软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(10):2312−2326 [doi: 10.3724/SP.J.1001.2013.04375]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

大样本领域自适应支撑向量回归机*

许 敏[1,2], 王士同[1], 顾 鑫[1,3], 俞 林[2]

[1](江南大学 数字媒体学院,江苏 无锡 214122)

[2](无锡职业技术学院 物联网技术学院,江苏 无锡 214121)

[3](无锡北方湖光光电有限公司 研发部,江苏 无锡 214035)

通讯作者: 王士同, E-mail: wxwangst@yahoo.com.cn

摘 要: 针对回归问题中存在采集数据不完整而导致预测性能降低的情况,根据支撑向量回归机(support vector regression,简称 SVR)等价于中心约束最小包含球(center-constrained minimum enclosing ball,简称 CC-MEB)以及相似领域概率分布差异只与两域各自的最小包含球中心点位置有关的理论新结果,提出了针对大数据集的领域自适应核心集支撑向量回归机(adaptive-core vector regression,简称 A-CVR).该算法利用源域 CC-MEB 中心点对目标域 CC-MEB 中心点进行校正,从而提高目标域的回归预测性能.实验结果表明,这种领域自适应算法可以弥补目标域缺失数据的不足,大大提高回归预测性能.

关键词: 领域自适应;支撑向量回归;核心集支撑向量机;中心约束最小包含球;大数据集

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 许敏,王士同,顾鑫,俞林.大样本领域自适应支撑向量回归机.软件学报,2013,24(10):2312−2326. http://www.jos.org.cn/1000-9825/4375.htm

英文引用格式: Xu M, Wang ST, Gu X, Yu L. Support vector regression for large domain adaptation. Ruan Jian Xue Bao/Journal of Software, 2013,24(10):2312−2326 (in Chinese). http://www.jos.org.cn/1000-9825/4375.htm

Support Vector Regression for Large Domain Adaptation

XU Min[1,2], WANG Shi-Tong[1], GU Xin[1,3], YU Lin[2]

[1](School of Digital Media, Jiangnan University, Wuxi 214122, China)

[2](School of Internet of Things Technology, Wuxi Institute of Technology, Wuxi 214121, China)

[3](Research and Development Department, Wuxi Northern Lake Optical Co., Ltd., Wuxi 214035, China)

Corresponding author: WANG Shi-Tong, E-mail: wxwangst@yahoo.com.cn

Abstract: Incomplete data collection in regression analysis would lead to low prediction performance, which aises the issue of domain adaptation. It is well known that support vector regression (SVR) is equivalent to center-constrained minimum enclosing ball (CC-MEB). Also in solving the problem of how to effectively transfer the knowledge between the two fields, new theorems reveal that the difference between two probability distributions from two similar domains only depends on the centers of the two domains' minimum enclosing balls. Based on these developments, a fast adaptive-core vector regression (A-CVR) algorithm is proposed for large domain adaptation. The proposed algorithm uses the center of the source domain's CC-MEB to calibrate the center of the target domain's in order to improve the regression performance of the target domain. Experimental results show that the proposed domain adaptive algorithm can make up for the lack of data and greatly improve the performance of the target domain regression.

Key words: domain adaptation; support vector regression (SVR); core vector machine (CVM); center-constrained minimum enclosing ball (CC-MEB); large data set

* 基金项目: 国家自然科学基金(61170122, 61272210); 江苏省研究生创新工程项目(CXZZ12-0759)
收稿时间: 2011-09-02; 修改时间: 2012-09-29; 定稿时间: 2013-01-25

机器学习大多由来自某一领域(源域)的训练集训练后得到完成某一任务的模型,用于完成相关领域(目标域)的同一任务.该模式是建立在训练集与测试集为同一概率分布前提下进行的.完成任务不仅依靠源域模型完成任务的性能,也依赖于两域的相似性.目前,国内外许多专家、学者提出的领域自适应算法都致力于找到两个领域间的联系,如陶剑文等人[1]提出多核局部学习技术,该算法在多核组合的再生核 Hilbert 空间构建一种有效的三段式领域迁移学习模型,从而降低领域间的差异.Brian 等人[2]基于结构风险最小化框架,提出大间隔直推式迁移学习方法,试图寻求一种特征变换,使得源域数据和目标域数据间分布差异最小.Pan 等人[3]提出迁移成分分析算法,该方法学习领域间共同的可迁移特征成分集,减小领域间特征分布差距.

在实际应用中,可能出现信息采集器或传感器的抗干扰能力达不到要求而导致数据采集出现扰动或噪音;设备稳定性出现问题,如短路状况而导致数据信息不全的现象.尤其是在回归问题中,若采集信息不完整,会使回归预测性能下降.例如,基于地区人口构成和住房市场状况的地区住房平均价格的预计,某些地区数据采集较为全面(源域),包含了各类房价数据,而某些地区采集的数据却未包含高端房源信息(目标域),如果利用目标域数据预计房源价格,可能会导致高端房源价格预计不准确.若将所有数据整合在一起进行训练,海量的数据会导致训练时间增加、效率极大降低,而且还有可能掩盖目标域城市房价数据集本身的特点及其与源域城市房价数据集间的差异性.因此本文提出的领域自适应算法所要解决的问题是:如何通过有效学习源域知识,使得目标域训练所得的回归函数既能很好地反映目标域数据提供的信息,又能有效地对缺失数据提供较为准确的预测.

针对上述问题,本文在 Tsang 等人[4]提出的核心集支撑向量回归机(core vector regression,简称 CVR)的基础上,提出了一种全新的针对大样本的领域自适应核心集支撑向量回归机(adaptive-core vector regression,简称 A-CVR).Tsang 等人[4]提出任何满足式 $\begin{cases}\max\limits_{\boldsymbol{\beta}} \boldsymbol{\beta}^T(diag(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta} \\ \text{s.t. } \boldsymbol{\beta}\geqslant 0, \boldsymbol{\beta}^T\mathbf{1}=1\end{cases}$ 的 QP 问题均可等价于中心约束最小包含球(center-constrained minimum enclosing ball,简称 CC-MEB)问题,然后证明支撑向量回归机(support vector regression,简称 SVR)等价于 CC-MEB 问题,并进一步利用基于最小包含球理论的核心集技术开发了 CVR 算法,将该算法应用于大样本支撑向量回归.在上述理论的基础上,本文首先提出并证明了相似领域概率分布差异只与两域各自的最小包含球中心点位置有关,且其上限与半径无关的理论,并在此基础上提出了 A-CVR 算法.

该算法的创新之处在于:当两域相似分布时,采用与两域各自等价的 CC-MEB 来进行领域自适应学习,这种学习不需要源域所有样本参与运算,只需将源域知识(模型参数,即源域 CVR 等价的 CC-MEB 球心)传递给目标域 CVR 等价的 CC-MEB.通过这种方式,在有效知识迁移的同时达到源域数据隐私保护的目的,且不增加目标域二次规划运算的规模.所得到的目标域回归函数既能很好地反映目标域数据提供的信息,又能有效地对缺失的数据提供较为准确的预测,提高回归预测的准确度,进而能够更有效地进行下一步应用.文中所提源域与目标域的分布差异可参考文献[5,6]加以量化.

本文第 1 节介绍中心约束最小包含球理论.第 2 节介绍 SVR 算法及其与 CC-MEB 之间的关系.第 3 节介绍 A-CVR 算法及其理论依据.第 4 节给出实验结果与分析.最后第 5 节总结全文.

1 CC-MEB

2002 年,Bădoiu 和 Clarkson 在文献[7]中提出了基于核心集的最小包含球(1+ξ)近似算法.核心集是指输入集的一个子集,在优化问题中,使用该子集可获得与原始输入集求解同样好的近似结果.他们还证明了核心集的大小与维度、样本规模均无关.Tsang 等人在文献[4,8,9]中提出最小包含球(minimum enclosing ball,简称 MEB)问题与许多核问题有关,特别是 MEB 方法等价于支持向量域描述(support vector domain description,简称 SVDD)算法[10],可用于异常点检测.SVDD 算法的思想是找到包含集合 S 中所有点$\varphi(\boldsymbol{x}_i)$的最小球,则属于该类的数据就在球中,不属于该类的数据就在球外.

1.1 MEB

给定训练样本 $S=\{(\varphi(\boldsymbol{x}_i))\}_{i=1}^m$,其中,$\boldsymbol{x}_i\in R^d$,$\varphi$为与给定核 $\boldsymbol{K}$ 相关的核映射.为了使球体包得更紧凑,引入核技

巧的 MEB 优化模型,如公式(1)所示:

$$\begin{cases}\min\limits_{c,R} R^2 \\ \text{s.t. } (\varphi(\boldsymbol{x}_i)-\boldsymbol{c})^2 \leqslant R^2\end{cases},\ i=1,2,...,m \tag{1}$$

通过拉格朗日法可得球心:

$$\boldsymbol{c}=\sum_{i=1}^{m}\beta_i\varphi(\boldsymbol{x}_i) \tag{2}$$

公式(1)相应对偶问题的矩阵形式为

$$\begin{cases}\max\limits_{\boldsymbol{\beta}} \boldsymbol{\beta}^T(diag(\boldsymbol{K}))-\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta} \\ \text{s.t. } \boldsymbol{\beta}\geqslant 0, \boldsymbol{\beta}^T\mathbf{1}=1\end{cases} \tag{3}$$

其中,$\boldsymbol{\beta}=[\beta_1,\ldots,\beta_m]^T$ 是 m 维拉格朗日乘子向量,$\boldsymbol{K}_{m\times m}=[k(\boldsymbol{x}_i,\boldsymbol{x}_j)]=[\varphi^T(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)]$是 $m\times m$ 维核矩阵.当 $k(\boldsymbol{x}_i,\boldsymbol{x}_i)=k$($k$ 为某一常数)时,上式演变为

$$\begin{cases}\max\limits_{\boldsymbol{\beta}} -\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta} \\ \text{s.t. } \boldsymbol{\beta}\geqslant 0, \boldsymbol{\beta}^T\mathbf{1}=1\end{cases} \tag{4}$$

1.2 CC-MEB

在上述研究的基础上,学者们进一步提出 CC-MEB 模型,即为每个 $\varphi(\boldsymbol{x}_i)$增加一维 δ_i,形成集合:

$$S=\{(\varphi^T(\boldsymbol{x}_i),\delta_i)\}_{i=1}^m.$$

将最后一维的中心点坐标设为 0,即中心点坐标为[$\boldsymbol{c}$,0],则找到包含集合 S 中所有样本的最小包含球的优化问题为

$$\begin{cases}\min\limits_{c,R} R^2 \\ \text{s.t. } (\varphi(\boldsymbol{x}_i)-\boldsymbol{c})^2+\delta_i^2 \leqslant R^2\end{cases},\ i=1,2,...,m \tag{5}$$

设 $\boldsymbol{\Delta}=[\delta_1^2,...,\delta_m^2]^T\geqslant 0$, 则公式(5)相应的对偶问题的矩阵形式为

$$\begin{cases}\max\limits_{\boldsymbol{\beta}} \boldsymbol{\beta}^T(diag(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta} \\ \text{s.t. } \boldsymbol{\beta}\geqslant 0, \boldsymbol{\beta}^T\mathbf{1}=1\end{cases} \tag{6}$$

其中,核矩阵 $\boldsymbol{K}_{m\times m}=[k(\boldsymbol{x}_i,\boldsymbol{x}_j)]=[\varphi^T(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)]$.

使用最优解$\boldsymbol{\beta}$,可得半径 R 和中心点 $\boldsymbol{c}$ 的值:

$$\begin{cases}R=\sqrt{\boldsymbol{\beta}^T(diag(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta}} \\ \boldsymbol{c}=\sum\limits_{i=1}^{m}\boldsymbol{\beta}_i\varphi(\boldsymbol{x}_i)\end{cases} \tag{7}$$

任意一点到中心点的距离公式为

$$\|\boldsymbol{c}-\varphi(\boldsymbol{x}_l)\|^2+\delta_l^2=\|\boldsymbol{c}\|^2-2(\boldsymbol{K}\boldsymbol{\beta})_l+k_{ll}+\delta_l^2 \tag{8}$$

因为$\boldsymbol{\beta}^T\mathbf{1}$=1,任意实数$\eta$加入公式(6),不会影响$\boldsymbol{\beta}$的取值.原对偶形式改为

$$\begin{cases}\max\limits_{\boldsymbol{\beta}} \boldsymbol{\beta}^T(diag(\boldsymbol{K})+\boldsymbol{\Delta}-\eta\mathbf{1})-\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta} \\ \text{s.t. } \boldsymbol{\beta}\geqslant 0, \boldsymbol{\beta}^T\mathbf{1}=1, \boldsymbol{\Delta}\geqslant 0\end{cases} \tag{9}$$

文献[4]指出,任意满足公式(9)的 QP 问题均能看作是最小包含球问题,可运用核心集快速算法进行求解.

2 CVR 算法

2.1 L2-SVR算法

文献[4,7]提出了用于大数据集支撑向量回归的 CVR 算法.设存在大规模训练样本集 $\{Z_i=(\boldsymbol{x}_i,y_i)\}_{i=1}^m$,其中,

$\boldsymbol{x}_i\in R^d$ 为输入值,$y_i\in R$ 为输出值,回归问题就是寻找一个从输入空间到输出空间的映射 $f:R^d\to R$,即寻求回归函数 $f(\boldsymbol{x})=\boldsymbol{w}^T\varphi(\boldsymbol{x})+b$,其中,$\boldsymbol{w},\boldsymbol{x}\in R^d,b\in R$,使得 $f(\boldsymbol{x})=y$.文献[4,7]提出的类似于 ν-SVR[11,12]的支撑向量回归最优化问题的原始公式为

$$\begin{cases}\min\limits_{\boldsymbol{w},b}\|\boldsymbol{w}\|^2+b^2+\dfrac{C}{\mu m}\sum\limits_{i=1}^{m}(\xi_i^2+\xi_i^{*2})+2C\bar{\varepsilon}\\ \text{s.t.}\quad y_i-(\boldsymbol{w}^T\varphi(\boldsymbol{x}_i)+b)\leqslant\bar{\varepsilon}+\xi_i\\ \qquad(\boldsymbol{w}^T\varphi(\boldsymbol{x}_i)+b)-y_i\leqslant\bar{\varepsilon}+\xi_i^*\end{cases}\tag{10}$$

其中,参数 $\mu>0$,其作用与 ν-SVR 中的 ν 相似,用于控制 $\bar{\varepsilon}$ 大小;偏移 b 是惩罚因子,式中自动满足 $\xi_i,\xi_i^*\geqslant 0$.

引入两组拉格朗日乘子,构造出公式(10)的拉格朗日函数:

$$L=\|\boldsymbol{w}\|^2+b^2+\frac{C}{\mu m}\sum_{i=1}^{m}(\xi_i^2+\xi_i^{*2})+2C\bar{\varepsilon}+\sum_{i=1}^{m}\alpha_i(y_i-(\boldsymbol{w}^T\varphi(\boldsymbol{x}_i)+b)-\bar{\varepsilon}-\xi_i)+\sum_{i=1}^{m}\alpha_i^*((\boldsymbol{w}^T\varphi(\boldsymbol{x}_i)+b)-y_i-\bar{\varepsilon}-\xi_i^*)\tag{11}$$

公式(11)相应的对偶问题的矩阵形式为

$$\begin{cases}\max\limits_{\boldsymbol{\alpha}}[\boldsymbol{\alpha}^T\ \boldsymbol{\alpha}^{*T}]\begin{bmatrix}\dfrac{2}{C}y\\ -\dfrac{2}{C}y\end{bmatrix}-[\boldsymbol{\alpha}^T\ \boldsymbol{\alpha}^{*T}]\tilde{K}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\alpha}^*\end{bmatrix}\\ \text{s.t. }[\boldsymbol{\alpha}^T\ \boldsymbol{\alpha}^{*T}]\mathbf{1}=1,\boldsymbol{\alpha},\boldsymbol{\alpha}^*\geqslant 0\end{cases}\tag{12}$$

其中,

$$y=\begin{bmatrix}y_1\\ \vdots\\ y_m\end{bmatrix},\boldsymbol{\alpha}=\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_m\end{bmatrix},\boldsymbol{\alpha}^*=\begin{bmatrix}\alpha_1^*\\ \vdots\\ \alpha_m^*\end{bmatrix},\tilde{\boldsymbol{K}}=[\tilde{k}(z_i,z_j)]=\begin{bmatrix}\boldsymbol{K}+\mathbf{1}\mathbf{1}^T+\dfrac{\mu m}{C}\boldsymbol{I} & -(\boldsymbol{K}+\mathbf{1}\mathbf{1}^T)\\ -(\boldsymbol{K}+\mathbf{1}\mathbf{1}^T) & \boldsymbol{K}+\mathbf{1}\mathbf{1}^T+\dfrac{\mu m}{C}\boldsymbol{I}\end{bmatrix}\tag{13}$$

求解得到各变量的值为

$$\begin{cases}\boldsymbol{w}=C\sum\limits_{i=1}^{m}(\alpha_i-\alpha_i^*)\varphi(\boldsymbol{x}_i)\\ b=C\sum\limits_{i=1}^{m}(\alpha_i-\alpha_i^*)\\ \xi_i=\alpha_i\mu m,\ \xi_i^*=\alpha_i^*\mu m\\ \bar{\varepsilon}=[\boldsymbol{\alpha}^T\ \boldsymbol{\alpha}^{*T}]\begin{bmatrix}y\\ -y\end{bmatrix}-C[\boldsymbol{\alpha}^T\ \boldsymbol{\alpha}^{*T}]\tilde{\boldsymbol{K}}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\alpha}^*\end{bmatrix}\end{cases}\tag{14}$$

此外,因为 $\sum\limits_{i=1}^{m}(\alpha_i+\alpha_i^*)=1$,故 $\mu=\sum\limits_{i=1}^{m}(\xi_i+\xi_i^*)\Big/m$,由此得到参数 μ 类似于 ν-SVR 中的 ν,可解释为期望误差.

2.2 SVR与CC-MEB之间的关系

公式(12)为 SVR 的 QP 形式,令 $\tilde{\boldsymbol{\alpha}}=[\boldsymbol{\alpha}^T\ \boldsymbol{\alpha}^{*T}]$,则

$$\boldsymbol{\Delta}=-diag(\tilde{\boldsymbol{K}})+\eta\mathbf{1}+\frac{2}{C}\begin{bmatrix}y\\ -y\end{bmatrix}\tag{15}$$

其中,实数 η 应足够大,以使 $\boldsymbol{\Delta}\geqslant 0$.这样,公式(12)就可写成如下形式:

$$\begin{cases}\max\limits_{\tilde{\boldsymbol{\alpha}}}\tilde{\boldsymbol{\alpha}}^T(diag(\tilde{\boldsymbol{K}})+\boldsymbol{\Delta}-\eta\mathbf{1})-\tilde{\boldsymbol{\alpha}}^T\tilde{\boldsymbol{K}}\tilde{\boldsymbol{\alpha}}\\ \text{s.t. }\tilde{\boldsymbol{\alpha}}^T\mathbf{1}=1\end{cases}\tag{16}$$

该形式用 $\tilde{\boldsymbol{\alpha}}$ 替换了公式(9)中的 $\boldsymbol{\beta}$,即等价于中心约束最小包含球问题,可使用核心集快速算法 CVR 求解.

根据公式(16),球心 $\boldsymbol{c}$ 可按 $\boldsymbol{c}=\sum\limits_{i=1}^{2\times m}\tilde{\alpha}_i\tilde{\varphi}(\boldsymbol{x}_i)$ 进行计算,其中,

- 当 i=1,…,m 时,$\tilde{\varphi}(\boldsymbol{x}_i)=\varphi(\boldsymbol{x}_i)$;

- 当 $i=m+1,\ldots,2m$ 时, $\tilde{\varphi}(\boldsymbol{x}_i)=-\varphi(\boldsymbol{x}_i)$.

由此推导可得:

$$\boldsymbol{c}=\sum_{i=1}^{2\times m}\tilde{\alpha}_i\tilde{\varphi}(\boldsymbol{x}_i)=\sum_{i=1}^{m}\alpha_i\varphi(\boldsymbol{x}_i)+\sum_{i=1}^{m}\alpha_i^*(-\varphi(\boldsymbol{x}_i))=\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)\varphi(\boldsymbol{x}_i).$$

公式(14)中的 $\boldsymbol{w}$ 就可简化为 $\boldsymbol{w}=C\boldsymbol{c}$.

3 A-CVR

在回归问题中,相似应用领域,若目标域采集到的数据不完整,如何充分利用已有相关领域的知识对目标域数据进行有效的预处理,是数据进一步应用的前提.针对该问题,本文提出了一种全新的大样本领域自适应学习算法 A-CVR,该算法可充分学习以前的知识,使现有回归函数更接近于真实回归函数,从而能够更有效地进行下一步应用.

3.1 A-CVR原理

A-CVR 的核心思想是:SVR 算法可化解成公式(16)形式的 CC-MEB,该球可由中心点和半径表示,如果源域数据和目标域数据为相似领域数据,则两域概率分布应接近,两域 SVR 各自等价的 CC-MEB 中心点应靠近,且半径相当.公式(17)为求解学习源域知识(源域半径为 r,中心点为 $\boldsymbol{c}_0$)的目标域 CC-MEB:

$$\begin{cases}\min\limits_{\boldsymbol{c},R}(R-r)^2+u\|\boldsymbol{c}-\boldsymbol{c}_0\|^2\\ \text{s.t. }(\varphi(\boldsymbol{x}_i)-\boldsymbol{c})^2+\delta_i^2\leqslant R^2\end{cases}\tag{17}$$

其中,u 为可调的常数项参数,$\boldsymbol{x}_i$ 为样本数据,$\varphi(\cdot)$为样本高维映射函数,δ_i 为每个$\varphi(\boldsymbol{x}_i)$增加的一维.公式(17)的第 1 部分表示目标域 CC-MEB 的 R 与源域 CC-MEB 的 r 尽量逼近,第 2 部分表示目标域 CC-MEB 的 $\boldsymbol{c}$ 和源域 CC-MEB 的 $\boldsymbol{c}_0$ 尽量逼近,常量参数 u 用于平衡两项的贡献.该优化公式的总含义可以归纳为:使目标域 CC-MEB 的中心点和半径与源域 CC-MEB 的中心点和半径尽量逼近.

下文通过定理证明:相似领域的概率分布差异可用两域各自的最小包含球中心点表示,且其上限与半径无关.其优化问题可进一步简化.

定理 1. 源域 D_{source} 与目标域 D_{target} 独立且近似分布,其中,D_{source} 的概率密度函数为 $p(\boldsymbol{x})$,D_{target} 的概率密度函数为 $\hat{p}(\boldsymbol{x})$, 两域概率差只与两域各自的中心点位置有关,且其上限与半径无关.

证明:设数据集 $S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}\in R^d$,则其核密度估计函数[13,14]为

$$\begin{cases}\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})=\sum\limits_{i=1}^{N}\gamma_iK_h(\boldsymbol{x},\boldsymbol{x}_i)\\ \sum\limits_{i=1}^{N}\gamma_i=1,\gamma_i\geqslant 0\end{cases}\tag{18}$$

其中,$\boldsymbol{\gamma}=[\gamma_1,\ldots,\gamma_N]^T$ 为权向量,$K_h(\boldsymbol{x},\boldsymbol{x}_i)$为给定的核函数,$h$ 为给定的核带宽.本文应用文献[13]所采用的最小积分平方误差(integrated squared error,简称 ISE)准则,使 $\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})$ 最优逼近真实密度函数 $p(\boldsymbol{x})$.定理证明分两部分:

i) 核密度估计函数可用最小包含球中心点表示:即 $\hat{p}(\boldsymbol{x})=\sum\limits_{i=1}^{N}\beta_iK(\boldsymbol{x},\boldsymbol{x}_i)=\boldsymbol{c}^T\varphi(\boldsymbol{x}_i)$, 其中,$K$ 为高斯核.

在 ISE 准则下,使 $\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})$ 尽量逼近 $p(\boldsymbol{x})$,即:

$$\begin{aligned}\hat{\gamma}&=\arg\min_{\boldsymbol{\gamma}}ISE(\boldsymbol{\gamma})\\ &=\arg\min_{\boldsymbol{\gamma}}\int_{R^d}\|p(\boldsymbol{x})-\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})\|^2\,\mathrm{d}x\\ &=\arg\max_{\boldsymbol{\gamma}}\{2E_{p(\boldsymbol{x})}[\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})]-\int_{R^d}\hat{p}^2(x;h,\boldsymbol{\gamma})\mathrm{d}x\}\end{aligned}\tag{19}$$

文献[13−16]中指出,核函数的选取对核估计好坏的影响远小于核带宽 h 的选取,而因高斯核具有如下性质: $\int G_h(\boldsymbol{x},\boldsymbol{x}_i)G_h(\boldsymbol{x},\boldsymbol{x}_j)\mathrm{d}x=G_{2h}(\boldsymbol{x}_i,\boldsymbol{x}_j)$, 可使计算复杂度大为减小,故本文选用高斯核,于是有,

$$\begin{aligned}\int_{R^d}\hat{p}^2(\boldsymbol{x};h,\boldsymbol{\gamma})\mathrm{d}x&=\int_{R^d}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j K_h(\boldsymbol{x},\boldsymbol{x}_i)K_h(\boldsymbol{x},\boldsymbol{x}_j)\mathrm{d}x\\&=\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j G_{2h}(\boldsymbol{x}_i,\boldsymbol{x}_j)\end{aligned}\tag{20}$$

若 $\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})$ 是 $p(\boldsymbol{x})$在 $p(\boldsymbol{x})$条件下的无偏估计量,则

$$E_{p(\boldsymbol{x})}[p(\boldsymbol{x};h,\boldsymbol{\gamma})]=E[p(\boldsymbol{x})]\tag{21}$$

将公式(20)和公式(21)代入公式(19),则有

$$\begin{cases}\hat{\gamma}=\arg\max\limits_{\boldsymbol{\gamma}}-\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m}\gamma_i\gamma_j G_{2h}(\boldsymbol{x}_i,\boldsymbol{x}_j)\\ \text{s.t.}\ \sum\limits_{i=1}^{N}\gamma_i=1,\gamma_i\geqslant 0,i=1,2,...,N\end{cases}\tag{22}$$

公式(22)与公式(4)等价,即 MEB 对偶形式的乘子$\boldsymbol{\beta}$向量可作核密度估计函数的权向量$\boldsymbol{\gamma}$,若用 $\hat{p}(\boldsymbol{x})$ 代替 $\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})$, 则有

$$\hat{p}(\boldsymbol{x})=\sum_{i=1}^{N}\beta_i\varphi^T(\boldsymbol{x}_i)\varphi(\boldsymbol{x}).$$

将公式(2)的 MEB 中心点 $\boldsymbol{c}$ 带入,可得:

$$\hat{p}(\boldsymbol{x})=\boldsymbol{c}^T\varphi(\boldsymbol{x})=\varphi^T(\boldsymbol{x})\cdot\boldsymbol{c}\tag{23}$$

ii) 使用第 i)部分的结论求两域概率差.

度量概率差可采用 KL 散度及 ISE 准则等度量方式,但为了方便推导和使用 CVM 技术解决大样本问题,本文根据文献[17,18]采用 ISE 准则.设源域数据集 $S=\{\boldsymbol{x}_1^*,\boldsymbol{x}_2^*,...,\boldsymbol{x}_M^*\}\in R^d$,目标域数据集 $T=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}\in R^d$.源域密度估计式为 $p(\boldsymbol{x})=\varphi^T(\boldsymbol{x}^*)\cdot\boldsymbol{c}_0=\varphi^T(\boldsymbol{x})\sum\limits_{i=1}^{M}\beta_i^*\varphi(\boldsymbol{x}_i^*)$;目标域密度估计式为 $\hat{p}(\boldsymbol{x})=\varphi^T(\boldsymbol{x})\cdot\boldsymbol{c}=\varphi^T(\boldsymbol{x})\sum\limits_{i=1}^{N}\beta_i\varphi(\boldsymbol{x}_i)$.

在 ISE 准则下,概率密度差可用 $\int(p(\boldsymbol{x})-\hat{p}(\boldsymbol{x}))^2\mathrm{d}x$ 表示.对该式进行推导:

$$\begin{aligned}\int(p(\boldsymbol{x})-\hat{p}(\boldsymbol{x}))^2\mathrm{d}x&=\int\left(\varphi^T(\boldsymbol{x})\sum_{i=1}^{M}\beta_i^*\varphi(\boldsymbol{x}_i^*)-\varphi^T(\boldsymbol{x})\sum_{j=1}^{N}\beta_j\varphi(\boldsymbol{x}_j)\right)^2\mathrm{d}x\\&=\int\varphi^T(\boldsymbol{x})\left(\sum_{i=1}^{M}\beta_i^*\varphi(\boldsymbol{x}_i^*)-\sum_{j=1}^{N}\beta_j\varphi(\boldsymbol{x}_j)\right)\left(\sum_{i=1}^{M}\beta_i^*\varphi(\boldsymbol{x}_i^*)-\sum_{j=1}^{N}\beta_j\varphi(\boldsymbol{x}_j)\right)^T\varphi(\boldsymbol{x})\mathrm{d}x.\end{aligned}$$

设 $\sum\limits_{i=1}^{M}\beta_i^*\varphi(\boldsymbol{x}_i^*)-\sum\limits_{j=1}^{N}\beta_j\varphi(\boldsymbol{x}_j)$ 为矩阵 $\boldsymbol{A}_{d\times d}$,可找到值$\varepsilon$大于矩阵 $\boldsymbol{A}$ 中每一个元素,则上式进一步推导可得:

$$\int\varphi^T(\boldsymbol{x})\boldsymbol{A}\boldsymbol{A}^T\varphi(\boldsymbol{x})\mathrm{d}x\leqslant\int\varphi^T(\boldsymbol{x})\boldsymbol{B}\boldsymbol{B}^T\varphi(\boldsymbol{x})\mathrm{d}x=\varepsilon^2\int\varphi^T(\boldsymbol{x})\boldsymbol{I}_{d\times d}\varphi(\boldsymbol{x})\mathrm{d}x,$$

其中,$\boldsymbol{I}$ 为全 1 矩阵.高斯核 $\int\varphi^T(\boldsymbol{x})\boldsymbol{I}_{d\times d}\varphi(\boldsymbol{x})\mathrm{d}x=$常数$k$, 于是有上式$\leqslant\varepsilon^2k$.定理得证. □

结论:核密度估计函数可用最小包含球中心点表示,且概率密度差上限只与两域各自的中心点位置有关.该定理给我们的启发是:学习源域知识时只需考虑源域中心点,而无需考虑半径.公式(17)可进一步简化为

$$\begin{cases}\min\limits_{\boldsymbol{c},R}R^2+u\|\boldsymbol{c}-\boldsymbol{c}_0\|^2\\ \text{s.t.}\ (\varphi(\boldsymbol{x}_i)-\boldsymbol{c})^2+\delta_i^2\leqslant R^2\end{cases}\tag{24}$$

其含义为求离源域中心点最近的目标域最小包含球.u 为自适应参数,u 值越大,源域中心点的影响就越大.

引入拉格朗日乘子变量,在约束条件下构造公式(24)的拉格朗日函数:

$$L=R^2+u\|\boldsymbol{c}-\boldsymbol{c}_0\|^2+\sum_{i=1}^{N}\gamma_i((\varphi(\boldsymbol{x}_i-\boldsymbol{c})^2+\delta_i^2-R^2)\tag{25}$$

由最优化理论可知,公式(25)在鞍点处取极值,在鞍点处 L 关于变量 $\boldsymbol{c}$ 和 R 的偏导数应满足:

$$\begin{cases}\dfrac{\partial L}{\partial R}=2R-2R\sum\limits_{i=1}^{N}\gamma_i=0\\ \dfrac{\partial L}{\partial \boldsymbol{c}}=2u\|\boldsymbol{c}-\boldsymbol{c}_0\|-2\sum\limits_{i=1}^{N}\gamma_i(\varphi(\boldsymbol{x}_i)-\boldsymbol{c})=0\end{cases} \tag{26}$$

由此可得:

$$\begin{cases}\sum\limits_{i=1}^{N}\gamma_i=1\\ \boldsymbol{c}=\dfrac{u\boldsymbol{c}_0+\sum\limits_{i=1}^{N}\gamma_i\varphi(\boldsymbol{x}_i)}{u+1}\end{cases} \tag{27}$$

将公式(27)代入公式(25),该问题的对偶形式为

$$\begin{cases}\max\sum\limits_{i=1}^{N}\left(\varphi^2(\boldsymbol{x}_i)-\dfrac{2u\boldsymbol{c}_0^T\varphi(\boldsymbol{x}_i)}{u+1}+\delta_i^2\right)\gamma_i-\dfrac{1}{u+1}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\gamma_i\gamma_j\varphi^T(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)\\ \text{s.t.}\sum\limits_{i=1}^{N}\gamma_i=1\end{cases} \tag{28}$$

3.2 A-CVR原理

第 3.1 节主要介绍中心点校正 CC-MEB 原理,本节将其应用到大样本,提出 A-CVR 新算法.

设存在目标域训练大样本集 $\{T_i=(\boldsymbol{x}_i,y_i)\}_{i=1}^{m}$,其中,$\boldsymbol{x}_i\in R^d$ 为输入值,$y_i\in R$ 为输出值,设$\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_m]$, $\boldsymbol{\alpha}^*=[\alpha_1^*,\ldots,\alpha_m^*]$,则公式(28)中,$N=2m$,$\boldsymbol{\gamma}=[\boldsymbol{\alpha}^T\ \boldsymbol{\alpha}^{*T}]$,$\varphi(\boldsymbol{x}_i)$取值为:

- 当 $i=1,\ldots,m$ 时,$\varphi(\boldsymbol{x}_i)=-\varphi(\boldsymbol{x}_i)$;
- 当 $i=m+1,\ldots,2m$ 时,$\varphi(\boldsymbol{x}_i)=-\varphi(\boldsymbol{x}_i)$.

$\boldsymbol{\Delta}_{old}=\{\delta_1^2,\ldots,\delta_N^2\}$, 其值由 CVR 公式(15)给出.公式(28)中,$\varphi(a)\varphi(b)$的形式均用核函数 $\tilde{\boldsymbol{K}}(a,b)$ 代替,其值由 CVR 公式(13)给出.

故公式(28)可推出 A-CVR 核矩阵 $\tilde{\tilde{\boldsymbol{K}}}=\dfrac{1}{u+1}\tilde{\boldsymbol{K}}$.

$$\boldsymbol{\Delta}_{new}=-diag(\tilde{\tilde{\boldsymbol{K}}})+diag(\tilde{\boldsymbol{K}})+\boldsymbol{\Delta}_{old}-\frac{2u\boldsymbol{c}_0^T\varphi(\boldsymbol{x}_i)}{u+1}+\eta_{new}\boldsymbol{1} \tag{29}$$

其中,η_{new} 为使 $\eta_{new}\geqslant 0$ 的实数;源域中心点 $\boldsymbol{c}_0=\sum\limits_{i=1}^{|St\_old|}\beta_i\varphi(\boldsymbol{x}_i^*),\boldsymbol{x}_i^*\in St\_old$,St_old 代表源域 CVR 核心集,$\boldsymbol{\beta}$为源域拉格朗日系数.

公式(28)可写成如下形式:

$$\begin{cases}\max\limits_{\boldsymbol{\gamma}}\boldsymbol{\gamma}^T(diag(\tilde{\tilde{\boldsymbol{K}}})+\boldsymbol{\Delta}_{new}-\eta_{new}\boldsymbol{1})-\boldsymbol{\gamma}^T\tilde{\tilde{\boldsymbol{K}}}\boldsymbol{\gamma}\\ \text{s.t.}\ \boldsymbol{\gamma}^T\boldsymbol{1}=1\end{cases} \tag{30}$$

比较公式(30)与公式(9),将$\boldsymbol{\gamma}$替换为$\boldsymbol{\beta}$, $\tilde{\tilde{\boldsymbol{K}}}$ 替换为 $\boldsymbol{K}$,$\boldsymbol{\Delta}_{new}$ 替换为$\boldsymbol{\Delta}$,η_{new} 替换为η,则两式等价,即 A-CVR 也是一个 CC-MEB 问题,可以使用核心集快速算法进行求解.

上文说明,由公式(14)得出的 $\boldsymbol{w}=C\sum\limits_{i=1}^{m}(\alpha_i-\alpha_i^*)\varphi(\boldsymbol{x}_i)$ 可化简成 $\boldsymbol{w}=C\boldsymbol{c}$,其中,$\boldsymbol{c}$ 为 CC-MEB 的中心点.自适应算法所得自适应后的中心点坐标由公式(27)给出,将其带入 $\boldsymbol{w}=C\boldsymbol{c}$ 即可获得 $\boldsymbol{w}$ 的值.将 $\boldsymbol{w}$,$\boldsymbol{x}$,y 的值带入 $y=\boldsymbol{w}^T\boldsymbol{x}+b$,即可获得 b 的值,其中,$\boldsymbol{x}$,y 为训练样本.在本文实验中,b 取所有训练样本的平均值.

3.3 A-CVR算法

下面介绍 A-CVR 算法的主要步骤:

输入:源域 CVR 的核心集 St_old,源域拉格朗日系数$\boldsymbol{\beta}$,目标域数据集 $\{Z_i=(\boldsymbol{x}_i,y_i)\}_{i=1}^m$,逼近参数$\xi$,实数$\eta$及$\eta_{new}$,以及高斯核宽$\sigma$.

$$\eta=\max\left(0,\max\left(diag(\tilde{\boldsymbol{K}})-\frac{2}{C}\begin{bmatrix}y\\-y\end{bmatrix}\right)\right),$$

$$\eta_{new}=\max\left(0,\max\left(diag(\tilde{\tilde{\boldsymbol{K}}})+\frac{2u\boldsymbol{c}_0^T\varphi(\boldsymbol{x}_i)}{u+1}-\eta\boldsymbol{1}-\frac{2}{C}\begin{bmatrix}y\\-y\end{bmatrix}\right)\right);$$

输出:核心集 St,目标域拉格朗日系数$\boldsymbol{\gamma}$.

训练步骤:

步骤 1. 随机产生初始核心集 Q_0,Q_0 中样本所生成 CC-MEB 的 $\boldsymbol{c}_0$ 和 R_0,并将迭代次数设为 0.

步骤 2. 如果没有样本 $\boldsymbol{x}$ 在 $CC\text{-}MEB(\boldsymbol{c}_t,(1+\xi)R_t)$球外,则进入步骤 7.

步骤 3. 按公式(8)找到离中心点 $\boldsymbol{c}_t$ 最远的点,并把该点加入核心集 $Q_{t+1}=Q_t\cup\{\boldsymbol{x}\}$.

步骤 4. 求解新的 CC-MEB,记为 $MEB(Q_{t+1})$,且 $\boldsymbol{c}_{t+1}=\boldsymbol{c}_{MEB}(Q_{t+1})$,$R_{t+1}=R_{MEB}(Q_{t+1})$.

步骤 5. $t=t+1$ 并返回步骤 2.

步骤 6. 终止训练,返回所需要的输出.

求 b 值步骤:

步骤 7. 将公式(27)自适应后中心点坐标带入下面公式,其中,$\boldsymbol{x}$ 为训练样本:

$$y'=\boldsymbol{w}^T\varphi(\boldsymbol{x})=C\boldsymbol{c}^T\varphi(\boldsymbol{x})=C\cdot\frac{u\boldsymbol{c}_0^T+\sum_{i=1}^N\gamma_i\varphi^T(\boldsymbol{x}_i)}{u+1}\cdot\varphi(\boldsymbol{x}).$$

步骤 8. $b=mean(y-y')$,y 为训练样本函数值,y'为步骤 7 获得值.

预测步骤:

步骤 9. 将测试样本 $\boldsymbol{x}_{test}$ 带入下式:

$$y\_test=\boldsymbol{w}^T\varphi(\boldsymbol{x}_{test})+b=C\boldsymbol{c}^T\varphi(\boldsymbol{x}_{test})+b=C\cdot\frac{u\boldsymbol{c}_0^T+\sum_{i=1}^N\gamma_i\varphi^T(\boldsymbol{x}_i)}{u+1}\cdot\varphi(\boldsymbol{x}_{test})+b.$$

4 实验与分析

本节使用 Benchmark 数据集和大规模真实回归数据集对 A-CVR 算法进行实验验证.实验环境为 Intel Core 2 2.40GHz CPU,2.39GHz 1.94GB RAM,Windows XP SP3,MATLAB 7.1 等.

本文采用均方根误差(root-mean-square error)[4]来评估 A-CVR 算法的性能,其定义为

$$RMSE=\frac{1}{\max_i y_i}\sqrt{\frac{1}{n}\sum_{i=1}^n(f(\boldsymbol{x}_i)-y_i)^2},i为测试样本大小.$$

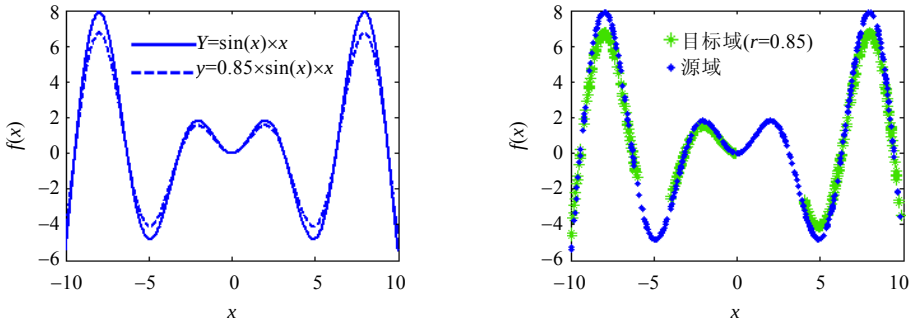
4.1 Benchmark数据集实验

为了更好地说明本文所提算法的自适应性,模拟数据集实验主要从以下 4 个方面展开:① 直接利用源域数据集(source domain dataset,简称 SD)训练回归函数对测试集进行测试;② 直接利用目标域缺失数据集(target domain dataset,简称 TD)训练获得回归函数对测试集进行测试;③ 将源域数据集与目标域数据集合并(source domain data set and target domain data sets together,简称 SDTD)训练获得回归函数对测试集进行测试;④ 使用文本方法基于目标域数据集和历史知识(target domain dataset and historical knowledge,简称 TDHK)训练获得回归函数对测试集进行测试.

4.1.1 源域数据集、目标域数据集及测试集的生成

由文献[19]数据构成方式生成源域、目标域数据集:根据函数 $Y=f(x)=\sin(x)\times x$,$x\in[-10,10]$生成 10 000 个源域数据集(SD);根据函数 $y=r\times Y=r\times f(x)$,$x\in[-10,10]$生成 10 000 个目标域数据集(TD)和 2 000 个测试集(TD_test),

其中,r 为源域与目标域相关性因子,本文分别取 0.7,0.75,0.8,0.85,0.9 进行实验,且在构造目标域训练集时在[−6,−4], [0,4]区间人为设置信息缺失.图 1(a)显示了当 r 值取 0.85 时的两域示意函数,为了显示得更加清晰,图 1(b)给出了两域各取 400 个点的采样数据集.



(a) 源域、目标域函数曲线 (b) 源域采样与目标域采样图

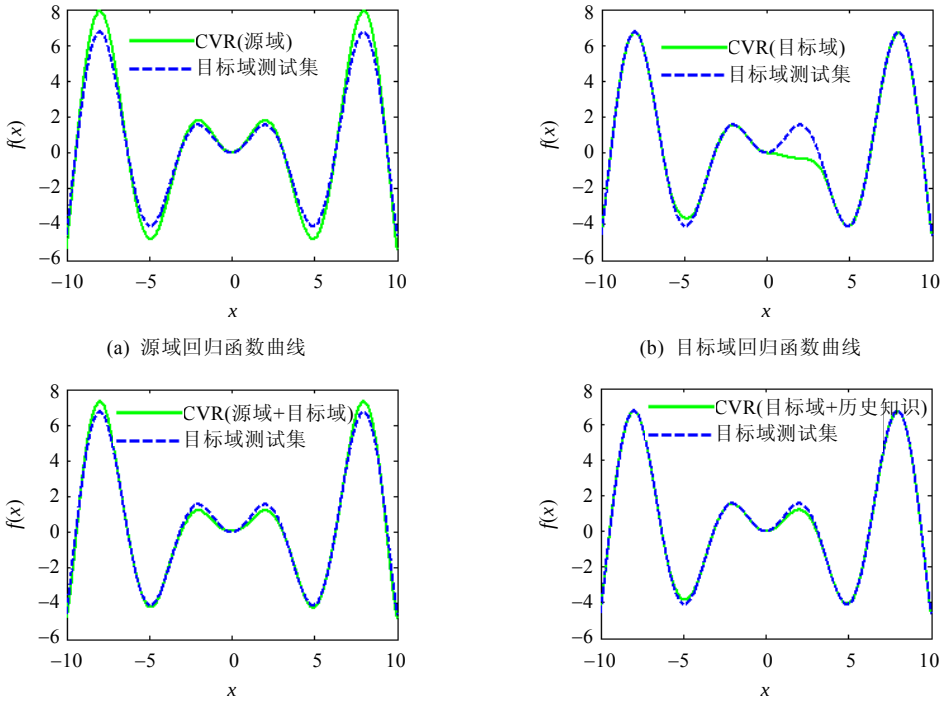
Fig.1 Sampling dataset from source domain and target domain with the correlation factor 0.85

图 1 相关因子为 0.85 时,源域和目标域示意函数及相应的采样数据集

通过上述方式设置的模拟数据集,使得源域与目标域既相似又存在差异,且目标域存在数据信息不全的问题.

4.1.2 实验结果分析

表 1 给出了不同相关因子情况下各种算法均方根误差的比较,图 2 给出了相关因子为 0.85 时,各种算法对目标域测试集(TD_test)测试时各回归函数曲线效果的比较.



(a) 源域回归函数曲线 (b) 目标域回归函数曲线

(c) 源域与目标域合并后的回归函数曲线 (d) A-CVR 自适应后的回归函数曲线

Fig.2 Performance comparison on the target test set (TD_test) with the correlation factor 0.85

图 2 相关因子为 0.85,对目标域测试集(TD_test)测试时各回归函数曲线效果比较

Table 1 RMSEs of several methods on the synthetic datasets

表 1 各种算法在模拟数据集上均方根误差比较

目标域信息缺失区间	源域与目标域相关性因子	SD	TD	SDTD	A-CVR	
					TDHK	u
[−6,−4] and [0,4]	0.9	0.003 8	0.005 6	0.002 0	0.001 7	9
	0.85	0.005 7	0.005 2	0.002 8	0.001 5	8
	0.8	0.007 6	0.004 9	0.003 7	0.001 1	8
	0.75	0.009 5	0.004 6	0.004 6	8.37E-04	8
	0.7	0.011 4	0.004 3	0.005 8	2.69E−04	14

由表 1 和图 2 可知:

① 本文提出的 A-CVR 算法与其他 CVR 算法相比有更好的性能;

② 图 2(a)中的实线表示源域回归函数曲线,虚线表示目标域测试集实际回归曲线.由图 2(a)可知,源域数据集与目标域数据集间存在偏移,若直接使用源域回归函数对目标域测试集进行预测,则无法达到与目标域测试集实际值逼近的效果.由表 1 可知,相关因子越小,误差越大;

③ 图 2(b)中的实线表示目标域回归函数曲线,虚线表示目标域测试集实际回归曲线.从图 2(b)可明显发现,该曲线在信息缺失区间[0,4]上存在明显的预测性能恶化现象,原因是已有方法只能实现向当前采样训练集逼近而不能实现信息弥补,因此,在信息缺失部分,必然存在缺陷而最终导致整个系统性能下降;

④ 图 2(c)中的实线表示源域、目标域合并训练后的回归函数曲线,虚线表示目标域测试集实际回归曲线.由图 2(c)可知,源域、目标域合并训练后生成的回归函数曲线预测结果好于以上两种方法,但其逼近效果仍不理想.造成这种结果的因素主要在于如图 2(a)所示源域与目标域存在一定的偏差,所获得的回归函数曲线只能取两者的折中,虽然可以对缺失部分进行弥补,降低了均方根误差,但在无数据缺失部分,所得回归曲线与实际曲线相比存在偏差.该方法的另一缺点是需要源域数据集全部参与运算,不但扩大了运算规模,且一些高度机密的源域数据通常难以获取.如果从源域仅能得到一些归纳出来的知识,如对应的 CVR 的参数,则该方法就变得不再可行;

⑤ 图 2(d)中的实线表示 A-CVR 自适应训练后的回归函数曲线,虚线表示目标域测试集实际回归曲线.由图 2(d)可知,本文方法与图 2(a)相比逼近效果更好;与图 2(b)相比,缺失部分得到了校正;与图 2(c)相比,不仅缺失部分得到校正,且未缺失部分仍能保持较好的性能,很好地体现了源域、目标域间的差异性.需要指出的是:本文的 A-CVR 方法只需目标域的数据和源域知识(模型参数)作为训练数据,而不需要所有源域数据参与运算,故在隐私保护方面也体现了较大的优势.

4.1.3 实验参数设置

A-CVR 算法与 CVR 算法相比,多了自适应参数 u.若参数 u 设置为 0,A-CVR 就退化为 CVR 算法.故算法参数设置分两步进行:

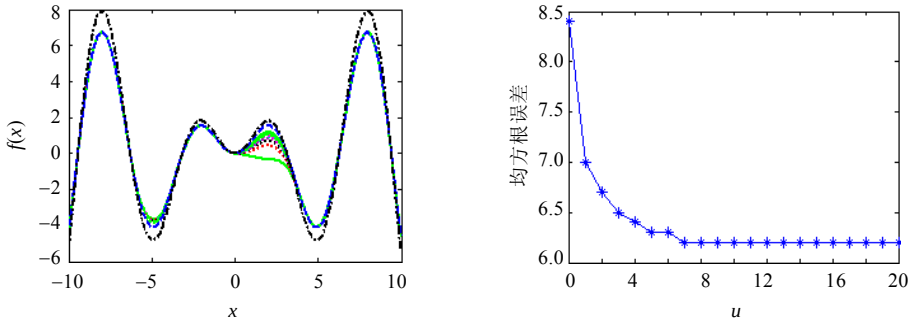
首先,将自适应参数 u 值固定为 0,训练当前目标域回归函数曲线,其参数设置如文献[4].核心集逼近参数 ξ 取 10^{-6};高斯核 $k(\boldsymbol{x},y)=\exp(-\|\boldsymbol{x}-y\|^2/h)$,其中,$h$ 先按公式 $h=(1/m^2)\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2$ 获得近似值,在此近似值附近搜索最优值,式中 m 表示数据规模[4].参数 h 与参数 C 的最优值通过交叉验证确定.

第 2 步,通过设置 u 学习源域知识,将缺失部分校正.

图 3(a)的点划线表示源域回归预测曲线;虚线表示目标域实际函数曲线;最下方实线表示目标域回归预测曲线,因存在数据缺失,缺失部分误差很大.由图 3(a)可知,因源域模型已确定,测试集在源域模型上的预测值固定不变.随着 u 值的增大,学习了源域知识,目标域回归曲线缺失部分向源域靠拢,未缺失部分变化不明显.故随着 u 参数的变化,测试集在目标域和源域上的回归预测值均方根误差减小,而这种减小趋势是由缺失部分回归曲线校正引起的.

因此,通过计算不同 u 值时,测试集在源域的预测值和在目标域的预测值间均方根误差,可判断 u 的取值.如图 3(b)所示.当 u 值取 1 时,误差降幅最大;随着 u 值的增大,误差降幅逐渐减小.图 3(a)从下往上依次显示不同 u

值时的校正回归曲线,与图 3(b)相对应,当 u 值为 1 时,校正幅度最大;u 值为 2 时,次之;当 u 值大于 8 时,均方根误差降幅缓慢,而回归曲线基本重合.本文下面讨论:u 值不是越大越好,增大 u 值会加强源域的作用,导致未缺失部分最终也向源域靠拢,当数据未缺失部分误差增幅大于缺失部分误差降幅时,总误差没有进一步下降,反而呈现上升趋势.故测试集在源域和目标域上的均方根误差变化趋于稳定的 u 值为最佳 u 值.本例 u 值可设为 8.



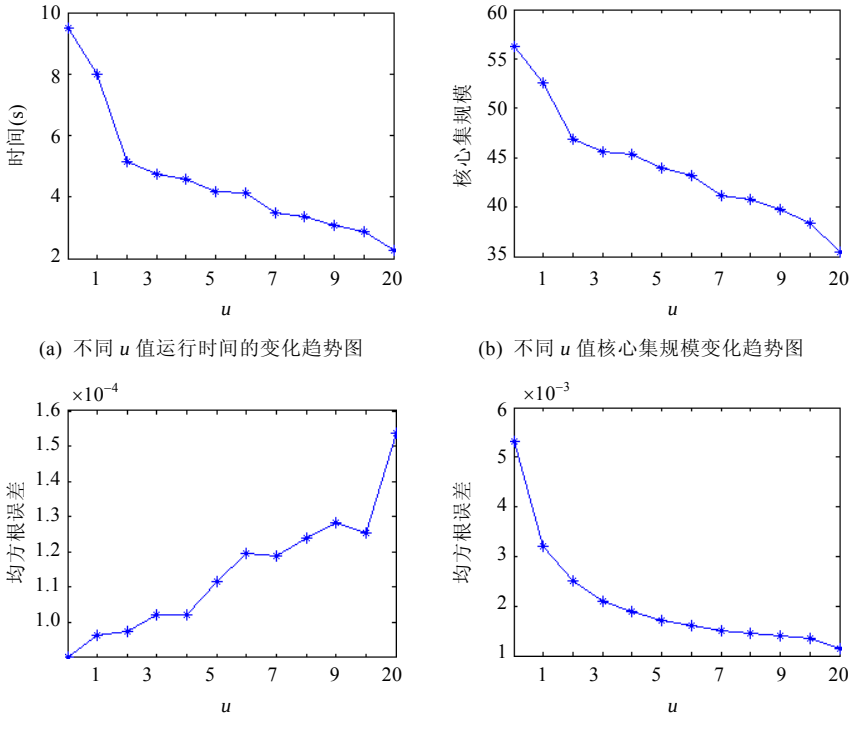
(a) 不同 u 值回归函数曲线的变化情况 (b) 不同 u 值测试集在源域与目标域间的均方根误差变化情况

Fig.3 Determination of u

图 3 u 值的确定方法

4.1.4 不同 u 值算法性能分析

图 4 所示运行结果均为程序运行 20 次的均值.



(a) 不同 u 值运行时间的变化趋势图 (b) 不同 u 值核心集规模变化趋势图

(c) 不同 u 值无缺失部分均方根误差变化趋势图 (d) 不同 u 值均方根误差变化趋势图

Fig.4 Performance of algorithm A-CVR with different u values

图 4 不同 u 值算法性能分析

由图 4(a)、图 4(b)可知,当增大 u 值时,核心集规模缩小,运行时间缩短.原因在于,学习源域知识后,把信息不全部分包含进球体,使所得 CC-MEB 比目标域训练集训练所得 CC-MEB 要大.即可理解为增大了目标域球体逼近参数ξ的值,从而使得核心集规模缩小,运行时间缩短.继续增大 u 值进行实验,最终目标域球体和源域球体无限接近,核心集趋于稳定.

由图 4(c)、图 4(d)可知,随着 u 值的增大,虽然总误差会呈下降趋势,但无缺失部分的误差也会有所升高,当 u 值为 1 时,总误差下降幅度最大;当 u 值大于 8 时,总误差降幅趋缓.无缺失部分的误差的升高是由于随着 u 值增大源域影响不断增大,回归预测曲线会向源域回归预测曲线靠拢,故 u 值不是越大越好.选择合适的 u 值,既保证无缺失部分的预测性能,又能使缺失部分得到校正,这一点至关重要.故模拟数据集实验较合理的 u 参数取值为 8.

4.2 真实数据集实验

本部分实验数据来自大规模回归数据集网站 http://www.liaad.up.pt/~ltorgo/Regression/DataSets.html.实验数据集具体情况见表 2.

Table 2 Large scale regression dataset

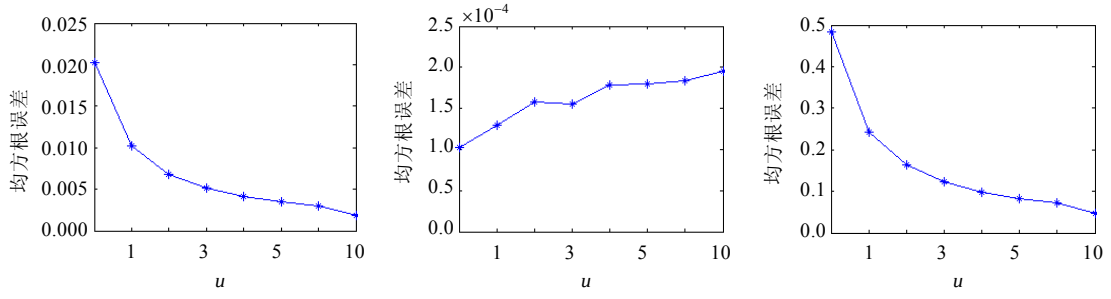
表 2 大规模回归数据集

数据集	数据集大小	属性个数
Census domains (8L)	22 784	8
Computer activity	8 192	21
Bank datasets (8FM)	8 192	8

4.2.1 Census Domains (8L)实验

Census Domains (8L)数据集为在美国人口普查局提供的数据基础上设计的一个数据库,其任务是根据该地区的人口构成和住房市场的状况,预测该地区房子的平均价格.在实验前将数据作如下处理:将 50%的数据作为源域数据,10%的数据作为测试数据,剩下 40%的数据作为目标域数据集.将数据集作适当调整,即将目标域房产平均价格大于 300 000 的样本数据全部放入源域数据,使得目标域数据集不包含高房价样本数据,而源域数据集包含较全面的高房价样本数据.调整 u 值,分 3 种情况对测试样本进行测试,分别是测试样本整体、房价小于 300 000 的测试样本及房价大于等于 300 000 的样本.测试结果如表 3、图 5 所示.

从图 5 及表 3 可总结出,测试样本的均方根误差较大是由于目标域训练集缺少 300 000 以上房源的样本数据引起的,而通过学习源域知识,对目标域 CC-MEB 进行中心点校正后,可以降低测试样本的均方根误差,尤其是当 u 值为 1 时,均方根误差由 0.020 3 迅速降为 0.010 2.由此可以得出结论:采用 A-CVR 算法,可以将源域知识传递给目标域,随着 u 值的增大,测试样本总体误差减小,缺失段样本的误差也随之减小;但通过图 5(b)可以看出,房价小于 300 000 的测试样本的误差会有所升高,故 u 值不是越大越好,过大的 u 值会过分增大源域的影响而降低目标域的作用.由第 4.1 节的 u 值选取方法,确定 u 值为 6 时,可取得较为满意的测试结果.



(a) 不同 u 值,所有测试集的均方根误差

(b) 不同 u 值,房价小于 300 000 的测试集的均方根误差

(c) 不同 u 值,房价大于 300 000 的测试集的均方根误差

Fig.5 Experimental results on dataset Census

图 5 Census 数据集实验结果

Table 3 RMSEs with different u values

表 3 不同 u 值下的均方根误差

RMSE	u=0	u=1	u=2	u=3
总误差	0.0203±4.83E−05	0.0102±0.00e−000	0.0068±9.14E−19	0.0051±4.83E−05
无缺失部分预测误差	1.03E−04±4.35E−05	1.29E−04±3.26E−05	1.58E−04±4.76E−05	1.55E−04±4.33E−05
缺失部分预测误差	0.4835±4.72E−04	0.2434±3.98E−04	0.1629±4.35E−04	0.1228±3.60E−04
RMSE	u=4	u=5	u=6	u=10
总误差	0.0041±0.00e−000	0.0035±3.16E−05	0.003±3.16E−05	0.0019±0.00e−000
无缺失部分预测误差	1.79E−04±4.74E−05	1.80E−04±2.39E−05	1.84E−04±5.82E−05	1.95E−04±3.31E−05
缺失部分预测误差	0.09856±3.88E−04	0.0825±2.25E−04	0.0711±3.73E−04	0.0462±1.84E−04

4.2.2 Computer activity 实验

Computer activity 数据集给出了一组计算机系统活动参数,其任务是预测在用户模式下的 CPU 运行时间.在实验前将数据作如下处理:将前 50%的数据作为源域数据,最后 10%的数据作为测试数据.源域对测试样本进行测试后的均方根误差 *RMSE* 值为 0.053 5.选用剩下的 40%数据中属性 8 的值小于等于 30 000 的 2 603 个样本进行训练.表 4 显示了不同 u 值下,分 3 种情况对测试样本进行测试的均方根误差,分别是测试样本整体均方根误差、属性 8 的值小于等于 30 000 的样本的均方根误差及属性 8 的值大于 30 000 的样本的均方根误差.测试结果如表 4 和图 6 所示.

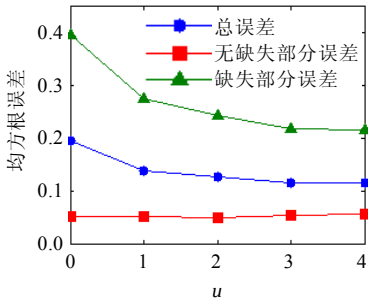


Fig.6 Experimental results on dataset Computer activity

图 6 Computer activity 数据集实验结果

从图 6、表 4 可知:当 u 值为 0 时,没有任何关于属性 8 的值大于 30 000 时的信息,误差较大;当 u 值为 1 时,测试样本的均方根误差的降幅最大;不断增大 u 的值,将源域知识传递给目标域,属性 8 的值大于 30 000 的测试样本的均方根误差减小,测试样本的总体均方根误差也随之减小,但属性 8 的值大于 30 000 的测试样本的均方根误差在 u 值为 3,4 时降幅缓慢,而属性 8 的值小于等于 30 000 的测试样本的均方根误差随着 u 值的增大呈上升趋势,且 u 值为 4 时,其误差增幅大于属性 8 的值大于 30 000 的测试样本的误差的降幅,从而使测试样本的总体均方根误差呈上升趋势.故 u 值不是越大越好,增大 u 值会加强源域的作用,而减弱目标域本身的作用.本例中,u 值取 3 时可获得最好的测试结果.

Table 4 RMSEs with different u values

表 4 不同 u 值下的均方根误差

RMSE	u=0	u=1	u=2	u=3	u=4
总误差	0.1962±0.0142	0.1399±0.0269	0.1276±0.0151	0.1144±0.023	0.1145±0.016
无缺失部分预测误差	0.0524±0.0037	0.0508±0.0045	0.0501±0.0041	0.0527±0.0037	0.0557±0.0044
缺失部分预测误差	0.3974±0.0306	0.2758±0.0588	0.2423±0.0323	0.2171±0.0529	0.2156±0.0347

4.2.3 Bank datasets 实验

Bank datasets (8FM)数据集是模拟银行客户如何选择银行的合成数据集,其任务是预测因为银行队列满而离开银行的可能性.在实验前将数据作如下处理:将 50%的数据作为源域数据,10%的数据作为测试数据,剩下 40%的数据作为目标域数据集.将目标域数据集作如下调整,即将目标域 Y 值介于 0.6~0.7 之间的数据去掉,形成缺失数据.调整 u 值,分 3 种情况对测试样本进行测试,分别是测试样本整体、Y 值小于 0.6 大于 0.7 的测试样本及 Y 值介于 0.6~0.7 之间的测试样本.测试结果如表 5、图 7 所示.

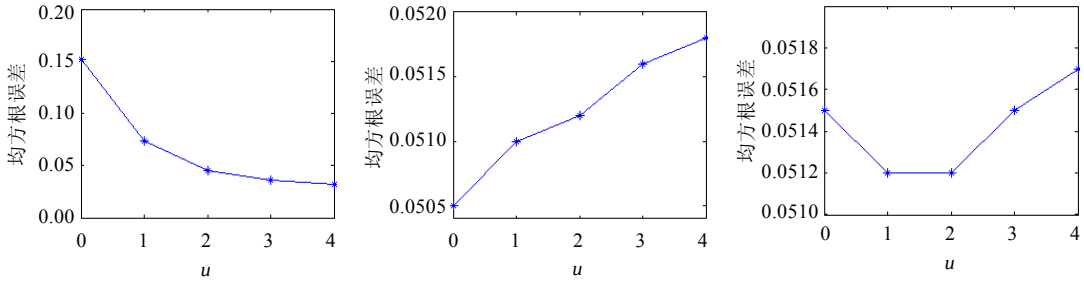
从表 5 和图 7(a)可以看出:随着 u 值的增大,Y 值介于 0.6~0.7 之间的测试样本的均方根误差呈不断减小的趋势,且在 u 值大于 2 后降幅趋缓;随着 u 值的增大,会增大源域的作用而使目标域作用降低,使得 Y 值小于 0.6 大于等于 0.7 的测试样本的均方根误差升幅明显(如图 7(b)所示),从而使所有测试样本的均方根误差在 u 值大于 2 后呈上升趋势(如图 7(c)所示).故 u 值不是越大越好,当 u 为 1 时,测试样本的均方根误差下降得最快.本实

验中,在 u 值取 1 或者 2 时就可得到较为满意的回归预测值.

Table 5 Regression performance on incomplete dataset with different u values

表 5 不同 u 值下缺失段数据集回归预测结果

真实值	0.634 1	0.608 4	0.602 8	0.617 5	0.663 5
u=0	0.473 1	0.581 3	0.581 4	0.467 1	0.629 6
u=1	0.555 5	0.599 3	0.589 3	0.549 5	0.632 2
u=2	0.586 7	0.602 5	0.600 4	0.581 4	0.631 9
u=3	0.600 4	0.600 4	0.608 6	0.595 6	0.630 4



(a) 不同 u 值,Y 值介于 0.6~0.7 之间的测试集的均方根误差

(b) 不同 u 值,Y 值小于 0.6 大于 0.7 的测试集的均方根误差

(c) 不同 u 值,所有测试集的均方根误差

Fig.7 Experimental results on dataset Bank

图 7 Bank 数据集实验结果

4.3 实验小结

本部分实验分为两个部分:(1) 通过 Benchmark 数据集,以图形的形式验证信息不全情况下 A-CVR 算法的有效性,且说明了参数设置的方法;(2) 通过真实大规模回归数据集进一步验证了 A-CVR 算法的有效性.实验结果表明,如果目标域数据集信息不全,则会导致相关测试样本预计不准确.采用 A-CVR 算法可以有效地学习源域知识,对目标域的 CC-MEB 进行中心点校正,随着 u 值的增大,缺失段测试样本的误差减小,所有测试样本的均方根误差也随之减小;但 u 值不是越大越好,过大的 u 值会增大源域的影响而降低目标域的作用,反而使测试样本的均方根误差呈上升趋势.可通过计算测试集在目标域的预测值与源域的预测值间的均方根误差来确定 u 值.在本文实验中,u 取 1 时,测试样本均方根误差降幅最明显.

5 结束语

在实际应用中,因时间、地点或设备不同,采集到的数据可能存在扰动或噪音,尤其是在回归问题中,存在采集数据不完整导致预测数据不准确的问题.针对该问题,本文提出了领域自适应的 A-CVR 算法.该算法不需要源域大量数据参与训练,仅需继承源域知识(中心点),使得本方法不仅能进行领域间的自适应学习,还能对源域起到隐私保护的作用.实验结果表明,该算法在两域相似的前提下,可提高信息不全情况下目标域模型的预测性能.

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] Tao JW, Wang ST. Multiple kernel local leaning-based domain adaptation. Ruan Jian Xue Bao/Journal of Software, 2012,23(9): 2297−2310 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4240.html [doi: 10.3724/SP.J.1001.2012.04240]

[2] 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]

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

[4] Tsang IW, Kwok JT, Zurada JM. Generalized core vector machines. IEEE Trans. on Neural Networks, 2006,17(5):1126−1140. [doi: 10.1109/TNN.2006.878123]

[5] Suzuki T, Sugiyama M, Tanaka T. Mutual information approximation via maximum likelihood estimation of density ratio. In: Proc. of the 2009 IEEE Int'l Symp. on Information Theory. Seoul, 2009. 463−467. [doi:10.1109/ISIT.2009.5205712]

[6] Suzuki T, Sugiyama M, Sese J, Kanamori T. Approximating mutual information by maximum likelihood density ratio estimation. In: Proc. of the JMLR Workshop and Conf. Antwerp, 2008. 5−20.

[7] Bădoiu M, Clarkson KL. Optimal core sets for balls. Computational Geometry: Theory and Applications, 2008,40(1):14−22. [doi: 10.1016/j.comgeo.2007.04.002]

[8] Tsang IW, Kwok JT, Cheung PM. Core vector machines: Fast SVM training on very large data sets. Journal of Machine Learning Research, 2005(6):363−392.

[9] Tsang IW, Kwok JT, Lai KT. Core vector regression for very large regression problems. In: Proc. of the 22nd Int'l Conf. on Machine Learning (ICML 2005). Bonn, 2005. 913−920. [doi: 10.1145/1102351.1102466]

[10] Tax D, Duin R. Support vector domain description. Pattern Recognition Letters, 1999,20(14):1191−1199. [doi: 10.1016/S0167-8655(99)00087-2]

[11] Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New support vector algorithms. Neural Computation, 2000,12(5):1207−1245. [doi: 10.1162/089976600300015565]

[12] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. Last updated, 2013. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] Mark G, He C. Probability density estimation from optimally condensed data samples. IEEE Trans. on PAMI, 2003,25(10): 1253−1264. [doi: 10.1109/TPAMI.2003.1233899]

[14] Deng ZH, Chung FL, Wang ST. FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation. Pattern Recognition, 2008,41:1363−1372. [doi: 10.1016/j.patcog.2007.09.013]

[15] Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximatie clustering and outlier detection in large datasets. IEEE Trans. on Knowledge and Data Engineering, 2003,15(5):1170−1187. [doi: 10.1109/TKDE.2003.1232271]

[16] Cressie NAC. Statistics for Spatial Data. New York: John Wiley and Sons, 1993.

[17] Marzio MZ, Taylor CC. Kernel density classification and boosting: An L2 analysis. Statistics and Computing, 2005,15(2):113−123. [doi: 10.1007/s11222-005-6203-8]

[18] JooSeuk K, Scott CD. L2 kernel classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(10):1822−1831. [doi: 10.1109/TPAMI.2009.188]

[19] Jiang YZ, Deng ZH, Wang ST. Mamdani-Larsen type transfer learning fuzzy system. Acta Automatica Sinica, 2012,38(9): 1393−1409 (in Chinese with English abstract).

附中文参考文献:

[1] 陶剑文,王士同.多核局部领域适应学习.软件学报,2012,23(9):2297−2310. http://www.jos.org.cn/1000-9825/4240.html [doi: 10.3724/SP.J.1001.2012.04240]

[19] 蒋亦樟,邓赵红,王士同.ML 型迁移学习模糊系统.自动化学报,2012,38(9):1393−1409.

许敏(1980－),女,江苏无锡人,博士生,讲师,主要研究领域为人工智能,模式识别.
E-mail: xum@wxit.edu.cn

顾鑫(1979－),男,博士生,工程师,主要研究领域为人工智能,模式识别.
E-mail: guxinbest@sina.com

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.
E-mail: wxwangst@yahoo.com.cn

俞林(1979－),男,讲师,主要研究领域为信息管理.
E-mail: yul@wxit.edu.cn