文章编号: 1001-0920 (2016) 11-2083-07 DOI: 10.13195/j.kzyjc.2015.1189

基于最大分布加权均值嵌入的领域适应学习

臧绍飞, 程玉虎, 王雪松

(中国矿业大学 信息与电气工程学院，江苏 徐州 221116)

摘 要: 最大均值差异忽略了单个样本对全局度量贡献的差异性. 为此, 提出一种最大分布加权均值差异度量方法, 采用白化余弦相似性度量为源域和目标域的所有样本设计相应的分布权重, 使得每个样本的分布差异信息在全局度量中均得以体现. 进一步, 结合联合分布调整思想, 提出一种基于最大分布加权均值嵌入的领域适应学习算法. 实验结果表明, 与典型的迁移学习和无迁移学习算法相比, 所提出算法在不同类型跨领域图片数据集上均具有较高的分类精度.

关键词: 领域适应学习；最大均值差异；分布权重；白化余弦相似性；联合分布调整

中图分类号: TP391.4 **文献标志码:** A

Domain adaptation learning based on maximum distribution weighted mean discrepancy

ZANG Shao-fei, *CHENG Yu-hu*, *WANG Xue-song*

(School of Information and Electrical Engineering，China University of Mining and Technology，Xuzhou 221116，China. Correspondent：CHENG Yu-hu，E-mail：chengyuhu@163.com)

Abstract: Maximum mean discrepancy neglects the difference of contribution of each sample to the global measure. Therefore, a kind of maximum distribution weighted mean discrepancy(MDWMD) methed is proposed, where the whitened cosine similarity is used to design distribution weights for all samples from the source and the target domains. Further, based on the idea of joint distribution adaptation, a domain adaptation learning algorighm based on MDWMD is proposed. Experimental results show that, compared with the typical transfer learning and non-transfer learning algorithms, the proposed algorithm has higher classification accuracy on different types of cross-domain image datasets.

Keywords: domain adaptation learning；maximum mean discrepancy；distribution weight；whitened cosine similarity; joint probaliality adaptation

0 引 言

在计算机视觉应用中, 伴随着互联技术的发展和数据更新速度的加快, 单一领域内可训练有价值的标记样本变得稀少, 而对每个领域内的数据进行人工标记, 成本又过高. 另一方面, 传统的机器学习方法由于假设训练数据与测试数据独立且服从相同的分布, 导致大量标记且不同分布的数据被抛弃, 造成浪费. 迁移学习打破了传统机器学习要求训练数据集与目标数据集必须遵守相同分布和相同特征空间的限制, 使得知识可以跨领域迁移[1-2].

领域适应学习 (DAL) 方法是迁移学习的一种特类, 其任务是传递和共享不同任务或域之间的知识[3]. 在DAL方法中, 一个最主要的计算性问题就是如何选取有效的度量来反映源域与目标域之间的分布差异. 目前, 常用的用于度量领域间分布差异的方法主要有Bregman差异[4]、基于熵的KL距离[5]和最大均值差异 (MMD)[6]. Bregman差异以失真函数为目标函数, 用于度量两个样本间的相似度差异. 由于一般采用梯度下降法求解目标函数, 从而导致Bregman差异度量计算产生较大的时间花销[7]. KL距离常用于度量两个概率分布 (如高斯分布[8]) 的距离, 是一种带参的估计方法, 在度量源域与目标域分布差异过程中需要不断地进行先验概率密度估计. MMD度量则是通过计算源域与目标域之间的均值差来反映两个区域

收稿日期: 2015-09-23；**修回日期:** 2015-12-21.

基金项目: 国家自然科学基金项目(61273143, 61472424)；中央高校基本科研业务费专项资金项目(2013RC10, 2013 RC12, 2014YC07).

作者简介: 臧绍飞 (1981−), 男, 博士生, 从事知识迁移学习的研究；程玉虎 (1973−), 男, 教授, 博士生导师, 从事机器学习、模式识别与智能系统等研究.

之间的分布差异, 是一种无参估计标准. 与Bregman差异和KL距离相比, MMD度量计算简单有效、直观且易于理解. 将MMD度量与传统PCA相结合, Pan等[9]提出了一种具有迁移学习能力的特征提取方法: 最大均值差异嵌入(MMDE). MMDE通过在源域上所获得的知识来构造适合目标域的特征提取技术, 实现了领域间的知识迁移. 由于MMDE在计算过程中需要求解半定规划问题, 会增加算法的计算开销. 为了克服MMDE在求解方面的不足, Pan等[10]提出了一个有效的特征提取算法: 迁移主成分分析(TCA). 由MMD度量的定义可知, MMD是使用源域与目标域的总体均值之差来表示两个领域之间分布差异的. 进一步, 由统计学理论可知, 总体均值一般反映的是样本空间总体的分布信息和全局结构信息. 因此, 从这一层面上讲, 可以认为MMDE和TCA是一种全局方法, 以致于在一定程度上忽视了样本空间内在的局部结构和局部信息. 在MMD的基础上, 皋军等[11-12]提出了具有局部学习能力的迁移学习度量, 该度量不仅能够有效地反映源域与目标域之间存在的局部分布差异, 而且也在一定程度上表明了区域内部存在的局部结构之间的差异. 为进一步提升MMD的度量性能, Long等[13-14]在目标域没有标记数据的场景下, 利用源域标记样本训练所得分类器给目标域未标记的样本赋予伪标签, 并将标签信息引入MMD中, 设计了一种联合分布调整算法(JDA), 能够同时度量并调整领域间的边缘分布差异和条件分布差异.

通过分析发现: 1) 文献[9-10,13-14]均忽略了样本个体在数据集中分布差异的因素, 从而影响了算法的性能; 2) 文献[11-12]虽然考虑了样本分布差异的因素, 但算法采用最近邻分块局部加权方法仅对源域和目标域部分样本的权重进行计算, 且样本分布差异权重受热核参数和近邻样本数影响较大, 同时, 提出的迁移学习框架未考虑领域间的条件分布差异. 受文献[15]中加权均值思想的启发, 本文通过设计每个样本的分布权重系数, 在MMD准则基础上提出一种跨领域分布差异度量方法: 最大分布加权均值差异(MDWMD), 不仅能够反映全局分布的差异性, 而且能够体现出数据集中单个样本对度量领域间差异度量的贡献差异性. 进一步, 在MDWMD基础上, 结合联合分布调整思想, 提出一种领域适应学习算法: 基于最大分布加权均值嵌入的联合分布调整(JDA-MDWMD). 实验结果表明了本文方法的有效性.

1 基于样本分布信息的最大分布加权均值差异

给定两组数据集, 共$n=n_S+n_T$个样本, 一组是已标记的高维样本源域数据集$\boldsymbol{D}_S=[\boldsymbol{x}_{S(1)},\cdots,\boldsymbol{x}_{S(n_S)}]$及其标签$\boldsymbol{Y}_S=[y_{S(1)},\cdots,y_{S(i)},\cdots,y_{S(n_S)}]$, 另一组是未标记目标域数据集$\boldsymbol{D}_T=\{(\boldsymbol{x}_{T(j)})\}|_{j=1}^{n_T}$, 目标域数据集的标签$\boldsymbol{Y}_T=[y_{T(1)},\cdots,y_{T(j)},\cdots,y_{T(n_T)}]$是未知的, 其中$\boldsymbol{x}_{S(i)}$, $\boldsymbol{x}_{T(j)}\in R^m$, $y_{S(i)},y_{T(j)}\in\{1,\cdots,c,\cdots,C\}$. 设这两个领域的边缘分布和条件分布分别为: $P_S(\boldsymbol{x}_S)$和$P_T(\boldsymbol{x}_T)$, $Q_S(y_S|\boldsymbol{x}_S)$和$Q_T(y_T|\boldsymbol{x}_T)$. 一般情况下$\boldsymbol{D}_S$与$\boldsymbol{D}_T$间会存在分布差异, 即$P_S(\boldsymbol{x}_S)\neq P_T(\boldsymbol{x}_T)$和$Q_S(y_S|\boldsymbol{x}_S)\neq Q_T(y_T|\boldsymbol{x}_T)$. 领域适应学习的目标是找到一个低维子空间来表示高维数据, 即

$$\boldsymbol{z}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}\in R^k. \tag{1}$$

其中: $\boldsymbol{z}$为高维数据$\boldsymbol{x}$的低维表示, $k\ll m$, 投影矩阵$\boldsymbol{W}=(w_1,w_2,\cdots,w_k)\in R^{m\times k}$. 通过$\boldsymbol{W}$使得$\boldsymbol{D}_S$和$\boldsymbol{D}_T$投影转换得到的数据$\boldsymbol{Z}_S$与$\boldsymbol{Z}_T$间的分布差异减小, 即$P_S(\boldsymbol{z}_S)\approx P_T(\boldsymbol{z}_T)$和$Q_S(y_S|\boldsymbol{z}_S)\approx Q_T(y_T|\boldsymbol{z}_T)$, 以便在$\{\boldsymbol{Z}_S,\boldsymbol{Y}_S\}$上训练得到的分类器对$\boldsymbol{Z}_T$分类时获得良好的效果.

由MMD[6]可知, 估计不同分布样本间的距离是通过最小化下面目标函数实现的:

$$\mathrm{dist}^2(\boldsymbol{D}_S,\boldsymbol{D}_T)=\left\|\frac{1}{n_S}\sum_{i=1}^{n_S}\varphi(\boldsymbol{x}_{S(i)})-\frac{1}{n_T}\sum_{j=1}^{n_T}\varphi(\boldsymbol{x}_{T(j)})\right\|_{\mathcal{H}}^2. \tag{2}$$

其中: $\|\cdot\|_{\mathcal{H}}^2$为Hilbert空间中的平方模, $\varphi(\boldsymbol{x})$是一个从原始输入空间到可再生高维Hilbert空间中的核映射函数. 同TCA[10]相似, 将MMD作为低维空间内跨领域边缘分布调整[13]时, 有

$$\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\left\|\frac{1}{n_S}\sum_{i=1}^{n_S}\boldsymbol{W}^{\mathrm{T}}\varphi(\boldsymbol{x}_{S(i)})-\frac{1}{n_T}\sum_{j=1}^{n_T}\boldsymbol{W}^{\mathrm{T}}\varphi(\boldsymbol{x}_{T(j)})\right\|_{\mathcal{H}}^2=$$

$$\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}). \tag{3}$$

其中: $\boldsymbol{K}=\varphi(\boldsymbol{X})^{\mathrm{T}}\varphi(\boldsymbol{X})$为数据集$\boldsymbol{X}$的核矩阵; $\boldsymbol{X}=\boldsymbol{D}_S\bigcup\boldsymbol{D}_T$; $\boldsymbol{L}_{n\times n}$为分布差异矩阵, 有

$$L_{ij}=\begin{cases}\dfrac{1}{n_Sn_S}, & \boldsymbol{x}_{S(i)},\boldsymbol{x}_{S(j)}\in\boldsymbol{D}_S;\\ \dfrac{1}{n_Tn_T}, & \boldsymbol{x}_{T(i)},\boldsymbol{x}_{T(j)}\in\boldsymbol{D}_T;\\ \dfrac{-1}{n_Sn_T}, & \text{其他}.\end{cases} \tag{4}$$

由式(4)可知, 传统的MMD认为, 单一数据集中每个样本对本数据集的内在结构信息和分布信息的贡献程度是一样的, 而事实上是不同的. 假定数据集$\boldsymbol{D}_S$和$\boldsymbol{D}_T$均满足高斯分布[15-16], 假设$\boldsymbol{\beta}_i$和$\boldsymbol{\beta}_j$分别为源域数据集中样本$\boldsymbol{x}_{S(i)}$和目标域数据集中样本$\boldsymbol{x}_{T(j)}$在本数据集中的分布权重, 它们可以反映对应样本对本数据集分布信息的贡献度, 一般可以通过$\boldsymbol{\beta}=\exp\left(-\dfrac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{h}\right)$求得. 其中: $\boldsymbol{\mu}$为数据集的均值, h为热核参数. 热核参数的选取方法主要有: 经验选择

法、实验试凑法和正交试验设计法. 经验选择要求对所研究问题拥有很好的经验和十足的知识, 否则, 并不容易获得合适的参数. 实验试凑是通过大量的数字仿真实验来获得较优的参数, 比较费时, 而且获得的参数也不一定是最优的. 正交试验设计法的缺点是计算量大. 白化余弦相似性度量是计算机视觉和模式识别中最常使用的相似性度量方法之一, 它反映了两个不同模式向量间的关系[17]. 因此, 本文考虑采用白化余弦相似性度量来求取样本的分布权重 $\boldsymbol{\beta}$, 其表达式为

$$\boldsymbol{\beta}=\frac{1}{\delta_{WC}(\boldsymbol{x},\boldsymbol{\mu})}=\frac{||A^{\mathrm{T}}\boldsymbol{x}||\cdot||\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}||}{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu})}. \tag{5}$$

其中: $\delta_{WC}(\boldsymbol{x},\boldsymbol{\mu})$ 为白化余弦相似性度量, $||\cdot||$ 表示取模, 白化转换矩阵 $\boldsymbol{A}$ 可以通过协方差矩阵计算. 采用 PCA 对所有样本的协方差矩阵 $\boldsymbol{\Sigma}=E[(\boldsymbol{x}-\boldsymbol{M}_0)\times(\boldsymbol{x}-\boldsymbol{M}_0)^{\mathrm{T}}]$ 进行分解, 可得 $\boldsymbol{\Sigma}=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$, 进一步可得 $\boldsymbol{A}=\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2}$. 其中: $\boldsymbol{M}_0=\mathrm{E}(\boldsymbol{x})$ 表示总均值矢量, $\mathrm{E}(\cdot)$ 表示期望, $\boldsymbol{V}$ 表示特征向量矩阵, $\boldsymbol{\Lambda}$ 表示对角特征值矩阵.

将源域和目标域数据集样本的分布权重代入式 (2), 可将 MMD 推广至线性最大分布加权均值差异 (MDWMD) 度量, 其目标函数为

$$\begin{aligned}&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{dist}^2(\boldsymbol{D}_S,\boldsymbol{D}_T)=\\&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\Big\|\sum_{i=1}^{n_S}\frac{\beta_{S(i)}\boldsymbol{W}^{\mathrm{T}}\varphi(\boldsymbol{x}_{S(i)})}{\sum\limits_{i=1}^{n_S}\beta_{S(i)}}-\\&\sum_{j=1}^{n_T}\frac{\beta_{T(j)}\boldsymbol{W}^{\mathrm{T}}\varphi(\boldsymbol{x}_{T(j)})}{\sum\limits_{j=1}^{n_T}\beta_{T(j)}}\Big\|_{\mathcal{H}}^2.\end{aligned} \tag{6}$$

如果将源域和目标域的数据权重分别按如下方式进行扩充:

$$\begin{aligned}&\boldsymbol{\beta}_S=\\&\Big(\frac{\beta_{S(1)}}{\sum\limits_{i=1}^{n_S}\beta_{S(i)}},\frac{\beta_{S(2)}}{\sum\limits_{i=1}^{n_S}\beta_{S(i)}},\cdots,\frac{\beta_{S(n_S)}}{\sum\limits_{i=1}^{n_S}\beta_{S(i)}},\underbrace{0,\cdots,0}_{n_T}\Big)^{\mathrm{T}},\end{aligned} \tag{7}$$

$$\begin{aligned}&\boldsymbol{\beta}_T=\\&\Big(\underbrace{0,\cdots,0}_{n_S},\frac{\beta_{T(1)}}{\sum\limits_{j=1}^{n_T}\beta_{T(j)}},\frac{\beta_{T(2)}}{\sum\limits_{j=1}^{n_T}\beta_{T(j)}},\cdots,\frac{\beta_{T(n_T)}}{\sum\limits_{j=1}^{n_T}\beta_{T(j)}}\Big)^{\mathrm{T}},\end{aligned} \tag{8}$$

则式 (6) 可写为

$$\begin{aligned}&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{dist}^2(\boldsymbol{D}_S,\boldsymbol{D}_T)=\\&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta}_S-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta}_T\|_{\mathcal{H}}^2=\\&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{\beta}_S\boldsymbol{\beta}_S^{\mathrm{T}}+\boldsymbol{\beta}_T\boldsymbol{\beta}_T^{\mathrm{T}}-\\&\boldsymbol{\beta}_S\boldsymbol{\beta}_T^{\mathrm{T}}-\boldsymbol{\beta}_S^{\mathrm{T}}\boldsymbol{\beta}_T)\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W})=\\&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{L}_0\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}).\end{aligned} \tag{9}$$

其中 $\boldsymbol{L}_0=\boldsymbol{\beta}_S\boldsymbol{\beta}_S^{\mathrm{T}}+\boldsymbol{\beta}_T\boldsymbol{\beta}_T^{\mathrm{T}}-\boldsymbol{\beta}_S\boldsymbol{\beta}_T^{\mathrm{T}}-\boldsymbol{\beta}_S^{\mathrm{T}}\boldsymbol{\beta}_T$ 为边缘分布差异权重矩阵, 有

$$(L_0)_{ij}=\begin{cases}\beta_{S(i)}\beta_{S(j)},\ \boldsymbol{x}_{S(i)},\boldsymbol{x}_{S(j)}\in\boldsymbol{D}_S;\\ \beta_{T(i)}\beta_{T(j)},\ \boldsymbol{x}_{T(i)},\boldsymbol{x}_{T(j)}\in\boldsymbol{D}_T;\\ -\beta_{S(i)}\beta_{T(j)},\ \text{其他}.\end{cases} \tag{10}$$

2　基于 MDWMD 的领域适应学习

目前, 跨领域学习算法大多是对边缘概率分布或条件概率分布单独进行调整, 同时对两种分布进行调整的算法较少[13]. 缩小领域间的边缘分布差异或条件分布差异均能实现跨领域的知识迁移, 但若对两种分布差异同时进行调整, 则能进一步提高领域适应算法的鲁棒性和分类精度. 基于 MDWMD 的领域适应学习的算法流程是: 首先, 对源域和目标域中的数据进行边缘概率分布调整; 其次, 对条件分布进行调整; 再次, 采用 PCA 进行特征提取, 求取最佳投影矩阵; 最后, 利用投影到子空间中的数据训练出一个分类器, 对目标域的未标记样本进行分类. 由于目标域数据集中样本都是未标记的, 算法在运算过程中先给目标域数据赋予伪标签, 然后再通过迭代精化其标签.

2.1　联合分布调整

联合分布调整需要对跨领域数据集间的边缘分布和条件分布同时进行调整. 对给定的不同但相关的跨领域数据集进行领域间的边缘分布调整时, 其目标函数为式 (9). 领域间的条件分布调整是通过在边缘分布调整的基础上加入标签信息来实现的, 其目标函数为

$$\begin{aligned}&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{dist}^2(\boldsymbol{D}_T,\boldsymbol{D}_S)=\\&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\sum_{c=1}^{C}\Big\|\sum_{i=1}^{n_S^{(c)}}\frac{\beta_{S(i)}^{(c)}\boldsymbol{W}^{\mathrm{T}}\varphi(\boldsymbol{x}_{S(i)}^{(c)})}{\sum\limits_{i=1}^{n_S^{(c)}}\beta_{S(i)}^{(c)}}-\\&\sum_{j=1}^{n_T^{(c)}}\frac{\beta_{T(j)}^{(c)}\boldsymbol{W}^{\mathrm{T}}\varphi(\boldsymbol{x}_{T(j)}^{(c)})}{\sum\limits_{j=1}^{n_T^{(c)}}\beta_{T(j)}^{(c)}}\Big\|_{\mathcal{H}}^2=\\&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\sum_{c=1}^{C}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}^{(c)}\boldsymbol{\beta}_S^{(c)}-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}^{(c)}\boldsymbol{\beta}_T^{(c)}\|_{\mathcal{H}}^2=\\&\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\sum_{c=1}^{C}\mathrm{tr}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}^{(c)}(\boldsymbol{\beta}_S^{(c)}(\boldsymbol{\beta}_S^{(c)})^{\mathrm{T}}+\\&\boldsymbol{\beta}_T^{(c)}(\boldsymbol{\beta}_T^{(c)})^{\mathrm{T}}-\boldsymbol{\beta}_S^{(c)}(\boldsymbol{\beta}_T^{(c)})^{\mathrm{T}}-\end{aligned}$$

$$(\boldsymbol{\beta}_S^{(c)})^{\mathrm{T}}\boldsymbol{\beta}_T^{(c)})(\boldsymbol{K}^{(c)})^{\mathrm{T}}\boldsymbol{W}\}=$$
$$\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\sum_{c=1}^{C}\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}^{(c)}\boldsymbol{L}_c(\boldsymbol{K}^{(c)})^{\mathrm{T}}\boldsymbol{W})=$$
$$\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{tr}\Big(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\sum_{c=1}^{C}\boldsymbol{L}_c\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}\Big). \tag{11}$$

其中: $\boldsymbol{K}^{(c)}$ 为 $\{\boldsymbol{D}_S^{(c)}\bigcup\boldsymbol{D}_T^{(c)}\}$ 的核映射矩阵, $\boldsymbol{D}_S^{(c)}$ 和 $\boldsymbol{D}_T^{(c)}$ 为源域和目标域中标签类别为 c 的样本子集, 而

$$\boldsymbol{\beta}_S^{(c)}=$$
$$\Big(\frac{\beta_{S(1)}^{(c)}}{\sum_{i=1}^{n_S^{(c)}}\beta_{S(i)}^{(c)}},\frac{\beta_{S(2)}^{(c)}}{\sum_{i=1}^{n_S^{(c)}}\beta_{S(i)}^{(c)}},\cdots,\frac{\beta_{S(n_S^{(c)})}^{(c)}}{\sum_{i=1}^{n_S^{(c)}}\beta_{S(i)}^{(c)}},\underbrace{0,\cdots,0}_{n_T^{(c)}}\Big)^{\mathrm{T}},$$

$$\boldsymbol{\beta}_T^{(c)}=$$
$$\Big(\underbrace{0,\cdots,0}_{n_S^{(c)}},\frac{\beta_{T(1)}^{(c)}}{\sum_{i=1}^{n_T^{(c)}}\beta_{T(i)}^{(c)}},\frac{\beta_{T(2)}^{(c)}}{\sum_{i=1}^{n_T^{(c)}}\beta_{T(i)}^{(c)}},\cdots,\frac{\beta_{T(n_S^{(c)})}^{(c)}}{\sum_{i=1}^{n_T^{(c)}}\beta_{T(i)}^{(c)}}\Big)^{\mathrm{T}},$$

$$\boldsymbol{L}_c=$$
$$\boldsymbol{\beta}_S^{(c)}(\boldsymbol{\beta}_S^{(c)})^{\mathrm{T}}+\boldsymbol{\beta}_T^{(c)}(\boldsymbol{\beta}_T^{(c)})^{\mathrm{T}}-\boldsymbol{\beta}_S^{(c)}(\boldsymbol{\beta}_T^{(c)})^{\mathrm{T}}-(\boldsymbol{\beta}_S^{(c)})^{\mathrm{T}}\boldsymbol{\beta}_T^{(c)}$$

为条件分布差异权重矩阵. 有

$$(L_c)_{ij}=\begin{cases}\beta_{S(i)}^{(c)}\beta_{S(\mathrm{j})}^{(c)},\ \boldsymbol{x}_{S(i)}^{(c)},\boldsymbol{x}_{S(j)}^{(c)}\in\boldsymbol{D}_S^{(c)};\\ \beta_{T(i)}^{(c)}\beta_{T(j)}^{(c)},\ \boldsymbol{x}_{T(i)}^{(c)},\boldsymbol{x}_{T(j)}^{(c)}\in\boldsymbol{D}_T^{(c)};\\ -\beta_{S(i)}^{(c)}\beta_{T(j)}^{(c)},\ \begin{cases}\boldsymbol{x}_{S(i)}^{(c)}\in\boldsymbol{D}_S^{(c)}\ \text{且}\ \boldsymbol{x}_{T(j)}^{(c)}\in\boldsymbol{D}_T^{(c)};\\ \boldsymbol{x}_{S(j)}^{(c)}\in\boldsymbol{D}_S^{(c)}\ \text{且}\ \boldsymbol{x}_{T(i)}^{(c)}\in\boldsymbol{D}_T^{(c)};\end{cases}\\ 0,\text{其他}.\end{cases} \tag{12}$$

这里: $\boldsymbol{x}_{S(i)}^{(c)}$ 和 $\boldsymbol{x}_{T(j)}^{(c)}$ 分别为 $\boldsymbol{D}_S^{(c)}$ 和 $\boldsymbol{D}_T^{(c)}$ 中的样本, $n_S^{(c)}$ 和 $n_T^{(c)}$ 分别表示 $\boldsymbol{D}_S^{(c)}$ 和 $\boldsymbol{D}_T^{(c)}$ 中的样本个数, $\beta_{S(i)}^{(c)}$ 和 $\beta_{T(j)}^{(c)}$ 分别为 $\boldsymbol{x}_{S(i)}^{(c)}$ 和 $\boldsymbol{x}_{T(j)}^{(c)}$ 所对应的样本分布权重, 有

$$\beta^{(c)}=\frac{1}{\delta_{WC}(\boldsymbol{x}^{(c)},\boldsymbol{\mu}^{(c)})}=\frac{||\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}^{(c)}||\cdot||\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}^{(c)}||}{(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}^{(c)})^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}^{(c)})}, \tag{13}$$

$\boldsymbol{x}^{(c)}$、$\boldsymbol{\mu}^{(c)}$ 分别为相应数据集中标签类别为 c 的样本和样本均值.

2.2 投影矩阵

为了求取最佳投影矩阵, 实现空间转换和降维的目的, 此处采用 PCA 算法来实现, 其目标函数为

$$\max_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{H}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}). \tag{14}$$

其中: $\boldsymbol{H}=\boldsymbol{I}-\dfrac{1}{n}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ 为中心矩阵, $\boldsymbol{u}\in R^n$ 为元素为 1 的列向量. 对式 (14) 进行特征值分解, 可得

$$\boldsymbol{K}\boldsymbol{H}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{W}\boldsymbol{\Phi}, \tag{15}$$

其中 $\boldsymbol{\Phi}=\mathrm{diag}(\phi_1,\cdots,\phi_k)\in R^{k\times k}$ 是由前 k 个最大特征值构成的对角矩阵, 其对应的特征向量矩阵即为投影矩阵 $\boldsymbol{W}$.

2.3 目标函数

为了对边缘概率分布和条件概率分布同时进行调整, 联合式 (9)、(11) 和 (14), 即可得到 JDA-MDWMD 的目标函数为

$$\min_{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{H}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I}}\mathrm{tr}\Big(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\sum_{c=0}^{C}\boldsymbol{L}_c\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}\Big)+\lambda\|\boldsymbol{W}\|_F^2, \tag{16}$$

其中 λ 是控制模型复杂度并保证优化问题适定的正则化平衡参数. 此处, 将边缘概率分布调整的情况视为条件概率分布调整 $c=0$ 的特殊形式. 当 $c=0$ 时, 可理解为标签信息未参与领域间的分布差异调整, 即边缘分布调整. 结合式 (15), 对式 (16) 进行特征值分解, 可得

$$\Big(\boldsymbol{K}\sum_{c=0}^{C}\boldsymbol{L}_c\boldsymbol{K}^{\mathrm{T}}+\lambda\boldsymbol{I}\Big)\boldsymbol{W}=\boldsymbol{K}\boldsymbol{H}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\Phi}. \tag{17}$$

2.4 算法步骤

输入: 源域数据集 $\boldsymbol{D}_S$ 及其标记 $\boldsymbol{Y}_S$, 目标域数据集 $\boldsymbol{D}_T$, 低维子空间维数 k, 正则化参数 λ, 基分类器 f, 最大迭代次数 $T_{\max}$;

输出: 低维投影矩阵 $\boldsymbol{W}$.

1) 根据式 (10), 构建边缘分布差异权重矩阵 $\boldsymbol{L}_0$;

2) $t=1$;

3) 求解式 (17), 得投影矩阵 $\boldsymbol{W}$;

4) 由式 (1) 将 $\boldsymbol{D}_S$ 和 $\boldsymbol{D}_T$ 分别映射到 k 维子空间中, 得到 $\boldsymbol{Z}_S$ 和 $\boldsymbol{Z}_T$;

5) 在数据集 $\{\boldsymbol{Z}_S,\boldsymbol{Y}_S\}$ 上训练基分类器 f, 并利用训练得到的分类器对 $\boldsymbol{Z}_T$ 进行标记, 得到目标域数据的标签集 $\tilde{\boldsymbol{Y}}_T$;

6) 采用数据集 $\{\boldsymbol{D}_S,\boldsymbol{Y}_S\}$ 和 $\{\boldsymbol{D}_T,\tilde{\boldsymbol{Y}}_T\}$, 根据式 (12) 构建条件分布差异矩阵 $\boldsymbol{L}_c$;

7) $t\leftarrow t+1$;

8) 若 $t<T_{\max}$, 则转 3).

3 实验结果及分析

3.1 实验数据集

实验数据集采用如表 1 所示的 USPS 与 MNIST 数字手写体、COIL20 物品类别、PIE 人脸以及 Office 与 Caltech 显示屏等 6 个图片数据集. 在这 6 个数据集上, 共设计 36 组迁移学习实验.

USPS 与 MNIST 均为描述手写体数字的图片数据集, 它们分布不同但相关, 共有 10 个数字类别. 实验过程中, 按下述方式构造 2 组实验数据 (USPS vs MNIST, MNIST vs USPS): 从 USPS 中随机抽取 1 800 幅图片分别作为源域和目标域数据集, 相应地, 从

表 1　实验数据集描述

数据集	数据类型	样本个数	特征维数	类别数	包含子集
USPS	Digit	1 800	256	10	USPS
MNIST	Digit	2 000	256	10	MNIST
COIL20	Object	1 440	1 024	20	COIL1, COIL2
PIE	Face	11 554	1 024	68	PIE1, $\cdots$, PIE5
Office	Object	1 410	800	10	A, W, D
Caltech	Object	1 123	800	10	C

MNIST 中随机抽取 2 000 幅图片作为目标域和源域数据集. USPS 和 MNIST 中的所有图片统一变换成像素为 16×16, 且将每幅图片变成由灰度值表示像素点的灰度图.

COIL20 包含 20 类物品, 共计 1 440 幅图片. 每个物品有 72 幅图片, 且被旋转 5 个角度. 每幅图片的像素规模为 32×32, 每个像素点为 0～256 的灰度值. 实验过程中, 将 COIL20 数据集划分成 COIL1 和 COIL2 两个子数据集: COIL1 含有角度方向为 $[0°, 85°] \bigcup [180°, 265°]$ (即第 1, 第 3 象限) 的所有图片: COIL1 含有角度方向为 $[90^0, 175^0] \bigcup [270^0, 355^0]$ (即第 2, 第 4 象限) 的所有图片, 则 COIL1 和 COIL2 满足了不同分布但相关的要求. 按下述方式构造两组实验数据 (COIL1 vs COIL2, COIL2 vs COIL1): 从 COIL1 中选取 720 幅图片分别作为源域和目标域数据集; 相应地, 从 COIL2 中选取 720 幅图片分别作为目标域和源域数据集.

PIE 是一个含有 68 个人的 41 368 幅人脸图片的标准人脸数据集, 每幅图片的像素规模为 32×32. 将 PIE 中的人脸图片划分为 5 个子集, 每个子集对应于不同的姿势, 即 PIE1 (C05, 左姿势), PIE2 (C07, 向上姿势), PIE3 (C09, 向下姿势), PIE4 (C27, 正脸姿势), PIE5 (C29, 右姿势). 每个子集中, 所有人脸均为处于不同灯光、光照和表情条件下获取得到的. 从这 5 个子集中随机抽取两个子集分别作为源域和目标域数据集, 则可构建 20 组跨领域人脸数据集, 即 PIE1 vs PIE2、PIE1 vs PIE3、…、以及 PIE5 vs PIE4. 由人脸不同姿势构建的 20 组数据集中源域和目标域数据集满足不同分布但相关的要求. 此外, 这些数据集中分布差异存在变化较多, 如左姿势子集和右姿势子集领域间的分布差异相比于左姿势子集和正脸姿势子集领域间的分布差异较大.

Office 是视觉跨领域学习常用数据集, 包含 3 个现实汇总的物品数据集: Amazon (由在线交易网站下载), Webcam (由低分辨率网络摄像头拍摄) 和 DSLR (由数码 SLR 高分辨率摄像机拍摄). 该数据集含有 31 个类别共 4 652 幅图片. Caltech 也是目标识别常用的标准数据集, 它含有 256 个类别共 30 607 幅图片. 采用 Gong 等在文献 [8] 中发布的 Office + Caltech 数据集, 其中含有 4 个领域 C (Caltech-256)、A (Amazon)、W (Webcam) 和 D (DSLR). 实验过程中, 随机抽取两个不同领域分别作为源域和目标域数据集, 则可构建 12 个跨领域目标数据集, 即 C→A, C→W, C→D, $\cdots$ 以及 D→W.

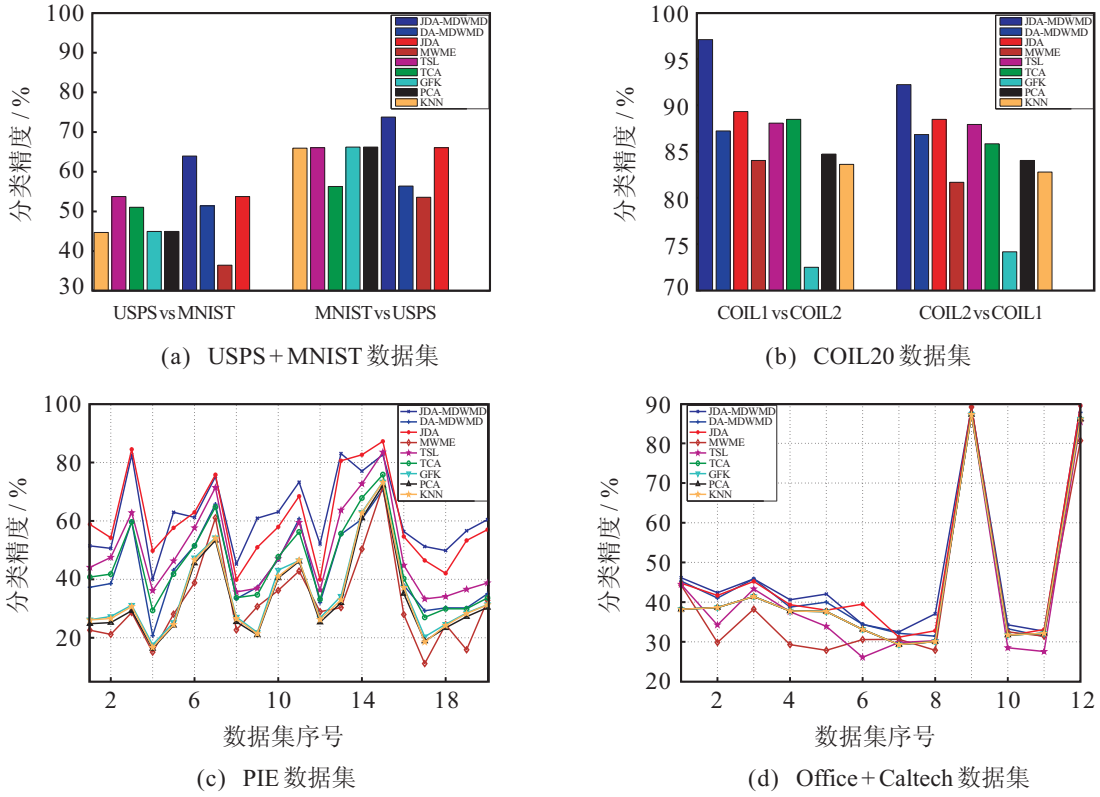
3.2　实验结果及分析

分别采用 KNN、PCA、TCA [10]、JDA[13]、TSL[7]、MWME[11]、GFK[8]、JDA-MDWMD 和基于最大分布加权均值嵌入的分布调整 (DA-MDWMD) 进行实验, 其中 DA-MDWMD 为 JDA-MDWMD 的特例, 即仅考虑边缘概率分布调整的 JDA-MDWMD. 为保证实验的公平性, 统一采用 1 近邻分类器作为基分类器进行监督分类, 每个实验均做 20 次, 取平均值. 实验过程中, 各算法的参数统一设置为: $k = 100$, $\lambda = 0.1$, $T_{\max} = 30$.

图 1 给出了各算法在 36 组跨领域图片数据集上的分类精度曲线, 表 2 中的结果则为各算法在 USPS + MNIST、COIL20、PIE 以及 Office + Caltech 等 4 种不同类型数据集上的分类精度平均值.

由图 1 和表 2 可以看出: 1) 在所有 36 组跨领域图片数据集的分类实验中, 除少数几组数据集外, JDA-MDWMD 的分类精度均高于其他 8 个分类算法. 另外, JDA-MDWMD 在 4 种不同类型数据集上的分类精度也是最高的. 2) 由于 JDA-MDWMD 同时对领域间的边缘分布和条件分布进行调整, 能够获得比只考虑边缘分布调整的 DA-MDWMD 更高的分类精度. 同理, JDA 比其余只考虑边缘分布调整的 GFK、MWME、TCA 以及 TSL 的分类精度要高. 3) 尽管 JDA 同时考虑了领域间的边缘分布差异和条件分布差异, 但它仅考虑了领域间的全局信息差异, 未考虑单个样本对领域间差异度量的贡献差异性, 而 JDA-MDWMD 通过为每个样本引入分布权重解决了这个问题, 从而使得其分类精度高于 JDA.

为了分别考察 JDA-MDWMD 的分类精度与迭代次数、子空间维数和正则化参数之间的关系, 图 2～图 4 给出了 JDA-MDWMD 在 USPS vs MNIST、COIL1 vs COIL2、PIE1 vs PIE2 以及 A→D 数据集上的实验结果. 由图 2～图 4 可以看出: 1) 随着迭代次数的增加, JDA-MDWMD 的分类精度会不断提高并逐渐趋于稳定. 2) JDA-MDWMD 在子空间维数时能取得较好的分类性能. 3) 当 $\lambda \to 0$ 时, JDA-MDWMD 目标函数的优化是不适定的; 当 $\lambda > 1$ 时, 随着 λ 的增大, JDA-MDWMD 的分类精度明显下降.



(a) USPS+MNIST 数据集 (b) COIL20 数据集

(c) PIE 数据集 (d) Office+Caltech 数据集

图 1 36 组跨领域图片数据集上的分类精度曲线

表 2 4 种类型跨领域图片数据集上的分类精度平均值对比 %

数据集	无迁移算法		迁移算法						
	KNN	PCA	TCA	JDA	TSL	MWME	GFK	JDA-MDWMD	DA-MDWMD
USPS+MNIST	55.32	55.59	56.84	45.01	53.67	59.91	63.47	68.87	53.90
COIL20	83.20	84.38	73.34	82.85	87.15	87.99	88.89	94.65	87.02
PIE	34.76	33.85	35.35	32.07	48.18	49.43	60.24	62.02	43.80
Office+Caltech	31.37	39.79	42.95	41.03	43.61	42.37	46.31	47.30	45.77

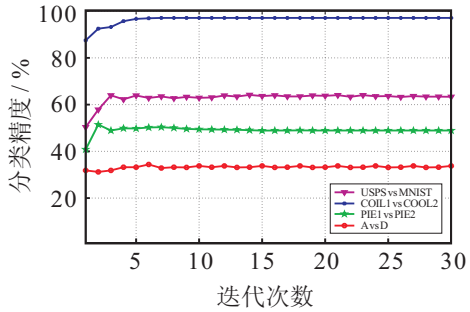


图 2 JDA-MDWMD 分类精度与迭代次数间关系曲线

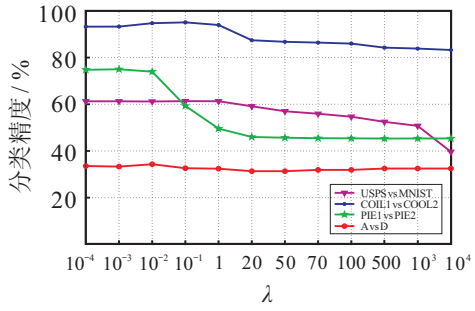


图 4 JDA-MDWMD 分类精度与正则化参数间关系曲线

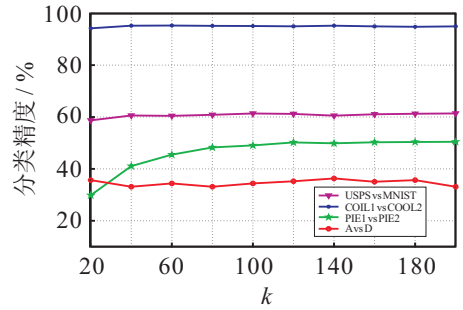


图 3 JDA-MDWMD 分类精度与子空间维数间关系曲线

4 结 论

最大均值差异 (MMD) 使用源域与目标域的总体均值之差来度量两个领域之间的分布差异, 具有计算简单有效、直观且易于理解的优点, 因此, 在领域适应学习中获得了广泛应用. 但是, MMD 仅注重对数据集全局的整体度量, 而忽略了单个样本对全局度量贡献的差异性, 从而影响了度量性能. 为此, 本文采用白化余弦相似性度量为源域和目标域的所有样本设计

相应的分布权重系数, 在MMD的基础上, 提出了一种最大分布加权均值差异度量方法(MDWMD), 使得每个样本的分布差异信息在全局度量中均得以体现. 为提高领域适应学习算法的分类精度, 在MDWMD基础上, 同时对源域和目标域中的数据进行边缘概率分布调整和条件分布调整, 提出了一种基于最大分布加权均值嵌入的领域适应学习算法(JDA-MDWMD). 多个跨领域学习数据集上的实验结果表明, JDA-MDWMD具有较高的分类精度.

参考文献(References)

[1] 许敏, 王士同, 顾鑫. TL-SVM: 一种迁移学习算法[J]. 控制与决策, 2014, 29(1): 141-146.
(Xu M, Wang S T, Gu X. TL-SVM: A transfer learning algorithm[J]. Control and Decision, 2014, 29(1): 141-146.)

[2] 倪彤光, 王士同. 适用于不确定类标签数据学习的迁移支持向量机[J]. 控制与决策, 2014, 29(10): 1751-1757.
(Ni T G, Wang S T. Transfer support vector machine for learning from data with uncertain labels[J]. Control and Decision, 2014, 29(10): 1751-1757.)

[3] 顾鑫, 王士同, 许敏. 领域自适应的最小包含球设计方法[J]. 控制与决策, 2013, 28(2): 177-183.
(Gu X, Wang S T, Xu M. Minimum enclosing ball for domain adaptation[J]. Control and Decision, 2013, 28(2): 177-183.)

[4] Ben-David S, Blitzer J, Crammer K, et al. Analysis of representations for domain adaptation[C]. Proc of Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2007: 137-144.

[5] Pérez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions[C]. Proc of IEEE Int Symposium on Information Theory. Piscataway: IEEE Press, 2008: 1666-1670.

[6] Borgwardt K M, Gretton A, Rasch M J, et al. Integrating structured biological data by kernel maximum mean discrepancy[J]. Bioinformatics, 2006, 22(14): 49-57.

[7] Si S, Tao D, Geng B. Bregman divergence-based regularization for transfer subspace learning[J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(7): 929-942.

[8] Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation[C]. Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 2066-2073.

[9] Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction[C]. Proc of the 23rd AAAI Conf on Artificial Intelligence and the 20th Innovative Applications of Artificial Intelligence Conf. Washington DC: AAAI Press, 2008: 667-682.

[10] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Trans on Neural Networks, 2011, 22(2): 199-210.

[11] 皋军, 黄丽莉. 最大局部加权均值差异嵌入[J]. 电子学报, 2013, 41(8): 1462-1468.
(Gao J, Huang L L. Maximum local weighted mean discrepancy embedding[J]. Acta Electronica Sinica, 2013, 41(8): 1462-1468.)

[12] 皋军, 黄丽莉, 孙长银. 一种基于局部加权均值的领域适应学习框架[J]. 自动化学报, 2013, 39(7): 1037-1052.
(Gao J, Huang L L, Sun C Y. A local weighted mean based domain adaptation learning framework[J]. Acta Automatica Sinica, 2013, 39(7): 1037-1052.)

[13] Long M S, Wang J M, Ding G G, et al. Transfer feature learning with joint distribution adaptation[C]. Proc of IEEE Int Conf on Computer Vision. Piscataway: IEEE Press, 2013: 2200-2207.

[14] Long M S, Wang J M, Ding G G, et al. Adaptation regularization: A general framework for transferlearning[J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(5): 1076-1089.

[15] Nakanishi J, Farrell J A, Schaal S. Composite adaptive control with locally weighted statistical learning[J]. Neural Networks, 2005, 18(1): 71-90.

[16] 皋军, 黄丽莉, 王士同. 基于局部子域的最大间距判别分析[J]. 控制与决策, 2014, 29(5): 827-832.
(Gao J, Huang L L, Wang S T. Local sub-domains based maximum margin criterion[J]. Control and Decision, 2014, 29(5): 827-832.)

[17] Liu C. The Bayes decision rule induced similarity measures[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1086-1090.

(责任编辑：李君玲)