文章编号: 1001-0920(2013)01-0125-06

基于RSDE的领域自适应概率密度估计方法

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 电子与信息技术学院，
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)

摘 要: 同一应用领域不同时间、地点或设备, 采集的样本数据可能存在扰动、噪音或缺失, 如何对样本数据集进行有效的预处理是其进一步应用的前提. 针对上述问题, 提出一种新的基于压缩集密度估计 (RSDE) 算法的领域自适应概率密度估计方法 A-RSDE, 通过学习源域 (训练域) 知识, 使目标域 (测试域) 概率密度估计更接近真实概率密度分布, 并用基于近似最小包含球的核心集快速算法求解 A-RSDE, 将其应用于大数据集密度估计. Benchmark 和 UCI 数据集上的实验表明, 该算法具有较好的性能.

关键词: 领域自适应；压缩集密度估计；最小包含球；核心集

中图分类号: TP391.4 **文献标志码:** A

Density estimation based on RSDE for domain adaption

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 Electronic and Information Technology，Wuxi Institute of Technology，Wuxi 214121，China；3. Research and Development Department，Wuxi Northern Lake Optical Co Ltd，Wuxi 214035，China．Correspondent：XU Min，E-mail：xum@wxit.edu.cn)

Abstract: Sample datasets are often collected from different times, places or devices. Due to the existence of the disturbance, noise and missing data, the collected datasets can not always keep the same distribution, and can even sometimes be required to concentrate them to reduce the computational burden, which can do the domain adaptation as the preprocessing step for the sample dataset before being fed into the next step. In order to achieve the above goal, a novel adaptive reduced set density estimator(A-RSDE) is proposed for adaptive probability density estimation by making full use of the source domain's (training dataset) knowledge of the probability density distribution, which lets the target domain's (testing dataset) probability density estimation be closer to the true probability density distribution. Meanwhile, the fast core-sets based minimum enclosing ball(MEB) approximation algorithm is introduced to develop the proposed algorithm. Finally, the experiment on the benchmark data sets and UCI data sets show that the proposed method has better performance.

Key words: domain adaption；reduced set density estimator；minimum enclosing ball；core-sets

0 引 言

概率密度估计是模式识别和机器学习领域的重要任务之一, 是分类、异常检测等的前提. Parzen[1]提出了 Parzen 窗法, 证明只要样本足够多便可保证其收敛于任何未知密度, 但该法需要完整的数据集参与计算, 对大数据集测试代价较高[2]. 对此, 张炤等[3]提出基于支持向量机的密度估计法, 利用支持向量机求解线性算子方程, 直接估计出与 Parzen 窗精度等级类似的稀疏解. Mark 等[4]提出了压缩集密度估计器 (RSDE), 只需权重因子不为 0 的数据样本参与计算, 在保证密度估计精度前提下, 达到数据浓缩目的. 该方法大大降低了计算时间和存储空间, 但其时间和空间复杂度仍为 $O(N^2)$. Deng 等[5]在 RSDE 基础上又提出了 FRSDE 算法, 揭示了 RSDE 与 CC-MEB 在计算几何上的相似性, 利用基于近似 MEB 的核心集快速算法求解 RSDE, 使其时间复杂度和样本规模呈线性关系, 空间复杂度不依赖于样本规模, 大大提高了大样本密度估计效率. 分类、回归等机器学习领域的前提是测试域和训练域为同一概率密度分布. 上述方法均未考虑同一应用领域因时间、地点或设备不

收稿日期: 2011-06-03；**修回日期:** 2011-10-09.

基金项目: 国家自然科学基金项目(60903100, 60975027, 61170122)；江苏省研究生创新工程项目(CXZZ12-0759).

作者简介: 许敏(1980−), 女, 讲师, 博士生, 从事人工智能与模式识别的研究；王士同(1964−), 男, 教授, 博士生导师, 从事人工智能与模式识别、生物信息等研究.

同, 采集的样本可能存在扰动、缺失和样本数较少不能反映真实概率密度分布等情况. 如何对测试域数据进行有效预处理是其进一步应用的前提, 如异常检测[6]、贝叶斯分类[6]、图像分割[7]等.

本文在RSDE的基础上, 提出一种新的领域自适应密度估计法A-RSDE(adaptive RSDE), 在保证目标域密度分布特点前提下, 能学习源域(训练域)已知的分布良好的密度估计结果, 并使目标域(测试域)密度估计与真实密度更接近. 利用基于最小包含球的核心集快速算法求解A-RSDE, 使该算法可用于大数据集.

1 压缩集密度估计

文献[4]首次提出RSDE, 文献[6]进一步将其应用于异常点检测与分类. 该方法试图使未知真实密度和RSDE估计密度间的最小累积平方误差达到最小, 利用该问题的最优值获得权重因子, 从而利用所有不为0的权重因子与数据集中对应的数据样本乘积之和获得密度估计值, 替代Parzen窗中所有样本, 既保证了密度估计精度, 又能达到数据浓缩的目的.

设数据集$S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N\}\in R^d$, RSDE密度估计函数为

$$\hat{p}(\boldsymbol{x};h,\boldsymbol{\gamma})=\sum_{i=1}^{N}\gamma_i K_h(\boldsymbol{x},\boldsymbol{x}_i).$$

其中: $\boldsymbol{\gamma}$为权向量, 且$\sum_{i=1}^{N}\gamma_i=1$, $\gamma_i\geqslant 0$; $K_h(\boldsymbol{x},\boldsymbol{x}_i)$为给定核函数; h为核宽. RSDE的对偶形式为

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_i\gamma_j\int_{R^d}K_h(\boldsymbol{x},\boldsymbol{x}_i)K_h(\boldsymbol{x},\boldsymbol{x}_j)\mathrm{d}x-\frac{2}{N}\sum_{i=1}^{N}\gamma_i\sum_{j=1}^{N}K_h(\boldsymbol{x}_i,\boldsymbol{x}_j).\tag{1}$$

2 A-RSDE

2.1 A-RSDE原理

设源域密度估计函数为$p(\boldsymbol{x};\theta_1)$, 目标域密度估计函数为$\hat{q}(\boldsymbol{x};\theta_2)$, 目标域真实密度函数为$q(\boldsymbol{x})$. 使用最小累积平方误差ISE[4], 使$\hat{q}(\boldsymbol{x};\theta_2)$最优逼近真实密度函数$q(\boldsymbol{x})$的同时, 与源域概率密度$p(\boldsymbol{x};\theta_1)$也最优逼近. 参数$\lambda$为源域对目标域的影响因子, λ值越大, 目标域密度分布与源域密度分布越靠近, 有

$$\underset{\theta}{\operatorname{argmin}}\int_{R^d}(q(\boldsymbol{x})-\hat{q}(\boldsymbol{x};\theta_2))^2\mathrm{d}x+\lambda\int_{R^d}(p(\boldsymbol{x};\theta_1)-\hat{q}(\boldsymbol{x};\theta_2))^2\mathrm{d}x.\tag{2}$$

当λ为0时, 式(2)简化为RSDE问题. 当目标域样本存在扰动、缺失、样本过少, 不能很好地体现密度分布时, 密度估计误差较大. 源域概率分布$p(\boldsymbol{x};\theta_1)$表示与目标域同一应用领域且已被验证能较好地反映密度分布的函数. 目标域与源域的密度分布应接近, 通过源域吸引可对目标域偏差进行校正.

设源域数据集$S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{N_0}\}\in R^d$, 目标域数据集$T=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{N_1}\}\in R^d$. 源域密度$p(\boldsymbol{x})$可以是任意已知真实密度, 也可以使用PW或RSDE等常用密度估计方法近似估计, 本文采用RSDE推导公式. 文献[8]指出, 不同形式的核函数对密度估计影响不大, 但高斯核的性质

$$\int G_h(\boldsymbol{x},\boldsymbol{x}_i)G_h(\boldsymbol{x},\boldsymbol{x}_j)\mathrm{d}x=G_{2h}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

可使计算复杂度大大降低. 将

$$p(\boldsymbol{x};\mathrm{ho},\boldsymbol{\alpha})=\sum_{i=1}^{N_0}\alpha_i K_{\mathrm{ho}}(\boldsymbol{x},\boldsymbol{x}_i),$$

$$\hat{q}(\boldsymbol{x};\mathrm{ht},\boldsymbol{\beta})=\sum_{j=1}^{N_1}\beta_j K_{\mathrm{ht}}(\boldsymbol{x},\boldsymbol{y}_j)$$

和高斯核函数$G_h(\boldsymbol{x},\boldsymbol{x}_i)$带入式(2), 得到A-RSDE对偶形式为

$$\underset{\beta}{\operatorname{argmin}}(1+\lambda)\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\beta_i\beta_j G_{2\mathrm{ht}}(\boldsymbol{y}_i,\boldsymbol{y}_j)-\frac{2}{N_1}\sum_{i=1}^{N_1}\beta_i\sum_{j=1}^{N_1}G_{\mathrm{ht}}(\boldsymbol{y}_i,\boldsymbol{y}_j)-2\lambda\sum_{i=1}^{N_0}\alpha_i\sum_{j=1}^{N_1}\beta_j G_{\mathrm{ho+ht}}(\boldsymbol{x}_i,\boldsymbol{y}_j).\tag{3}$$

其中: $\sum_{i=1}^{N_0}\alpha_i=1$, $\sum_{j=1}^{N_1}\beta_j=1$, N_0为源域数据集规模, $\boldsymbol{x}_i$为源域样本, ho为源域核宽, N_1为目标数据集规模, $\boldsymbol{y}_i$为目标域样本, ht为目标域核宽.

2.2 A-RSDE快速算法A-FRSDE

本节主要介绍如何使用基于MEB的核心集算法对A-RSDE进行大样本快速求解.

2.2.1 A-RSDE和CC-MEB之间的关系

Tsang等[9]揭示了满足一定条件的二次规划问题等价于MEB问题, 可利用Badoiu等[10]提出的计算近似MEB的迭代算法进行大数据集求解. Tsang等[11]进一步提出了中心约束最小包含球(CC-MEB)问题, 下面对CC-MEB作简单介绍.

1) CC-MEB.

设数据集$S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}\in R^d$, 对核空间任意样本点$\varphi(\boldsymbol{x}_i)$设置参数$\delta_i\in R$, 形成新样本$[\varphi(\boldsymbol{x}_i)\ \ \delta_i]^{\mathrm{T}}$, 同时将增加维的中心点固定在原点$[\boldsymbol{c}\ \ 0]^{\mathrm{T}}$处, $\boldsymbol{c}$为未扩展核空间中对应MEB的球心向量, 优化问题如下:

$$\underset{\boldsymbol{c},R}{\operatorname{argmin}}R^2;\quad \mathrm{s.t.}\ ||\varphi(\boldsymbol{x}_i)-c||^2+{\delta_i}^2\leqslant R^2,\ i=1,2,\cdots,N.\tag{4}$$

转换成相应的对偶形式成为QP问题, 即

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, \boldsymbol{\alpha}^{\mathrm{T}}(\operatorname{diag}(K)+\boldsymbol{\Delta})-\boldsymbol{\alpha}^{\mathrm{T}}K\boldsymbol{\alpha};$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \alpha_i\geqslant 0,\ \forall i. \tag{5}$$

其中: $\boldsymbol{\Delta}=[\delta_1^2,\delta_2^2,\cdots,\delta_N^2]^{\mathrm{T}}\geqslant 0$, 任意 $\eta\in R$. 下式与式 (5) 等价:

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, \boldsymbol{\alpha}^{\mathrm{T}}(\operatorname{diag}(K)+\boldsymbol{\Delta}+\eta\mathbf{1})-\boldsymbol{\alpha}^{\mathrm{T}}K\boldsymbol{\alpha};$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \alpha_i\geqslant 0,\ \forall i. \tag{6}$$

CC-MEB 使不满足 $K(\boldsymbol{x}_i,\boldsymbol{x}_i)$ 为常数的其他核方法也可以转换成 MEB 问题进行快速求解.

2) A-RSDE 与 CC-MEB 的关系.

设

$$C(\boldsymbol{y}_i,\boldsymbol{y}_j)=(1+\lambda)G_{2\mathrm{ht}}(\boldsymbol{y}_i,\boldsymbol{y}_j),$$
$$p(\boldsymbol{y}_i)=\frac{1}{N}\sum_{j=1}^{N_1}G_{\mathrm{ht}}(\boldsymbol{y}_i,\boldsymbol{y}_j)+\lambda\sum_{k=1}^{N_0}\alpha_k G_{\mathrm{ho+ht}}(\boldsymbol{x}_k,\boldsymbol{y}_i),$$

式 (3) 可以转换成如下形式:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{p}-\boldsymbol{\beta}^{\mathrm{T}}C\boldsymbol{\beta}. \tag{7}$$

A-RSDE 可以转换成如下 CC-MEB 形式:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \boldsymbol{\beta}^{\mathrm{T}}(\operatorname{diag}(K)+\boldsymbol{\Delta}+\eta\mathbf{1})-\boldsymbol{\beta}^{\mathrm{T}}K\boldsymbol{\beta};$$
$$\text{s.t. } \boldsymbol{\beta}^{\mathrm{T}}\mathbf{1}=1,\ \beta_i\geqslant 0,\ \forall i. \tag{8}$$

其中: $\Delta=-\operatorname{diag}(C)+2\boldsymbol{p}+\eta\mathbf{1},\ \eta\geqslant 0,\ \Delta\geqslant 0$.

比较式 (6) 和 (8), 用 C 替换 K, $\boldsymbol{\beta}$ 替换 $\boldsymbol{\alpha}$, 可见两表达式等价.

2.2.2 A-FRSDE 算法

A-RSDE $O(N^2)$ 的时间复杂度对于大数据集而言计算开销仍相当可观. 但观察式 (8) 能够发现, A-RSDE 可以视作 CC-MEB 问题, 使用基于近似 MEB 的快速核心集技术求解该问题[10,12].

A-FRSDE 算法的主要步骤如下.

输入: 源域密度估计 RSDE 法, 核心集 St_old, 权向量 $\boldsymbol{\alpha}$, 核宽 ho 或 PW 法, 数据集 Ref_old, 窗宽 hp, 目标域数据集 Ref_new, 逼近参数 ε, 正数 η, 核宽 ht, 参数 λ;

输出: 核心集 Q, 权向量 $\boldsymbol{\beta}$.

Step 1: 初始化 Q_0, c_0 和 R_0, 将迭代次数设为 1;

Step 2: 若无点 $\boldsymbol{y}$ 在球 $B(c_t,(1+\varepsilon)R_t)$ 外, 则转至 Step 6;

Step 3: 寻找离中心点 c_t 最远的点 $\boldsymbol{y}$, 将该点加入核心集 $Q_{t+1}=Q_t\bigcup\{\boldsymbol{y}\}$;

Step 4: 求解新的 CC-MEB(Q_{t+1}), 且 $c_{t+1}=c_{\mathrm{MEB}}(Q_{t+1})$, $R_{t+1}=R_{\mathrm{MEB}(Q_{t+1})}$;

Step 5: 令 $t=t+1$, 返回 Step 2;

Step 6: 终止训练, 返回所需要的输出.

主要步骤实现细节如下.

Step 1 中, 随机寻找一个点 $\boldsymbol{z}$, 找到与该点距离最远的点 $\boldsymbol{z}_a$ 和与 $\boldsymbol{z}_a$ 最远的点 $\boldsymbol{z}_b$, 构成初始集 $Q_0=\{\boldsymbol{z}_a,\boldsymbol{z}_b\}$. 对 Q_0 求 CC-MEB 得到 c_0 和 R_0.

Step 3 中, $\tilde{\varphi}(\boldsymbol{y}_l)=[\varphi(\boldsymbol{y}_l)\ \ \delta_l]^{\mathrm{T}}$ 到中心点 c_t 的距离为

$$||\tilde{\varphi}(\boldsymbol{y}_l)-c_t||^2=||\varphi(\boldsymbol{y}_l)-c_t'||^2+\delta_l^2=$$
$$C(\boldsymbol{y}_l,\boldsymbol{y}_l)-2\sum_{\boldsymbol{y}_i\in Q_t}\beta_i C(\boldsymbol{y}_i,\boldsymbol{y}_l)+$$
$$\sum_{\boldsymbol{y}_i,\boldsymbol{y}_j\in Q_t}\beta_i\beta_j C(\boldsymbol{y}_i,\boldsymbol{y}_j)+\delta_l^2, \tag{9}$$

其中 $\delta_l=-C(\boldsymbol{y}_l,\boldsymbol{y}_l)+2p(\boldsymbol{y}_l)+\eta$.

计算所有样本集 S 中点到中心点 c_t 的距离, 第 t 次迭代的时间复杂度为 $O(|Q_t|^2+|S||Q_t|)$, 当样本规模大时较为耗时. 实验采用 Smola 等提出的加速方法[13], 该方法指出, 在样本 S 中随机找到一个样本子集 S', 在子集 S' 中寻找离中心点 c_t 最远的点近似代替 S 中的最远点, 并证明当子集大小为 59 时, 最远点包含在 S' 中的可能性为 95 %, 时间复杂度降为 $O(|Q_t|^2+|S'||Q_t|)$, 本文实验中令 $|S'|=59$.

Step 4 中, MEB(Q_{t+1}) 通过式 (8) 的 QP 问题求得, 核心集规模远小于样本总体规模, 解决子问题的时间复杂度远小于解决所有样本 QP 问题的时间复杂度.

A-FRSDE 是将 CC-MEB 理论作用于大数据集 A-RSDE 的算法, 其系统开销可参考 CVM 的计算复杂度[10-11], 并具有以下性质.

性质 1 对于给定的 ε, 核心集规模上界为 $O(1/\varepsilon)$, 故浓缩集规模上界不会超过 $O(1/\varepsilon)$.

性质 2 对于给定的 ε, 时间复杂度上界为 $O(N/\varepsilon^2+1/\varepsilon^4)$, 与样本规模 N 呈线性关系.

性质 3 对于给定的 ε, 空间复杂度上界为 $O(1/\varepsilon^2)$, 可使用存储核心集代替所有样本.

3 实验分析

本文实验分为两部分, 第 1 部分按文献 [4] 提出的 Benchmark 数据集从扰动、缺失、数据样本少 3 个方面进行 A-RSDE 自适应验证、A-FRSDE 时间复杂度验证和浓缩验证; 第 2 部分进行 UCI 真实数据集验证.

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 等.

本文核函数均为高斯核, 快速算法中逼近参数 ε 均取 1e-6, 误差公式[4]为

$$L_1\mathrm{error}=\frac{1}{N}\sum_{i=1}^{N}|q(\boldsymbol{y}_i)-\hat{q}(\boldsymbol{y}_i)|.$$

其中: q 为真实密度估计函数, $\hat{q}$ 为 A-RSDE 密度估计函数, $\boldsymbol{y}_i$ 为测试样本.

3.1 Benchmark 数据集实验

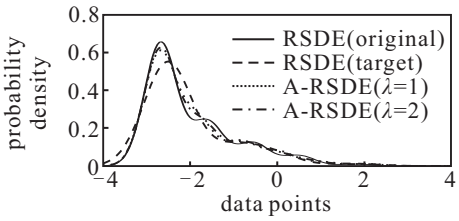
3.1.1 A-RSDE 验证

1) 一维高斯混合模型实验.

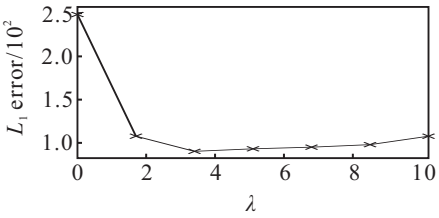
根据文献 [4] 提出的一维高斯混合模型

$$p(x)=\frac{1}{8}\sum_{i=0}^{7}G_{h_i}(u_i,x),\ h_i=\left(\frac{2}{3}\right)^i,\ u_i=3(h_i-1)$$

生成 200 个源域样本; 重新生成 200 个样本, 加上均值为 0、方差为 1 的扰动作为目标域样本; 另生成 10 000 个测试样本. 模型如图 1 所示.



(a) 目标域自适应图



(b) 不同λ值时的L_1误差

图 1 一维高斯混合模型实验

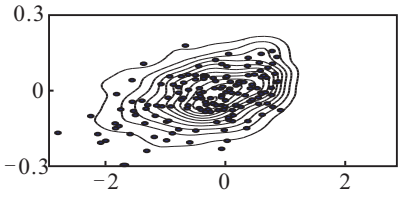
图 1 中, 实线为源域样本密度估计曲线, 高斯核宽为 0.37, L_1 误差为 0.011 7; 虚线为有扰动的目标域 RSDE 密度估计曲线, 高斯核宽为 0.55, L_1 误差为 0.025 1. 可知, 源域样本可以较好地反映样本密度, 而目标域样本密度估计曲线存在较大偏差. 由图 1 可见, 随着 λ 值的增大, 该算法可有效利用源域训练结果对目标域密度估计曲线进行校正. 当 λ 值为 1 时, 校正幅度最大, 误差由 0.025 1 迅速降为 0.010 6; 当 λ 值为 2 时, 误差达到最小 0.008 8, 且小于源域 L_1 误差; 随着 λ 继续增大, 误差越来越接近源域误差 0.011 7. 实验还发现, 由于目标域样本不能较好反映真实密度分布, λ 值固定时, 目标域核宽采用源域核宽, 校正效果最好.

2) 二维高斯模型实验.

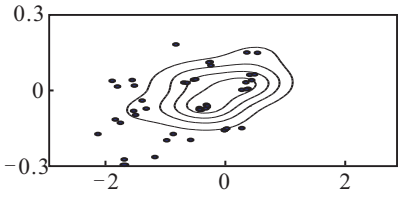
以文献 [4] 提出的中心点为原点, 协方差矩阵为

$$C_{ij}=\begin{cases}1, & i=j;\\ 0.5, & i\neq j.\end{cases}\tag{10}$$

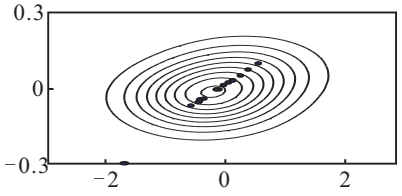
随机生成二维源域样本 300 个, 测试样本 10 000 个, RSDE 核宽为 0.82, L_1 误差为 0.010 2. 按相同方式生成二维样本 900 个, 去除第 1 维值大于 1 的样本, 形成 751 个样本组成目标域缺失数据集. 由于存在缺失, PW 和 RSDE 均不能精确估计样本的真实密度分布 (如图 2(a)和图 2 (b)). 其中: PW 窗宽为 0.242 1, L_1 为误差 0.019 0; RSDE 核宽为 0.46, L_1 误差为 0.023 2. 由图 2(c) 和图 2(d) 可见, 随着 λ 值增大, 可有效利用源域知识对目标域密度估计进行校正. 当 λ 值为 1 时校正幅度最大, 误差由 0.023 2 迅速降为 0.011 8; 当 λ 值为 4 时, 误差达到最小 0.009 5, 且小于源域的 L_1 误差; 随着 λ 继续增大, 误差又呈上升趋势, 越来越接近于源域误差 0.010 2. 图 2(a) 中所画点为所有点的 1/5, 其余图中所画点为权重因子不为 0 的点. 自适应时, 目标域核宽采用源域核宽.



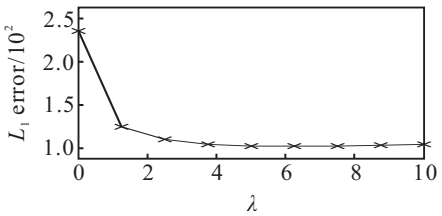
(a) PW密度估计等高线



(b) RSDE密度估计等高线



(c) λ=1密度估计等高线



(d) 不同λ值时的L_1误差

图 2 二维高斯模型实验

3) 五维高斯模型实验.

以文献 [4] 提出的中心点为原点, 协方差矩阵为式 (10) 五维高斯分布生成源域样本 700 个, RSDE 核宽为 0.82, 使用 10 000 个测试样本, L_1 误差为 0.001 4. 另生成 90 个同分布的目标域数据集, RSDE 核宽为 1, L_1 误差为 0.002 246. 当 λ 取不同值时, 使用 A-RSDE 算法校正后的 L_1 误差和不为 0 的权重因子比例见表 1. 表 1 表明, 当目标域样本数较少且不能最好体现样

本密度分布时, 使用A-RSDE算法可对目标域密度估计进行校正. 当λ值为1时, 校正幅度最大; 当λ值为5时, 误差不降反增. 不为0的权重因子比例在18%以内, 达到数据浓缩的目的.

表 1　五维高斯模型实验

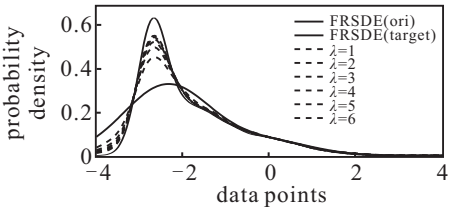
λ	1	2	3	4	5	6
L_1误差/10^{-3}	1.788	1.756	1.748	1.743	1.745	1.747
浓缩率/%	15.56	16.67	17.78	15.56	15.56	15.56

利用文献[4]提出的一维、二维、多维数据集, 对因扰动、缺失、样本数量少等引起的密度估计偏差采用A-RSDE算法进行校正, 实验需确定目标域样本核宽和λ值. 结果表明, 由于源域与目标域为同一应用领域, 其最佳核宽应基本相同, 但目标域样本不能很好地体现真实密度分布, 将目标域核宽设置为源域核宽校正效果最好. 选择合适的λ值能保留目标域数据集的特点; 同时, 学习源域知识也能对其偏差进行校正, λ值为1时, 校正幅度最大, 但λ值继续增大, 会使误差反增. 上述实验λ取值均在5以内.

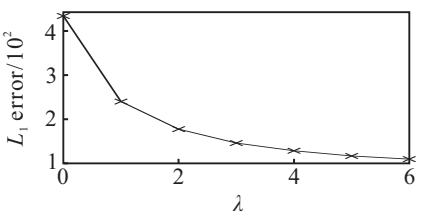
3.1.2　A-FRSDE验证

1) 自适应验证实验.

按文献[4]生成一维高斯混合源域数据集10 000个, 如式(10)所示. 使用FRSDE[5]进行密度估计, 图3(a)最上方实线为FRSDE概率密度分布图, 核宽为0.34, L_1误差为0.0057, 核心集规模为26. 重新生成10 000个该概率分布下的随机数, 并加上均值为0、方差为0.5的扰动作为目标域数据. 图3(a)下方实线显示为目标域FRSDE概率密度分布图, 核宽为1, L_1误差为0.044 1, 核心集规模为15. 可知, 扰动后的概率分布曲线与真实分布相差较大. 图3(a)虚线为A-FRSDE算法自适应后的概率密度分布图, 随着λ值的增大, 曲线向源域靠拢. 图3(b)显示不同λ值时的



(a)　目标域自适应图



(b)　不同λ值时的L_1误差

图 3　一维高斯混合模型大数据集实验

L_1误差. 由图3可见, 随着λ值增大, 目标域概率分布曲线向源域靠近. 当λ值为1时, 误差减幅最大; 当λ值为5时, 误差减幅趋缓, 可获得较理想的校正值. 因源域能较好地反映样本概率密度, 本文自适应实验中, 目标域核宽采用源域核宽.

2) 浓缩及时间复杂度验证.

根据

$$\text{浓缩率(CR)} = \frac{\text{浓缩集大小}}{\text{整个训练集大小}} \times 100\%,$$

将训练集大小由10 000~90 000递增, 表2为λ值是5时相应的浓缩率、训练时间和误差, 图4为算法时耗与不同样本容量的关系. 由表2可知, 使用A-FRSDE算法不仅能找到与真实密度接近的密度估计, 而且使用核心集代替所有样本达到了数据浓缩目的. 图4表明, A-FRSDE时间复杂度与样本容量N呈线性关系.

表 2　A-FRSDE算法的浓缩率及训练时间

样本规模	浓缩率/%	训练时间/s	误差/%
10 000	0.208 0±0.017 4	11.618 0±0.892 7	1.137 5±0.018 0
30 000	0.072 3±0.005 9	24.514 1±1.952 4	1.133 0±0.012 6
50 000	0.042 1±0.003 4	35.604 7±2.722 8	1.132 5±0.009 7
70 000	0.030 6±0.002 6	49.423 4±4.041 2	1.129 5±0.009 4
90 000	0.024 2±0.002 0	62.601 6±4.926 7	1.128 5±0.008 1

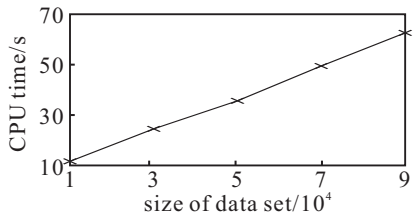


图 4　算法时耗与不同样本容量的关系

综上, A-RSDE的有效性、自适应性, A-FRSDE的线性时间复杂度、浓缩率均在benchmark数据环境中得到证实. 下面在贝叶斯分类应用中验证其性能.

3.2　UCI数据集实验

文献[6]指出, 在贝叶斯分类器中采用RSDE算法估算类条件概率密度适用于多分类, 实验数据来自UCI真实数据集[14].

1) A-RSDE实验.

选用数据集前70%作为源域数据, 后30%中随机选取4/5作为目标域数据、1/5作为测试数据进行5次实验. 采用RSDE与A-RSDE两种方法训练目标域数据, λ取值和分类精度如表3所示.

表 3　两种方法分类精度比较

数据集	RSDE	A-RSDE	λ
Iris	0.966 7±0.04	0.983 3±0.03	2
Wine	0.96±0.05	0.98±0.04	5

2) A-FRSDE实验.

实验数据来自UCI数据集 Waveform 21. 选取大数据集前70%作为源域数据, FRSDE核宽0.4, 分类精度0.77 ± 0.01; 选取24%的数据并将各属性加上均值为0, 方差为0.4的扰动作为目标域数据, FRSDE核宽为0.4, 分类精度为0.75 ± 0.025; 其余6%作为测试数据.

采用A-FRSDE自适应算法, 当λ值为1时, 分类精度为0.76 ± 0.013; 当λ值为2时, 分类精度为0.77 ± 0.011.

真实数据集实验表明, 采用A-RSDE(A-FRSDE)算法, 可充分利用源域概率分布知识对目标域概率密度估计进行校正, 从而使目标域能更好地进行下一步应用, 如贝叶斯分类等.

4 结 论

本文在RSDE算法的基础上, 提出了A-RSDE算法, 并在Benchmark数据集和贝叶斯分类实验中验证了该算法的有效性. 本文主要研究成果可归纳为两点: 1) 在目标域存在扰动、缺失、样本少而不能获得精确概率密度分布情况下, 利用该算法可以充分学习源域知识并予以校正; 2) 发现该算法与CC-MEB在计算几何上等价, 并可使用基于近似MEB的核心集快速算法进行求解, 应用于大数据集密度校正, 效果较好.

参考文献(References)

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

[2] Jeon B, Landgrebe A. Fast Parzen density estimation using clustering based branch and bound[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1994, 16(9): 950-954.

[3] 张炤, 张素, 章琛曦, 等. 基于支持向量机的概率密度估计方法[J]. 系统仿真学报, 2005, 17(10): 2355-2357.
(Zhang Z, Zhang S, Zhang C X, et al. Density estimation based on support vector machines[J]. J of System Simulation, 2005, 17(10): 2355-2357.)

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

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

[6] He C, Girolami M. Novelty detection employing an L_2 optimal nonparametric density estimator[J]. Pattern Recognition Letters, 2004, 25(12): 1389-1397.

[7] 钱鹏江, 王士同, 邓赵红. 大数据集快速均值漂移谱聚类算法[J]. 控制与决策, 2010, 25(9): 1307-1312.
(Qian P J, Wang S T, Deng Z H. Fast mean shift spectral clustering on large data sets[J]. Control and Decision, 2010, 25(9): 1307-1312.)

[8] Kollios G, Gunopulos D. Efficient biased sampling for approximatie clustering and outlier detection in large datasets[J]. IEEE Trans on Knowledge and Data Engineering, 2003, 15(5): 1170-1187.

[9] Tsang I, Kwok J, Cheung P. Core vector machines: Fast SVM training on very large data sets[J]. J of Machine Learning Research, 2005, 6(4): 363-392.

[10] Badoiu M, Clarkson K L. Optimal core sets for balls[J]. Computational Geometry: Theory and Applications, 2008, 40(1): 14-22.

[11] Tsang I, Kwok J, Zurada J. Generalized core vector machines[J]. IEEE Trans on Neural Networks, 2006, 17(5): 1126-1139.

[12] Badoiu M, Har-Peled S, Indyk P. Approximate clustering via core-sets[C]. Proc of 34th Annual ACM Symposium on Theory of Computing. Montreal: ACM Press, 2002: 250-257.

[13] Smola A, Schölkopf B. Sparse greedy matrix approximation for machine learning[C]. Proc of the 17th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2000: 911-918.

[14] Blake C, Merz C. UCI repository of machine learning databases[EB/OL]. http://archive.ics.uci.edu/ml/datasets.html, 2011-2-16.

(上接第124页)

[17] Tang M Z, Yang C H, Gui W H, et al. Data-based process fault detection using active cost-sensitive learning[C]. The 2nd Int Conf on Information Science and Engineering. Hangzhou, 2010: 1110-1113.

[18] Aleix M Martinez, Avinash C K. PCA versus LDA[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(2): 228-233.