软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(5):977−1000 [doi: 10.13328/j.cnki.jos.004575]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

稀疏标签传播:一种鲁棒的领域适应学习方法*

陶剑文[1], Fu-Lai CHUNG[2], 王士同[2,3], 姚奇富[4]

[1](浙江大学 宁波理工学院 信息科学与工程学院,浙江 宁波 315100)

[2](香港理工大学 电子计算学系,香港)

[3](江南大学 数字媒体学院,江苏 无锡 214122)

[4](浙江工商职业技术学院 电子与信息工程学院,浙江 宁波 315012)

通讯作者: 陶剑文, E-mail: jianwen_tao@aliyun.com

摘 要: 稀疏表示因其所具有的鲁棒性,在模式分类领域逐渐得到关注.研究了一种基于稀疏保留模型的新颖领域适应学习方法,并提出一种鲁棒的稀疏标签传播领域适应学习(sparse label propagation domain adaptation learning,简称 SLPDAL)算法.SLPDAL 通过将目标领域数据进行稀疏重构,以实现源领域数据标签向目标领域平滑传播.具体来讲,SLPDAL 算法分为 3 步:首先,基于领域间数据分布均值差最小化准则寻求一个优化的核空间,并将领域数据嵌入到该核空间;然后,在该嵌入核空间,基于 l_1-范最小化准则计算各领域数据的核稀疏重构系数;最后,通过保留领域数据间核稀疏重构系数约束,实现源领域数据标签向目标领域的传播.最后,将 SLPDAL 算法推广到多核学习框架,提出一个 SLPDAL 多核学习模型.在鲁棒人脸识别、视频概念检测和文本分类等领域适应学习任务上进行比较实验,所提出的方法取得了优于或可比较的学习性能.

关键词: 领域适应学习;稀疏表示;标签传播;最大均值差;多核学习

中图法分类号: TP181

中文引用格式: 陶剑文,Chung FL,王士同,姚奇富.稀疏标签传播:一种鲁棒的领域适应学习方法.软件学报,2015,26(5):977−1000. http://www.jos.org.cn/1000-9825/4575.htm

英文引用格式: Tao JW, Chung FL, Wang ST, Yao QF. Sparse label propagation: A robust domain adaptation learning method. Ruan Jian Xue Bao/Journal of Software, 2015,26(5):977−1000 (in Chinese). http://www.jos.org.cn/1000-9825/4575.htm

Sparse Label Propagation: A Robust Domain Adaptation Learning Method

TAO Jian-Wen[1], Fu-Lai CHUNG[2], WANG Shi-Tong[2,3], YAO Qi-Fu[4]

[1](School of Information Science and Engineering, Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China)

[2](Department of Computing, Hong Kong Polytechnic University, Hong Kong, China)

[3](School of Digital Media, Jiangnan University, Wuxi 214122, China)

[4](School of Information Engineering, Zhejiang Business Technology Institute, Ningbo 315012, China)

Abstract: Sparse representation has received an increasing amount of interest in pattern classification due to its robustness. In this paper, a domain adaptation learning (DAL) approach is explored based on a sparsity preserving model, which assumes that each data point can be sparsely reconstructed. The proposed robust DAL algorithm, called sparse label propagation domain adaptation learning (SLPDAL), propagates the labels from labeled points in the source domain to the unlabeled dataset in the target domain using those sparsely reconstructed objects with sufficient smoothness. SLPDAL consists of three steps. First, it finds an optimal kernel space in which all samples from both source and target domains can be embedded by minimizing the mean discrepancy between these two domains. Then, it

* 基金项目: 教育部人文社会科学研究规划基金(13YJAZH084); 浙江省自然科学基金(LY14F020009); 宁波市自然科学基金(2013A610065, 2013A610072); 香港理工大学基金(G-UA68)

收稿时间: 2013-01-21; 修改时间: 2013-09-04; 定稿时间: 2014-01-10; jos 在线出版时间: 2014-08-19

CNKI 网络优先出版: 2014-08-19 14:39, http://www.cnki.net/kcms/doi/10.13328/j.cnki.jos.004575.html

computes the best kernel sparse reconstructed coefficients for each data point in the kernel space by using l_1-norm minimization. Finally, it propagates the labels of source domain to the target domain by preserving the kernel sparse reconstructed coefficients. The paper also derives an easy way to extend SLPDAL to out-of-sample data and multiple kernel learning respectively. Promising experimental results have been obtained for several DAL problems such as face recognition, visual video detection and text classification tasks.

Key words: domain adaptation learning; sparse representation; label propagation; maximum mean discrepancy; multiple kernel learning

在数据挖掘和机器学习领域,为非独立且同分布(independent and identically distributed,简称 IID)数据构建学习模型是近年来出现的热门研究主题[1].为了有效地解决非 IID 数据学习问题,研究者提出了领域适应学习(domain adaptation learning,简称 DAL)[2]方法.该方法利用某些不同但相关的源(或辅助)领域数据的有监督学习来实现目标领域数据分类.DAL 在许多实际应用中有较广泛的需求,并逐渐得到研究者的大量关注[2−9].目前已提出大量学习方法以解决视频概念检测[4]、文本分类[3]、人脸识别和图像标注[7]等领域适应问题,其中大多数方法侧重于统计分类器的适应,即所谓归纳学习,而对于演绎适应学习,鲜有相关研究工作开展.直观上来说,演绎 DAL 方法在具体实践中可取得更优的效能.另外,现有 DAL 方法同样需要来自源领域充足的标签训练数据以实现知识迁移,换句话说,若标签训练数据不足或有限,DAL 性能在具体应用中则会在一定程度上有所下降.

在过去的几年里,基于图的半监督学习(semi-supervised learning,简称 SSL)方法因其优雅的数学形式和独特的效能,已发展成为机器学习领域热门研究主题之一[10].这些 SSL 方法一般都假设训练数据和测试数据抽取自某个相同的特征分布或特征空间,而当数据分布发生变化时,这些利用先验信息习得的模型需要再次采用新的训练数据进行重构.在 DAL 中,如果忽略领域差异,将源领域数据作为带标签的训练样本和目标领域无标签数据作为测试样本,DAL 则降级为 SSL 问题.直观上来说,SSL 算法可被直接用于领域适应问题,其间的细微差别在于:(1) 在 SSL 中带标签的训练数据量相对于 DAL 中的较小;(2) SSL 中训练数据和测试数据均取自某个相同但未知的数据分布,而 DAL 则来自不同但相关的数据分布.已有几项研究工作将 SSL 扩展为 DAL 方法:Dai 等人[11]提出一种基于期望最大的 DAL 算法,除了利用领域间 Kullback-Leibler 散度[6]最小化来估计标签样本和无标签样本间平衡参数外,该方法等价于半监督期望最大算法[12];文献[13]中,Xing 等人采取在最近邻图上传播标签的方法,提出一种桥接精炼 DAL 算法,其相似于基于图的 SSL 算法.但是研究结果得知,基于图的 SSL 算法性能在一定程度上依赖于图边的权值,其通常采用 *k*-最近邻(*k*-nearest neighbor,简称 *k*-NN)或 Gaussian 核相似性(Gaussian kernel similarity,简称 GKS)[14,15]等方法计算取得,而且,传统 SSL 方法通常对噪声(或桥接点)[14]和缺失数据(如人脸数据中的遮罩)[16]较敏感,会导致将标签错误地传播到不同类.为此,Wang 等人[14]提出一种线性邻居传播算法(linear neighbor propagation,简称 LNP),该方法虽然能通过修正数据点与其 *k*-NN 间的权值来改善传统 *k*-NN 方法的性能,但是其仍然依赖传统的欧式距离来预定义数据点的 *k*-NN,换句话说,LNP 依然未能彻底解决传统 SSL 方法中如何有效确定数据点的邻居数这个根本问题.

近年来,稀疏表示(sparse representation,简称 SR)[16]技术在机器学习和模式识别(特别是在人脸识别[16−19]和图像分类[20])等领域得到有效应用,并展现出其独特的鲁棒性能.针对上述传统图构建存在的问题,Fan 等人[18]提出一种稀疏正则化半监督分类算法;文献[17]提出采用稀疏表示来度量样本间的相似性,以实现样本间的标签传播;与文献[17]思想相似,文献[21]提出一种基于稀疏表示的 SSL 方法,通过新构建的稀疏图实现标签的有效传播;Cheng 等人[22]明确提出一种鲁棒的稀疏(或 l_1-范最小)驱动的有向图(或 l_1-图),并将该图分别应用于普聚类、子空间学习和 SSL,取得了良好的效果.

本文基于图 SSL 模型,利用核稀疏表示技术,提出了一种稀疏标签传播 DAL 算法(sparse label propagation domain adaptation learning,简称 SLPDAL):首先,基于领域间数据分布均值差最小化准则寻求一个优化的核空间,并将领域数据嵌入到该核空间;然后,在该嵌入核空间,基于 l_1-范最小化准则计算各领域数据的核稀疏重构系数;最后,通过保留领域数据间核稀疏重构系数约束,实现源领域数据标签向目标领域传播.从上述分析来看,在数据图的构建上,虽然本文方法与文献[17,18,21−23]具有相似之处,但明显不同的是:现有方法都是针对 IID 数据在原始输入空间(如文献[17,18,21,22])或某个核空间(如文献[23])构建一个用以标签传播的稀疏图;而本文方法是针对非 IID 数据学习问题,在一个基于分布差最小优化的再生核 Hilbert 空间(reproduced kernel hilbert

space,简称 RKHS)构建该稀疏图.特别地,与传统方法相比,SLPDAL 方法所具有的优势在于:

(1) 在继承和发展传统的基于图的 SSL 方法优点的基础上,利用核稀疏表示技术和领域数据分布差最小化准则,提出一种鲁棒有效的全局和局部一致正则化 DAL 方法 SLPDAL,实现源领域数据标签向目标领域传播,并将该方法拓展为多核学习模型 MKSLP.

(2) 由于稀疏表示能保留自然的判别信息,使得 SLPDAL 方法在解决 DAL 问题时只需要相对更少的带标签数据,并且通过使用稀疏集中索引(sparsity concentration index,简称 SCI)[16],SLPDAL 方法能够自动消除噪声数据和恢复缺失数据.

(3) 在 SLPDAL 方法中,标签传播图模型的构建准则为领域数据的核稀疏表示,而该准则一般优于传统 SSL 方法中采用的最近邻准则,尤其是对于高维小样本数据集的学习.与传统 SSL 方法中数据近邻和图边权值的计算分开完成不同,本文方法中邻居大小和图边权值通过 l_1-范优化过程来一步确定,使得领域中不同样本具有不同的近邻数,能够增强本文方法对复杂数据分布学习的自适应性.

(4) 由于所提出的方法框架模型所具有的一般性,在一定的条件变换下,许多现有的 SSL 方法和 DAL 方法能被该模型所恢复.如在忽略领域差条件下,该方法模型能够简单地应用于基于图的 SSL 问题.

由于基于图 Laplacian 正则化 SSL[3]和本文方法的紧密关系,下节将首先讨论基于图的 SSL 模型.

1 图 Laplacian 正则化 SSL

一般来说,存在两种 SSL 任务:1) 演绎学习(或直推学习(transductive learning)),其仅旨在预测无标签顶点的分类标签[14];2) 归纳学习,其试图归纳出一个在整个样本空间具有最低误差率的决策函数[15].显然,归纳学习困难且复杂[14],因此,本文重点关注半监督演绎学习模型.研究结果表明,SSL 问题的关键是先验一致性[15],也称聚类假设或流形假设,其假设相邻数据点可能具有相同标签或具有相同结构(如聚类或子流形[24])的数据具有相同标签,前者为局部假设,而后者为全局假设.近年来,在 SSL 方面取得的一个杰出成就就是基于图的 SSL 模型,其将整个数据集建模为一个图结构 $G=(V,E)$,其中,V 为顶点集,E 为边集.每个边 $e_{ij}\in E$ 被赋予一个非负权值 $w_{ij}\geqslant 0$,以反映数据点对 i 和 j 间的相似性.图 G 可以为有向(即 $w_{ij}\neq w_{ji}$)或无向(即 $w_{ij}=w_{ji}$),本文仅关注无向图的情况.

给定数据集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_l,\boldsymbol{x}_{l+1},\ldots,\boldsymbol{x}_n\}\in\mathbb{R}^d$ 和标签集 $L=\{1,2,\ldots,c\}$,其中,前 l 个点 $\boldsymbol{x}_i(1\leqslant i\leqslant l)$标签为 $y_i\in L$,余下的 $n-l$ 个数据点 $\boldsymbol{x}_u(l+1\leqslant u\leqslant n)$无标签,每个数据点 $\boldsymbol{x}_i$ 均采样自某个固定但未知的分布,则基于图的 SSL 旨在基于由数据集 X 构成的图 G 寻求 c 个优化的分类函数 $f^j(1\leqslant j\leqslant c)$,且满足:1) 优化函数的输出应接近图中带标签顶点的标签值;2) 优化函数的输出应在整个图上平滑.从而,基于图的 SSL 一般框架旨在最小化:

$$J(f)=\lambda\sum_{j=1}^{c}\sum_{i=1}^{l}\delta(f^j(\boldsymbol{x}_i),y_i)+\beta\sum_{j=1}^{c}R(f^j) \tag{1}$$

其中,$\delta(\cdot,\cdot)$表示损失函数(如 hinge 损失函数或平方损失函数),以度量标签数据的预测值和期望值间的不一致性;$R(f^j)$为惩罚正则项,以约束函数 f^j 在本质数据流形上的平滑性;λ和β为两个正则化参数,以分别控制损失和平滑项间的平衡.惩罚正则项可采用如下一般形式:

$$R(F)=tr(F^TQF) \tag{2}$$

其中,$F=(\boldsymbol{F}^1,\boldsymbol{F}^2,\ldots,\boldsymbol{F}^c)\in\mathbb{R}^{n\times c}$ 为类指示矩阵,$\boldsymbol{F}^j=[f^j(\boldsymbol{x}_1),f^j(\boldsymbol{x}_2),\ldots,f^j(\boldsymbol{x}_n)]^T$,$Q$ 为一个 $n\times n$ 平滑矩阵.

定义 1(图 Laplacian 正则化). 如果 $Q=L$ 且 $Q\boldsymbol{e}=0$,则 $R(f)$称为一阶图 Laplacian 正则化,其中 $L=D-W\in\mathbb{R}^{n\times n}$ 为图 Laplacian,$W=[w_{ij}]_{n\times n}$ 为图边权值矩阵,$D=diag(d_1,d_2,\ldots,d_n)$为一对角度矩阵,其中,$d_i=\sum_j w_{ij}$,$\boldsymbol{e}$ 为一 n-维全 1 向量;如果 $Q=(I-W)^T(I-W)$,且 $Q\boldsymbol{e}=0$,则称 $R(f)$为二阶图 Laplacian 正则化.

如果选$\delta(\cdot,\cdot)$为平方损失函数,$R(f)$为基于 Laplacian 算子的图 Laplacian 正则化,则公式(1)可描述为

$$\min\ \lambda tr((F-Y)^TC(F-Y))+\beta tr(F^TQF) \tag{3}$$

其中,$Y\in\mathbb{R}^{n\times c}$ 为类标签矩阵,且若 $\boldsymbol{x}_i$ 被标识为 $y_i=j$,则 $Y_{ij}=1$;否则,$Y_{ij}=0$.$Q\in\mathbb{R}^{n\times n}$ 称为图 Laplacian,$C\in\mathbb{R}^{n\times n}$ 为对角矩阵,其前 l 个对角元素 $C_{ii}=C_l>0(1\leqslant i\leqslant l)$,余下对角元素 $C_{ii}=C_u\geqslant 0(l+1\leqslant i\leqslant n)$,$C_l$ 和 C_u 为两个参数.可以很容易地

推导出公式(3)的解析解为

$$F=\lambda(\lambda C+\beta Q)^{-1}CY \tag{4}$$

从而 $\boldsymbol{x}_i(l+1\leqslant i\leqslant n)$的预测标签由下式确定:

$$y_i=\arg\max_{1\leqslant j\leqslant c}F_{ij},\ l+1\leqslant i\leqslant n \tag{5}$$

图 Laplacian Q 和(或)对角矩阵 C 在不同设置条件下,现有大多数基于图的 SSL 方法能够被统一到公式(3)框架,例如在文献[25]中,Q 设置为一个组合一阶图 Laplacian 且 $C_l=\infty$,$C_u=0$;在文献[26]中,Q 被设置为规范化一阶图 Laplacian 且 $C_l=C_u=1$.以上两种图 Laplacian 均基于聚类假设,而在文献[27]中,Q 被设置为基于局部学习假设的二阶图 Laplacian 且 $C_l=1$,$C_u=0$;在文献[14]中,Q 被设置为基于局部线性嵌入思想的二阶图 Laplacian 且 $C_l=C_u=1$;在文献[28]中,Q 被设置为混合图 Laplacian.

2 稀疏标签传播 DAL

2.1 问题描述

对于 DAL 问题,本文将训练领域定义为源领域,其中有充足的带标签训练数据;将测试领域定义为目标领域,其中带标签的数据不存在或非常有限.对于一个模式分类问题,一个领域 D 由某个潜在的真实数据分布 $P(\boldsymbol{x},y)$给出,其中,$\boldsymbol{x}\in X$ 为样本集,$y\in Y$ 为相应的类标签集.对于 DAL,无标签的测试数据集 $X^t=\{(\boldsymbol{x}_j^t)\}_{j=1}^m,\boldsymbol{x}_j^t\in X$,抽取自目标领域 D^t,带标签的训练数据集 $X^s=\{(\boldsymbol{x}_i^s,y_i^s)\}_{i=1}^n,\boldsymbol{x}_i^s\in X,y_i^s\in Y$ 抽取自某个与 D^t 不同但相关的源领域 D^s.令源领域数据分布 $P^s(\boldsymbol{x},y)=P^s(y|\boldsymbol{x})\cdot P^s(\boldsymbol{x})$和目标领域数据分布 $P^t(\boldsymbol{x},y)=P^t(y|\boldsymbol{x})\cdot P^t(\boldsymbol{x})$是两个潜在的真实数据分布,且 $P^s(\boldsymbol{x},y)\neq P^t(\boldsymbol{x},y)$.事实上,绝大多数 DAL 方法均假设存在两个不同但高度相关的领域[3,5].DAL 的关键思想是,通过某种分布变换技术来减小 $P^t(\boldsymbol{x},y)$和 $P^s(\boldsymbol{x},y)$间的差异[2,3,6].在 DAL 研究中,常用的分布距离度量方法包括基于熵概念的 Kullback-Leibler 散度[5]和基于统计概念的最大均值差(maximum mean discrepancy,简称 MMD)[29]等.现有研究结果显示,MMD 度量准则能够更有效地估计在某个再生核 Hilbert 空间两个分布间的距离.

通过某个非线性映射 $\phi:\mathbb{R}^d\to\boldsymbol{H}$,可将原始空间问题变换为再生核 Hilbert 空间(RKHS) $\boldsymbol{H}$ 中的问题[30].对于某个恰当选择的映射 ϕ,在空间 $\boldsymbol{H}$ 中,内积$\langle\cdot,\cdot\rangle$算子定义为$\langle\phi(\boldsymbol{x}_1),\phi(\boldsymbol{x}_2)\rangle_H=K(\boldsymbol{x}_1,\boldsymbol{x}_2)$,其中,$\boldsymbol{x}_1,\boldsymbol{x}_2\in X$,且 $K(\cdot,\cdot):X\times X\to\mathbb{R}$为一半正定核函数.在 RKHS 中,度量两个分布间距离的 MMD 可定义如下:

定义 2(MMD)[29]. 设 p 和 q 为定义于领域 D 上的分布,令 F 为某个函数类,且 $f(\in F):X\to\mathbb{R}$.

给定观测集 $X^s=\{(\boldsymbol{x}_i^s,y_i^s)\}_{i=1}^n$ 和 $X^t=\{(\boldsymbol{x}_j^t)\}_{j=1}^m$, MMD 及其经验估计定义为

$$\left.\begin{aligned}MMD[F,p,q]&=\sup_{f\in F}(E_{X^s\in p}[f(\boldsymbol{x}^s)]-E_{X^t\in q}[f(\boldsymbol{x}^t)])\\MMD[F,X^s,X^t]&=\sup_K\left(\frac{1}{n}\sum_{i=1}^n\phi(\boldsymbol{x}_i^s)-\frac{1}{m}\sum_{j=1}^m\phi(\boldsymbol{x}_j^t)\right)\end{aligned}\right\} \tag{6}$$

基于 MMD 准则和稀疏保留假设[19],本文提出 SLPDAL 算法,其将传统的基于图的 SSL 算法有效扩展到 DAL 领域.

2.2 SLPDAL算法

本节将形式化地提出 SLPDAL 算法.令 $X^s=\{\boldsymbol{x}_1^s,\boldsymbol{x}_2^s,...,\boldsymbol{x}_n^s\}$ 和 $X^t=\{\boldsymbol{x}_1^t,\boldsymbol{x}_2^t,...,\boldsymbol{x}_m^t\}$ 为分别来自源领域和目标领域的两个数据点集.设 $X=\{\{\boldsymbol{x}_i^s\}_{i=1}^n,\{(\boldsymbol{x}_j^t\}_{j=1}^m\}$ 代表$\mathbb{R}^d$ 空间 $n+m$ 个数据点,$\overline{L}=\{+1,-1\}$ 为标签.数据点 $\boldsymbol{x}_i^s\in X^s$ $(1\leqslant i\leqslant n)$标记为 $L_i\in\overline{L}$,其余数据点 $\boldsymbol{x}_j\in X(n+1\leqslant j\leqslant n+m)$无标签.SLPDAL 的目标是,试图预测数据点 $\boldsymbol{x}_j$ 的标签值.实现该任务需要 3 个步骤:(1) 领域分布核均值匹配;(2) 数据的核稀疏表示;(3) 稀疏标签传播以实现源领域标签向目标领域迁移.

2.2.1 领域分布核均值匹配

当样本数据映射到高维甚至无限维空间时,定义 2 中的 MMD 能够捕捉到数据的高阶统计特征[4].基于此,

Gretton 等人[29]提出核函数 f 的选取原则,即,f 为 RKHS 中的单位球.这样,两个领域分布距离度量可以简单地表示为 RKHS 中数据分布的均值差,即,源领域和目标领域间最小分布距离为

$$\min_K dist(X^s,X^t)=\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i^s)-\frac{1}{m}\sum_{i=1}^{m}\phi(\boldsymbol{x}_i^t)\right\|_H^2 \tag{7}$$

进一步简化为

$$\min_K dist(X^s,X^t)=tr(X^{\phi}\Pi_{st}(X^{\phi})^T)=tr(\Pi_{st}K_{XX}) \tag{8}$$

其中, $X^{\phi}=[\phi(\boldsymbol{x}_1^s),\phi(\boldsymbol{x}_2^s),...,\phi(\boldsymbol{x}_n^s),\phi(\boldsymbol{x}_1^t),\phi(\boldsymbol{x}_2^t),...,\phi(\boldsymbol{x}_m^t)],K_{XX}=(X^{\phi})^TX^{\phi},\Pi_{st}\in\mathbb{R}^{(n+m)\times(n+m)}$ 定义为

$$\Pi_{st}(i,j)=\begin{cases}\dfrac{1}{n^2}, & \text{if } \boldsymbol{x}_i,\boldsymbol{x}_j\in X^s\\ \dfrac{1}{m^2}, & \text{if } \boldsymbol{x}_i,\boldsymbol{x}_j\in X^t\\ -\dfrac{1}{nm}, & \text{otherwise}\end{cases} \tag{9}$$

现有研究指出:高斯核映射能够提供一个有效的 RKHS 嵌入,使得领域间分布距离的一致性度量得以实现[30].为此,本文采用高斯核函数 $k_{\sigma}(\boldsymbol{x},\boldsymbol{z})=\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{z}\|^2\right)$ 作为 Hilbert 空间映射的再生核函数,其中,$\boldsymbol{x},\boldsymbol{z}\in X$,$\sigma$指核带宽.

定理 1[31]. 假设 A 为一个对称矩阵且 $A=\boldsymbol{P}\Sigma\boldsymbol{P}^T$,其中,$\boldsymbol{P}$ 包含矩阵 A 的正交特征向量列,$\Sigma=diag(\sigma_1,\sigma_2,\ldots,\sigma_n)$ 为相应特征值构成的对角矩阵,b 为一个正常数,则公式(10)中半定规划问题和公式(11)中线性规划问题具有相同的优化解,即,$K^*=\boldsymbol{P}\Gamma\boldsymbol{P}^T$,其中,$\Gamma=diag(\gamma_1,\gamma_2,\ldots,\gamma_n)$.

$$\begin{cases}\min\limits_K tr(AK)\\ \text{s.t. } 0\preceq K\preceq\boldsymbol{I},tr(K)=b\end{cases} \tag{10}$$

$$\begin{cases}\min\limits_{\gamma_i}\sum\limits_{i=1}^{n}\gamma_i\sigma_i\\ \text{s.t. } 0\leqslant\gamma_i\leqslant 1,\sum\limits_{i=1}^{n}\gamma_i=b,1\leqslant i\leqslant n\end{cases} \tag{11}$$

证明:关键步骤是证明矩阵 A 和 K 能被联合对角化,详细过程可参见文献[31]. □

令 $A=\Pi_{st}$ 和 $K=K_{XX}$,我们可以通过求解公式(11)中线性规划问题得到公式(8)的优化解 K^*,这可以利用现存的半定规划软件包来高效地实现.

2.2.2 数据的核稀疏表示

本阶段关注如何通过数据的稀疏表示来构建一个加权图 $G=\{X,S\}$,其中,X 为数据集构成的顶点集,S 为边权值,每条边 $s_{ij}\in S$ 代表数据点对 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间的稀疏关系.如下两条理由能够说明为何数据的稀疏表示适于图的构建:

1) 在典型的 k-邻居图构建中,稀疏性具有重要地位:一方面,稀疏性刻画了数据分布的全局性;另一方面,稀疏性能够有效节省计算成本和存储空间.但是,传统的基于 k-NN 和高斯函数构建的 k-邻居图的稀疏性依赖于人工设定的邻居数和高斯核参数σ.文献[14]研究指出:在只有少量标签数据的情况下,难以可靠地选取模型参数,即,难以确定优化的参数σ.为此,需要寻求一种更可靠、更稳定的方法来构建图模型 G.
2) 最稀疏的表示自然地具有判别性.因为我们的最终目标是对目标领域数据实现分类,所以我们期望图数据包含尽可能多的判别信息,即,来自相同类的两个数据点通过边连接起来.对于典型的 k-NN 图,上述所期望的属性严重依赖近邻准则在原始空间实施效果的好坏[14],然而,对于原始高维数据(如人脸图像数据),最近邻准则通常不能取得较好的性能.相比之下,近年来研究[16]显示:稀疏表示具有自然的判别力并能在高维数据环境下取得较好的性能,而且该判别力仅与类数紧密相关,而与样本

数无关.因此,基于 SR 构建的图模型在无需源领域大量标签数据的情况下能够包含更多的判别信息.

基于以上原因,本文试图避开传统的基于图的SSL方法中所采用的点对间关系度量方法,而采用SR来重构各数据点 $\boldsymbol{x}_i \in X$.为此,我们首先在RKHS中通过求解如下修改的 l_1-范最小化问题来为各数据点 $\boldsymbol{x}_i$ 寻求一个稀疏重构权值向量 $\boldsymbol{s}_i$:

$$\min_{s_i} C \| \boldsymbol{s}_i \|_1 + \| \phi(\boldsymbol{x}_i) - X^{\phi} \boldsymbol{s}_i \|_2^2 = \min_{s_i} C \| \boldsymbol{s}_i \|_1 + L(\boldsymbol{s}_i, K) \tag{12}$$

其中,

- 核 K 为公式(8)中寻求的优化解;
- C 为正则化参数,以控制重构稀疏性和重构补偿间的平衡;
- $L(\boldsymbol{s}_i, K) = 1 + \boldsymbol{s}_i K_{XX} \boldsymbol{s}_i - 2\boldsymbol{s}_i^T K_X(\boldsymbol{x}_i)$, 其中,$K_{XX}$ 为一个$(n+m)\times(n+m)$矩阵,其中元素 $\{K_{XX}\}_{ij}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$;$K_X(\boldsymbol{x}_i)$是一个$(n+m)\times 1$ 向量,其中元素 $\{K_X(\boldsymbol{x}_i)\}_j=K(\boldsymbol{x}_j,\boldsymbol{x}_i)$;$\boldsymbol{s}_i=[s_{i1},s_{i2},\ldots,s_{i(i-1)},0,s_{i(i+1)},\ldots,s_{i(n+m)}]^T$ 是一个$(n+m)$-维列向量,其中,第 i 个元素等于 0 表示 $\boldsymbol{x}_i$ 从 X 中移除,$s_{ij}(j\neq i)$表示样本 $\boldsymbol{x}_j$ 对 $\boldsymbol{x}_i$ 的重构贡献度.本文进一步约束 $\sum_{j\neq i} s_{ij} = 1$ 且 $s_{ij} \geqslant 0$.

可以通过 KOMP 算法[32]来求解公式(12)中核稀疏表示问题.在求得所有数据点的重构稀疏向量 $\hat{\boldsymbol{s}}_i$ $(1 \leqslant i \leqslant n+m)$后,即可构建稀疏权值矩阵 $S = [\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2, \ldots, \hat{\boldsymbol{s}}_{n+m}]$, 进而可构建一个稀疏图模型 $G=\{X,S\}$,其中,X 为训练样本集,S 为边权值矩阵.值得说明的是,S 中元素 s_{ij} 并非数据点对 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间简单的相似性度量,矩阵 S 本质上有别于传统的图正则化算法(如 LPP(locality preserving projection)[33])中的权值矩阵.对于基于图的 DAL 问题,采用稀疏矩阵 S 作为图权值矩阵具有如下的有效属性:(1) 在稀疏矩阵 S 中,各权值向量 $\boldsymbol{s}_i$ 均遵从重要的对称性,即旋转不变性(满足公式(12)约束)和转换不变性(满足约束 $\mathbf{1}_{n+m}^T \boldsymbol{s}_i = 1$, 其中,$\mathbf{1}_{n+m}$ 代表$(n+m)\times 1$ 维列向量),使得权值矩阵 S 能够在一定程度上反映数据的本质几何属性;(2) 即使在无类标签的情况下,权值矩阵 S 中也能自然地保留数据的判别信息.

2.2.3 从标签数据到无标签数据的稀疏标签传播

本节,我们将利用公式(12)构建的核稀疏图和一个迭代过程来有效地解决源领域数据 $\boldsymbol{x}_i \in X^s$ 的标签向目标领域数据 $\boldsymbol{x}_u \in X^t$ 传播的问题.设 F 表示定义于样本集 X 上的分类函数集,且 $\forall f \in F$,则可赋予每个数据点 $\boldsymbol{x}_i$ 一个实值 f,无标签数据 $\boldsymbol{x}_u$ 的标签由 $f_u=f(\boldsymbol{x}_u)$ 的符号确定.在每次迭代中,使每个数据从其稀疏重构对象中"吸收"部分标签信息,且保留其初始状态的部分标签信息.这样,在第 t+1 次迭代时,$\boldsymbol{x}_i$ 的标签为

$$f_i^{t+1} = \alpha \sum_{j \neq i} M_{ij} f_j^t + (1-\alpha) y_i \tag{13}$$

其中,$0<\alpha<1$ 控制 $\boldsymbol{x}_i$ 从其重构对象"吸收"标签信息部分,$M_{ij}=(S+S^T-S^TS)_{ij}$. 令 $\boldsymbol{y}=(y_1,y_2,\ldots,y_{n+m})^T$ 且 $y_i=L_i(1 \leqslant i \leqslant n)$, $y_u=0(n+1 \leqslant u \leqslant n+m)$. $\boldsymbol{f}^t = (f_1^t, f_2^t, \ldots, f_{n+m}^t)^T$ 为在第 t 次迭代的预测标签向量,$\boldsymbol{f}^0=\boldsymbol{y}$.公式(13)迭代方程重写为

$$\boldsymbol{f}^{t+1}=\alpha M \boldsymbol{f}^t+(1-\alpha)\boldsymbol{y} \tag{14}$$

本文将采用公式(14)来更新各数据对象的标签直至收敛,即,数据的预测标签在经过几次迭代后不再发生变化.

定理 2. 公式(14)中计算的序列 $\{\boldsymbol{f}^t\}$ 收敛于下式:

$$\boldsymbol{f}^*=(1-\alpha)(I-\alpha M)^{-1}\boldsymbol{y} \tag{15}$$

证明:由公式(13)和初始条件 $\boldsymbol{f}^0=\boldsymbol{y}$ 可得:

$$\boldsymbol{f}^t = (\alpha M)^{t-1}\boldsymbol{y} + (1-\alpha)\sum_{i=0}^{t-1}(\alpha M)^i \boldsymbol{y} \tag{16}$$

显然,矩阵 M 的谱半径满足 $\rho(M) \leqslant 1$,同时,$0<\alpha<1$,从而可得 $\lim_{t\to\infty}(\alpha M)^{t-1} = 0$,进而,

$$\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha M)^i = (I-\alpha M)^{-1},$$

其中,I 为 n 阶指示矩阵(identity matrix).显然,序列 $\{\boldsymbol{f}^t\}$ 收敛于

$$\boldsymbol{f}^* = \lim_{t\to\infty} \boldsymbol{f}^t = (1-\alpha)(I-\alpha M)^{-1}\boldsymbol{y} \tag{17}$$

由于在分类中我们仅用 f_i 的符号去确定数据点 $\boldsymbol{x}_i$ 的标签,因此,衡量 $1-\alpha>0$ 不影响 $\boldsymbol{f}^*=(I-\alpha M)^{-1}\boldsymbol{y}$ 的符号变化,从而定理 2 得证. □

根据定理 2,以 $\boldsymbol{f}^*$ 作为分类函数使得 SLPDAL 成为"一站式"算法,即,只需一步就能预测所有数据标签.

下面,将 SLPDAL 算法拓展为多分类问题:设有 c 个分类,标签集为 $\bar{L}=\{1,2,...,c\}$,令 P 为 $(n+m)\times c$ 矩阵集,其中,矩阵元素为非负的实数值.任意矩阵 $F=[\boldsymbol{F}_1^T,\boldsymbol{F}_2^T,...,\boldsymbol{F}_{n+m}^T]\in P$ 对应 X 上的一个特定的分类,即,数据 $\boldsymbol{x}_i$ 分类为 $y_i=\arg\max_{j\leqslant c}F_{ij}$.因此,$F$ 也可看成一个标签函数.初始地,设 $F_0=Y$,其中:如果 $\boldsymbol{x}_i$ 标记为 j,则 $Y_{ij}=1$;否则 $Y_{ij}=0$,对于无标签数据点 $\boldsymbol{x}_u$,$Y_{uj}=0(1\leqslant j\leqslant c)$.同样地,对于多分类情况,只要将公式(17)中的 $\boldsymbol{y}$ 简单地替换为 Y,即可得到如下的多分类预测函数:

$$F^*=(1-\alpha)(I-\alpha S)^{-1}Y \tag{18}$$

则每个数据对象的标签可由 $y_i=\arg\max_{j\leqslant c}F_{ij}^*$ 确定.

2.3 算法步骤及其复杂度分析

SLPDAL 算法的主要步骤描述见算法 1.

算法 1. SLPDAL.

输入: 标签样本 $\boldsymbol{x}_i\in X^s$,无标签样本 $\boldsymbol{x}_u\in X^t$,参数 γ,α 以及初始化标签向量 $\boldsymbol{y}$.

输出: 目标领域无标签数据的预测标签.

Step 1: 通过公式(8)寻求优化核函数 K.

Step 2: 通过求解公式(12)中 l_1-范最小化问题,构建稀疏图 $G=\{X,S\}$.构建传播矩阵 $M=S+S^T-S^TS$.

Step 3: 重复 $F^{t+1}=\alpha MF^t+(1-\alpha)\boldsymbol{y}$ 直到收敛.

Step 4: 令 F^* 为序列 F^t 的极限,输出各数据点 $\boldsymbol{x}_i$ 的标签 $y_i=\arg\max_{j\leqslant c}F_{ij}^*$.

令 $N=n+m$,其中,n 和 m 分别代表源领域和目标领域数据集大小,k 表示算法迭代次数.SLPDAL 方法整体计算复杂度包括 3 个部分:优化核矩阵和 MMD 矩阵计算复杂度 $O(N^2)$;采用 KOMP 算法[32],样本的核稀疏表示计算复杂度 $O(N^2)$;基于稀疏图的标签传播需要复杂度近似为 $O(kN)$[34],则该算法的总体计算复杂度为 $O(2N^2+kN)$.因此,大样本数据集将会明显增大算法的复杂度,所幸的是,由于稀疏表示所具有自然判别能力,使得上述算法在解决实际 DAL 问题时只需要相对较少的源领域数据,即可取得可比较的效率和精度(如下文实验结果显示).另外,在实际应用中,为了进一步改善本文方法在大规模目标数据集上的处理效率,在 SLPDAL 算法第 1 步结束后,可利用传统的支持向量机[4]对目标领域数据进行初始划分,然后采用我们在文献[35]中的相似做法(文献[35]第 1.6 节),选取目标领域数据集的一个有效子集(设大小为 $t<<m$),再继续进行 SLPDAL 算法的第 2 步~第 4 步的学习.该做法使得 SLPDAL 算法所需处理的数据大小变为 $N'(=n+t)<<N$,从而有望提升算法执行效能.

3 SLPDAL 正则化框架

3.1 稀疏保留正则化

首先,基于上述稀疏重构图 $G=\{X,S\}$ 提出一个稀疏保留正则项.根据以上描述,稀疏权值矩阵 S 能够在一定程度上反映数据的本质几何特征且包含了数据的自然判别信息,另外,根据聚类假设,即,如果数据点对 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 相近,则其对应标签值 f_i 和 f_j 也应该彼此接近,我们因此可期望通过稀疏表示来保留类标签空间与核特征空间相似的几何特性.但是,对于新建立的稀疏图 G,其边权值 $\hat{s}_{i,j}$ 不是一个严格的相似性度量,我们不能按照基于图的框架来构建稀疏保留正则项.注意到,数据点对 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间关系刻画为 $\boldsymbol{x}_i=\sum_{j=1}^{n+m}\hat{s}_{i,j}\boldsymbol{x}_j$,其并非简单的亲近性,因此,我们期望该数据点对对应的类标签 f_i 和 f_j 也尽量保留这种关系,即,分别按照 LLE(locally linear embedding)[24]

和 LPP[34]算法思想,刻画为 $\min\sum_i\left\|f_i-\sum_{j=1}^{n+m}\hat{s}_{i,j}f_j\right\|_2^2$ 或 $\min\sum_{i,j=1}^{n+m}\hat{s}_{i,j}(f_i-f_j)$. 因此,可通过最小化如下目标函数来构建稀疏保留正则项,其最好地保留了优化权值向量 $\hat{\boldsymbol{s}}_i$,从而有效地实现稀疏标签从源领域到目标领域的平滑传播.

$$J_1(F)=\min_F\sum_{i,j=1}^{m+n}\hat{s}_{i,j}\|\boldsymbol{f}_i-\boldsymbol{f}_j\|^2 \tag{19}$$

$$J_2(F)=\min_F\sum_i\left\|\boldsymbol{f}_i-\sum_{j=1}^{m+n}\hat{s}_{i,j}\boldsymbol{f}_j\right\|_2^2 \tag{20}$$

其中,$F=(\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_c)\in\mathbb{R}^{(n+m)\times c}$ 为类指示矩阵,$\boldsymbol{f}_j=[f_j(\boldsymbol{x}_1),f_j(\boldsymbol{x}_2),\ldots,f_j(\boldsymbol{x}_{n+m})]^T$.$J_1(F)$和 $J_2(F)$之间的差别在于"加法"算子的顺序.虽然公式(19)和公式(20)中的正则项能够潜在地集成到许多 SSL 框架,但本文仅关注 DAL 框架的构建.

定理 3. $J_1(F)$和 $J_2(F)$分别为一阶和二阶图 Laplacian 正则化.

证明:利用简单的代数计算可得:

$$J_1(F)=tr(FL_1F^T),J_2(F)=tr(FL_2F^T),$$

其中,$L_1=D-S,L_2=(I-S)^T(I-S),S=[s_{ij}]_{(n+m)\times(n+m)},D=diag(d_1,d_2,\ldots,d_{n+m})$为对角矩阵, $d_i=\sum_j s_{ij}$.

另外,根据 $\sum_j s_{ij}=1,\forall i$, 则可得:

$$L_2\boldsymbol{e}=\sum_j[(I-S)^T(I-S)]_{ij}=\sum_j(I-S-S^T+S^TS)_{ij}=1-\sum_j s_{ij}-\sum_j s_{ji}+\sum_j\sum_k s_{ki}s_{kj}=1-1-\sum_j s_{ji}+\sum_k s_{ki}=0,$$

其中,$\boldsymbol{e}$ 为一$(n+m)$-维全 1 向量.

按照以上同样推导,可得 $L_1\boldsymbol{e}=0$.根据定义 1,$J_1(F)$和 $J_2(F)$可分别称为一阶和二阶图 Laplacian 正则化,且 L_1 和 L_2 分别为一阶和二阶图 Laplacian. □

根据定理 3,$J_1(F)$和 $J_2(F)$能够被进一步统一为

$$J(F)=tr(FLF^T) \tag{21}$$

其中,$L=L_1$ 或 $L=L_2$.

定理 4. SLPDAL 通过公式(15)计算的预测结果可通过如下正则化框架导出:

$$Q(F)=\min_F tr(FLF^T)+\frac{\mu}{2}tr((F-Y)^T(F-Y)) \tag{22}$$

其中,$Y\in\mathbb{R}^{(n+m)\times c}$ 为原始标签矩阵,如果 $\boldsymbol{x}_i$ 标记为 $y_i=j$,则 $Y_{i,j}=1$;否则,$Y_{i,j}=0$.

证明:针对变量 F 对 $Q(F)$求导数可得:

$$\frac{\partial Q(F)}{\partial F}=LF+\mu(F-Y) \tag{23}$$

令 $L=I-M$,对于一阶图 Laplacian $M=S$,对于二阶图 Laplacian $M=S+S^T-S^TS$.令公式(23)等于 0,可得最小化 $Q(F)$的近似解,即

$$(I-M)F+\mu(F-Y)=0.$$

进一步整理为

$$F-\frac{1}{1+\mu}MF-\frac{\mu}{1+\mu}Y=0.$$

引入两个新变量 $\alpha=\frac{1}{1+\mu}$ 和 $\beta=\frac{\mu}{1+\mu}$, 注意到$\alpha+\beta=1$,则$(I-\alpha M)F=\beta Y$,因为$(I-\alpha M)$可逆,故可得:

$$F=\beta(I-\alpha M)^{-1}Y \tag{24}$$

很容易看到,公式(24)和公式(15)相等.换句话说,SLPDAL 能够从正则化框架公式(22)导出,因此,目标领域数据点 $\boldsymbol{x}_u\in X(n+1\leqslant u\leqslant n+m)$的预测标签可由下式确定:

$$f_u=\arg\max_{1\leqslant j\leqslant c}F_{u,j},n+1\leqslant u\leqslant n+m \tag{25}$$

证毕. □

公式(22)中,$Q(F)$的第1项称为平滑项,其描述了相对于稀疏重构结构的数据标签的总体变化;第2项称为拟合项,其度量预测标签和原始标签的拟合性能.

3.2 全局和局部一致正则化

以全局的方式选择一个好的函数,可能不是一个好的策略,因为函数集可能没包含一个适于全局数据集的学习函数[27].然而从函数集中,可更容易地选出一些适于输入空间局部区域的好的预测函数,因此,可将整个输入空间分割成多个局部区域,然后针对各个局部区域实现更有效的最小化局部成本函数.但是,采用纯粹的局部学习算法也可能存在问题,因为在各局部区域可能不具备用以训练局部学习函数的数据[28],因此在局部学习正则化的基础上,还应该应用一个全局平滑项,以根据本质数据分布来平滑预测数据标签,使得预测标签更合理、更精确.

从第3.1节的讨论可知:公式(24)中,稀疏保留正则化虽然能比传统的基于图的SSL方法更好地捕捉数据的全局判别信息,但是它却不能捕捉局部数据的判别信息.因此,为了捕捉局部判别信息,本文采用局部核岭回归函数对每个数据模式实施一次回归运算.令矩阵 $X=[\boldsymbol{x}_1^s,\boldsymbol{x}_2^s,...,\boldsymbol{x}_n^s,\boldsymbol{x}_1^t,\boldsymbol{x}_2^t,...,\boldsymbol{x}_m^t]\in\mathbb{R}^{d\times(n+m)}$ 代表包含源领域和目标领域的数据集,为简单起见,假定 $X\mathbf{1}_{n+m}=0$,其中,$\mathbf{1}_{n+m}$ 表示$(n+m)\times1$ 向量,定义总体散度矩阵 S_t、类间散度矩阵 S_b 和类内散度矩阵 S_w 分别为 $S_t=XX^T,S_b=XGG^TX^T,S_w=XX^T-XGG^TX^T$,其中,$G=Y(Y^TY)^{-\frac{1}{2}}$ 为一个加权类指示矩阵,且 $G^TG=I_c$,则有如下定理:

定理5[36]. 如果 $rank(S_b)=c-1$ 和 $rank(S_t)=rank(S_w)+(S_b)$,则真实的类指示矩阵能被某个数据低维线性投影所表示,即,存在 $W\in\mathbb{R}^{d\times c}$,使得 $Y=X^TW$.

定理5中的条件在实际应用中,通常对于高维、小样本问题是满足的(如人脸识别应用)[36].

令 $X_i=\{\boldsymbol{x}_j\}_{j=1}^k\in\mathbb{R}^{d\times k}\subset X$ 代表数据 $\boldsymbol{x}_i$ 的核稀疏重构对象集,其中,k 代表数据点 $\boldsymbol{x}_i$ 的稀疏重构对象数(根据经验,为了获得最好的有判别的 k 个重构模式,本文设置 $s_{ij}\geqslant\varepsilon,\varepsilon\in(0,1)$为某个用户定义的相对较小的阈值).由定理5,可以定义如下局部正则项:

$$\min_{F\in\mathbb{R}^{(n+m)\times c}}\sum_{l=1}^{c}\|\boldsymbol{f}^l-\boldsymbol{o}^l\|^2 \tag{26}$$

其中,c 为待分的类数,$F=[\boldsymbol{f}^1,\boldsymbol{f}^2,...,\boldsymbol{f}^c]$为类指示矩阵且 $\boldsymbol{f}^l=[f_1^l,f_2^l,...,f_{n+m}^l]^T\in\mathbb{R}^{(n+m)},\boldsymbol{o}^l=\{o_i^l(\boldsymbol{x}_i)\}_{i=1}^{n+m}\in\mathbb{R}^{n+m}$ 代表在数据 $\boldsymbol{x}_i$ 的低维线性投影上的局部输出,即,$o_i^l(\boldsymbol{x}_i)=X_i^TW_i$. 为了获得公式(26)中 $o_i^l(\boldsymbol{x}_i)$ 的分析解,本文采用核岭回归算法[27],在某个具有核映射 ϕ 的RKHS中,公式(26)可重写为

$$\min_{f_i,W_i}\sum_{l=1}^{c}\sum_{i=1}^{n+m}(\|(X_i^\phi)^TW_i-f_i\|^2+\eta\|W_i\|^2) \tag{27}$$

其中,$X_i^\phi=\{\phi(\boldsymbol{x}_j)\}_{j=1}^k\subset X^\phi,X^\phi=[\phi(\boldsymbol{x}_1^s),\phi(\boldsymbol{x}_2^s),...,\phi(\boldsymbol{x}_n^s),\phi(\boldsymbol{x}_1^t),\phi(\boldsymbol{x}_2^t),...,\phi(\boldsymbol{x}_m^t)],\eta>0$ 为一个正则参数.

根据 Representer Theorem[15],可得 $W_i=\sum_{j=1}^{k}v_{ij}^l\phi(\boldsymbol{x}_j)$. 对于任意重构对象 $\boldsymbol{x}_j\in X_i$,其系数向量 $\boldsymbol{v}_i^l\in\mathbb{R}^{k\times1}$ 中元素为 v_{ij}^l.从而,$o_i^l(\boldsymbol{x}_i)=K_i\boldsymbol{v}_i^l$,$K_i\in\mathbb{R}^{k\times k}$ 为核矩阵,包含元素 $K(\boldsymbol{x}_u,\boldsymbol{x}_v),\boldsymbol{x}_u,\boldsymbol{x}_v\in X_i$.替换公式(27)中的 W_i 可得:

$$\min_{\boldsymbol{v}_i^l\in\mathcal{H}^k}\sum_{l=1}^{c}\sum_{i=1}^{n+m}(\|K_i\boldsymbol{v}_i^l-f_i^l\|^2+\eta(\boldsymbol{v}_i^l)^TK_i\boldsymbol{v}_i^l) \tag{28}$$

求解公式(28)后,可得 $\boldsymbol{v}_i^l=(K_i+\eta I)^{-1}f_i^l$,则 $o_i^l(\boldsymbol{x}_i)$ 的分析解可表示为

$$o_i^l(\boldsymbol{x}_i)=\boldsymbol{k}_i^T(K_i+\eta I)^{-1}f_i^l=\Omega_i^Tf_i^l \tag{29}$$

其中,$\boldsymbol{k}_i\in\mathbb{R}^k$ 表示包含元素 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$的向量,其中,$\boldsymbol{x}_j\in X_i$,$\Omega_i=\boldsymbol{k}_i^T(K_i+\eta I)^{-1}$.公式(26)可被重写为如下紧凑形式:

$$\min_{F\in\mathcal{H}^{(n+m)\times c}}\sum_{l=1}^{c}\|\boldsymbol{f}^l-A\boldsymbol{f}^l\|^2=\sum_{l=1}^{c}(\boldsymbol{f}^l)^TL_o\boldsymbol{f}^l=tr(F^TL_oF) \tag{30}$$

其中,$L_o=(I-A)^T(I-A)$,且 I 为一$(n+m)$-维单位矩阵,矩阵 $A=[a_{ij}]\in\mathbb{R}^{(n+m)\times(n+m)}$的构造方法是:对于$\forall\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j(1\leqslant i,j\leqslant$

$n+m$),如果 $\boldsymbol{x}_j \in X_i$,则 $a_{ij}=\Omega_{ij}$;否则,$a_{ij}=0$.在公式(21)中,用 $(1-\lambda)L+\lambda L_o$ 替换 L,其中,$\lambda \in [0,1]$为一平衡参数,从而得到一个基于全局和局部一致视角的混合图 Laplacian 正则化[26]形式:

$$J_{mix}(F)=tr\{F[(1-\lambda)L+\lambda L_o]F^T\} \tag{31}$$

最后,可导出基于混合图 Laplacian 正则化的 SLPDAL 框架形式:

$$Q_{mix}(F)=\min_F J_{mix}(F)+\frac{\mu}{2}tr((F-Y)^T(F-Y)) \tag{32}$$

根据公式(31),当 $\lambda=1$ 时,$J_{mix}(F)$降级为基于局部正则化的 DAL 形式.

4 算法性质和扩展

4.1 对样本外(out-of-sample)数据的推理

第 3 节论述了 SLPDAL 算法的主要演绎学习过程,本节介绍如何将 SLPDAL 算法推广到样本外数据学习.按照文献[14]的做法,为了将 SLPDAL 泛化到样本外数据学习需要做两件事情:(1) 对于新的样本外测试数据点 $\boldsymbol{x}_u$,使用和公式(23)相同的平滑准则类型;(2) 确保样本外数据 $\boldsymbol{x}_u$ 的加入不会影响训练数据集的原始 $Q(F)$值.

对于新的测试数据点 $\boldsymbol{x}_u$,平滑准则定义为

$$Q(f(\boldsymbol{x}_u))=\sum_{j:x_j\in X} s_{uj}(f(\boldsymbol{x}_u)-f_j)^2 \text{或} Q(f(\boldsymbol{x}_u))=\left(f(\boldsymbol{x}_u)-\sum_{j:x_j\in X} s_{uj}f_j\right)^2 \tag{33}$$

因为 $Q(f(\boldsymbol{x}_u))$关于 $f(\boldsymbol{x}_u)$为凸函数,故在公式(34)条件下,其能被最小化:

$$f(\boldsymbol{x}_u)=\sum_{j:x_j\in X} s_{uj}f_j \tag{34}$$

有趣的是,公式(34)正好是当数据点 $\boldsymbol{x}_u$ 的标签能够被训练数据集内重构数据对象的标签优化重构时的公式,即为公式(35)的优化解:

$$f(\boldsymbol{x}_u)=\min_{f(x_u)} \| f(\boldsymbol{x}_u)-\textstyle\sum_j s_{uj}f_j \|_2^2 \tag{35}$$

4.2 SLPDAL的鲁棒性

当数据集中有无效测试数据时,SLPDAL 算法存在潜在问题,因此在实施标签传播前,须先确定无标签样本是否有效.在实际应用场景中,检测并排斥无效测试样本(或离群点),是模式分类方法的关键能力.

对于每个类 $i(1\leqslant i\leqslant c)$,令 $\Psi_i:\mathbb{R}^{n+m}\to\mathbb{R}^{n+m}$ 为选取与第 i 类相关的稀疏重构系数的特征函数.对于一个稀疏重构系数向量 $\boldsymbol{s}\in\mathbb{R}^{n+m}$,$\Psi_i(\boldsymbol{s})\in\mathbb{R}^{n+m}$ 为一个新的向量,其非零元素为 $\boldsymbol{s}$ 中与第 i 个类相关的元素.

定义 3(稀疏集中索引(sparsity concentration index,简称 SCI)[16]). 稀疏重构系数向量 $\boldsymbol{s}\in\mathbb{R}^{n+m}$ 的 SCI 定义为

$$SCI(\boldsymbol{s})=\frac{c\cdot\max_i \|\Psi_i(\boldsymbol{s})\|_1/\|\boldsymbol{s}\|_1-1}{c-1} \tag{36}$$

根据定义 3,对于由公式(12)取得的优化解 $\hat{\boldsymbol{s}}_i$:如果 $SCI(\hat{\boldsymbol{s}}_i)=1$,则表示测试样本 $\boldsymbol{x}_i$ 仅由同一类的对象表示;如果 $SCI(\hat{\boldsymbol{s}}_i)=0$,则表示稀疏重构系数均分于所有类.

从而,可以选择一个适当的阈值 $\tau\in(0,1)$,使得当 $SCI(\hat{\boldsymbol{s}}_i)\geqslant\tau$ 时,则认为测试样本 $\boldsymbol{x}_i$ 为有效.因此,在稀疏图构造之前,我们可实施一个预处理过程,即,通过 SCI 测试方法来排除数据集中某些噪声或离群点,从而增强 SLPDAL 算法的鲁棒性能.

4.3 多核学习扩展

本文所提出的核方法在一定程度上严重依赖核函数的选择,然而,对于某个特定领域的适应学习任务,最合适的核函数事先往往是无法预知的.而且,在某个用户事先定义的核函数池中进行优化函数的穷尽搜索也将是非常耗时的,因此,为了增强该方法的鲁棒性,关键在于如何有效学习一个适当的核函数.目前,已有一些多核学

习(multiple kernel learning,简称 MKL)方法[37,38]被提出来,然而,这些方法均假设训练数据和测试数据来自相同领域,导致其不能基于来自不同领域的样本数据学习一个优化的核函数,从而使得这些 MKL 方法在源领域的学习性能不能有效地迁移到目标领域.但是,在某些约束条件(如领域间分布距离最小化)下[4],上述 MKL 方法能够明显改善 DAL 性能.由于核函数在本文方法中的中心地位,选取一个好的核函数是必须的,因此,本节将重点讨论在 SLPDAL 算法的领域分布核均值匹配阶段,如何利用多核学习技术来学习一个有效的集成核函数,从而将上述 SLPDAL 方法推广到多核学习框架.图 1 显示了 SLPDAL 的多核学习模型.

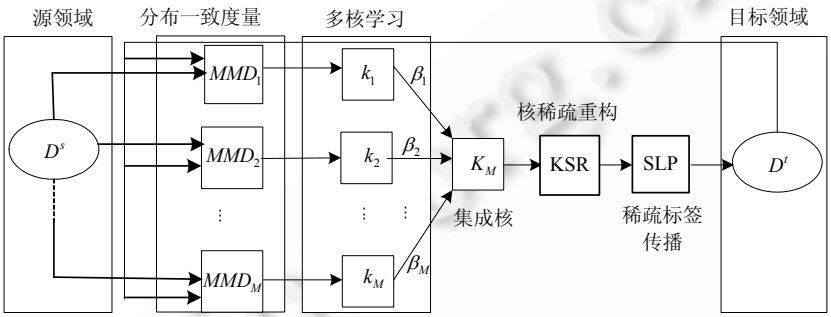


Fig. 1 Multiple kernel learning schema for SLPDAL

图 1 SLPDAL 的多核学习模型

根据 MKL 技术,即,将 M 个核 $k_1,k_2,\ldots,k_M$ 及其核诱导特征映射 $\phi_1,\phi_2,\ldots,\phi_M$ 构成一个凸组合[37,38],基于领域数据的多核 K_M 是 M 个核 $\{k_h\}_{h=1}^M$ 的凸组合,其中,k_h 与公式(8)中定义的 K 相同.

采用文献[37,38]中的 MKL 形式,可得:

$$K_M=\sum_{h=1}^{M}\mu_h k_h,\ \mu_h\geqslant 0,\ \sum_{h=1}^{M}\mu_h=1.$$

从而,公式(8)中的单核函数被推广为多核函数 K_M.

公式(8)可直接利用现有的 MKL 软件包(如 SimpleMKL[37])求解,更多细节可参见文献[37].为有所区别,下文将基于 MKL 的 SLPDAL 算法称为多核稀疏标签传播(multiple kernel sparse label propagation,简称 MKSLP).

4.4 领域间分布一致控制

定理 6[39]. 给定一个高斯核函数类 $K_g=\{e^{-\|x-z\|_2^2/2\sigma^2},\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^d:\sigma\in[\sigma_0,\infty)\}$,其中,$\sigma_0>0$.对于任意 $k_\sigma,k_\tau\in K_g$ 且 $0<\tau<\sigma<\infty$,则有 $\gamma_{k_\sigma}(P,Q)\geqslant\gamma_{k_\tau}(P,Q)$.

定理 6 说明:核带宽越大,RKHS 嵌入领域的分布距离将越大,从而导致 SLPDAL(或 MKSLP)算法收敛速度降低.为了实验研究核带宽对 SLPDAL(或 MKSLP)的性能影响,本文特将高斯核带宽参数化,即,高斯核函数泛化为

$$k_{\sigma/\gamma}(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2(\sigma/\gamma)^2}\right),$$

其中,γ为一个可调参数.根据下文实验结果:随着γ值的增加,领域内样本分布呈现强的内聚性,从而导致领域内不同类样本出现交叠,这将会严重影响学习性能;另一方面,随着γ值的减小,将会导致 SLPDAL(或 MKSLP)算法收敛率下降.因此,本文约束参数$\gamma\in[1,\gamma_0]$,其中,γ_0 为一用户指定的阈值.实验结果显示:通过协调参数γ,能够进一步增强所提出算法的领域适应性能.

5 实验结论

为了评价 SLPDAL 方法在 DAL 问题上的有效性,本文在一系列人造数据集和几个实际领域适应数据集上

将该方法与几个代表性的算法进行比较,其中,实际领域适应应用包括:跨领域人脸识别、跨领域图像标注、可视化视频概念检测、跨领域文本分类.

对于所有数据集,源领域数据和部分目标领域数据真实标签已知,来自源领域和目标领域的标签数据用于训练数据集,目标领域无标签数据用于测试数据集.

与所提出的方法进行比较的学习算法包括基线方法 LLGC[26]、基于稀疏重构技巧的 S-RLSC[18]以及基于线性邻居传播模型的 LNP[14].这些方法都是经典的 SSL 算法,它们在许多不同的 SSL 任务中表现出良好的鲁棒性,但是,它们不能直接有效应用于 DAL 任务.为此,本文还重点比较了几种代表性的 DAL 方法,如 LMPROJ[6], CD-SVM[9],KMM[40]和 DTSVM[4].

在所有实验中,对于几个相关的核学习方法 CD-SVM,KMM 和 LMPROJ,本文采用标准的高斯核函数:

$$k_\theta(\boldsymbol{x},\boldsymbol{z})=\exp(-\theta\|\boldsymbol{x}-\boldsymbol{z}\|^2),$$

其中,θ设置为 $1/d$(d 为数据维数).对于多核学习方法 DTSVM,按照文献[4]的设置,令核参数为 $1.2^\delta\theta$,其中,δ分别设置为{0,0.5,1,1.5,2,2.5,3,3.5},从而为 DTSVM 构造 8 个基核.对于 SLPDAL 方法,采用参数化高斯核函数:

$$k_{\sigma/\gamma}(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2(\sigma/\gamma)^2}\right),$$

其中,核参数σ按照文献[29]的做法,通过最小化 MMD 准则获得.根据经验,对于二元分类问题,本文首先选取核参数σ为训练数据平均范数的平方根;对于多类划分,则选取核参数为 $\sigma\sqrt{c}$ (c 为分类数).

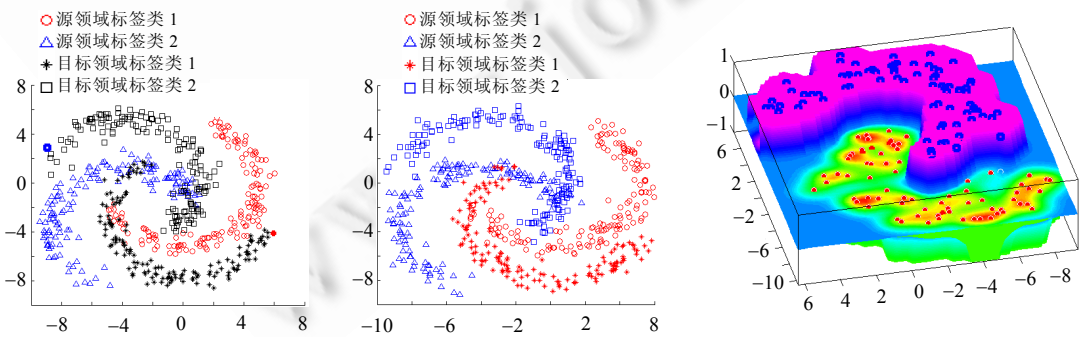
对于多核方法 MKSLP,为了公平起见,本文也按照文献[4]的设置,选取核参数为 $1.2^\delta\sigma$,其中,δ的设置与 DTSVM 相同.在该方法利用 SCI 测试的预处理阶段,本文按照经验选取阈值τ,以过滤 2%的潜在噪声点.

目前,在核学习方法中如何有效地选择模型参数,仍然属于一个具有挑战性的公开问题.本文采用五重交叉验证法来选择各算法参数,实验结果的平均值用于算法性能评价,所有算法代码实现均在 Matlab2010a 上完成.

5.1 人造数据集实验

5.1.1 标签传播效果

本节将采用一个二维人造数据集来显示所提出方法的学习性能,以更好地理解该方法在特定的领域适应问题上的标签传播过程.首先,人工生成一个包含 300 个样本的刻画两个信息类别的“双月形”二维数据集作为源领域(source domain,简称 SD)数据,其中,每个类包含 150 个样本对象;然后,将该数据集逆时针旋转 30°作为目标领域(target domain,简称 TD)数据,并在其中标识 2 个标签数据点(分别以实心*和□表示),如图 2(a)所示(即,当迭代次数 t=0 时仅有 2 个标签样本).上述旋转操作,使得两个领域数据呈现不同分布.从图 2(b)(即,当迭代次数 t=15 时,目标领域所有样本被标识)可以清楚地看出本文所提出的方法在 DAL 上的标签传播有效性.



(a) 初始状态 (b) 收敛状态 (c) 目标领域归纳学习结果

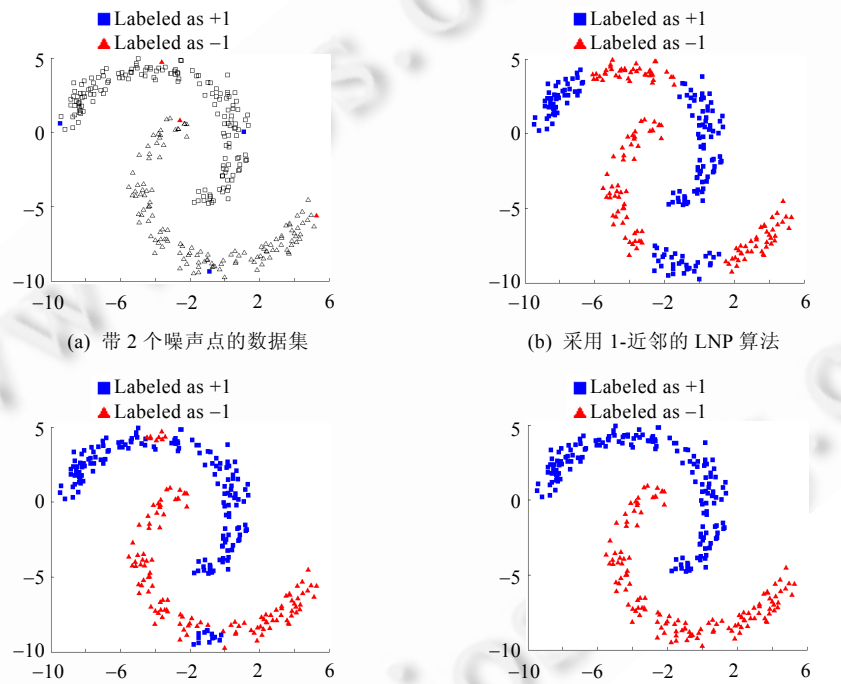
Fig.2 Label propagation results of SLPDAL

图 2 SLPDAL 标签传播结果

为了进一步显示所提出的方法对样本外数据的学习效能,继续采用图 2(a)中的数据集,利用 SLPDAL 预测目标领域无标签样本,通过公式(35)来推理目标领域在区域$\{(x,y)|x\in[-10,7],y\in[-10,5]\}$内样本的标签,图 2(c)显示了所有样本标签的推理结果,图中 z 轴显示数据的预测标签值.从图 2(c)结果可直观地看出,推理结果的边界和样本内数据的预测边界的本质结构几乎相同.

5.1.2 对噪声数据的鲁棒性

在某些情况下,由于数据标注的疏忽,使得标签样本中可能包含某些噪声,其将直接导致标签传播算法有效性的明显下降.而检测这些噪声标签需要大量人力和时间,因此在实际应用中,有必要设计一个鲁棒的分类器.图 3 展现了本文所提方法对噪声标签的鲁棒性能,在图 3(a)中,一个包含两类的“双月形”数据集直观上应被划分为 2 个聚类,其中包含 2 个划分异常的数据点,或可称为噪声标签(即,目标领域每个类包含 1 个噪声点).图 3(b)~图 3(d)分别显示 1-近邻 LNP 算法、无 SCI 预处理的 SLPDAL 和具有 SCI 预处理的 SLPDAL 算法,经比较可以看出,只有经过 SCI 预处理的 SLPDAL 算法正确地划分了所有数据点.



(a) 带 2 个噪声点的数据集 (b) 采用 1-近邻的 LNP 算法

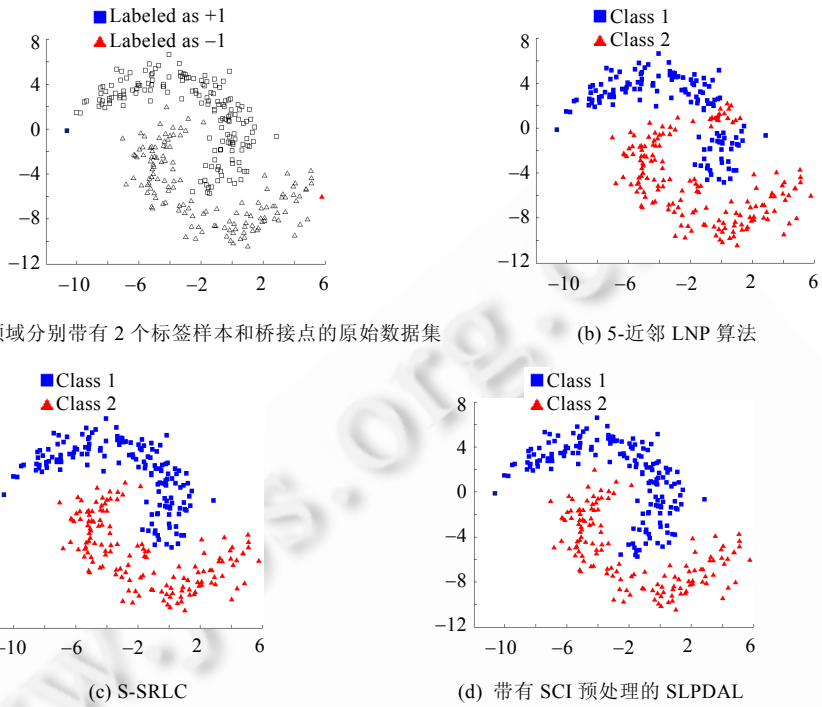
(c) 无 SCI 过滤预处理的 SLPDAL 算法 (d) 带有 SCI 过滤预处理的 SLPDAL 算法

Fig.3 Label propagation of SLPDAL on noisy dataset

图 3 SLPDAL 在噪声数据集上的标签传播

5.1.3 对桥接数据的鲁棒性

在 SLPDAL 算法中存在的另一个潜在问题是,数据集中存在的桥接数据(即,在某些复杂的类相互交连的数据集中那些连接不同类数据的样本点)也将导致标签传播性能下降.如图 4(a)所示的“双月形”数据集,该数据集的上半月和下半月距离较近,导致上半月右端点和下半月的左端点出现一定程度的交织,从而产生了某些桥节点.图 4(b)~图 4(d)分别显示 5-近邻 LNP 算法、S-SRLC 算法以及基于 SCI 预处理的 SLPDAL 算法分类结果.从这些结果可以看出:针对具有桥接点的数据集,SLPDAL 方法经过 SCI 预处理后,取得了优于 LNP 和 S-SLRC 方法的学习性能.由于 LNP 算法无法有效判别类间交叠数据点,使得其明显对桥节点较敏感[36].另外,值得注意的是,与 LNP 算法相比,S-SRLC 算法取得了较好的鲁棒性.可能的解释是,SR 具有自然的判别能力,因而即使在桥架点存在的情况下,基于 SR 的 S-SRLC 算法依然运行良好.



(a) 目标领域分别带有 2 个标签样本和桥接点的原始数据集 (b) 5-近邻 LNP 算法

(c) S-SRLC (d) 带有 SCI 预处理的 SLPDAL

Fig.4 Label propagation of SLPDAL on a more complex toy dataset with bridged data

图 4 SLPDAL 在具有桥接点的更复杂的人造数据集上的标签传播

5.2 领域适应人脸识别

5.2.1 数据集描述

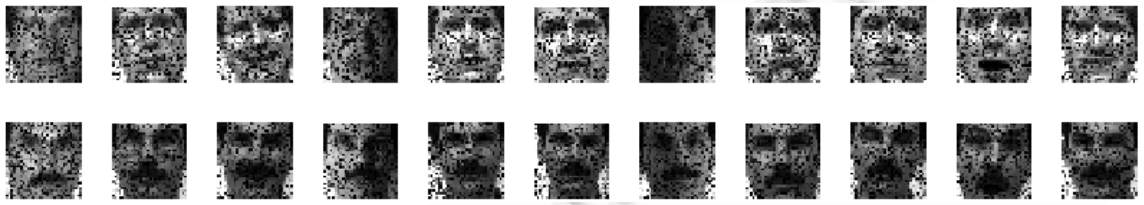
为了评价所提出的方法在跨领域人脸识别应用上的有效性,本文采用两个具有代表性的人脸数据库,即,ORL 和 YALE(两个数据集均可从网站 http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html 下载).

ORL 数据库包含 40 个不同对象的 400 幅人脸图像,根据时间、灯照条件、面部表情、面部细节等不同特征,分别有 10 幅图像刻画每个对象,实验中,将每幅原始图像的大小缩减为 40×40 像素;YALE 数据库包含 15 个不同对象的 165 幅人脸图像,根据不同的面部表情和面部配置,分别有 11 幅图像刻画每个对象,并将原始图像缩减为 32×32 像素大小.

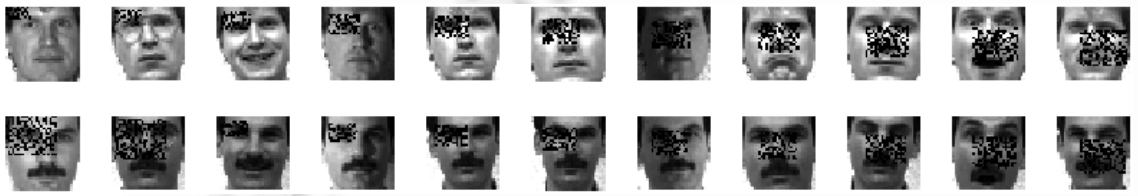
5.2.2 实验设置

为了评价所提出的方法在人脸识别应用上的鲁棒性,本文从 YALE 和 ORL 数据库中分别对每个对象随机选取 8 幅图像作为源领域数据集(如图 5(a)所示).目标领域数据集通过将源领域数据逆时针旋转 30°来获得,该旋转操作使得源领域和目标领域数据具有不同的分布.实验中,为了测试所提出方法的鲁棒性,在样本数据中加入了随机噪声和遮罩信息.第 1 次,在目标领域图像样本中逐渐增加高斯白噪声百分比,以逐渐减小信噪比(signal-to- noise ratio,简称 SNR),如图 5(b)所示.第 2 次,不同大小的黑色方块随机覆盖在目标领域样本图像上,从而产生遮罩(或缺失数据)现象,如图 5(c)所示.对于大小为 32×32 像素的 YALE 图像,黑色方块大小分别为 6×6,10×10,18× 18 和 22×22 像素;对于大小为 40×40 像素的 ORL 图像,黑色方块大小分别为 8×8,14×14,22×22 和 26×26 像素.

(a) YALE 数据库中的两个人脸对象



(b) 带白噪声的人脸对象



(c) 带遮罩的人脸对象

Fig.5 Face images in YALE with noise and occlusions

图 5 YALE 中带有噪声和遮罩的人脸图像

5.2.3 实验结果

上述数据集的实验结果记录于表 1 和表 2,表中分别详细记录了每种算法在各数据集上 10 次最好分类精度率(ACC%)的平均值和标准差,其中,黑体数据表示经 *t*-测试证实相应算法明显优于其他算法.由表 1 和表 2 可看出:在样本数据完好(无缺损数据和噪声信息)或接近完好的情况下,SLPDAL(或 MKSLP)和 DTSVM 算法均取得了优于其他算法的分类性能,这说明 SR 和 MKL 技术可用于改善 DAL 泛化性能.同时,几种 SSL 算法取得了最差的分类性能,这也证实了 SSL 算法不适于 DAL 任务.但值得指出的是,本文方法 SLPDAL 和 MKSLP 在绝大多数情况下均取得了最好或可比较的识别性能,这说明基于稀疏重构和 MMD 思想的 SLPDAL 和 MKSLP 算法是有效的.另外,随着噪声或遮罩信息的增加,所有算法的识别性能均表现出了不同程度的下降趋势,而本文所提出的方法 SLPDAL(或 MKSLP)和 S-SRLC 性能下降相对缓慢.同时看到,基于多核技术的 MKSLP 方法几乎总是稍好于 SLPDAL 方法,这也进一步说明了基于稀疏重构技术或(和)MKL 技术的图像识别方法对于噪声或缺失数据具有更强的鲁棒性.

Table 1 Means (%) and standard deviations of classification accuracies (ACC) of all algorithms with different levels of white Gaussian noise in the target domain datasets

表 1 在具有不同白高斯噪声等级的目标领域数据上所有算法分类精度的平均值(%)和标准差

人脸数据库	白噪声等级	LLGC	S-RLSC	LNP	KMM	LMPROJ	DTSVM	SLPDAL	MKSLP
Yale	无噪声	58.72 (±1.40)	60.63 (±0.48)	60.14 (±0.30)	65.78 (±1.44)	68.82 (±1.65)	71.62 (±2.24)	71.49 (±0.45)	**72.45** (±0.20)
	20db	56.32 (±0.44)	60.12 (±0.02)	59.47 (±0.41)	63.52 (±0.37)	66.90 (±1.24)	71.06 (±1.13)	71.28 (±0.6)	**72.36** (±0.12)
	15db	56.02 (±0.54)	60.12 (±1.22)	59.36 (±0.20)	62.84 (±0.06)	66.28 (±0.34)	70.87 (±0.40)	71.06 (±0.52)	**72.21** (±0.02)
	10db	54.21 (±1.14)	59.56 (±0.01)	57.89 (±0.62)	61.32 (±1.24)	64.72 (±2.14)	68.26 (±1.26)	70.68 (±0.60)	**71.82** (±0.34)
	5db	52.29 (±2.12)	59.28 (±0.41)	55.62 (±0.06)	59.60 (±1.22)	61.65 (±1.08)	65.01 (±0.50)	70.37 (±0.04)	**71.40** (±0.00)
ORL	无噪声	76.30 (±1.00)	82.91 (±0.46)	83.82 (±0.4)	84.55 (±1.35)	84.84 (±0.00)	86.28 (±0.04)	86.56 (±0.44)	**87.64** (±0.36)
	20db	75.72 (±0.20)	82.91 (±0.02)	83.11 (±0.61)	83.10 (±0.6)	83.79 (±0.13)	85.60 (±0.01)	86.42 (±0.02)	**87.46** (±0.04)
	15db	73.46 (±0.42)	82.52 (±0.44)	82.01 (±0.53)	81.69 (±0.40)	81.46 (±0.18)	84.46 (±0.35)	86.10 (±0.08)	**87.25** (±0.10)
	10db	69.82 (±0.45)	82.02 (±0.44)	80.96 (±0.03)	81.65 (±0.24)	82.54 (±1.04)	82.29 (±1.03)	86.36 (±0.48)	**86.92** (±0.14)
	5db	69.27 (±0.6)	81.65 (±1.05)	79.88 (±1.42)	78.98 (±1.42)	80.42 (±1.11)	81.66 (±0.04)	86.22 (±0.18)	**86.56** (±0.32)

Table 2 Means (%) and standard deviations of classification accuracies (ACC) of all algorithms with different size of occlusion in the target domain datasets

表 2 在具有不同遮罩大小的目标领域数据上所有算法分类精度的平均值(%)和标准差

人脸数据	遮罩大小	LLGC	S-RLSC	LNP	KMM	LMPROJ	DTSVM	SLPDAL	MKSLP
Yale	无遮罩	58.72 (±1.40)	60.63 (±0.48)	60.14 (±0.30)	65.78 (±1.44)	68.82 (±1.65)	71.62 (±2.24)	71.49 (±0.45)	**72.45** (±0.20)
	6×6	57.46 (±0.4)	60.52 (±0.14)	59.56 (±0.4)	64.45 (±0.03)	67.69 (±0.22)	71.46 (±0.22)	71.20 (±0.04)	**72.14** (±0.30)
	10×10	56.00 (±0.42)	69.29 (±0.05)	59.25 (±0.00)	62.73 (±0.16)	65.88 (±0.30)	70.60 (±0.40)	71.00 (±0.02)	**71.96** (±0.25)
	18×18	54.01 (±0.72)	59.38 (±0.00)	57.46 (±0.08)	60.62 (±0.52)	64.12 (±0.16)	68.18 (±0.56)	70.85 (±0.24)	**71.76** (±0.42)
	22×22	52.26 (±1.32)	59.08 (±0.56)	55.25 (±0.10)	60.04 (±1.02)	61.21 (±2.24)	64.67 (±0.70)	70.42 (±0.36)	**71.40** (±0.06)
ORL	无遮罩	76.30 (±1.00)	82.91 (±0.46)	83.82 (±0.4)	84.55 (±1.35)	84.84 (±0.00)	86.28 (±0.04)	86.56 (±0.44)	**87.64** (±0.36)
	6×6	75.94 (±0.22)	82.78 (±0.33)	83.40 (±0.01)	83.20 (±0.6)	84.27 (±0.12)	85.82 (±0.4)	86.38 (±0.52)	**87.32** (±0.14)
	10×10	73.66 (±0.46)	82.54 (±0.72)	82.56 (±0.64)	82.48 (±0.08)	82.66 (±0.80)	84.59 (±0.44)	86.41 (±0.51)	**87.15** (±0.01)
	18×18	70.24 (±0.22)	82.11 (±0.48)	80.66 (±0.53)	81.34 (±0.66)	81.62 (±0.64)	83.17 (±0.82)	86.06 (±0.40)	**86.78** (±0.10)
	22×22	68.80 (±0.67)	81.92 (±0.08)	79.58 (±0.52)	78.16 (±0.22)	79.68 (±1.04)	81.23 (±0.24)	85.94 (±0.58)	**86.41** (±0.34)

5.3 跨领域Web图像标注

5.3.1 数据集及其设置

本部分将在一个实际 Web 图像标注数据库 NUS-WIDE[41]上进行实验,以验证所提出的方法在图像标注问题上的领域适应学习性能.NUS-WIDE 数据库包含来自 81 个不同概念的 269 648 幅带标签的 Web 图像,样本特征为 500 维.为了模拟 DAL 环境,实验选取 12 个动物概念,包括熊猫、猴子、猫、斑马、老虎、鸟、狗、蛙、马、蝴蝶、蛇、长颈鹿,如图 6 所示,从中随机选取 6 个带标签的概念作为源领域数据集,采用所提出的算法将稀疏标签传播到其他 6 个概念.

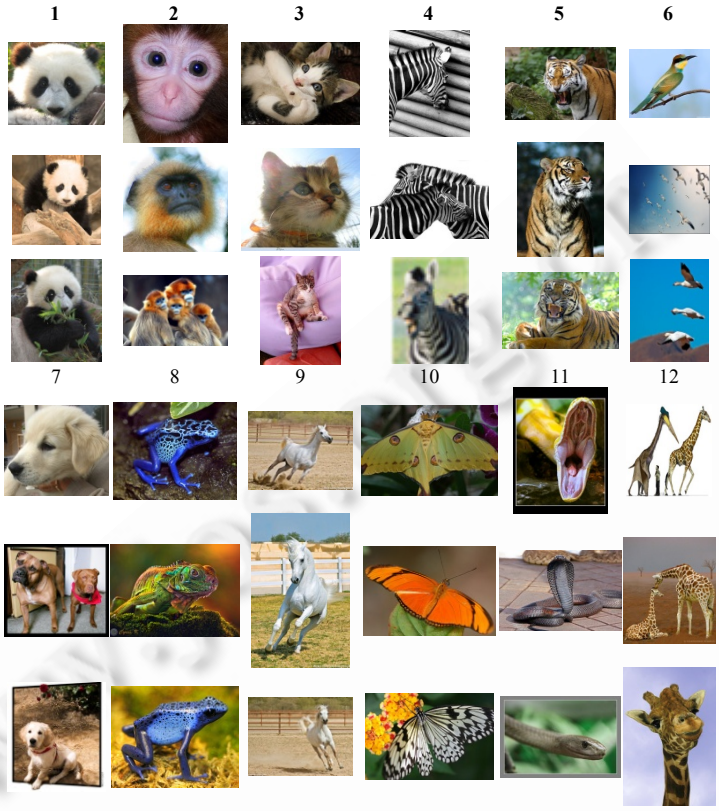


Fig.6 Sample images of 12 animal categories of NUS-WIDE dataset

图 6 NUS-WIDE 数据集中 12 个动物类别样本图像

5.3.2 实验结果

实验中对 12 个概念进行了 6 次随机划分,每次划分的图像概念标注性能如图 7 所示.

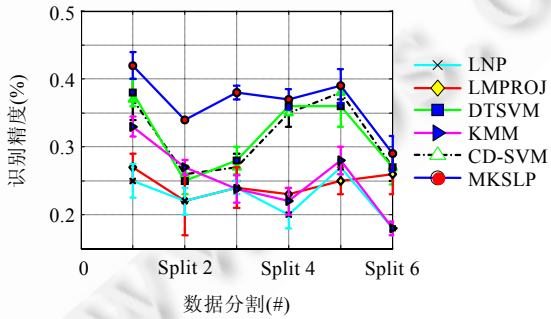


Fig.7 Recognition rate of different adaptation settings in NUS-WIDE dataset

图 7 NUS-WIDE 数据集中不同适应设置的识别率

从图 7 可以看出:本实验取得了与人脸识别相似的结论,唯一例外的是 CD-SVM 方法在本数据集上取得了相对较好的性能,但由于其未能明确考虑有效控制领域间分布的一致性,使得 CD-SVM 方法依然达不到满意的效果.值得指出的是:本文基于 MKL 的方法 MKSLP 在所有划分上的图像概念标注性能均优于其他方法,甚至比人脸识别中优势更明显.对此可能的解释是:由于不同概念图像的差异性较大,导致领域间分布距离较大,在此条件下,MKSL 与其他方法相比,能够迁移更多的判别信息至目标领域.而且,本实验结果也再次显示出:MKSLP

算法通过领域间分布一致正则化来学习一个多核空间,对于跨领域图像标注任务是有效的.

5.4 视频概念检测

本节将进一步验证 MKSLP 方法在大规模视频数据集 TRECVID 上的有效性和效率.

5.4.1 数据集描述

TRECVID 视频数据库(http://www-nlpir.nist.gov/projects/trecvid)是目前供研究所用的最大的带标注视频数据集之一,其中,TRECVID 2005 数据集包含 61 901 个关键帧,分别抽取自 6 个广播频道 108 小时的视频节目;TRECVID 2007 数据集包含 21 532 个关键帧,分别抽取自 60 小时的新闻杂志、科学新闻、纪录片以及教育节目等视频数据[4,9].如文献[4]所展示,由于 TRECVID 2007 数据集和 TRECVID 2005 数据集在节目结构和制作规格等方面存在较大的差异,使得在 TRECVID 数据集上进行领域适应学习是一项艰难挑战.实验从 LSCOM-lite 词库[4]中挑选 36 个语言概念,以覆盖在广播新闻视频出现的 36 个主要的可视概念,这 36 个概念已被予以手工标注,以描述在 TRECVID 2005 和 TRECVID 2007 数据集中关键帧的可视内容.抽取 3 个低级全局特征(即,网格颜色矩(225-维)、Gabor 纹理(48-维)和边缘方向直方图(73-维))以表示关键帧的不同内容,然后,将这 3 类特征连接起来,使得每个关键帧均成为一个 346-维的特征向量.

5.4.2 实验设置

本实验将系统地比较所提出的方法 MKSLP 和基线方法 S-SRLC 以及几种跨领域学习方法(包括 CD-SVM, DTSVM 以及 KMM)在视频概念识别应用上的性能.MKSLP 和 S-SRLC 使用源领域 D^s 或源领域加上目标领域 $D^s \cup D^t$ 带标签数据作为训练样本集,这里, D_l^t 代表目标领域带标签数据,为了有所区别,将在这两种情况下训练的 MKSLP 和 S-SRLC 方法分别称为 MKSLP_A,MKSLP_AT,S-SRLC_A 和 S-SRLC_AT.几种跨领域学习方法 CD-SVM,DASVM,KMM 和 DTSVM 也同时采用两个领域标签数据集 $D^s \cup D_l^t$ 作为训练数据集.对于 CD-SVM, KMM,DTSVM 和 MKSLP 方法,从目标领域随机选出 4 000 个无标签样本作为无标签目标领域数据集 D_u^t. 按照文献[4]的设置,根据在 TRECVID 2007 数据集中正标注样本的频率,实验中将 36 个概念分成 3 个组:第 1 组由 12 个具有高频率(正标注样本频率超过 0.05)的概念组成;第 2 组由 11 个具有中等频率(0.01≤正标注样本频率≤0.05);第 3 组由剩下的 13 个低频率概念组成(正标注样本频率小于 0.01).实验采用非差值平均精度(non-interpolated average precision,简称 AP)[4]作为性能评价标准,并记录每种算法最好的实验结果.

5.4.3 实验结论

表 3 记录了所有参与比较的算法分别在 3 个概念组和总体概念上的平均 AP 率.

Table 3 Performance comparison (AP rate) of different related algorithms in three concept groups of the video annotation dataset TRECVID

表 3 在视频标注数据集 TRECVID 中 3 个概念组上的性能(AP 率(%))比较

	S-SRLC_A	S-SRLC_AT	CD-SVM	KMM	DTSVM	MKSLP_A	MKSLP_AT
Group-1	37.02% (±0.25)	39.79% (±0.14)	40.06% (±0.30)	40.22% (±0.35)	44.98% (±0.46)	45.33% (±0.28)	**45.99%** (±0.36)
Group-2	12.32% (±0. 28)	12.91% (±0.32)	12.60% (±0.25)	12.83% (±0.18)	15.29% (±0.05)	15.36% (±0.00)	**15.94%** (±0.16)
Group-3	13.20% (±0.45)	14.95% (±0.36)	13.18% (±0.20)	12.93% (±0.48)	16.91% (±0.24)	17.05% (±0.22)	**17.51%** (±0.15)
Group-All	20.85% (±0.32)	22.55% (±0.27)	21.95% (±0.25)	21.99% (±0.34)	25.73% (±0.25)	25.91% (±0.17)	**26.48%** (±0.22)

由表 4 可以得出如下几个有意义的结论:

(1) 在所有概念组上,S_SRLC_A 方法均逊色于其他方法.这表明仅利用源领域训练数据的 S_SRLC 分类器不能在目标领域取得较好性能,这也说明了在不同年代收集的 TRECVID 数据集的数据分布间差异较大.

(2) S-SRLC_A 和 S-SRLC_AT 方法在 Group-1 中的 AP 值一定程度上高于在 Group-3 中的 AP 值.这说

明在Group-1中,概念具有大量来自两个领域的正样本数据.直观上来说,当两个领域中存在足够的正样本时,样本将在特征空间中分布紧密.在此情况下,领域样本分布可能出现交叠[4],从而使得来自源领域的训练数据有助于目标领域视频概念检测;相反地,在Group-3中,来自两个领域的中样本在特征空间的分布较稀疏,导致两个领域数据分布间交叠较少,因此对于Group-3中的概念识别,来自源领域的训练数据将在一定程度上降低目标领域视频概念检测性能.

(3) 在所有概念组上,方法S-SRLC_AT,CD-SVM和KMM均优于方法S-SRLC_A.这表明在S-SRLC_AT,CD-SVM和KMM等方法中,利用目标领域信息能够有效改善目标领域学习性能.从Group-ALL中的结果可知,KMM和CD-SVM方法比S-SRLC_AT方法稍差.可能的解释是:CD-SVM方法采用来自目标领域k-NN来定义源领域数据权重,而当目标领域正标注训练样本数非常有限时(如本实验中,每个概念10个正标注样本),上述学习得到权值是不可靠的;相似地,KMM方法采用无监督的方式学习源领域数据集权重,可能使得其识别性能在一定程度上弱于其他跨领域识别方法.

(4) 在所有概念组上,DTSVM和MKSLP方法明显好于S-SRLC_AT和两种DAL方法CD-SVM和KMM.这说明,DTSVM和MKSLP方法通过有效组合多基核函数,能够成功地减小两个领域间的分布差距.而且,本文带有稀疏保留特性的MKSLP_A和MKSLP_AT方法在Group-ALL概念组上的识别性能优于DTSVM方法,由此可知,MKSLP方法在所有概念组上的识别性能优于所有其他方法.

(5) 值得注意的是,在所有概念组上,MKSLP_AT方法均稍好于MKSLP_A方法,这进一步证实了目标领域先验信息能够有效增强MKSLP方法的领域适应能力.

在定理2中,我们已系统地证明了SLPDAL方法的收敛性.下面,我们将采用分别来自上述3个不同概念组中的3个概念,即Person,Office和Charts的数据作为实验样本,以实验证实MKSLP算法的收敛性能.如图8所示,MKSLP的目标值(迭代变化值)在少于10次内即可收敛,相似结果在TRECVID数据集中的其他概念上也能观测得出.

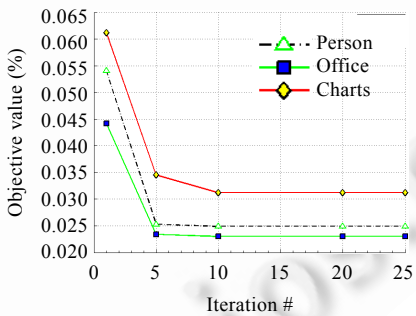


Fig.8 Convergence of MKSLP

图8 MKSLP收敛性能

5.5 跨领域文本分类

本节将采用两个实际文本数据集20 Newsgroups和垃圾邮件过滤[3,6]来深入展现所提方法在跨领域文本分类任务上的普遍有效性.

5.5.1 数据集描述

上述两个文本数据集的简单描述汇总于表4,表中数据集序号标识了各个实验任务,如序号3表示20 Newsgroup文本集中“Rec.”(源领域)和“Sci.”(目标领域)数据作为实验样本集,其中,源领域(Rec.)中分别包含1 984个正的训练样本和1 977个负训练样本;而目标领域(Sci.)分别包含1 993个正的训练样本和1 972个负训练样本.

Table 4 Datasets in cross-domain text classification tasks

表 4 跨领域文本分类任务中的数据集

任务		数据集	源领域训练样本		目标领域测试样本	
			正类	负类	正类	负类
1	20 Newsgroups	Comp vs. Sci	1 958	1 972	2 923	1 977
2		Rec vs. Talk	1 993	1 568	1 984	1 658
3		Rec vs. Sci	1 984	1 977	1 993	1 972
4		Sci vs. Talk	1 971	1 403	1 978	1 850
5		Comp vs. Rec	2 916	1 993	1 965	1 984
6		Comp vs. Talk	2 914	1 568	1 967	1 685
7	垃圾邮件过滤	User 1 vs. User2	User 1 的邮件		User 2 的邮件	
8		User 2 vs. User3	User 2 的邮件		User 3 的邮件	
9		User 3 vs. User1	User 3 的邮件		User 1 的邮件	

5.5.2 实验设置

20 Newsgroups 和垃圾邮件过滤是两个公开的跨领域文本分类任务,现有的许多 DAL 方法[3,6]通常采用它们来评价算法性能.20 Newsgroups 数据集包含 20 个新闻组类别,每个组大约包含 1 000 份文档.对于该文本分类任务,其主要目标是利用各顶层类别下的子类文档分别作为训练文档和测试文档来正确区分顶层类别下的新闻文档(如区分"Sci."文档和"talk"文档),其中,各子类文档集代表一个不同的领域.

在垃圾邮件过滤数据集中[3]有 3 个邮件子集,分别表示为 User 1,User 2 和 User 3,分别代表 3 个不同用户.在该 DAL 问题中,主要任务是识别出正常邮件和垃圾邮件.由于在各子集中正常和垃圾邮件由不同用户标识,使得各子集的数据分布相关但不同.各子集包含 2 500 邮件,其中一半为正常邮件(标签为 1),另一半为垃圾邮件(标签为−1).按照文献[6]的设置,本实验考虑 3 种情况:1) User 1(源领域) & User 2(目标领域);2) User 2(源领域) & User 3(目标领域);3) User 3(源领域) & User 1(目标领域).

在各实验中,训练数据集包含所有来自源领域的标签样本和随机采自目标领域的 2l 个标签样本(每类 l 个标签样本),目标领域中剩余样本用于测试数据,其中,对于 20 Newsgroups 数据集,设置 l=3;对于垃圾邮件数据集,设置 l=5.

5.5.3 实验结论

对于各数据集,每种算法均运行 5 次并记录最好的结果,并将最终的平均值和标准差记录于表 5.

Table 5 Means (%) and standard deviations of classification accuracies of all algorithms in text classification

表 5 在文本分类上所有算法分类精度的平均值(%)和标准差

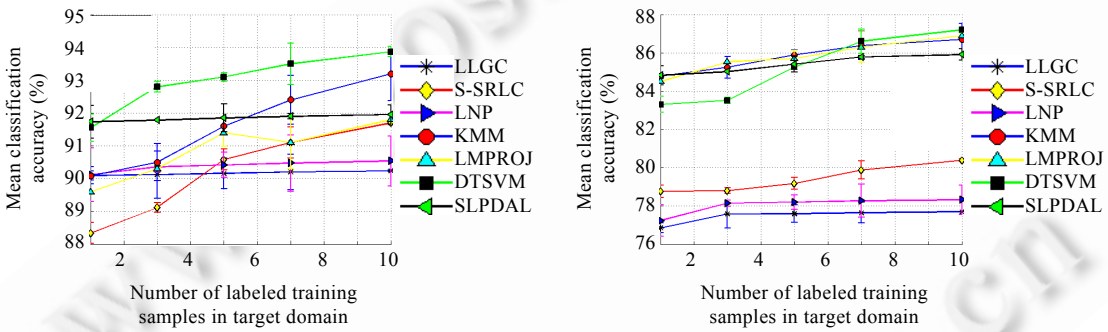
Datasets		Algorithms						
		LLGC	S-SRLC	LNP	KMM	LMPROJ	DTSVM	SLPDAL
20 newsgroups	1	77.58 (±0.14)	78.80 (±0.28)	78.15 (±3.56)	85.24 (±1.81)	**85.52** (±0.86)	83.52 (±2.74)	85.02 (±1.08)
	2	75.76 (±5.10)	77.86 (±1.06)	77.11 (±2.26)	78.60 (±0.34)	79.30 (±2.76)	80.5 (±2.82)	**82.67** (±0.30)
	3	82.17 (±3.22)	84.66 (±3.20)	84.52 (±2.69)	87.20 (±2.10)	86.34 (±3.10)	**90.23** **(±0.82)**	88.81 (±0.6)
	4	78.20 (±0.46)	79.84 (±0.60)	79.41 (±0.27)	75.32 (±0.47)	84.68 (±2.11)	84.84 (±1.49)	**85.37** **(±0.80)**
	5	90.12 (±0.72)	89.12 (±0.14)	90.36 (±0.16)	90.50 (±0.56)	90.30 (±0.01)	**92.80** (±0.16)	91.78 (±0.00)
	6	88.74 (±0.70)	89.20 (±0.16)	90.50 (±1.74)	94.00 (±2.06)	93.43 (±0.79)	94.13 (±0.4)	**96.78** **(±0.20)**
Email spam filtering	7	91.24 (±0.25)	93.24 (±0.42)	93.18 (±0.11)	93.51 (±0.40)	93.21 (±0.52)	96.89 (±0.1)	**97.19** **(±0.06)**
	8	89.0 (±2.30)	92.04 (±1.46)	92.52 (±0.18)	98.74 (±0.4)	94.0 (±0.00)	97.65 (±0.20)	**97.78** (±0.41)
	9	89.32 (±2.55)	90.12 (±1.40)	89.82 (±1.20)	88.78 (±0.01)	88.79 (±0.24)	93.50 (±0.60)	**93.85** **(±0.10)**

从表 6 中的结果能够得出如下几个有意义的结论:

1) 跨领域学习算法 LMPROJ 和 KMM 经常取得与几种 SSL 算法 LLGC,S-SRLC 和 LNP 相似的学习性能,可能的解释为:由于两个领域的数据分布非常相关或相近时,现有的 DAL 方法难以进一步提升在目标领域的学习性能.KMM 方法在两个数据集上的分类性能普遍差于 LMPROJ 方法,其原因可能是:基于转导学习(transductive learning)框架的 LMPROJ 方法比 KMM 方法更适于这些文本数据集分类.另外,在绝大多数情况下,采用 MKL 策略的 DTSVM 方法均优于其他方法(SLPDAL 方法除外),从而可以推断,MKL 技术在某些特定环境下确实能够有效地改进 DAL 性能.

2) 本文所提出的方法 SLPDAL 在两个数据集上的分类性能一般均优于其他方法,这是因为 SLPDAL 方法明确而综合性地考虑了 3 个方面的特征:(1) 领域数据分布的差异匹配;(2) 数据分布的整体与局部本质几何特征的一致性;(3) 同时利用源领域和目标领域先验信息.

为了进一步评价目标领域标签数据个数变化对所提出的方法鲁棒性的影响,令 l 值在区域 $l\in\{0,1,3,5,7,10\}$ 内变化,并在 20 Newsgroups 数据集中的任务 5 和任务 1 上进行重新实验,实验结果如图 9(a)、图 9(b)所示.从图 9 中可看出:目标领域标签样本数的增加,能够明显增强所有方法(尤其是 DAL 方法)的学习能力.这说明充分利用目标领域先验信息,可明显改善 DAL 方法的学习性能.



(a) Comp. vs. Rec. (b) Comp. vs. Sci.

Fig.9 Classification accuracies of all algorithms with different number of labeled training samples (m) from target domain

图 9 不同个数(m)的目标领域带标签训练数据下所有算法的分类精度

5.6 参数敏感性分析

本实验从表 5 中选取几个 DAL 任务(包括任务 3、任务 6 和任务 7),以明确阐述模型参数对所提出方法的分类性能影响.根据定理 6,实验设置γ_0=10,图 10(a)~图 10(d)分别显示,模型参数μ,λ,γ和η对所提方法的分类性能影响,图中的参数变化曲线是在固定其他 3 个最优参数的情况下绘制的.

图 10(a)~图 10(c)显示了所提出的方法在任务 6 和任务 7 上的参数敏感性,图 10(d)显示了所提出的方法在任务 3 上的参数敏感性.从这些图示曲线的变化可以得出如下结论:

(1) 图 10(a)显示了控制参数α的正则参数μ对所提方法的敏感性.根据定理 4 的证明可知,参数μ起到平衡预测损失和平滑性的作用.图中结果显示,该参数虽然从某种意义上来说对所提出的方法的分类精度起到一定的影响,但通过对 y 轴的观测发现,该影响效果非常有限.

(2) 图 10(b)显示,当λ=0 时,即,忽略领域间局部一致性时,所提出的方法不能取得最优学习性能.在一定的参数值区域内,随着λ的增加,所提出方法的性能逐渐缓慢提升,直到收敛于某个最大性能值;当λ=1,即,忽略领域间全局一致性(由稀疏保留正则项控制)时,所提出的方法在一定程度上呈现下降趋势.从以上分析可知,基于转导框架模型的 DAL 方法仅考虑领域数据分布的全局或局部特征是不完全的,应综合考虑领域数据的全局和局部本质分布特征的一致性,才可能获得最优的领域适应分类性能.

(3) 图 10(c)显示,一方面,参数γ的值越小(如$\gamma\in[1,3)$),即高斯核带宽越大,使得领域内数据分布散度就越

大,从而导致领域间分布差最小化的收敛速率减小;另一方面,参数γ的值越大(如$\gamma\in[6,+\infty)$),即高斯核带宽越小,导致领域内类间数据分布出现交叠.上述两种情况均可能导致所提出的方法分类性能的明显下降,而只有在一个适度的参数值域范围内(e.g. $\gamma\in[3,6)$),所提出的方法才可能获得最优的性能.

(4) 从图 10(d)可观察到:当参数值$\eta<4$ 时,SLPDAL 不能取得优化性能;但是,随着参数η值的增加,SLPDAL 分类精度能够得到稳步提升.

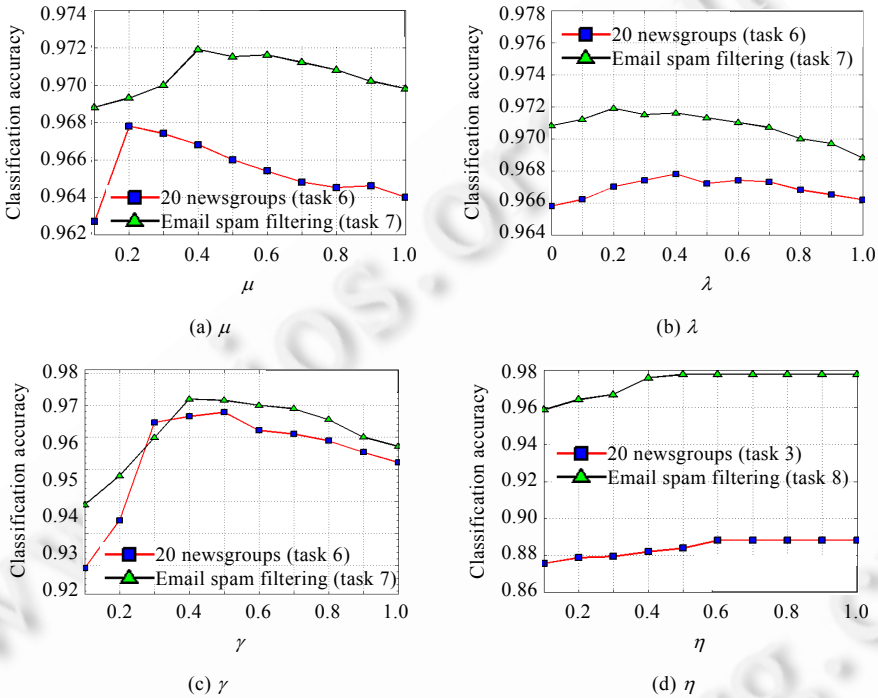


Fig.10 Sensitivity of different parameters in text classification datasets

图 10 不同参数在文本分类数据集上的敏感性

6 结束语

本文提出一种领域适应学习方法,即,稀疏标签传播领域适应学习(SLPDAL).基于领域间数据分布的最小化 MMD 准则,寻求一个 RKHS H,在 H 中,采用核稀疏表示技术来重构领域数据并构造一个稀疏图,并据此完成 SLPDAL 方法的标签传播过程,从而实现最终的跨领域学习任务.理论分析结果显示:基于领域数据分布的全局和局部一致正则化,SLPDAL 所预测的数据标签具有充分平滑性.在人造和实际 DAL 任务上的实验结果验证了所提出方法的鲁棒性和有效性.对于基于领域间数据分布最小化 MMD 准则的 DAL 方法,源领域数据集选择的有效性是决定其成败的一个非常重要的因素.现有研究成果表明:多源领域的集成有利于避免单一源领域可能导致的所谓“负迁移”现象[42],但是多个源领域的集成势必造成学习算法的计算复杂度增加.因此,基于多源领域有效集成的 SLPDAL 模型的构建,是本文值得进一步研究的一个方向.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是本文的各位审稿专家表示衷心的感谢.

References:

[1] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010,22(10):1345−1359. [doi: 10.1109/TKDE.2009.191]

[2] Pan SJ, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. IEEE Trans. on Neural Networks, 2011, 22(2):199−210. [doi: 10.1109/TNN.2010.2091281]

[3] Xiang EW, Cao B, Hu DH, Yang Q. Bridging domains using world wide knowledge for transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010,22(6):770−783. [doi: 10.1109/TKDE.2010.31]

[4] Duan LX, Tsang IW Xu D. Domain transfer multiple kernel learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,34(3):465−479. [doi: 10.1109/TPAMI.2011.114]

[5] Bruzzone L, Marconcini M. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(5):770−787. [doi: 10.1109/TPAMI.2009.57]

[6] Quanz B, Huan J. Large margin transductive transfer learning. In: Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM). New York: ACM Press, 2009. 1327−1336. [doi: 10.1145/1645953.1646121]

[7] Geng B, Tao D, Xu C. DAML: Domain adaptation metric learning. IEEE Trans. on Image Process, 2011,20(10):2980−2989. [doi: 10.1109/TIP.2011.2134107]

[8] Ling X, Dai WY, Xue GR, Yang Q, Yu Y. Spectral domain transfer learning. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2008. [doi: 10.1145/1401890.1401951]

[9] Jiang W, Zavesky E, Chang SF, Loui A. Cross-Domain learning methods for high-level visual concept classification. In: Proc. of the 15th IEEE Int'l Conf. on Image Processing. San Diego: IEEE Press, 2008. 161−164. [doi: 10.1109/ICIP.2008.4711716]

[10] Zhu X. Semi-Supervised learning literature survey. Technical Report, 1530, University of Wisconsin-Madison, 2005.

[11] Dai WY, Xue GR, Yang Q, Yu Y. Transferring Naive Bayes classifiers for text classification. In: Proc. of the 22nd AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2007. 540−545. http://www.aaai.org/Papers/AAAI/2007/AAAI07-085.pdf

[12] Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine Learning, 2000,39(2-3):103−134. [doi: 10.1023/A:1007692713085]

[13] Xing DK, Dai WY, Xue GR, Yu Y. Bridged refinement for transfer learning. In: Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases. Warsaw: PKDD Press, 2007. 324−335. [doi: 10.1007/978-3-540-74976-9_31]

[14] Wang F, Zhang C. Label propagation through linear neighborhoods. IEEE Trans. on Knowledge and Data Engineering, 2008,20(1): 55−67. [doi: 10.1109/TKDE.2007.190672]

[15] Belkin M, Niyogi P, Sindhwani V, Bartlett P. Manifold regularization: A geometric framework for learning from examples. Journal of Machine Learning Research, 2006,7(1):2399−2434.

[16] Wright J, Yang A, Sastry S, Ma Y. Robust face recognition via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(2):210−227. [doi: 10.1109/TPAMI.2008.79]

[17] Cheng H, Liu ZC, Yang J. Sparsity induced similarity measure for label propagation. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV). 2009. [doi: 10.1109/ICCV.2009.5459267]

[18] Fan M, Gu N, Qiao H, Zhang B. Sparse regularization for semi-supervised classification. Pattern Recognition, 2011,44(8):1777−1784. [doi: 10.1016/j.patcog.2011.02.013]

[19] Qiao L, Chen S, Tan X. Sparsity preserving projections with applications to face recognition. Pattern Recognition, 2010,43(1): 331−341. [doi: 10.1016/j.patcog.2009.05.005]

[20] Zheng M, Bu J, Chen C, Wang C, Zhang LJ, Qiu G, Cai D. Graph regularized sparse coding for image representation. IEEE Trans. on Image Processing, 2011,20(5):1327−1336. [doi: 10.1109/TIP.2010.2090535]

[21] Yan SC, Wang H. Semi-Supervised learning by sparse representation. In: Proc. of the SIAM Int'l Conf. on Data Mining (SDM). 2009.

[22] Cheng B, Yang JC, Yan SC, Fu Y, Huang TS. Learning with l_1-graph for image analysis. IEEE Trans. on Image Process, 2010, 19(4):858−866. [doi: 10.1109/TIP.2009.2038764]

[23] Xiao L, Dai B, Fang YQ, Wu T. Kernel l_1 graph for image analysis. In: Proc. of the CCPR 2012, CCIS 321. 2012. 447−454. [doi: 10.1007/978-3-642-33506-8_55]

[24] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290:2323−2326. [doi: 10.1126/science.290.5500.2323]

[25] Zhu X, Ghahramani Z, Lafferty J. Semi-Supervised learning using Gaussian fields and harmonic functions. In: Proc. of the 20th Int'l Conf. on Machine Learning. 2003.

[26] Zhou D, Bousquet O, Lal T, Weston J, Schölkopf B. Learning with local and global consistency. In: Proc. of the Advances in Neural Information Processing Systems 16. 2004.

[27] Wu MR, Schölkopf B. Transductive classification via local learning regularization. In: Proc. of the 11th Int'l Conf. on Artificial Intelligence and Statistics. Cambridge: MIT Press, 2007. 624−631.

[28] Wang F, Li T, Wang G, Zhang C. Semi-Supervised classification using local and global regularization. In: Proc. of the 23rd AAAI Conf. on Artificial Intelligence (AAAI). Chicago, 2008.

[29] Gretton A, Harchaoui Z, Fukumizu K, Sriperumbudur BK. A fast, consistent kernel two-sample test. In: Proc. of the Advances in Neural Information Processing Systems 22. MIT Press, 2010. 673−681.

[30] Sriperumbudur BK, Gretton A, Fukumizu K, Schölkopf B, Lanckriet GRG. Hilbert space embeddings and metrics on probability measures. Journal of Machine Learning Research, 2010,11(3):1517−1561.

[31] Gu QQ, Zhou J. Transductive classification via dual regularization. In: Proc. of the 19th European Conf. on Machine Learning (ECML). Bled, 2009. 439−454. [doi: 10.1007/978-3-642-04180-8_46]

[32] Li HX, Gao YS, Sun J. Fast kernel sparse representation. In: Proc. of the 2011 Int'l Conf. on Digital Image Computing: Techniques and Applications. Washington: IEEE Computer Society Press, 2011. 72−77. [doi: 10.1109/DICTA.2011.20]

[33] He X, Yan S, Hu Y, Niyogi P, Zhang HJ. Face recognition using Laplacian faces. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(3):328−340. [doi: 10.1109/TPAMI.2005.55]

[34] Liu W, Wang J, Chang SF. Robust and scalable graph-based semisupervised learning. Proc. of the IEEE, 2012,100(9):2624−2638. [doi: 10.1109/JPROC.2012.2197809]

[35] Tao JW, Wang ST. Kernel distribution consistency based local domain adaptation learning. Acta Automatica Sinica, 2013,39(8): 1295−1309 (in Chinese with English abstract).

[36] Nie F, Zeng Z, Tsang IW, Xu D, Zhang C. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. IEEE Trans. on Neural Networks, 2011,22(11):1796−1808. [doi: 10.1109/TNN.2011.2162000]

[37] Rakotomamonjy A, Bach FR, Canu S, Grandvalet Y. SimpleMKL. Journal of Machine Learning Research, 2008,9:2491−2521.

[38] Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B. Large scale multiple kernel learning. Journal of Machine Learning Research, 2006,7:1531−1565.

[39] Sriperumbadur BK, Fukumizu K, Gretton A, Lanckriet GRG, Schölkopf B. Kernel choice and classifiability for RKHS embeddings of probability distributions. In: Proc. of the Advances in Neural Information Processing Systems 22. MIT Press, 2010. 1750−1758.

[40] Huang J, Smola A, Gretton A, Borgwardt KM, SchÄolkopf B. Correcting sample selection bias by unlabeled data. In: Proc. of the 20th Annual Conf. on Neural Information Processing Systems. 2006.

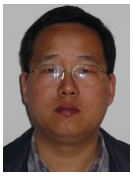
[41] Chua TS, Tang JH, Hong RC, Li HJ, Luo ZP, Zheng YT. NUS-WIDE: A real-world Web image database from National University of Singapore. ACM Int'l Conf. on Image and Video Retrieval, 2009,48(9):1−48.

[42] Yang J, Tong W, Hauptmann AG. A framework for classifier adaptation for large-scale multimedia data. Proc. of the IEEE, 2012, 100(9):2639−2657. [doi: 10.1109/JPROC.2012.2204009]

附中文参考文献:

[35] 陶剑文,王士同.核分布一致局部领域适应学习.自动化学报,2013,39(8):1295−1309.

陶剑文(1973−),男,湖北武汉人,博士,教授,主要研究领域为模式识别,数据挖掘.



王士同(1964−),男,教授,博士生导师,主要研究领域为人工智能,机器学习.

Fu-Lai CHUNG(1965−),男,博士,副教授,主要研究领域为人工智能,模糊神经网络.

姚奇富(1965−),男,教授,博士生导师,主要研究领域为数据挖掘.