核分布一致局部领域适应学习

陶剑文 [1, 2] 王士同 [1]

摘 要 针对领域适应学习 (Domain adaptation learning, DAL) 问题, 提出一种核分布一致局部领域适应学习机 (Kernel distribution consistency based local domain adaptation classifier, KDC-LDAC), 在某个通用再生核 Hilbert 空间 (Universally reproduced kernel Hilbert space, URKHS), 基于结构风险最小化模型, KDC-LDAC 首先学习一个核分布一致正则化支持向量机 (Support vector machine, SVM), 对目标数据进行初始划分; 然后, 基于核局部学习思想, 对目标数据类别信息进行局部回归重构; 最后, 利用学习获得的类别信息, 在目标领域训练学习一个适于目标判别的分类器. 人造和实际数据集实验结果显示, 所提方法具有优化或可比较的领域适应学习性能.

关键词 领域适应学习, 核分布一致, 局部学习, 模式分类, 最大平均差

引用格式 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

DOI 10.3724/SP.J.1004.2013.01295

Kernel Distribution Consistency Based Local Domain Adaptation Learning

TAO Jian-Wen[1, 2] WANG Shi-Tong[1]

Abstract In allusion to domain adaptation learning (DAL) problems, this paper proposes a novel so-called kernel distribution consistency based local domain adaptation classifier (KDC-LDAC). Firstly, in some universally reproduced kernel Hilbert space (URKHS), the KDC-LDAC trains a kernel distribution consistency regularized domain adaptation support vector machine (SVM) based on the structure risk minimization model, which extends the formulation of classical SVMs to the domain adaptation learning schema. And secondly, according to the idea of local learning, the proposed method predicts the label of each data point in target domain based on its neighbors and their labels in the URKHS. The last but not least, the KDC-LDACs learning a discriminant function to classify the unseen data in target domain with training data well predicted in the local learning procedure. Experimental results on artificial and real world problems show the advantages or comparable effectiveness of the proposed approach compared to related approaches.

Key words Domain adaptation learning (DAL), kernel distribution consistency (KDC), local learning, pattern classification, maximum mean discrepancy (MMD)

Citation Tao Jian-Wen, Wang Shi-Tong. Kernel distribution consistency based local domain adaptation learning. *Acta Automatica Sinica*, 2013, **39**(8): 1295−1309

目前的机器学习算法在 Web 挖掘研究中存在着一个关键的问题[1−2], 即在一些新出现的 Web 应用领域中, 大量的标签化训练数据非常难得, 这需要对每个领域都标注大量训练数据, 这将会耗费大量的人力与物力, 而标签数据的缺少会严重影响学习性能, 从而使得很多与学习相关研究与应用无法开展. 相反, 即使我们有了大量的、在不同分布下的训练数据, 针对新的兴趣领域, 完全丢弃这些训练数据而重新构建训练数据也是非常浪费的. 另外, 传统的机器学习假设训练数据与测试数据独立且服从相同的分布 (Identically and independently distributed, IID), 而非 IID 数据在许多应用中自然存在[1−4], 这些应用领域存在的主要问题是精确标注的任务特定数据较少, 而任务相关的数据却大量存在, 且通常可能发生训练数据过期的情况, 这又需要我们去重新标注大量的训练数据以满足训练的需要.

针对上述问题, 迁移学习 (Transfer learning, TL)[5] 得以提出, 旨在将从一个环境中学到的知识用来帮助新环境中的学习任务. 作为迁移学习的一个特例, 领域适应学习 (Domain adaptation learning, DAL)[6] 旨在利用源领域 (Source domain, SD) 中的训练数据来解决目标领域 (Target domain, TD) 中的学习问题, SD 和 TD 中的数据分布可以相同或不同. 近年来, 在机器学习、数据挖掘、多任务学习等应用领域中, DAL 吸引了越来越多研究

收稿日期 2011-10-28 录用日期 2012-03-19
Manuscript received October 28, 2011; accepted March 19, 2012
国家自然科学基金 (60975027, 60903100), 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 (LY13F020011) 资助
Supported by National Natural Science Foundation of China (60975027, 60903100), Humanities and Social Sciences Research Fund of Ministry of Education (13YJAZH084), and Natural Science Foundation of Zhejiang Province (LY13F020011)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 江南大学数字媒体学院 无锡 214122 2. 浙江工商职业技术学院电子与信息工程学院 宁波 315012
1. School of Digital Media, Southern Yangtze University, Wuxi 214122 2. School of Electronics and Information Engineering, Zhejiang Business Technology Institute, Ningbo 315012

者的关注和研究[1−2, 6−12]. 在 DAL 中的一个主要计算问题是如何减小 SD 和 TD 中数据的分布差距, 其关键在于确保有效分类性能的情况下, 如何通过给定的目标函数来实现不同分布之间的距离度量. Ben-David 等[8] 分析指出, 最好性能的超平面分类器应提供一种能较好度量不同数据表示之间分布距离的方法. 同样, Sriperumbudur 等[13] 和 Gretton 等[14] 也分析指出, 两个不同分布之间的距离可通过某种特定的函数类来进行度量, 且在再生核 Hilbert 空间 (Reproduced kernel Hilbert space, RKHS) 中, 能明显简化这种分布距离度量的计算复杂度, 基于此, Gretton 等[14] 提出了一种称为最大均值差 (Maximum mean discrepancy, MMD) 的分布距离度量方法. 概率分布 RKHS 嵌入之间的距离度量 (MMD) 方法相较于传统的方法具有计算简单、收敛快和有限样本估计低偏差等优点. 近来, Quanz 等[7] 基于正则风险最小化和 MMD 方法的思想, 提出一种基于特征空间的大间隔直推式迁移学习方法 (Large margin projection, LMPROJ), 其核心思想在于: 基于经验风险正则化分类框架, 通过寻求一个特征变换使得训练数据和测试数据之间的分布距离最小化, 从而实现迁移学习.

近来, 文献 [6] 针对领域适应学习问题, 提出一种基于迭代思想的领域适应支持向量机 (Domain adaptation support vector machine, DASVM) 学习框架, DASVM 的学习过程主要包括三步: 1) 按照经典的支持向量机 (Support vector machine, SVM)[15] 学习模型, 利用源领域数据来初始化学习一个 SVM; 2) 对初始 SVM 进行迭代学习, 核心思想是在源领域训练数据集中迭代增加目标领域数据集, 以更新训练集, 同时逐步消除源领域数据, 直至最后训练集中只包含目标领域数据集; 3) 迭代收敛, 通过目标领域数据集来学习一个判别函数以对目标数据进行决策判别. 关于 DASVM 的详细描述见文献 [6].

虽然 DASVM 在某些特定数据集的学习下, 取得了较好的适应学习性能, 但是, 通过对文献 [6] 的研究分析发现, DASVM 的学习性能在某种程度上依赖于初始 SVM 的学习能力, 而 DASVM 在第 1) 步中仅采取传统的 SVM 框架模型来学习一个领域适应判别函数, 即 DASVM 的初始 SVM 学习在一定程度上仅考虑了源领域数据, 虽然可能会在源领域取得较好的初始判别界面, 但是当领域间数据分布距离较大时, 将会导致目标领域数据判别偏向单一类方向. 两个领域间数据分布的差异是导致传统 SVM 不能适用于领域适应学习的关键原因所在, 从这层意义上来说, DASVM 一开始就未能充分考虑领域间数据分布的间隔差异, 从而一定程度上可能会限制 DASVM 在某些具体应用上的学习性能. 为此, 本文将从充分考虑数据分布距离的角度, 基于局部学习[16−17] 思想, 提出一种新颖的核分布一致局部领域适应学习机 (Kernel distribution consistency based local domain adaptation classifier, KDC-LDAC). 与 DASVM 的学习过程相似, 在某个通用再生核 Hilbert 空间 (Universal reproducing kernel Hilbert space, URKHS)[18−20], KDC-LDAC 的学习过程也包括三步: 1) 核分布一致领域适应学习, 即通过寻求某个特征变换 ϕ, 使得在确保训练数据的最大分割的同时, 实现领域间分布距离充分最小化, 从而初步实现领域适应学习; 2) 局部领域适应学习, 即在核空间, 基于局部学习的思想, 分别重构出目标领域数据标签; 3) 利用第 2) 步学习的目标领域数据标签信息, 训练学习一个用于目标领域的判别决策函数. 本文方法 KDC-LDAC 通过核映射, 在某个 URKHS 中, 充分考虑了领域间分布的间隔一致性, 使得领域适应学习性能明显优于 DASVM 方法.

1 KDC-LDAC 方法

1.1 相关概念与问题描述

一个领域 (Domain) D 由特征空间 χ 和边沿概率分布 $P(X)$ 两个部分组成, 即 $D = \{\chi, P(X)\}$, 其中, $X = \{\boldsymbol{x}_i\}_{i=1}^{N} \in \chi$. 对于给定的领域 $D = \{\chi, P(X)\}$, 一个学习任务 (Task) T 由两个部分组成, 即标签空间 Y 和目标预测函数 $f(\cdot)$, 即 $T = \{Y, f(\cdot)\}$, 其中, $f(\cdot)$ 可通过训练数据对集 $\{x_i, y_i\}$ 学习获得, 其中, $\boldsymbol{x}_i \in X$, $y_i \in Y$. 函数 $f(\cdot)$ 能用以预测测试数据 $\boldsymbol{x}$ 的相应标签 $f(\boldsymbol{x})$. 从概率角度来说, $f(\boldsymbol{x}) = P(y|\boldsymbol{x})$.

给定一个源领域 D_s 及其学习任务 T_s 和一个目标领域 D_t 及其学习任务 T_t, 领域适应学习旨在利用从 D_s 和 T_s 中所学习到的知识来帮助增强 D_t 和 T_t 中目标预测函数 $f_t(\cdot)$ 的学习性能, 其中, $D_s \neq D_t$ 或 $T_s \neq T_t$. 当 $D_s = D_t$ 且 $T_s = T_t$ 时, 领域适应学习退化为传统的机器学习问题.

核方法 (Kernel tricks) 是一种广泛使用的构建非线性算法的有效途径[18], 本文利用核技巧将两个概率分布嵌入到一个 URKHS, 从而获得一种处理概率分布高阶统计特征的新方法[19−20]. 设 H 为函数族 F 的完备内积空间 (即 Hilbert 空间), 且对于 $f \in F$ 有 $f : X \to \mathbf{R}$, 其中, X 为一个非空紧凑集, 如果对于所有 $\boldsymbol{x} \in X$, 线性点函数映射 $f \to f(\boldsymbol{x})$ 存在且连续, 则 H 可称为一个再生核 Hilbert 空间 (RKHS), 在此条件下, $f(\boldsymbol{x})$ 可表示为一个内积: $f(\boldsymbol{x}) = \langle f, \phi(\boldsymbol{x})\rangle_H$, 其中, $\phi : X \to H$ 为从 $\boldsymbol{x}$ 到 H 的特征空间映射, 且两个特征映射的内积称为核 (Kernel), $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}')\rangle_H$. 基于核技术, 引出如下概念:

定义 1[19] **(概率分布的 RKHS 嵌入距离度量).** 设定义于拓扑空间 M 上的所有 Borel 概率度量集 Θ 与函数集的 RKHS(H,k) 及其再生核 k, H 代表 RKHS 空间, 任意 $P\in\Theta$ 表示为 $Pk=\int_M k(\cdot,\boldsymbol{x})\mathrm{d}P(\boldsymbol{x})$. 如果 k 是可测且有界的, 则 P 在 H 中的嵌入定义为 $Pk\in H$. 两个概率度量 P, $Q\in\Theta$ 在 RKHS 中的嵌入距离定义为

$$\gamma_k(P,Q)=||Pk-Qk||_H \tag{1}$$

概率分布的 RKHS 嵌入距离度量方法相较于传统的方法具有计算简单、收敛快和有限样本估计低偏差等优点[20]. 如果映射 $P\mapsto Pk$ 是内射 (Injective) 的, 则称 k 为特征核 (Characteristic kernel, CK)[13, 19], 此时, 当且仅当 $P=Q$ 时, $\gamma_k(P,Q)=0$, 即 γ_k 为 Θ 上的距离度量. 当 k 不是 CK 时, 概率分布在 RKHS 中的嵌入不可分辨, 从而会导致嵌入距离度量失败. 因此, k 是否为 CK 是 RKHS 嵌入距离度量成功的关键条件. 幸运地是, 目前许多流行的核函数 (如多项式核函数, Gaussian 核函数等) 都是通用的 CK 型[13, 19]; 另外, 值得说明的是, 文献 [19] 从理论上分析指出, 高斯型核函数簇为概率分布距离度量的一致性估计提供了一个有效的 RKHS 嵌入空间, 详细论证见文献 [13, 19]. 故此, 本文以下所有核函数均采用高斯型核函数 $k_\sigma(\boldsymbol{x},\boldsymbol{y})=\exp(-\frac{1}{2\sigma^2}||\boldsymbol{x}-\boldsymbol{y}||^2)$, 其中, σ 为核带宽.

根据定义 1, 特别地, 对于领域适应学习问题, 引出如下定义:

定义 2 (领域分布的 RKHS 嵌入距离度量). 设 $p,q\in\Theta$, 线性函数 f: $f(\boldsymbol{x})=\langle\boldsymbol{w},\phi(\boldsymbol{x})\rangle$, $\boldsymbol{w}$ 为投影向量, 则领域间分布的 RKHS 嵌入均值距离度量定义为

$$\begin{aligned}&\gamma_{KM}(p,q)^2=\\&\left\|\int_{X_s}f_{\boldsymbol{x}\sim p}(\boldsymbol{x})\mathrm{d}p-\int_{X_t}f_{\boldsymbol{z}\sim q}(\boldsymbol{z})\mathrm{d}q\right\|^2:=\\&\gamma_{KM}(p_n,q_m)^2\end{aligned} \tag{2}$$

其中, $\boldsymbol{x}\in X_s,\boldsymbol{z}\in X_t$,$\gamma_{KM}(p_n,q_m)$ 为 $\gamma_{KM}(p,q)$ 的经验估计, 定义为

$$\begin{aligned}&\gamma_{KM}(p_n,q_m)^2=\\&\left\|\int_{X_s}\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})\mathrm{d}p_n-\int_{X_t}\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{z})\mathrm{d}q_m\right\|^2:=\\&\boldsymbol{w}^{\mathrm{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i)-\frac{1}{m}\sum_{j=1}^{m}\phi(\boldsymbol{z}_j)\right\|^2\boldsymbol{w}\end{aligned} \tag{3}$$

对于采样自 $X\times Y$ 的服从某种 (未知) 概率分布 $P_s(\boldsymbol{x}_s,y_s)$ 的 n 个训练样本 $D_s=\{(\boldsymbol{x}_{s1},y_{s1}),\cdots,(\boldsymbol{x}_{sn},y_{sn})\}$ 和采样自 X 的服从某种 (未知) 概率分布 $P_t(\boldsymbol{x}_t,y_t)$ 的 m 个测试样本 $D_t=\{\boldsymbol{x}_{t1},\cdots,\boldsymbol{x}_{tm}\}$, 其中, 对应的输出 y_t 未知或隐蔽而需要学习预测, 假设两个数据集内数据的采样均分别服从 IID, 传统的大间隔学习机假设 $P_s(\boldsymbol{x}_s,y_s)=P_t(\boldsymbol{x}_t,y_t)$, 而领域适应学习则在 $P_s(\boldsymbol{x}_s,y_s)\neq P_t(\boldsymbol{x}_t,y_t)$ 的假设下, 学习一个能精确预测无标签测试数据的输出的分类机. 若无特别说明, 下文以 $X_s,X_t\subset X$ 分别代表源领域和目标领域数据集, $Y_s,Y_t\subset Y$ 分别代表源领域和目标领域类标签集, δ_i 代表数据点 x_i 的 k 近邻集, $N_i=|\delta_i|$ 指 δ_i 的集合基数. 本文研究基于如下几点假设:

1) 所研究对象仅包含一个源领域 D_s 和一个目标领域 D_t[1−2, 6−8], 其中, 源领域数据 $D_s=\{(\boldsymbol{x}_{s1},y_{s1}),\cdots,(\boldsymbol{x}_{sn},y_{sn})\}$, $\boldsymbol{x}_{si}\in X_s$ 为源领域 D_s 中的实例数据, $y_{si}\in Y_s$ 为相应的类标签, 源领域 D_s 的实例数据集、特征空间和概率分布分别为 X_s、χ_s 和 $P_s(X)$; 目标领域数据 $D_t=\{\boldsymbol{x}_{t1},y_{t1}),\cdots,(\boldsymbol{x}_{tm},y_{tm})\}$, $\boldsymbol{x}_{ti}\in X_t$ 为目标领域 D_t 中的实例数据, 目标领域 D_t 的实例数据集、特征空间和概率分布分别为 X_t、χ_t 和 $P_t(X)$, 其中对应的输出 y_{ti} 为未知或隐蔽而需要学习预测, $0\leq m\ll n$.

2) 两个不同领域 (源领域和目标领域) 共享同一个特征空间, 而它们的边沿概率分布不同, 即 $P_s(X)\neq P_t(X)$, 且 $P_s(y_s|\boldsymbol{x}_s)\neq P_t(y_t|\boldsymbol{x}_t)$.

3) 源领域和目标领域数据均带有标签信息, 但是与目标领域相关的先验信息仅用于学习方法分类性能的客观量化评价.

给定一个源领域 D_s 及其学习任务 T_s 和一个目标领域 D_t 及其学习任务 T_t, 设两个数据集 X_s 和 X_t 的分布 $P_s(X)$ 和 $P_t(X)$ (简写为 P_s 和 P_t) 分别满足 IID 要求, 则 KDC-LDAC 算法的实现过程主要包括三步: 1) 初始化阶段: 核分布一致正则化支持向量机学习; 2) 局部回归学习阶段: 基于核稀疏表示和核局部学习思想, 重构目标领域数据标签信息; 3) 目标领域判别函数学习阶段: 利用带标签信息的目标领域数据学习一个判别决策函数. 图 1 显示了 KDC-LDAC 三段式领域适应学习模型结构.

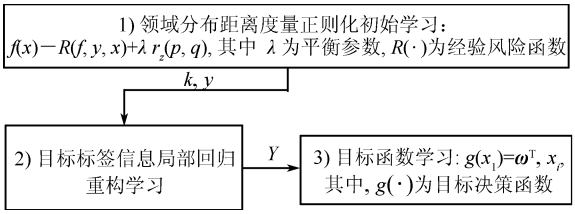


图 1　KDC-LDAC 三段式领域适应学习模型

Fig. 1　Three-stage domain adaptation learning schema in KDC-LDAC

图 1 中, k, y 分别代表第 1) 步完成后获得的核函数和初始划分类标签矩阵, Y 指第 2) 步完成后所习得的两个领域样本的标签.

1.2 核分布一致正则化领域适应初始化

对于领域适应学习问题, 在通过高斯核映射的通用再生核 Hilbert 特征空间中, 本文初始化阶段旨在寻求一个线性特征变换 $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})$, 其中 $\boldsymbol{w}$ 为 URKHS 中待求的线性投影向量, 使得最小化领域间分布距离的同时, 使得分类决策函数的经验风险最小化. 其核心思想为: 基于统计模式识别的大间隔方法思想, 在高斯型再生核 Hilbert 空间 (RKHS), 通过同时正则化最小化训练数据和测试数据之间的分布距离和学习经验风险, 学习一个用于领域适应学习的大间隔核分类机. 本文确保在源领域学习性能最大化的前提下, 力求最小化源领域和目标领域的数据分布差, 从而实现从源领域学习到目标领域学习的最大可能的迁移, 即 KDC-LDAC 的初始化目标函数描述为

$$\begin{aligned} f = & \min_{\boldsymbol{w}\in H_K} C_1\sum_{i=1}^{n}\xi_i + (1-\lambda)||\boldsymbol{w}||_K^2 + \lambda\gamma_K(P_s,P_t)^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b) \geq 1-\xi_i, \\ & \xi_i \geq 0,\ i=1,\cdots,n \end{aligned} \tag{4}$$

其中, $\boldsymbol{w}$ 为投影向量, K 为特征映射核, H_K 为核空间函数集, λ 为平衡参数, $\gamma_K(P_s,P_t)$ 为源领域和目标领域间分布的 RKHS 嵌入距离度量.

为了有效求解优化问题 (4), 下面引出领域适应学习 Representer 定理:

定理 1[21] **(领域适应 Representer 定理).** 令 $\sum : [0,\infty) \to \mathbf{R}$ 表示为一个严格单调递增函数, $X = X_s \cup X_t$ 为一个数据集合, $c : (X\times\mathbf{R}^2)^n \to \mathbf{R}\cup\{\infty\}$ 为一任意损失函数[22], 正则风险函数定义为

$$R(f) = c((\boldsymbol{x}_i, y_i, f(\boldsymbol{x}_i))_{i=1}^n) + \sum(||f||_H^2)$$

其中, $f \in H$ 可以表示为如下形式:

$$f(\boldsymbol{x}) = \sum_{i=1}^{m}\beta_i k(\boldsymbol{x}_i,\boldsymbol{x}) + \sum_{j=1}^{n}\beta_j k(\boldsymbol{z}_j,\boldsymbol{x})$$

其中, k 为核映射, $\boldsymbol{x}_i \in X_s$, $y_i \in Y_s$, $z_j \in X_t$, β_i 为系数.

则由定理 1, 可得出如下结论:

定理 2. 原始优化问题 (4) 可表示为

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\xi},b} f = & \tfrac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} + C_1\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y_i\left(\sum_{j=1}^{n+m}\beta_j k_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_j)+b\right) \geq 1-\xi_i, \\ & i=1,\cdots,n \end{aligned} \tag{5}$$

其中, $\boldsymbol{x}_i \in X_s$, $\boldsymbol{x}_j \in X_s\cup X_t$, Ω 为一半正定核函数, $\xi_i \geq 0$.

证明. 给定一个非空数据集 $X = (\{\boldsymbol{x}_i\}_{i=1}^n, \{\boldsymbol{z}_j\}_{j=1}^m)$, $\boldsymbol{x}_i \in X_s$, $\boldsymbol{z}_j \in X_t$, 考虑一个非线性映射函数 $\phi : \boldsymbol{x} \to \phi(\boldsymbol{x})$, 将原始空间中的数据点 $\boldsymbol{x}$ 映射到特征空间 H 中的点 $\phi(\boldsymbol{x})$. 则原始空间数据矩阵在特征空间的表示为: $\phi(\boldsymbol{X}) = (\{\phi(\boldsymbol{x}_i)\}_{i=1}^n, \{\phi(\boldsymbol{z}_j)\}_{j=1}^m)$, 通过对核空间线性函数 $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})$ 的决策超平面法向量 $\boldsymbol{w}$ 的分析得知, 法向量 $\boldsymbol{w}$ 与源领域样本和目标领域样本有关, 并结合定理 1, 可以将特征空间中决策平面法向量表示为

$$\boldsymbol{w} = \sum_{i=1}^{n}\beta_i\phi(\boldsymbol{x}_i) + \sum_{j=1}^{m}\beta_j\phi(\boldsymbol{z}_j) \tag{6}$$

其中, $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_m,\cdots,\beta_{m+n})^{\mathrm{T}}$ 表示权值向量, 则 $\boldsymbol{w} = \phi(X)\boldsymbol{\beta}$. 从而 $\|\boldsymbol{w}\|^2 = \boldsymbol{\beta}^{\mathrm{T}}\phi(X)^{\mathrm{T}}\phi(X)\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{T}}K\boldsymbol{\beta}$, 其中, K 为包含两个领域数据的 $(n+m)\times(n+m)$ 对称半正定核矩阵.

同理, 式 (3) 变为

$$\begin{aligned} & \gamma_{KM}(p_n,q_m)^2 = \\ & \quad \boldsymbol{w}^{\mathrm{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(\boldsymbol{z}_j)\right\|^2\boldsymbol{w} = \\ & \quad \left\|\frac{1}{n}\sum_{j=1}^{n+m}\beta_j^{\mathrm{T}}\phi(X_j)^{\mathrm{T}}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i) - \right. \\ & \quad \left.\frac{1}{m}\sum_{i=1}^{n+m}\beta_i^{\mathrm{T}}\phi(X_i)^{\mathrm{T}}\sum_{j=1}^{m}\phi(\boldsymbol{z}_j)\right\|^2 = \boldsymbol{\beta}^{\mathrm{T}}\Omega_1\boldsymbol{\beta} \end{aligned} \tag{7}$$

其中, Ω_1 为一个 $(n+m)\times(n+m)$ 对称半正定核矩阵[7], 定义为

$$\begin{aligned} \Omega_1 = & \frac{1}{n^2}K_s[1]^{n\times n}K_s^{\mathrm{T}} + \frac{1}{m^2}K_t[1]^{m\times m}K_t^{\mathrm{T}} - \\ & \frac{1}{nm}(K_s[1]^{n\times m}K_t^{\mathrm{T}} + K_t[1]^{m\times n}K_s^{\mathrm{T}}) \end{aligned} \tag{8}$$

其中, K_s 为训练数据的 $(n+m)\times n$ 核矩阵, K_t 为测试数据的 $(n+m)\times m$ 核矩阵, $[1]^{k\times l}$ 为 $k\times l$ 的全 1 矩阵. 令 $\Omega = (1-\lambda)K + \lambda\Omega_1$, 定理可得. □

定理 3. 原始优化问题 (5) 的对偶问题为

$$\begin{aligned} \min_{\alpha} & \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}H^{\phi}\boldsymbol{\alpha} - 1^{\mathrm{T}}\boldsymbol{\alpha} \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C_1,\ i=1,\cdots,n \\ & \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{9}$$

其中, $H^{\phi} = \widetilde{Y}K_s^{\mathrm{T}}(\Omega)^{-1}K_s\widetilde{Y}$, $\widetilde{Y} = \mathrm{diag}\{y_1,\cdots,y_n\}$, $y_i \in Y_s$.

证明. 该定理可由文献 [23] 方法证得. □

1.3 目标标签局部重构学习

对于给定的目标领域数据点集 $X^t = \{\boldsymbol{x}_i\}_{i=1}^m$, 初始判别函数 $f^{(0)}(\boldsymbol{x})$ 可将 X^t 划分为 C ($C \geq 2$) 个类别, 分类结果可由一个类别标签矩阵 $P = [p_{ic}] \in \{0,1\}^{m\times C}$ 表示, 且当数据点 $\boldsymbol{x}_i$ 属于第 c ($1 \leq c \leq C$) 类时, $p_{ic} = 1$; 否则 $p_{ic} = 0$. 本文依据文献 [17] 做法对类别标签矩阵 P 进行如下伸缩处理:

$$Y = P(P^{\mathrm{T}}P)^{-\frac{1}{2}} = [y^1, y^2, \cdots, y^C] \tag{10}$$

其中, $\boldsymbol{y}^c = [y_1^c, y_2^c, \cdots, y_m^c]^{\mathrm{T}} \in \mathbf{R}^m$ 为 Y 的第 c 列向量, $y_i^c = p_{ic}/\sqrt{N_c}$ 表示 $\boldsymbol{x}_i$ 属于第 c 类的置信度, 其中, N_c 为第 c 类的数据点数. 对式 (10) 进行一定的数学整理可得:

$$Y^{\mathrm{T}}Y = I \tag{11}$$

其中, $I \in \mathbf{R}^{m\times m}$ 为一个单位矩阵.

根据局部学习思想, 目标领域数据点的标签信息可由其邻居数据点集的标签信息经过回归学习所得, 即对于任意目标领域数据点 $\boldsymbol{x}_i \in X^t$ ($1 \leq i \leq m$), 设 δ_i 为 $\boldsymbol{x}_i$ 的 k-近邻集 (k-NN), 且 $\boldsymbol{x}_i \notin \delta_i$, 则局部学习模型可由数据集 $\{(\boldsymbol{x}_j, y_j^c)\}_{\boldsymbol{x}_j\in\delta_i} (1 \leq c \leq C, 1 \leq j \leq m)$ 在 URKHS 中训练习得, 模型的输出形式为

$$f_i^c(\boldsymbol{x}_i) = \sum_{\boldsymbol{x}_j\in\delta_i} \beta_{ij}^c K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{12}$$

其中, $\boldsymbol{x}_i \in X^t$, $K: X \times X \to \mathbf{R}$ 为一正定核函数, β_{ij}^c 为扩张系数. 为了计算 β_{ij}^c, 需要求解如下核岭回归问题:

$$\min_{\beta_i^c\in\mathbf{R}^{N_i}} \sum_{i=1}^m (\boldsymbol{\tau}_i \left\| K_i\beta_i^c - y_i^c \right\|^2 + \eta(\beta_i^c)^{\mathrm{T}} K_i \beta_i^c) \tag{13}$$

其中, $\boldsymbol{\tau}_i = \{\tau_{ij}\}$ 为局部邻接权重向量, $K_i = [k(\boldsymbol{x}_u, \boldsymbol{x}_v)] \in \mathbf{R}^{n_i\times n_i}$ 为核矩阵, n_i 为 $\boldsymbol{x}_i$ 的 k-近邻集大小, $\boldsymbol{x}_u, \boldsymbol{x}_v \in \delta_i$, $\beta_i^c = [\beta_{ij}^c]^{\mathrm{T}}$, $\eta > 0$ 为平衡参数. 按照文献 [17] 的推导过程, 可得数据点 $\boldsymbol{x}_i$ 的局部回归标签信息为

$$f_i^c(\boldsymbol{x}_i) = k_i^{\mathrm{T}}(\boldsymbol{\tau}_i K_i + \eta I)^{-1} y_i^c \tag{14}$$

其中, $k_i = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]^{\mathrm{T}} \in \mathbf{R}^{n_i}$, $\boldsymbol{x}_j \in \delta_i$, 从而可得局部学习的一般化形式:

$$\boldsymbol{f}^c = A\boldsymbol{y}^c \tag{15}$$

其中, $\boldsymbol{f}^c = [f_1^c(\boldsymbol{x}_1), f_2^c(\boldsymbol{x}_2), \cdots, f_m^c(\boldsymbol{x}_m)]^{\mathrm{T}}$, $\alpha_i^{\mathrm{T}} = k_i^{\mathrm{T}}(\boldsymbol{\tau}_i K_i + \eta I)^{-1}$, $A = [a_{ij}] \in \mathbf{R}^{m\times m}$, 且

$$a_{ij} = \begin{cases} \alpha_{ij}, & \text{若 } \boldsymbol{x}_j \in \delta_i \\ 0, & \text{其他} \end{cases}$$

根据式 (15), 估计出所有目标领域数据点的标签信息. 为了最大可能地接近真实标签信息, 须使得所有目标数据点的估计误差最小化, 即满足如下优化形式:

$$\begin{aligned} &\min_{Y\in\mathbf{R}^m} \left\| A\boldsymbol{y}^c - \boldsymbol{y}^c \right\|^2 = \mathrm{tr}(Y^{\mathrm{T}}TY) \\ &\text{s.t.} \quad Y^{\mathrm{T}}Y = I \end{aligned} \tag{16}$$

其中, $T = (I - A)^{\mathrm{T}}(I - A)$. 按照文献 [17] 的做法, 最终的目标领域数据点的类别标签信息矩阵 P 可通过离散化矩阵 Y 获得.

1.4 目标领域判别函数学习

为了简单起见, 本文主要考虑二类分类的情况, 即 $c = 2$. 对于多类分类问题, 可根据 OAO (One against one) 或 OAA (One against all) 策略, 分解为多个二类分类问题来求解. 为了适于传统 SVM 的训练学习, 首先, 对获得的目标领域标签信息矩阵 P 进行适当的处理, 即使得目标领域数据类别标签根据数据所属的类分别取值为 -1 和 $+1$, 分别指示数据所归属的类别, 从而目标领域中训练数据的类别标签变为一个 m 维列向量 $y_t = [y_{tj}]^{\mathrm{T}} \in \mathbf{R}^m$, $\boldsymbol{y}_{tj} \in \{-1, +1\}$ ($1 \leq j \leq m$). 然后, 在目标领域, 依据传统的 SVM 的训练方法, 学习一个用于目标领域数据判别的决策函数:

$$g(\boldsymbol{x}) = \sum \psi_j K(\boldsymbol{x}, \boldsymbol{x}_j) + b^\phi \tag{17}$$

其中, ψ_j 为权值系数, $\boldsymbol{x}_j \in X_t$, b^ϕ 为偏置变量.

1.5 KDC-LDAC 完整算法

核分布一致局部领域适应学习完整算法描述如下:

算法 1. 基核分布一致局部领域适应学习算法

输入. $X_s = \{\boldsymbol{x}_{si}\}_{i=1}^n$, $X_t = \{\boldsymbol{x}_{tj}\}_{j=1}^m$, 正则参数 C_1, 分布平衡参数 λ, 近邻数 k, 局部平衡参数 η, 核带宽控制参数 γ, 热核参数 t;

输出. 目标领域判别函数: $g(\boldsymbol{x}) = \sum \psi_j K(\boldsymbol{x}, \boldsymbol{x}_j) + b^\phi$.

步骤 1. 选择参数 γ, 分别计算训练数据 X_s 的 $(n+m)\times n$ 核矩阵 K_s 和测试数据 X_t 的 $(n+m)\times m$ 核矩阵 K_t;

步骤 2. 根据式 (8) 计算 Ω_1. 选择参数 λ, 进而构建矩阵 $\Omega = K + \lambda\Omega_1$;

步骤 3. 选择参数 C_1, 根据定理 3 求解 Lagrange 乘子向量 $\boldsymbol{\alpha}$, 根据式 (16) 求得 $\boldsymbol{\beta}$, 再根据式 (17) 和式 (18) 分别计算初始决策超平面法向量 $\boldsymbol{w}^{(0)}$ 和偏置变量 $b^{(0)}$, 进而求得核分布一致正则化分类决策函数:

$$f^{(0)}(\boldsymbol{x}) = \boldsymbol{w}^{(0)} \cdot \phi(\boldsymbol{x}) + b^{(0)}$$

步骤 4. 利用 $f^{(0)}(\boldsymbol{x})$ 对所有数据进行初始划分, 根据式 (19), 求得矩阵 Y;

步骤 5. 选择参数 t 和 k, 构建数据集 X_t 的 k-近邻加权连接图 G, 并根据式 (27) 计算邻接图加权矩阵 τ;

步骤 6. 选择参数 η, 根据式 (25), 求得矩阵 Y, 进而获得目标领域数据的标签信息矩阵 P;

步骤 7. 在目标领域, 按照有监督学习方法, 训练学习一个适于目标领域数据判别的决策函数 $g(\boldsymbol{x}) = \sum \psi_j K(\boldsymbol{x}, \boldsymbol{x}_j) + b^{\phi}$.

1.6 讨论: 核局部学习目标领域数据集选择

所提方法 KDC-LDAC 在进行核局部领域适应学习阶段, 需要对所有目标领域数据的类别标签信息进行核岭回归学习, 其算法复杂度为 $\mathrm{O}(m^3)$[16−17], 这对于目标领域大样本情况效率较低. 为此, 定义目标领域数据子集:

$$M = \{\boldsymbol{x} | -1 \le f(\boldsymbol{x}) \le 1, \boldsymbol{x} \in X_t\}$$

令 $l = |M|$ 为集合 M 的基数, 则在核分布一致正则化学习后, 有 $l \ll m$. 本文依照文献 [6] 假设, 即距离分割超平面越近的数据点被初始误分的可能性越大, 或者说距离分割超平面愈远的点被初始划分正确的可能性愈大, 则 M 中的数据点被误分的可能性大于其他区域的数据点. 故此, 本文在核局部领域适应学习阶段, 选取数据子集 M 作为目标领域适应学习数据集, 从而使得核局部领域适应学习算法复杂度变为 $\mathrm{O}(l^3) \ll \mathrm{O}(m^3)$, 提高了本文算法的可扩展性.

2 讨论

2.1 局部邻接图构造

给定数据集 X, 设 $G(V, E)$ 代表一个无向加权图, 其中, $V = \{\boldsymbol{v}_i\}$ 为图顶点集, $\boldsymbol{v}_i$ 对应为某个数据点 $\boldsymbol{x}_i \in \boldsymbol{X}$, E 为图边集. 令 τ 为一个对称的权值矩阵, 其中, τ_{ij} 为数据点对间的邻接权重, 定义为

$$\tau_{ij} = \begin{cases} \exp\left(-\dfrac{d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{t}\right), \\ \qquad \text{若 } \boldsymbol{x}_i \in \delta(\boldsymbol{x}_j) \quad \text{或} \quad \boldsymbol{x}_j \in \delta(\boldsymbol{x}_i) \\ 0, \quad \text{否则} \end{cases} \tag{18}$$

其中, $t > 0$ 为热核参数, $\delta(\boldsymbol{x}_k)$ 代表数据点 $\boldsymbol{x}_k$ 的 k-近邻集.

2.2 核带宽的协调

为了说明高斯核带宽对样本的 RKHS 嵌入分布影响, 首先引出如下定理:

定理 4[19]**.** 对于一个高斯核函数类 $K_g = \{k_\sigma = \mathrm{e}^{-||\boldsymbol{x}-\boldsymbol{z}||_2^2/2\sigma^2}, \boldsymbol{x}, \boldsymbol{z} \in \mathbf{R}^d : \sigma \in [\sigma_0, \infty)\}$, $\sigma_0 > 0$, 对于任意 $k_\sigma, k_\tau \in K_g$, $0 < \tau < \sigma < \infty$, 则 $\gamma_{k\sigma}(P, Q) \ge \gamma_{k\tau}(P, Q)$.

由定理 4 可知, 核带宽越大, 领域分布的 RKHS 嵌入距离越大, 从而使得 KDC-LDAC 收敛速度减慢. 为了进一步研究高斯核带宽对 KDC-LDAC 方法的性能影响, 将高斯核带宽进行参数化, 即泛化高斯核函数定义为

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

其中, γ 为可调参数, 从下文的实验分析可知, 当 γ 太大时, 领域内样本高度内聚, 导致正负类在一定程度上出现了交叠, 不利于模式的有效分类; 而当 γ 太小时, 可能在一定程度上导致 KDC-LDAC 算法收敛缓慢, 故本文限制 $\gamma \in [1, \gamma_0]$, 其中, γ_0 为一足够大的待调正参数. 由此, 可得矩阵 Ω 的新形式:

$$\widetilde{\Omega} = K^{\frac{\sigma}{\gamma}} + \lambda \Omega_1^{\frac{\sigma}{\gamma}} \tag{19}$$

其中, $(\cdot)^{(\eta)}$ 指核矩阵 $(\cdot)$ 中核带宽为 η. 式 (19) 显示了 KDC-LDAC 中 Ω 的取值可由参数 γ 进行协调控制, 从而进一步增强了所提方法的自适应能力, 使得领域间样本分布差度量的收敛速度加速, 进一步提升了算法的执行效率.

2.3 算法复杂度

所提方法的步骤 1 计算复杂度为 $\mathrm{O}(n^3)$; 步骤 2 核局部学习算法的复杂度至多为 $\mathrm{O}(m^3)$, 另外需要额外计算目标领域数据的 k-近邻, 需要时间复杂度为 $\mathrm{O}(m^2)$; 步骤 3 训练 SVM 的复杂度为 $\mathrm{O}(m^3)$, 则 KDC-LDAC 整体计算复杂度至多为 $\mathrm{O}(n^3 + 2m^3 + m^2)$. 为了在一定程度上提高本文方法的执行效率, 在训练高维数据时, 首先采用 QR 分解方法对核矩阵进行变换处理, 以提高所提方法的执行效率.

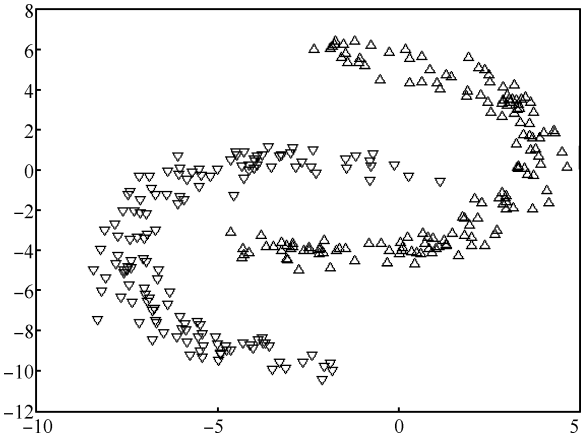
3 实验分析

为了说明所提方法 KDC-LDAC 在领域适应学习问题上的有效性, 本节将在几个不同类型数据集上进行实验: 1) 一系列具有不同复杂度的人造二维数据集 (双月形数据集[24]); 2) 多个针对不同领域适应问题的实际数据集, 包括跨领域文本分类数据集 (20 Newsgroups Reuters[1] 和情感评论[2]) 和人脸数据集[25]. 将 KDC-LDAC 与相关的方法进行比较, 以显示本文方法的优化性能.

所有实验均通过网格搜索的方式来确定优化的实验参数. 所有方法中高斯核函数带宽 σ 取值为训练样本平均范数的平方根; 热核参数 t 通过 10 重交叉验证确定; 正则参数 η 和近邻数 k 分别在网格 $\eta \in \{0.1, 1, 10\}$ 和 $k \in \{5, 10, 20, 50, 80\}$ 中搜索选取, 其中, $n_{\min} = \min\{n^+, n^-\}$, $n^{\pm}$ 指源领域正或负类样本数; 正则参数 C 采取 10 重交叉验证法来选取. 实验中, 每个数据集重复实验 10 次, 取其平均精度值作为度量所提方法的学习性能. 实验中, SVM 算法由 LIBSVM[26] 软件实现, 其他算法均在 Matlab 2009B 环境下实现.

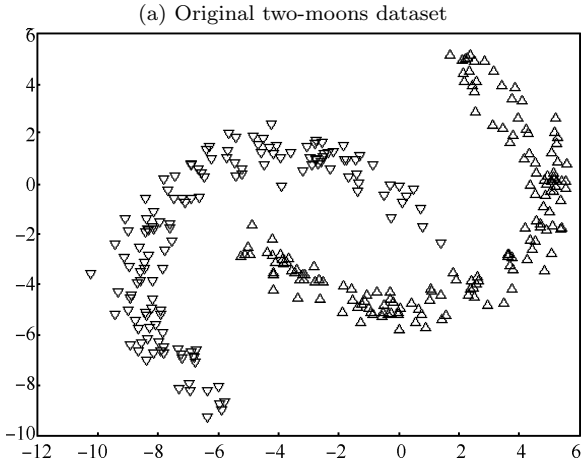
3.1　人造数据实验

人工生成一个包含 600 个样本的双月形二维样本集作为源领域数据集, 每个半月包含 300 个样本, 分别代表正类和负类样本, 如图 2 (a) 所示. 将源领域数据分别按逆时针旋转 11 次, 产生 11 个复杂度不同的目标领域数据集, 从而使得源领域和目标领域数据呈现不同分布, 而且旋转角度越大, 产生的领



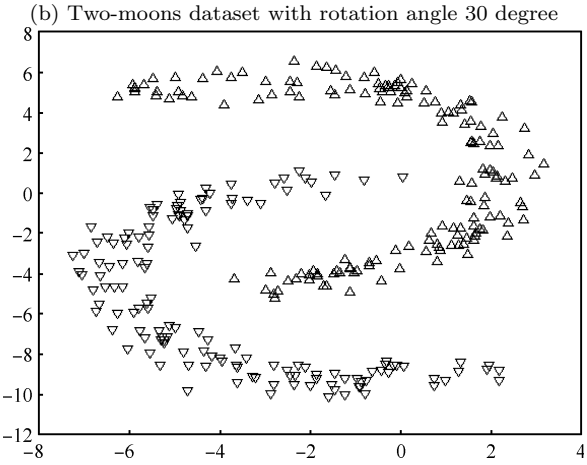
(a) 原始双月形数据集

(a) Original two-moons dataset



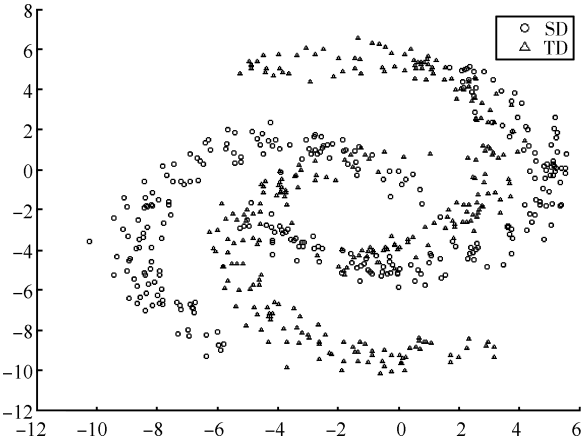
(b) 旋转 30 度后双月形数据集

(b) Two-moons dataset with rotation angle 30 degree



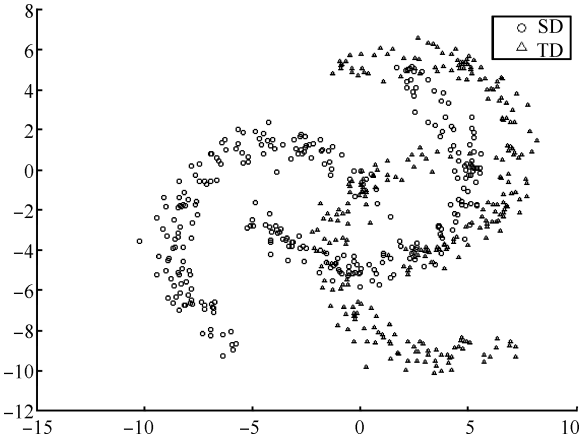
(c) 旋转 60 度后双月形数据集

(c) Two-moons dataset with rotation angle 60 degree



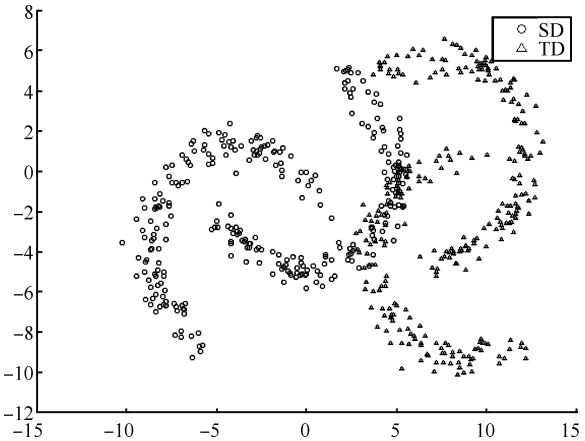
(d) 旋转 60 度且平移 1 单位后数据集

(d) Rotation angle 60 degree and horizontal movement of one unit



(e) 旋转 60 度且平移 5 单位后数据集

(e) Rotation angle 60 degree and horizontal movement of five units

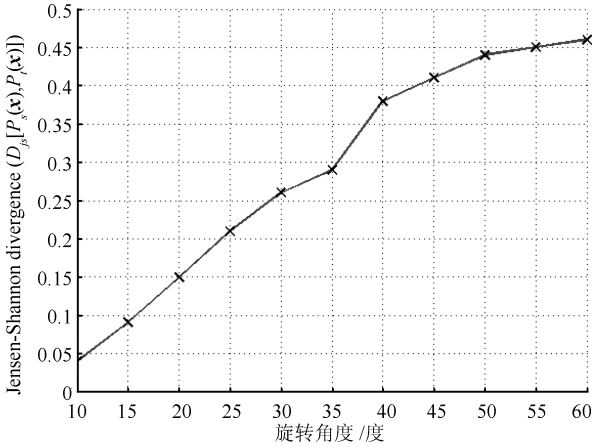


(f) 旋转 60 度且平移 10 单位后数据集

(f) Rotation angle 60 degree and horizontal movement of ten units

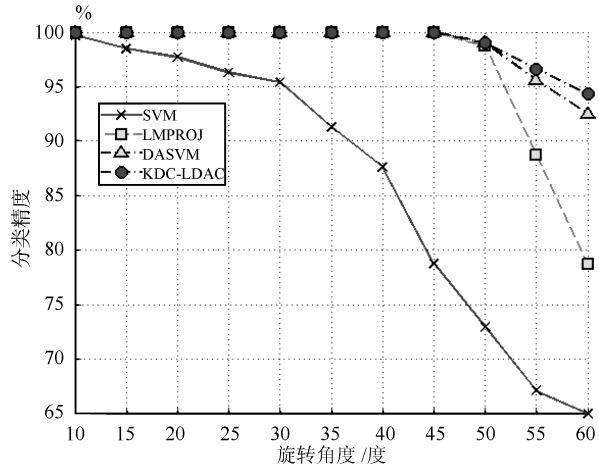
图 2　Two-moons 数据集

Fig. 2　Two-moons datasets



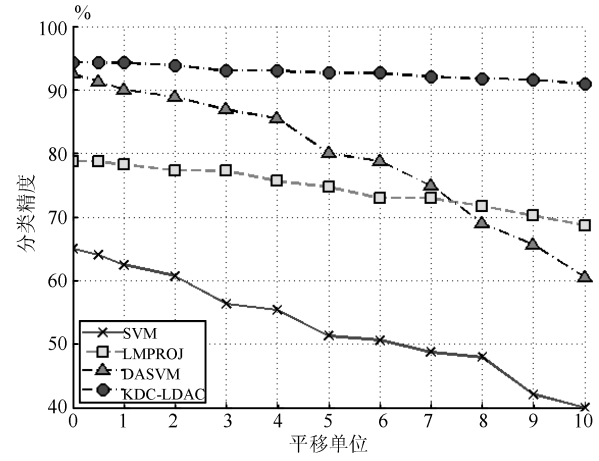
(a) 数据集分布复杂度变化

(a) Distribution variation for varied rotation angles



(b) 旋转变化对分类精度影响

(b) Classification accurancies with varied rotation degrees



(c) 分布均值距离变化对分类精度影响

(c) Classification accurancies with varied MMD

图 3 不同复杂度变化下四种方法分类精度比较

Fig. 3 Classification accurancies of four methods on datasets with varied complexity

域适应问题越复杂. 本文按照文献 [6] 做法, 采用 Jensen-Shannon 散度 (D_{JS}) 来度量领域分布之间的差异程度, 从而说明领域适应问题的复杂度. 各目标数据集以 D_{JS} 度量的复杂度值如图 3 (a) 所示. 本部分实验主要比较所提方法 KDC-LDAC 与 DASVM、SVM、LMPROJ 等方法的领域适应学习性能. 图 2 (b) 和 (c) 分别显示旋转 30 度和 60 度后的目标数据, 图 3 (b) 显示各方法在上述 11 种目标数据集上的领域适应学习性能比较.

为了进一步说明所提方法在分布间隔增大的情况下所具有的相对优势, 将原始双月形数据集先做旋转 60 度变换, 然后再做均值平移变换, 以 Δm 表示变化的均值 (不失一般性, 可令 $\Delta m \geq 0$), 在一系列均值平移变换后, 形成了多个目标领域数据集, 从而产生一系列领域适应学习问题, 图 2 (d) ∼ (f) 显示源数据旋转 60 度后分别平移 1 单位、5 单位和 10 单位后的 DAL 数据集. 图 3 (c) 显示不同方法在这些领域适应学习问题上的学习泛化误差比较曲线. 从图 3 (b) 和 (c) 显示结果可得如下几点结论:

1) 从图 3 (b) 可知, SVM 在所有旋转数据集下的分类性能均逊于其他方法, 这也进一步说明传统的基于结构风险最小化的学习方法不适于跨领域学习; 随着领域适应问题的复杂度上升, 所有方法的学习性能均呈下降趋势, 且 SVM 方法下降尤为明显. 但是本文方法 KDC-LDAC 从考虑保持领域间样本的分布的一致性的角度, 采取三段式学习模型, 使得其相较于其他方法具有一定程度的鲁棒学习性能. 另外, 在一定的旋转角度范围 (10 度 ∼ 50 度) 内, 本文方法和 DASVM 均保持着相对较高的且较为一致的学习性能, LMPROJ 方法也呈现了相当的分类性能, 而 SVM 一直呈现下降趋势. 但当旋转角度上升到 50 度以上时, 所有方法均下降明显, 这也说明当领域适应问题的复杂度上升到一定程度时, 使得源领域与目标领域间的分布呈现几乎完全无关的状态, 从而超出领域适应学习的应用条件[6].

2) 从图 3 (c) 可看出, 在一定的领域分布复杂度下, 随着分布距离的增大, 本文方法 KDC-LDAC 具有最好的一致分类性能, LMPROJ 方法也呈现出较为一致的分类性能, 而随着分布间隔的增大, DASVM 和 SVM 的学习性能明显下降, 这说明充分考虑领域数据分布距离一致性, 能有效增强领域适应学习泛化性能.

3.2 文本数据集实验

为了评价所提方法的普遍有效性, 本节将在多个不同的文本数据集上进行一系列领域适应学习实验. 与所提方法进行比较的对象, 除了两个基线方法 SVM 和 TSVM[3] 外, 还包括文献 [27] 中所分析比较的几种典型的领域适应学习方法: CDCS (Cross domain spectral classifier)[9]、LWE (locally-weighted

ensemble)[27]、LMPROJ 和 DASVM, 将这些方法分别与本文方法 KDC-LDAC 进行学习性能比较. 需要说明的是, 本文对上述所有方法均进行了参数协调学习, 并取最优的学习性能进行比较, 从而可能导致所记录的学习精度高于相关文献中报告的采取缺省策略所取得的精度值.

3.2.1 数据集设置

1) Reuters 和 20Newsgroups 分类. 首先, 从文献 [27] 中抽取两个常用的文本分类数据集 Reuters 和 20Newsgroups 来评价所提方法与相关方法的领域适应性能. 为了有效比较所提方法与相关方法的分类性能, 本文采用与文献 [7] 相同的实验设置, 即对于 20Newsgroups 和 Reuters 数据集, 分别从顶层大类中抽取 6 个和 3 个大类以构建学习数据集, 其中, 每 2 个大类分别选作正类和负类, 数据基于子类进行分割, 不同的子类认为是不同的领域. Reuters 和 20Newsgroups 数据集的详细信息如表 1 所示.

2) 情感评论分类. 情感评论领域适应数据集由 Amazon 中 4 种不同的产品类型的产品评论数据组成, 这 4 种产品类型包括书籍、DVD、电子产品和厨房设备用品. 每个评论内容包括: 产品评论分值 (范围从 0 到 5), 评论者名字和所在位置, 评论的产品 3 名、评论标题和时间以及评论内容. 评论分值高于 3 的评论标记为正, 评论分值低于 3 的评论标记为负, 其余模棱两可的评论予以删除. 表 2 汇总了情感评论数据集的详细信息. 为了进行领域适应学习性能比较, 按照文献 [2] 的设置, 构建 12 个跨领域情感评论分类任务, 如使用来自领域 A 的评论作为训练数据, 通过学习所得模型来预测来自领域 B 的评论情感类别.

3.2.2 实验结论

文本数据集跨领域学习实验的最优结果分别记录于表 3 和表 4, 其中, Avg 列表示各方法在同一数据集上的平均分类精度值. 可得实验结论如下:

1) 基线方法 SVM 因不能有效地迁移到多领域学习, 故在所有数据集上的分类性能均低于其他领域适应学习方法; 另外一个基线方法 TSVM 虽然在

表 1 跨领域文本分类数据 20Newsgroups 和 Reuters

Table 1 The cross-domain text classification datasets 20Newsgroups and Reuters

	任务	数据集	训练样本数		测试样本数	
			正类	负类	正类	负类
1	Reuters	Orgs vs. People	子类集合所有文档		子类集合所有文档	
2		Orgs vs. Place				
3		People vs. Place				
4	20Newsgroups	Comp vs. Sci	1 958	1 972	2 923	1 977
5		Rec vs. Talk	1 993	1 568	1 984	1 658
6		Rec vs. Sci	1 984	1 977	1 993	1 972
7		Sci vs. Talk	1 971	1 403	1 978	1 850
8		Comp vs. Rec	2 916	1 993	1 965	1 984
9		Comp vs. Talk	2 914	1 568	1 967	1 685

表 2 Web 文本 (情感评论) 分类数据集

Table 2 The cross-domain web text classification datasets (sentiment reviews)

	任务	数据集	训练样本数	测试样本数
10	Amazon sentiment reviews	Books (B)	1 000	1 000
11		DVD (D)	1 000	1 000
12		Electronics (E)	1 000	1 000
13		Kitchen (K)	1 000	1 000

表 3 20Newsgroups 和 Reuters 文本分类精度 (%) 比较

Table 3 Classification accuracy comparison on 20Newsgroups and Reuters datasets (%)

方法	Reuters				20Newsgroups						
	1	2	3	Average	4	5	6	7	8	9	Average
SVM	80.20	71.35	65.36	72.30	72.53	70.10	75.40	78.00	83.80	92.70	78.76
TSVM	81.84	75.80	69.80	75.81	76.75	73.40	83.90	81.20	85.24	88.74	81.54
CDCS	**88.50**	73.90	64.00	75.47	69.80	**82.92**	64.00	70.84	82.72	90.20	76.75
LWE	83.42	69.70	68.52	73.88	85.24	78.60	87.20	75.32	88.30	94.00	84.78
LMPROJ	84.63	80.20	70.80	78.54	82.52	79.30	86.34	84.68	85.40	93.43	85.28
DASVM	87.34	84.18	81.46	84.33	**88.25**	85.24	86.79	**86.66**	91.31	94.22	88.75
KDC-LDAC	86.83	**87.53**	**82.62**	**85.66**	87.17	**87.02**	**89.61**	85.48	**93.55**	**96.61**	**89.91**

表 4 情感评论分类精度 (%)

Table 4 The classification accurancies on sentiment reviews datasets (%)

任务	SVM	TSVM	CDCS	LWE	LMPROJ	DASVM	KDC-LDAC
B-D	71.29	71.52	69.33	74.54	76.71	80.47	**82.10**
E-B	67.89	68.27	72.30	69.10	71.29	**74.32**	73.61
K-B	67.72	69.96	74.34	77.60	75.80	**78.89**	77.71
B-D	75.69	74.84	82.30	78.70	83.65	84.29	**86.33**
E-D	67.81	64.51	73.00	69.40	73.20	78.92	**83.62**
K-D	71.77	71.63	80.66	78.21	81.14	84.66	**84.87**
B-E	68.83	71.13	76.80	78.90	75.20	**81.53**	79.40
D-E	69.31	70.52	70.36	75.67	77.33	76.22	**82.05**
K-E	78.77	80.16	84.07	**84.73**	81.66	80.68	82.76
B-K	71.13	76.14	79.02	78.79	80.04	81.52	**85.13**
D-K	73.68	76.57	80.58	83.19	85.33	84.72	**87.82**
E-K	80.46	80.99	**86.47**	79.89	81.38	82.37	84.44
Avg	72.20	73.02	77.44	77.39	78.56	80.80	**82.49**

部分数据集 (如 20Newgroups 和 Reuters) 上取得了一定的分类性能, 但是在情感评论数据集上的分类性能一般. 所提方法 KDC-LDAC 和 DASVM 在所有文本数据集上的分类性能在大多数情况下均优于其他方法.

2) 虽然 KDC-LDAC 和 DASVM 方法在 3 个数据集上的平均分类性能相当, 但是值得指出的是, 所提方法的分类性能在大多数情况下均优于 DASVM, 另外, LMPROJ 也在所有文本数据集上取得了较好的平均分类性能, 这充分说明, 考虑领域间分布距离能在一定程度上提升领域适应学习性能.

3.3 人脸数据集实验

为了评价所提方法 KDC-LDAC 在高维数据集下的多类分类性能, 分别选取 2 个标准人脸数据库 (Yale 和 ORL)[23, 25] 作为源领域数据集. Yale 人脸数据库包括 15 张人脸的 165 个灰度级图像, 每张人脸由 11 幅图像组成, 在每次实验中每人分别随机选取 8 幅图像作为训练样本; ORL 人脸数据库有 40 张脸, 每张脸包括 10 幅图像, 在每次实验中每人分别随机选取 8 幅图像作为训练样本. 实验前, 对上述图像集进行预处理, 使其缩放到 32 像素 ×32 像素大小, 且每个像素为 256 灰度级, 则在图像空间, 每张图像由一个 1 024 维向量表示.

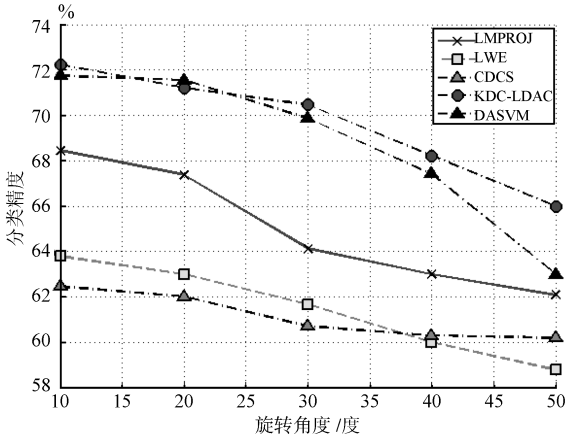
实验重点将本文方法 KDC-LDAC 与 4 个典型的领域适应方法 CDCS、LWE、LMPROJ、DASVM 进行比较, 4 种参与比较的方法采取传统的将多类划分为多个二类分类 ("一对一") 的策略进行模式划分. 记录最好的模式分类性能.

实验 1. 图像旋转实验. 本实验主要为了评价所提方法在不同数据复杂度下的领域适应学习性能. 对上述各人脸数据集样本分别进行一系列逆时针旋转, 旋转角度范围 $\theta \in [10°, 50°]$, 以分别形成目标领域数据集, 用于领域适应学习的测试样本, 从而产生多个领域适应学习数据集. 实验结果如图 4 (a) 和 (b) 所示.

实验 2. 图像旋转且平移实验. 本实验主要是为了进一步测试不同方法在一定的复杂度 (旋转角度) 下, 增加数据集领域分布间隔后的人脸识别适应能力, 将 ORL 数据集旋转 30 度后, 再分别进行平移 Δm 个单位大小, 且 $\Delta m \in [0, 5]$, 以分别形成新的目标领域数据集, 相应产生多个领域适应学习问题. 实验结果如图 4 (c) 所示.

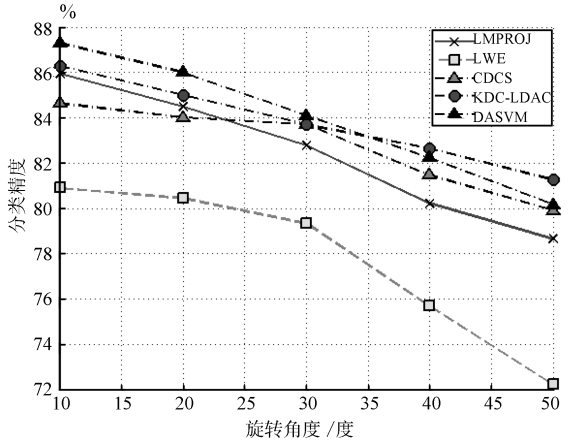
从图 4 (a) ∼ (c) 显示结果可得出如下几点结论:

1) 从图 4 (a) 和 (b) 可看出, LWE 在 2 个高维多类跨领域数据集上的模式分类性能明显低于其他领域适应学习方法; CDCS 方法在 2 个人脸数据集上的识别性能波动较大, 这说明基于谱技术的 CDCS 方法在一定程度上依赖于实验数据集. KDC-LDAC 和 DASVM 在 2 个数据集上的性能均明显优于其他方法; 随着旋转角度的增加, 即领域适应问题的复杂度上升, 所有方法的分类性能均在一定程度上呈现下降趋势, KDC-LDAC 方法下降幅度在一定程度上小于其他方法; 例外的是, CDCS 方法在一定程度上呈现较强的稳定性[7], 这进一步说明谱方法在具有流形结构的数据集 (如人脸数据) 上具备较强的



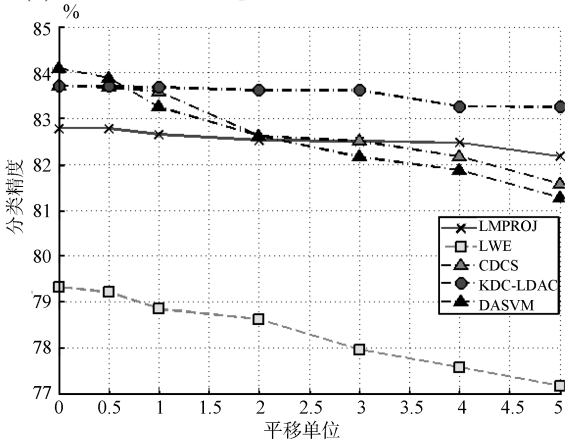
(a) 不同旋转角度下 Yale 数据集实验比较

(a) Classification comparisons on Yale with rotation



(b) 同旋转角度下 ORL 数据集实验比较

(b) Classification comparisons on ORL with rotation



(c) 旋转 60 度后双月形数据集

(c) Classification comparisons on ORL with rotation and movement

图 4　不同旋转角度或 (和) 平移单位下分类性能比较

Fig. 4　Classification comparisons on Yale and ORL

学习能力[23, 25]; 所提方法 KDC-LDAC 在 2 个数据集上的识别性能在整体上优于所有其他方法.

2) 从图 4 (c) 可看出, 随着领域分布间隔的增大, 所有方法的识别性能均在不同程度上出现了下滑, 所提方法 KDC-LDAC 呈现出了最强的稳定性, LMPROJ 也具备了一定的鲁棒性, 而 DASVM 方法下降尤为明显, 这说明 DASVM 的识别性能在较大程度上受初始 SVM 学习效果的影响, 而从充分考虑领域分布间隔差的角度, KDC-LDAC 和 LMPROJ 均在不同程度上显示了较强的学习优势.

3.4　大规模数据集实验

根据第 2.3 节讨论可知, 所提方法 KDC-LDAC 在一定程度上具有较高的计算复杂度, 为此, 本节将进一步测试 KDC-LDAC 在大规模数据集 TRECVID[1] 视频数据库上的学习性能, 并与 2 个基线方法 SVM、TSVM 和 4 个迁移学习方法 LWE, KMM[28]、LMPROJ 和 DASVM 进行比较.

1) 数据描述: TRECVID 是目前用于研究测试的最大的带标注的视频数据库之一. TRECVID 2005 数据集包括 61 901 个关键帧, 分别抽取自 6 个不同的广播频段 (英语、阿拉伯语和汉语) 的 108 小时长的视频节目. TRECVID 2007 数据集包含 21 532 个关键帧, 其分别采自新闻杂志、科学新闻、档案和教育等类节目的 60 小时长的视频. 由于 TRECVID 2007 数据集和 TRECVID 2005 数据集中视频节目结构的巨大差异性, TRECVID 数据集对于领域适应学习具有一定的挑战性[29].

2) 实验设置: 按照文献 [29] 的设置, 选取 36 个语义概念, 其覆盖了广播新闻中的 36 个主要的可视概念, 并被手工标注以分别描述 TRECVID 2007 和 TRECVID 2005 数据集中关键帧的可视内容. 本文选取 TRECVID 2005 数据集中中文频道 CCTV4 作为源领域, 以 TRECVID 2007 数据集为目标领域. 数据集 D_s 包含源领域所有标签化样本, 即 CCTV4 频道中 10 896 个关键帧视频. 根据文献 [29] 设置, 领域中关键帧的数据维数为 346. 从目标领域中随机选取 4 000 个无标签样本用于学习模型选择. 本文将采用非插值平均精度 (Non-interpolated average precision, AP)[30] 作为性能评价标准. 根据源领域中正样本的频率 (或称正频率) 大小, 将 36 个概念分割成 3 个组, 即高频组、中频组和低频组, 每个组分别包含 12 个概念. 另外, 为了评价本文方法在大规模数据集上的学习复杂度, 本文进一步比较了按照第 1.6 节策略进行学习的 KDC-LDAC, 为了区别, 实验中将该方法标记为 KDC-LDAC(l).

3) 实验结论: 所有方法在三个组 (包含 36 个概念) 以及总体上的平均分类精度记录于表 5. 表 6 记录了所有方法的总体运行时间. 从实验结果比较可得出如下几点结论:

a) 从表 5 可看出, 由于收集于不同年份的 TRECVID 数据集的分布差异较大, 基线方法 SVM 和 TSVM 在所有 36 个概念数据集上的性能均逊于其他方法; SVM、TSVM、KMM 和 LMPROJ 在高频组性能优于在低频组的性能, 可能的解释是, 在高频组的概念通常包含领域中大量正模式, 直观上来说, 在两个领域中存在大量正模式时, 在特征空间, 样本分布较密, 在此情况下, 来自两个领域的样本分布会出现彼此交叠的现象[29], 从而导致源领域数据有助于目标领域的学习; 另一方面, 对于低频组中的概念, 来自两个领域的正样本在特征空间的分布较稀, 两个领域的数据分布间隔可能较大, 因此, 在源领域的样本学习会降低目标领域的学习性能.

[1] http://www-nlpir.nist.gov/projects/trecvid

表 5 所有方法分别在 TRECVID 数据集三个组 36 个视频概念上的平均精度的总均值 (%)

Table 5 Performance comparisons in terms of mean AP from concepts of three groups on the TRECVID dataset including 36 concepts (%)

	SVM	TSVM	LWE	KMM	LMPROJ	DASVM	KDC-LDAC	KDC-LDAC(l)
高频组	40.05	41.56	45.77	44.2	44.91	46.67	**48.72**	46.7
中频组	10.11	14.22	14.61	12.66	16.09	19.49	**21.14**	18.57
低频组	4.62	17.13	13.85	16.02	16.23	**20.14**	20.10	18.70
所有组	18.26	24.30	24.74	24.29	25.74	28.77	**29.99**	27.99

表 6 所有方法分别在 TRECVID 数据集上的平均训练和测试时间 (秒) (表中 2 个数字分别表示 KDC-LDAC 和 DASVM 在第一阶段和第三阶段的训练时间)

Table 6 Average training and test time (in second) comparisons of all methods on the TRECVID dataset (The two numbers represent the average training for the frist stage and third stage, respectively, in KDC-LDAC and DASVM, respectively.)

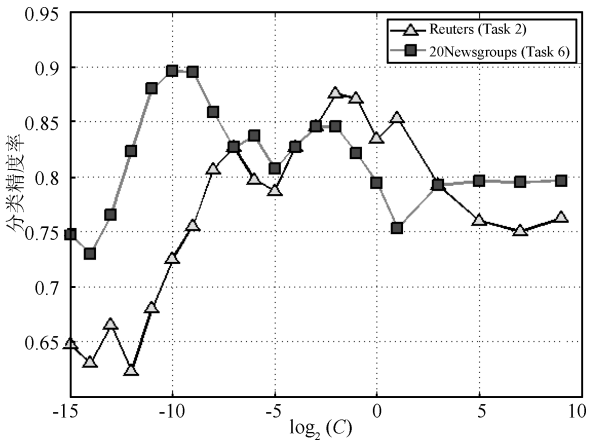
	SVM	TSVM	LWE	KMM	LMPROJ	DASVM	KDC-LDAC	KDC-LDAC(l)
训练	1 450	1 478	1 642	4 218	1 669	1 576 + 178	1 883 + 135	1 585 + 135
测试	443	318	107	508	61	74	65	65

b) 另外, 从表 5 还可看出, KMM 和 LMPROJ 的检测性能在大多数情况下优于 LWE, 这说明通过明确考虑两个领域间分布一致性, 能提升领域适应学习整体性能. 另外, DASVM 和 KDC-LDAC 具有可比较的视频检测性能, 但是, 值得强调的是, 所提方法 KDC-LDAC 在三组 36 个概念数据集上的 AP 值均优于其他方法, 这充分说明, KDC-LDAC 通过明确考虑领域分布间隔最小化和核空间局部学习技巧, 能成功有效地最小化领域间分布不匹配性和目标邻域中学习函数的结构风险; KDC-LDAC(l) 方法也取得了与 KDC-LDAC 可比较的学习性能.

c) 从表 6 可看出, SVM 训练时间最快, TSVM 次之, 因为它们只利用了源领域的标签数据进行学习; KMM 方法学习时间最长; KDC-LDAC 训练时间高于 DASVM, 但是 KDC-LDAC(l) 与 DASVM 训练时间相当; 从测试时间上来说, KDC-LDAC 或 KDC-LDAC(l) 明显低于 DASVM, 可能的解释是 DASVM 在测试阶段需要一定的时间用于循环校验, 导致其收敛速度较慢; KDC-LDAC 学习最耗时的部分主要体现在局部学习阶段, 另外, 值得说明的是, KDC-LDAC(l) 的训练时间在一定程度上明显低于 KDC-LDAC, 这说明第 1.6 节讨论的策略能在一定程度上降低所提方法的时间复杂度. 因此, 综合检测精度和时间复杂度两个因素, KDC-LDAC(l) 方法在实际应用中具备一定程度的竞争性.

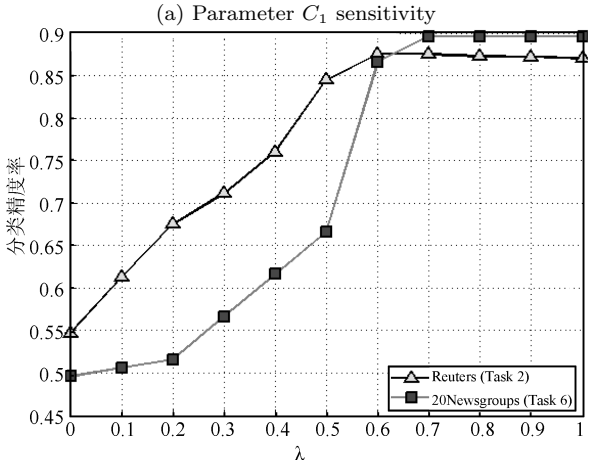
3.5 参数敏感实验

所提方法的算法实现需要协调 5 个实验参数: $C_1, \lambda, \eta, k, \gamma$, 实验中取 $\gamma_0 = 10$, 在评价某个参数的性能影响时, 先固定其他参数的最优值. 分别采用第 2 个 Reuters (任务 2) 数据集和第 3 个 20Newsgroups (任务 6) 数据集作为实验数据, 图 5 (a) ∼ (e) 分别显示了上述 5 个参数对所提方法的性能影响曲线, 由此可得如下结论:



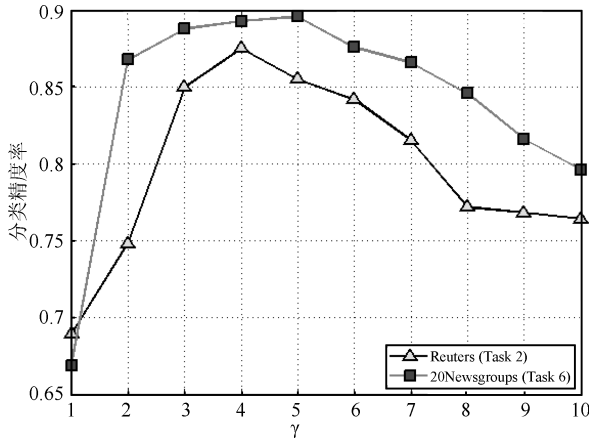
(a) 参数 C_1 的影响

(a) Parameter C_1 sensitivity



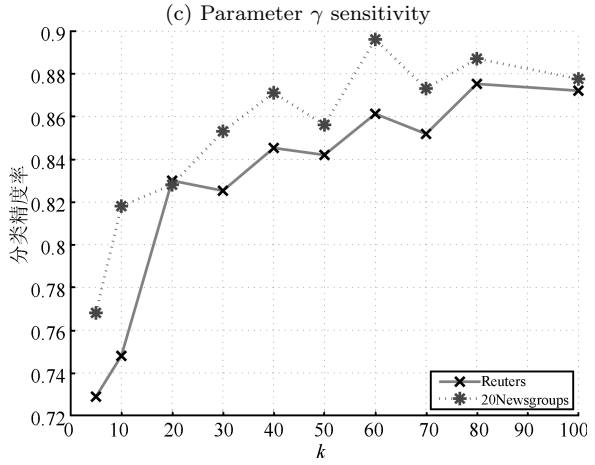
(b) 参数 λ 的影响

(b) Parameter λ sensitivity



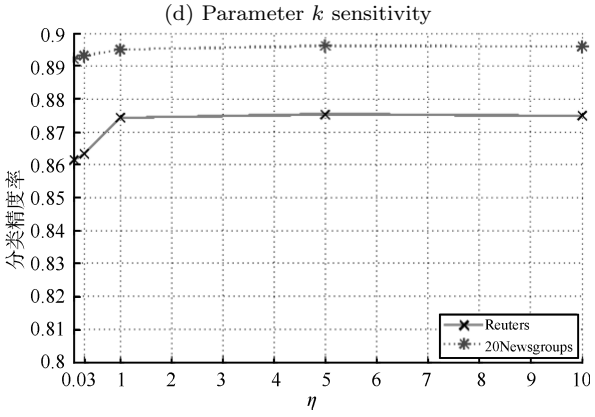
(c) 参数 γ 的影响

(c) Parameter γ sensitivity



(d) 参数 k 的影响

(d) Parameter k sensitivity



(e) 参数 η 的影响

(e) Parameter η sensitivity

图 5　参数敏感性实验

Fig. 5　Parameters sensitivity

1) 从图 5 (a) 可看出, 由于本文方法基于结构风险最小化学习模型, 故对正则化参数 C_1 具有较大程度上的敏感性, 即 C_1 在一定范围内的不同取值明显影响所提方法的泛化性能, 这也进一步说明了参数 C_1 协调的重要性.

2) 由图 5 (b) 可看出, 当 $\lambda = 0$ 时, 即忽略领域间分布差时, 所提方法不能取得最优的分类性能, 随着 λ 逐渐增加, 所提方法分类性能有所上升, 当 λ 接近 1 时, 所提方法的性能趋于稳定 (略显下降), 因此, 在实际应用中, 在不明显降低学习性能的情况下, 可固定参数 $\lambda = 1$.

3) 从图 5 (c) 可看出, 高斯核函数的带宽大小对所提方法的性能影响较突出, 由定理 4 可知, γ 越小 (如 $\gamma \in [1, 2]$), 高斯核带宽越大, 领域内数据分布散度越大, 导致领域间分布距离最小化过程收敛缓慢, 从而使得所提方法的学习性能下降; 反之, 随着 γ 值增加 (如 $\gamma \in (5, 10]$), 高斯核带宽变小, 领域内数据分布逐渐内聚, 导致正负类数据逐渐出现交叠现象, 从而使得模式分类性能下降. 只有在一定的核带宽范围内 (如 $\gamma \in [2, 5]$), 所提方法取得了较优的学习性能.

4) 从图 5 (d) 可看出, 局部近邻数 k 对所提方法的影响较大, 整体上来说, 在较大 k 值情况下, 所提方法能取得较好的学习性能, 因此, 在实际情况下, 可令近邻数 k 等于所属类的样本数.

5) 从图 5 (e) 可看出, 参数 $\eta < 1$ 时, 所提方法难以取得最优性能, 随着 η 值增加, 所提方法学习性能在一定程度上呈现平稳的上升趋势.

综上分析, 对于上述 5 个模型参数, 可以固定参数 λ 和 k, 而只需协调其他三个参数, 以此尽量减少参数选择对所提方法的性能影响, 从而进一步提高所提方法的学习效率.

4　结束语

对于领域适应学习问题, 最大化地缩小领域间样本分布差是领域适应学习成功的关键. 本文从最小化领域间分布距离的新颖视角, 在某个再生核 Hilbert 空间, 构建了一种有效的三段式领域适应学习模型, 在人造数据集和实际数据集上的系列实验结果显示了所提方法的鲁棒学习性能. 核空间的学习在一定程度上影响所提方法的学习能力, 因此对核函数的选择及其学习是本文需要进一步研究的问题. 另外, 对领域适应中高维数据的降维处理也是一个值得进一步研究的问题.

References

1 Pan S J, Tsang I W, Kwok J T, Yang Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011, **22**(2): 199−210

2 Xiang E W, Cao B, Hu D H, Yang Q. Bridging domains using world wide knowledge for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(6): 770−783

3 Joachims T. Transductive inference for text classification using support vector machines. In: Proceedings of the 16th In-

ternational Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1999. 200−209

4 Ozawa S, Roy A, Roussinov D. A multitask learning model for online pattern recognition. *IEEE Transactions on Neural Networks*, 2009, **20**(3): 430−445

5 Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345−1359

6 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

7 Quanz B, Huan J. Large margin transductive transfer learning. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM). New York, NY, USA: ACM, 2009. 1327−1336

8 Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. In: Proceedings of the 2006 Conference on Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press, 2007. 137−144

9 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. New York, NY, USA: ACM, 2008. 488−496

10 Dai W Y, Xue G R, Yang Q, Yu Y. Co-clustering based classification for out-of-domain documents. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA: ACM, 2007. 210−219

11 Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. 120−128

12 Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL′07). 2007. 440−447

13 Sriperumbudur B K, Gretton A, Fukumizu K, Schölkopf B, Lanckriet G R G. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 2010, **11**(3): 1517−1561

14 Gretton A, Harchaoui Z, Fukumizu K, Sriperumbudur B K. A fast, consistent kernel two-sample test. In: Proceedings of Advances in Neural Information Processing Systems 22, the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009). Red Hook, NY: MIT Press, 2010. 673−681

15 Vapnik V N. *Statistical Learning Theory*. New York: John Wiley and Sons, 1998

16 Zeng H, Cheung Y M. Feature selection and kernel learning for local learning-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(8): 1532−1547

17 Wu M R, Schölkopf B. Transductive classification via local learning regularization. In: Proceedings of the 11th International Conference Artificial Intelligence and Statistics. Cambridge, MA: MIT Press, 2007. 628−635

18 Hofmann T, Schölkopf B, Smola A J. Kernel methods in machine learning. *Annals of Statistics*, 2008, **36**(3): 1171−1220

19 Sriperumbadur B K, Fukumizu K, Gretton A, Lanckriet G, Schoelkopf B. Kernel choice and classifiability for RKHS embeddings of probability distributions. In: Proceedings of Advances in Neural Information Processing Systems 22, the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009). Red Hook, NY: MIT Press, 2010. 1750−1758

20 Smola A J, Gretton A, Song L, Schölkopf B. A Hilbert space embedding for distributions. In: Proceedings of the 18th International Conference on Algorithmic Learning Theory. Sendai, Japan: Springer-Verlag, 2007. 13−31

21 Schölkopf B, Herbrich R, Smola A J. A generalized representer theorem. In: Proceedings of the 14th Annual Conference on Computational Learning Theory COLT′2001. Berlin Heidelberg: Springer Press, 2001. 416−426

22 Wu Y C, Liu Y F. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 2007, **102**(479): 974−983

23 Tao Jian-Wen, Wang Shi-Tong. Locality-preserved maximum information variance v-support vector machine. *Acta Automatica Sinica*, 2012, **38**(1): 79−108
(陶剑文, 王士同. 局部保留最大信息差 v-支持向量机. 自动化学报, 2012, **38**(1): 79−108)

24 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006, **7**: 2399−2434

25 Cai D, He X F, Han J W, Zhang H J. Orthogonal Laplacianfaces for face recognition. *IEEE Transactions on Image Processing*, 2006, **15**(11): 3608−3614

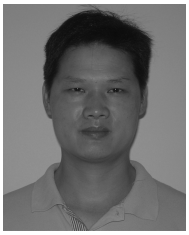
26 Chang C C, Lin C J. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm, March 4, 2013

27 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. New York, USA: ACM, 2008. 283−291

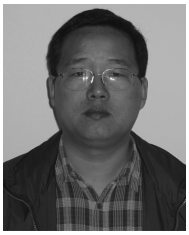
28 Huang J Y, Smola A J, Gretton A, Borgwardt K M, Schölkopf B. Correcting sample selection bias by unlabeled data. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006. 601−608

29 Jiang W, Zavesky E, Chang S F, Loui A. Cross-domain learning methods for high-level visual concept classification. In: Proceedings of the 15th IEEE International Conference on Image Processing. San Diego, CA: IEEE, 2008. 161−164

30 Xu D, Chang S F. Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(11): 1985−1997



陶剑文 江南大学数字媒体学院博士研究生. 主要研究方向为模式识别与数据挖掘技术. 本文通信作者
E-mail:jianwen_tao@yahoo.com.cn
(**TAO Jian-Wen** Ph. D. candidate with the School of Digital Media, Southern Yangtze University. His research interest covers pattern recognition and data mining.)



王士同 江南大学数字媒体学院教授. 主要研究方向为人工智能, 机器学习.
E-mail: wxwangst@yahoo.com.cn
(**WANG Shi-Tong** Professor with the School of Digital Media, Southern Yangtze University. His research interest covers artificial intelligence and machine learning.)