文章编号: 1001-0920 (2014) 08-1363-09 DOI: 10.13195/j.kzyjc.2013.0376

面向共享数据的迁移组概率学习机

倪彤光[1,2], 王士同[1], 史荧中[1], 张景祥[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 常州大学 信息科学与工程学院，江苏 常州 213164)

摘 要: 为了解决机器学习中的主观信息缺失问题, 提出一种新的面向共享数据的迁移组概率学习机 (TGPLM-CD). 该方法基于结构风险最小化模型, 将源领域所含知识和目标领域的类标签组概率信息, 特别是领域间的共享数据纳入学习框架中, 实现了源领域和目标领域的知识迁移, 在待研究领域数据信息不足的情况下提高了分类精确度. 大量数据集上的实验结果验证了所提出方法的有效性.

关键词: 迁移学习；分类；支持向量机；共享数据

中图分类号: TP391 **文献标志码:** A

Transfer learning machine based on group probabilities toward common data

NI Tong-guang[1,2], *WANG Shi-tong*[1], *SHI Ying-zhong*[1], *ZHANG Jing-xiang*[1]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. School of Information Science and Technology，Changzhou University，Changzhou 213164，China. Correspondent：NI Tong-guang，E-mail：hbxtntg-12@163.com)

Abstract: To address the problem of man-made information scarcity in the machine learning, a novel transfer learning machine based on group probabilities toward the common data, called TGPLM-CD, is proposed. The proposed method is based on the structure risk minimization model, and both knowledges of the source domain and class label group probabilities are considered as well as the common data between the source domain and the target domain in the learning process, which realizes knowledge transfer between the source domain and the target domain. Experiment results on extensive datasets show the effectiveness of the proposed method.

Key words: transfer learning；classification；support vector machine；common data

0 引 言

近年来, 随着机器学习的实际应用不断加强, 传统机器学习方法面临着一些新的问题. 传统的机器学习方法都是假定源数据域数据信息是充分的[1], 而实际上, 信息采集缺失造成数据信息残缺或不充分的情况大量存在. 通过研究发现, 信息采集缺失主要可分为客观缺失和主观缺失两种情况: 1) 客观信息缺失主要涉及新兴事物或概念, 人们无法在短时间内完成足够数量的新兴领域数据的采集[2]; 2) 主观信息缺失主要指人们为了保护隐私信息而刻意地使用不完整或加密的数据信息来进行系统建模[3-5], 比如在政治选举中, 选民的选票是投给哪位候选人, 疾病诊断中, 病人是否患有疾病等. 基于法律或道德伦理等原因, 在进行机器学习系统建模时, 必须不用或少用这些数据的类标签信息, 这就造成了信息的主观缺失. 针对信息缺失的情景, 众多学者已经探讨了比传统方法更为智能的迁移学习方法[1]. 纵观近几年该理论的发展, 其已成功运用于解决客观信息缺失的许多场景中, 如文献 [6-9] 所述. 而针对主观信息缺失的场景的相关研究还较少且不全面[10-12], 比较有代表性的是文献 [12] 所提出的组概率支持向量机 (IC-SVM), 利用数据的组概率来训练分类器模型, 由于无需使用已标定的数据即可进行模式分类任务, 在一定程度上可以解决主观信息缺失的问题.

本文从迁移学习的角度出发, 提出一种新颖的面向共享数据的迁移组概率学习机 (TGPLM-CD). 首先, 鉴于支持向量机 (SVM)[13]的各种优点, 将其作为基础研究对象来构造迁移学习机; 然后引入迁移学习

收稿日期: 2013-04-01；**修回日期:** 2013-06-27.

基金项目: 国家自然科学基金项目(61272210, 61170122)；江苏省自然科学基金项目(BK2012552).

作者简介: 倪彤光(1978−), 男, 讲师, 博士生, 从事模式识别、人工智能的研究；王士同(1964−), 男, 教授,博士生导师, 从事模式识别、人工智能等研究.

机制, 结合源领域中所学知识与目标领域组概率信息进行分类学习, 同时考虑领域间共享数据对分类决策的影响, 从而构造出新的目标函数分类器, 并证明了所得到的新分类器的求解过程仍然是一个二次规划(QP) 问题. 值得指出的是, 本文方法训练决策模型时仅利用了少量共享的已标定样本及源领域知识(支持向量机参数) 和目标领域数据的类标签概率信息, 所以在主观信息缺失的场景下, 所提出的方法能够在暴露较少数据信息的条件下获得令人满意的分类结果. 这不仅对源数据和目标领域都具有较好的隐私保护性, 同时提高了对目标领域的泛化性能. 因此, 在新场景新问题不断涌现的当今社会, 本文所提出的迁移学习方法具有较高的研究及应用价值.

1 组概率支持向量机 (IC-SVM)

1.1 传统支持向量机

考虑线性可分的分类问题. 设训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}\in(\boldsymbol{x}\times Y)^l$, 其中 $y_i\in Y=\{1,-1\}$, $i=1,2,\cdots,l$. 传统的二类支持向量机[15]的优化目标函数式为

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i. \tag{1}$$

$$\text{s.t. } y(\boldsymbol{w}^{\mathrm{T}}x_i+b)-1+\xi_i\geqslant 0;$$

$$\xi_i\geqslant 0, i=1,2,\cdots,l.$$

其中: C 为正则化参数, ξ_i 为松弛变量.

由式 (1) 可以看出, 只有采集了足够量的已标记样本作为训练集, 上述传统支持向量机的分类精度才会达到满意的效果, 但此时也暴露了数据样本的类标签信息.

1.2 组概率支持向量机

基于上述传统SVM模式分类框架, 文献[12]提出了组概率学习机方法IC-SVM. IC-SVM效仿文献[14]的方法利用Platt模型来标定SVM的输出, 原始公式为

$$p(y=1|\boldsymbol{x})=1/(1+\exp(-Af(\boldsymbol{x})+B)). \tag{2}$$

其中: $\boldsymbol{x}$ 为样本特征向量; y 为样本标签, 且 $y\in\{-1,1\}$; $p(y=1|\boldsymbol{x})$ 为标签为正的概率; A 和 B 为参数, 通过最小交叉熵获得. 在文献[12]中, 取 $A=1$, $B=0$, 得

$$p=\sigma(y)=\frac{1}{1+\exp(-y)}. \tag{3}$$

由反函数变换的相关原理, 式 (3) 变形后可得

$$y=\sigma^{-1}(p)=-\log\left(\frac{1}{p}-1\right), \tag{4}$$

其中 p 为标签为正类样本的概率. 实际应用时为了避免出现无效的 y 值, 文献[12]取 p 为正类标签的组概率, 且限定 $p\in[\varepsilon,1-\varepsilon]$, ε 为分类估计器精度.

实际情况下很难获取每个样本数据所对应的类标签概率, 所以更合理的方式是用组 S_i 中类标签估计的平均值来逼近分组类标签的预测值, 即采用下式所示形式.

$$\forall i:\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\approx y_i. \tag{5}$$

其中: $(\boldsymbol{w},b)$ 为样本分类超平面 $f(x)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j+b$ 的参数对, d 为样本的分组个数. 利用式 (5) 可构建适用于类标签隐藏的分类器优化目标函数, 表示为

$$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}(\xi_i+\xi_i^*). \tag{6}$$

$$\text{s.t. }\forall_{i=1}^{m}:\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j+b)\geqslant y_i-\varepsilon_i-\xi_i;$$

$$\forall_{i=1}^{m}:\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j+b)\leqslant y_i+\varepsilon_i+\xi_i;$$

$$\forall_{i=1}^{m}:\xi_i,\xi_i^*\geqslant 0.$$

上述优化目标函数的解法与传统SVM类似, 详细描述见文献[12].

2 面向共享数据的迁移组概率学习机

本文以结构风险最小化模型为基础, 融合源领域知识, 目标领域组概率信息及领域间共享数据构造了迁移组概率学习方法模型, 其原理如图1所示.

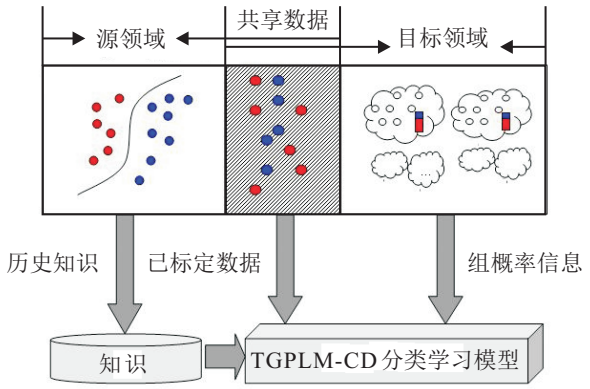


图 1 面向共享数据的迁移组概率学习方法 (TGPLM-CD) 的构造原理

由图1可以看出, 所提方法在构建目标领域分类模型时, 利用从已标定的源领域数据学习而来的分类知识, 目标领域数据类标签组概率信息及少量共享数据共同指导分类器进行决策分类, 最终分类模型从源领域知识, 共享数据和目标领域数据3个方面都抽取对分类效果有益的信息, 因而提高了在主观信息缺失场景下分类器的泛化效果, 同时保护了源领域和目标领域大部分数据的隐私性.

2.1 问题定义

给定源领域数据集SD和目标领域数据集TD, SD和TD包含一部分共享数据CD, 记 $\text{SD}=\{(\boldsymbol{x}_i,y_i)|i=1,2,\cdots,N\}$, $\text{TD}=\{(\boldsymbol{x}_i)|i=1,2,\cdots,M\}$, CD =

$\{(\boldsymbol{x}_i, y_i)|i=1,2,\cdots,n\}$, $\boldsymbol{x}$ 为样本特征向量, y 为分类标签, M、N、n 为样本数量. 本文考虑二分类, $y \in \{-1,1\}$. 设源领域样本服从概率分布 P_s, 记为 $(X_s, Y_s) \sim P_s(X,Y)$; 目标领域样本分布 P_t 未知, 但可确定由 TD 划分成的 d 个子集; $p_k = \dfrac{|\{i \in S_k : y_i = 1\}|}{|S_k|}$ $(k=1,2,\cdots,d)$ 为类条件概率 $P(Y=1|S_k)$ 的概率估计; 源领域和目标领域对应的边缘分布和条件分布记为 $P_s(X), P_t(X)$, P_s 和 P_t 相关但不相同.

2.2　目标函数构造

本文的研究重点是最基本的二元分类问题, 选用 L2-SVMs[15]为所提算法的基本模型. 对于支持向量机, 一类有用的知识可以描述为该支持向量分类机对应的分类超平面参数 $(\boldsymbol{w}, b)$. 因此, 对于某源领域数据受训得到的支持向量机模型, 可把其对应的 $(\boldsymbol{w}_s, b_s)$ 作为已有的可用源领域知识, 也可以作为相似领域间差异的一种度量[16]. 为了从源领域知识和共享数据进行有效地知识迁移, 构造面向共享数据的迁移组概率学习框架下的新目标函数为

$$\begin{aligned}
&\min_{\boldsymbol{w}_t, b_t} \frac{1}{2}\|\boldsymbol{w}_t\|^2 + \frac{C_1}{2}\sum_{i=1}^{n}\eta_i^2 + \\
&\qquad \frac{C_2}{2}\sum_{i=n+1}^{n+d}((\xi_i)^2 + (\xi_i^*)^2) + \frac{\lambda}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_s\|^2. \quad (7)\\
&\text{s.t. } \boldsymbol{w}_t^{\mathrm{T}}\tilde{\boldsymbol{x}}_i + b_t = \boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i + b_s - \eta_i,\ i=1,2,\cdots,n;\\
&\qquad \forall_{i=1}^{d}: \frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j + b_t) \geqslant z_i - \varepsilon_i - \xi_i,\\
&\qquad\qquad i = n+1, n+2, \cdots, n+d;\\
&\qquad \forall_{i=1}^{d}: \frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j + b_t) \leqslant z_i + \varepsilon_i + \xi_i^*,\\
&\qquad\qquad i = n+1, n+2, \cdots, n+d.
\end{aligned}$$

其中: $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)(i=1,2,\ldots,n)$ 表示共享数据; d 表示目标领域数据所分组数, 本文参考文献 [12] 的相关分组策略, 每组样本个数相同; z_i $(i=n+1,n+2,\cdots,n+d)$ 表示由式 (4) 求出的对应每一分组的反向标定输出值; $\boldsymbol{\eta} = [\eta_1, \eta_2, \cdots, \eta_n]^{\mathrm{T}}$ 和 $\boldsymbol{\xi} = [\xi_1, \xi_2, \cdots, \xi_d, \xi_1^*, \xi_2^*, \cdots, \xi_d^*]^{\mathrm{T}}$ 分别表示共享数据和目标领域数据的松弛向量; C_1 和 C_2 表示共享数据和目标领域数据正则化参数(惩罚误差程度).

为了进一步阐述上述优化目标函数的机理, 给出如下的分析与说明.

1) $\dfrac{1}{2}\|\boldsymbol{w}_t\|^2$、$\dfrac{C_1}{2}\sum_{i=1}^{n}\eta_i^2$、$\dfrac{C_2}{2}\sum_{i=n+1}^{n+d}((\xi_i)^2 + (\xi_i^*)^2)$ 分别表示目标领域数据的结构风险项、经验风险项和共享数据的经验风险项.

2) $\dfrac{\lambda}{2}(\|\boldsymbol{w}_t - \boldsymbol{w}_s\|^2)$ 表示目标领域和源领域的差异项, 其大小反映了两个相似领域数据分布的差异程度, 数值越大, 表示分类器之间的差异越大; 反之, 差异越小. 惩罚的程度通过参数 λ 来控制, λ 取较大值时, 源领域与目标领域的分类超平面非常接近; λ 取较小的值时, 源领域与目标领域的分类超平面将相对独立. 这里需要说明的是, 此差异项不包含分类超平面的参数 b, 因为第 1 个约束项同时包含参数 $\boldsymbol{w}$ 和 b, 所以本优化目标函数可以使得源领域和目标领域分类面的参数 b 达到接近的效果, 并且简化了目标函数表达式的复杂性.

3) 第 1 个约束项是为了保证对于源领域和目标领域间的共享数据的分类结果尽可能相同, 后两个约束项表示在整个目标领域中关于数据子集 S_i 的类标签平均概率与由组概率 p_i 得到的反向标定值尽可能接近, 以保证在类标签信息缺失的情况下, 目标领域内分类器尽量保证正确的分类.

4) 因为第 2 和第 3 个约束项中的 ε_i 反映 z_i 的逼近精度, 而实际上组类标签 z_i 是由目标领域中子集 S_i 类标签概率反向标定而得来的, 所以制定一个与目标领域数据类标签概率相关的精度更加恰当. 本文采用与文献 [2] 相同的方法, 取 $\varepsilon_i = \dfrac{\varepsilon'}{p_i(1-p_i)}$. 其中: p_i 为 S_i 中标签为正的数据的组概率, ε' 为一个较小的正常数.

2.3　相关定理推导和证明

根据相关优化理论, 本节对式 (7) 所示的原始问题提出如下几个定理.

定理 1　面向共享数据的迁移组概率学习机 TGPLM-CD 的原始优化问题的对偶问题为

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\beta} + \tilde{\boldsymbol{e}}\boldsymbol{\beta};\\
&\text{s.t.} \boldsymbol{f}^{\mathrm{T}}\boldsymbol{\beta} = 0.
\end{aligned} \quad (8)$$

其中

$$\begin{aligned}
&\boldsymbol{\beta} = [\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*]^{\mathrm{T}};\\
&0 \leqslant \boldsymbol{\beta} \leqslant [\underbrace{C_1, \cdots, C_1}_{n}, \underbrace{C_2, \cdots, C_2}_{d}, \underbrace{C_2, \cdots, C_2}_{d}];\\
&\boldsymbol{f}^{\mathrm{T}} = [\underbrace{1, \cdots, 1}_{n}, \underbrace{1, \cdots, 1}_{d}, \underbrace{-1, \cdots, -1}_{d}];\\
&\tilde{\boldsymbol{e}} = [\boldsymbol{h}, \boldsymbol{\varepsilon} - \boldsymbol{z} + \boldsymbol{g}, \boldsymbol{\varepsilon} + \boldsymbol{z} - \boldsymbol{g}];\\
&h_i = -\frac{1}{1+\lambda}\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i - b_s,\ i=1,2,\cdots,n;\\
&g_i = \frac{\lambda}{1+\lambda}\cdot\frac{1}{|S_i|}\sum_{j\in S_i}\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_j,\\
&\qquad i = n+1, n+2, \cdots, n+d;\\
&\tilde{\boldsymbol{K}} = \frac{\begin{bmatrix} \boldsymbol{K}_{1,1} & -\boldsymbol{K}_{1,2} & \boldsymbol{K}_{1,2} \\ -\boldsymbol{K}_{1,2}^{\mathrm{T}} & \boldsymbol{K}_{2,2} & -\boldsymbol{K}_{2,2} \\ \boldsymbol{K}_{1,2}^{\mathrm{T}} & -\boldsymbol{K}_{2,2} & \boldsymbol{K}_{2,2} \end{bmatrix}_{(n+2d)\times(n+2d)}}{2(1+\lambda)} +
\end{aligned}$$

$$\begin{bmatrix} \dfrac{\delta_{ij}}{C_1} & 0 & 0 \\ 0 & \dfrac{\delta_{ij}}{C_2} & 0 \\ 0 & 0 & \dfrac{\delta_{ij}}{C_2} \end{bmatrix}_{(n+2d)\times(n+2d)};$$

$$\boldsymbol{K}_{1,1} = k(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)_{\substack{i=1,2,\cdots,n \\ j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{1,2} = \Big(\frac{1}{|S_j|}\sum_{j'\in S_j} k(\tilde{\boldsymbol{x}}_i, \boldsymbol{x}_{j'})\Big)_{\substack{i=1,2,\cdots,n \\ j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{2,2} = \Big(\frac{1}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'\in S_j} k(\boldsymbol{x}_{i'}, \boldsymbol{x}_{j'})\Big)_{\substack{i=1,2,\cdots,n \\ j=1,2,\cdots,d}}.$$

这里 $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$

证明 最小值问题 (7) 的拉格朗日函数为

$$L(\boldsymbol{w}_t, b_t, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\xi}^*, \tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*) =$$
$$\frac{1}{2}\|\boldsymbol{w}_t\|^2 + \frac{C_1}{2}\sum_{i=1}^{n}\eta_i^2 +$$
$$\frac{C_2}{2}\sum_{i=n+1}^{n+d}(\xi_i^2 + (\xi_i^*)^2) + \frac{\lambda}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_s\|^2 -$$
$$\sum_{i=1}^{n}\tilde{\alpha}_i(\boldsymbol{w}_t^{\mathrm{T}}\tilde{\boldsymbol{x}}_i + b_t - \boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i - b_s + \eta_i) -$$
$$\sum_{i=n+1}^{n+d}\alpha_i\Big(\frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j + b_t) - z_i + \varepsilon_i + \xi_i\Big) -$$
$$\sum_{i=n+1}^{n+d}\alpha_i^*\Big(z_i + \varepsilon_i + \xi_i^* - \frac{1}{|S_i|}\sum_{j\in S_i}(\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_j + b_t)\Big). \quad (9)$$

其中: $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \tilde{\alpha}_2, \cdots, \tilde{\alpha}_n)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_d)$, $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_d^*)$ 是拉格朗日乘子.

根据 KKT (karush-kuhn-tucker)[17]条件

$$\frac{\partial L}{\partial \eta_i} = 0 \Rightarrow C_1\eta_i = \tilde{\alpha}_i, \quad (10)$$

$$\frac{\partial L}{\partial \xi_i^*} = 0 \Rightarrow C_2\xi_i^* = \alpha_i^*, \quad (11)$$

$$\frac{\partial L}{\partial \boldsymbol{w}_t} = 0 \Rightarrow \boldsymbol{w}_t + \lambda(\boldsymbol{w}_t - \boldsymbol{w}_s) =$$
$$\sum_{i=1}^{n}\tilde{\alpha}_i\tilde{\boldsymbol{x}}_i + \sum_{i=n+1}^{n+d}\frac{\alpha_i - \alpha_i^*}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j, \quad (12)$$

$$\frac{\partial L}{\partial b_t} = 0 \Rightarrow \sum_{i=1}^{n}\tilde{\alpha}_i + \sum_{i=n+1}^{n+d}(\alpha_i - \alpha_i^*) = 0. \quad (13)$$

将式 (10)～(13) 代入 (9), 化简后可得其对偶问题, 表示为

$$\min_{\tilde{\boldsymbol{\alpha}},\boldsymbol{\alpha},\boldsymbol{\alpha}^*}\frac{1}{2(1+\lambda)}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\alpha}_i\tilde{\alpha}_j\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\tilde{\boldsymbol{x}}_j +$$
$$\sum_{i=n+1}^{n+d}\sum_{j=n+1}^{n+d}\frac{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)}{|S_i||S_j|}\sum_{i^{\in}S_i}\sum_{j^{=}S_j}\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{x}_j +$$
$$\sum_{i=1}^{n}\sum_{j=n+1}^{n+d}\frac{\tilde{\alpha}_i(\alpha_j - \alpha_j^*)}{|S_j|}\sum_{k\in S_j}\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{x}_k +$$
$$\sum_{i=1}^{n}\sum_{j=n+1}^{n+d}\frac{\tilde{\alpha}_i(\alpha_j - \alpha_j^*)}{|S_j|}\sum_{k\in S_j}\boldsymbol{x}_k^{\mathrm{T}}\tilde{\boldsymbol{x}}_i\Big) -$$
$$\frac{\sum_{i=1}^{n}\tilde{\alpha}_i\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}{1+\lambda} + \frac{\lambda}{1+\lambda}\sum_{i=n+1}^{n+d}\frac{(\alpha_i - \alpha_i^*)}{|S_i|}\sum_{j\in S_i}\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_j +$$
$$\frac{1}{2C_1}\sum_{i=1}^{n}(\tilde{\alpha}_i)^2 + \frac{1}{2C_2}\sum_{i=n+1}^{n+d}(\alpha_i)^2 +$$
$$\frac{1}{2C}\sum_{i=n+1}^{n+d}(\alpha_i^*)^2 - \sum_{i=1}^{n}\tilde{\alpha}_i b_s + \sum_{i=n+1}^{n+d}\alpha_i(\varepsilon_i - z_i) +$$
$$\sum_{i=n+1}^{n+d}\alpha_i^*(\varepsilon_i + z_i) - \frac{\lambda}{2(1+\lambda)}\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{w}_s. \quad (14)$$
$$\text{s.t. } \tilde{\boldsymbol{\alpha}} \geqslant 0,\ \boldsymbol{\alpha} \geqslant 0,\ \alpha^* \geqslant 0;$$
$$\sum_{i=1}^{n}\tilde{\alpha}_i + \sum_{i=n+1}^{n+d}(\alpha_i - \alpha_i^*) = 0.$$

其中 $\frac{\lambda}{2(1+\lambda)}\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{w}_s$ 为常数项, 不影响对偶问题的极值求解. 因此, 式 (14) 可进一步化为

$$\min_{\tilde{\boldsymbol{\alpha}},\boldsymbol{\alpha},\boldsymbol{\alpha}^*}\frac{1}{2(1+\lambda)}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\alpha}_i\tilde{\alpha}_j\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\tilde{\boldsymbol{x}}_j +$$
$$\sum_{i=n+1}^{n+d}\sum_{j=n+1}^{n+d}\frac{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'=S_j}\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{x}_j +$$
$$\sum_{i=1}^{n}\sum_{j=n+1}^{n+d}\frac{\tilde{\alpha}_i(\alpha_j - \alpha_j^*)}{|S_j|}\sum_{k\in S_j}\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{x}_k +$$
$$\sum_{i=1}^{n}\sum_{j=n+1}^{n+d}\frac{\tilde{\alpha}_i(\alpha_j - \alpha_j^*)}{|S_j|}\sum_{k\in S_j}\boldsymbol{x}_k^{\mathrm{T}}\tilde{\boldsymbol{x}}_i\Big) -$$
$$\frac{\sum_{i=1}^{n}\tilde{\alpha}_i\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i}{1+\lambda} + \frac{\lambda}{1+\lambda}\sum_{i=n+1}^{n+d}\frac{(\alpha_i - \alpha_i^*)}{|S_i|}\sum_{j\in S_i}\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_j +$$
$$\frac{1}{2C_1}\sum_{i=1}^{n}(\tilde{\alpha}_i)^2 + \frac{1}{2C_2}\sum_{i=n+1}^{n+d}(\alpha_i)^2 +$$
$$\frac{1}{2C}\sum_{i=n+1}^{n+d}(\alpha_i^*)^2 - \sum_{i=1}^{n}\tilde{\alpha}_i b_s +$$
$$\sum_{i=n+1}^{n+d}\alpha_i(\varepsilon_i - z_i) + \sum_{i=n+1}^{n+d}\alpha_i^*(\varepsilon_i + z_i). \quad (15)$$
$$\text{s.t. } \tilde{\boldsymbol{\alpha}} \geqslant 0\,, \alpha \geqslant 0, \boldsymbol{\alpha}^* \geqslant 0;$$
$$\sum_{i=1}^{n}\tilde{\alpha}_i + \sum_{i=n+1}^{n+d}(\alpha_i - \alpha_i^*) = 0.$$

为了将式 (15) 化为标准的二次规划形式, 令

$$\boldsymbol{\beta} = [\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*]^{\mathrm{T}};$$
$$0 \leqslant \boldsymbol{\beta} \leqslant [\underbrace{C_1, \cdots, C_1}_{n}, \underbrace{C_2, \cdots, C_2}_{d}, \underbrace{C_2, \cdots, C_2}_{d}];$$
$$\boldsymbol{f}^{\mathrm{T}} = [\underbrace{1, \cdots, 1}_{n}, \underbrace{1, \cdots, 1}_{d}, \underbrace{-1, \cdots, -1}_{d}];$$

$$\tilde{\boldsymbol{e}}=[\boldsymbol{h},\boldsymbol{\varepsilon}-\boldsymbol{z}+\boldsymbol{g},\boldsymbol{\varepsilon}+\boldsymbol{z}-\boldsymbol{g}];$$

$$h_i=-\frac{1}{1+\lambda}\boldsymbol{w}_s^{\mathrm{T}}\tilde{\boldsymbol{x}}_i-b_s,\ i=1,2,\cdots,n;$$

$$g_i=\frac{\lambda}{1+\lambda}\cdot\frac{1}{|S_i|}\sum_{j\in S_i}\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_j,$$

$$i=n+1,n+2,\cdots,n+d;$$

$$\tilde{\boldsymbol{K}}=\frac{\begin{bmatrix}\boldsymbol{K}_{1,1}&-\boldsymbol{K}_{1,2}&\boldsymbol{K}_{1,2}\\-\boldsymbol{K}_{1,2}^{\mathrm{T}}&\boldsymbol{K}_{2,2}&-\boldsymbol{K}_{2,2}\\\boldsymbol{K}_{1,2}^{\mathrm{T}}&-\boldsymbol{K}_{2,2}&\boldsymbol{K}_{2,2}\end{bmatrix}_{(n+2d)\times(n+2d)}}{2(1+\lambda)}+$$

$$\begin{bmatrix}\dfrac{\delta_{ij}}{C_1}&0&0\\0&\dfrac{\delta_{ij}}{C_2}&0\\0&0&\dfrac{\delta_{ij}}{C_2}\end{bmatrix}_{(n+2d)\times(n+2d)};$$

$$\boldsymbol{K}_{1,1}=(\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\tilde{\boldsymbol{x}}_j)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{1,2}=\Big(\frac{1}{|S_j|}\sum_{j'\in S_j}\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{x}_{j'}\Big)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{2,2}=\Big(\frac{1}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'\in S_j}\boldsymbol{x}_{i'}^{\mathrm{T}}\boldsymbol{x}_{j'}\Big)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}}.$$

其中

$$\delta_{ij}=\begin{cases}1,\ i=j;\\0,\ i\neq j.\end{cases}$$

式 (15) 可化成标准的二次规划形式, 如下式所示:

$$\min_{\boldsymbol{\beta}}\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\beta}+\tilde{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{\beta}.$$
$$\text{s.t. }\boldsymbol{f}^{\mathrm{T}}\boldsymbol{\beta}=0.\tag{16}$$

一般情况下, 真实的样本空间很难准确划分, 为此需要进行核化, 其实质是找到一个合适的映射 $\varphi:\boldsymbol{x}_i\in R^d\to\varphi(\boldsymbol{x}_i)\in R^D(d<<D)$, 并用核函数 $k(\mu,v)$ 表示映射后的内积 $\varphi(\mu)^{\mathrm{T}}\varphi(v)$, 令

$$\boldsymbol{K}_{1,1}=k(\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_j)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{1,2}=\Big(\frac{1}{|S_j|}\sum_{j'\in S_j}k(\tilde{\boldsymbol{x}}_i,\boldsymbol{x}_{j'})\Big)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{2,2}=\Big(\frac{1}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'\in S_j}k(\boldsymbol{x}_{i'},\boldsymbol{x}_{j'})\Big)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}}.$$

式 (16) 核化后可得式 (8). □

由式 (8) 可知, 所提迁移学习系统的优化问题可转化成一个二次规划问题. 根据优化理论, 式 (8) 中的核函数 $\tilde{\boldsymbol{K}}(\cdot)$ 只有在保证是 Mercer 核时, 才能保证其是二次凸规划, 使得其求得的解为全局最优解. 为了验证这一问题, 本文给出如下定理.

引理 1[18]　令 X 是 R^n 上的一个紧集, $\varphi(x,z)$ 是 Mercer 核, 当且仅当 $\varphi(x,z)$ 是 $X\times X$ 上的连续对称函数, 且关于任意的 $x_1,x_2,\cdots,x_n\in X$ 的 Gram 矩阵半正定.

定理 2　式 (8) 所表达的核函数是 Mercer 核.

证明　由式 (8) 的定义可知, $\tilde{\boldsymbol{K}}$ 矩阵是一个对称矩阵.

若要证明形如 $\tilde{\boldsymbol{K}}$ 的核是 Mercer 核, 则需证明 $\tilde{\boldsymbol{K}}$ 为半正定矩阵. 可以看出, 若式 (17) 所示矩阵 $\boldsymbol{H}$ 为半正定矩阵, 则 $\tilde{\boldsymbol{K}}$ 即为半正定矩阵.

$$\boldsymbol{H}=\begin{bmatrix}\boldsymbol{K}_{1,1}&-\boldsymbol{K}_{1,2}&\boldsymbol{K}_{1,2}\\-\boldsymbol{K}_{1,2}^{\mathrm{T}}&\boldsymbol{K}_{2,2}&-\boldsymbol{K}_{2,2}\\\boldsymbol{K}_{1,2}^{\mathrm{T}}&-\boldsymbol{K}_{2,2}&\boldsymbol{K}_{2,2}\end{bmatrix}_{(n+2d)\times(n+2d)}.\tag{17}$$

其中

$$\boldsymbol{K}_{1,1}=(\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_j)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{1,2}=\Big(\frac{1}{|S_j|}\sum_{j'\in S_j}(\tilde{\boldsymbol{x}}_i,\boldsymbol{x}_{j'})\Big)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}};$$

$$\boldsymbol{K}_{2,2}=\Big(\frac{1}{|S_i||S_j|}\sum_{i'\in S_i}\sum_{j'\in S_j}(\boldsymbol{x}_{i'},\boldsymbol{x}_{j'})\Big)_{\substack{i=1,2,\cdots,n\\j=1,2,\cdots,d}}.$$

令

$$\boldsymbol{Q}=\Big(y_1\tilde{\boldsymbol{x}}_1,\cdots,y_n\tilde{\boldsymbol{x}}_n,-\frac{1}{|S_1|}\sum_{i\in S_1}\boldsymbol{x}_i,\cdots,-\frac{1}{|S_d|}\sum_{i\in S_d}\boldsymbol{x}_i,\frac{1}{|S_1|}\sum_{i\in S_1}\boldsymbol{x}_i,\cdots,\frac{1}{|S_d|}\sum_{i\in S_d}\boldsymbol{x}_i\Big),$$

则 $\boldsymbol{H}=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}$, 所以 H 是半正定矩阵. 再由引理 1 的描述, 可得形如式 $\tilde{\boldsymbol{K}}$ 的核是 Mercer 核. □

引理 2[18]　假设二次规划中的 Gram 矩阵为半正定矩阵, 则该二次规划为凸二次规划.

引理 3[18]　假设二次规划为凸二次规划, 则 KKT 条件也是充分条件, 因此得到的二次规划的解为全局最优解.

定理 3　设 $\hat{\boldsymbol{\beta}}=[\hat{\tilde{\boldsymbol{\alpha}}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\alpha}}^*]$ 是对偶问题 (8) 的解, 则在线性条件下, TGPLM-CD 的原始优化问题 (7) 对于 $\boldsymbol{w}_t$ 和 b_t 的解存在全局最优解, 并可以表示为

$$\boldsymbol{w}_t^*=\frac{1}{1+\lambda}\Big(\sum_{i=1}^{n}\hat{\tilde{\alpha}}_i\tilde{\boldsymbol{x}}_i+\sum_{i=n+1}^{n+d}\frac{\hat{\alpha}_i-\hat{\alpha}_i^*}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j+\lambda\boldsymbol{w}_s\Big),\tag{18}$$

$$b_t^*=z_i-\frac{\lambda}{1+\lambda}\sum_{j\in S_i}\frac{\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{x}_j}{|S_i|}-\frac{1}{1+\lambda}\sum_{j=1}^{n}\frac{\hat{\tilde{\alpha}}_j}{|S_i|}\sum_{k\in S_i}\tilde{\boldsymbol{x}}_j^{\mathrm{T}}\boldsymbol{x}_k-\frac{1}{1+\lambda}\sum_{j=n+1}^{n+d}\frac{\hat{\alpha}_j-\hat{\alpha}_j^*}{|S_j||S_i|}\sum_{l\in S_j}\sum_{k\in S_i}\boldsymbol{x}_l^{\mathrm{T}}\boldsymbol{x}_k.\tag{19}$$

证明　由引理 2 以及定理 2 的证明可知式 (8) 为凸二次规划, 又由引理 3 的满足条件可知该二次规划的解为全局最优解. 因此, 若 $\hat{\boldsymbol{\beta}}=[\hat{\tilde{\boldsymbol{\alpha}}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\alpha}}^*]$ 为式 (8)

的解, 则在线性情况下, 由式(12)可得如(18)所示$\boldsymbol{w}_t$的最优解$\boldsymbol{w}_t^*$. 选取n个$\hat{\tilde{\alpha}}_i$位于开区间$(0, C_1)$, d个$\hat{\alpha}_i$、$\hat{\alpha}_j^*$位于开区间$(0, C_2)$的分量$(\hat{\tilde{\alpha}}_j, \hat{\alpha}_j, \hat{\alpha}_j^*)^{\mathrm{T}}$, 由此可以计算得到$b_t$的最优解$b_t^*$. □

对于式(18)和(19)所给出的最优解同时包含了从源领域、目标领域和共享数据中获取的知识, 如: ${\boldsymbol{w}_t}^*$中$\dfrac{\lambda}{1+\lambda}\boldsymbol{w}_s$部分为从源领域中学习得到的知识; $\dfrac{1}{1+\lambda}\displaystyle\sum_{i=1}^{n}\hat{\tilde{\alpha}}_i\tilde{\boldsymbol{x}}_i$部分为从共享数据中学习得到的知识; $\dfrac{1}{1+\lambda}\displaystyle\sum_{i=n+1}^{n+d}\frac{\tilde{\alpha}_i-\tilde{\alpha}_i^*}{|S_i|}\sum_{j\in S_i}\boldsymbol{x}_j$部分为从目标领域中学习得到的知识.

在非线性情况下, 根据Represent Theorems[17]可将源领域决策超平面法向量$\boldsymbol{w}_s$表示为

$$\boldsymbol{w}_s=\sum_{i=1}^{N}\alpha_i^s\varphi(\boldsymbol{x}_i).$$

其中: $\boldsymbol{\alpha^s}=[\alpha_1^s,\cdots,\alpha_N^s]^{\mathrm{T}}$为权值矢量, 非线性函数$\varphi(\cdot)$为源领域样本空间到特征空间的映射函数. 因此, 非线性情况下原始问题(7)的解可表示为

$$\begin{aligned}\boldsymbol{w}_t^*=\frac{1}{1+\lambda}\Big(&\sum_{i=1}^{n}\hat{\tilde{\alpha}}_i\varphi(\tilde{\boldsymbol{x}}_i)+\lambda\sum_{i=1}^{N}\alpha_i^s\varphi(\boldsymbol{x}_i)+\\&\sum_{i=n+1}^{n+d}\frac{\hat{\alpha}_i-\hat{\alpha}_i^*}{|S_i|}\sum_{j\in S_i}\varphi(\boldsymbol{x}_j)\Big),\end{aligned}\tag{20}$$

$$\begin{aligned}{b_c}^*=z_i-&\frac{\lambda}{1+\lambda}\sum_{j=1}^{N}\frac{\alpha_j^s}{|S_i|}\sum_{k\in S_i}k(\boldsymbol{x}_j,\boldsymbol{x}_k)-\\&\frac{1}{1+\lambda}\sum_{j=1}^{n}\frac{\hat{\tilde{\alpha}}_j}{|S_i|}\sum_{k\in S_i}k(\tilde{\boldsymbol{x}}_j,\boldsymbol{x}_k)-\\&\frac{1}{1+\lambda}\sum_{j=n+1}^{n+d}\frac{\hat{\alpha}_j-\hat{\alpha}_j^*}{|S_j||S_i|}\sum_{l\in S_j}\sum_{k\in S_i}k(\boldsymbol{x}_l,\boldsymbol{x}_k).\end{aligned}\tag{21}$$

2.4 TGPLM-CD算法流程

由2.3节所作的推导和分析可得到TGPLM-CD方法的具体步骤, 表示如下.

Input: N个有标号的源领域样本$\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$; M个无标号的目标领域样本$\{\boldsymbol{x}_j\}_{j=N+1}^{N+M}$; 目标领域数据分为$d$个组, 目标领域组概率$\{(S_k,p_k\}_{k=1}^d$; 源领域和目标领域共有的$n$个有标号数据$\{(\tilde{\boldsymbol{x}}_i,\tilde{y}_i)\}_{i=1}^n$.

Output: 目标领域的分类决策函数$f(x)$.

1) 源领域知识总结阶段.

Step 1: 设置核函数带宽σ_s和正则参数C_s;

Step 2: 利用SVM得到源领域数据的分类模型;

Step 3: 求解拉格朗日系数$\boldsymbol{\alpha}^s$.

2) 目标领域迁移学习阶段.

Step 1: 由式(3)和$\{S_k,p_k\}_{k=1}^d$计算当前领域数据的分组的反向标定输出$z_k, k=1,2,\cdots,d$;

Step 2: 由$\{(\tilde{\boldsymbol{x}}_i)\}_{i=1}^n$、$\{\boldsymbol{x}_j\}_{j=N+1}^{N+M}$和$\{(S_k,p_k\}_{k=1}^d$计算核矩阵$\boldsymbol{K}_{1,1}$, $\boldsymbol{K}_{1,2}$, $\boldsymbol{K}_{2,2}$;

Step 3: 由定理1构造矩阵$\tilde{\boldsymbol{K}}$, 求解拉格朗日系数$\boldsymbol{\beta}$.

3) 迁移学习机生成阶段.

Step 1: 由式(21)计算偏移量b_t;

Step 2: 输出分类决策函数$f(\boldsymbol{x})=\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{x}_i+b_t$.

3 实验与分析

为了讨论本文所提方法TGPLM-CD在针对不同领域迁移分类学习问题上的有效性, 本节将在几个不同类型的数据集上进行实验: 1) 具有共享数据的人造双月型二维数据集; 2) 人脸图像分类数据集PIE[19]; 3) 不同应用领域的真实UCI数据集[20].

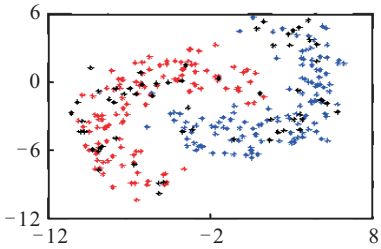
在实验中主要引入SVM[10]、TSVM[21]、LWE[9] TrSVM[7]、LMPROJ[8]、IC-SVM[12]六种算法与本文所提方法进行比较: 与SVM和TSVM的比较用以验证所提方法在迁移学习问题上与传统基于IID假设的支持向量机和直推式支持向量机分类方法的优势; TrSVM、LWE和LMPROJ均为迁移学习分类方法, 用以说明本文所提方法在目标领域类标签保护的前提下与其他领域自适应方法具有可比较的性能; IC-SVM为仅考虑由目标领域中所获信息的针对类标签保护数据集的分类方法, 二者的比较说明本文所提方法TGPLM-CD在通过迁移学习和融合共享数据带来的性能提升.

在本文方法与其他方法进行学习能力比较时, 以目标域数据分类的精度为评价指标, 具体的指标为Accuracy $=\dfrac{|\{\boldsymbol{x}|\boldsymbol{x}_t\in D_t\cap f(\boldsymbol{x}_i)=y_t\}|}{|\{\boldsymbol{x}|\boldsymbol{x}_t\in D_t\}|}$, 核函数中的$2\sigma^2$选择以源领域样本的平均2范数的平方$s$为基准, 并在网格$\{s/64, s/32, s/16, s/8, s/4, s/2, s, 2s, 4s, 8s, 16s, 32s, 64s\}$中搜索, 直至最优; TGPLM-CD的正则化参数C_1和C_2在网格$\{2^{-10}, 2^{-9}, 2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}\}$中搜索, 直至最优; 平衡参数$\lambda$在区间$\{10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}\}$中搜索, 直至最优. 本文所有实验均通过网格搜索的方式来确定优化的实验参数. 实验中, 在采用网格搜索方法选择参数时, 判定参数性能的标准采用文献[12]相同的策略, 在训练集上以10倍交叉验证时所得最优参数下的分类精度为评价指标. 所有实验均在Intel Core 2, 2.0 GHz主频, 3G RAM, Windows XP系统, Matlab 2009a平台实现. 实验中, SVM算法由Libsvm[22]软件实现, 其他算法均在Matlab R2009A环境下实现.

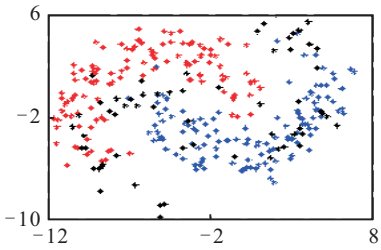
对于所有数据集, 源域和目标域数据均具有标签信息, 但目标域标签信息仅用于学习方法分类性能的客观量化评价.

3.1　人造双月型数据集

构造均值为 0, 标准差为 1, 正负类各 150 个样本的原始双月型数据集来建立源领域和目标领域数据集, 选取全部 150 个负类样本和 150 个正类样本构成源领域数据集, 如图 2(a) 所示. 将该源领域数据集围绕所有样本的中心顺时针方向旋转 4 次, 每次旋转 10°, 从而得到 4 个不同分布的目标领域数据集. 图 2(b) 描述了源领域数据集在旋转 30° 后的形状. 可以看出, 旋转角度越大, 目标领域与源领域分布差异越大. 针对本文所设场景, 需要特别说明的是源领域和目标领域包含相同的 20% 数据, 见图 2 中深黑色点, 此部分点不参与旋转, 用以模拟共享数据.



(a) 旋转 0° 源领域双月数据集



(b) 转 30° 的目标领域双月数据集

图 2　源领域和目标领域双月型数据集 (黑色部分为共享数据, 所占比例为 20%)

表 1 为共享数据分别占 10% 和 20% 的比例时不同类型方法在 4 个目标领域上的分类精度, 目标领域分组数为 50; 图 3 为共享数据占 10% 比例且目标领域分组数为 50 时, 在旋转 30° 的目标领域上分类精度随平衡参数 $\lambda(10^x)$ 的变化趋势.

表 1　不同方法在双月型数据集上的性能比较 (粗体表示最优结果)

算法	10%					20%				
	0	10	20	30	40	0	10	20	30	40
SVM	**100.00**	98.50	90.50	81.20	70.50	**100.00**	98.80	91.30	82.20	73.80
TSVM	**100.00**	**99.50**	90.80	81.60	70.80	**100.00**	**99.60**	91.90	84.90	73.90
LMPROJ	**100.00**	**99.50**	94.10	87.10	81.10	**100.00**	**99.60**	94.66	88.15	81.67
TrSVM	**100.00**	**99.50**	93.80	87.00	80.00	**100.00**	**99.60**	95.00	88.10	80.30
LWE	**100.00**	**99.50**	93.60	88.20	81.00	**100.00**	**99.60**	94.80	88.75	81.10
IC-SVM	89.86	89.86	89.87	89.86	89.86	89.90	89.88	89.90	89.90	89.89
TGPLM-CD	**100.00**	**99.50**	**95.15**	**91.25**	**89.95**	**100.00**	**99.60**	**96.00**	**92.80**	**90.10**

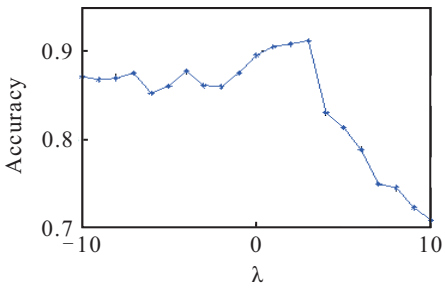


图 3　平衡参数的敏感性

1) 从表 1 可以看出: 由于源领域和目标领域呈现不同分布, 迁移方法的实验效果明显优于基于 IID 假设的 SVM, 特别的, 由于 IC-SVM 仅使用目标领域的类标签组概率信息, 其精度仅与目标领域的分组数存在联系, 所以源领域的变化对其精度几乎没有影响; 所提方法融合了共享数据和源领域知识, 其分类性能优于其他几类迁移方法; 随着目标领域旋转角度的不断增大, 5 类算法的学习性能均呈现下降趋势, 这说明源领域与目标领域分布的差异程度将严重影响目标领域的分类效果; 领域间共享的数据量越大, 意味着源领域与目标领域越相似, 所以几种迁移算法分类性能都有所提高, 这也说明共享数据对于迁移学习的有效性, 且正是因为所提方法在构造目标函数时考虑了共享数据的积极作用, 从而拥有相比其他迁移方法更高的性能提升. 由表 1 所提方法与 IC-SVM 的比较还可以看出, 本文所采用的迁移学习框架在主观信息缺失情况下是非常有效的.

2) 从图 3 可以看出, 平衡参数 λ 对双月型数据集分类精度有影响. 当 λ 取值较小时, 所提方法只依赖目标领域的类标签概率信息和共享数据进行知识学习, 所以精度只比 IC-SVM 略高; 随着 λ 取值的增大, 所提方法在借鉴目标领域的类标签组概率信息的同时学习了源领域的辅助知识, 从而获得了最优分类效果; 当 λ 取值很大时, 所提方法过分依赖了源领域的辅助知识, 忽视了源领域与目标领域数据分布的差异, 使得目标领域和源领域的决策超平面被强制性趋同, 导致了负迁移, 使得新方法的分类性能下降. 实验中还发现, $2\sigma^2$ 控制在区间 $[s, 32s]$, 正则参数 C_1 和 C_2 取相同数值, 且在 $[2^1, 2^{11}]$ 时, 算法能取得较好的效果. 具体的实验结果见表 1.

3.2 真实数据集

为了进一步说明所提方法的优越性, 本节将在PIE人脸数据集和UCI数据集上与相关算法进行性能比较, 以充分验证本文方法的有效性. 实验为模拟主观信息缺失的场景, 将目标领域数据集组内样本数目定为8, 对应的最终得到的组数由各个数据集采用下取整策略确定. 所有实验中目标领域均包含源领域样本个数的20% 作为共享数据.

1) PIE人脸数据集. PIE数据库包含41 368幅属于68个人的人脸灰度图像, 随机选取1名男性和1名女性每人各170幅的人脸图像, 构成一个二类数据集进行实验. 对上述图像数据选取样本, 并分别进行逆时针旋转10°、30°、50°, 以形成变化的3个迁移学习目标领域图像数据集. 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, 则每幅图像在图像空间中由一个1 024维的向量表示. 图4(a)、图4(b)分别显示了旋转前后的部分图像, 对应的男性标签为1号, 女性标签为35号. 表2给出了不同算法在PIE数据集上的实验结果.

(a) 源领域人脸样本

(b) 逆时针旋转10° 的目标领域人脸样本

图 4 基于PIE人脸数据库构造的源领域及目标领域样本

表 2 PIE数据集上各方法的性能比较

PIEDataset	SVM	TSVM	TrSVM	LWE	LMPROJ	IC-SVM	TGPLM-CD
10°	73.59	74.84	75.24	75.86	77.63	72.02	**81.33**
30°	67.31	69.80	71.17	72.87	73.20	71.98	**78.76**
50°	57.66	63.80	64.72	67.61	68.80	71.96	**72.55**

2) UCI数据集. 本节将在UCI数据集上测试所提方法的分类精度. 抽取6个2分类数据集Australia、Breast-cancer、Vote、Ionoshere、Heart-c、Sonar来分别测试所提方法与比较方法的分类效果. 数据集详细信息如表3所示. 实验中随机抽取样本总数的各30% 作为源领域和目标域领域数据. 表4给出了不同算法在各个UCI数据集上的实验结果.

表 3 UCI数据集描述

学习任务	数据集	样本总数	属性数
1	Australia	690	14
2	Breast-cancer	683	10
3	Vote	435	16
4	Ionoshere	351	34
5	Heart-c	303	23
6	Sonar	208	60

表 4 各种方法在不同UCI数据集上的性能比较

Task	SVM	TSVM	TrSVM	LWE	LMPROJ	IC-SVM	TGPLM-CD
1	73.59	75.84	81.21	81.16	81.63	80.16	**82.83**
2	80.31	86.80	90.15	91.03	90.90	90.87	**91.76**
3	87.66	88.80	90.22	90.31	90.40	89.61	**90.55**
4	89.35	89.89	91.92	92.11	91.60	90.33	**94.01**
5	80.05	85.55	86.12	85.56	86.10	86.00	**88.04**
6	78.67	83.15	85.52	84.61	85.90	85.24	**88.93**

根据6类方法在真实数据集上的实验结果, 可得到如下结论:

1) 因为源领域和目标领域样本数占样本总数的比例较低, 传统的SVM由于仅考虑在源领域分类最优, 所以无法在目标领域上得到较高的分类精度, 因此分类精度均低于其余几类方法;

2) 由于充分考虑了共享数据及由源领域中获取的知识, 虽然存在类标签隐藏这种主观信息缺失, TGPLM-CD的分类精度均优于其他几类迁移学习方法.

3) 由于所提方法采用了迁移学习框架, 在主观信

息缺失的情况下, 与IC-SVM相比所提方法依然可以达到可利用的精度, 并且在共享数据所占比例很少的情况下, 所提方法优势依然较为明显.

4 结　论

针对当前领域仅已知类标签概率信息的模式分类问题, 本文将源领域的知识和共享数据同时纳入目标决策函数的构造, 同时结合结构风险最小化模型, 提出了一种有效的迁移组概率学习机TGPLM-CD. 通过在人工和真实数据集实验上的结果可以看出, TGPLM-CD方法具备较好的泛化能力, 可在一定程度上弥补主观信息缺失带来的精度损失. 本文方法将在多分类问题和大样本问题这两个方面展开进一步研究.

参考文献(References)

[1] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[2] 蒋亦樟, 邓赵红, 王士同. ML型迁移学习模糊系统[J]. 自动化学报, 2012, 38(9): 1393-1409
(Jiang Y Z, Deng Z H, Wang S T. Mamdani-Larsen type transfer learning fuzzy system[J]. Acta Automatica Sinica, 2012, 38(9): 1393-1409.)

[3] 胡文军, 王士同. 隐私保护的SVM快速分类方法[J]. 电子学报, 2012, 40(2): 280-286.
(Hu W J, Wang S T. Fast classification approach of support vector machine with privacy preservation[J]. Acta Electronica Sinica. 2012, 40(2): 280-286.)

[4] 张战成, 王士同, 钟富礼. 具有隐私保护功能的协作式分类机制[J]. 计算机研究与发展, 2011, 48(06): 1018-1029.
(Zhang Z C, Wang S T, Zhong F L. Collaborative classification mechanism for privacy-preserving[J]. J of Computer Research and Development, 2011, 48(6): 1018-1028.)

[5] Quadrianto N, Smola A J, Caetano T S, et al. Estimating labels from label proportions[J]. J of Machine Learning Research, 2009, (10): 2349-2374.

[6] Tao J W, Chung F L, Wang S T. On minimum distribution discrepancy support vector machine for domain adaptation[J]. Pattern Recognition, 2012, 45(11): 3962-3984.

[7] 洪佳明, 印鉴, 黄云, 等.一种基于领域相似性的迁移学习算法[J]. 计算机研究与发展, 2011, 48(10): 1823-1830.
(Hong J M, Yin J, Huang Y, et al. TrSVM: A transfer learning algorithm using domain similarity[J]. J of Computer Research and Development, 2011, 48(10): 1823-1830.)

[8] Quanz B, Huan J. Large margin transductive transfer learning[C]. Proc of the 18th ACM Conf on Information and Knowledge Management. New York, 2009: 1327-1336.

[9] Gao J, Fan W, Jiang J, et al. Knowledge transfer via multiple model local structure mapping[C]. Proc of the 14th ACM SIGKDD Inter Conf on Knowledge Discovery and Data Mining. New York, 2008: 283-291.

[10] Stolpe M, Morik K. Learning from label proportions by optimizing cluster model selection[C]. Proc of Machine Learning and Knowledge Discovery in Databases-European Conference 2011. Berlin: Heidelberg, 2011: 349-364.

[11] Quadrianto N, Smola A J, Caetano T S, et al. Estimating labels from label proportions[C]. The 25th Int Conf on Machine Learning. Omnipress, 2008: 776-783.

[12] Rüping S. SVM classifier estimation from group probabilities[C]. The 27th Int Conf on Machine Learning. Haifa, 2010: 911-918.

[13] Vapnik V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995: 123-167.

[14] Platt J C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[C]. Advances in Large Margin Classifiers. Cambridge: MIT Press, 1999: 61-74.

[15] Tsang I W , Kwok J T, Zurada J M. Generalized core vector machines[J]. IEEE Trans Neural Network, 2006, 17(5): 1126-1140

[16] Guillermo L G, Lucas C U, Alejandro C H, et al. Solving nonstationary classification problems with coupled support vector machines[J]. IEEE Trans on Knowledge on Neural Network, 2011, 22(1): 37-51

[17] Scholkopf B, Herbrich R, Smola A J. A generalized representer theorem[C]. Proc of Conf on Learning Theory. Amsterdam: Springer Press, 2001: 416-426.

[18] 邓乃杨, 田英杰. 数据挖掘的新方法—支持向量机[M]. 北京: 科学出版杜, 2004.
(Deng N Y, Tian Y J. New method in data mining: Support vector machine[M]. Beijing: Science Press, 2004.)

[19] He X F, Cai D, Partha N. Laplacian score for feature selection[J]. Advances in Neural Information Proc System, 2006(18): 507-514 .

[20] Asuncion A, Newman D J. UCI machine learning repository[DB/OL]. [2008-11-01]. http://archive.ics.uci.edu/ml/.

[21] Joachims T. Transductive inference for text classification using support vector machines[C]. Proc of 16th Int Conf on Machine Learning. San Francisco, Morgan Kaufmann Publishers, 1999: 200-209.

[22] Chang C C, Lin C J. LIBSVM: A library for support vector machines[EB/OL]. [2008-11-04]. http://www.csie.ntu.edu.tw/cjlin/libsvm.

(责任编辑: 齐　霁)