DOI:10.3969/j.issn.1673-4785.201312041
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20150317.1025.002.html

一种新颖的领域自适应概率密度估计器

许敏[1,2],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)

摘 要:传统概率密度估计法建立好密度估计模型后,无法将源域知识传递给相关目标域密度估计模型。提出用无偏置 v-SVR 的回归函数来表示传统概率密度估计法获得密度估计信息,并说明无偏置 v-SVR 等价于中心约束最小包含球及概率密度回归函数可由中心约束最小包含球中心点表示。在上述理论基础上提出中心点知识传递领域自适应概率密度估计法,用于解决因目标域信息不足而无法建立概率密度函数的场景。实验表明,此种领域自适应方法进行领域间知识传递的同时,还能达到源域隐私保护的目的。

关键词:概率密度函数;无偏置 v-SVR ;中心约束最小包含球;核心集;领域自适应
中图分类号: TP391.4 **文献标志码**:A **文章编号**:1673-4785(2015)02-0221-06

中文引用格式:许敏,俞林. 一种新颖的领域自适应概率密度估计器[J]. 智能系统学报, 2015, 10(2): 221-226.
英文引用格式:XU Min, YU Lin. A probability density estimator for domain adaptation[J]. CAAI Transactions on Intelligent Systems, 2015, 10(2): 221-226.

A probability density estimator for domain adaptation

XU Min[1,2], 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)

Abstract: This paper proposes that the density information received from the traditional probability density estimation method can be represented by no bias v-SVRregression function. It addresses the problem that after the source domain´s probability density estimation model is established using the traditional probability density estimation method its source domain knowledge can not be transferred to the relevant target domain´s density estimation model. In this paper, no bias v-SVR is equivalent to the center-constrained minimum enclosing ball (CC-MEB) and the probability density regression function is constrained by CC-MEB´s center point is described. On the basis of the above theory, an adaptive probability density evaluation method for transferring knowledge through the center point was put forward to solve the problem that an accurate probability density estimation model can not be established because of the lack of information of the target domain. The experiments showed that this adaptive method can reach the goals of knowledge transfer between domains and privacy protection in the source domain.

Keywords: probability density estimation; no bias v-SVR ; center-constrained minimum enclosing ball(CC-MEB); core set; domain adaptation

概率密度估计常见的做法是根据所得数据建立概率密度函数(probability density function, PDF),在机器学习和模式识别中具有非常重要的作用[1],如聚类分析[2]等。通常概率密度估计法分参数估计和非参数估计 2 类。因真实数据概率密度分布不可知,故非参数核密度估计法(kernel density estima-

收稿日期:2013-12-20. 网络出版日期:2015-03-17.
基金项目:江苏省高校自然科学研究资助项目(13KJB520001);江苏省高校哲学社会科学基金资助项目(2012SJB880077);江苏省研究生创新工程资助项目(CXZZ12-0759).
通信作者:许敏. E-mail:xum@ wxit.edu.cn.

tion, KDE)[3]是采用较广泛的方法。因 KDE 需要所有样本参与计算且需存储所有数据,故压缩集概率密度估计器[4]和快速压缩集概率密度估计器[5]被提出以解决存储空间和运行效率问题。上述传统的概率密度估计法效果显著但均未考虑领域间自适应学习的问题。在实际应用中存在这样的场景,已有源域数据集数据量大、密度估计精确;但相关目标域数据集由于隐私保护或数据遗失等原因只获得少量数据,这些数据是目标域真实信息但却不足以建立目标域 PDF。如何既保证目标域已知数据对建立目标域 PDF 的作用,又能利用源域知识对目标域信息不足部分加以弥补是本文研究的重点。

1 DADE 模型

1.1 DADE 模型理论依据

领域自适应概率密度估计器的应用前提是存在两相关领域,两域通过传统密度估计法,如 Parzen 窗法获得概率密度估计值,形成 $(\boldsymbol{x},y)$ 对。其中,$\boldsymbol{x}$ 是输入向量,y 是概率密度估计值。源域 $(\boldsymbol{x},y)$ 对足以构建概率密度函数,而出于隐私保护或数据遗失等原因,一些高度机密的数据无法获得,所得少量目标域 $(\boldsymbol{x},y)$ 信息精确,但不足以构建目标域概率密度函数。

传统密度估计法本身不能进行领域间知识传递,本文的贡献在于使用无偏置 v-SVR 回归函数表示概率密度函数,这样做的优势在于:

1)无偏置 v-SVR 等价于 CC-MEB 的特性,可使用核心集[6-8]代替源域所有数据建立概率密度函数,提高密度估计效率;

2)密度回归函数 $f(\boldsymbol{x})$ 可由 CC-MEB 中心点表示,提出中心点知识传递模型[9],实现相似领域间领域自适应概率密度器的建立,若使用源域核心集代替所有源域样本表示源域中心点,还可起到源域隐私保护的目的。

1.2 DADE 模型架构

设训练集 $T=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}$,其中输入向量 $\boldsymbol{x}_i\in R^n$,输出向量 $y_i\in Y=R$ 为概率密度估计值,$i=1,2,\cdots,l$。本文用无偏置支持向量回归函数 $y=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x})$ 建立概率密度估计函数,与传统 v-SVR 相比,没有 b 项,文章下面部分介绍无偏置 v-SVR。

1.2.1 无偏置 v-SVR

无偏置 v-SVR 试图寻找 R^n 上的一个实值函数 $g(\boldsymbol{x})$,以便使用 $y=g(\boldsymbol{x})$ 来推断任一输入 $\boldsymbol{x}$ 所对应的输出值 y。通常训练集在输入空间线性不可分,故引入映射函数 $\boldsymbol{\varphi}(\boldsymbol{x})$ 将 $\boldsymbol{x}_i$ 映射到高维空间 $\boldsymbol{\varphi}(\boldsymbol{x}_i)$ 中。无偏置 v-SVR 原始优化问题如下:

$$\begin{aligned}&\min\frac{1}{2}\|\boldsymbol{w}\|^2+\lambda\left(v\varepsilon+\frac{1}{l}\sum_{i=1}^{l}(\xi_i+\xi_i^*)\right)\\&\text{s.t.}\,\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_i)-y_i\leqslant\varepsilon+\xi_i\\&y_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_i)\leqslant\varepsilon+\xi_i^*\\&\xi_i^{(*)}\geqslant 0\end{aligned}\tag{1}$$

式中:$(*)$ 表示向量有 $*$ 号和无 $*$ 号 2 种情况。为导出原始问题(1)的对偶问题,引入拉格朗日函数:

$$\begin{aligned}&L(\boldsymbol{w},\xi^{(*)},\boldsymbol{\alpha}^{(*)},\boldsymbol{\eta}^{(*)})=\\&\frac{1}{2}\|\boldsymbol{w}\|^2+\lambda\left(\varepsilon+\frac{1}{vl}\sum_{i=1}^{l}(\xi_i+\xi_i^*)\right)-\\&\sum_{i=1}^{l}(\eta_i\xi_i+\eta_i^*\xi_i^*)-\sum_{i=1}^{l}\alpha_i(\varepsilon+\xi_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_i)+y_i)-\\&\sum_{i=1}^{l}\alpha_i^*(\varepsilon+\xi_i^*+\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_i)-y_i)\end{aligned}\tag{2}$$

式中:$\boldsymbol{\alpha}^{(*)}=[\alpha_1\ \ \alpha_1^*\ \ \cdots\ \ \alpha_l\ \ \alpha_l^*]^{\mathrm{T}}$,$\boldsymbol{\eta}^{(*)}=[\eta_1\ \ \eta_1^*\ \ \cdots\ \ \eta_l\ \ \eta_l^*]^{\mathrm{T}}$ 是拉格朗日乘子向量。

为了使式(2)最小化,对 L 关于向量 $\boldsymbol{w}$ 和变量 ε、$\xi_i^{(*)}$ 求偏导数,得

$$\partial L/\partial\boldsymbol{w}=0\Rightarrow\boldsymbol{w}=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)\boldsymbol{\varphi}(\boldsymbol{x}_i)\tag{3}$$

$$\partial L/\partial\varepsilon=0\Rightarrow\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)=\lambda\tag{4}$$

$$\partial L/\partial\xi_i^{(*)}=0\Rightarrow\frac{\lambda}{vl}-\eta_i^{(*)}-\alpha_i^{(*)}=0\tag{5}$$

将式(3)、(4)带入式(2),可得对偶优化问题:

$$\begin{aligned}&\min\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\\&\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i\\&\text{s.t.}\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)=\lambda\\&0\leqslant\alpha_i^{(*)}\leqslant\frac{\lambda}{vl}\end{aligned}\tag{6}$$

最终所得回归函数:

$$g(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x})=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x})\tag{7}$$

概率密度函数 $p(\boldsymbol{x})$ 需满足 $p(\boldsymbol{x})\geqslant 0$,$\int_{-\infty}^{+\infty}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$ 的条件,但无偏置 v-SVR 进行概率密度估计时不能满足上述条件,故需添加约束 $\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=1$,且核函数的选择满足 $K(\boldsymbol{x},\boldsymbol{x}')\geqslant 0$,$\int_{-\infty}^{+\infty}K(\boldsymbol{x},\boldsymbol{x}')\mathrm{d}\boldsymbol{x}=1$。

1.2.2　无偏置 v-SVR 与 CC-MEB

1) CC-MEB

Tsang 等在文献[6]中介绍了最小包含球(minimum enclosing ball, MEB)与中心约束最小包含球(center-constrained MEB, CC-MEB)。设 $S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$,其中 $\boldsymbol{x}_i\in R^d$,MEB 的思想是找到包含集合 S 所有样本 $\boldsymbol{\varphi}(\boldsymbol{x}_i)$ 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。为每个 $\boldsymbol{\varphi}(\boldsymbol{x}_i)$ 增加一维 δ_i,形成集合 $S'=\{(\boldsymbol{\varphi}(\boldsymbol{x}_i)',\delta_i)\}_{i=1}^m$,将最后一维中心点坐标设为 0,即中心点坐标$(c,0)$,则找到包含集合 S' 中所有样本的最小超球最优化问题为

$$\min_{c,R} R^2$$
$$\text{s.t.}\ \|\boldsymbol{\varphi}(\boldsymbol{x}_i)-\boldsymbol{c}\|^2+\delta_i^2\leqslant R^2,\ i=1,2,\cdots,m \tag{8}$$

设 $\boldsymbol{\Delta}=[\delta_1^2\quad\delta_2^2\quad\cdots\quad\delta_m^2]'\geqslant 0$,式(8)对应对偶问题的矩阵形式为

$$\max_{\boldsymbol{\beta}}\boldsymbol{\beta}^{\mathrm{T}}(\operatorname{diag}(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta}$$
$$\text{s.t.}\ \boldsymbol{\beta}\geqslant 0,\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{1}=1 \tag{9}$$

式中:核矩阵 $\boldsymbol{K}_{m\times m}=[k(\boldsymbol{x}_i,\boldsymbol{x}_j)]=[\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\quad\boldsymbol{\varphi}(\boldsymbol{x}_j)]$。

使用最优解 $\boldsymbol{\beta}$,可得到半径 R、中心点 $\boldsymbol{c}$ 的值:

$$R=\sqrt{\boldsymbol{\beta}^{\mathrm{T}}(\operatorname{diag}(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta}}$$
$$\boldsymbol{c}=\sum_{i=1}^m\boldsymbol{\beta}_i\boldsymbol{\varphi}(\boldsymbol{x}_i) \tag{10}$$

因为 $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{1}=1$,任意实数 η 加入公式,不会影响 $\boldsymbol{\beta}$ 的取值。原对偶形式改为

$$\max_{\boldsymbol{\beta}}\boldsymbol{\beta}^{\mathrm{T}}(\operatorname{diag}(\boldsymbol{K})+\boldsymbol{\Delta}-\eta\boldsymbol{1})-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta}$$
$$\text{s.t.}\boldsymbol{\beta}\geqslant 0,\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{1}=1\ ,\boldsymbol{\Delta}\geqslant 0 \tag{11}$$

文献[6]指出,任意满足式(11)的 QP 问题均能看作 CC-MEB 问题,可运用核心集快速算法求解。把整个数据集合 S 的求解转化成对 S 的一个子集 Q 的求解,可得到一个精确有效的近似解,其中 Q 被称为核心集。具体方法参见文献[6]。

2) 无偏置 v-SVR 与 CC-MEB 间关系

令 $\alpha_i^{(*)\prime}=\dfrac{\alpha_i^{(*)}}{\lambda}$,以满足 $\sum_{i=1}^{l}(\alpha_i'+\alpha_i^{*\prime})=1$,式(12)与式(6)等价。

$$\min\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^{*\prime}-\alpha_i')(\alpha_j^{*\prime}-\alpha_j')\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\frac{1}{\lambda}\sum_{i=1}^{l}(\alpha_i^{*\prime}-\alpha_i')y_i$$
$$\text{s.t.}\sum_{i=1}^{l}(\alpha_i'+\alpha_i^{*\prime})=1$$
$$0\leqslant\alpha_i^{(*)}\leqslant\frac{1}{vl} \tag{12}$$

令 $\tilde{\boldsymbol{\alpha}}=[\boldsymbol{\alpha}^{*\prime\mathrm{T}}\quad\boldsymbol{\alpha}'^{\mathrm{T}}]$,式(12)式相应的矩阵形式:

$$\min_{\tilde{\boldsymbol{\alpha}}}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{K}}\tilde{\boldsymbol{\alpha}}-\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\begin{bmatrix}\frac{2}{\lambda}\boldsymbol{Y}\\-\frac{2}{\lambda}\boldsymbol{Y}\end{bmatrix}$$
$$\text{s.t.}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{1}=1,0\leqslant\tilde{\boldsymbol{\alpha}}\leqslant\frac{1}{\lambda vl} \tag{13}$$

式中:$\tilde{\boldsymbol{K}}=[\tilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_j)]=\begin{bmatrix}\boldsymbol{K}&-\boldsymbol{K}\\-\boldsymbol{K}&\boldsymbol{K}\end{bmatrix}$。

式(13)为无偏置 v-SVR 的 QP 形式,与式(11)相比较,求 $\boldsymbol{\Delta}$ 的值:

$$\boldsymbol{\Delta}=-\operatorname{diag}(\tilde{\boldsymbol{K}})+\eta\boldsymbol{1}+\frac{2}{\lambda}\begin{bmatrix}\boldsymbol{Y}\\-\boldsymbol{Y}\end{bmatrix} \tag{14}$$

式中:实数 η 足够大,以使 $\boldsymbol{\Delta}\geqslant 0$。式就可以写成

$$\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}(\operatorname{diag}(\tilde{\boldsymbol{K}})+\boldsymbol{\Delta}-\eta\boldsymbol{1})-\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{K}}\tilde{\boldsymbol{\alpha}}$$
$$\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{1}=1 \tag{15}$$

该形式用 $\tilde{\boldsymbol{\alpha}}$ 替换了 $\boldsymbol{\beta}$ 与式(11)等价,是 CC-MEB 问题,可使用核心集快速法求解。

按式(15)求解,球心 $\boldsymbol{c}$ 可按下面公式计算:

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

式中 $i=1,2,\cdots,m$ 时 $\tilde{\boldsymbol{\varphi}}(\boldsymbol{x}_i)=\boldsymbol{\varphi}(\boldsymbol{x}_i)$,$i=m+1,m+2,\cdots,2m$ 时,$\tilde{\boldsymbol{\varphi}}(\boldsymbol{x}_i)=-\boldsymbol{\varphi}(\boldsymbol{x}_i)$,由此可得:

$$\boldsymbol{c}=\sum_{i=1}^{2*m}\tilde{\alpha}_i\tilde{\boldsymbol{\varphi}}(\boldsymbol{x}_i)=\sum_{i=1}^{m}\alpha'_i\boldsymbol{\varphi}(\boldsymbol{x}_i)+\sum_{i=1}^{m}\alpha^{*\prime}{}_i(-\boldsymbol{\varphi}(\boldsymbol{x}_i))=\sum_{i=1}^{m}(\alpha'_i-\alpha^{*\prime}{}_i)\boldsymbol{\varphi}(\boldsymbol{x}_i) \tag{16}$$

式(3)中的 $\boldsymbol{w}$ 就可简化为 $\boldsymbol{w}=\lambda\boldsymbol{c}$。故

$$g(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x})=\lambda\,\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x})=\lambda\sum_{i=1}^{m}(\alpha^{*\prime}{}_i-\alpha'_i)\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x})=\lambda\sum_{i=1}^{m}(\alpha^{*\prime}{}_i-\alpha'_i)\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{17}$$

由式(17)可获得以下两结论:

1)无偏置 v-SVR 等价于 CC-MEB,故可用核心集技术进行快速求解;

2)概率密度回归曲线可由其二次规划形式等价的 CC-MEB 的中心点表示。

1.2.3　DADE 模型

从 1.2.2 节分析可知,无偏置 v-SVR 等价于 CC-

MEB,概率密度函数由 CC-MEB 中心点表示。在此理论基础上,本文提出通过学习源域中心点将源域知识传递给目标域,构造学习源域知识且与目标域无偏置 v-SVR 等价的 CC-MEB,此 CC-MEB 的中心点可用于目标域概率密度函数的建立。

学习源域中心点的 CC-MEB 原始问题如下:

$$\min_{c,R} R^2 + \mu \| \boldsymbol{c} - \boldsymbol{c}_0 \|^2$$
$$\text{s.t.} \| \boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{c} \|^2 + \delta_i^2 \leqslant R^2 \quad (18)$$

引入拉格朗日乘子变量,在约束条件下构造式(18)的拉格朗日函数:

$$L = R^2 + \mu \| \boldsymbol{c} - \boldsymbol{c}_0 \|^2 + \sum_{i=1}^{l} \gamma_i (\| \boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{c} \|^2 + \delta_i^2 - R^2) \quad (19)$$

由最优化理论可知,式(19)在鞍点处取极值,在鞍点处 L 关于变量 $\boldsymbol{c}$ 和 R 的偏微分:

$$\frac{\partial L}{\partial R} = 2R - 2R\sum_{i=1}^{N}\gamma_i = 0 \Rightarrow \sum_{i=1}^{N}\gamma_i = 1$$

$$\frac{\partial L}{\partial \boldsymbol{c}} = 2\mu \| \boldsymbol{c} - \boldsymbol{c}_0 \| + 2\sum_{i=1}^{N}\gamma_i(\boldsymbol{\varphi}(\boldsymbol{x}_i) - \boldsymbol{c}) = 0$$

$$\Rightarrow \boldsymbol{c} = \frac{\mu \boldsymbol{c}_0 + \sum_{i=1}^{l}\gamma_i\boldsymbol{\varphi}(\boldsymbol{x}_i)}{\mu + 1} \quad (20)$$

将(20)代入(19),该问题的对偶形式为:

$$\max_{\gamma} \sum_{i=1}^{N} (\| \boldsymbol{\varphi}(\boldsymbol{x}_i) \|^2 - \frac{2\mu \boldsymbol{c}_0^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_i)}{\mu + 1} + \delta_i^2)\gamma_i - \frac{1}{\mu + 1}\sum_{i=1}^{l}\sum_{j=1}^{l}\gamma_i\gamma_j\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_j)$$
$$\text{s.t.} \sum \gamma_i = 1 \quad (21)$$

式中:$\boldsymbol{\varphi}(\boldsymbol{x}_i)\boldsymbol{\varphi}(\boldsymbol{x}_j) = \tilde{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$,$\boldsymbol{c}_0$ 由源域无偏置 v-SVR 训练按式(16)获得,δ_i 由目标域样本按式(14)获得。求解式(21)二次规划,按式(20)获得中心点带入式(22)即可获得目标域概率密度回归函数:

$$g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}) = \lambda \boldsymbol{c}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}) = \lambda \frac{\mu \boldsymbol{c}_0^{\mathrm{T}} + \sum_{i=1}^{l}\gamma_i\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}}{\mu + 1}\boldsymbol{\varphi}(\boldsymbol{x}) \quad (22)$$

2 实验与分析

2.1 实验设置

本文实验将本文所提算法与如下 3 个方面的回归函数进行性能对比:1)直接使用源域数据构建概率密度回归函数;2)直接使用包含少量信息的目标域数据构建概率密度回归函数;3)使用源域、目标域数据共同构建概率密度回归函数。从而来体现本文所提算法的优势。

实验中将 DADE 方法与上述相关的方法进行性能比较,以目标域测试集概率密度估计精度作为评价指标,具体为:error = $\frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x}_i) - \hat{f}(\boldsymbol{x}_i))^2$,其中 $\boldsymbol{x}_i$ 表示目标域测试集元素,$f(\boldsymbol{x}_i)$ 表示 $\boldsymbol{x}_i$ 的真实密度值,$\hat{f}(\boldsymbol{x}_i)$ 表示各算法所得 $\boldsymbol{x}_i$ 概率密度估计值,N 值为 500。实验通过网格搜索方式确定最优参数,高斯核函数的方差 h 在网格 $\{\bar{\boldsymbol{x}}/2\sqrt{2}, \bar{\boldsymbol{x}}/2, \bar{\boldsymbol{x}}/\sqrt{2}, \bar{\boldsymbol{x}}, \sqrt{2}\bar{\boldsymbol{x}}, 2\bar{\boldsymbol{x}}, 2\sqrt{2}\bar{\boldsymbol{x}}\}$ 中搜索选取,其中 $\bar{\boldsymbol{x}}$ 为训练样本平均 2 范数的平方根;λ 参数在网格{1,2,3,4,5,6,7,8,9,10}中搜索选取;v 参数在网格$\{1\times10^{-4}, 1\times10^{-3}, 1\times10^{-2}, 1\times10^{-1}, 1\}$中搜索选取;$\mu$ 参数在网格{1,2,3,4,5,6,7,8,9,10,15,20}中搜索选取。

实验环境为:Intel Core 2 2.40 GHz CPU, 2.39 GHz、1.94 GB RAM,Windows XP SP3, MATLAB 7.1。

2.2 实验结果与分析

为了利用源域知识弥补当前场景下信息过少造成受训系统泛化能力下降之缺陷,模拟数据集的构造需遵循以下原则:1)源域和目标域之间既有很大相似性,又存在区别;2)已知的目标域数据集 $(\boldsymbol{x}, y)$ 是精确的,但由于样本过少,不能构建出概率密度估计回归函数。

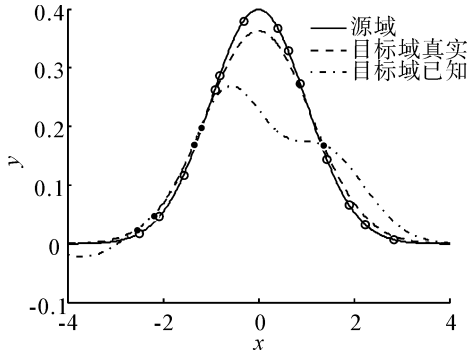
为了表征上述原则,首先生成样本数较多且能精确表示概率密度分布均值为 0、方差为 1 的源域数据集,需指出的是文章 1.2.2 节说明无偏置 v-SVR 与 CC-MEB 等价且概率密度函数可由 CC-MEB 中心点组成,若源域有数据隐私保护的需要,还可通过核心集技术,求得源域数据集的核心集,由少量核心集元素表示源域 CC-MEB 的中心点,进行迁移学习。另一方面,为了表示目标域与源域相近但不同,目标域设置时对均值、方差进行漂移,分均值、方差、均值方差均漂移 3 种情况,如表 1 所示。

表 1 数据源描述

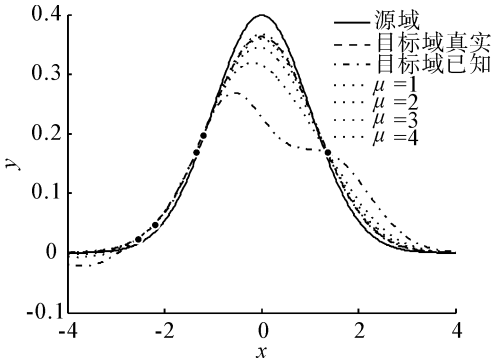
Table 1 Description of the data source

数据源	均值	方差
源域	0	1
目标域(均值漂移)	0.1	1
目标域(方差漂移)	0	1.1
目标域(均值、方差漂移)	0.1	1.1

由于隐私保护等原因,目标域获得信息量少且精确,但不足以构建目标域概率密度函数。图 1(a)虚线显示了均值为 0、方差为 1.1 时目标域真实概率密度分布图,图 1(b)显示了此种情况下目标域自适应学习效果图。图 2 将本文所提算法与另外 3 种训练方法进行比较。



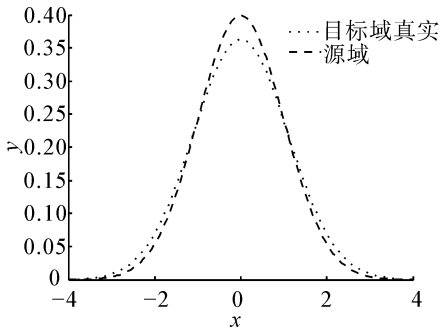
(a) 源域、目标域概率密度分布图



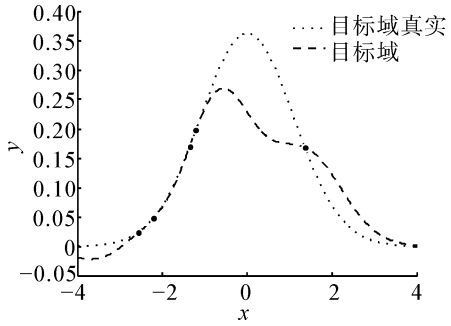
(b) 自适应学习效果图

图 1　均值为 0、方差为 1.1 自适应学习效果图

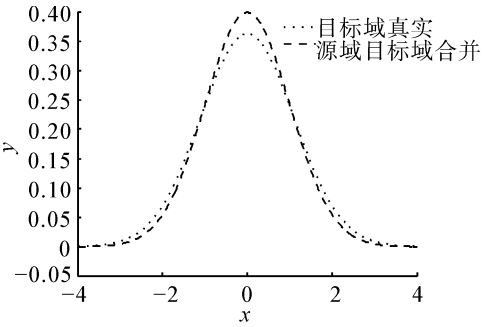
Fig.1　Charts of adaptive learning on the data set with mean 0, variance 1.1



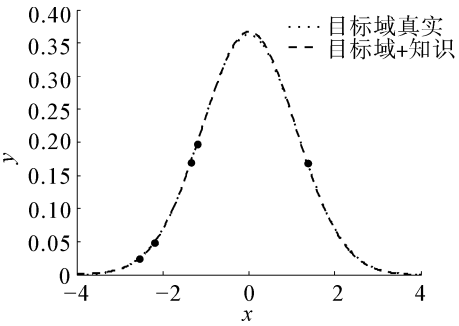
(a) 源域性能



(b) 目标域性能



(c) 源域目标域合并性能



(d) 自适应学习性能

图 2　原始图像和退化仿真图像

Fig.2　Performance comparison charts of different algorithms

表 2 列出了设置目标域不同均值方差后各算法的性能。

表 2　不同算法性能比较

Table 2　Performances comparison of different algorithms

均值	方差	源域 密度估计性能	目标域 密度估计性能	源域+目标域 密度估计性能	源域知识+目标域 密度估计性能	μ 值
0	1.1	$2.282\,1\times10^{-4}$	0.003 2	$2.279\,2\times10^{-4}$	$3.125\,5\times10^{-6}$	4
0	1.2	$7.959\,2\times10^{-4}$	0.002 9	$7.959\,1\times10^{-4}$	$1.609\,5\times10^{-6}$	2
0.1	1	$1.757\,4\times10^{-4}$	0.003 2	$1.\,757\,3\times10^{-4}$	$3.763\,5\times10^{-7}$	8
0.2	1	$7.003\,2\times10^{-4}$	0.003 5	$7.003\,0\times10^{-4}$	$6.646\,1\times10^{-5}$	20
0.1	1.1	$3.795\,2\times10^{-4}$	0.002 5	$3.257\,6\times10^{-4}$	$1.451\,1\times10^{-6}$	5
0.2	1.2	0.001 3	0.004 1	0.001 3	$5.445\,3\times10^{-6}$	7

生成均值为 0、方差为 1 源域样本 10 000 个,如图 1(a)所示,实线表示源域概率密度函数曲线,使用核心集技术获得源域的核心集由 13 个空心圆表示,源域知识只需知道模型参数和这 13 个样本点即可获得。虚线表示均值为 0、方差为 1.1 的目标域真实概率密度函数曲线。由图 1(a)可以看出,源域、目标域分布近似但不相同。图 1(a)中 5 实点表示目标域已知信息,为了体现数据隐私保护的目的,文

中实验选取的 5 个样本均在[-1,1]之外。点划线表示由这 5 个点获得的目标域概率密度函数曲线。由图可知,虽然已知信息精确,但信息过少不能反映目标域真实概率密度分布。图 1(b)显示了不同 μ 值自适应学习效果图,随着 μ 值的增大,目标域概率密度曲线向目标域真实分布靠拢。此种自适应学习的优势在于,既可保证目标域已知信息精确表示,又可通过源域知识对未知信息进行自适应学习,极大提高目标域概率密度估计性能。

根据表 2 和图 2,可给出如下的观察:

1) 从表 2 可知,本文提出的 DA-PDF 算法充分利用目标域已知信息的同时,学习了源域知识,较之于两域各自训练、合并训练所得概率密度估计函数具有更好的性能。

2) 对图 2(a)可知,若直接使用源域概率密度估计函数对现有测试集进行密度估计,效果不理想,其原因在于目标域与源域密度分布已发生变化(源域方差为 1,目标域方差为 1.1),这种变化导致若继续使用源域模型进行预测,其预测性能不好,无法达到与目标域实际情况逼近的效果。

3) 对图 2(b)可知,由于在当前场景下采集的数据数量较少,虽然这些数据真实可靠,但对于构建整个概率密度估计函数信息量过少,故密度估计性能低下。

4) 对图 2(c)可知,使用源域数据与目标域数据结合后生成的概率密度估计函数,其性能提升不明显。原因在于源域数据较之目标域收集到的数据,数据量大,因此在模型训练时,其所占的比重也大,故得到的概率密度估计函数最终更偏向于源域数据所得模型。合并训练另一缺点是需要源域所有数据参与模型的建立,但一些高度机密的历史数据通常难以获取,若源域有数据隐私保护的需要,此种方法则无法实现。

5) 从图 2(d)可知:本文方法较之图 2(a)有更好的逼近效果;与图 2(b)相比,可利用源域知识较好地弥补目标域信息不足的缺陷;与图 2(c)相比,不仅逼近程度有明显改进,且本文方法只需要历史知识(历史模型参数)以及目标域数据,并不需要源域数据作为训练数据,因而在隐私保护方面也体现了较大优势。

3 结束语

本文采用无偏置 v-SVR 对已知概率密度 $(\boldsymbol{x},y)$ 对进行概率密度函数建模,并证明无偏置 v-SVR 等价于 CC-MEB 且概率密度回归函数可由 CC-MEB 中心点表示,以此为前提,提出中心点领域自适应学习的概率密度估计函数建模思想,解决多领域相关联且某一领域信息较少无法构建概率密度函数的问题。本文所提方法不需要大量源域数据的支持,仅是继承历史知识(源域中心点),且允许当前领域信息较少,不但能够根据历史知识进行当前领域的信息补偿,又能对源域数据进行隐私保护,这些特性是传统概率密度估计方法所不具备的。通过合成数据的仿真实验表明本文方法较之于传统方法具有更好的适应性。

参考文献:

[1] VAPNIK V N. Statistical learning theory [M]. New York: John Wiley and Sons, 1998: 35-41.

[2] 吉根林, 姚瑶. 一种分布式隐私保护的密度聚类算法[J].智能系统学报, 2009, 4(2):137-141.
JI Genlin, YAO Yao. Density-based privacy preserving distributed clustering algorithm[J]. CAAI Transactions on Intelligent Systems, 2009, 4(2):137-141.

[3] PARZEN E. On estimation of a probability density function and mode[J]. The Annals of Mathematical Statistics, 1962, 33(3): 1065-1076.

[4] GIROLAMI M, HE C. Probability density estimation from optimally condensed data samples[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25 (10): 1253-1264.

[5] DENG Z H, CHUNG F L, WANG S T. FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation[J]. Pattern Recognition, 2008, 41(4):1363-1372.

[6] TSANG I W, KWOK J T, ZURADA J M. Generalized core vector machines [J]. IEEE Transactions on Neural Networks, 2006, 17(5): 1126-1140.

[7] TSANG I W, KWOK J T, CHEUNG P M. Core vector machines: fast SVM training on very large data sets[J]. Journal of Machine Learning Research, 2005(6): 363-392.

[8] CHU C S, TSANG I W, KWOK J K. Scaling up support vector data description by using core-sets[C]//IEEE International Joint Conference on Neural Networks. Budapest, Hungary: 2004: 425-430.

[9] 许敏,王士同. 基于最小包含球的大数据集域自适应快速算法[J]. 模式识别与人工智能, 2013, 26(2): 159-168.
XU Min, WANG Shitong. A fast learning algorithm based on minimum enclosing ball for large domain adaptation[J]. Pattern Recognition and Artificial Intelligence, 2013, 26(2): 159-168.

作者简介:

许敏:女,1980 年生,讲师,博士,主要研究方向为模式识别、人工智能。