \theta_i = 1, x_i \in \Psi, \theta_i \geq 0\}$. Therefore, we have $L \in Conv(\Theta)$, which is also a graph Laplacian. Particularly, when $p = 1$, L is the traditional sparse graph Laplacian matrix.

Under this strategy, the optimal SFSR-graph estimation is turned into the problem of learning the optimal linear combination of the SFSR-graphs set. Because each SFSR-graph Laplacian L_p represents a certain SFSR of the given target samples, the ensemble SFSR-graph framework can be understood geometrically as follows: first, compute all possible approximated SFSRs, each of which corresponds to a "guess" at the intrinsic data representation of the target data, and then learn to linearly combine them for an optimal composite. To minimize the learning complexity over the ensemble SFSR-graph Laplacian, we introduce a new SFSR-graph regularization term,

$$J(F, W, P, Q) = \min \widetilde{\Omega}(W, P, Q, F) + tr\left(F\left(\sum_{p=1}^{g} \mu_p L_p\right)F^T\right) + \eta \|\mu\|_2^2,$$
$$s.t. \quad P^T P = I_r, \sum_{p=1}^{g} \mu_p = 1, \ \mu_p \geq 0, for p = 1, ...,g \tag{38}$$

where $\mu = [\mu_1, ...,\mu_g]^T$, $\widetilde{\Omega}(W, F, P, Q) = \|X^T W - F\|_{2,1} + \alpha \|W\|_{2,1} + \beta \|W - PQ\|_F^2 + tr((F - Y)^T U (F - Y)) + \lambda \left\|F - \sum_{p=1}^{a} \gamma_s F_s\right\|^2$.

Because (38) contains a weighted combination of multiple SFSR-graph regularization terms, we name this new regularization as ensemble SFSR-graph regularization (ESFSR-graph). Note that the works in [35] and [36] also adopt an ensemble graph Laplacians approach to learn an optimal graph combination for inferring the data manifold. However, our ESFSR-graph regularization is intrinsically different from those methods in that our method aims to learn an optimal SFSR-graphs combination via constructing a more informative feature space representation for target learning.

For a fixed μ, (38) degenerates to (18), with $L = \sum_{p=1}^{g} \mu_p L_p$. On the other hand, for a fixed $\widetilde{\Omega}(W, F, P, Q)$, (38) is simplified to:

$$J(\mu) = \min tr\left(F\left(\sum_{p=1}^{g} \mu_p L_p\right)F^T\right) + \eta \|\mu\|_2^2, \quad s.t. \ \sum_{p=1}^{g} \mu_p = 1, \mu_p \geq 0, \text{for} p = 1, ...,g. \tag{39}$$

where η is a regularization parameter. Let $\omega = (\omega_1, \omega_2, ...,\omega_g)^{\mathrm{T}}$ and $\omega_p = tr(\mu_p(F)^T L_p F)$. If we set $\eta = \infty$, the solution is the uniform weight $\mu = 1/g$. On the other hand, if we set $\eta = 0$, the solution is $\mu = e_{p_{\min}}$, where e_k is the kth unit vector (i.e., a vector with a one in the kth element and zeros elsewhere) and $p_{\min} = \arg\min_p \omega_p$. As reported in [36], between these two extremes $\eta \in (0, \infty)$, we can obtain sparse weighting coefficients.

For a fixed $\widetilde{\Omega}(W, F, P, Q)$, the optimal μ is the solution of the following problem:

$$\min_{\mu} \omega^{\mathrm{T}} \mu + \frac{\eta}{2} \|\mu\|_2^2 s.t.\ \ \mu^T 1 = 1, \mu \geq 0. \tag{40}$$

Without loss of generality, we may assume that $\{\omega_p\}_{p=1}^{g}$ are stored in increasing order: $\omega_1 \leq \omega_2 \leq ... \leq \omega_g$. Then, the following theorem gives analytic representations of the optimal solution for (40).

Theorem 4. [36]. *The optimal solution to the problem (40) is given by the following equations:*

$$\mu_p = \frac{\varsigma - \omega_p}{\eta}, \ \text{for} p = 1, 2, ...,\xi \ \text{and} \ \mu_p = 0 \ \text{for} p = \xi + 1, ...,g, \tag{41}$$

where $\varsigma = \frac{\eta + \sum_{p=1}^{g} \omega_p}{\xi}$, and $\xi = |\{p | \varsigma - \omega_p > 0, p = 1, 2, ...,g\}|$.

Theorem 4 shows that the optimal μ has only ξ nonzero entries. This makes this method have sparse solutions in terms of the SFSR-graphs Laplacian weights. In other words, this scheme can automatically select the important SFSR-graphs and ignore the irrelevant or noisy SFSR-graphs for discrimination. Since one can compute the optimal μ by given the optimal ξ, Karasuyama et al. [36] further proposed a naïve algorithm to find the optimal ξ, which is verified to be efficient enough if the number of given SFSR-graphs is kept at a moderate size. In the Section 6, we will evaluate the effectiveness of ensemble SFSR-graph Laplacians.

5.2. Generalization error bound

To demonstrate the effectiveness and superiority of the proposed framework, we further derive in this section the generalization error bound of it using the tool of transductive Rademacher complexity for general function classes [30], which generally measures the richness of a class of real-valued functions with respect to a probability distribution.

Theorem 5. [30]. *Given a fixed sample set $X = \{x_1, x_2, ...,x_n\}$ generated by a distribution P_χ on a set χ and a real-valued function class Ψ with domain χ, and a class of real-value functions Ψ mapping from $\chi \times \Gamma$ to $[0, 1]$, fix $\delta \in (0, 1)$, for any fixed sample set $\{(x_i, y_i)\}_{i=1}^{n}$ with probability $1 - \delta$ over random draws of a subsample of size n_l, every function $f \in \Psi$ satisfies*

$$err(f) \leq \widehat{err}(f) + \widehat{R}_n(\Psi) + c_0 Q \sqrt{\min(n_l, n_u)} + \sqrt{2Q \ln(1/\delta)}, \tag{42}$$

where $c_0 = \sqrt{\frac{32 \ln(4e)}{3}}$, $\widetilde{Q} = \frac{1}{n_l} + \frac{1}{n_u}$, n_l and n_u ($n_l + n_u = n$) denote the number of the labeled and unlabeled samples, respectively, $err(f)$ is the expected error on the unlabeled data, $\widehat{err}(f)$ is the empirical error on the labeled data, and $\widehat{R}_n(\Psi)$ is the empirical transductive Rademacher complexity of Ψ.

The above error bound is quite general and applicable to various transductive learning algorithms if an empirical transductive Rademacher complexity $\widehat{R}_n(\Psi)$ of the function class Ψ can be found efficiently. It also implies that in order to prove the generalization error bound for SFSR-MSAL, it is sufficient to give an estimation of the empirical transductive Rademacher complexity for the following function class definition.

Definition 1. The function class of SFSR-MSAL is defined as:

$$\psi_{SFSR-MSAL} = \{\mathbf{F} = \mathrm{Y}^{-1}(UY + \lambda F^s), \|UY + \lambda F^s\|_F \leq \mathbb{C}\}, \tag{43}$$

where $\mathrm{Y} = B - \overline{D} X^T N^{-1} X \overline{D}$ with $B = L + U + \lambda I_d + \overline{D}$, and $\mathbb{C}$ is some constant.

Theorem 6. *The empirical transductive Rademacher complexity of the function class $\psi_{SFSR-MSAL}$ is upper bounded as*

$$\widehat{R}_n(\psi_{SFSR-MSAL}) \leq \mathbb{C} \sqrt{\frac{2}{n_l n_u} tr(\mathrm{Y}^{-2})}. \tag{44}$$

Proof. The detailed proof of Theorem 6 can be found in Appendix B.

Using Theorems 4 and 5, we obtain the following generalization error bound for SFSR-MSAL.

Theorem 7. [30]. *Fix $\delta \in (0, 1)$, and let the function class of SFSR-MADL $\psi_{SFSR-MSAL}$ be a class of functions mapping from $\chi \times \Gamma$ to [0, 1]. Let $c_0 = \sqrt{\frac{32\ln(4e)}{3}}$ and $\widetilde{Q} = \frac{1}{n_l} + \frac{1}{n_u}$. For any fixed sample set $\{(x_i, y_i)\}_{i=1}^{n}$, with probability $1 - \delta$ over random draws of a subsample of size n_l, every $f \in \psi_{SFSR-MADL}$ satisfies*

$$err(f) \le \widehat{err}(f) + \mathbb{C}\sqrt{\frac{2}{n_l n_u} tr(\Upsilon^{-2})} + c_0 Q\sqrt{\min(n_l, n_u)} + \sqrt{2Q\ln(1/\delta)}. \tag{45}$$

Note that SFSR-MSAL is performed in some latent subspace P, where the target classifier is adapted by leveraging multiple prior source classifiers, $N = M - \beta PP^T$ and $M = X\overline{D}X^T + \beta I_d + \alpha\widetilde{D}$. Theorem 7 states that the target classifier f can be guaranteed by an error bound in the target domain. In other words, by effectively utilizing both the limited labeled information and large number of unlabeled structure information particularly the sparse feature space representation information from target domain in some latent spaces, SFSR-MSAL can implicitly minimize the empirical transductive Rademacher complexity of the function class $\psi_{SFSR-MSAL}$. This is also verified by our experiments later, in which our SFSR-MSAL can achieve comparable or outperformed performance compared with several state-of-the-art methods in most cases of our trials.

5.3. Adaptation weights

As we may know, one of major problems in the DA challenge is how to reduce the distribution divergent between source and target domains by a certain distribution transform technique [15]. There have witnessed several research works aiming to measure the distance between distributions of both domains. For instance, Gretton et al. [37] showed that for a given class of functions, the measure could be simplified by computing the discrepancy between two means of the distributions in a reproducing kernel Hilbert space (RKHS), thus resulting in the maximum mean discrepancy (MMD) measure criterion. MMD is an effective nonparametric criterion to compare data distributions based on the distance between the means of samples of two domains in the sth RKHS H_s induced by a kernel k [38], namely, the empirical measure of MMD between the target domain and the sth source domain can be well estimated by

$$M^s = Dist(X^s, X) = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_i^s) - \frac{1}{n}\sum_{j=1}^{n}\phi(x_j) \right\|_{H_s}. \tag{46}$$

The MMD criterion is able to capture the high order statistics of the data when the samples are mapped into a high dimensional or even infinite dimensional space [37] and has recently been shown to be both efficient and effective for estimating the distance between two distributions [17].

Notice that finding the optimal value for the elements of the weight vector $\gamma = [\gamma_1, ..., \gamma_a]^T$ corresponds to ranking the prior source models and decide from where and how much to adapt. We here propose to choose γ in order to minimize MMD. Specifically, we use the MMD between the target domain and the sth source domain to define the adaptation weight γ_s as below:

$$\gamma_s = \frac{\exp(-\delta M^s)}{\sum_{s=1}^{a}\exp(-\delta M^s)}, \quad s = 1, ..., a. \tag{47}$$

where $\delta > 0$ is the bandwidth parameter to control the spread of M^s.

6. Experiments

Currently, how to choose the optimal model parameters still keeps an open and hot topic. In general, the parameters are manually set. In order to evaluate the performance of the algorithms, a common strategy is that a set of prior parameters is first given and then the best cross-validation mean rate among the set is used to estimate the generalized accuracy. In this work, this strategy is also adopted. The fivefold cross-validation is used on the training set for the optimal parameters selection. Finally, the mean of the experimental results on testing data is used for performance evaluation. The overall accuracy rate (i.e., the percentage of correctly labeled samples over the total number of samples) or error rate (i.e., the percentage of wrongly labeled samples over the total number of samples) are used as the reference classification performance measure. For all the datasets, true labels are available for instances from source domains. All the labeled samples from both the source and target domains are selected for training while the unlabeled samples from only the target domain are selected for testing. We use classification Accuracy on test data as the evaluation metric as follows:

$$Acc\% = \frac{|x: x \in X \wedge f(x) = y_x|}{|x: x \in X|},$$

where $f(x)$ is the label predicted by the classification algorithm, y_x is the truth label of x. The overall accuracy $Acc\%$ is used as the reference classification measure.

6.1. Semi-supervised learning

In this section, we present a serial of experiments where we use our method for several SSL tasks to evaluate the effectiveness of SFSR-GSSL. We compare our SFSR-GSSL against several representative learning algorithms as follows.

- 1-NN [39];
- GFHF [2];
- LapRLS [3] which is based on manifold regularization;
- LNP [8] which is based on linear neighbor propagation scheme;
- S-RLSC [18] which is based on sparse reconstruction trick.

These algorithms except 1-NN are the state-of-the-art GSSL algorithms that have shown their effectiveness in various semi-supervised learning tasks.

For the kernel methods GFHF and LapRLS, we use the standard Gaussian kernel function $k_\theta(x, z) = \exp(-\theta\|x - z\|^2)$, where θ is set to $1/d$ (where d is the number of features). The diagonal element of Laplacian graph matrix in GFHF method is set as 0.

In LapRLS and S-RLSC, there are another two main parameters to be tuned, i.e., the regularization parameters $\gamma_A$and γ_I. For fairness, we set them as the same as those in reference [3].

In LNP, there are two main parameters to be tuned, i.e., the nearest neighbor number and one regularization parameter. In our experiments, we set manually the number of the nearest neighbor is 7 and the regularization parameter $\alpha = 0.9$.

In our method SFSR-GSSL, the regularization parameters are first empirically set as: $\lambda = 10^3$, $\alpha = 1$, $\beta = 10^3$. For simplicity, in this experiment, we just set $p = 2$. In our experiments, it is observed that the performance of SFSR-GSSL (resp. SFSR-MSAL) is not sensitive to the dimensionality r of the optimal latent subspace. We may set $r = 7 \times \lfloor\frac{c-1}{7}\rfloor$ for simplicity without loss of performance,[2] where $\lfloor o \rfloor$ denotes the largest integer not greater than o. This strategy is also effectively adopted in [13]. Based on the selected features, we then can

[2] In real applications where the number of classes is smaller than 8, we can still select the r smallest eigenvectors corresponding to the r smallest eigenvalues to construct the projection matrix, which is similar to PCA.

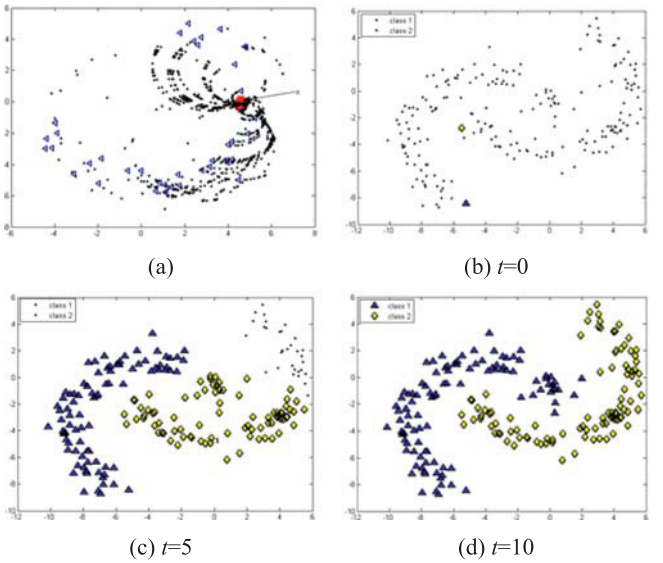


Fig. 3. Label propagation results of SFSR-GSSL on the toy dataset: (a) feature space embedding of one original point (annotated in red); (b) on the initial state, i.e., the iterative time $t = 0$, only two samples are labeled; (c) the state after 5 iterations; (d) on the convergent state, that is, $t = 10$, all points are correctly labeled.

train a classifier for different tasks.

6.1.1. Label propagation on toy data

In this subsection, a two-dimensional toy dataset is used to show the efficiency and effectiveness of the proposed algorithm, and to understand the label propagation processes for a GSSL problem. The learning dataset is characterized by a synthetic dataset composed of 100 samples generated according to a 2-dimensional pattern of two intertwining moons denoting two different information classes (50 samples each). The dataset includes two labeled samples annotated in yellow and blue, respectively, as shown in Fig. 3(b).

Fig. 3(a) shows the feature space embedding projection points (marked with black "*") of one point x (marked with red "☆") from one class (marked with blue "△"). We can see from Fig. 3(a) that comparing with other original data points, the feature space embedding projection points of x have richer representation information. It means that the feature space embedding projections can augment the representation space of original data to some extent and have more complete discrimination information. This can obviously strengthen the recognition performance. Fig. 3(c) directly shows the effectiveness of label propagation in double-moon datasets after $t = 5$ iterations. Fig. 3(d) reports the result of label propagation by SFSR-GSSL, which demonstrates the effectiveness of SFSR-GSSL in solving this kind of problem.

6.1.2. Face recognition

In order to evaluate the robustness and effectiveness of our proposed algorithm, four face datasets are adopted in this experiment, namely, YALE B, CMU PIE, ORL, and YALE, which are shown in Fig. 4(a)–(d).

The expanded Yale (YALE B) face database contains 2114 front face images of 38 distinct objects and each subject has 60 images. Each image is cropped to the size of 32 × 32 pixels, thus being represented by a 1024-dimensional vector; the PIE face database contains 41,368 face images with 68 volunteers and each subject contains 170 images taken at different gestures, illumination conditions, and face expressions (open/closed eyes, smiling/not smiling). In the test, the images in the database is preprocessed to zoom out to 32 × 32 pixel and each pixel has 256 Gy degrees. Then in the image space, each image is represented by a 1024-dimensional vector; the ORL dataset contains 400 face images with 40 distinct subjects and each subject has 10 images taken at different times, lighting conditions, facial expressions (open/closed eyes, smiling/not smiling), and facial details (with glasses/no glasses) against a dark homogeneous background. The images used in our experiments are of size 40 × 40 pixels; the YALE dataset consists of 165 images from 15 individuals, each of which has 11 images with different facial expressions or configurations. Each image from this dataset was cropped to the size of 32 × 32 pixels in our experiments.

For each face dataset X, we randomly select m samples as the labeled training data. For 1-NN algorithm, the training set only contains one labeled data. For LNP, GFHF, LapRLS, S-RLSC and SFSR-GSSL methods, each training set contains all labeled and unlabeled data in the dataset X. Fig. 5(a)–(d) show the experimental results of face recognition in four face datasets with the afore-mentioned methods, respectively. The horizontal axis shows the number of face images labeled randomly in each subject. The vertical axis shows the corresponding recognition error rate of all the algorithms. From the results as shown in Fig. 5, we can observe that the recognition performance of SFSR-GSSL is better than the other methods on all face datasets. With the increasing number of labeled face samples, SFSR-GSSL, S-RLSC and LNP methods have comparable recognition performance.

6.1.3. Object recognition

In this part, we apply the SFSR-GSSL method to object recognition task. The image library of Columbia University (COIL 20) is adopted as the training dataset. This library contains the gray object image sets of 20 classes. For each class, there are 72 images of size 28 × 28. Fig. 6(a) shows the object images of 20 classes. Fig. 6(b) shows the recognition performance of different methods on COIL 20 with different number of labeled objects. We can see from Fig. 6(b) that SFSR-GSSL method is still better than the traditional methods, where it has the comparable performance with S-RLSC. With the number of label samples increases, the recognition performances of LNP, S-RLSC and SFSR-GSSL get close.

It can be seen from the plot that all methods manifest the same trend of upgrade of performance with the increase of the number of labeled samples. The LNP method obtains more significant improvement of performance when the number of labeled samples gradually increases. The performance of our method and S-RLSC can be steadily improved when the number of the labeled training samples increases. This demonstrates two folds: on one hand, it is beneficial to utilize the labeled data in GSSL to improve the learning performance; on the other hand, sparse representation in object recognition can bring more

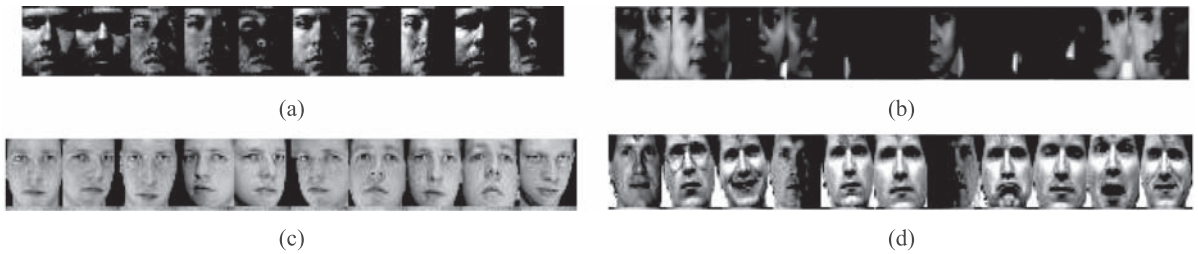


Fig. 4. face objects in (a)YALE B; (b)PIE; (c)ORL and (d)YALE.

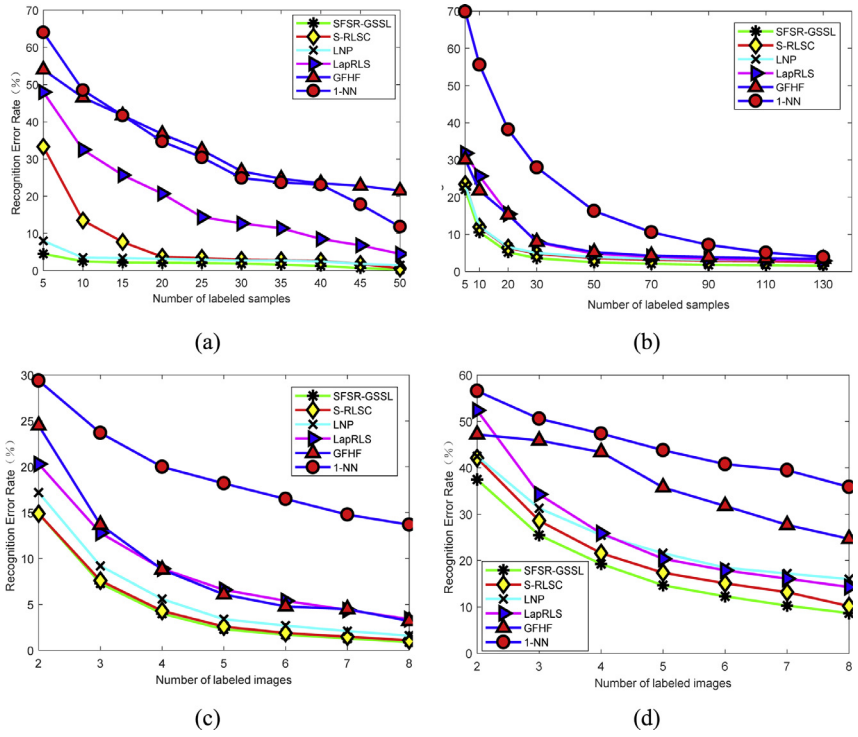


Fig. 5. Face recognition results on (a)YALE B; (b)PIE; (c) ORL and (d)YALE.

discriminative ability as well as robustness. However, our method still is the best while it has the comparable performance with S-RLSC. Moreover, we can observe that our method changes smoothly. In other words, even using only a relatively small fraction of labeled data, our method can still obtain a higher classification accuracy rate. However, other GSSL methods, especially GFHF, may only achieve satisfactory performance when the number of labeled training samples is relatively large.

6.1.4. Digit recognition

Next, we further evaluate our method on digit recognition by using the real-world digit dataset USPS (http://www.kernel-machines.org/data.html). The dataset contains 10 classes of size 16 × 16 handwritten digital images. We choose the first 200 images of digit 1, 2, 3, and 4 as the training set and the 200 images afterwards as the out-of-sample set. Fig. 7(a) shows the samples of 4 digits. In the trial, the average value of 10 independent experimental results is recorded as the recognition precision rate of each algorithm. Fig. 7(b) plots the performance comparison of all methods. We can see from the results shown in Fig. 7(b) the performance superiority of our method. With the increase of randomly labeled samples, LapRLS, LNP and S-RLSC methods have comparable and consistent recognition performance. Note that with the increase of randomly labeled samples, the recognition performance of the SFSR-GSSL method is very stable. In other words, even with a few labeled samples, SFSR-GSSL can have high recognition performance. Since the digit images of the same class are in the same sub-manifold structure [7], the experimental results show that our SFSR-GSSL method can reveal the intrinsic structures of digits of different classes, which is attributed to the sparse representation technology, thus correctly recognizing those unlabeled digits even with a few labeled samples. Besides, due to the sparse feature space representation technology adopted in our method, the discrimination information is augmented to some extent, thus improving the discriminative performance of SFSR-GSSL.

To further illustrate the effectiveness of the proposed algorithm for out-of-sample data, the digit recognition problem with 1% labeled samples in the training set is used. We perform transduction learning on the first 200 digits with the variable number of labeled samples. We here use the formula (36) to directly predict the label matrix of the out-of-sample for inductive learning. Fig. 7(c) shows the results of transduction learning and out-of-sample inductive learning, respectively. We can see from the plot that the SFSR-GSSL method can get high inductive learning performance.

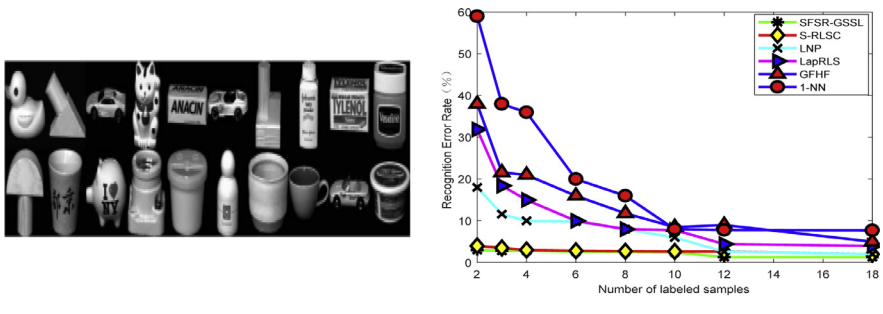
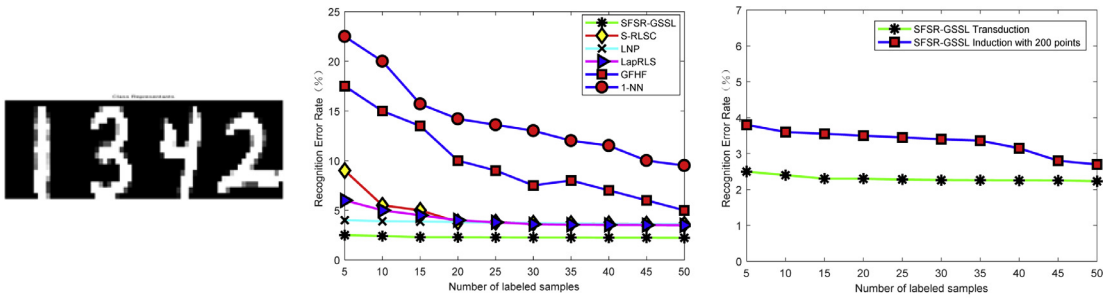


Fig. 6. Object recognition: (a) 20 samples from COIL 20; (b) experimental results of all methods.



(a) (b) (c)

Fig. 7. Digit recognition on the USPS dataset. (a) A subset containing digits 1 to 4; (b) The recognition accuracies of different algorithms. (c) The recognition accuracies with 200 points selected for training (transduction) and following 200 data points for testing (induction).

6.2. Multi-source adaptation

In this subsection, we will further evaluate the effectiveness and the robustness of our method SFSR-MSAL on multi-source adaptation tasks with two visual benchmark datasets widely adopted to evaluate computer vision and patter recognition algorithms, i.e., object dataset Caltech-256, and video dataset TRECVID 2005. The selection of these classification (recognition) problems is to demonstrate the wide applicability of our proposed multi-source adaptation framework to various tasks.

6.2.1. Data description

The Caltech-256 dataset contains images of 256 object classes plus a clutter category that can be used as negative class in object vs. background problems. The objects are organized in a hierarchical ontology that makes it easy to identify the related and unrelated categories. We download[3] the pre-computed features of [20] and we select SIFT Appearance Descriptors for image descriptors. They are all computed in a spatial pyramid and we consider just the first level (i.e. information extracted from the whole image). We perform the experiments on this dataset in a leave-one-class-out approach, which is considered in turn each class as target and all the others as sources. The number of source training samples is kept fixed while the number of target training samples increases in subsequent steps till reaching the same amount of the source set. The samples are extracted randomly 10 times for an equal number of experimental runs.

The TRECVID video corpus[4] is one of the largest annotated video benchmarking dataset for research purposes. The TRECVID 2005 dataset contains 61,901 key-frames extracted from 108 hours of video programs from six different broadcast channels, including three English channels (CNN, MSNBC, and NBC), two Chinese channels (CCTV and NTDTV), and one Arabic channel (LBC). The total number of keyframes in each channel is listed in Table 1. Thirty-six semantic concepts were chosen from the LSCOM-lite lexicon [17,18], which covers 36 dominant visual concepts present in broadcast news videos including objects, locations, people, events, and programs. The 36 concepts have been manually annotated to describe the visual content of the key-frames in the TRECVID 2005 dataset. As shown in [17] and [18], the data distributions of 6 channels are quite different, making it suitable for evaluating domain adaptation learning methods. In practice, 4000 samples from the target domain are randomly sampled as testing dataset, which are used as unlabeled samples during the learning process. Three low-level global features, namely, Grid Color Moment (225-dimension), Gabor Texture (48-dimension) and Edge Direction Histogram (73-dimension), are extracted to represent the diverse contents of key-frames, due to their consistently good performances reported in TRECVID. Moreover, three types of global features can be efficiently extracted, and the previous works [17] also shows that the cross-domain issue exists when using these global features. We further concatenate the three types of features to form a 346-dimensional feature vector for each key-frame.

6.2.2. Baseline methods

Here we compare our method SFSR-MSAL with several representative state-of-the-art multi-source adaptation methods for visual classification, as shown below.

- FastDAM [18][5];
- Multi-KT [20],[6] in which the l_2-norm constraint on p is used;
- A-SVM [19][7].

In visual video conception detection problems, for performance evaluation, we use non-interpolated average precision (AP) [18] which has been used as the official performance metric in TRECVID since 2001. AP is related to the multi-point average precision value of a precision-recall curve, and incorporates the effect of recall when AP is computed over the entire classification results.

6.2.3. Experimental setup

In the paper, we focus mainly on multiple source domains adaptation settings. In order to evaluate the performance of the algorithms, the mean of the experimental results on testing data is used for performance comparison. We present the best performance for each algorithm over a range of parameters, and we center this range on the best performing parameters reported in the corresponding papers.

For all benchmark methods, each prior classifier model is built with standard SVM[8] on each source domain and we use the Gaussian kernel $k(x, x') = \exp(-\gamma\|x-x'\|^2)$ on the both source and target domains for all the experiments. Regarding the parameters in FastDAM, Multi-KT, and A-SVM, a unique common value for γ is chosen for all the kernels by cross validation on the source sets. And in these method, the value of C is instead determined as the one producing the best result when learning from scratch. There is no guarantee that the obtained C value is the best for the transfer approach; still in this way we compare against the best performance that can be obtained by learning only on the available training samples, without exploiting the source knowledge. We use this setup for all the experiments; specific differences are otherwise mentioned. In FastDAM, we also need to determine the

[3] http://files.is.tue.mpg.de/pgehler/projects/iccv09/.
[4] http://www-nlpir.nist.gov/projects/trecvid.
[5] The MATLAB code is available from http://vc.sce.ntu.edu.sg/transfer_learning_domain_adaptation_data/DAM-TNNLS2012.html.
[6] We use the code available from http://homes.esat.kuleuven.be/~ttommasi/source_code_CVPR10.html.
[7] The MATLAB code is available online http://www.robots.ox.ac.uk/~vgg/software/tabularasa/.
[8] We use the LIBSVM tool downloaded from http://www.csie.ntu.edu.tw/~cjlin/.

Table 1
Description of the TRECVID 2005 dataset.

Channel	CNN_ENG	NBC_ENG	MSNBC_ENG	CCTV_CHN	NTDTV_CHN	LBC_ARB
# key-frames	11,025	9322	8905	10,896	6481	15,272

weight γ_i for the ith base classifier. For fair comparison, we set γ_i as the same as that in our method and empirically set $\delta = 100$.

In our method SFSR-MSAL, there are three main model parameters to be tuned, i.e., λ, α, and β. To determine the appropriate parameters setting for our method, we vary the values of them throughout the experiments. In the coming sections, we will provide empirical analysis on parameter sensitivity, which verifies that proposed methods can achieve stable performance under a wide range of parameter values. When comparing with the baseline methods, we first empirically use the following parameter settings: $\lambda = 10^3$, $\alpha = 1$, $\beta = 10^3$. In the following experiments, we set $p = 3$ in our SFSR-MSAL algorithm.

In our experiments, the randomly selected 50% data is treated as the training dataset and the remaining 50% data is used as the test dataset if without special expression. Among the training data, we randomly label n_l samples per target domain and treat the other target samples as unlabeled data. The above setting (referred to as semi-supervised setting) has been used in [20].

6.2.4. Object recognition

In this trial, we show a benchmark evaluation of our SFSR-MSAL method against A-SVM, Multi-KT, and FastDAM on object recognition task. In A-SVM, we use the average of all the prior models as source information, i.e., $W = \frac{1}{a}\sum_{s=1}^{a} W_s$ with $\overline{\gamma} = 1$. We adopt the SIFT features and two randomly extracted sets of 10 and 20 classes from Caltech-256. In particular, the second set is obtained by adding an extra random group of 10 classes to the first one. From Fig. 8 it is clearly observed that in both experiments SFSR-MSAL significantly outperforms other methods. Moreover, for very few samples, properly weighting each prior source model with SFSR-MSAL, FastDAM, and Multi-KT is better than averaging over all the prior models as done by A-SVM. In the case of 20 classes, the difference among all methods is not statistically significant when the number of target training samples is increased to 9.

For any open-ended learning system, the number of known object categories is expected to grow in time. An increasing number of categories (or sources) may give rise to a scalability problem in domain adaptation learning due to the necessity of checking the reliability of each prior model for one new task. Specifically, for more than 100 sources the domain adaptation methods described above become extremely expensive in computational costs [20]. We perform experiments with 150 and 256 object classes from Caltech-256 dataset, reporting the results of SFSR-MSAL and Multi-KT with weighted prior model, FastDAM and A-SVM with averaged prior model in Fig. 9, respectively. In both cases, properly choosing the weights to assign to each source pays off with respect to average over all the sources for very few training samples: SFSR-MSAL and Multi-KT outperform FastDAM and A-SVM for less than three target training samples. This demonstrates that with enough training samples and a rich prior domain set, the best choice is to consider any source information. Besides, our method still achieves the best performance than all of others in most cases, which manifest sound scalability of proposed method.

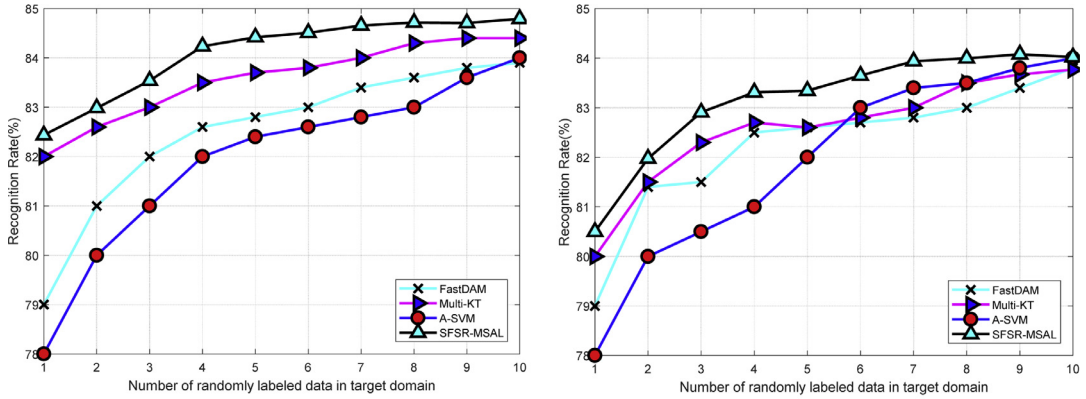
6.2.5. Visual concept recognition

We here choose five channels (i.e., three Chinese channels and two English channels) as the trial domain set. The training dataset consists of all labeled samples from source domains as well as 10 randomly labeled samples from the target domain. The remaining samples from the target domain are used as the test dataset. In this trial, each of the five channels is considered in turn as target while the others represent the source domains. The number of training samples increases in subsequent steps till reaching 10. The samples are extracted randomly 10 times for an equal number of experimental runs.

The mean average recognition precisions (MAPs) of all methods on different channel over 36 concepts are given in Fig. 10. From the Figure, we can also observe that our method SFSR-MSAL is clearly better than others in terms of MAPs on all channels. From Fig. 10(a), A-SVM and Multi-KT achieve statistically similar performance over all channels. An interesting observation is that FastDAM statistically gets a significant superiority than Multi-KT and A-SVM on the channel CCTV_CHN. A possible explanation is that FastDAM can learn a more robust target classifier for domain adaptation by leveraging a set of prelearned base classifiers in such complex scenarios of multi-source adaptation.

Besides, the plot in Fig. 10(b) shows the comparison of SFSR-MSAL with other methods in terms of the average recognition rates of all channels when increasing the number of target training samples. It is clear that for relatively small number of target training samples, our method SFSR-MSAL still outperforms all other methods, while all methods approach asymptotic performance when the number of target training samples increases progressively.

Remark 1. Tommasi et al. [20] pointed out that it is possible to define



(a) 10 classes (b) 20 classes

Fig. 8. Recognition rate as a function of the number of training samples from target domain. In multi-source adaptation settings, we consider in turn one of the domains as target and the others as sources and the final results correspond to average recognition rate over the individuals.

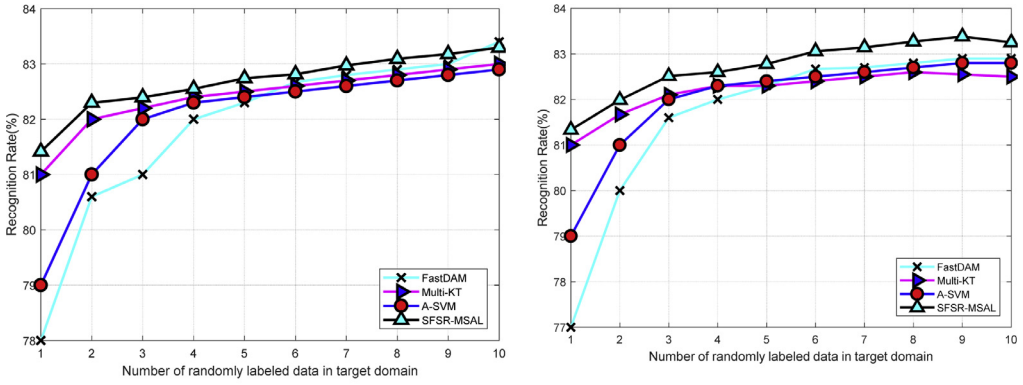


Fig. 9. Recognition performance for high number of source sets when varying the number of prior known object categories.

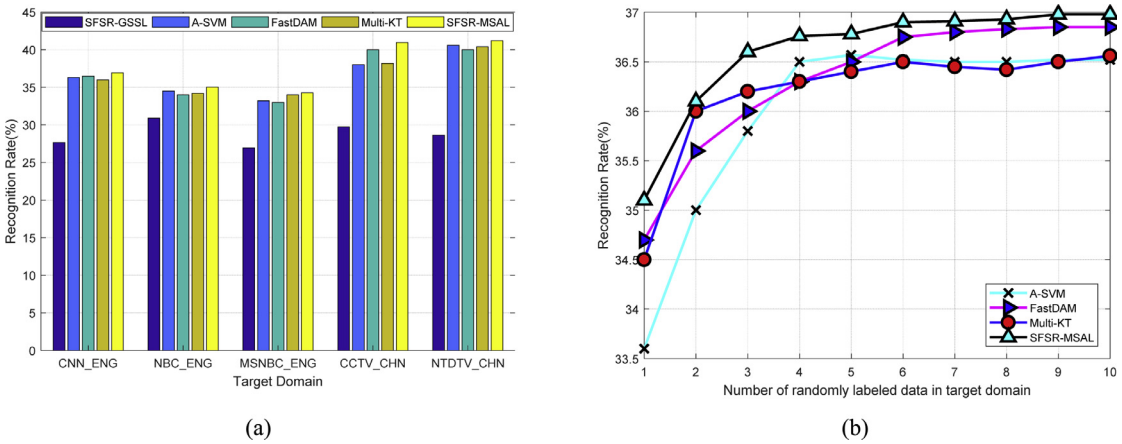


Fig. 10. Recognition performance (AP rate) of all algorithms over 36 concepts for (a) individual channel, and (b) the average recognition rate over the concepts when varying the number of prior known concepts.

three measures by which a proposed domain adaptation method may improve the effectiveness of learning, i.e., (1) Higher start: the initial performance achievable on the target domain is much better than those traditional arts; (2) Higher slope: this indicates a shorter amount of time needed to fully learn the target task, given the transferred knowledge, in comparison with learning from scratch; (3) Higher asymptote: in the long run, the final performance level achievable over the target domain may be higher compared to the final level without transfer. From the experimental results shown in Figs 8–10, it can be interestingly found that our proposed method completely meets all these measures. Hence, we argue that the propose domain adaptation framework is effective on improving the performance of learning in target task.

6.3. Robustness of SFSR-MSAL

In this subsection, we will once again evaluate on Caltech-256 datasets the robustness and effectiveness of our proposed SFSR-MSAL, and l_2 − MSAL which replaces the $l_{2,1}$-norm in the framework SFSR-MSAL with l_2-norm the same as that defined in FME [32], i.e., $\overline{D} = I_c$ and $\widetilde{D} = I_c$. In the trials, we compare the obtained results over two groups of data that differ in the level of relatedness among source and target knowledge. Specifically, we extracted 6 unrelated classes[9] (harp, microwave, fire-truck, cowboy-hat, snake, bonsai) and 6 related classes (all vehicles: bulldozer, fire-truck, motorbikes, school-bus, snowmobile, car-side) from Caltech-256, and is augmented with $b \in \{1, 5, 10, 15, 20, 25, 30\}$ web images per category that are collected through textual search using Bing. For each class used as target, we extracted 20 training and 100 testing samples with half positive and half negative data. Since the web images are randomly obtained from Internet and thereby noises or outliers may abound in the target training data by nature.

The experimental results of the aforementioned datasets are reported in Fig. 11, where each result corresponds to the averaged classification accuracies of an algorithm in 10 tests. It can be seen from Fig. 11 that our methods SFSR-MSAL and l_2 − MSAL always outperform other methods in all cases due to the consideration of augmenting the discrimination space of the target data. l_2 − MSAL obtains comparable performance with SFSR-MSAL when the number of noise samples is small, e.g., less than 5. However, the proposed SFSR-MSAL is always stable and robust with the changing noise as a result of the robustness of $l_{2,1}$-norm to noise/outlers. It can be also observed that with the increase of noisy or outlier data, the classification accuracy of all methods degrade to some extent. However, our method SFSR-MSAL degrades more slowly than others. This indicates that our method is more robust to noise, thus yielding less performance degradation in some specific learning tasks in which noises or/and outliers may abound. The robustness of the target classifier may be very important for the improvements of generalization performance in such scenario.

Remark 2. The robustness is an important property to encode the effective domain knowledge thus helpful to improve the robust capability of the target model. From a statistical point of view, the $l_{2,1}$-norm based method SFSR-MSAL is more robust than the l_2-norm based l_2 − MSAL, which has been verified by the results from Fig. 11.

[9] We here refer to a class as the combination of 80 object and 80 background images.

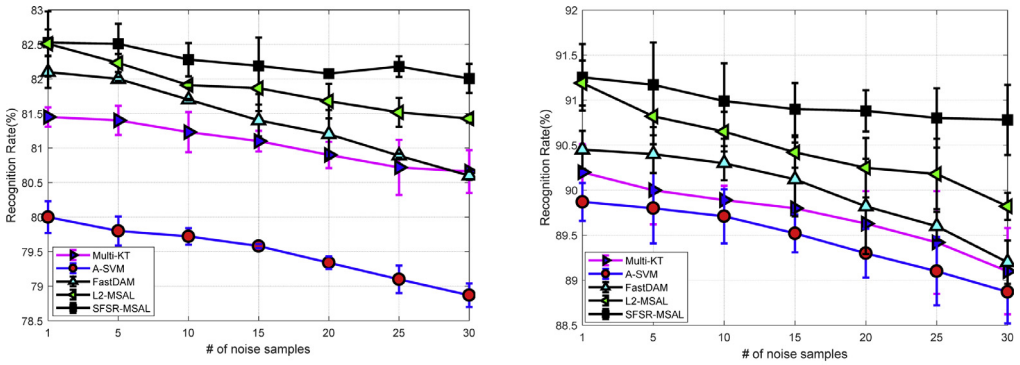


Fig. 11. Recognition rate for different algorithms with different sizes of occlusion and white Gaussian noise on Caltech-256.

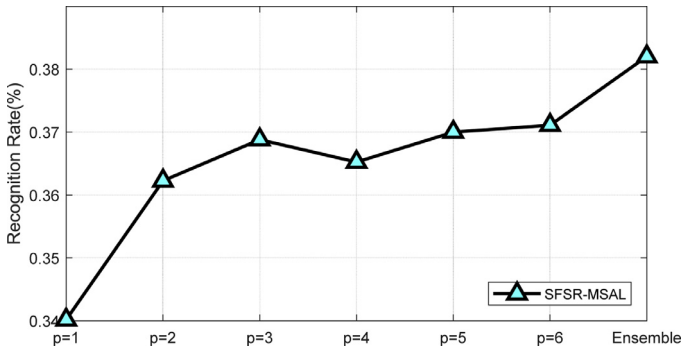


Fig. 12. Average video concept recognition rate with different SFSR-graph Laplacians.

The reason is that the l_2-norm based algorithm is prone to suffering from the noise/outliers compared with the $l_{2,\ 1}$-norm-based one since the effect of noise/outliers may be more exaggerated by the l_2-norm. However, according to the celebrated "No Free Lunch" theorem [39], the proposed SFSR-MSAL has somewhat higher computational complexity than $l_2 -$ MSAL due to the time-consuming iteration procedure in SFSR-MSAL algorithm, which may be a challenging issue in our method when the number of target data is very large.

6.4. Effectiveness of ensemble SFSR-graph Laplacians

In this subsection, we consider to analyze the effectiveness of our ensemble SFSR-graph Laplacian regularization. We give the experimental results in Fig. 12 corresponding to different values of p in SFSR on video recognition in terms of the average recognition rates of all channels. From Fig. 12, we can see that while the performance of SFSR-MSAL could not be consistently improved when the value of p is increased, which is consistent with that reported in [7], the prominently superior performance can be surely obtained when the ensemble SFSR-graph Laplacian regularization is adopted in our method. This further proves the importance of ensemble multiple SFSR-graph Laplacian in our SFSR-MSAL.

Besides, Fig. 13 shows the regularization path for the parameter η (the optimized μ as a function of η) on video concept recognition tasks. We can see that μ has the selection effect for different SFSR-graph Laplacians with different η. In other words, we can obtain relatively sparse solutions in which only a part of weights has very small elements. We also can see that the original sparse graph Laplacian (i.e., $p = 1$) [12] always has the least weight value while the SFSR-graph with $p = 6$ has the largest weight in most cases for different η. This means that sparse graph cannot effectively enlarge the discrimination space of the target data, while our method can provide more discriminative information with $p \geq 2$. Furthermore, from Theorem 4, we can present a clear interpretation of the mechanism of selection property in our trials.

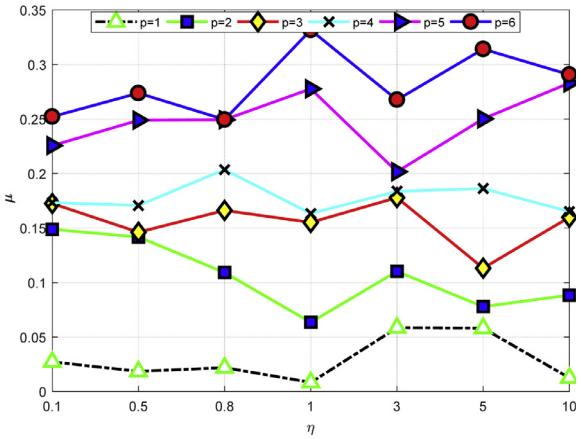


Fig. 13. Regularization path for η on object recognition task with six SFSR-graphs (i.e., $\mu \in \mathbb{R}^6$).

Because $\{\omega_p\}_{p=1}^g$ are sorted in the increasing order, a sequence of non-zero weights $\{\mu_p\}_{p=1}^g$ should be in decreasing order by the definition of the optimal μ_p, given by Theorem 4. Because small ω_p $(= tr(\mu_p(F)^T L_p F))$ means that the score F is smooth on the graph L_p, larger weights are assigned to the smoother graphs and the rest $g - \xi$ graphs with relatively small weight values are relatively not smooth, comparing to the first ξ graphs.

6.5. Convergence study

Since SFSR-MSAL is an iterative algorithm, we empirically check its convergence property still using the above-mentioned visual dataset, where the object recognition dataset is composed of 10 classes. All experimental results are computed on a laptop with 3.0 GHz Intel(R) XEON(TM) E3-1505 M V6 CPU and 24 GB memory, and the codes are

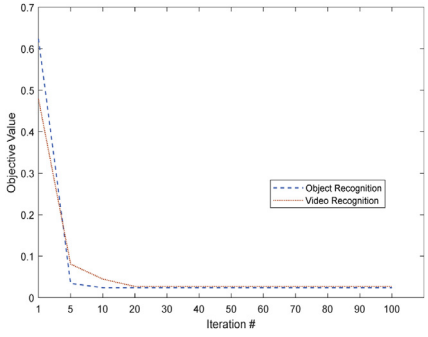


Fig. 14. Convergence of SFSR-MSAL.

written in MATLAB. Empirically, we observe from experiments that the Algorithm 1 performs well in terms of convergence. In Fig. 14, we experimentally demonstrate the convergence, where the objective values (averaged by #examples) of SFSR-MSAL usually converge in less than 20 iterations. Similar observations can be made for other visual tasks in the experiments.

6.6. Parameter sensitivity

In this Section, we conduct empirical analysis on parameter sensitivity using three groups of data extracted from Caltech-256, which differ in the level of relatedness among source and target knowledge, i.e., 6 unrelated classes (harp, microwave, fire-truck, cowboy-hat, snake, bonsai), 6 related classes (all vehicles: bulldozer, fire-truck, motorbikes, school-bus, snowmobile, car-side) and 10 mixed classes (motorbikes, dog, cactus, helicopter, fighter, car-side, dolphin, zebra, horse, goose). For each class used as target, we extracted 20 training and 100 testing samples with half positive and half negative data. The experimental results in Fig. 15 validate that our method SFSR-MSAL can achieve optimal performance under a wide range of parameter values.

- ***Performance variations w.r.t. the regularization parameter*** λ. We run SFSR-MSAL with varying values of the parameter λ. Theoretically, λ controls the weight of source adaptation regularization, and larger values of λ will make the source adaptation more important in SFSR-MSAL. An extreme case is $\lambda \to 0$, where SFSR-MSAL will degenerate to its no transfer variant SFSR-GSSL, which cannot leverage the prior source information with discriminating power. We plot the classification accuracy w.r.t. different values of λ in Fig. 15(a). From this plot we can see that all curves show a sharp upgrade in performance around $\lambda = 0$, which shows the importance of multi-source adaptation in our framework. The performance slowly waves when the value of λ increases. Therefore, we can empirically choose $\lambda \in [10^2, 10^4]$ in the experiments.
- ***Performance variations w.r.t. the regularization parameter*** β. We run SFSR-MSAL with varying values of the parameter β. For the parameter β, we can see from Fig. 15(b) that it should not be small. Large β makes the learned W satisfy the expected properties, which guarantees that better results are achieved. We therefore empirically choose $\beta \in [10^2, 10^5]$.
- ***Performance variations w.r.t. the trade-off parameter*** α. We run SFSR-MSAL with varying values of the parameter α. We plot the classification accuracy w.r.t. different values of α in Fig. 15(c). From this plot we can see that SFSR-MSAL is not sensitive to α when it is in the range of $[10^2, 10^6]$.

7. Conclusion and future work

The scarcity of discriminative information in training samples is a common characteristic of many visual applications [49]. For example, in image classification, large amounts of images can be easily accessed while labeled images are fairly expensive to obtain because they require human effort. Meanwhile, since few labeled visual data do not allow to cover the high intraclass variability, the traditional GSSL methods would provide very few guarantees. This motivates GSSL to extract more discriminative information useful for generalization from training data. To address this issue in specific visual tasks, this paper exploits the feature space embeddings of the target data as well as multi-source prior models to augment the discrimination space for the target function learning. Specifically, the paper proposes a novel SFSR-graph regularization multi-source adaptation framework, which is universal and can be easily degraded into semi-supervised learning by just tuning the regularization parameter. Moreover, to deal with the selection issue of SFSR-graphs, the ensemble SFSR-graph Laplacians is also introduced into SFSR-MSAL. It is worthy to note that our framework also is very general for some specific applications. In other words, the proposed methods including semi-supervised and DA learning could be easily applied to the other related application areas such as multi-labeled learning [50], labeled multiple attribute decision making [51] and Arabic handwritten characters recognition [52,53], just to list a few.

While most of the proposed domain adaptation learning algorithms has achieved significant success in visual classification with different techniques, how to scale up the adaptation learning problem is still an open challenging issue. Note that in SFSR the size of all feature space embedding projection sets of all data will be $n^*C_p^{n-1}$, C_p^{n-1} times of the size of the original dataset, which means the computational complexity of our method on big datasets would be very high. Hence, an important work to be worth studying is about the scalability of the proposed method due to the increasing number of training examples.

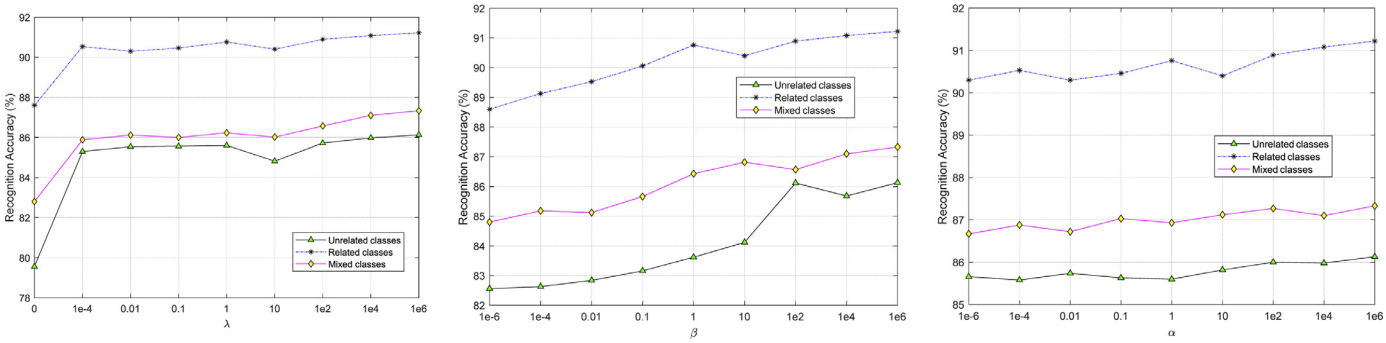


Fig. 15. The sensitivity of proposed model to the choice of parameters.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 61673318, 61703331, and 61502385.

Appendix A. Proof of Theorem 1

Proof. The following conditions can be used for verifying the correctness of conclusion.

(1) When $p = 1$, the object functions (9) and (10) can be represented as the Laplacian regularization formulations [6], respectively.

(2) When $p = 2$, formulas (9) and (10) change into the similarity measure from each point to its label lines. The calculation of distance from the label line $\overline{f_a f_b}$ to f_i is$\|f_i - L^{(2)}_{a,b}(f_i)\|$. The project point can be calculated by $L^{(2)}_{a,b}(f_i) = f_a + \pi_{a,b}(f_a - f_b)$ and $\pi_{a,b} + \pi_{b,a} = 1(i \neq a \neq b)$, then we have

$$\begin{aligned}
&F_1 = \sum_i \|\sum_{a\neq b}(f_i - L^{(2)}_{a,b}(f_i))s^{(2)}_{a,b}(x_i)\|^2 = \sum_i \|\sum_{a\neq b}(f_i - \pi_{b,a}f_a - \pi_{a,b}f_b)s^{(2)}_{a,b}(x_i)\|^2 \\
&= \sum_i \left\|f_i - \sum_j M_{i,j}f_j\right\|^2 = tr(F(I-M)^T(I-M)F^T) \\
&= tr(F(D-\overline{W})F^T) = tr(FLF^T)
\end{aligned} \tag{48}$$

where, $M_{i,b} = \sum_a s^{(2)}_{a,b}(x_i)q_{a,b}$, $q_{a,b} = (x_i - x_a)^T(x_a - x_b)/(x_a - x_b)^T(x_a - x_b)$ and $\sum_j M_{i,j} = 1$. According to the conclusion of [6], the matrix $\overline{W}$ is set as: when $i \neq j$, $\overline{W}_{i,j} = (M + M^T - M^TM)_{i,j}$, $\overline{W}_{i,j} = 0$ otherwise. Then the object function (9) can be represented as a Laplacian regularization.

On the other hand, the object function (10) can be divided into C_2^{n-1} elements and each element represents the quadratic sum of distance from each label object f_i to its $1 \le k \le C_2^{n-1}$ label lines. Considering the first item, the matrix $M_{i,\ j}(1)$ represents the relationship matrix of x_i to the feature line $\overline{x_a x_b}(i,a,\ b = 1,\ \ldots,\ n,\ and\ i \neq a \neq b)$. Two non-zero items$M_{i,\ b}(1)andM_{i,\ a}(1)$exist in each row of matrix $M_{i,\ j}(1)$ and $\sum_j M_{i,j}(1) = 1$. Generally speaking, if the feature line $\overline{x_a x_b}$ is the sparse reconstruction object of pointx_i, i.e. $s^{(2)}_{a,b}(x_i) > 0$, then $M_{i,b}(k) = s^{(2)}_{a,b}(x_i)q_{a,b}$, $M_{i,a}(k) = s^{(2)}_{a,b}(x_i)q_{b,a}(i \neq a \neq b)$, or $M_{i,b}(k) = M_{i,a}(k) = 0$. The above C_2^{n-1} elements can be represented as the Laplacian regularization formulation as follows.

$$\begin{aligned}
&F_2 = \sum_i \sum_{a\neq b} \|f_i - L^{(2)}_{a,b}(f_i)\|^2 s^{(2)}_{a,b}(f_i) = \sum_i \sum_{a\neq b} \|f_i - \pi_{b,a}f_a - \pi_{a,b}f_b\|^2 s^{(2)}_{a,b}(f_i) \\
&= \sum_i \left\|f_i - \sum_j M_{i,j}(1)f_j\right\|^2 + \sum_i \left\|f_i - \sum_j M_{i,j}(2)f_j\right\|^2 + \cdots + \sum_i \left\|f_i - \sum_j M_{i,j}(C_2^{n-1})f_j\right\|^2 \\
&= tr(F(I-M(1))^T(I-M(1))F^T) + tr(F(I-M(2))^T(I-M(2))F^T) \\
&+ \cdots + tr(F(I-M(C_2^{N-1}))^T(I-M(C_2^{n-1}))F^T) \\
&= tr(F(D(1)-\overline{W}(1))F^T) + tr(F(D(2)-\overline{W}(2))F^T) + \cdots + tr(F(D(2)-\overline{W}(C_2^{n-1}))F^T) \\
&= tr(F(D-\overline{W})F^T) = tr(FLF^T)
\end{aligned} \tag{49}$$

where, $\overline{W}_{i,j}(k) = (M(k) + M(k)^T - M(k)^TM(k))_{i,j}$, and $D = \frac{1}{C_2^{n-1}}(D(1) + D(2) + \cdots D(C_2^{n-1}))$.

Then the object function (10) also can be represented as a Laplacian regularization.

(3) When$p = 3$, the formulas (9) and (10) represent the distance from the object f_i to its label surfaces, i.e. $f_i - L^{(3)}_{r,a,b}(f_i)$. And the label surfaces are generated by objectsf_r,$f_a$andf_b. The embedding (or projection) object $L^{(3)}_{r,a,b}(f_i)$of the label surface is got from the following formula:

$$\begin{aligned}
&L^{(3)}_{r,a,b}(f_i) = F_{r,a,b}(F^T_{r,a,b}F_{r,a,b})^{-1}F^T_{r,a,b}(f_i - f_r) + f_r \\
&= [(f_a - f_r)\ \ (f_b - f_r)][\lambda_a\ \ \lambda_b]^T + f_r \\
&= (1 - \pi_a - \pi_b)f_r + \pi_a f_a + \pi_b f_b \\
&= \pi_r f_r + \pi_a f_a + \pi_b f_b
\end{aligned} \tag{50}$$

where, $F_{r,a,b} = [(f_a - f_r)\ \ (f_b - f_r)]$ is a $r \times 2$ matrix. Matrix $(F^T_{r,a,b}F_{r,a,b})^{-1}F^T_{r,a,b}(f_i - f_r)$ is represented as $[\pi_a\ \ \pi_b]^T$, and $\pi_r + \pi_a + \pi_b = 1$. Then the projection point $L^{(3)}_{r,a,b}(f_i)$ is represented as the liner combination of objectsf_r,$f_a$andf_b. By the same way with $p = 2$, formulas (9) and (10) can be represented as $\sum_i \left\|f_i - \sum_j M_{i,j}f_j\right\|^2$ and the assignment of weights ($M_{i,\ r}$, $M_{i,\ a}$, $M_{i,\ b}$) in matrix M is:

$(M_{i,r}, M_{i,a}, M_{i,b}) = (q_r s^{(3)}_{r,a,b}(x_i),\ q_a s^{(3)}_{r,a,b}(x_i),\ q_b s^{(3)}_{r,a,b}(x_i))$,

$[q_a\ \ q_b]^T = (X^T_{r,a,b}X_{r,a,b})^{-1}X^T_{r,a,b}(x_i - x_r)$,

$X_{r,a,b} = [(x_a - x_r)\ \ (x_b - x_r)]$,

$q_r + q_a + q_b = 1$.

Then we can derive the Laplacian regularization formulations of formula (15) and (16) with the same way as the case of$p = 2$.

(4) Generally, for the condition of$p > 3$, formulas (9) and (10) represent the distance of object f_i to its label spaces, i.e. $f_i - L^{(p)}_{1:p}(f_i)$. The projection point $L^{(p)}_{1:p}(f_i)$ of object f_i in the label space is represented as the liner combination of p objects off_1, f_2, ..., f_p.

$$L^{(p)}_{1:p}(f_i) = F_{(1:p)}(F^T_{(1:p)}F_{(1:p)})^{-1}F^T_{(1:p)}(f_i - f_1) + f_1 = \sum_{j=1}^{p} \pi_j f_j, \tag{51}$$

where, $F_{(1:p)} = [(f_2 - f_1)\ \ (f_3 - f_1)\ \ \ldots\ \ (f_p - f_1)]$ is a $c \times (p-1)$ matrix and $\sum_{j=1}^{p} \pi_j = 1$. The distance from object f_i to the projection point of its label spaces is:

$$\sum_i \|f_i - L_{1:p}^{(p)}(f_i)\|^2 s_{1:p}^{(p)} = \sum_i \left\| f_i - \sum_j M_{i,j} f_j \right\|^2, \tag{52}$$

Then the object functions (9) and (10) also can be represented as the Laplacian regularization formulation, respectively. In sum, when$p \geq 1$, the object functions (9) and (10) can be represented as a Laplacian regularization.

Appendix B. Proof of Theorem 6

Proof. Let $\overline{\mathbf{Y}} = UY + \lambda F^s$. According to Definition 1, we have $\|\overline{\mathbf{Y}}\|_H \leq \mathbb{C}$. The empirical Rademacher complexity of the function class $\psi_{SFSR-MSAL}$ is computed as

$$\begin{aligned}
&\widehat{R}_n(\psi_{SFSR-MSAL}) = \left(\tfrac{1}{n_l} + \tfrac{1}{n_u}\right) E_\sigma \left[\sup_{f:\|\overline{\mathbf{Y}}\|_H \leq \mathbb{C}} \overline{\mathbf{Y}}^T (\mathrm{Y})^{-1} \sigma \right] \\
&\leq \left(\tfrac{1}{n_l} + \tfrac{1}{n_u}\right) E_\sigma \left[\sup_{f:\|\overline{\mathbf{Y}}\|_H \leq \mathbb{C}} \|\overline{\mathbf{Y}}\|_F \, \|(\mathrm{Y})^{-1}\sigma\|_2 \right] \\
&\leq \left(\tfrac{1}{n_l} + \tfrac{1}{n_u}\right) \mathbb{C} E_\sigma \left[\sqrt{\textstyle\sum_{i,j=1}^n \sigma_i \sigma_j ((\mathrm{Y})^{-2})_{ij}} \right] \\
&\leq \left(\tfrac{1}{n_l} + \tfrac{1}{n_u}\right) \mathbb{C} \sqrt{\tfrac{2 n_l n_u}{n^2} tr((\mathrm{Y})^{-2})} = \mathbb{C} \sqrt{\tfrac{2}{n_l n_u} tr(\mathrm{Y})^{-2}}
\end{aligned} \tag{53}$$

where the first inequality holds due to the Cauchy–Schwarz's inequality and the third inequality holds due to the Jensen's inequality.

References

[1] M. Ghifary, D. Balduzzi, W.B. Kleijn, et al., Scatter component analysis: a unified framework for domain adaptation and domain generalization, IEEE Trans. Pattern Anal. Mach. Intell. 99 (2017) 1-1.
[2] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, Proc. 20th Int. Conf. on Machine Learning, 2003.
[3] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from examples, J. Mach. Learn Res. 7 (2006) 2399–2434.
[4] D. Zhou, O. Bousquet, T. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, Adv. Neural Inf. Process. Syst. 16 (2004).
[5] W. Liu, J. Wang, S.-F. Chang, Robust and scalable graph-based semi-supervised learning, Proc. IEEE 100 (9) (2012) 2624–2638.
[6] S. Yan, D. Xu, B. Zhang, et al., Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40 51.
[7] Y.-N. Chen, C.-C. Han, C.-T. Wang, et al., Face recognition using nearest feature space embedding, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1073–1086.
[8] F. Wang, C. Zhang, Label propagation through linear neighborhoods, IEEE Trans. Knowl. Data Eng. 20 (1) (2008) 55–67.
[9] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[10] J. Wright, A. Yang, S. Sastry, Y. Ma., Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[11] H. Cheng, Z. Liu, J. Yang, Sparsity induced similarity measure for label propagation, IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 2009 Sep. 29-Oct. 2.
[12] M. Fan, N. Gu, H. Qiao, B. Zhang, Sparse regularization for semi-supervised classification, Pattern Recogn. 44 (8) (2011) 1777–1784.
[13] J. Tao, S. Wen, W. Hu, L1-norm locally linear representation regularization multi-source adaptation learning, Neural Netw. 69 (2015) 80–98.
[14] L. Bruzzone, M. Marconcini, Domain adaptation problems: a DASVM classification technique and a circular validation strategy, IEEE Trans. PAMI 32 (5) (2010) 770–787.
[15] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[16] V.M. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: a survey of recent advances, IEEE Signal Process. Mag. 32 (3) (2015) 53–69.
[17] L. Duan, I.W. Tsang, D. Xu, Domain transfer multiple kernel learning, IEEE Trans. Pattern Anal. Mach. Intell. (Mar. 2012) 465–479.
[18] L. Duan, D. Xu, I.W. Tsang, Domain adaptation from multiple sources: a domain-dependent regularization approach, IEEE Trans. Neural Netw. Learn. Syst. (Mar. 2012) 504–518.
[19] J. Yang, R. Yan, A.G. Hauptmann, Cross-domain video concept detection using adaptive SVMs, Proc. ACM Int. Conf. on Multimedia, 2007, pp. 188–197.
[20] T. Tommasi, F. Orabona, B. Caputo, Learning categories from few examples with multi model knowledge transfer, IEEE Trans. Pattern Anal. Mach. Intell. 36 (5) (2014) 928–941.
[21] B. Geng, D. Tao, C. Xu, DAML: Domain adaptation metric learning, IEEE Trans. Image Process. 20 (10) (2011) 2980–2989.
[22] M. Long, J. Wang, G. Ding, S.J. Pan, S.Y Philip, Adaptation regularization: a general framework for transfer learning. IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1076–1089.
[23] J. Tao, K.F.-L. Chung, S. Wang, On minimum distribution discrepancy support vector machine for domain adaptation, Pattern Recogn. 45 (11) (2012) 3962–3984.
[24] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, Proc. ECCV, 2010, pp. 213–226.
[25] R. Gopalan, R. Li, R. Chellappa, Domain adaptation for object recognition: an unsupervised approach, Proc. ICCV, 2011, pp. 999–1006.
[26] H. Wang, F. Nie, H. Huang, Robust and discriminative self-taught learning, ICML (3) (2013) 298–306.
[27] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 2012.
[28] B. Kulis, K. Saenko, T. Darrell, What you saw is not what you get: domain adaptation using asymmetric kernel transforms, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[29] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Netw. 22 (2) (2011) 199–210.
[30] R. El-Yaniv, D. Pechyony, Transductive rademacher complexity and its applications, J. Artif. Intell. Res. (JAIR) 35 (2009) 193–234.
[31] Z. Li, J. Liu, J. Tang, et al., Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2085–2098.
[32] F. Nie, D. Xu, I.W.-H. Tsang, C. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction, IEEE Trans. Image Process. 19 (7) (2010) 1921–1932.
[33] S. Gao, I.W. Tsang, L.T. Chia, Kernel sparse representation for image classification and face recognition, Proc. 11th European Conference on Computer Vision (ECCV 2010), Crete, Greece, September 2010.
[34] F.R.K. Chung, Spectral graph theory, Regional conference series in mathematics, 2011.
[35] B. Geng, D. Tao, C. Xu, et al., Ensemble manifold regularization, IEEE Trans. Pattern Anal. Mach. Intell. 34 (6) (2012) 1227–1233.
[36] M. Karasuyama, H. Mamitsuka, Multiple graph label propagation by sparse integration, IEEE Trans. Neural Netw. Learn Syst. 24 (12) (2013) 1999–2012.
[37] A. Gretton, Z. Harchaoui, K. Fukumizu, B. Sriperumbudur, A fast, consistent kernel two-sample test, Advances in Neural Information Processing Systems 22 MIT Press, 2010, pp. 673–681.
[38] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.
[39] R. Duda, P. Hart, D. Stork, Pattern Classification, Second edn., Wiley, New York, 2001.
[40] L. Peel, Graph-based semi-supervised learning for relational networks, SDM (2017) 435–443.
[41] X. Wang, W. Huang, Y. Cheng, et al., Multisource domain attribute adaptation based on adaptive multi-kernel alignment learning, IEEE Trans. Syst. Man. Cybern. Syst. 99 (2018) 1–12.
[42] J. Tao, D. Zhou, B. Zhu, Multi-source adaptation embedding with feature selection by exploiting correlation information, Knowl.-Based Syst. 143 (2018) 208–224.
[43] J. Tao, D. Song, S. Wen, W. Hu, Robust multi-source adaptation visual classification using supervised low-rank representation, Pattern Recogn. 61 (2017) 47–65.
[44] J. Tao, S. Wen, W. Hu, Multi-source adaptation learning with global and local regularization by exploiting joint kernel sparse representation, Knowl.-Based Syst. 98 (2016) 76–94.
[45] AS Mozafari, M. Jamzad, Cluster-based adaptive SVM: a latent subdomains discovery method for domain adaptation problems, Comput. Vis. Image Und. 162 (2017) 116–134.
[46] T. Yao, Y. Pan, C.W. Ngo, et al., Semi-supervised domain adaptation with subspace learning for visual recognition, Comput. Vis. Pattern Recogn. IEEE, (2015) 2142–2150.

[47] B. Chen, W. Lam, I.W. Tsang, et al., Discovering low-rank shared concept space for adapting text mining models, IEEE Trans. Pattern Anal. Mach. Intell. 35 (6) (2013) 1284–1297.
[48] L. Wen, X. Zheng, X. Dong, et al., Domain generalization and adaptation using low rank exemplar SVMs, IEEE Trans. Pattern Anal. Mach. Intell. (99) (2017) 1-1.
[49] H. Tao, C. Hou, F. Nie, et al., Scalable multi-view semi-supervised classification via adaptive regression, IEEE Trans. Image Process. A Publication of the IEEE Signal Process. Society (99) (2017) 1-1.
[50] J. Xuan, J. Lu, G. Zhang, et al., A Bayesian nonparametric model for multi-label learning, Mach. Learn. 106 (11) (2017) 1–29.
[51] M. Suo, B. Zhu, Y. Zhang, et al., Fuzzy Bayes risk based on Mahalanobis distance and Gaussian kernel for weight assignment in labeled multiple attribute decision making, Knowl.-Based Syst. (2018), http://dx.doi.org/10.1016/j.knosys.2018.04.002.
[52] Y. Boulid, A. Souhar, M.E. Elkettani, Handwritten character recognition based on the specificity and the singularity of the Arabic language, Int. J. Inter. Multi. Art. Intell. 4 (4) (2017) 45–53.
[53] A. Souhar, Y. Boulid, E. Ameur, M.M. Ouagague, segmentation of Arabic handwritten documents into text lines using watershed transform, Int. J. Inter. Multi. Art. Intell. 4 (6) (2017) 96–102.
[54] C. Wang, M. Shao, Q. He, et al., Feature subset selection based on fuzzy neighborhood rough sets, Knowl.-Based Syst. 111 (2016) 173–179.