软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(9):2297−2310 [doi: 10.3724/SP.J.1001.2012.04240]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

多核局部领域适应学习*

陶剑文[1,3], 王士同[1,2+]

[1](江南大学 信息工程学院,江苏 无锡 214122)

[2](香港理工大学 电子计算学系,香港)

[3](浙江工商职业技术学院 信息工程学院,浙江 宁波 315012)

Multiple Kernel Local Leaning-Based Domain Adaptation

TAO Jian-Wen[1,3], WANG Shi-Tong[1,2+]

[1](School of Information Engineering, Southern Yangtze University, Wuxi 214122, China)

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

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

+ Corresponding author: E-mail: wxwangst@yahoo.com.cn

Tao JW, Wang ST. Multiple kernel local leaning-based domain adaptation. *Journal of Software*, 2012,23(9): 2297−2310 (in Chinese). http://www.jos.org.cn/1000-9825/4240.htm

Abstract: Domain adaptation (or cross domain) learning (DAL) aims to learn a robust target classifier for the target domain, which has none or a few labeled samples, by leveraging labeled samples from the source domain (or auxiliary domain). The key challenge in DAL is how to minimize the maximum distribution distance among different domains. To address the considerable change between feature distributions of different domains, this paper proposes a three-stage multiple kernel local learning-based domain adaptation (MKLDA) scheme:1) MKLDA simultaneously learns a reproduced multiple kernel Hilbert space and a initial support vector machine (SVM) by minimizing both the structure risk functional and the maximum mean discrepancy (MMD) between different domains, thus implementing the initial separation of patterns from target domain; 2) By employing the idea of local learning-based method, MKLDA predicts the label of each data point in target domain based on its neighbors and their labels in the kernel Hilbert space learned in 1); And 3) MKLDA learns a robust kernel classifier to classify the unseen data in target domain with training data well predicted in 2). Experimental results on real world problems show the outperformed or comparable effectiveness of the proposed approach compared to related approaches.

Key words: domain adaptation learning; multiple kernel learning; local learning; pattern classification; maximum mean discrepancy

摘 要: 领域适应(或跨领域)学习旨在利用源领域(或辅助领域)中带标签样本来学习一种鲁棒的目标分类器,其关键问题在于如何最大化地减小领域间的分布差异.为了有效解决领域间特征分布的变化问题,提出一种三段式多核局部领域适应学习(multiple kernel local leaning-based domain adaptation,简称 MKLDA)方法:1) 基于最大均值差(maximum mean discrepancy,简称 MMD)度量准则和结构风险最小化模型,同时,学习一个再生多核 Hilbert 空间和一

* 基金项目: 国家自然科学基金(60975027, 60903100); 宁波市自然科学基金(2009A610080)
收稿时间: 2011-10-27; 定稿时间: 2012-04-05

© 中国科学院软件研究所 http://www.jos.org.cn

个初始的支持向量机(support vector machine,简称 SVM),对目标领域数据进行初始划分;2) 在习得的多核 Hilbert 空间,对目标领域数据的类别信息进行局部重构学习;3) 最后,利用学习获得的类别信息,在目标领域训练学习一个鲁棒的目标分类器.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;多核学习;局部学习;模式分类;最大均值差

中图法分类号: TP181 **文献标识码:** A

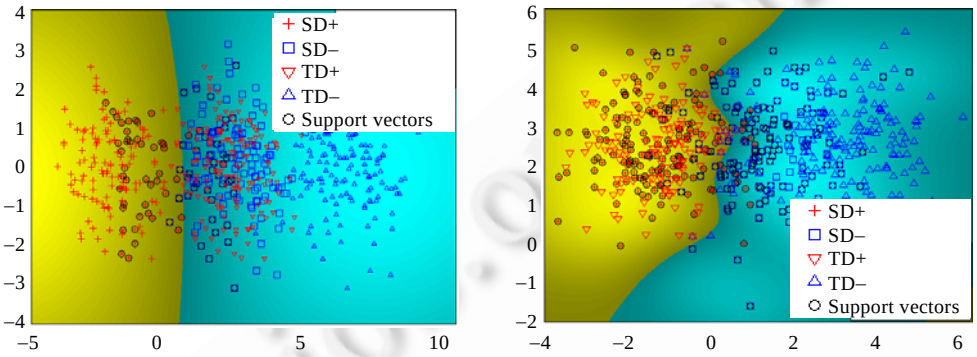
传统的机器学习方法通常要求训练数据和测试数据服从相同的概率分布,带标签的训练数据的缺少会严重影响学习性能.在一些应用领域,收集一定数量的带标签训练样本需要花费很多的时间和人力[1−3],从而在一定程度上阻碍了许多与学习相关的研究与应用的开展.近年来提出的领域适应学习(domain adaptation learning,简称 DAL)[4]旨在利用源领域(source domain,简称 SD)中的训练数据来解决目标领域(target domain,简称 TD)中的学习问题,SD 和 TD 中的数据分布可以相同或不同.在机器学习、数据挖掘、多任务学习等应用领域中,DAL 吸引了越来越多研究者的关注和研究[1,2,4−8].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 等人[6]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 等人[9,10]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进行度量,且在再生核 Hilbert 空间(reproduced kernel hilbert space,简称 RKHS)中能够明显简化这种分布距离度量的计算复杂度.基于此,Gretton 等人[10]提出了最大均值差(maximum mean discrepancy,简称 MMD)的分布距离度量方法.近来,Brian 等人[5]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 transductive support vector machine,简称 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.同样,基于 MMD 准则,文献[1]提出一种名为 TCA(transfer components analysis)的领域迁移特征变换方法,其主要学习一个领域间共同的可迁移的特征成分集,减小了领域间特征分布差距;Huang 等人[11]提出一种名为核均值匹配(kernel mean matching,简称 KMM)的两步式迁移学习方法.多核学习(multiple kernel learning,简称 MKL)方法[12]在机器学习领域得到广泛应用,基于 MKL 框架,Duan 等人[13]提出一种新颖的多核领域迁移学习方法 DTMKL(domain transfer multiple kernel learning),并在大规模视频数据检测应用中取得了明显的学习效果.文献[4]针对 DAL 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,简称 DASVM)学习框架,DASVM 的学习过程主要包括 3 步:第 1 步按照经典的支持向量机(support vector machine,简称 SVM)[14]的学习模型,利用源领域数据来初始化学习一个 SVM;第 2 步是对初始 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;第 3 步是迭代收敛,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.上述 DAL 方法在不同的具体应用领域均在一定程度上取得了较好的学习性能,但是这些方法至少尚存在如下几个问题:

(1) 实验分析得知,MMD 准则的收敛效率和效果严重依赖于核函数的选择,然而针对某个具体应用,事先无法确定最优的核函数,即使在某个预定义的足够大小的核函数集中进行穷尽搜索以寻求最优核,也需要花费大量时间.因此,优化核空间的学习在一定程度上影响了现有的基于 MMD 准则的领域迁移方法的学习性能.尽管 DTMKL 方法采用多核学习策略在一定程度上提升了领域迁移的学习性能,但是 DTMKL 是基于全局学习(global learning-based)的视角来预测目标领域样本类属,其在一定程度上可能忽略领域内样本的局部分布特征,可能会限制领域迁移的最大化学习性能;

(2) 虽然 DASVM 在某些特定数据集的学习下取得了较好的领域适应学习性能,但是通过对文献[4]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,在领域间数据分布距离较大时,DASVM 的初始 SVM 的学习在一定程度上易产生明显的“过拟合学习”问题,从而导致目标领域数据判别偏向单一方向(如图 1(a)所示).两个

© 中国科学院软件研究所 http://www.jos.org.cn

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,在一定程度上可能会限制 DASVM 在某些具体应用上的学习性能.为了说明本文方法与 DASVM 方法的差别,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 中样本点数均为 300,SD 中样本均值为[0 2.9497],TD 中样本均值为[4 2.9497],如图 1 所示,其中,SD+、SD–(分别以“+”和“∇”标识)和 TD+、TD–(分别以“□”和“Δ”标识)分别代表源领域和目标领域中正类和负类样本,如图 1(a)所示.本文方法与 DASVM 的领域适应性能比较如图 1(a)、图 1(b)所示,从图 1(a)可以看出,DASVM 方法在源领域的学习出现“过拟合”现象,而目标领域数据均明显地偏向一方;在图 1(b)中,本文方法通过核映射,在某个再生核 Hilbert 空间(reproduced kernel hilbert space,简称 RKHS),充分可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.



(a) DASVM 误差率 11.8%　　(b) 本文方法误差率 3.7%

Fig.1 Error rate comparison between the proposed method and DASVM

图 1 本文方法与 DASVM 的学习误差率比较

针对上述问题,本文受 MKL 和局部学习[15]思想的启发,从充分考虑数据分布距离的角度,提出一种新颖的三段式多核局部领域适应学习方法 MKLDA(multiple kernel local learning-based domain adaptation).与现有相关方法相比,MKLDA 方法的创新之处在于:

(1) 针对 DAL 问题,首次创新性地融合多核学习和局部学习思想,在充分考虑领域分布距离的基础上,同时学习一个组合式多核空间和一个鲁棒的目标分类函数;

(2) 在局部二次回归优化问题中引入等价稀疏正则化项,在一定程度上同时实现了稀疏多核学习和高维特征提取;

(3) 通过引入一个控制核矩阵带宽的可调参数,使得领域间分布距离差在一定的可控范围内平滑下降,提升了算法收敛速度.

1 MKLDA 方法

1.1 相关概念与问题描述

对于一个模式分类问题,设数据领域为 D,领域数据概率分布为 $P(x,y)$,$x\in X$,$y\in Y$,其中,X 和 Y 分别指领域内数据实例及其对应的类标签,分类器为一个映射函数 $f(x):X\to Y$,其将实例 $x\in X$ 映射为相应的类标签 $y\in Y$.对于 DAL 问题,设源领域和目标领域分别为 D^s 和 D^t,D^s 中有所有带标签的数据集为 $X^s=\{(x_i^s,y_i^s)\}_{i=1}^n, x_i^s\in X, y_i^s\in Y$;$D^t$ 中样本集为 $X^t=\{x_i^t,y_i^t\}_{i=1}^m\subset X$,其中,$y_i^t$ 未知.

本文利用核技巧将两个概率分布嵌入到一个 RKHS[16,17],从而获得一种处理概率分布高阶统计特征的新方法[18].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 $f:X\to\mathfrak{R}$,其中,X 为一个非空紧凑集.如

© 中国科学院软件研究所 http://www.jos.org.cn

果对于所有 $x\in X$,线性点函数映射 $f\to f(x)$存在且连续,则 H 可称为一个再生核 Hilbert 空间(RKHS).在此条件下,$f(x)$可表示为一个内积:$f(x)=\langle f,\phi(x)\rangle_H$,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):$k(x,x')=\langle\phi(x),\phi(x')\rangle_H$.

定义 1(领域分布的最大均值差(MMD)). 设 $F=\{f:X\to Y\}$为一个定义于 RKHS 的函数集,且 RKHS 的特征映射为 $\phi\in X\to H$;令 $x_i\in X^s$ 和 $z_j\in X^t(1\leqslant i\leqslant n,1\leqslant j\leqslant m)$为分别采样自分布 P 和 Q 的样本,且 P 和 Q 分别为 Borel 概率度量.则概率分布的最大均值差及其经验估计分别定义为

$$\left.\begin{aligned} DIST_k(D^s,D^t)&=\|E_{X^s\sim P}[\phi(X^s)]-E_{X^t\sim Q}[\phi(X^t)]\|_H\\ DIST_k(D^s,D^t)&=\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)-\frac{1}{m}\sum_{j=1}^{m}\phi(z_j)\right\|_H \end{aligned}\right\} \tag{1}$$

其中,$E_{x\sim u}[\cdot]$指服从概率分布 u 的数学期望算子.

MMD 方法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].值得说明的是,文献[18]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,详细论证可见文献[9,18].因此,本文以下所有核函数均采用高斯型核函数 $k_\sigma(x,y)=\exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$,其中,$\sigma$为核带宽.

与 DASVM 方法相同,MKLDA 方法也包括 3 步:

1) 领域适应多核学习:基于 MMD 度量准则和结构风险最小化模型,同时学习一个再生多核 Hilbert 空间和一个初始的支持向量机(support vector machine,简称 SVM),对目标领域数据进行初始划分;
2) 多核局部学习:在习得的多核 Hilbert 空间,对目标领域数据的类别信息进行局部重构学习;
3) 目标分类器学习:最后利用习得的类别信息,在目标领域训练学习一个鲁棒的目标分类器.

1.2 领域适应多核学习

对于某个领域适应学习问题,领域适应多核学习阶段旨在同时寻求一个高斯核 Hilbert 空间和一个初始分类函数 $f^{(0)}(x)=w^T\phi(x)+b$,其中,w 为 RKHS 中待求的线性投影向量,在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,通过同时正则化领域间 MMD 和结构经验风险,同时学习一个多核组合的 Hilbert 空间和一个用于初始化分割目标领域数据的大间隔核分类机.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地迁移.即,MKLDA 的初始化目标函数描述为

$$\left.\begin{aligned} [k,f]&=\min_{w\in H_K}\psi(DIST_k^2(D^s,D^t))+\lambda\left(C_1\sum_{i=1}^{n}\xi_i+\frac{1}{2}\|w\|_k^2\right)\\ \text{s.t.}\quad & y_i(w^T\phi(x_i)+b)\geqslant 1-\xi_i,\\ & \xi_i\geqslant 0,i=1,\ldots,n \end{aligned}\right\} \tag{2}$$

其中,w 为投影向量,k 为特征映射核,H_K 为核空间函数集,C_1,λ为平衡参数,$\psi(\cdot)$为某个单调递增函数,$DIST_k^2(D^s,D^t)$ 为源领域和目标领域间分布的 MMD.

首先定义一个包含 $n+m$ 个项的列向量 s,其中,前 n 个项为 $1/n$,剩下的 m 个项为$-1/m$.

另设 $\phi(X)=[\phi(x_1^s)\phi(x_2^s)\ldots\phi(x_n^s)\phi(x_1^t)\ldots\phi(x_m^t)]$ 为特征映射后的核矩阵,则有

$$\frac{1}{n}\sum_{i=1}^{n}\phi(x_i^s)-\frac{1}{m}\sum_{j=1}^{m}\phi(x_j^t)=\phi(X)s \tag{3}$$

由公式(1)和公式(3)可得

$$DIST_k^2(D^s,D^t)=\|\phi(X)s\|^2=tr(KS) \tag{4}$$

其中,$S=ss^T\in\mathfrak{R}^{(n+m)\times(n+m)}$. $K=\phi(X)^T\phi(X)=\begin{bmatrix}K_s & K_{st}\\ K_{ts} & K_t\end{bmatrix}\in\mathfrak{R}^{(n+m)\times(n+m)}$,$K_s\in\mathfrak{R}^{n\times n}$,$K_s\in\mathfrak{R}^{m\times m}$ 和 $K_{st}\in\mathfrak{R}^{(n\times m)}$(或 $K_{ts}\in\mathfrak{R}^{(m\times n)}$)

© 中国科学院软件研究所 http://www.jos.org.cn

分别为定义于源领域、目标领域以及源和目标领域(或称交叉领域)的核矩阵,$tr(\cdot)$为矩阵迹运算,$(\cdot)^T$为矩阵或向量的转置算子.

根据 MKL 思想,对于优化问题(2),本文考虑核 k 为某个基核集合 $\{k_p\}_{p=1}^P$ 的线性组合($P\geqslant 1$ 为基核的数量),即 $k=\sum_{p=1}^P \gamma_p k_p$,其中,$\sum_{p=1}^P \gamma_p=1$,且$\gamma_p\geqslant 0$.进一步定义公式(2)中递增函数$\psi(\cdot)$为

$$\psi(tr(KS))=\frac{1}{2}(tr(KS))^2=\frac{1}{2}\left(tr\left(\sum_{p=1}^P \gamma_p K_p S\right)\right)=\frac{1}{2}\tilde{\gamma}^T qq^T\tilde{\gamma},$$

其中,$q=[q_1,q_2,\ldots,q_P]^T$,$q_p=tr(K_pS)$,$K_p=[k_p(x_i,x_j)]\in\Re^{(n+m)\times(n+m)}$,且 $\tilde{\gamma}=[\gamma_1,\gamma_2,\ldots,\gamma_P]^T$.则有 $f^{(0)}(x)=\sum_{p=1}^P \gamma_p w_p^T\phi_p(x)+b$,其中,$w_p=\sum_{i=1}^n \alpha_i\phi(x_i)$.从而,优化问题(2)可重写为

$$\left.\begin{aligned}&\min_{\tilde{\gamma}\in\Gamma}\min_{w,b,\xi}\frac{1}{2}\tilde{\gamma}^T qq^T\tilde{\gamma}+\lambda\left(\frac{1}{2}\sum_{p=1}^P\gamma_p\|w_p\|^2+C_1\sum_{i=1}^n\xi_i\right)\\&\text{s.t.}\quad y_i\left(\sum_{p=1}^P\gamma_p w_p^T\phi_p(x_i)+b\right)\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,\ldots,n\end{aligned}\right\}\tag{5}$$

其中,$\Gamma=\{\tilde{\gamma}\mid\tilde{\gamma}\geqslant 0,\tilde{\gamma}^T 1_P=1\}$为优化变量 $\tilde{\gamma}$ 的可行域.由于在不等式约束中存在γ_p 和 w_p 的乘积项,使得优化问题(5)是非凸的.按照文献[13]中所提方法,引入变换 $v_p=\gamma_p w_p$,则公式(5)可改写为

$$\left.\begin{aligned}&\min_{\gamma\in\Gamma}\min_{v_p,b,\xi}\frac{1}{2}\tilde{\gamma}^T qq^T\tilde{\gamma}+\underbrace{\lambda\left(\frac{1}{2}\sum_{p=1}^P\frac{\|w_p\|^2}{\gamma_p}+C_1\sum_{i=1}^n\xi_i\right)}_{J(\tilde{\gamma})}\\&\text{s.t.}\quad y_i\left(\sum_{p=1}^P v_p^T\phi_p(x_i)+b\right)\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,\ldots,n\end{aligned}\right\}\tag{6}$$

其中,$J(\tilde{\gamma})=\max_{\alpha\in\Delta}1_n^T\alpha-\frac{1}{2}(\alpha\circ y)^T\left(\sum_{p=1}^P\gamma_p K_p\right)(\alpha\circ y)$ 为 $\tilde{\gamma}$ 的线性函数,$\Delta=\{\alpha\mid\alpha^T y=0,0_n\leqslant\alpha\leqslant C1_n\}$为优化变量$\alpha$的可行域,$y=[y_1,y_2,\ldots,y_n]$为标签向量,$K_p=[k_p(x_i,x_j)]=[\phi_p(x_i)^T\phi_p(x_j)]\in\Re^{n\times n}$ 为标签样本的第 p 个基核矩阵.值得说明的是,对于任意的 p,当$\gamma_p=0$ 时,令 $v_p=0$.

定理 1[13]. 对于优化变量γ,v_p,b,ξ,问题(6)为凸优化问题.

根据定理 1,优化问题(6)可以通过梯度下降算法求得全局最优解,具体算法过程参见文献[13].设γ^*,α^*,b^*为优化问题(6)的优化解,则对于某个测试样本 $x\in X^t$,其目标决策值为

$$f^{(0)}(x)=\sum_{i:\alpha_i^*\neq 0}^n\alpha_i^* y_i\sum_{p=1}^P\gamma_p^* k_p(x_i,x)+b^*\tag{7}$$

1.3 多核局部学习

对于给定的目标领域数据集 $X^t=\{x_i^t\}_{i=1}^m,x_i^t\in\Re^d$,初始判别函数$f^{(0)}(x)$可将 X^t划分为 C($C\geqslant 2$)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 x_i 属于第 c($1\leqslant c\leqslant C$)类时,$p_{ic}=1$;否则,$p_{ic}=0$.本文依据文献[15]的做法对类别标签矩阵 P 进行如下伸缩处理:

$$Y=P(P^TP)^{-\frac{1}{2}}=[y^1,y^2,\ldots,y^C]\tag{8}$$

其中,$y^c=[y_1^c,y_2^c,\ldots,y_m^c]^T\in\Re^m$ ($1\leqslant c\leqslant C$)为 Y 的第 c 列向量,$y_i^c=\frac{p_{ic}}{\sqrt{N_c}}$ 表示 x_i 属于第 c 类的置信度,其中,N_c 为第 c 类的数据点数,$\sum_{c=1}^C N_c=m$.对公式(8)进行一定的数学整理可得:

$$Y^TY=I\tag{9}$$

其中,$I\in\Re^{m\times m}$ 为一个单位矩阵.

© 中国科学院软件研究所 http://www.jos.org.cn

设 $\tilde{\gamma}=[\gamma_p]_{p=1}^P$ 为学习得到的 P 个核函数的组合系数向量,且 $\{K^{(p)}\}_{p=1}^P$ 为习得的 P 个基核函数,令组合核函数为 $K^{\tilde{\gamma}}(x,z)=\sum_{p=1}^P \gamma_p K^{(p)}(x,z)$,相应的特征映射为 ϕ,多核组合再生核 Hilbert 空间为 $\tilde{H}=\oplus_{p=1}^P H^{(p)}$,其中,$H^{(p)}$为基核映射的再生核 Hilbert 空间.在 $\tilde{H}$ 空间,数据点对间距离为

$$d_{\tilde{\gamma}}(x_1,x_2)=\|\phi(x_1)-\phi(x_2)\|_{\tilde{\gamma}}^2=K^{\tilde{\gamma}}(x_1,x_1)+K^{\tilde{\gamma}}(x_2,x_2)-2K^{\tilde{\gamma}}(x_1,x_2) \tag{10}$$

根据局部学习思想[15],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,对于目标领域任意数据点 $x_i\in X^t(1\leqslant i\leqslant m)$,设 δ_i 为 x_i 的 k-近邻集(k-NN),且 $x_i\notin\delta_i$,则局部学习模型可由数据集 $\{(x_j,y_j^c)\}_{x_j\in\delta_i}$ ($1\leqslant c\leqslant C, 1\leqslant j\leqslant N_c$)在 $\tilde{H}$ 中训练习得,即 x_i 的局部判别函数为

$$f_i^c(\phi(x_i))=\phi(x_i)^T w_i^c+b_i^c,$$

其中,$\phi(x_i)=[\phi_1(x_i)\phi_2(x_i)...\phi_P(x_i)]^T\in\Re^d$,$\phi_p(x_i)\in\Re^{d_p}$ 为第 p 个核函数映射样本,d 和 d_p 分别为 P 个核组合映射空间和第 p 个核映射空间的数据特征维数,且 $d=\sum_{p=1}^P d_p$.从而,回归系数 $w_i^c\in\Re^d$ 和偏置变量 $b_i^c\in\Re$ 通过求解如下加权 l_2 范正则化最小平方问题获得:

$$\min_{w_i^c,b_i^c}\sum_{c=1}^C\sum_{i=1}^m\left[\sum_{x_j\in\delta_i}\tau_i(y_i^c-\phi(x_j)^T w_i^c-b_i^c)^2+\eta(w_i^c)^T\Lambda_{\tilde{\gamma}}^{-1}w_i^c\right] \tag{11}$$

其中,$\tau_i=[\tau_{ij}]$为局部邻接权重向量,$\Lambda_{\tilde{\gamma}}\in\Re^{d\times d}$ 为一对角矩阵,对角元素为 $(\underbrace{\gamma_1,...,\gamma_1}_{d_1},...,\underbrace{\gamma_P,...,\gamma_P}_{d_P})^T$.对于优化问题(11),按照文献[19]的推导过程可得,数据点 x_i 属于第 $c(1\leqslant c\leqslant C)$类的预测值为

$$y_i^c=f_i^c(\phi(x_i))=\phi(x_i)^T w_i^c+b_i^c=\alpha_i^T y_i^c \tag{12}$$

其中,

$$\alpha_i^T=\tau_i\left(k_i^{\tilde{\gamma}}-\frac{1}{n_i}e_i^T K_i^{\tilde{\gamma}}\right)\Pi_i(\eta I_i+\tau_i\Pi_i K_i^{\tilde{\gamma}}\Pi_i)^{-1}\Pi_i+\frac{1}{n_i}e_i^T \tag{13}$$

其中,$\Pi_i=I_i-\frac{1}{n_i}e_i e_i^T$ 为中心投影矩阵,e_i 为 $n_i\times 1$ 的全 1 向量,I_i 为 $n_i\times n_i$ 的单位矩阵.K_i^{γ} 为定义于数据集$\{x_j|x_j\in\delta_i\}$上的组合核矩阵,即 $K_i^{\tilde{\gamma}}=[K^{\tilde{\gamma}}(x_u,x_v)], x_u,x_v\in\delta_i, k_i^{\tilde{\gamma}}\in\Re^m$ 指向量$[K^{\tilde{\gamma}}(x_i,x_j)]^T, x_j\in\delta_i$.由公式(12)、公式(13)可得目标领域多核局部学习的一般化形式:

$$f^c=A^c y^c \tag{14}$$

其中,$f^c=[f_1^c(\phi(x_1)), f_2^c(\phi(x_2)),..., f_m^c(\phi(x_m))]^T$,$A^c=[a_{ij}]\in\Re^{m\times m}$,且 $a_{ij}=\begin{cases}\alpha_{ij}, & \text{if } x_j\in\delta_i\\ 0, & \text{otherwise}\end{cases}$.

根据公式(14),可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:

$$\min_{Y\in R^N}\|A^c y^c-y^c\|^2=trace(Y^T TY)\quad \text{s.t.}\quad Y^T Y=I \tag{15}$$

其中,$T=(I-A^c)^T(I-A^c)$.公式(15)中,优化解 Y 矩阵由矩阵 T 的最小的 C 个特征值所对应的 C 个特征向量构成.按照文献[20]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 one against one(OAO)或 one against all(OAA)策略进行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 $\boldsymbol{P}$ 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:

$$y_i^t=[y_{ij}^t]^T\in\Re^m, y_{ij}^t\in\{-1,+1\},\ 1\leqslant i\leqslant m.$$

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:

© 中国科学院软件研究所 http://www.jos.org.cn

$$g(x)=\sum \psi_i K^{\tilde{\gamma}}(x,x_i)+b^{\phi} \tag{16}$$

其中,ψ_i 为权值系数,$x,x_i\in \boldsymbol{X}^t$,$b^{\phi}$为偏置变量.

1.5 复杂度分析

所提方法 MKLDA 的第 1 步主要采用梯度下降法来迭代地更新基核组合系数,同时学习一个初始的领域适应分类器.整个优化过程主要包括一系列的核机训练,其中每次迭代过程的训练成本实质上等同于标准的 SVM 训练成本.经验分析可知,SVM 的训练复杂度为 $O(n^{2.3})$[22],则 MKLDA 的训练复杂度为 $O(T_{\max}\times n^{2.3})$,其中, $T_{\max}$ 为最大迭代次数,n 为源领域样本数;MKLDA 在进行多核局部重构学习阶段,需要对所有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 $O(m^3)$[19];MKLDA 算法第 3 步训练目标领域 SVM 的复杂度为 $O(m^{2.3})$.综上,MKLDA 整体训练复杂度至多为 $O(T_{\max}\times n^{2.3}+m^3+m^{2.3})$.

1.6 核局部学习目标领域数据集选择

由第 1.5 节分析可知,MKLDA 算法的第 2 步计算复杂度为 $O(m^3)$,这对于目标领域大样本情况的执行效率较低,为此,定义目标领域数据子集:

$$M=\{x|-1\leqslant f(x)\leqslant 1,x\in \boldsymbol{X}^t\}.$$

令 $l=|M|$为集合 M 的基数,则在核分布一致正则化学习后,有 $l<<m$.本文依照文献[4]的假设,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,从而使得多核局部重构学习算法复杂度变为 $O(l^3)<<O(m^3)$,则 MKLDA 的总体计算复杂度变为 $O(T_{\max}\times n^{2.3}+l^3+m^{2.3})<<O(T_{\max}\times n^{2.3}+m^3+m^{2.3})$,从而提升了 MKLDA 算法的可扩展性.

2 方法讨论与分析

2.1 核空间局部邻接图构造

给定数据集 $\boldsymbol{X}$,设 $G(V,E)$代表一个无向加权图,其中:$V=\{v_i\}$为图顶点集,v_i 对应为某个数据点 $x_i\in \boldsymbol{X}$;E 为图边集,且有 $E=\{v_iv_j|v_i$ 为 v_j 邻居或 v_j 为 v_i 邻居,$i\neq j\}$.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

$$\tau_{ij}=\begin{cases}\exp\left(-\dfrac{d^2(x_i,x_j)}{t}\right), & \text{if } x_i\in\delta(x_j) \text{ or } x_j\in\delta(x_i)\\ 0, & \text{otherwise}\end{cases},$$

其中,$t>0$ 为热核参数,$\delta(x_k)$代表数据点 x_k 的 k 近邻集.

2.2 等价稀疏性分析

由公式(11)和 $K^{\tilde{\gamma}}(x,z)$ 的定义可知,当$\gamma_p(1\leqslant p\leqslant P)$为 0 时,第 p 个核和样本特征将被排除,这说明矩阵 $\Lambda_{\tilde{\gamma}}$ 具有信息稀疏的特性,从而有如下定理:

定理 2[19]. 带有单纯形(simplex)约束的加权 l_2 范正则化等价于 l_1 范稀疏正则化,即

$$\inf_{\sum_p \gamma_p=1,\gamma_p\geqslant 0}\sum_p \frac{\|\tilde{W}_p\|_2^2}{\gamma_p}=\left(\sum_p \|\tilde{W}_p\|_1\right)^2, \text{ where } \|\tilde{W}_p\|_1=\sqrt{\sum_{c=1}^{C}\sum_{i=1}^{N}(w_i^c)_p^2}.$$

定理 2 说明,本文带有单纯形加权(γ)的 l_2 范正则化项能够产生至少与 $\left(\sum_p \|\tilde{W}_p\|_1\right)^2$ 正则化项相同的稀疏效果,该特性使得本文方法能在一定程度上有效解决高维数据的领域迁移问题.

2.3 学习风险误差分析

本文方法 MKLDA 的学习风险主要由两个部分组成:初始的领域适应学习风险和局部重构学习误差风险.

© 中国科学院软件研究所 http://www.jos.org.cn

2.3.1 领域适应学习风险

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 的分布概率为 $P(x)$,$x\in X$ 以及标签函数 f:$X\to[0,1]$,定义于实例空间 X 的假设函数类 $\tilde{H}:X\to\{0,1\}$,则假设函数 $h\in\tilde{H}$ 与标签函数 f 之间的差(或假设风险函数$\varepsilon_s(h,f)$)[21]定义为

$$\varepsilon_s(h,f)=E_{x\sim P}[|h(\boldsymbol{x})-f(\boldsymbol{x})|].$$

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

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

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 $\varepsilon_{\tilde{\alpha}}(\bar{h})=\tilde{\alpha}\varepsilon_t(\bar{h})+(1-\tilde{\alpha})\varepsilon_s(\bar{h})]$,对应地,令 $\varepsilon_{\tilde{\alpha}}(h)$ 为真实的组合风险,则领域适应学习经验风险界由定理 3 确定.

定理 3[21]. 设 $\tilde{H}$ 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 $\tilde{\beta}s$ 个样本和源领域 D_s 的 $(1-\tilde{\beta})s$ 个样本,源领域和目标领域标签函数分别为 f^s,f^t.若 $\bar{h}\in\tilde{H}$ 为组合经验风险 $\varepsilon_{\tilde{\alpha}}(\bar{h})$ 在 S 上的经验最小量,且 $h_t^*=\min_{h\in\tilde{H}}\varepsilon_t(h)$ 为目标风险最小量,则至少以概率 $1-\delta$满足下式:

$$\varepsilon_t(\bar{h})\leqslant\varepsilon_t(h_t^*)+2\sqrt{\frac{\tilde{\alpha}^2}{\tilde{\beta}}+\frac{(1-\tilde{\alpha})^2}{1-\tilde{\beta}}}\sqrt{\frac{d\log(2s)-\log\delta}{2s}}+$$
$$2(1-\tilde{\alpha})\left(DIST_k(D^s,D^t)+4\sqrt{\frac{2d\log(2s)+\log(4/\delta)}{s}}+\lambda^*(h)\right).$$

2.3.2 核局部领域适应估计误差风险

假设给定数据点 x_i 及其映射值 f_i,可以通过核岭回归算法来学习一个映射函数,将 x_i 映射为 f_i.根据局部学习思想,f_i 可由 x_i 的邻居点的映射值进行加权重构获得,其局部重构估计误差为

$$E_{local}(f)=\sum_i(f_i-o_i(x_i))^2 \tag{17}$$

其中,$o_i(\cdot)$代表由 x_i 的邻居集 $\{(x_j,f_j)\}_{x_j\in\delta(x_i)}$ 利用核岭回归算法训练学习的某个核机的输出.

2.3.3 MKLDA 学习风险

根据定理 3 和公式(17),所提方法 MKLDA 的目标领域学习误差风险 $\varepsilon_t(\bar{h})$ 至少以概率 $1-\delta$满足下式:

$$\varepsilon_t(\bar{h})\leqslant\varepsilon_t(h_t^*)+2\sqrt{\frac{\tilde{\alpha}^2}{\tilde{\beta}}+\frac{(1-\tilde{\alpha})^2}{1-\tilde{\beta}}}\sqrt{\frac{d\log(2s)-\log\delta}{2s}}+$$
$$2(1-\tilde{\alpha})\left(DIST_k(D^s,D^t)+4\sqrt{\frac{2d\log(2s)+\log(4/\delta)}{s}}+\lambda^*(h)\right)+E_{local}(f).$$

2.4 基核带宽的调节

为了说明高斯核带宽对样本的 RKHS 嵌入分布的影响,首先引出如下定理:

定理 4[18]. 对于一个高斯核函数类 $K_g=\{k_\sigma=\mathrm{e}^{-\|x-z\|_2^2/2\sigma^2},x,z\in R^d,\sigma\in[\sigma_0,\infty)\}$,$\sigma_0>0$,对于任意的 $k_\sigma,k_\tau\in K_g$,$0<\tau<\sigma<\infty$,则 $\gamma_{k_\sigma}(P,Q)\geqslant\gamma_{k_\tau}(P,Q)$.

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 MKLDA 收敛速度减慢.为了进一步研究高斯核带宽对 MKLDA 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

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

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

© 中国科学院软件研究所 http://www.jos.org.cn

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MKLDA算法收敛缓慢.故本文限制$\gamma\in[1,\gamma_0]$,其中,γ_0为一个足够大的待调正参数.由此,MKLDA算法的领域分布差的最小化收敛速度可由参数γ进行协调控制,进一步增强了所提方法的自适应能力.

3 实验分析

为了说明MKLDA方法的有效性,本节将在几个不同类型领域适应学习问题上进行实验:1) 跨领域文本数据集分类;2) 跨领域人脸识别;3) TRECVID视频检测.

实验中,所有源领域均包含有标签数据集,所有目标领域均只包含无标签数据集.将MKLDA与相关的方法进行比较,以显示本文方法的优越性能.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LMPROJ,DASVM和DTMKL,以及文献[23]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)[7]和LWE(locally-weighted ensemble)[23].

作为多核学习方法MKLDA和DTMKL,本文选择4种基核函数:

- 高斯型核函数 $k_{\sigma/\gamma}(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2(1.2^{\delta+3}\sigma/\gamma)^2}\right)$;
- Laplacian核函数 $k_\theta(x_i,x_j)=\exp(-\sqrt{1.2^{\delta+3}\theta}\|x_i-x_j\|)$;
- 多项式核函数 $k_a(x_i,x_j)=(x_i^Tx_j+1)^a$;
- 线性核函数 $k(x_i,x_j)=x_i^Tx_j$.

其中,σ为训练样本平均范数的平方根,$\delta\in\{-3,-2.5,\ldots,2.5,3\}$,$a$=1.5,1.6,…,2.0, $\theta=\frac{1}{d}$(d为数据的特征维数).以此构建33个基核函数.其他核方法均采用标准的RBF核函数$k_\theta(\boldsymbol{x},\boldsymbol{z})=\exp(-\theta|\boldsymbol{x}-\boldsymbol{z}\|^2)$.

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.所有参与比较的方法均通过网格搜索的方式来确定优化的模型参数.

对于MKLDA方法,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$k\in\{5,10,20,50,n_{\min}-1\}$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.

根据经验[35],设置$\lambda=10^{-5}$,正则参数C_1取值范围为{0.1,0.2,0.5,1,2,5,10,20,50,100},采取10重交叉验证法来选取最优值.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方法的学习性能.SVM算法由LIBSVM[22]软件实现,其他算法均在Matlab2009B环境下实现.

3.1 跨领域文本数据分类

本节将在两个文本数据集20Newsgroups(20NG)(http://people.csail.mit.edu/jrennie/20Newsgroups/)和email spam(http://www.ecmlpkdd2006.org/challenge.html)上进行领域适应分类实验,以比较所提方法与相关方法的学习性能.

3.1.1 数据集描述与设置

(1) 20NG数据集:从文献[23]中抽取文本分类数据集20NG.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[5]相同的实验设置.即对于20NG数据集,分别从顶层大类中抽取6个大类以构建学习数据集,其中每两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.20NG数据集的详细信息见表1.

(2) Email Spam数据集:在email spam数据集中有3个email子集(分别表示为User1,User2和User3),以代表3个不同用户.学习的任务是划分出spam邮件和非spam邮件.由于数据集中不同用户的spam邮件和非spam邮件是不同的,因此3个email数据集的数据分布是不同但相关的.每个数据集包含2 500封邮件,其中一半邮件为非spam邮件(类标签为1),另一半spam邮件类标签为–1.根据文献[13]的设置,实验中考虑3种设置,见表1.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 1 Description of the cross domain text datasets

表 1 跨领域分本分类数据描述

	Data sets	Task	Source domain samples		Target domain samples	
			Positive class	Negative class	Positive class	Negative class
1	20NG	Comp vs. Sci	1 958	1 972	2 923	1 977
2		Rec vs. Talk	1 993	1 568	1 984	1 658
3		Rec vs. Sci	1 984	1 977	1 993	1 972
4		Sci vs. Talk	1 971	1 403	1 978	1 850
5		Comp vs. Rec	2 916	1 993	1 965	1 984
6		Comp vs. Talk	2 914	1 568	1 967	1 685
7	Email Spam	User1 vs. User2	User1's emails		User2's emails	
8		User2 vs. User3	User2's emails		User3's emails	
9		User3 vs. User1	User3's emails		User1's emails	

3.1.2 实验结论

文本数据集跨领域学习实验的最优结果记录于表 2.

Table 2 Average classification accuracies (%) of all methods on the text datasets

表 2 所有方法在文本数据集上的分类精度(%)比较

	20NG						Email Spam		
	1	2	3	4	5	6	7	8	9
SVM	72.53	70.10	75.40	78.00	83.80	92.70	96.08	96.89	91.7
TSVM	76.75	73.40	83.90	81.20	85.24	88.74	96.21	97.0	91.80
CDSC	69.80	82.92	64.00	70.84	82.72	90.20	83.28	92.14	90.02
LWE	82.24	78.60	87.20	75.32	88.30	94.00	93.51	98.74	88.78
LMPROJ	82.52	79.30	86.34	84.68	85.40	93.43	93.21	94.0	88.79
DASVM	82.91	81.10	**87.83**	84.55	87.00	94.73	96.89	97.65	94.50
DTMKL	83.93	**84.02**	86.86	85.78	**91.80**	94.30	96.90	**97.70**	94.00
MKLDA	**84.12**	83.40	87.49	**86.50**	89.92	**95.47**	**97.38**	97.34	**94.65**

由表 2 所示结果可知:

- 基线方法 SVM 和 TSVM 因不能有效地迁移到其他领域学习,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;
- 本文方法 MKLDA 和 DTMKL 以及 DASVM 方法在大多数学习任务上均取得了相当的且优于其他方法的学习性能,其中两个多核学习方法 MKLDA 和 DTMKL 学习性能最优;
- 在部分学习任务上,LMPROJ 方法也取得了和 DASVM 方法相当的学习性能.可能的原因是,充分考虑领域间分布距离,能在一定程度上提升领域迁移性能;
- 总体来看,本文所提方法 MKLDA 在两个文本数据集的大多数学习任务上均取得了最优性能.

由此说明,在跨领域学习中,在充分考虑领域间分布距离的基础上采用局部学习思想和多核学习技术,能在一定程度上明显增强跨领域学习性能.

3.2 跨领域人脸识别

为了评价所提方法 MKLDA 在高维数据集下的多类分类性能,分别选取 3 个标准人脸数据库(Yale,FERET 和 ORL)[16,24],按照文献[24]的设置,交叉设置源领域和目标领域数据集,分别实现不同人脸数据集间的迁移学习(如从 ORL 到 YALE 迁移、从 FERET 到 YALE 迁移、从 FERET 到 ORL 迁移、从 YALE 到 FERET 迁移等).

3.2.1 数据集描述与设置

FERET 数据集包含图像大小、姿态、照明和表情不同的 13 539 个人脸图像,分别采样自 1 565 张人脸;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;ORL 人脸数据库有 40 张脸,每张脸包括 10 幅图像.实验前,对上述图像集进行预处理,使其缩放到 32×32 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.对于每个实验任务,在源领域数据集中分别对每人随机选取 8 幅图像作为训练样本集,其他作为测试样本集.根据定理 2,在 Λ_7 矩阵的约束下,所提方法 MKLDA 的正则项具有对

© 中国科学院软件研究所 http://www.jos.org.cn

高维特征提取的功能.为了验证该特性,本节实验中,为了区别设 $\Lambda_{\tilde{\gamma}} = \boldsymbol{I}_{d\times d}$ 对应的方法为 *n*MKLDA(即无稀疏加权正则项方法),而公式(11)对应的方法为 MKLDA(即本文所提带有稀疏加权正则项方法).

3.2.2 实验结论

实验中,对于多类人脸分类问题,本文采取传统的将多类划分为多个二类分类(一对一)的策略,记录最好的识别性能于表 3,其中,AVG 表示在所有识别任务上的平均识别精度率.

Table 3 Recognition rate over different adaptation settings on the face datasets (%)

表 3 对于不同的领域适应设置的人脸数据集上的识别率 (%)

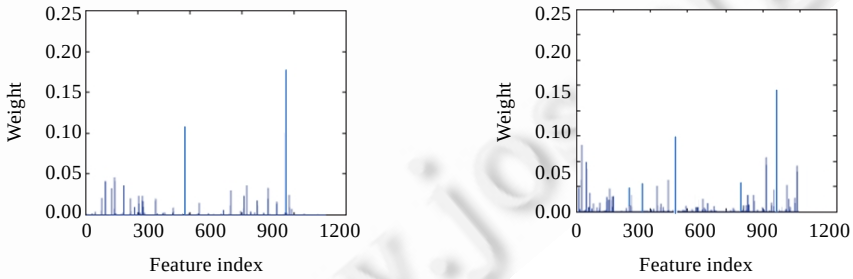
	Task	TSVM	CDCS	LWE	LMPROJ	DASVM	DTMKL	nMKLDA	MKLDA
10	ORL to YALE	28.20	45.00	41.20	44.64	46.78	46.59	48.10	**48.86**
11	FERET to YALE	39.40	42.24	37.82	43.07	43.28	43.78	44.00	**44.26**
12	FERET to ORL	46.21	64.90	44.04	67.72	67.89	**69.56**	69.22	69.40
13	YALE to FERET	27.10	**29.48**	26.92	28.49	28.56	28.98	29.25	29.25
	AVG	35.23	45.41	37.50	45.98	46.63	47.23	47.64	**47.94**

从表 3 可以看出:

- 所有方法的人脸识别误差率均较大,其中,TSVM 和 LWE 方法识别性能最差.可能原因是,3 个人脸数据集间的数据分布差距较大,导致跨领域学习的复杂度增加;
- CDCS 方法在部分识别任务上的性能表现较突出.可能的解释为,谱技术能在一定程度上改善跨领域人脸识别性能;
- DASVM 方法虽然在有些识别任务上的性能与 MKLDA(或 *n*MKLDA)和 DTMKL 相当,但在个别任务上的识别性能甚至略逊于 LMPROJ 方法.可能的原因在于,不同人脸数据库间分布差距较大,导致忽略领域分布间隔度量的 DASVM 方法识别性能下降;
- MKLDA(或 *n*MKLDA)和 DTMKL 在几乎所有识别任务上的性能均明显优于其他方法.

值得指出的是,由于不同人脸间的领域分布差距大小不同,所有方法的识别性能均在不同程度上产生了波动,所提方法 MKLDA(或 *n*MKLDA)和 DTMKL 呈现出了较强的一致性,而 DASVM 方法变化较为明显.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,尤其是 DASVM 在初始阶段就忽略了领域间的分布差距,从而导致识别性能呈现一定的不稳定性;而从充分考虑领域分布间隔的角度,基于多核局部学习思想,所提方法 MKLDA 在所有学习任务上的平均识别率(AVG)明显优于其他方法.

图 2(a)、图 2(b)分别显示在两个人脸识别任务(FERET to ORL 和 ORL to YALE)中,MKLDA 学习到的所有样本特征(1 024 个)所对应的权重.



(a) FERET to ORL (b) ORL to YALE

Fig.2 Feature weight learned by MKLDA

图 2 MKLDA 方法在人脸识别上学习到的特征权重

从图 2 可以看出,对应于大多数特征的权值均接近零值,充分显示了 MKLDA 的稀疏特性,这也进一步证实了定理 2 的有效性.从表 3 也能够看出,MKLDA 方法识别性能在所有情况下均优于或相当于 *n*MKLDA 方法.

© 中国科学院软件研究所 http://www.jos.org.cn

这说明在 MKLDA 方法模型中加入带约束的 l_2 稀疏正则项,能在一定程度上较好地改善领域适应学习性能.

3.3 TRECVID视频检测

3.3.1 数据集描述与设置

TRECVID(http://www-nlpir.nist.gov/projects/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[13],TRECVID 数据集对于领域适应学习具有一定的挑战性[13].按照文献[13]的设置,选取 TRECVID 2005 数据集中的中文频道 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,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,简称 AP)[13]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,为了评价所提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 MKLDA.为了区别,实验中将该方法标记为 MKLDA (l).

3.3.2 实验结论

所有方法在 3 个组(包含 36 个概念)以及总体上的平均分类精度记录于表 4.

Table 4 Performance comparison of different methods in terms of mean AP (%)

表 4 不同方法的 AP(%)性能比较

	SVM	TSVM	LWE	LMPROJ	DASVM	DTMKL	MKLDA	MKLDA (l)
高频组	40.05	41.56	45.77	44.91	46.67	48.50	**48.84**	47.70
中频组	10.11	14.22	14.61	16.09	19.49	19.10	**21.31**	19.52
低频组	4.62	17.13	13.85	16.23	20.14	**20.40**	20.24	19.62
所有组	18.26	24.30	24.74	25.74	28.77	29.33	**30.13**	28.95

从表 4 结果比较可知: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 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,当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[11],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疏,两个领域的数据分布间间隔可能较大,因此,在源领域的样本学习会降低目标领域的学习性能.

另外,从表 4 还可以看出,LMPROJ 的检测性能在大多数情况下优于 LWE,且在个别情况下与 DASVM 相当.这进一步说明,通过明确考虑两个领域间分布一致性,能提升视频概念检测性能.另外,DASVM,DTMKL 和 MKLDA 具有可比较的视频检测性能.但是值得强调的是,所提方法 MKLDA 在 3 组 36 个概念数据集上的 AP 值均优于其他方法.这充分说明,MKLDA 通过明确考虑领域分布间隔最小化和多核局部学习技巧,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;MKLDA(l)方法也取得了与 MKLDA 可比较的学习性能,但是在实验分析中发现,MKLDA(l)的训练时间明显少于 MKLDA 方法.由此,从学习精度和计算复杂度两个方面综合评价,MKLDA(l)方法具备一定的应用性能优势.

3.4 参数敏感实验

所提方法的算法实现需要协调 4 个实验参数:C_1,η,k,γ,实验中取 γ_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 3 个文本(20NG)分类任务和第 10 个人脸识别任务(即 ORL to YALE)作为实验数据,图 3(a)~图 3(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:

(1) 从图 3(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,C_1 值变大,MKLDA 的性能随之趋于上升;

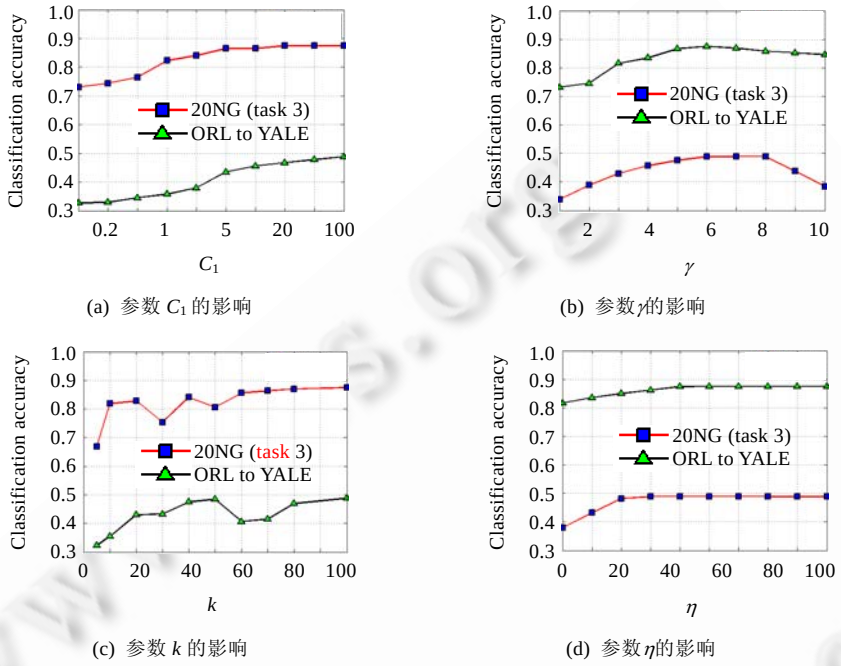
(2) 由图 3(b)可看出,高斯核函数的带宽大小对所提方法的性能影响较突出.由定理 3 可知,γ越小,高斯核带宽越大,领域内数据分布散度越大,导致领域间分布距离最小化过程收敛缓慢,从而使得所提方法

© 中国科学院软件研究所 http://www.jos.org.cn

的学习性能下降;反之,随着γ值的增加,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,所提方法取得了较优的学习性能;

(3) 从图 3(c)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,所提方法能够取得较好的学习性能;

(4) 从图 3(d)可看出,参数η较小时,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.



(a) 参数 C_1 的影响 (b) 参数γ的影响

(c) 参数 k 的影响 (d) 参数η的影响

Fig.3 Parameters sensitivity

图 3 参数敏感性

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,基于多核局部学习技术,在某个多核组合的再生核 Hilbert 空间构建了一种有效的三段式领域迁移学习模型,在不同类型的数据集上的系列实验结果显示了所提方法的优良学习性能.多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,如何优化选取不同的基核函数是本文需要进一步研究的问题.

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

References:

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

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

[3] Joachims T. Transductive inference for text classification using support vector machines. In: Bratko I, Dzeroski S, eds. Proc. of the 16th Int'l Conf. on Machine Learning (ICML'99). Morgan Kaufmann Publishers, 1999. 200−209.

© 中国科学院软件研究所 http://www.jos.org.cn

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

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

[6] Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. In: Proc. of the NIPS. MIT Press, 2007.

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

[8] Dai W, Xue GR, Yu Y. Co-Clustering based classification for out-of-domain documents. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Jose: ACM Press, 2007. 210−219. [doi: 10.1145/1281192.1281218]

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

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

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

[12] Bach FR, Lanckriet GRG, Jordan M. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proc. of the Int'l Conf. on Machine Learning. 2004. [doi: 10.1145/1015330.1015424]

[13] Duan LX, Tsang IW, Xu D. Domain transfer multiple kernel learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012. 465−479. [doi: 10.1109/TPAMI.2011.114]

[14] Vapnik V. Statistical Learning Theory. John Wiley and Sons, 1998.

[15] Wu M, Schölkopf B. A local learning approach for clustering. In: Schölkopf B, Platt J, Hoffman T, eds. Advances in Neural Information Processing Systems 19. Cambridge: MIT Press, 2007. 1529−1536.

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

[17] Hofmann T, Schölkopf B, Smola AJ. Kernel methods in machine learning. Annals of Statistics, 2007,36:1171−1220. [doi: 10.1214/009053607000000677]

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

[19] Zeng H, Cheung YM. Feature selection and kernel learning for local learning-based clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(8):1532−1547. [doi: 10.1109/TPAMI.2010.215]

[20] Yu SX, Shi J. Multiclass spectral clustering. In: Raedt LD, Wrobel S, eds. Proc. of the Int'l Conf. on Computer Vision. ACM Press, 2003. [doi: 10.1109/ICCV.2003.1238361]

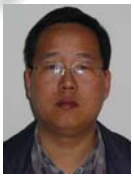
[21] Blitzer J, Crammer K, Kulesza A, Pereira F, Wortman J. Learning bounds for domain adaptation. In: Proc. of the NIPS. MIT Press, 2007.

[22] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[23] Gao J, Fan W, Jiang J, Han J. Knowledge transfer via multiple model local structure mapping. In: Proc. of the 14th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. 2008. [doi: 10.1145/1401890.1401928]

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

陶剑文(1973－),男,湖北武汉人,博士生,副教授,主要研究领域为模式识别,Web挖掘.



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

© 中国科学院软件研究所 http://www.jos.org.cn