

一种基于局部加权均值的领域适应学习框架

皋军^{1,2,3} 黄丽莉⁴ 孙长银¹

摘要 最大均值差异 (Maximum mean discrepancy, MMD) 作为一种能有效度量源域和目标域分布差异的标准已被成功运用。然而, MMD 作为一种全局度量方法一定程度上反映的是区域之间全局分布和全局结构上的差异。为此, 本文通过引入局部加权均值的方法和理论到 MMD 中, 提出一种具有局部保持能力的投影最大局部加权均值差异 (Projected maximum local weighted mean discrepancy, PMLWD) 度量, 结合传统的学习理论提出基于局部加权均值的领域适应学习框架 (Local weighted mean based domain adaptation learning framework, LDAF), 在 LDAF 框架下, 衍生出两种领域适应学习方法: LDAF_MLC 和 LDAF_SVM。最后, 通过测试人工数据集、高维文本数据集和人脸数据集来表明 LDAF 比其他领域适应学习方法更具优势。

关键词 迁移学习, 领域适应学习, 局部加权均值, 投影最大局部加权均值差异, 基于局部加权均值的领域适应学习框架

引用格式 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. 自动化学报, 2013, 39(7): 1037–1052

DOI 10.3724/SP.J.1004.2013.01037

A Local Weighted Mean Based Domain Adaptation Learning Framework

GAO Jun^{1,2,3} HUANG Li-Li⁴ SUN Chang-Yin¹

Abstract Maximum mean discrepancy (MMD), as a criterion effectively and efficiently measuring the distribution discrepancy between source domains and target ones, has been successfully used. But it is a global measuring algorithm and to some extent only reflects the global distribution discrepancy between domains and the global structural difference. Therefore, we propose projected maximum local weighted mean discrepancy (PMLWD) scheme by with locality preserving ability integrating the theory and method of local weighted mean into the MMD. At the same time, we formulate in theory that the PMLWD is one of generalized algorithms of the MMD. Furthermore, on the basis of the PMLWD and by integrating classical learning theories, we present local weighted mean based domain adaptation learning framework (LDAF). Following the LDAF, we propose local weighted mean based multi-label classification domain adaptation learning algorithm (LDAF_MLC) and local weighted mean based domain adaptation supporting vector machine (LDAF_SVM). At last, tests on artificial data sets, high dimensional text data sets and face data sets show the LDAF methods are superior to other domain adaption ones.

Key words Transfer learning (TL), domain adaptation learning (DAL), local weighted mean (LWM), projected maximum local weighted mean discrepancy (PMLWD), local weighted mean based domain adaptation learning framework (LDAF)

Citation Gao Jun, Huang Li-Li, Sun Chang-Yin. A local weighted mean based domain adaptation learning framework. *Acta Automatica Sinica*, 2013, 39(7): 1037–1052

收稿日期 2012-10-23 录用日期 2013-01-15
Manuscript received October 23, 2012; accepted January 15, 2013

国家自然科学基金 (61272210, 60903100), 江苏省自然科学基金 (BK2011417), 苏州大学江苏省计算机信息处理技术重点实验室开放课题 (KJS1126), 江苏省新型环保重点实验室开放课题 (AE201068), 江苏省高校优秀青年教师和校长境外研修计划资助

Supported by National Natural Science Foundation of China (61272210, 60903100), Jiangsu Provincial Natural Science Foundation (BK2011417), Open Project of Jiangsu Provincial Key Laboratory for Computer Information Processing Technology (KJS1126), Open Project of Key Laboratory for Advanced Technology in Environmental Project of Jiangsu Province (AE201068), and Oversea Education Plan for Young Outstanding Teachers and Presidents of University in Jiangsu Province
本文责任编辑 刘成林

Recommended by Associate Editor LIU Cheng-Lin
1. 东南大学自动化学院 南京 210096 2. 盐城工学院信息工程学院 盐城 224001 3. 苏州大学江苏省计算机信息处理重点实验室 苏州

传统的统计学习技术一般都猜想训练样本和测试样本服从于相同的概率分布, 即是独立同分布的 (Identically and independently distributed, I.I.D)。而当分布发生改变时, 绝大部分传统意义上的智能学习模型必须进行重构。同时在一些具体应用中也会碰到一些非独立同分布 (non-I.I.D) 的识别问题, 比如语际 (Cross-language) 文本挖掘、生物信息学、社会网络研究和多任务学习^[1-2] 等。因此, 为了解

215006 4. 安徽理工大学电气与信息工程学院 淮南 232001
1. School of Automation, Southeast University, Nanjing 210096
2. School of Information Engineering, Yancheng Institute of Technology, Yancheng 224001 3. Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006 4. School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232001

决这一非独立同分布学习问题, 迁移学习 (Transfer learning, TL) 方法被提出^[3]. 迁移学习的目的就是运用在源域 (Source-domain) 上所获得的知识去有效构造适合目标域 (Target-domain) 的统计学习模型^[4], 该类方法并不能简单地理解成是跨问题 (Cross-problems) 学习方法的泛化形式, 而应该理解为在不同任务^[2]、不同视角^[5-6] 和不同区域上的知识的迁移. 从这个意义上讲, 迁移方法明显不同于传统意义上的有监督和无监督方法. 这是因为传统的有监督和无监督方法所处理的训练数据和测试数据具有相同的分布, 而迁移学习则可以处理具有不同分布的源域数据和目标域数据.

作为迁移学习方法的一种特类, 领域适应学习 (Domain adaptation learning, DAL) 方法旨在通过学习分布不同但相关的源域和目标域上的数据来获取相应的统计学习模型. 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去反映源域和目标域之间的分布差异. 近来研究者已经提出了一些适合评判两个区域分布差异的度量, 比如基于熵的 Kullback-Leibler 距离 (Kullback-Leibler distance, KL-distance)^[7]、最大均值差异 (Maximum mean discrepancy, MMD)^[8] 等. KL-distance 度量是一种带参的估计方法, 在度量源域和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, 而 MMD 度量则是一种无参估计标准, 该度量是通过计算源域与目标域之间的均值差来反映两个区域之间的分布差异. 因此, 该度量计算简单有效, 同时直观含义明显. 基于 MMD 度量并结合直推式支持向量机 (Transductive support vector machine, TSVM)^[9]、多标签分类 (Multi-label classification, MLC)^[10]、特征提取等这些传统的统计学习方法, 来重构一些具有一定领域适应学习能力的统计学习方法.

不管是 MMD 度量还是 PMMD 度量, 都是使用源域和目标域或两个区域对应的嵌入子空间的总体均值之差来表示不同分布区域之间的分布差异. 而从统计理论上讲^[11], 一个区域的总体均值或方差作为一个有效的统计特征往往反映的是该区域的总体分布情况, 而且我们还知道, 从几何直观意义上讲, 区域的总体均值一般更能反映团状数据的分布特征, 也就是说能更好地反映那些服从高斯分布的数据的统计特征. 因此, MMD 度量和 PMMD 度量一定程度上反映的是源域与目标域之间存在的总体分布以及全局结构信息上的差异, 而且更适合用来反映具有明显高斯分布特征的区域之间的分布差异. 从而说明上述那些基于 MMD 度量的领域适应学习方法^[4-5, 11-15] 一定程度上属于全局方法, 而忽略了不同区域之间的局部分布和局部信息上的差异.

正是这一原因使得一些领域适应学习方法^[15-16] 采用线性核函数形式的 MMD 度量去计算具有均值相似特征的两个区域分布差异时, 存在一定的不适应性^[17]. 可喜的是, 根据流形学习理论, 呈现非高斯分布或流形分布的数据, 可以被分解成若干局部分块, 而每一个分块可以被看成满足局部高斯分布或可以被视为满足局部欧几里德的^[18]. 根据该理论, 可以把任意分布的源域和目标域按照一定的方法划分成若干个小的局部分块, 然后, 在这些小的局部分块上构造新的度量, 而这一新的度量可以在一定程度上反映两个区域之间内在的局部分布差异, 从而解决 MMD 度量局部学习能力缺失的问题. 同时我们也知道, 普通意义上的均值作为统计特征, 一般认为该区域上的所有样本对反映区域分布和保持区域内在几何机构的能力是相同的, 但实际上区域中的每一个样本所带有的分布信息以及对区域内结构的贡献程度是不一样的.

近来局部加权方法成为统计学习方法^[19-24] 研究的热点问题之一, 不同于在文献 [22-24] 中使用局部加权的方式来表明不同分类器在局部分块上的分类能力的大小, 文献 [19-21] 中则使用的局部加权均值 (Local weighted mean, LWM) 来反映局部分块内不同样本对维持局部结构的贡献程度的差异, 而这一点正是本文必须依据的出发点. 因此, 本文中通过引入 LWM 的技术和理论并结合 MMD 度量的基本原理, 提出一种新颖的能有效度量源域与目标域局部分布差异的标准: 投影最大局部加权均值差异 (Projected maximum local weighted mean discrepancy, PMLWD), 并基于 PMLWD 度量提出一种具有较强局部学习能力的领域适应学习框架: 基于局部加权均值的领域适应学习框架 (Local weighted mean based domain adaptation learning framework, LDAF), 最后, 在 LDAF 框架下分别使用 MLC 和支持向量机 (Support vector machine, SVM) 衍生出两种具体的领域适应学习分类方法: LDAF_MLC 和 LDAF_SVM. 本文提出的方法与其他方法相比具有如下优势:

1) PMLWD 度量不但能有效计算不同分布区域上局部分块之间的局部分布, 还能通过局部分布差异的累积去反映区域之间的总体分布上的差异. 我们也通过理论来说明 PMLWD 度量是 MMD 和 PMMD 度量的泛化形式.

2) 由于 LWM 的概念和方法的引入, 我们定义了一个新颖的最近邻局部分块的概念, 并通过该定义合理地解决了如何计算源域和目标域之间局部分布差异的问题.

3) 基于 PMLWD 度量构造具有一定局部学习能力的领域适应学习框架: LDAF. 基于 LDAF 框

架, 衍生出两种领域适应学习方法. 第一种方法: LDAF_MLC, 该方法既可以作为一种多标号分类器, 同时还可以在在一定程度上实现原始输入空间的低维嵌入; 第二种方法: LDAF_SVM 方法既可以作为一种大间距的领域支持向量机, 同时也可以作为一种局部学习分类器. LDAF 框架作为一种领域适应学习框架也可以用来处理经典的有监督或半监督智能识别问题.

4) 通过在具有明显局部流形特征的人工数据集、高维文本数据集和人脸图像数据集的测试表明 PMLWD 度量和 LDAF 框架具有上述优势.

本文的组织结构如下: 第 1 节介绍相关工作; 第 2 节详细讨论本文的 PMLWD 度量; 第 3 节研究和探讨本文的 LDAF 框架以及基于该框架的 LDAF_MLC 和 LDAF_SVM 方法; 第 4 节使用相应的实验来验证本文提出的方法; 第 5 节总结全文并提出未来的研究方法.

1 相关工作

为了便于讨论本文的 PMLWD 度量和 LDAF 框架, 首先简单介绍一下 MMD 度量和 LWM 的概念.

1.1 最大均值差异

MMD 被现有的领域适应方法广泛运用的主要原因在于该度量计算简单、含义直观. 下面给出 MMD 度量的定义.

定义 1^[5]. 假设分别存在一个满足分布为 P , 源域 $D_s = \{\mathbf{x}_{si}\}_{i=1}^{n_s}$ 和一个满足分布为 Ψ , 目标域 $D_t = \{\mathbf{z}_{tj}\}_{j=1}^{n_t}$, 则在再生核希尔伯特空间 (Reproducing kernel Hilbert space, RKHS) 中源域 D_s 与目标域 D_t 的 MMD 可以用表示为

$$\text{dist}(D_s, D_t) = \|E_{\mathbf{x} \sim P}[\phi(\mathbf{x})] - E_{\mathbf{z} \sim \Psi}[\phi(\mathbf{z})]\|_{\mathcal{H}} \quad (1)$$

其中, $\phi(\cdot)$ 是一个从原始输入空间到高维 Hilbert 空间 \mathcal{H} 上的非线性映射, 且当 $\Psi = P$ 时, 上述区域之间的分布差为 0. 基于 RKHS 空间的一个基本事实, 评价函数可以被写为 $f(\mathbf{x}) = \langle \phi(\mathbf{x}), f \rangle$, 则式 (1) 可以被无偏地改写为如下的经验公式 (详细推导见文献 [8]):

$$\text{dist}^2(D_s, D_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(\mathbf{x}_i) - \frac{1}{n_t} \sum_{i=1}^{n_t} \phi(\mathbf{z}_i) \right\|_{\mathcal{H}}^2 \quad (2)$$

从式 (1) 和 (2) 可以看出 MMD 是在找一个关于数据的统计量 f , 用该统计量期望之差的上界来描述两个分布. 同时也反映了在高维 RKHS 空间, 两个区域之间的分布差异等价于该空间中样本均值之差.

1.2 局部加权均值

标准均值表明区域中每个样本都被平等看待, 但实际上区域中的每一个样本所带有的分布信息以及对内在结构的贡献程度是不一样的, 为了说明这一点, 下面介绍一种被许多局部学习方法^[20-21, 25] 成功使用的均值概念: 局部加权均值.

定义 2^[21]. 假设 $D_{1q} = \{\mathbf{x}_{1q}^i\}_{i=1}^k$ 和 $D'_{1q} = \{\mathbf{y}_{1q}^i\}_{i=1}^k$ 分别表示一个局部区域和该局部区域对应的嵌入子空间, 那么局部区域 D_{1q} 和 D'_{1q} 的 LWM 可以分别写为: $\sum_{i=1}^k \frac{\beta_{qi} \mathbf{x}_{1q}^i}{\sum_{p=1}^k \beta_{qp}}$ 和 $\sum_{i=1}^k \frac{\beta_{qi} \mathbf{y}_{1q}^i}{\sum_{p=1}^k \beta_{qp}}$, 其中, $0 \leq \beta \leq 1$ 是只与局部区域 D_{1q} 上的样本相关的权值参数.

定义 2 中的权值是用来表明在局部区域不同样本对该区域内在的几何结构贡献程度的大小, 而如果令定义中的权值相等, 那么 LWM 就退化为标准均值, 从而说明 LWM 是标准均值概念的泛化形式.

为了便于描述后续内容, 我们用图 1 来表示本文学习框架的总体结构.

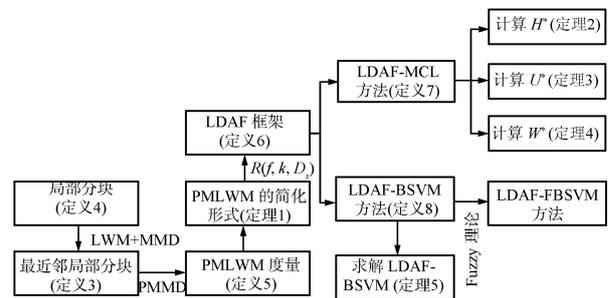


图 1 本文学习框架的总体结构

Fig. 1 Structure of the learning framework in this paper

2 投影最大局部加权均值差异

根据文献 [18] 提出的流形学习理论, 对于任意一个非高斯或流形分布的数据可以被分成若干个分块, 而每一个分块可以被认为是存现局部高斯的, 由此, 在本节将引入局部加权均值的理论和方法到 MMD 方法中, 提出一种能有效度量区域之间局部分布差异的新度量: PMLWD.

定义 3. 假设 $D_1 = \{D_{1q}\}_{q=1}^N$ 和 $D_2 = \{D_{2d}\}_{d=1}^M$ 是具有一定分布差异的两个区域, 其中, D_{1q} 和 D_{2d} 分别表示是上述两个区域上满足局部高斯分布 P_q 和 Ψ_d 的局部分块, 对于 $\forall D_{1q}$, 如果存在一个 $D_{2d'} \subseteq D_2$ 使得下式成立, 那么就称 $D_{2d'}$ 是 D_{1q} 在区域 D_2 上的最近邻局部分块.

$$\text{dist}(D_{1q}, D_{2d'}) = \|E_{\mathbf{x} \sim P_q}[f(\mathbf{x})] - E_{\mathbf{z} \sim \Psi_{d'}}[f(\mathbf{z})]\|_{\mathcal{H}} = \min_{d=1, \dots, M} \|E_{\mathbf{x} \sim P_q}[f(\mathbf{x})] - E_{\mathbf{z} \sim \Psi_d}[f(\mathbf{z})]\|_{\mathcal{H}} \quad (3)$$

需要说明的是, 如果某个数据域是满足非高斯或流形分布的, 那么当将这一数据域划分成若干个局部区域后, 这些局部区域一般不会具有相同的高斯分布. 这是因为如果所有的局部区域都满足同一个高斯分布, 那么将导致原数据域上的所有数据都同时满足这一高斯分布, 从而说明该数据域是满足高斯分布的, 这与前提就相互矛盾了. 这就说明定义 3 中对每一个局部分块定义不同的高斯分布是合理的.

如果使用定义 2 中的 LWM 而不是使用标准均值的形式替换等式 (3) 期望, 同时使用 $\phi(\mathbf{x}) = \mathbf{x}$, 可以得到基于局部加权均值的最近邻局部分块 (Local weighted mean based the nearest local patch, LWNP), 那么 LWNP 对应的估计式为

$$\begin{aligned} \text{dist}_{\text{LWMP}}(D_{1q}, D_{2d'}) = & \left\| \sum_{c_1=1}^{k_1} \frac{\beta_{1q}^{c_1} v_{1q}^{c_1}}{\sum_{p_1=1}^{k_1} \beta_{1q}^{p_1}} - \sum_{c_2=1}^{k_2} \frac{\beta_{2d'}^{c_2} u_{2d'}^{c_2}}{\sum_{p_2=1}^{k_2} \beta_{2d'}^{p_2}} \right\|_2 = \\ & \min_{d=1, \dots, M} \left\| \sum_{c_1=1}^{k_1} \frac{\beta_{1q}^{c_1} v_{1q}^{c_1}}{\sum_{p_1=1}^{k_1} \beta_{1q}^{p_1}} - \right. \\ & \left. \sum_{c_2=1}^{k_2} \frac{\beta_{2d}^{c_2} u_{2d}^{c_2}}{\sum_{p_2=1}^{k_2} \beta_{2d}^{p_2}} \right\|_2 \end{aligned} \quad (4)$$

在处理实际领域适应学习问题时, 如果根据如上定义去使用局部加权均值 LWM, 那么还必须解决两个关键的问题: 1) 如何在特定区域上有效地划分局部分块; 2) 如何根据局部分块中的数据去构造每一个样本对应的局部权值. 幸运的是, 文献 [19, 26] 中的方法尽管不能实现领域适应学习, 但也为我们提供一个思路. 为此, 通过如下定义给出局部分块的划分方法和局部权值构造过程.

定义 4. 假设 $D_s = \{\mathbf{x}_{si}\}_{i=1}^{n_s}$ 和 $D_t = \{\mathbf{z}_{tj}\}_{j=1}^{n_t}$ 分别表示源域和目标域, 对于 $\forall \mathbf{x}_{si} \in D_s$ 和 $\forall \mathbf{z}_{tj} \in D_t$, 分别把 \mathbf{x}_{si} 对应的 k_1 个最近邻样本组成的局部区域 $D_{si} = \{\mathbf{x}_{si}\}_{i=1}^{k_1} \subseteq D_s$ 、 \mathbf{z}_{tj} 对应的 k_2 个最近邻样本组成的局部区域 $D_{tj} = \{\mathbf{z}_{tji}\}_{i=1}^{k_2} \subseteq D_t$ 称为源域、目标域上的局部分块.

同时如果存在一个投影映射 ψ 且令 $D'_{si} = \psi(D_{si})$, $D'_{tj} = \psi(D_{tj})$, 那么 D_{si} , D_{tj} , D'_{si} 和 D'_{tj} 分块的 LWM 可以分别被写为 $\sum_{c_1=1}^{k_1} \frac{\beta_{si}^{(c_1)} \mathbf{x}_{si}^{(c_1)}}{\sum_{p_1=1}^{k_1} \beta_{si}^{p_1}}$, $\sum_{c_2=1}^{k_2} \frac{\beta_{tj}^{(c_2)} \mathbf{z}_{tj}^{(c_2)}}{\sum_{p_2=1}^{k_2} \beta_{tj}^{p_2}}$,

$\sum_{c_1=1}^{k_1} \frac{\beta_{si}^{(c_1)} \psi(\mathbf{x}_{si}^{(c_1)})}{\sum_{p_1=1}^{k_1} \beta_{si}^{p_1}}$ 和 $\sum_{c_2=1}^{k_2} \frac{\beta_{tj}^{(c_2)} \psi(\mathbf{z}_{tj}^{(c_2)})}{\sum_{p_2=1}^{k_2} \beta_{tj}^{p_2}}$. 其中, $\beta_{si}^{(c_1)} = \exp\left(-\frac{\|\mathbf{x}_{si} - \mathbf{x}_{si}^{(c_1)}\|^2}{h_1}\right)$ 和 $\beta_{tj}^{(c_2)} = \exp\left(-\frac{\|\mathbf{z}_{tj} - \mathbf{z}_{tj}^{(c_2)}\|^2}{h_2}\right)$ 分别表示样本 $\mathbf{x}_{si}^{(c_1)}$ 在局部分块 D_{si} 和 $\mathbf{z}_{tj}^{(c_2)}$ 在局部分块 D_{tj} 上的权值, h_1, h_2 是热核函数 $\exp(-\frac{d^2}{h})$ 上的热核参数.

由此, 可以得到本文的 PMLWD 度量.

定义 5. 假设 $D_s = \{D_{si}\}_{i=1}^{n_s}$, $D_t = \{D_{tj}\}_{j=1}^{n_t}$ 分别表示源域和目标域, 其中, $\forall D_{si} \in D_s$, $\forall D_{tj} \in D_t$ 分别是样本 $\mathbf{x}_{si} = (x_{si_1}, \dots, x_{si_n})^T$ 、 $\mathbf{z}_{tj} = (z_{tj_1}, \dots, z_{tj_n})^T$ 所对应的局部分块, 同时令 D'_s, D'_t 分别表示 D_s, D_t 对应的嵌入子空间, 则 PMLWD 度量可以定义为

$$\begin{aligned} \text{dist}_{\text{PMLWD}}^2(D'_s, D'_t) = & \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \text{dist}_{\text{PMLWD}}^2(D'_{si}, D'_{tj}) = \\ & \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} r_{ij} \left\| \sum_{c_1=1}^{k_1} \frac{\beta_{si}^{c_1} \psi(\phi(\mathbf{x}_{si}^{c_1}))}{\sum_{p_1=1}^{k_1} \beta_{si}^{p_1}} - \right. \\ & \left. \sum_{c_2=1}^{k_2} \frac{\beta_{tj}^{c_2} \psi(\phi(\mathbf{z}_{tj}^{c_2}))}{\sum_{p_2=1}^{k_2} \beta_{tj}^{p_2}} \right\|_F^2 \end{aligned} \quad (5)$$

其中, $r_{ij} = \begin{cases} 1, & D_{si} \text{ is the LWNP of } D_{tj} \text{ in the } D_s \text{ or} \\ & D_{tj} \text{ is the LWNP of } D_{si} \text{ in the } D_t \\ 0, & \text{否则} \end{cases}$

为局部分块相关系数, ψ 是低维的嵌入映射, $\|\cdot\|_F$ 表示 F 范数.

根据以上定义, 本文的 PMLWD 通过累积局部分块之间的分布差异来表示两个区域之间的总体分布差异, 从而表明 PMLWD 是通过局部学习来最终实现全局度量学习, 这一理念在许多局部学习方法中得到了成功的运用, 而这一点明显不同于 MMD. 同时对于定义 5, 如果令局部分块中任意一个样本所对应的局部权值相等, 且 $k_1 = n_s$, $k_2 = n_t$, 则本文的 PMLWD 简化为 PMMD, 由此还可以进一步简化为 MMD. 因此, 从这一层面上讲, 本文的 PMLWD 不但没有使得区域之间的分布差异的表达受到任何影响, 同时在一定程度上是对 MMD 和 PMMD 的泛化.

定理 1. 式 (5) 对应的 PMLWD 度量估计式可以简化为

$$\text{dist}_{\text{PMLWD}}^2 = \boldsymbol{\alpha}^T K L K^T \boldsymbol{\alpha} \quad (6)$$

其中, $K = \begin{pmatrix} K_{ss} & K_{st} \\ K_{ts} & K_{tt} \end{pmatrix} \in \mathbf{R}^{(n_s+n_t) \times (n_s+n_t)}$,

K_{ss} 、 K_{tt} 和 K_{st} 是分别定义在源域、目标域和跨域 (Across-domain) 的核矩阵, $\alpha = (\alpha_{s1}, \dots, \alpha_{sn_s}, \alpha_{t1}, \dots, \alpha_{tn_t})^T$ 是由 $n_s + n_t$ 个系数组成的列向量, L 为全局分布差异权矩阵.

证明. 如果令 $X = D_s \cup D_t = \{\mathbf{x}_{s1}, \dots, \mathbf{x}_{sn_s}, \mathbf{z}_{t1}, \dots, \mathbf{z}_{tn_t}\}$, 同时令式 (5) 中的

$\psi(\phi(\mathbf{x})) = \omega^{\phi T} \phi(\mathbf{x})$, 对于任意两个 $D_{si} \subseteq D_s$, $D_{tj} \subseteq D_t$ 局部分块, 将定义在上述局部分块上的权值扩充到整个数据集 X , 则有:

$$\beta_{si} = \left(\underbrace{\frac{\beta_{si}^{(1)}}{\sum_{p_1=1}^{n_s} \beta_{si}^{(p_1)}}, \dots, \frac{\beta_{si}^{(n_s)}}{\sum_{p_1=1}^{n_s} \beta_{si}^{(p_1)}}}_{n_s}, \underbrace{0, \dots, 0}_{n_t} \right) \quad (7)$$

$$\beta_{tj} = \left(\underbrace{0, \dots, 0}_{n_s}, \underbrace{\frac{\beta_{tj}^{(1)}}{\sum_{p_2=1}^{n_t} \beta_{tj}^{(p_2)}}, \dots, \frac{\beta_{tj}^{(n_t)}}{\sum_{p_2=1}^{n_t} \beta_{tj}^{(p_2)}}}_{n_t} \right) \quad (8)$$

则可以用 $\beta_{si}^T \phi(X)^T \omega^\phi$, $\beta_{tj}^T \phi(X)^T \omega^\phi$ 去替换式 (5) 中的 $\sum_{c_1=1}^{k_1} \frac{\beta_{si}^{c_1} (\omega^{\phi T} \phi(\mathbf{x}_{si}^{c_1}))}{\sum_{p_1=1}^{k_1} \beta_{si}^{p_1}}$, $\sum_{c_2=1}^{k_2} \frac{\beta_{tj}^{c_2} (\omega^{\phi T} \phi(\mathbf{x}_{tj}^{c_2}))}{\sum_{p_2=1}^{k_2} \beta_{tj}^{p_2}}$, 则式 (5) 可以转化为

$$\text{dist}_{\text{PMLWD}}^2(D'_s, D'_t) = \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} r_{ij} \|\beta_{si}^T \phi(X)^T \omega^\phi - \beta_{tj}^T \phi(X)^T \omega^\phi\|_F^2 \quad (9)$$

根据 $\|A\|_F^2 = \text{tr}(A^T A)$, 则式 (9) 可以改写为

$$\sum_{i=1}^{n_s} \sum_{j=1}^{n_t} r_{ij} \|\beta_{si}^T \phi(X)^T \omega^\phi - \beta_{tj}^T \phi(X)^T \omega^\phi\|_F^2 = \text{tr}(\omega^{\phi T} \phi(X) \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} r_{ij} (\beta_{si} \beta_{si}^T + \beta_{tj} \beta_{tj}^T - 2\beta_{si} \beta_{tj}^T) \phi(X)^T \omega^\phi) \quad (10)$$

如果令 $L_{ij} = \beta_{si} \beta_{si}^T + \beta_{tj} \beta_{tj}^T - 2\beta_{si} \beta_{tj}^T$ 为局部分块分布差异权值矩阵, $R_{ij} = \text{diag} \left\{ \underbrace{(r_{ij}, \dots, r_{ij})}_{n_s+n_t} \right\}$ 为

局部分块关联系数矩阵, 则式 (10) 可以被写为

$$\sum_{i=1}^{n_s} \sum_{j=1}^{n_t} r_{ij} \|\beta_{si}^T \phi(X)^T \omega^\phi - \beta_{tj}^T \phi(X)^T \omega^\phi\|_F^2 = \text{tr}(\omega^{\phi T} \phi(X) \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} R_{ij} L_{ij} \phi(X)^T \omega^\phi) = \text{tr}(\omega^{\phi T} \phi(X) L \phi(X)^T \omega^\phi) \quad (11)$$

其中, $L = \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} R_{ij} L_{ij}$ 称为源域与目标域全局分布差异权值矩阵.

根据 Represent theory^[27], 则式 (11) 中的非线性投影变换 ω^ϕ 可以表示为 $\phi(X)\alpha$. 同时定义一个基于内积的核函数 $k(x_i, x_j) = (\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$, 则式 (11) 可以改写式 (6) 的形式. \square

根据式 (6), PMLWD 的核心思想主要集中在源域与目标域所对应的全局分布差异权值矩阵 L 上, 为此便于后续方法具有一定的可操作性, 我们单独对权值矩阵 L 构造如下算法.

算法 1. 构造源域与目标域全局分布差异权值矩阵 L

步骤 1. 输入源域 D_s 和目标域 D_t , 且令 $X = D_s \cup D_t$, 输入热核参数 h_1 和 h_2 , 样本 k -NN 参数 k_1 和 k_2 ;

步骤 2. 根据定义 4 构造 D_s 和 D_t 的局部分块, $D_s = \{D_{si}\}_{i=1}^{n_s}$ 和 $D_t = \{D_{tj}\}_{j=1}^{n_t}$;

步骤 3. 对于 $\forall D_{si}$ 和 $\forall D_{tj}$:

1) 分别使用式 (7) 和 (8) 各自计算相对于数据集 X 权值 β_{si} 和 β_{tj} ;

2) 构造局部区域 D_{si} 和 D_{tj} 分布差异权值矩阵 $L_{ij} = \beta_{si} \beta_{si}^T + \beta_{tj} \beta_{tj}^T - 2\beta_{si} \beta_{tj}^T$;

3) 根据式 (4) 分别计算 D_{si} 和 D_{tj} 的局部最近邻局部分块, 如果 D_{si} 是 D_{tj} 的局部最近邻局部分块或 D_{tj} 是 D_{si} 的局部最近邻局部分块, 则定义局部分块关联系数矩阵 $R_{ij} = \text{diag} \left\{ \underbrace{1, \dots, 1}_{n_s+n_t} \right\}$, 否则

$$R_{ij} = \text{diag} \left\{ \underbrace{0, \dots, 0}_{n_s+n_t} \right\};$$

步骤 4. 计算源域和目标域全局分布差异权值矩阵 $L = \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} R_{ij} L_{ij}$.

3 基于局部加权均值的领域学习框架

通过以上分析得知, 无论是 MLCF, 还是大间距投影领域适应学习框架^[4], 以及 DTMKL^[12-13] 等, 一定程度上可以被当作全局迁移学习框架, 从而导致它们不能有效反映不同区域内部的局部结构和局部信息. 由此, 本文在 PMLWD 的基础上提出具有

一定局部学习能力的领域适应学习框架。

定义 6. 假设源域 D_s 和目标域 D_t 对应的嵌入子空间分别为 D'_s 和 D'_t , 则 LDAF 框架可以表示为

$$\arg \min_{f \in \mathcal{H}} J(f) = \arg \min_{f \in \mathcal{H}} (R(f, k, D_s) + \lambda \text{dist}_{\text{PMLWD}}^2(D'_s, D'_t)) \quad (12)$$

其中, 等式左边的 f 为与核函数相关的高维 Hermit 空间 \mathcal{H} 中的决策函数; 等式右边第一项 $R(f, k, D_s)$ 是一个只依赖源域 D_s 中有标号样本的风险函数, 右边第二项为 D'_s 和 D'_t 之间的局部分布距离, 参数 $\lambda > 0$ 是一个平衡参数, 主要是为了提高两个区域的数据分布与损失函数之间的匹配度。

我们知道, 本文的 PMLWD 可以作为 PMMD 度量的泛化形式, 因此, 本文的 LDAF 更具泛化能力. 为此, 在 LDAF 框架下, 我们分别基于 MLC 和 SVM 方法提出两种具有一定泛化能力的领域适应学习方法: LDAF_MLC 和 LDAF_SVM.

3.1 LDAF_MLC

定义 7. 假设源域 $D_s = X_s = \{\mathbf{x}_{si}\}_{i=1}^{n_s} \in \mathbf{R}^{n \times n_s}$ 可以分成 m 个不同类, $Y_s \in \mathbf{R}^{n_s \times m}$ 包含了 n_s 个样本相对于 m 个不同类的类别信息, 目标域为 $D_t = X_t = \{\mathbf{z}_{tj}\}_{j=1}^{n_t} \in \mathbf{R}^{n \times n_t}$, 且 D_s 和 D_t 对应的嵌入子空间分别为 D'_s 和 D'_t , 其中, $\forall \mathbf{x}_{si} \in \mathbf{R}^n$ 如果该样本属于第 p 类, 那么令 $Y_{s_{ip}} = 1$, 否则令 $Y_{s_{ip}} = 0$, 则 LDAF_MLC 的目标函数为

$$\arg \min_{W^T W = I_l} J(f_p, W) = \arg \min_{W^T W = I_l} \left(\sum_{p=1}^m \left(\sum_{i=1}^{n_s} l(\mathbf{x}_{si}, f_p, Y_{s_{ip}}) + \theta \Omega(f_p) \right) + \lambda \text{tr}(W^T X L X^T W) \right) \quad (13)$$

其中, f_p 是第 p 类分类决策函数, $W = [\omega_1, \dots, \omega_l] \in \mathbf{R}^{n \times n_s}$ 是满足等式 (13) 对应的优化问题的正交投影变换矩阵, I_l 为 l 阶单位阵, $\text{tr}(W^T X L X^T W)$ 为 PMLWD 度量的线性形式。

沿用文献 [10] 中的方法, 则可令第 p 类分类决策函数 $f_p(\mathbf{x}) = u_p^T \mathbf{x} = g_p^T \mathbf{x} + h_p^T W^T \mathbf{x}$, 其中, u_p 为决策函数权矢量, $g_p \in \mathbf{R}^n$, $h_p \in \mathbf{R}^l$ 分别为原始输入空间、嵌入子空间中的权矢量, 并将损失函数 l 使用最小二乘误差来替代, 即: $l(\mathbf{x}_{si}, f_p, Y_{il}) = (u_p^T \mathbf{x}_{si} - Y_{il})^2$. 则式 (13) 中的结

构化风险函数可以表示为

$$\sum_{p=1}^m \left(\sum_{i=1}^{n_s} l(\mathbf{x}_{si}, f_p, Y_{s_{ip}}) + \theta \Omega(f_p) \right) = \frac{1}{n_s} \|X_s^T U - Y_s\|_F^2 + \theta \|U - WH\|_F^2 + \eta \|U\|_F^2 \quad (14)$$

其中, $U = [u_1, \dots, u_m]$, $H = [h_1, \dots, h_m]$.

将式 (14) 代入到式 (13) 即可得到 LDAF_MLC 方法目标函数的最终形式:

$$\arg \min_{W^T W = I_l} J(H, U, W) = \arg \min_{W^T W = I_l} \left(\frac{1}{n_s} \|X_s^T U - Y_s\|_F^2 + \theta \|U - WH\|_F^2 + \eta \|U\|_F^2 + \lambda \text{tr}(W^T X L X^T W) \right) \quad (15)$$

为了有效求解式 (15), 本文采用类似于文献 [10, 18] 的方法来求解满足该优化问题的解 H^* , U^* 以及 W^* .

3.1.1 计算 H^*

根据如下定理, 可以证明式 (15) 对应的优化问题解 H^* 可以表示为关于 U 和 W 的形式。

定理 2. 如果 U 和 W 固定, 那么式 (15) 存在局部最优解必要条件为

$$H^* = U^T W \quad (16)$$

证明. 根据拉格朗日乘法法, 可以构造式 (15) 对应的拉格朗日函数, 并令此拉格朗日函数对 H 偏导数为 0, 则有: $W^T(U - WH) = 0$, 并依据条件 $W^T W = I_l$. \square

3.1.2 计算 U^*

根据 $\|A\|_F^2 = \text{tr}(A^T A)$, 则可以将式 (15) 中的 F -范数转化为迹运算的形式, 所以有:

$$\frac{1}{n_s} \|X_s^T U - Y_s\|_F^2 + \theta \|U - WH\|_F^2 + \eta \|U\|_F^2 = \text{tr} \left(\frac{1}{n_s} Y_s^T Y_s - 2U^T \left(\frac{1}{n_s} X_s Y_s + \theta WH \right) + U^T \left((\theta + \eta) I_n + \frac{1}{n_s} X_s X_s^T \right) U + H^T H \right) \quad (17)$$

由于式 (15) 中最后一项并没有显式地包含矩阵 U , 因此, 为了有效求解 U , 只需使用式 (17) 对矩阵 U 求偏导并令求到后的表达式等于 0, 则可以得到满足

式 (15) 优化问题解 U^* 为

$$U^* = \left((\theta + \eta)I_n + \frac{1}{n_s}X_sX_s^T \right)^{-1} \times \left(\frac{1}{n_s}X_sY_s + \theta WH \right) \quad (18)$$

由此可以得到如下定理:

定理 3. 如对于固定的 W 和 H , 式 (15) 对应的优化问题存在局部最优解的必要条件为式 (18) 成立.

然而如果想有效地求解等式 (18), 那么必须降低求解 $((\theta + \eta)I_n + \frac{1}{n_s}X_sX_s^T)^{-1}$ 的时间复杂度 ($O(n^{2.5})$), 为此本文通过引入文献 [28] 提出的 (Sherman-Woodbury-Morrison, SWM) 范式理论, 可以将计算该矩阵的时间复杂度降为 ($O(n_s^{2.5})$), 特别是在处理高维小样本数据集时, 一般有 $n_s \ll n$, 从而使得可以将计算该矩阵所需的时间忽略不计. 根据 SWM 范式, 重写 $((\theta + \eta)I_n + \frac{1}{n_s}X_sX_s^T)^{-1}$ 为 $\frac{1}{(\theta + \eta)}I_n - \frac{1}{(\theta + \eta)}X_s(n_s(\theta + \eta)I_{n_s} + X_s^T X_s)^{-1}X_s^T$, 则式 (18) 可以改写为

$$U^* = \frac{1}{n_s}(I_n - X_s(n_s(\theta + \eta)I_{n_s} + X_s^T X_s)^{-1}X_s^T) \times \left(\frac{1}{n_s}X_sY_s + \theta WH \right) \quad (19)$$

3.1.3 计算嵌入集 W^*

定理 4. 假设固定 H 和 U , 则优化的 W^* 可以通过对如下的目标函数进行奇异值分解 (Singular value decomposition, SVD) 得到.

$$\arg \min_{W^T W = I_l} \lambda \text{tr}(W^T X L X^T W) - \theta \text{tr}(W^T U U^T W)$$

证明. 根据式 (15) 可以看出, 该式中只有两项与变换矩阵 W 直接相关, 也就是说式 (15) 可以精简为: $\arg \min_{W^T W = I_l} (\theta \|U - WH\|_F^2 + \lambda \text{tr}(W^T X L X^T W))$, 且根据式 (16), 有:

$$\begin{aligned} \arg \min_{W^T W = I_l} (\theta \|U - WH\|_F^2 + \lambda \text{tr}(W^T X L X^T W)) = \\ \arg \min_{W^T W = I_l} \theta \text{tr}(U^T U - 2U^T W W^T U + U^T \times \\ W W^T W W^T U) + \lambda \text{tr}(W^T X L X^T W) \quad (20) \end{aligned}$$

由于约束条件 $W^T W = I_l$ 并结合迹运算性质 $\text{tr}(AB) = \text{tr}(BA)$, 则上式可以再次简化为

$$\arg \min_{W^T W = I_l} (\lambda \text{tr}(W^T X L X^T W) - \theta \text{tr}(W^T U U^T W)) \quad (21)$$

对式 (21) 应用拉格朗日乘数法构造相应的拉格朗日函数, 并令该拉格朗日函数对投影矩阵 W 的偏

导数为 0, 则有: $(\lambda X L X^T - \theta U U^T) = \lambda_1 W$, 则我们对 $\lambda X L X^T - \theta U U^T$ 矩阵进行 SVD 分解, 即可构成正交投影矩阵. \square

根据上述定义和定理, 可以得到如下的 LDAF_MLC 算法.

算法 2. LDAF_MLC 算法

步骤 1. 源域 $D_s = \{\mathbf{x}_{si}, y_{si}\}_{i=1}^{n_s}$ 和目标域 $D_t = \{\mathbf{z}_{tj}\}_{j=1}^{n_t}$, 设定参数 λ, θ, η , 最大迭代次数 maxIter 和阈值 ε ;

步骤 2. 使用算法 1 构造源域与目标域全局分布差异权值矩阵 L ;

步骤 3. 给 $U \leftarrow I_{n \times m}, W \leftarrow I_{d \times l}$, 计数器 $p = 0$;

步骤 4. 使用定理 2 求解 H ;

步骤 5. 使用式 (19) 求解 U ;

步骤 6. 使用定理 4 求解 W ;

步骤 7. 当 $|J(H, U, W)^{(p+1)} - J(H, U, W)^p| \leq \varepsilon$ 或 $p > \text{maxIter}$, 退出程序并输出 H^*, U^*, W^* , 否则令 $p = p + 1$, 转到步骤 2.

3.2 LDAF_SVM

近来基于 SVM 与 PMMD 相结合提出了一系列具有领域适应学习能力的非线性支持向量机^[4-5, 12-14]. 不同于上述已有的方法, 本文提出的 LDAF_SVM 则是在 LDAF 框架下通过引入 LS-SVM 分类器^[28] 而构成的, 从而表明 LDAF_SVM 方法更具泛化性. 在本文中提出两种 LDAF_SVM 方法: 1) 二分类方法 LDAF_BSVM; 2) 模糊多分类方法 LDAF_FBSVM.

3.2.1 LDAF_BSVM

定义 8. 如果使用 LS-SVM 的目标函数作为式 (12) 中的结构化风险函数 $R(f, k, D_s)$, 那么本文的 LDAF_BSVM 分类器可以通过最小化如下优化问题得到:

$$\begin{aligned} \min_{\omega^\phi, b} J(\omega^\phi, b) = \min_{\omega^\phi, b} \frac{C}{2} \left(\|\omega^\phi\|^2 + \sum_{i=1}^{n_s} e_{si}^2 \right) + \\ \lambda \text{tr}(\alpha K L K^T \alpha) \\ \text{s.t. } y_{si}(\omega^{\phi T} \phi(\mathbf{x}_{si}) + b) = 1 - e_{si}, \quad i = 1, \dots, n_s \quad (22) \end{aligned}$$

其中, ω^ϕ 是决策超平面法向量, b 是偏移量, e_{si} 为残差.

如果令 $\omega^\phi = \phi(X)\alpha$, 则上述等式中的 $\|\omega^\phi\|^2$ 和 $\omega^\phi \phi(\mathbf{x}_{si})$ 可以分别表示为 $\|\omega\|^2 = \alpha^T K \alpha$, $\omega^T \phi(\mathbf{x}_{si}) \alpha^T \phi(X)^T \phi(\mathbf{x}_{si}) = \alpha^T K_{si}$, 则等式 (22) 可

以被重写为

$$\begin{aligned} \min_{\alpha, b} J(\alpha, b) &= \min_{\alpha, b} \alpha^T \left(\lambda K L K + \frac{C}{2} K \right) \alpha + \\ &\frac{C}{2} \sum_{i=1}^{n_s} e_{si}^2 \\ \text{s.t. } y_{si}(\alpha^T K_{si} + b) &= 1 - e_{si}, \quad i = 1, \dots, n_s \end{aligned} \quad (23)$$

对应式 (23) 需要说明的是, 由于 α 是一个一维矢量, 所以式 (23) 中的 $\text{tr}(\alpha^T K L K \alpha)$ 可以直接表示为 $\alpha^T K L K \alpha$, 使用类似于 LS-SVM 方法的求解过程, 可以得到如下定理.

定理 5. 对于式 (26) 所对应的 LDAF_BFSVM 分类器, 可以通过下列线性系统求解得到:

$$\begin{pmatrix} 0 & -\mathbf{Y}_{n_s}^T \\ \mathbf{Y}_{n_s} & \Pi + C^{-1} I_{n_s} \end{pmatrix} \begin{pmatrix} b \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{1}_{n_s} \end{pmatrix} \quad (24)$$

其中, $\mathbf{1}_{n_s} = \underbrace{(1, \dots, 1)}_{n_s}^T$, $\Pi_{ij} = y_{si} y_{sj} K_{sj}^T (2\lambda K L K + C K)^{-1} K_{si}$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{n_s})^T$ 拉格朗日系数, $\mathbf{Y}_{n_s} = (y_{s1}, \dots, y_{sn_s})^T$.

3.2.2 LDAF_BFSVM

相对于多分类问题, 在 LDAF_BFSVM 方法的基础上并使用 (One-against-one, OAO) 方法, 特别是使用文献 [29–30] 中模糊分类方法的基本原理来构造本文的模糊多分类方法: LDAF-FBSVM. 该方法不但能实现多标号领域适应学习, 同时还能较好地解决不可分区域 (Unclassified region) 样本的分类问题.

对于有 m 个不同类别标号的多分类问题, 如果使用 OAO 方法来实施分类, 那么需要产生 $m(m-1)/2$ 个不同的分类决策函数. 因此, 基于 LDAF_BFSVM 方法, 可以假设第 p 类和第 q 类的分类决策函数为

$$f_{pq}(\mathbf{x}) = \alpha_{pq}^T K_{\mathbf{x}} + b_{pq} \quad (25)$$

其中, \mathbf{x} 为输入样本, $K_{\mathbf{x}} = \phi(X)^T \phi(\mathbf{x})$ 且 $f_{pq}(\mathbf{x}) = -f_{qp}(\mathbf{x})$. 那么输入样本 \mathbf{x} 通过如下公式计算相对于分类函数 $f_{pq}(\mathbf{x})$ 的类别号:

$$f_p(\mathbf{x}) = \sum_{p \neq q, p=1}^m \text{sgn}(f_{pq}(\mathbf{x})) \quad (26)$$

其中, $\text{sgn}(f(\mathbf{x})) = \begin{cases} 1, & f(\mathbf{x}) > 0 \\ 0, & f(\mathbf{x}) \leq 0 \end{cases}$. 因此, 这个输

入样本 \mathbf{x} 最终可以被赋予如下类别号:

$$\arg \max_{p=1, \dots, m} f_p(\mathbf{x}) \quad (27)$$

然而, 根据式 (27), 如果样本 \mathbf{x} 被同时赋予多个类别标号时, 那么该样本被称为不可分样本^[29], 而有这些不可分样本组成的区域就叫不可分类区域.

为了一定程度上解决这一问题, 我们采用类似于 FLS-SVM^[29] 的方法设定一个一维模糊隶属度函数:

$$\pi_{pq}(\mathbf{x}) = \begin{cases} 1, & \text{若 } f_{pq}(\mathbf{x}) \geq 1 \\ f_{pq}(\mathbf{x}), & \text{否则} \end{cases} \quad (28)$$

使用最小化算子, 不可分样本 \mathbf{x} 属于第 p 类的模糊隶属度函数 $\pi_p(\mathbf{x})$ 可以根据下式计算:

$$\pi_{\mathbf{x}} = \arg \max_{q=1, \dots, m} \pi_{pq}(\mathbf{x}) \quad (29)$$

基于式 (32), 该不可分样本 \mathbf{x} 最后被赋予的类别号根据下式得到:

$$\arg \max_{p=1, \dots, m} \pi_p(\mathbf{x}) \quad (30)$$

由此, 可以得到 LDAF_FBSVM 方法.

算法 3. LDAF_FBSVM 算法

步骤 1. 源域 $D_s = \{\mathbf{x}_{si}, y_{si}\}_{i=1}^{n_s}$ 和目标域 $D_t = \{\mathbf{z}_{tj}\}_{j=1}^{n_t}$, 设定参数 λ, C ;

步骤 2. 使用算法 1 构造源域与目标域全局分布差异权值矩阵 L ;

步骤 3. 使用式 (25) 计算 $m(m-1)/2$ 个最优决策分类函数 $f_{pq} = -f_{qp}$ ($p, q = 1, \dots, m$ 且 $p \neq q$);

步骤 4. 对于 $\forall \mathbf{z}_{tj} \in D_t$ 做如下操作:

1) 使用式 (25) 计算 $f_{pq}(\mathbf{z}_{tj})$ 并令 $f_{pq}(\mathbf{z}_{tj}) = -f_{qp}(\mathbf{z}_{tj})$;

2) 使用式 (26) 计算 $f_p(\mathbf{z}_{tj})$;

3) 计算式 (27) 的值, 如果该值唯一, 则该值就作为测试样本 \mathbf{z}_{tj} 的类别标号并赋值给 y_{tj} , 转到步骤 4; 否则转到 4);

4) 使用式 (28) 计算隶属函数 $\pi_{pq}(\mathbf{z}_{tj})$;

5) 使用式 (29) 计算最终模糊隶属度函数 $\pi_p(\mathbf{z}_{tj})$;

6) 计算式 (30) 并将结果赋值给 \mathbf{z}_{tj} 的类别标号 y_{tj} , 转到步骤 4;

步骤 5. 输出目标域 D_t 的类别标号集 $Y_{n_t} = (y_{t1}, \dots, y_{tn_t})^T$.

3.3 算法时间复杂度分析

3.3.1 LDAF_MLC 时间复杂度

在 LDAF_MLC 算法中, 算法的时间复杂度主要集中在步骤 6, 即求解投影矩阵 W . 从理论上讲全

局区域分布差异权值矩阵 L 的秩最大为 $n_s n_t$, 然而, 在实际计算时, 矩阵 W 的秩实际最大为 $n_s + n_t$, 所以矩阵 XLX^T 的秩最大为 $n_s + n_t$. 而且, 矩阵 $U^T U$ 的秩至多为 $\min(n, m)$, 那么矩阵 $\lambda XLX^T - \theta U^T U$ 的秩至多为 $\min(n, m + n_s + n_t)$. 假设处理的数据集具有高维特征, 即 $n \gg n_s + n_t$, 那么在算法步骤 6 过程中求解投影矩阵 W 所对应的时间复杂度为 $O(n^2(m + n_s + n_t))$. 因此, LDAF_MLC 的时间复杂度为 $O(n^2(m + n_s + n_t))$.

3.3.2 LDAF_SVM 时间复杂度

对于本文的 LDAF_BSVM 方法来说, 时间复杂度主要表现在两个方面: 1) 计算 n_s 阶矩阵的逆矩阵; 2) 求解 $(2\lambda K L K + C K)^{-1}$. 我们知道在计算 n_s 阶矩阵的逆矩阵时所对应的时间复杂度为 $O(n_s^{2.5})$, 而在计算 $(2\lambda K L K + C K)^{-1}$ 时时间复杂度则为 $O((n_s + n_t)^{2.5})$. 然而在构造式 (24) 的系数矩阵时, 可以预先计算矩阵 $(2\lambda K L K + C K)^{-1}$, 从而使得 LDAF_BSVM 的时间复杂度还可以看成是 $O(n_s^{2.5})$. 那么本文的 LDAF_FBSVM 的时间复杂度就为 $O(n_s^{2.5} m^2)$, 这和其他 OAO 系列的 SVM 方法的时间复杂度相当.

4 实验

为了说明本文方法的有效性, 我们分别使用三类不同的数据集来进行测试: 1) 人造 Two-moons 数据集; 2) 高维文本 20Newsgroups、Reuter^[31] 数据集; 3) 人脸识别 ORL、Yale 数据集 (<http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html>). 同时结合与其他相似的领域适应学习方法进行比较来说明本文方法的优越性.

通过测试 Two-moons 数据集来说明两个问题: 1) 说明本文的局部领域适应学习框架具有较强的局部学习能力; 2) 说明平衡参数 λ 对本文所提学习框架的学习能力的影响. 测试两个真实的高维文本数据集来说明在处理实际问题时的学习能力, 同时说明近邻参数对迁移学习的影响. 通过测试人脸数据集来说明本文方法处理高维数据时所具有的多分类迁移学习能力. 本测试过程在 Intel Core 2, 2.0 GHz 主频, 2 GB RAM, Vista 系统, Matlab2007 平台上实现.

4.1 测试人造数据集

人造 Two-moon 数据集具有典型的局部流形结构, 因此经常被用来测试相应方法的局部学习能力. 在本阶段为了测试本文提出的线性的 LDAF_MLC 方法和非线性 LDAF_BSVM 方法在实现迁移学习的同时, 还尽可能地保持样本的局部结构, 我们使

用一个包含有 300 样本的 Two-moon 数据集作为源域, 该数据集可以分为正、负 2 类, 每一类包含有 150 个样本, 见图 2(a). 将源域数据分别按逆时针旋转 10 次, 可以得到 10 个分布不同但相关的目标域. 图 2(b)~2(d) 分布表示旋转 30° 、 60° 和 80° 得到的目标域. 从图 2 可以看出, 当旋转的角度越大, 所产生的目标域与源域的分布差异越大, 从而说明相对应的领域适应问题就越复杂. 为此, 我们在该测试过程中分别使用 MLC 方法^[10]、MCDA^[15] 和本文的 LDAF_MLC 方法进行测试比较, 同时使用 SVM、TSVM、LMPROJ 和本文的非线性方法 LDAF_BSVM 进行测试比较.

4.1.1 实验设计

由于该实验过程需要测试两种方法, 因此, 本测试实验可以设计为:

1) 在测试比较线性 LDAF_MLC 方法的过程中, 令本文方法中的 $k_1 = k_2 = [2, 4, 6, 8]$, $h_1 = h_2 = 2^i$ ($i = -10, -8, \dots, 8, 10$), $\lambda = 2^i$ ($i = -7, -5, -3, -2, -1, 0, 1, 3, 5, 7$). θ, η 与 MLC^[10] 方法、文献 [15] 中的 MCDA 方法的参数 α, β 设置相同.

2) 在测试比较本文的非线性 LDAF_BSVM 方法的过程中, 所有的测试比较方法都使用高斯核函数, 其中高斯核函数的带宽 σ 统一设定为训练样本平均范数的平方根^[14]. 同时令 LDAF_BSVM 方法中的 $h_1, h_2, k_1, k_2, \lambda$ 同于 1) 过程中参数的设定范围, C 取值范围设定为 2^i ($i = -2, 0, 2, 4, 6, 8$). 而 TSVM 和 LMPROJ 方法中参数设定等同于文献 [4].

3) 为了反映平衡参数 λ 对本文提出的 LDAF_MLC、LDAF_BSVM 方法局部学习能力的影响, 我们随机抽取一种测试情况来加以说明, 而这一测试过程指的是当上述两种方法在其他参数固定的情况下, 表明平衡参数 λ 对两种方法所对应的分类效果的影响, 具体结果见图 3.

4) 该测试过程使用 10-折交叉验证, 精度使用正确分类的样本数与测试样本数的比值来表示; 运行时间使用机器时间, 单位为秒. 实验结果见表 1 和表 2.

4.1.2 实验分析

根据上述实验结果, 可以得到如下结论:

1) 从表 1 和表 2 可以看出, 随着旋转的角度增大, 上述几种测试方法对应的分类效果随之降低, 这一趋势是符合实际的. 因为随着旋转角度的变化, 目标域的复杂差异层度也发生了变化, 旋转角度增大, 目标域与源域的分布差异就越大, 从而导致了算法的适应性变差. 同时还可以看出, MLC 和 SVM 方

法的分类精度明显低于其他两种方法, 特别是在分布差异较大的情况下, 上述两种方法的分类精度明显低于其他方法, 从而一定程度上说明传统方法不太适合解决迁移学习问题, 还可以从表 1 和表 2 中看出本文两种方法的精度要比其他方法要高, 这也

一定程度上说明本文的局部学习框架提高了方法的学习能力. 同时从上述两个表中的算法运行的时间上来看, 本文的算法同其他方法相比都一定程度上偏高, 因此如何更为有效地提高本文算法的运行效率将是我们以后亟待解决的一个问题.

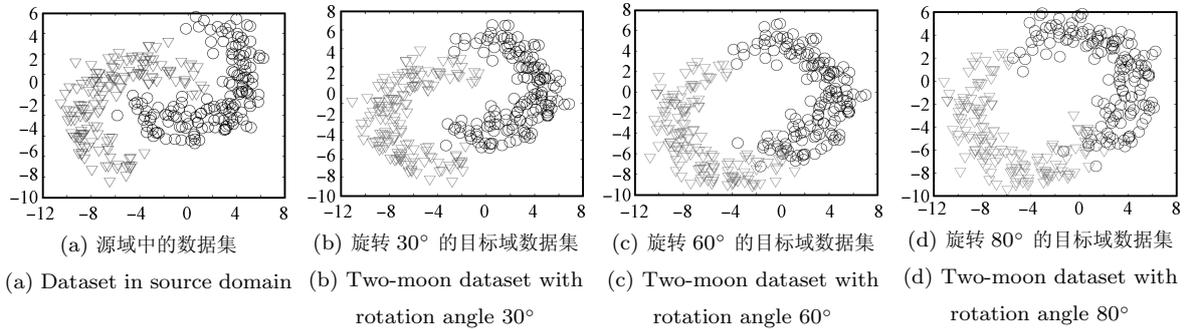


图 2 基于 Two-moon 型数据集构造的源域和目标域

Fig. 2 Two-moon dataset structure based source domain and target domain

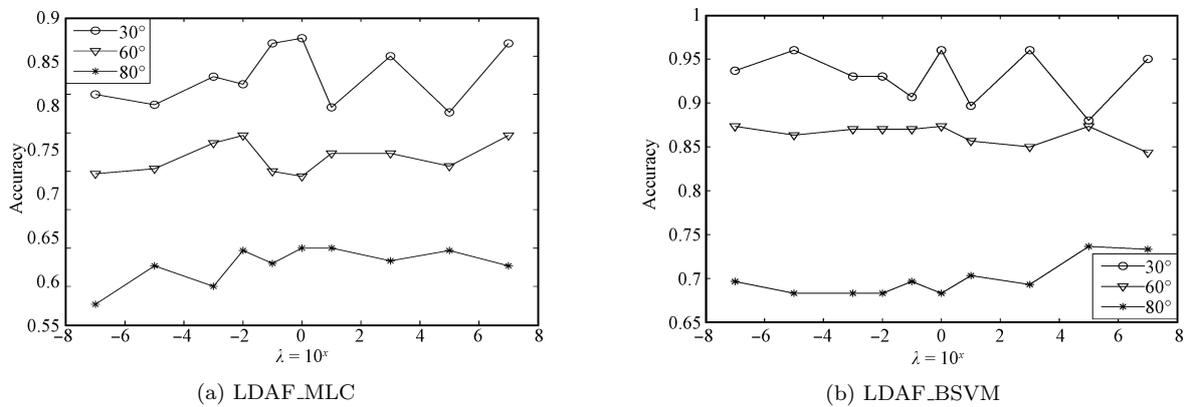


图 3 Two-moon 关联参数 λ 对分类精度的影响

Fig. 3 Influence of two-moon correlation parameter λ on classification accuracy

表 1 3 种线性方法对 10 种不同分布的 Two-moon 数据集的测试比较

Table 1 Comparison of testing results of 3 linear algorithms on Two-moon datasets with 10 different distribution

Target domain (rotation angle)	10°	15°	20°	25°	30°	40°	50°	60°	70°	80°
Algorithm	Accuracy (time)									
MLC	0.8833 (0.6754)	0.8633 (0.5373)	0.8233 (0.3780)	0.8 (0.4199)	0.7833 (0.2962)	0.7133 (0.6218)	0.70 (0.5543)	0.6733 (0.4815)	0.5433 (0.5377)	0.45 (0.4982)
MCDA	0.9267 (0.8795)	0.92 (0.9012)	0.8933 (1.006)	0.87 (0.8997)	0.85 (0.9103)	0.8067 (1.0893)	0.7867 (0.8796)	0.7333 (0.8923)	0.65 (0.9038)	0.5133 (0.8978)
LDAF_MLC	0.9433 (1.7936)	0.9267 (1.6855)	0.9167 (1.7973)	0.9067 (2.0498)	0.8733 (1.9939)	0.8167 (2.0156)	0.7867 (1.8978)	0.7533 (2.1098)	0.6867 (1.9923)	0.6 (2.0142)

表 2 5 种非线性方法对 10 种不同分布的 Two-moon 数据集的测试比较

Table 2 Comparison of testing results of 5 non-linear algorithms on Two-moon datasets with 10 different distributions

Target domain (rotation angle)	10°	15°	20°	25°	30°	40°	50°	60°	70°	80°
Algorithm	Accuracy (time)	Accuracy (time)	Accuracy (time)	Accuracy (time)	Accuracy (time)	Accuracy (time)				
SVM	1 (4.2744)	0.9833 (4.2967)	0.9667 (5.0271)	0.9533 (4.9376)	0.9367 (3.9121)	0.87 (5.3982)	0.7567 (4.9736)	0.7133 (5.9219)	0.6000 (4.6194)	0.5767 (5.0239)
TSVM	1 (6.7869)	0.9967 (7.9210)	0.9900 (7.3347)	0.9800 (6.5728)	0.9567 (6.9826)	0.9233 (7.9935)	0.8833 (8.0177)	0.7833 (7.9645)	0.6967 (7.7855)	0.6133 (7.5761)
LMPORJ	1 (7.8598)	1 (8.1943)	0.9967 (8.8683)	0.9767 (7.9991)	0.96 (8.6366)	0.9567 (9.0326)	0.9167 (8.5387)	0.8467 (7.9485)	0.7567 (9.2847)	0.7033 (8.7543)
LDAF_BSVM	1 (8.9376)	1 (9.5638)	1 (10.8769)	0.9867 (9.8847)	0.9800 (10.0189)	0.9667 (8.9903)	0.9200 (9.6983)	0.8733 (9.9143)	0.8233 (9.4565)	0.7467 (9.8922)

2) 图 3 表明的是本文两种方法在赋予关联参数 λ 不同值时, 对分类精度的影响. 从该图可以看出, 在测试本文两种方法过程中如果关联参数 λ 取值不同, 则所对应的分类精度存在一定差异, 即随着关联参数 λ 从小到大, 识别精度则先由低变高的过程. 但图 3 也反映出并不是参数 λ 的值越大, 精度就越高, 这一点是充分说明了参数 λ 对迁移学习能力的影响. 从另外一个角度上看, 参数 λ 的值的变化远远超出了识别精度的变化, 这说明本文的局部领域适应学习框架具有一定的适应性和稳定性.

4.2 测试高维文本数据集

20Newsgroups、Reuter 作为两个比较典型的文本数据集经常被用来测试领域适应方法^[4-5, 12]的性能. 为了有效测试本文方法的迁移学习效果, 我们依据文献 [5] 给出的子类组合原则, 分别从 20Newsgroups、Reuter 数据集的顶层大类中分别抽取 4 个 (comp、rec、sci、talk) 和 3 个大类 (orgs、people、places) 来构建学习数据集. 由于上述两个数据集非常庞大, 为了便于测试本文方法的所具有的基本的分类能力, 首先, 随机抽取每一个子类 25% 的样本作为测试样本, 同时人为地将上述数据集中的任意两个大类定义为正类和负类, 形成一个只包含两个类别标号的测试子集, 在此测试子集上我们再根据样本之间的相关性形成源域和目标域. 比如, 在使用 20Newsgroups 数据集来测试本文方法时, 我们首先随机抽取每个子类 25% 的样本作为测试样本, 同时我们将 comp、rec 这两个大类组合成一个只有两个类别标号的数据子集: 20Ng-1. 在形成数据子集 20Ng-1 之后, 可以将 comp.graphics、comp.os.ms-

windows.misc 这两个子类组合成源域中正类, 而将 rec.auto、rec.motorcycles 这两个子类组合成源域中的负类. 按照同样的原则可以组合成对应的目标域. 为此, 我们在该测试过程中使用两种有监督方法: MLC、SVM; 一种半监督方法: TSVM; 五种领域适应学习方法: MCDA、CDCS^[31]、LWE^[30]、LMPROJ、DASVM^[32] 与本文的两种局部领域学习方法: LDAF_MLC、LDAF_BSVM 进行测试比较.

4.2.1 实验设计

1) 在测试过程中, MLC、MCDA、SVM、TSVM、LMPROJ 和本文的 LDAF_MLC、LDAF_BSVM 方法的参数设计采用第 4.1.1 节中的参数设定方法. CDCS 方法中的参数设定就采用文献 [31] 中的设定原则, LWE 方法的测试环境也使用文献 [30] 中的测试环境. 同时为了说明本文局部领域适应学习框架中参数 k_1 、 k_2 对框架学习能力的影响, 我们抽取本文两种局部学习方法在测试 20Ng-2、20Ng-2 和 Rut-1 三个数据集过程中某一个具体测试结果来加以说明, 而这一具体测试结果是在固定其他参数的情况下, 并令 $k_1 = k_2$ 且设定参数值范围为 [1 ~ 10, 12, 14, 16, 18, 20] 时得到的.

2) 在测试过程中, 为了一定程度上提高测试效率, 我们对 20Newsgroups、Reuter 数据集分别随机抽取每类样本的 25%、60% 作为训练样本和测试样本 (见表 3). 本测试过程使用 10-折交叉验证. 实验结果见表 4 和图 4.

4.2.2 实验分析

从表 4、图 4 可以得到如下结论:

1) 从表 4 对应的分类精度来看, 两种传统的有

监督方法: MLC、SVM 同其他方法比较来看, 在绝大部分的测试集上表现不佳, 而 TSVM 由于使用了部分目标域中的无标号数据参加训练, 从而一定程度上提高分类精度. 但从总体上讲, 上述方法所对

应的分类能力明显低于其他几种领域适应学习方法. 同时根据该表, 本文线性方法 LDAF_MLC 与同样具有一定局部领域学习能力的 LWE 方法相比识别能力相当, 这在一定程度上表明 LDAF_MLC 方法

表 3 测试用 20Newsgroups、Reuter 数据集
Table 3 Testing datasets-20Newsgroups and Reuter

Datasets	Sub-datasets	Number of sample in source domain		Number of sample in target domain		Number of feature		
		Positive class	Negative class	Positive class	Negative class			
20Newsgroups	20Ng-1 (comp+rec)	489	495	484	497	26 214		
	20Ng-2 (comp+sci)	489	492	484	494			
	20Ng-3 (comp+talk)	484	462	489	493			
	20Ng-4 (rec+sci)	495	492	497	494			
	20Ng-5 (rec+talk)	497	462	495	193			
	20Ng-6 (sci+talk)	495	462	497	493			
	Reuter	Rut-1 (orgs+people)	353	389	372		350	4 771
		Rut-2 (orgs+places)	353	257	350		273	4 415
Rut-3 (people+places)		389	257	372	273	4 562		

表 4 10 种线性方法对 20Newsgroups、Reuter 数据集识别效果的比较
Table 4 Comparison of recognition results of 10 linear algorithms on 20Newsgroups and Reuter

Datasets	20Ng-1	20Ng-2	20Ng-3	20Ng-4	20Ng-5	20Ng-6	Rut-1	Rut-1	Rut-1
Algorithm	Accuracy								
MLC	0.7512	0.7035	0.8416	0.6226	0.6890	0.7522	0.7410	0.7014	0.6636
SVM	0.8287	0.7638	0.9164	0.7013	0.7209	0.80	0.8186	0.7801	0.6822
TSVM	0.8419	0.7587	0.8914	0.7921	0.7427	0.8217	0.8324	0.7913	0.7256
CDCS	0.8563	0.7038	0.9134	0.6539	0.8023	0.7203	0.8726	0.7416	0.6620
MCDA	0.8122	0.7382	0.8577	0.6619	0.7006	0.7536	0.7618	0.7512	0.6698
LWE	0.8828	0.7803	0.9054	0.7028	0.7380	0.7521	0.8269	0.7303	0.6932
LMPROJ	0.8644	0.8537	0.9530	0.8496	0.7980	0.8623	0.8629	0.8026	0.7116
DASVM	0.8695	0.8507	0.9516	0.8658	0.8125	0.8536	0.8726	0.8186	0.8202
LDAF_MLC	0.8154	0.7985	0.8622	0.7013	0.7406	0.7565	0.7881	0.7640	0.6945
LDAF_BSVM	0.8736	0.8885	0.9530	0.8729	0.8110	0.8797	0.8920	0.8016	0.8310

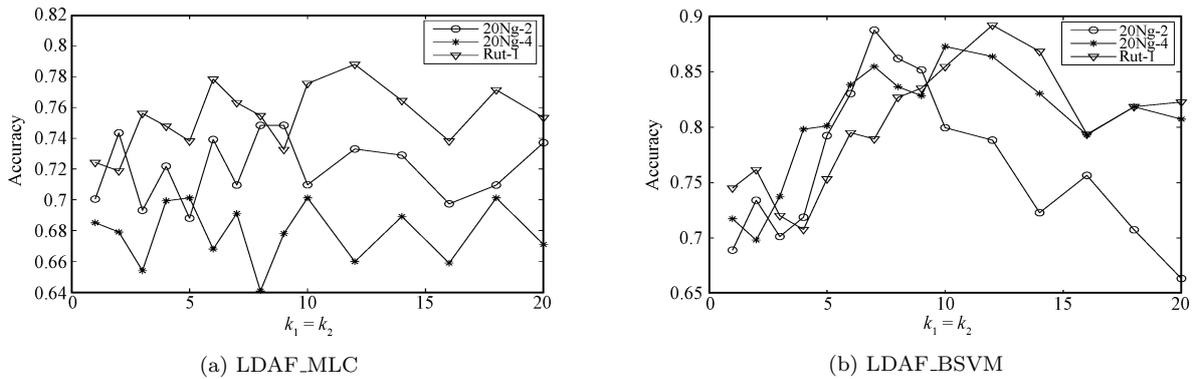


图4 近邻参数 k_1, k_2 对 LDAF_MLCL 和 LDAF_BSVM 方法的局部学习能力的影响

Fig. 4 Influence of nearest neighbor parameters k_1, k_2 on local study ability of LDAF_MLCL and LDAF_BSVM

的有效性,而非线性方法 LDAF_BSVM 方法则具有比LWE方法更强的迁移学习能力,从而一定程度反映本文局部迁移学习框架具有较强的局部学习能力。

2) 从图4可以看出,本文的局部领域学习框架的学习能力一定程度上依赖于参数 k_1, k_2 ,也就是说,本文的学习框架所具有的学习能力会受到所划分得到的局部分块大小的影响,然而这一点也是局部学习方法所共有的问题。同时也可以说明本文的LDAF_MLCL、LDAF_BSVM可以看出MCDA、LMPROJ方法的泛化。

4.3 测试人脸数据集

人脸数据集 ORL 和 Yale 经常被用来测试多分类方法所具有的分类效果。ORL 人脸数据包含 40 个类别的人脸图像,每一类别包含有 10 幅不同人脸表情的图像(见图5);Yale 人脸数据则包含有 15 类不同类别的人脸图像,每一类别包含有 11 幅不同人脸表情的图像(见图6)。然而,为了测试领域适应学习方法,我们有必要对上述两个数据集中数据进行相应的处理。

4.3.1 实验设计

1) 构造源域数据集和目标域数据集。在 ORL 数据集和 Yale 数据集中分别随机选取每类的 5 个样本组成源域数据集,同时将所剩余的数据分布添加满足一定条件的高斯分布、泊松分布和伽马分布的噪声,这样对于上述两个数据集就可以各自生成 3 个与源域数据分布不同的目标域,我们分别将加噪后的目标域数据集记为 ORL_Gauss、ORL_Poiss、ORL_Gam、Yale_Gauss、Yale_Poiss 和 Yale_Gam。其中上述三种噪声数据是分别通过如下的三个 Matlab 命令实现的: normrnd、poissrnd 和 gamrnd。ORL 数据集和 Yale 数据集中的某一类样本原始数据和相应加噪声的数据见

图5和图6。

2) 本测试阶段为了测试本文学习框架的多分类的学习能力,我们继续使用MLC、SVM、TSVM、MCDA、CDCS、LWE、LM PROJ、DASVM与本文提出的LDAF_MLCL、LDAF_BSVM进行测试比较。由于SVM、TSVM、CDCS、LWE、LM PROJ和DASVM方法不能直接处理多分类问题,所以在测试过程中将上述6中方法的多分类问题分解成多个二分类问题,并采用“一对一”策略进行分类测试,同时上述方法的参数设定同于第4.2.1节。对于其他的多分类方法,除了在本文的LDAF_MLCL、LDAF_BSVM方法中令 $k_1 = k_2 = [2, 4, 6, 8]$,而MLC、MCDA方法的参数设定范围也同于第4.2.1节。

3) 本测试过程使用10折交叉验证,分类精度见表5。

4.3.2 实验分析

从以上实验结果可以看出:

1) 根据上述实验结果看出,本文的线性方法LDAF_MLCL同其他几种线性方法相比具有较好的多分类能力,特别是和也具有局部迁移学习能力的LWE方法相比,尽管两种方法的基本构造原理不一样,但从局部学习性能上讲,多分类能力还是比较相当的。

2) 对于非线性方法,由于本文的LDAF_BSVM方法使用了模糊理论,使得在处理传统支持向量机方法很难处理的不可分区域时,更具适应性,这一点可以从表5中得到验证。同时由于本文的模糊多分类支持向量机在考虑最大间隔的基础上,还一定程度上还结合了局部学习的思想,从而使得同非线性的SVM、TSVM、LM PROJ、DASVM这些方法相比更有效性。

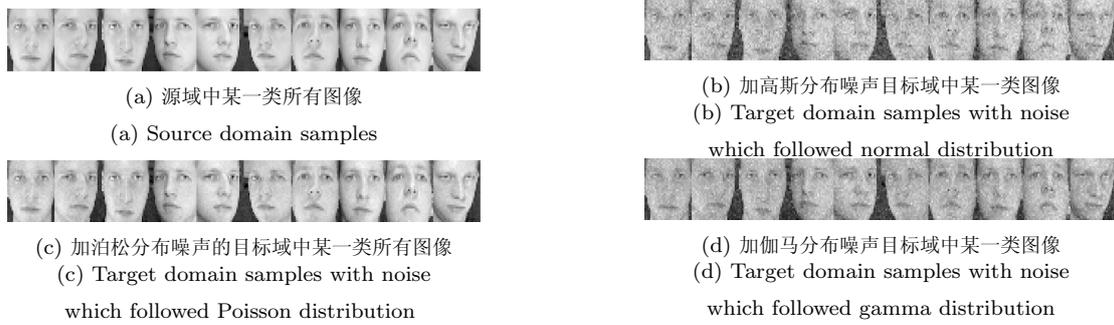


图 5 基于 ORL 数据集构造的源域和目标域样本

Fig. 5 ORL dataset structure based source domain samples and target domain samples

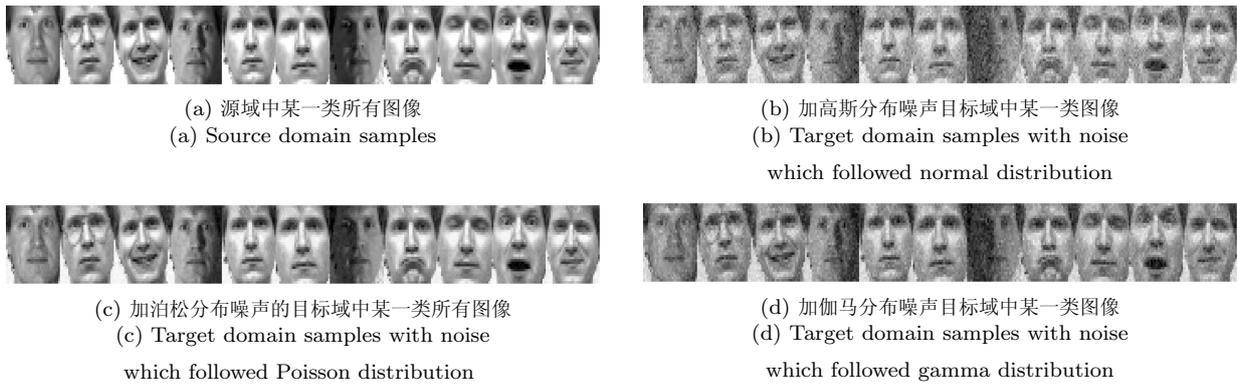


图 6 基于 Yale 数据集构造的源域和目标域样本

Fig. 6 Yale dataset structure based source domain samples and target domain samples

表 5 10 种方法对 ORL 和 Yale 数据集测试效果比较

Table 5 Comparison of testing results of 10 algorithms on ORL and Yale datasets

Datasets	ORL_Gauss	ORL_Poiss	ORL_Gam	Yale_Gauss	Yale_Poiss	Yale_Gam
Algorithm	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
MLC	0.6800	0.7000	0.6950	0.5444	0.6111	0.6111
SVM	0.7100	0.7350	0.7250	0.6111	0.6556	0.7111
TSVM	0.7200	0.7450	0.7600	0.6667	0.6889	0.7222
CDCS	0.7000	0.7550	0.7050	0.5889	0.6445	0.7000
MCDA	0.7000	0.7150	0.7100	0.6111	0.6556	0.7111
LWE	0.7300	0.8400	0.7150	0.6222	0.6222	0.7222
LMPROJ	0.8050	0.8350	0.8650	0.7222	0.7333	0.7333
DASVM	0.8450	0.8750	0.8700	0.7222	0.7333	0.7667
LDAF_MLC	0.7400	0.7200	0.7350	0.5889	0.6333	0.7111
LDAF_FBSVM	0.8500	0.8600	0.8850	0.7333	0.7333	0.8000

5 总结

本文在 MMD 的基础上提出具有局部学习能力的迁移学习度量 PMLWD, 该度量不但能有效地反映源域和目标域之间存在的局部分布差异, 而且一定程度上表明了区域内部存在的局部结构之间的差异. 在 PMLWD 度量的基础上提出一种具有一定局部迁移学习能力统计学习框架: LDAF, 在此框架下衍生出两种具有较强泛化能力的迁移学习方法: LDAF_MLC 和 LDAF_SVM. 最后, 通过测试人工数据和真实数据都表明上述两种方法具有较强的迁移学习能力. 然而为了提高本文算法的运行效率, 从理论上讲可以使用完全覆盖的理论来得到特定区域上的局部分块, 但如何获得合理的完全覆盖可能也是我们面临的一个亟待解决的问题.

References

- Ozawa S, Roy A, Roussinov D. A multitask learning model for online pattern recognition. *IEEE Transactions on Neural Networks*, 2009, **20**(3): 430–445
- Xu Z J, Sun S L. Multi-source Transfer Learning with Multi-view Adaboost [Online], available: <http://www.cst.ecnu.edu.cn/~slsun/pubs/MvTransfer.pdf>, November 7–9, 2006
- Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345–1359
- Quanz B, Huan J. Large margin transductive transfer learning. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). New York, USA: ACM, 2009. 1327–1336
- Zhang D, He J R, Liu Y, Si L, Lawrence R D. Multi-view transfer learning with a large margin approach. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, USA: ACM, 2011. 1208–1216
- Xu Z J, Sun S L. Multi-view Transfer learning with adaboost. In: Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI). New York, USA: IEEE, 2011. 399–402
- Perez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions. In: Proceedings of the 2008 IEEE International Symposium on Information Theory (ISIT) 2008. New York, USA: IEEE, 2008. 1666–1670
- Borgwardt K M, Gretton A, Rasch M J, Kriegel H P, Schölkopf B, Smola A J. Integrating structured biological data by kernel maximum mean discrepancy. In: Proceedings of the 14th International Conference on Intelligent Systems for Molecular Biology (ISMB). California, USA: ISCB, 2006. e49–e57
- Joachims T. Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning (ICML). San Francisco, CA: Morgan Kaufmann Publishers, 1999. 200–209
- Ji S W, Tang L, Yu S P, Ye J P. Extracting shared subspace for multi-label classification. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, USA: ACM, 2008. 381–389
- Vapnik V N. *Statistical Learning Theory*. New York: Wiley, 1998. 88
- Duan L X, Tsang I W, Xu D. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(3): 465–479
- Duan L X, Xu D, Tsang I W. Domain adaptation from multiple sources: a domain-dependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(3): 504–518
- Tao J W, Chung F L, Wang S T. On minimum distribution discrepancy support vector machine for domain adaptation. *Pattern Recognition*, 2012, **45**(11): 3962–3984
- Chen B, Lam W, Tsang I W, Wong T L. Extracting discriminative concepts for domain adaptation in text mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, USA: ACM, 2009. 179–188
- Zhang Z H, Zhou J. Multi-task clustering via domain adaptation. *Pattern Recognition*, 2012, **45**(1): 465–473
- Quanz B, Huan J, Mishra M. Knowledge transfer with low-quality data: a feature extraction issue. *IEEE Transactions on Knowledge and Data Engineering*, 2012, **24**(10): 1789–1802
- Lee J M. *Riemannian Manifolds: An Introduction to Curvature*. Berlin: Springer-Verlag, 2003. 1–4
- Zhao D L, Lin Z C, Xiao R, Tang X O. Linear Laplacian discrimination for feature extraction. In: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA: IEEE, 2007. 1–7
- Wang Y Y, Chen S C, Zhou Z H. New semi-supervised classification method based on modified cluster assumption. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(5): 689–702
- Atkeson C G, Moore A W, Schaal S. Locally weighted learning. *Artificial Intelligence Review*, 1997, **11**(1–5): 11–73
- Woods K, Kegelmeyer W P, Bowyer J. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, **19**(4): 405–410
- Sun S L. Local within-class accuracies for weighting individual outputs in multiple classifier systems. *Pattern Recognition Letters*, 2010, **31**(2): 119–124
- Sun S L, Zhang C S. Subspace ensembles for classification. *Physica A: Statistical Mechanics and its Applications*, 2007, **385**(1): 199–207

- 25 Bregler C, Omohundro S M. Surface learning with applications to lipreading. In: Proceedings of the 1993 Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 1993. 43–50
- 26 Zhang W, Wang X G, Zhao D L, Tang X O. Graph degree linkage: agglomerative clustering on a directed graph. In: Proceedings of the 12th European Conference on Computer Vision (ECCV). Berlin: Springer-Verlag, 2012. 428–441
- 27 Deng Nai-Yang, Tian Ying-Jie. *The New Method of Data Mining — Support Vector Machine*. Beijing: Science Press, 2004. 73–150
(邓乃阳, 田英杰. 数据挖掘中的新方法 — 支持向量机. 北京: 科学出版社, 2004. 73–150)
- 28 Kanamori T, Hido S, Sugiyama M. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 2009, **10**: 1391–1445
- 29 Wang Z, Chen S C. New least squares support vector machines based on matrix patterns. *Neural Processing Letters*, 2007, **26**(1): 41–56
- 30 Gao J, Fan W, Jiang J, Han J W. Knowledge transfer via multiple model local structure mapping. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, USA: ACM, 2008. 283–291
- 31 Ling X, Dai W Y, Xue G R, Yang Q, Yu Y. Spectral domain-transfer learning. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). New York, USA: ACM, 2008. 488–496
- 32 Bruzzone L, Marconcini M. Domain adaptation problems: a DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(5): 770–787



皋 军 东南大学自动化学院博士后, 盐城工学院信息工程学院副教授. 主要研究方向为人工智能, 模式识别, 数据挖掘, 模糊系统. E-mail: gjxllin@yahoo.cn

(**GAO Jun** Postdoctor at the School of Automation, Southeast University and associate professor at the School of Information Engineering, Yancheng Institute of Technology. His research interest covers artificial intelligence, pattern recognition, data mining, and fuzzy system.)



黄丽莉 安徽理工大学电气与信息工程学院硕士研究生. 主要研究方向为人工智能, 模式识别.

E-mail: llhuang135@163.com

(**HUANG Li-Li** Master student at the School of Electrical & Information Engineering, Anhui University of Science & Technology. Her research interest covers artificial intelligence and pattern recognition.)



孙长银 东南大学自动化学院教授. 主要研究方向为人工智能, 神经网络, 智能控制理论与方法, 模式识别. 本文通信作者. E-mail: cysun@seu.edu.cn

(**SUN Chang-Yin** Professor at the School of Automation, Southeast University. His research interest covers artificial intelligence, neural networks, theory and design of intelligent control systems, and pattern recognition. Corresponding author of this paper.)