领域适应核支持向量机

陶剑文 [1, 2] 王士同 [1]

摘 要 领域适应学习是一种新颖的解决先验信息缺少的模式分类问题的有效方法, 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因素之一, 而仅考虑领域间分布均值差最小化, 使得在具体领域适应学习问题上存在一定的局限性. 对此, 在某个再生核 Hilbert 空间, 在充分考虑领域间分布的均值差和散度差最小化的基础上, 基于结构风险最小化模型, 提出一种领域适应核支持向量学习机 (Kernel support vector machine for domain adaptation, DAKSVM) 及其最小平方范式, 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模式分类性能.

关键词 领域适应学习, 支持向量机, 模式分类, 最大均值差, 最大散度差

引用格式 陶剑文, 王士同. 领域适应核支持向量机. 自动化学报, 2012, **38**(5): 797−811

DOI 10.3724/SP.J.1004.2012.00797

Kernel Support Vector Machine for Domain Adaptation

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

Abstract Domain adaptation learning is a novel effective technique to address pattern classification, in which the prior information for training a learning model is unavailable or insufficient. To minimize the distribution discrepancy between the source domain and target domain is one of the key factors. However, domain adaptation learning may not work well when only considering to minimize the distribution mean discrepancy between source domain and target domain. In the paper, we design a novel domain adaptation learning method based on structure risk minimization model, called DAKSVM (kernel support vector machine for domain adaptation) with respect to support vector machine (SVM) and least-square DAKSVM (LSDAKSVM) with respect to least-square SVM (LS-SVM), respectively to effectively minimize both the distribution mean discrepancy and the distribution scatter discrepancy between source domain and target domain in some reproduced kernel Hilbert space, which is then used to improve the classification performance. Experimental results on artificial and real world problems show the superior or comparable effectiveness of the proposed approach compared to related approaches.

Key words Domain adaptation learning, support vector machine (SVM), pattern classification, maximum mean discrepancy, maximum scatter discrepancy

Citation Tao Jian-Wen, Wang Shi-Tong. Kernel support vector machine for domain adaptation. *Acta Automatica Sinica*, 2012, **38**(5): 797−811

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的数据较少, 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我们训练的需要.

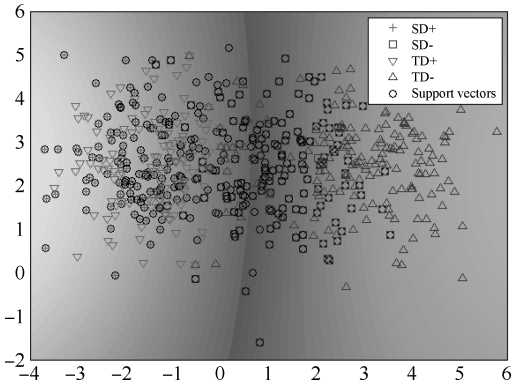
针对上述问题, 迁移学习 (Transfer learning, TL)[5] 得以提出, 旨在将从一个环境中学到的知识用来帮助新环境中的学习任务. 作为迁移学习的一个特例, 领域适应学习 (Domain adaptation learning, DAL)[6] 旨在利用源领域 (Source domain, SD) 中的训练数据来解决目标领域 (Target domain, TD) 中的学习问题, SD 和 TD 中的数据分布可以相同或不同. 近年来, 在机器学习、数据挖掘、多

收稿日期 2011-06-27 录用日期 2011-11-20
Manuscript received June 27, 2011; accepted November 20, 2011
国家自然科学基金 (60975027, 60903100), 宁波市自然科学基金 (2009A610080) 资助
Supported by National Natural Science Foundation of China (60975027, 60903100) and Natural Science Foundation of Ningbo City (2009A610080)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 江南大学信息工程学院 无锡 214122 2. 浙江工商职业技术学院信息工程学院 宁波 315012
1. School of Information Engineering, Southern Yangtze University, Wuxi 214122 2. School of Information Engineering, Zhejiang Business Technology Institute, Ningbo 315012

任务学习等应用领域中, DAL 吸引了越来越多研究者的关注和研究[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能提供一种较好度量不同数据表示之间分布距离的方法. 同样, Gretton 等[13−14] 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] 提出了一种名为最大均值差 (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, Brian 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 projected transductive support vector machine, 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.

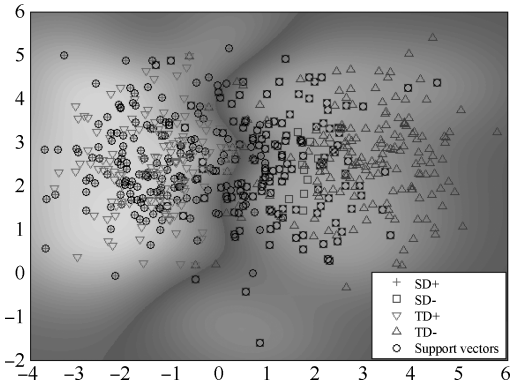
近来, 文献 [1] 从特征降维的角度分析指出, 仅仅考虑领域间样本分布的均值差在一定程度上不能充分度量领域间样本的分布距离. 另外, 根据概率论和统计学习理论[15] 可知, 均值 (或期望) 和方差 (或散度) 是描述数据分布的两个主要数学特征, 其分别度量数据分布的一阶和二阶统计特征. 从这个角度来说, 已有的领域适应学习方法仅考虑了数据分布的均值或一阶统计特征, 而未能充分考虑样本的方差或二阶统计特性. 本文认为, 为了更有效地度量领域间样本分布的距离, 在分布距离度量上应充分考虑领域间样本分布的均值和方差 (或散度) 特征. 基于此思想, 本文从充分考虑数据分布的一阶和二阶统计特征 (即均值和散度) 的角度, 基于经典支持向量机 (Support vector machine, SVM)[15] 的统计学习模型, 提出一种新颖的领域适应核支持向量机 (Kernel support vector machine for domain adaptation, DAKSVM), 在一个通用再生核 Hilbert 空间 (Universal reproducing kernel Hilbert space, URKHS)[16], DAKSVM 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, 实现领域间不同分布之间的距离充分最小化, 从而实现领域适应学习.

为了进一步说明基于 MMD 的领域适应学习方法 LMPROJ 的不足, 人工生成两个分别服从不同高斯分布的二类 2-D 样本集, 分别代表源领域 (SD) 和目标领域 (TD), SD 和 TD 中样本数为 300, 如图 1 所示. SD 中样本均值为 [−0.1345 2.9497], 方差为 [13.5742 14.0050], TD 中样本均值为 [0.1419 2.9497], 方差为 [4.8217 0.9409], 其中 SD+、SD− 和 TD+、TD− 分别代表源领域和目标领域中正类和负类样本. 本文所提方法与 SVM 和 LMPROJ 的



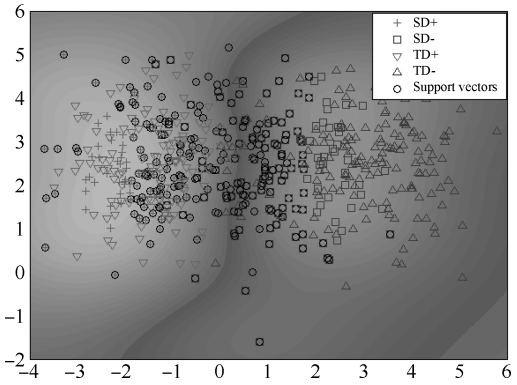
(a) SVM 分类精度: 93 %

(a) Classification accuracy of SVM: 93 %



(b) LMPROJ 分类精度: 94.6 %

(b) Classification accuracy of LMPROJ: 94.6 %



(c) 本文方法分类精度: 97.4 %

(c) Classification accuracy of the proposed method: 97.4 %

图 1 本文方法与 SVM 和 LMPROJ 的领域适应学习性能比较

Fig. 1 Comparison among SVM, LMPROJ, and DAKSVM

领域适应性能比较如图 1 (a) ∼ 1 (c) 所示, 对于这两个领域的适应学习, 从直观上来看, 水平轴方向为分割正负类的合理判别方向, 而垂直轴方向为某个具有较小方差的噪声方向, 且在该方向, 两个领域间的样本均值差最小 (或接近一致). 从图 1 看出, 基于最小化 MMD 思想的领域适应学习方法 LMPROJ 的模式分割的判别方向在一定程度上偏向于垂直轴方向 (如图 1 (b)), 而本文所提方法由于充分考虑了领域间样本分布的均值和散度差, 使得模式分割的判别方向明显优于 LMPROJ (如图 1 (c)), 而 SVM 由于仅考虑源领域正负类样本的分割间隔最大化, 从而使得领域适应性能最差 (如图 1 (a)). 由此可知, 基于 MMD 思想的领域适应学习方法在度量分布距离上仅考虑了数据分布的一阶 (均值) 统计特性, 而在一定程度上未能充分考虑数据分布的二阶 (散度) 特征, 从而在一定程度上将会限制这类学习方法在具体领域适应学习中的泛化性能.

所提方法 DAKSVM 可看成是 LMPROJ 方法的一个泛化, 另外, 本文方法思想相似于流形正则化[17] 方法, 流形正则化方法认为数据分布于某个嵌入在高维空间的低维流形空间, 流形学习的核心思想就在于寻求该低维流形子空间, 从而实现数据的特征变换, 据此, 学习的任务是寻求一个具有较低复杂度的用于较好地分割相异类别数据的决策函数, 且其在该流形空间平滑变化. 相较于现有相关方法, 本文方法的创新之处在于:

1) 承袭了基于经验风险最小化框架的大间隔分类机 SVM 易于实现的优点, 且将其引入到解决领域适应学习问题, 利用 Representer theorem 的严格数学推导, 使得所提方法凸优化解的实现只需多项式级运行时间.

2) 针对领域适应学习问题, 首次创新性地在大间隔学习机中充分引入核分布距离度量正则项, 从而在一定程度上使得同时基于均值 (或期望) 和方差距离最小化的核空间嵌入分布之间的距离度量更充分, 最大程度地减小了领域间数据分布的间隙.

3) 通过引入一个控制散度核矩阵带宽的可调参数 γ, 使得领域间分布距离差在一定的可控范围内平滑下降.

4) 对 DAKSVM 进行扩展, 提出了一种最小平方 DAKSVM 方法 (Least squared DAKSVM, LS-DAKSVM), 使得所提方法在多元分类问题上也具有一定的优化学习性能.

1 DAKSVM 方法

1.1 相关概念与问题描述

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. 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 LMPROJ 的优化形式为

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

其中, $\boldsymbol{w}$ 为投影向量, ξ_i 为松弛变量, b 为偏置变量, K 为特征映射核, H_K 为核空间函数集, λ 为平衡参数, $d_{\boldsymbol{w},K}(P_s, P_t)$ 为源领域和目标领域间的分布距离度量, 定义为

$$\begin{aligned} & d_{\boldsymbol{w},K}(P_s, P_t) = \\ & \quad \boldsymbol{w}^{\mathrm{T}} \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i) - \frac{1}{m} \sum_{j=1}^{m} \phi(\boldsymbol{z}_j) \right\|^2 \boldsymbol{w} = \\ & \quad \boldsymbol{w}^{\mathrm{T}} \left(\frac{1}{n^2} \left(\sum_{i=1}^{n} \phi(\boldsymbol{x}_i) \right)^2 + \frac{1}{m^2} \left(\sum_{j=1}^{m} \phi(\boldsymbol{z}_j) \right)^2 - \right. \end{aligned}$$

$$\frac{2}{nm}\sum_{i,j=1}^{n,m}\phi(\boldsymbol{x}_i)\phi(\boldsymbol{z}_j)\Bigg)\boldsymbol{w}$$

其中, $\boldsymbol{x}_i \in X_s$, $\boldsymbol{z}_j \in X_t$.

由式 (1) 可看出, LMPROJ 通过最小化源领域和目标领域间的样本分布的均值差来学习一个具有迁移能力的核分类机, 但式 (1) 也在一定程度上说明了 LMPROJ 方法存在的缺陷, 即 LMPROJ 没有充分考虑保持领域间样本的散度分布特征, 从而在一定程度上导致 LMPROJ 在具体的模式分类上存在 "过学习" 问题. 为此, 本文通过引入领域分布的 RKHS 嵌入距离度量的概念, 在传统 LMPROJ 方法的基础上, 提出一种鲁棒的领域适应核支持向量分类方法 (DAKSVM), DAKSVM 在选择决策超平面时充分考虑领域间数据分布的均值和散度差信息, 实验结果证实 DAKSVM 具有优于或等同于传统方法的跨领域学习性能. 为了简单起见, 本文首先主要考虑领域内二元分类任务, 接着基于最小平方范式, 对 DAKSVM 方法进行扩展提出一种最小平方领域适应核支持向量机 LSDAKSVM, 以适应多类分类问题.

1.2 DAKSVM 目标函数

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 $f : X \to R$, 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 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 (概率分布的 RKHS 嵌入距离度量)[19]. 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 $Pk = \int_M k(.,\boldsymbol{x})\mathrm{d}P(\boldsymbol{x})$. 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: $\gamma_k(P, Q) = \|Pk - Qk\|_H$.

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[10]. 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. 幸运的是, 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, 详细论证可参见文献 [13] 和 [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 嵌入均值距离度量). 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

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

其中, $\boldsymbol{x} \in X_s$, $\boldsymbol{z} \in X_t$, $\gamma_{KM}(p, q)$ 的一个无偏经验估计为

$$\gamma_{KM}(F, X_s, X_t)^2 = \left\|\int_{X_s} \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})\mathrm{d}p - \int_{X_t} \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{z})\mathrm{d}q\right\|^2 = \boldsymbol{w}^{\mathrm{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(\boldsymbol{z}_j)\right\|^2\boldsymbol{w} \tag{3}$$

定义 3 (领域分布的 RKHS 嵌入散度距离度量). 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 嵌入散度距离度量定义为

$$\gamma_{KS}(p, q) = \Big|\int_{X_s} f_{\boldsymbol{x}\sim p}(x) f_{\boldsymbol{x}\sim p}(\boldsymbol{x})^{\mathrm{T}}\mathrm{d}p - \int_{X_t} f_{\boldsymbol{z}\sim q}(\boldsymbol{z}) f_{\boldsymbol{z}\sim q}(\boldsymbol{z})^{\mathrm{T}}\mathrm{d}q\Big| \tag{4}$$

其中, $\boldsymbol{x} \in X_s$, $\boldsymbol{z} \in X_t$, $\gamma_{ks}(p, q)$ 的一个无偏经验估计为

$$\gamma_{KS}(F, X_s, X_t) = \Big|\int_{X_s} \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{w}\mathrm{d}p - \int_{X_t} \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{z})\phi(\boldsymbol{z})^{\mathrm{T}}\boldsymbol{w}\mathrm{d}q\Big| \tag{5}$$

定义 4 (领域分布的 RKHS 嵌入距离度量). 领域间样本概率分布 $p, q \in P$ 的 RKHS 嵌入距离

度量及其无偏估计分别定义为

$$\gamma_{KMS}(p,q)=(1-\lambda)\gamma_{KM}(p,q)+\lambda\gamma_{KS}(p,q),$$
$$\gamma_{KMS}(F,X_s,X_t)=$$
$$(1-\lambda)\gamma_{KM}(F,X_s,X_t)+\lambda\gamma_{KS}(F,X_s,X_t) \tag{6}$$

其中, $\lambda\in[0,1]$. 参数 λ 起平衡领域间数据分布均值差和散度差的作用, 当 λ 增大时, 偏向于保持领域间数据分布的散度一致性, 反之, 则偏向于保持领域间分布均值一致性, 特别地, 当 $\lambda=0$ 时, $\gamma_{KMS}(p,q)=\gamma_{KM}(p,q)$, DAKSVM 变为 LMPROJ. 从而, 在适当的 λ 值下, 本文方法既能较好地保持领域间数据分布一致性, 又能保持较强的领域内模式判别能力.

当 F 为 URKHS 中的单位球时, 下列定理保证 $\gamma_{KMS}(p,q)$ 能检测到两个概率分布 p 和 q 之间的差异.

定理 1[13]**.** 设 F 为定义在 URKHS 空间 H 上的一个单位球, H 及其核 $k(\cdot,\cdot)$ 定义在紧凑度量空间, 另设 X 为度量空间的一个紧凑子集, p 和 q 为定义在 X 上的 Borel 概率度量, 则当且仅当 $p=q$ 时, $\gamma_{KMS}(F,p,q)=0$.

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迁移, 即 DAKSVM 的目标函数描述为

$$\min f=\gamma_{KMS}(p,q)+C\sum_{i=1}^{n}V(\boldsymbol{x}_i,y_i,f) \tag{7}$$

其中, $\boldsymbol{x}_i\in X$ 为训练样本集, $y_i\in Y_s$ 为对应于训练集的类标签集, C 为正则化参数, V 为正则风险函数, 通常采用 Hinge 损失函数[21]: $V=(1-y_if(\boldsymbol{x}_i))_+$, 如果 $x\geq 0$, $(x)_+=x$, 否则 $x=0$. 对于线性函数 f, 式 (7) 变为

$$\arg\min_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\xi}} f=\gamma_{KMS}(p,q)+C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.}\quad y_i(\langle\boldsymbol{w},\phi(\boldsymbol{x}_i)\rangle+b)\geq 1-\xi_i,\quad i=1,\cdots,n \tag{8}$$

1.3 DAKSVM 算法

为了有效求解优化问题 (8), 下面引出领域适应学习 Representer 定理:

定理 2 (领域适应 Representer 定理)[22]**.** 令 $\sum:[0,\infty)\to\mathbf{R}$ 表示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, 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:

$$f(x)=\sum_{i=1}^{m}\beta_ik(\boldsymbol{x}_i,\boldsymbol{x})+\sum_{j=1}^{n}\beta_jk(\boldsymbol{z}_j,\boldsymbol{x}) \tag{9}$$

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

由定理 2, 可得出如下结论:

定理 3. DAKSVM 的原始优化问题形式为

$$\min_{\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{b}} f=\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta}+C\sum_{i=1}^{N}\xi_i \tag{10}$$

s.t.

$$y_i(\sum_{j=1}^{n+m}\beta_jk_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_j)+b)\geq 1-\xi_i,\quad i=1,\cdots,n$$

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: $\phi(X)=(\{\phi(\boldsymbol{x}_i)\}_{i=1}^n,\{\phi(\boldsymbol{z}_j)\}_{j=1}^m)$, 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 2, 可以将特征空间中决策平面法向量表示为

$$\boldsymbol{w}=\sum_{i=1}^{n}\beta_i\phi(\boldsymbol{x}_i)+\sum_{j=1}^{m}\beta_j\phi(\boldsymbol{z}_j) \tag{11}$$

其中, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_m,\cdots,\beta_{m+n})^{\mathrm{T}}$ 表示权值向量, 则 $\boldsymbol{w}=\phi(X)\boldsymbol{\beta}$. 从而式 (3) 变为

$$\gamma_{KM}(F,X_s,X_t)^2=$$
$$\boldsymbol{w}^{\mathrm{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i)-\frac{1}{m}\sum_{j=1}^{m}\phi(\boldsymbol{z}_j)\right\|^2\boldsymbol{w}=$$

$$\left\| \frac{1}{n}\sum_{j=1}^{n+m}\beta_j^{\mathrm{T}}\phi(X_j)^{\mathrm{T}}\sum_{i=1}^{n}\phi(\boldsymbol{x}_i) - \frac{1}{m}\sum_{i=1}^{n+m}\beta_i^{\mathrm{T}}\phi(X_i)^{\mathrm{T}}\sum_{j=1}^{m}\phi(\boldsymbol{z}_j)\right\|^2 = \boldsymbol{\beta}^{\mathrm{T}}\Omega_1\boldsymbol{\beta} \tag{12}$$

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

$$\Omega_1 = \frac{1}{n^2}K_s[1]^{n\times n}K_s^{\mathrm{T}} + \frac{1}{m^2}K_t[1]^{m\times m}K_t^{\mathrm{T}} - \frac{1}{nm}(K_s[1]^{n\times m}K_t^{\mathrm{T}} + K_t[1]^{m\times n}K_s^{\mathrm{T}}) \tag{13}$$

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.

同理, 由式 (5) 可得:

$$\begin{aligned}\gamma_{KS}(F,X_s,X_t) &= \left|\int_{X_s}\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{w}\mathrm{d}p - \int_{X_t}\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{z})\phi(\boldsymbol{z})^{\mathrm{T}}\boldsymbol{w}\mathrm{d}q\right| = \\ &\left|\int_{X_s}\sum_{j,k=1}^{n+m}\beta_j^{\mathrm{T}}\phi(\boldsymbol{x}_j)^{\mathrm{T}}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}}\beta_k\phi(\boldsymbol{x}_k)\mathrm{d}p - \int_{X_t}\sum_{j,k=1}^{n+m}\beta_j^{\mathrm{T}}\phi(\boldsymbol{x}_j)^{\mathrm{T}}\phi(\boldsymbol{z})\phi(\boldsymbol{z})^{\mathrm{T}}\beta_k\phi(\boldsymbol{x}_k)\mathrm{d}q\right| = \\ &\left|\frac{1}{n}\sum_{j,k=1}^{n+m}\beta_j^{\mathrm{T}}\beta_k\sum_{i=1}^{n}k_\sigma(\boldsymbol{x}_j,\boldsymbol{x}_i)k_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_k) - \frac{1}{m}\sum_{j,k=1}^{n+m}\beta_j^{\mathrm{T}}\beta_k\sum_{i=1}^{m}k_\sigma(\boldsymbol{x}_j,\boldsymbol{z}_i)k_\sigma(\boldsymbol{z}_i,\boldsymbol{x}_k)\right| = \\ &\boldsymbol{\beta}^{\mathrm{T}}\left|\frac{1}{n}K_sK_s^{\mathrm{T}} - \frac{1}{m}K_tK_t^{\mathrm{T}}\right|\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{T}}\Omega_2\boldsymbol{\beta}\end{aligned} \tag{14}$$

其中, $\Omega_2 = |\frac{1}{n}K_sK_s^{\mathrm{T}} - \frac{1}{m}K_tK_t^{\mathrm{T}}|$ 为一个 $(n+m)\times(n+m)$ 对称半正定核矩阵. 令 $\Omega = (1-\lambda)\Omega_1 + \lambda\Omega_2$, 定理可得. □

从式 (12) 和式 (14) 可看出, 所提方法 DKSVM 将输入空间的领域间样本分布差的度量转化为 RKHS 嵌入空间的多个核函数项的简单代数运算, 从而使得领域间分布差的计算简单, 方便.

定理 4. DAKSVM 原始优化问题 (10) 的对偶问题为

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}H^{\phi}\boldsymbol{\alpha} - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\alpha} \tag{15}$$

$$\text{s.t.}\quad 0\le\alpha_i\le C,\ \ i=1,\cdots,n \tag{16}$$

$$\sum_{i=1}^{n}\alpha_iy_i = 0 \tag{17}$$

$$\boldsymbol{\beta} = (\Omega)^{-1}K_s\widetilde{Y}\boldsymbol{\alpha} \tag{18}$$

其中, $\mathbf{1}_n$ 表示元素为全 1 的向量, $H^{\phi} = \widetilde{Y}K_s^{\mathrm{T}}(\Omega)^{-1}\times K_s\widetilde{Y}$, $\widetilde{Y} = \mathrm{diag}\{y_1,\cdots,y_n\}$, $y_i\in Y_s$.

定理 4 可根据 Lagrange 对偶定理不难证得, 限于篇幅, 详细证明省略.

根据线性方法同样原理可得 DAKSVM 方法的偏置变量 b^{ϕ} 为

$$b^{\phi} = -\frac{1}{2}\left(\frac{1}{|X_{s+}|}\sum_{\boldsymbol{x}\in X_{s+}}\sum_{j=1}^{n+m}\beta_jk_\sigma(\boldsymbol{x}_j,\boldsymbol{x}) + \frac{1}{|X_{s-}|}\sum_{\boldsymbol{x}\in X_{s-}}\sum_{j=1}^{n+m}\beta_jk_\sigma(\boldsymbol{x}_j,\boldsymbol{x})\right) \tag{19}$$

其中, X_{s-} 和 X_{s+} 分别表示源领域负类样本集和正类样本集, $|\cdot|$ 表示集合基数. 综上, DAKSVM 算法描述如下:

算法 1. 基于领域分布核空间嵌入距离度量的领域适应学习

输入. 学习数据集矩阵 $X = (\{\boldsymbol{x}_i,y_i\}_{i=1}^n, \{\boldsymbol{z}_j\}_{j=1}^m)$, $\boldsymbol{x}_i\in X_s$, $y_i\in Y_s$, $\boldsymbol{z}_j\in X_t$, 高斯型核函数 $k_{\sigma/\gamma}(\boldsymbol{x},\boldsymbol{x}_i) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2(\sigma/\gamma)^2})$, 参数 γ 可调.

输出. 领域适应决策函数 $f(\boldsymbol{x})$.

步骤 1. 选择参数 γ,σ, 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 (13) 和式 (14) 分别计算 Ω_1 和 Ω_2. 选择参数 λ, 进而构建矩阵 $\Omega = (1-\lambda)\Omega_1+\lambda\Omega_2$;

步骤 3. 根据定理 4 求解 Lagrange 乘子向量 $\boldsymbol{\alpha}$, 根据式 (18) 求得 $\boldsymbol{\beta}$, 再根据式 (11) 和式 (19) 分别计算决策超平面法向量 $\boldsymbol{w}$ 和偏置变量 b;

步骤 4. 输出分类决策函数 $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b^{\phi}$.

1.4 方法扩展

本文根据最小平方支持向量机 (LS-SVM)[23] 的思想, 将上述方法 DAKSVM 的不等式约束改为等式约束形式, 从而得到 DAKSVM 方法的最小平方版 (LSDAKSVM):

$$\begin{aligned}&\min f = \gamma_{KMS}(p,q) + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2\\ &\text{s.t.}\quad (\boldsymbol{w},\phi(\boldsymbol{x}_i)) + b = y_i - \xi_i,\ \ i=1,\cdots,n\end{aligned} \tag{20}$$

原始优化形式:

$$\min_{\boldsymbol{\beta},\xi,b} f = \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \tag{21}$$

s.t.

$$\sum_{j=1}^{n+m}\beta_j k_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_j) + b = y_i - \xi_i, \xi_i \geq 0, i = 1,\cdots,n \tag{22}$$

定理 5 (二元分类解). 给定参数 $\lambda \in [0,1]$, 对于二元分类问题, 式 (21) 和式 (22) 的优化解等价于求解如下关于变量 $\boldsymbol{\alpha}$ 的线性方程组:

$$\begin{bmatrix} 0 & \mathbf{1}_n^{\mathrm{T}} \\ \mathbf{1}_n & \widetilde{\Omega} \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{Y}_s \end{bmatrix} \tag{23}$$

其中, $\mathbf{1}_n = [1,\cdots,1]^{\mathrm{T}}$, $\boldsymbol{\alpha} = [\alpha_1,\cdots,\alpha_n]^{\mathrm{T}}$, $\boldsymbol{Y}_s = [y_1,\cdots,y_n]^{\mathrm{T}}$, $\widetilde{\Omega} = K_s^{\mathrm{T}}(\Omega)^{-1}K_s + \frac{I_n}{C}$, I_n 为 n 维单位矩阵.

证明. 优化问题 (21) 和 (22) 的 Lagrange 方程为

$$L_\lambda(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\alpha}) = \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 + \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\Omega\boldsymbol{\beta} - \sum_{i=1}^{n}\alpha_i\left(\sum_{j=1}^{n+m}\beta_j k_\sigma(\boldsymbol{x}_j,\boldsymbol{x}_i) + b + \xi_i - y_i\right) \tag{24}$$

其中, α_i 为 Lagrange 乘子. 分别对各优化变量进行求偏导数并令所得各偏导方程为 0 可得:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0 \rightarrow \boldsymbol{\beta} = (\Omega)^{-1}K_s\boldsymbol{\alpha} \tag{25}$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{n}\alpha_i = 0 \tag{26}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow \xi_i = \frac{\alpha_i}{C} \tag{27}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \boldsymbol{\beta}^{\mathrm{T}}K_s + b + \xi_i = y_i \tag{28}$$

由式 (25) ∼ (28) 并消除变量 $\boldsymbol{\beta}$ 和 ξ_i 后可得:

$$\begin{bmatrix} 0 & \mathbf{1}_n^{\mathrm{T}} \\ \mathbf{1}_n & \widetilde{\Omega} \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{Y}_s \end{bmatrix} \tag{29}$$

从而定理得证. □

对于多元分类问题, 传统的做法是通过构造多个二元分类来实现多元划分, 如 "一对一" 法 (One against one, OAO) 和 "一对多" 法 (One against all, OAA) 等, 这些方法明显的缺点是计算复杂度较高且存在类不平衡问题. 为了提出 LSDAKSVM 方法对多元分类问题的求解模型, 根据文献 [24] 做法, 引入向量标签概念, 即对于包含 c 个类的训练集, 如果训练样本 $\boldsymbol{x}_i$ $(i = 1,\cdots,n)$ 属于第 k 类, 则 $\boldsymbol{x}_i$ 的类标签为 $\boldsymbol{Y}_i = [\underbrace{0,\cdots,1}_{k},\cdots,0]^{\mathrm{T}} \in \mathbf{R}^c$. 这样导致 LSDAKSVM 的计算复杂度独立于分类类数, 从而使得 LSDAKSVM 在多类分类上的计算量与单一的二元分类器的计算量相同[24], 而且不但不会损失 LSDAKSVM 的分类性能, 反而在某些情况下还会增强多类分类性能. 对于多元分类问题, LSDAKSVM 的优化问题形式描述为

$$\min_{\boldsymbol{\beta},\boldsymbol{\xi},b} f = \frac{1}{2}\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\Omega\widetilde{\boldsymbol{\beta}} + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \tag{30}$$

$$\text{s.t.} \quad \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}K_s + b = \boldsymbol{Y}_i - \xi_i, \ \ i = 1,\cdots,n \tag{31}$$

其中, $\widetilde{\boldsymbol{\beta}} \in \mathbf{R}^{n\times c}$, $b \in \mathbf{R}^c$.

定理 6 (多元分类解). 给定参数 $\lambda \in [0,1]$, 对于多元分类问题, 式 (30) 和式 (31) 的优化解等价于求解如下关于变量 $\boldsymbol{\alpha}$ 的线性方程组:

$$\begin{bmatrix} b & \boldsymbol{\alpha} \end{bmatrix}\begin{bmatrix} 0 & \mathbf{1}_n^{\mathrm{T}} \\ \mathbf{1}_n & \widetilde{\Omega} \end{bmatrix} = \begin{bmatrix} 0_c & \widetilde{Y}_s \end{bmatrix} \tag{32}$$

其中, $0_c = [0,\cdots,0]^{\mathrm{T}}$, $\boldsymbol{\alpha} = [\alpha_1,\cdots,\alpha_n]^{\mathrm{T}}$, $\widetilde{Y}_s = [\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_n]^{\mathrm{T}}$, $\widetilde{\Omega} = K_s^{\mathrm{T}}(\Omega)^{-1}K_s + \frac{I_n}{C}$.

证明. 与定理 5 类似证明可得. □

定理 5 和定理 6 分别提供了 LSDAKSVM 方法对于二元和多元分类问题的实现算法. 从式 (23) 和式 (32) 的形式上可看出, LSDAKSVM 对二元分类和多元分类具有相同的框架模型, 从而使得 LSDAKSVM 方法的分类学习复杂度独立于样本类别数, 即 LSDAKSVM 对于二元分类和多元分类问题具有相当的计算复杂度.

2 讨论

2.1 领域适应学习经验风险界

对于一个面向二元模式分类的领域适应学习问题, 设领域中的实例数据集 X 的分布概率 $P(\boldsymbol{x})$, $\boldsymbol{x} \in X$, 以及标签函数 $f: X \rightarrow [0,1]$, 定义于实例空间 X 的假设函数类 $\widetilde{H}: X \rightarrow \{0,1\}$, 则假设函数 $h \in \widetilde{H}$ 与标签函数 f 之间的差 (或假设风险函数 $\varepsilon(h,f)$)[25] 定义为

$$\varepsilon(h,f) = \mathrm{E}_{\boldsymbol{x}\sim P}\left[|h(\boldsymbol{x}) - f(\boldsymbol{x})|\right]$$

为了简单起见, $\varepsilon(h,f)$ 表示为 $\varepsilon(h)$, 对应的经验风险函数为 $\overline{\varepsilon}(h)$. 则分别对应源领域和目标领域风险及其经验风险函数为 $\varepsilon_s(h)$, $\varepsilon_s(\overline{h})$, $\varepsilon_t(h)$, $\varepsilon_t(\overline{h})$. 则

领域适应学习中理想的假设风险应为同时最小化 $\varepsilon_s(h)$ 和 $\varepsilon_t(h)$:

$$h^* = \arg\min_{h\in\widetilde{H}}[\varepsilon_s(h)+\varepsilon_t(h)]$$

令 $\lambda^*(h)=\varepsilon_s(h^*)+\varepsilon_t(h^*)$, 对于领域适应学习, 我们期望 $\lambda^*(h)$ 最小, 从而可以利用源领域风险和领域间分布距离来近似目标经验风险. 设组合经验风险 $\varepsilon_{\widetilde{\alpha}}(\overline{h})=\widetilde{\alpha}\varepsilon_t(\overline{h})+(1-\widetilde{\alpha})\varepsilon_s(\overline{h})$, $\widetilde{\alpha}\in[0,1]$, 对应地, 令 $\varepsilon_{\widetilde{\alpha}}(h)$ 为真实的组合风险, 则领域适应学习经验风险界由如下定理确定:

定理 7[25]**.** 设 $\widetilde{H}$ 为 VC-维 d 的一个假设空间, U_s, U_t 分别为抽取自 D_s, D_t 的大小为 s 的无标签样本, 另设大小为 s 的随机标签样本集 S, 分别抽取自目标领域 D_t 的 $\widetilde{\beta}s$ 个样本和源领域 D_s 的 $(1-\widetilde{\beta})s$ 个样本, 源领域和目标领域标签函数分别为 f_s, f_t. 若 $\overline{h}\in\widetilde{H}$ 为组合经验风险 $\varepsilon_{\widetilde{\alpha}}(\overline{h})$ 在 S 上的经验最小量, 且 $h_t^*=\min_{h\in\widetilde{H}}\varepsilon_t(h)$ 为目标风险最小量, 则至少以概率 $1-\delta$ 满足下式:

$$\varepsilon_t(\overline{h})\le\varepsilon_t(h_t^*)+2\sqrt{\frac{\widetilde{\alpha}^2}{\widetilde{\beta}}+\frac{(1-\widetilde{\alpha})^2}{1-\widetilde{\beta}}}\sqrt{\frac{d\log(2s)-\log\delta}{2s}}+2(1-\widetilde{\alpha})\Big(\gamma_{KMS}+4\sqrt{\frac{2d\log(2s)+\log(\frac{4}{\delta})}{s}}+\lambda^*(h)\Big)$$

2.2 矩阵奇异问题

当矩阵 Ω 为奇异矩阵时, 其逆矩阵不可得, 为了克服该问题, 在算法实现中, 可对 Ω 增加一个较小的正则项, 即令 $\Omega=(1-\lambda)\Omega_1+\lambda\Omega_2+\lambda_0 I$, 其中, I 为一个 $(n+m)\times(n+m)$ 的单位矩阵, $\lambda_0\ge 0$, 从而实现算法可解.

2.3 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 8[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 8 可知, 核带宽越大, 领域分布的 RKHS 嵌入距离越大, 从而使得 DAKSVM 收敛速度减慢. 为了进一步研究高斯核带宽对 DAKSVM 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

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

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 DAKSVM 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:

$$\widetilde{\Omega}=(1-\lambda)\Omega_1^{(\frac{\sigma}{\gamma})}+\lambda\Omega_2^{(\frac{\sigma}{\gamma})} \tag{33}$$

其中, $\Omega_i^{(\eta)}$ $(i=1,2)$ 指核矩阵 Ω_i 中核带宽为 η. 式 (33) 显示 DAKSVM 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.

以上分析说明, DAKSVM 方法的散度差度量正则项不但可以约束领域间分布散度尽量一致, 同时还能在一定核带宽范围内分别降低两个领域内样本分布的散度, 从而使得领域间样本分布散度差度量的收敛速度加速, 进一步提升了算法的执行效率.

2.4 执行效率问题

另外, 由于 DAKSVM 算法实现要计算矩阵 Ω 及其逆矩阵, 使得其与传统的大间隔方法相比具有较高的空间复杂度 ($\mathrm{O}(d^2)$) 和时间复杂度 ($\mathrm{O}(d^3)$)[26], 特别是在处理较高维数据时尤为明显, 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 分解方法对矩阵 Ω 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 DAKSVM 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[17]); 2) 多个针对不同领域适应问题的实际数据集, 包括跨领域分本分类数据集 (20 Newsgroups 和 Reuters[26]) 和 Web 查询分类数据集[16]; 3) 领域内多类分类数据集 (人脸识别数据库 Yale 和 ORL[27−28]). 将 DAKSVM 或 LS-DAKSVM 与相关的方法进行比较, 以显示本文方法的优化性能.

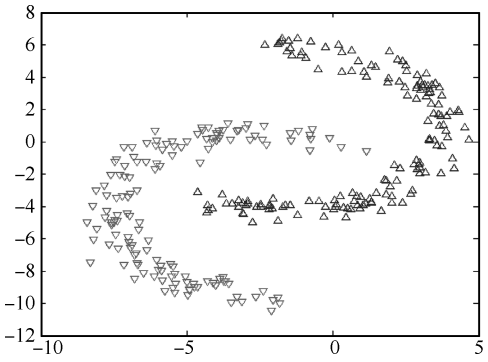
通过测试人造数据来说明 DAKSVM 算法在不同学习复杂度的可控环境下处理领域适应问题的能力, 以有助于较好地理解所提方法优化抉择领域适应超平面的运作过程与条件; 测试真实数据集主要说明本文方法在同时保持领域间数据分布的均值和散度信息一致的情况下的模式分类性能以及参数的影响; 领域内多类分类实验主要测试 LSDAKSVM 在多类分类领域的优化性能.

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; 对于 SVM, 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[29] 软件实现, 其他算法均在 Matlab 2009 B 环境下实现.

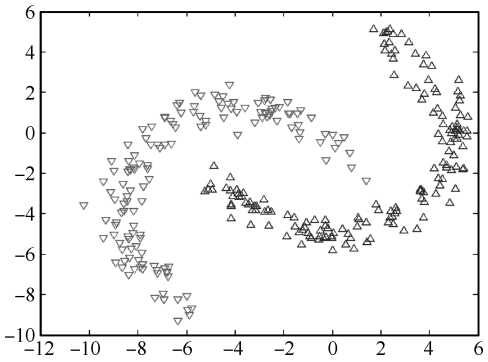
3.1　人造数据实验

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 4 (a) 所示. 本部分实验主要比较 DAKSVM 与 SVM, LMPROJ 等方法的领域适应学习性能. 图 2 (b) 和图 2 (c) 分别显示旋转 30° 和 60° 后的目标数据, 图 3 (a) 和图 3 (b), 图 3 (c) 和图 3 (d), 图 3 (e) 和图 3 (f) 分别显示不同方法在旋转 30° 和 60° 数据集上的领域适应学习效果, 图 4 (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, 由此可得如下几点结论:



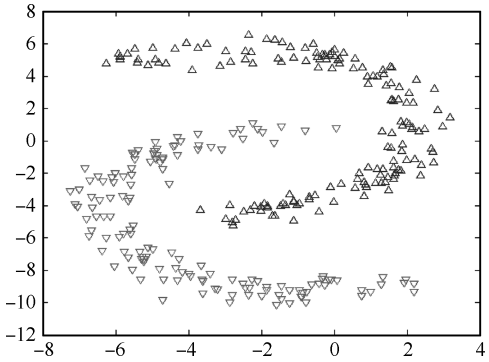
(a) 原始双月形数据集

(a) Original two-moon dataset



(b) 旋转 30° 后双月形数据集

(b) Two-moon dataset with rotation angle 30°



(c) 旋转 60° 后双月形数据集

(c) Two-moon dataset with rotation angle 60°

图 2　不同旋转角度的双月形图数据集

Fig. 2　Three two-moon datasets with different rotation degrees

1) 从图 3 (a) ∼ 3 (f) 可知, SVM 在所有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, 但是所提方法 DAKSVM 由于充分考虑保持领域间样本的分布的一致性, 使得其相较于 SVM 和 LMPROJ 方法具有一定程度的鲁棒学习性能.

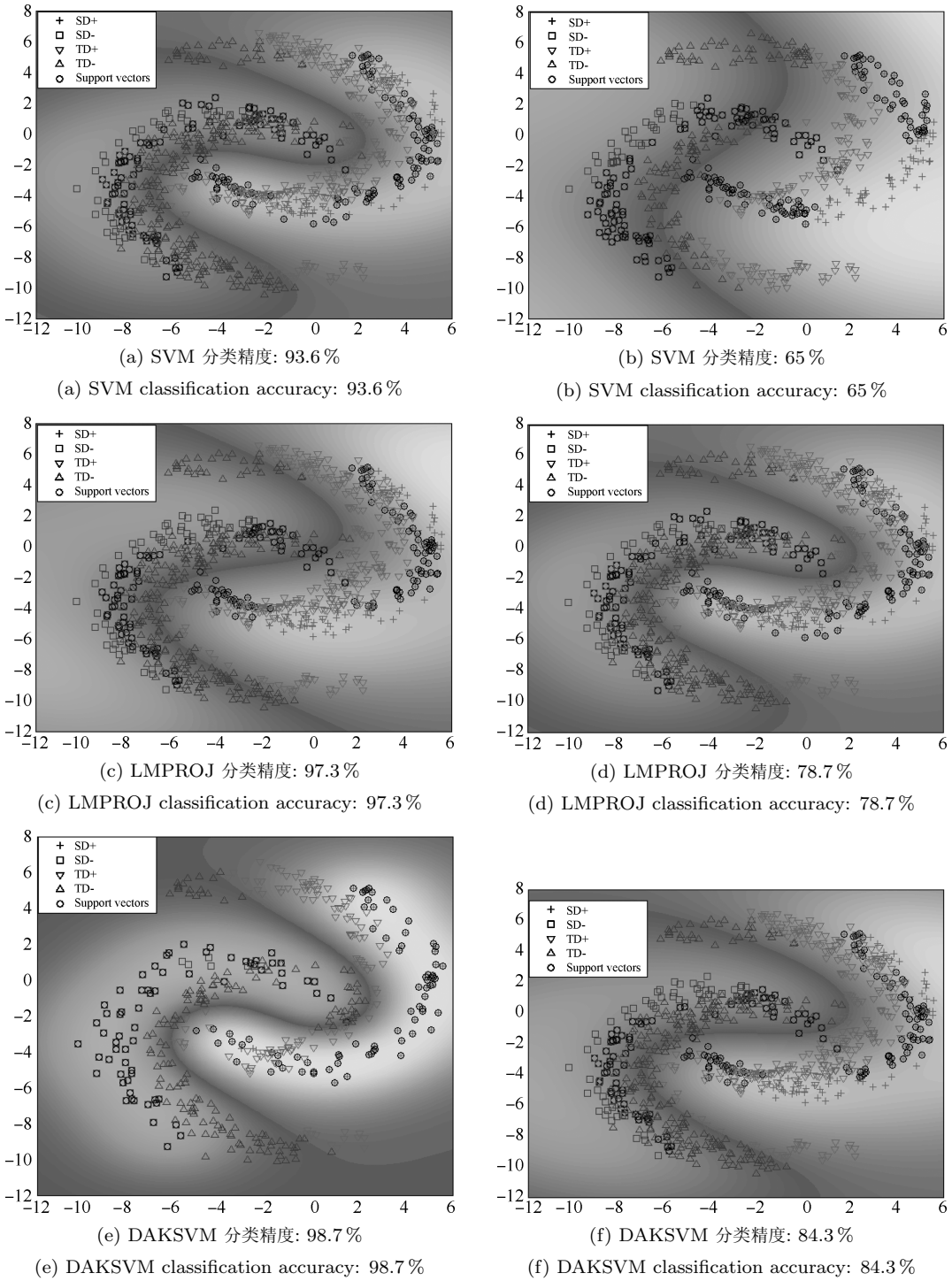
2) 从图 4 (b) 可看出, 在一定的旋转角度范围 (10° ∼ 50°) 内, 所提方法均保持着相对较高的且较为一致的学习性能, LMPROJ 方法也呈现了相当的分类性能, 而 SVM 一致呈现下降趋势. 但当旋转角度上升到 50° 以上时, 所有方法均下降明显, 这也说明当领域适应问题的复杂度上升到一定程度时, 使得源领域与目标领域间的分布呈现几乎完全无关的状态, 从而超出领域适应学习的应用条件[6].

3.2　实际数据实验

3.2.1　数据集设置

1) Reuters 和 20 Newsgroups 分类. 首先从文献 [26] 中抽取两个常用的文本分类数据集 Reuters 和 20 Newsgroups 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 20 Newsgroups 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, 不同的子类认为不同的领域. Reuters 和 20 Newsgroups 数据集的详细信息如表 1 所示.

2) Web 查询分类. 本实验旨在构建一个面向搜索引擎的跨领域查询分类. 将某个 Google 搜索结果的内容摘要集作为实验训练数据集, 某个无标签查询集作为测试集, 详细过程描述参见文献 [16]. 使



(a) SVM 分类精度: 93.6 %

(a) SVM classification accuracy: 93.6 %

(b) SVM 分类精度: 65 %

(b) SVM classification accuracy: 65 %

(c) LMPROJ 分类精度: 97.3 %

(c) LMPROJ classification accuracy: 97.3 %

(d) LMPROJ 分类精度: 78.7 %

(d) LMPROJ classification accuracy: 78.7 %

(e) DAKSVM 分类精度: 98.7 %

(e) DAKSVM classification accuracy: 98.7 %

(f) DAKSVM 分类精度: 84.3 %

(f) DAKSVM classification accuracy: 84.3 %

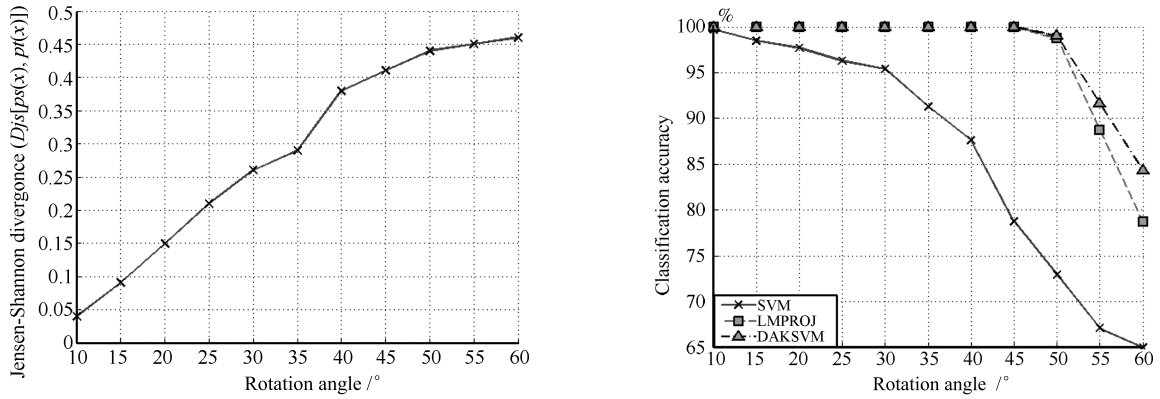
图 3 在两种双图形数据集上的决策边界

Fig. 3 The discriminant boundaries in the two two-moon datasets with 30° and 60° rotation, respectively

用 AOL 的带标签查询集作为评测数据[30] (http://grepgsadetsky.com/aol-data). 本实验主要考虑 5 个类别的查询 (Business, Computer, Entertainment, Health 和 Sports), 分别作为训练集和测试集, 从而形成 10 个二元查询分类任务. Web 查询数据集的详细信息如表 2 所示.

3.2.2 实验结论

实际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, 实验结论如下:



(a) 不同旋转角度的数据集分布复杂度变化　　(b) 不同旋转角度数据集的分类精度

(a) Two-moons complexity with different rotation degrees　　(b) Classification accuracy comparison with rotation angle

图 4　双月形人造数据集复杂度变化和分类精度曲线

Fig. 4　The two-moon datasets complexities and classification accuracies with different rotation degrees

表 1　跨领域文本分类数据 20 Newsgroups 和 Reuters

Table 1　The cross-domain text classification datasets 20 Newsgroups 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				
2	Reuters	Orgs vs. Place	子类集合所有文档		子类集合所有文档	
3	Reuters	People vs. Place				
4	20 Newsgroups	Comp vs. Sci	1 958	1 972	2 923	1 977
5	20 Newsgroups	Rec vs. Talk	1 993	1 568	1 984	1 658
6	20 Newsgroups	Rec vs. Sci	1 984	1 977	1 993	1 972
7	20 Newsgroups	Sci vs. Talk	1 971	1 403	1 978	1 850
8	20 Newsgroups	Comp vs. Talk	2 914	1 568	1 967	1 685

表 2　Web 文本分类数据 AOL 查询

Table 2　The cross-domain web text classification datasets AOL query

	任务	数据集	训练样本数	测试样本数
9	Web query	Business (B)	1 500	1 200
10	Web query	Computers (C)	1 500	1 000
11	Web query	Education (E)	2 210	2 500
12	Web query	Health (H)	1 180	1 190
13	Web query	Sports (S)	1 420	660

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

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

方法	Reuters			20 Newsgroups				
	1	2	3	4	5	6	7	8
SVM	80.20	71.35	65.36	72.53	70.10	75.40	78.00	92.70
TSVM	81.84	75.80	69.80	76.75	73.40	83.90	81.20	88.74
CDCS	**88.50**	73.90	64.00	69.80	**82.92**	64.00	70.84	90.20
LWE	83.42	69.70	68.52	**85.24**	78.60	87.20	75.32	94.00
LMPROJ	84.63	80.20	70.80	82.52	79.30	86.34	84.68	93.43
DAKSVM	84.78	**82.12**	**71.40**	83.73	81.40	**86.70**	**84.92**	**94.80**

表 4 Web 查询分类精度 (%)

Table 4 Classification accuracies on web query datasets (%)

数据集	SVM	TSVM	CDCS	LWE	LMPROJ	DAKSVM
B-C	82.52	82.64	84.34	83.26	**84.68**	**84.68**
B-H	85.93	85.42	82.86	86.72	86.48	**87.32**
B-H	90.94	91.19	**96.44**	93.20	94.82	95.34
B-S	87.25	82.60	85.40	82.56	85.78	**88.32**
C-E	87.34	83.35	78.70	85.80	88.52	**88.78**
C-H	93.19	89.70	91.28	**93.66**	92.00	93.40
C-S	92.01	84.68	89.40	84.20	86.12	**96.46**
E-H	92.55	93.11	92.76	94.40	95.38	**96.12**
E-S	81.59	77.93	**85.55**	79.46	82.70	83.20
H-S	93.45	80.84	91.25	94.71	95.00	**96.43**

表 5 人脸数据库 (Yale 和 ORL) 平均识别精度 (%)

Table 5 Classification accuracies on face databases Yale and ORL (%)

人脸数据集		LS-SVM	LMPROJ	LWE	CDCS	LSDAKSVM
Yale	10°	61.78	68.45	63.78	62.47	**70.24**
	30°	58.37	64.13	61.66	60.70	**66.47**
	50°	52.29	62.08	58.78	60.20	**63.00**
ORL	10°	76.30	85.94	80.90	84.64	**86.28**
	30°	70.72	82.00	79.33	**83.71**	83.10
	50°	65.70	78.65	72.22	**79.91**	79.25

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 虽然在部分数据集 (如 20 Newsgroups 和 Reuters) 上取得了一定的分类性能, 但是在 Web 查询文本数据集上的分类性能不佳.

2) 虽然四种领域适应学习方法在数据集 20 Newsgroups 和 Reuters 上的分类性能在整体上相当, 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或相当于 LMPROJ, 这也说明, 充分考虑领域间分布的均值差和散度差能在一定程度上提升领域适应学习的性能.

3) 在所有实际数据集上的实验结果显示, 所提方法具有较强的鲁棒性, 即所提方法在所有数据集上的分类性能均优于或相当于其他相关方法, 尤其是在 Web 查询数据上的分类性能明显优于其他五种方法.

3.3 多类分类实验

为了评价扩展方法 LSDAKSVM 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)[27−28] 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. 对上述各图像数据集训练样本分别进行逆时针旋转 10°, 30°, 50°, 以形成目标领域图像数据集, 用于领域适应学习的测试样本, 从而产生 3 个领域适应学习数据集. 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.

最小平方支持向量机 (LS-SVM) 具有与 LSDAKSVM 相似的优化结构模型, 故本文重点将所提方法 LSDAKSVM 与基线方法 LS-SVM 和 3 个领域适应方法 CDCS, LWE, LMPROJ 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于表 5. 从表 5 结果可看出:

1) LS-SVM 在所有高维多类跨领域适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,

该结论与 SVM 一致.

2) 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, LSDAKSVM 方法下降幅度相较于其他方法略小; 例外的是, CDCS 方法在一定程度上呈现较强的稳定性[7], 尤其是在较复杂的 ORL 数据集上的学习性能略优于其他方法, 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) 上具备较强的学习能力[27−28].

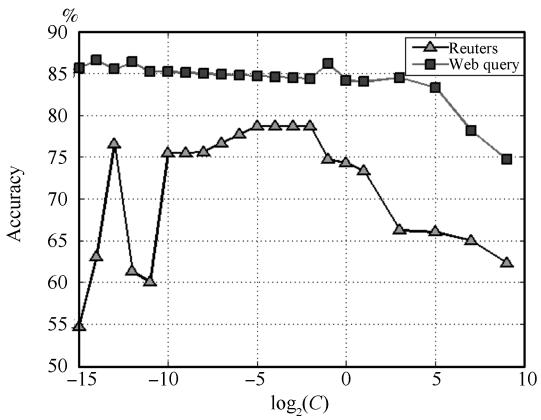
3) 从所有数据集的实验结果可看出, 所提方法 LSDAKSVM 在整体上取得了相当的或优化的分类性能, 由此可得出结论: LSDAKSVM 在多元分类上的整体性能 (如计算复杂度, 分类效率和性能等方面) 均在一定程度上具有较强的学习优势.

3.4　参数敏感实验

所提方法的算法实现需要协调三个实验参数: C, λ, γ, 其中, C 为结构风险正则化参数, 另外两个需要协调的参数是领域间数据分布的均值差和散度差平衡参数 λ 和影响高斯核函数带宽的参数 γ, 取 $\gamma_0 = 10$. 在评价某个参数的性能影响时, 先固定另两个参数的最优值. 分别采用第 2 个 Reuters 数据集和第 1 个 Web 查询数据集作为实验数据, 图 5 (a) ∼ 5 (c) 分别显示了上述 3 个参数对所提方法的性能影响曲线, 由此可得如下结论:

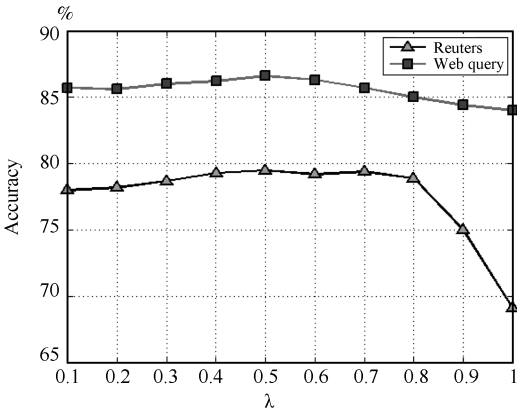
1) 由图 5 (a) 可看出, 由于本文方法基于结构风险最小化学习模型, 故对正则化参数 C 具有较大程度上的敏感性, 即 C 在一定范围内的不同取值明显影响所提方法的泛化性能, 这也进一步说明了参数 C 协调的重要性.

2) 由图 5 (b) 可看出, 当 $\lambda = 0$ 时, 即忽略领域间分布散度差时, 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, 但当 $\lambda = 1$ 时, 即忽略领域间分布均值差时, 所提方法的性能明显下降, 由此可看出, 在领域适应学习问题上, 单独考虑领域间数据分布的均值差或散度差均不能获得最优的领域适应学习性能, 而只有平衡考虑数据分布的均值差和散度差才有可能获得最优学习性能.



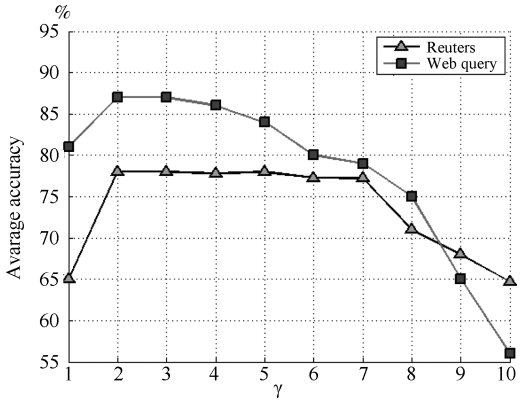
(a) 参数 C 的精度影响

(a) Sensitivities of parameter C



(b) 参数 λ 的精度影响

(b) Sensitivities of parameter λ



(c) 参数 γ 的精度影响

(c) Sensitivities of parameter γ

图 5　参数敏感性

Fig. 5　Sensitivities of parameters

3) 由图 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 9 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, 导致领域间分布散度差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 $\gamma \in [4, +\infty)$), 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 $\gamma \in [1.5, 3.5]$), 所提方法取得了较优的学习性能.

4　结论

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, 而现有基

于结构风险最小化思想的领域适应学习方法仅考虑了最小化领域间分布的均之差, 使得其在具体领域适应学习问题上存在一定的局限性, 对此, 本文从同时最小化领域间分布均值差和散度差的新颖视角, 基于结构风险最小化模型思想, 提出一种有效的领域适应学习方法 DAKSVM, 进而提出了该方法的最小平方范式 LSDAKSVM, 经过分析指出所提方法在二元分类和多元分类问题上具有一致的学习模型架构, 从而使得所提方法在领域内多类分类问题上的分类性能得到一定程度的提高. 需要进一步研究的问题: 1) 核函数的选择对所提方法的性能影响; 2) 理论上分析并提出高斯核函数带宽的变化区间及其影响函数; 3) 领域间分布差距较大时的桥接技术的进一步研究.

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 16th International Conference on Machine Learning (ICML-99). San Francisco, CA: 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, 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 Neural Information Processing Systems (NIPS) 2006. Cambridge, MA: MIT Press, 2007

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, USA: ACM, 2008

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 in Natural Language Processing. Sydney, Australia: 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). Prague, CZ: Association for Computational Linguistics, 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, Fukumizu K, Harchaoui Z, 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 Phan X H, Nguyen M L, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web (WWW′08). New York, USA: ACM, 2008. 91−100

17 Belkin M, Niyogi P, Sindhwani V, Bartlett P. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006, **7**(1): 2399−2434

18 Hofmann T, Schölkopf, Smola A J. Kernel methods in machine learning. *Annals of Statistics*, 2007, **36**(3): 1171−1220

19 Sriperumbadur B K, Fukumizu K, Gretton A, Lanckriet G R G, Schölkopf B. Kernel choice and classifiability for RKHS embeddings of probability distributions. In: 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, 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 Wu Y C, Liu Y F. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 2007, **102**(479): 974−983

22 Schölkopf B, Herbrich R, Smola A J. A generalized representer theorem. In: Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory (COLT′2001). Amsterdam, UK: Springer Press, 2001. 416− 426

23 Kanamori T, Hido S, Sugiyama M. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 2009, **10**(1): 1391−1445

24 Szedmak S, Shawe-Taylor J. Multiclass Learning at One-class Complexity. Technical Report No: 1508, School of Electronics and Computer Science, Southampton, UK, 2005

25 Blitzer J, Crammer K, Kulesza A, Pereira F, Wortman J. Learning bounds for domain adaptation. In: Proceedings of the Neural Information Processing Systems (NIPS) 2006. Cambridge, MA: MIT Press, 2007

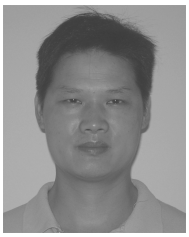
26 Gao J, Fan W, Jiang J, Han J W. Knowledge transfer via multiple model local structure mapping. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008

27 Gao Jun, Wang Shi-Tong, Deng Zhao-Hong. Global and local preserving based semi-supervised support vector machine. *Acta Electronica Sinica*, 2010, **38**(7): 1626−1633
(皋军, 王士同, 邓赵红. 基于全局和局部保持的半监督支持向量机. 电子学报, 2010, **38**(7): 1626−1633)

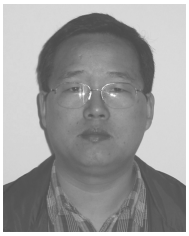
28 Cai D, He X F, Han J W, Zhang H J. Orthogonal Laplacianfaces for face recognition. *IEEE Transactions on Image Processing*, 2006, **15**(11): 3608−3614

29 Chang C C, Lin C J. LIBSVM: a library for support vector machines. *Science*, 2001, **2**(3): 1−39

30 Beitzel S M, Jensen E C, Frieder O, Lewis D D, Chowdhury A, Kolcz A. Improving automatic query classification via semi-supervised learning. In: Proceedings of the 5th IEEE International Conference on Data Mining (ICDM′05). Washington DC, USA: IEEE Computer Society, 2005. 42−49



陶剑文 江南大学信息工程学院博士研究生. 主要研究方向为模式识别与数据挖掘技术.
E-mail: jianwen_tao@yahoo.com.cn
(**TAO Jian-Wen** Ph. D. candidate at Southern Yangtze University. His research interest covers pattern recognition and data mining.)



王士同 江南大学信息工程学院教授. 主要研究方向为人工智能, 机器学习. 本文通信作者.
E-mail: wxwangst@yahoo.com.cn
(**WANG Shi-Tong** Professor at Southern Yangtze University. His research interest covers artificial intelligence and machine learning. Corresponding author of this paper.)