文章编号: 1001-0920 (2012) 02-0221-06

隐私团校准的模糊MEB学习

胡文军[1,2], 王士同[1]

(1. 江南大学 信息工程学院，江苏 无锡 214122；2. 湖州师范学院 信息与工程学院，浙江 湖州 313000)

摘　要: 在一定条件下, 基于最小累积平方误差 (ISE) 准则的高斯核密度估计与最小包含球 (MEB) 等价. 在此基础上提出了一种含团状隐私数据保护的MEB 学习方法, 称为隐私团校准的 MEB (PCC-MEB) 方法; 同时, 通过引入模糊隶属度函数将 PCC-MEB 拓展为模糊的 PCC-MEB(FPCC-MEB), 从而解决二类及多类问题中区域不可分问题. 人造和真实数据集上的实验结果表明, 所提出方法具有较好的性能.

关键词: 最小包含球；核密度估计；隐私数据团；核方法；模糊

中图分类号: TP391.4　　　**文献标识码:** A

Privacy cloud calibration fuzzy learning for MEB

HU Wen-jun[1,2], *WANG Shi-tong*[1]

(1. School of Information Engineering，Jiangnan University，Wuxi 214122，China；2. School of Information and Engineering，Huzhou Teachers College，Huzhou 313000，China．Correspondent：HU Wen-jun，E-mail：hoowenjun@yahoo.com.cn)

Abstract: Under given conditions, Gaussian kernel density estimate with minimum integrated square error(ISE) criterion can be equivalent to the minimum enclosing ball(MEB). Based on this conclusion, a learning method of MEB with privacy cloud data is proposed, called privacy cloud calibration MEB(PCC-MEB). Meanwhile, PCC-MEB is extended to fuzzy privacy cloud calibration MEB(FPCC-MEB) by introducing a fuzzy membership function, which can resolve unclassifiable zones among classes. Experimental results on the artificial and real-word data sets show the effectiveness of presented method.

Key words: minimum enclosed ball(MEB)；kernel density estimator；privacy data cloud；kernel method；fuzzy

1 引　言

最近, 在机器学习领域, 通过对已知可能性概率数据团的学习而解决分类问题受到了较大的关注[1-4]. 虽然数据团中样本的类标签不确定, 但数据团属于某类的概率可能性已知, 即已知数据团的类标签频率分布, 因此这类机器学习是一种介于监督和无监督学习之间的方法. 在现实中, 含数据团的机器学习例子很多, 如文献 [1] 中的投票选举, 对于一个地区 (或区域) 每张选票结果不知道, 但该地区 (或区域) 的选票结果是清楚的, 则选票结果与投票人的收入、家庭类型等存在着某种关系, 这种关系往往反应到每张选票结果的分布; 又如文献 [1] 中鉴别骗子的例子, “某个或某些人是骗子”的结论与“某些人中有 5 人是骗子的可能性概率为 p”的结论相比, 前者比后者触犯当地法律的风险大得多; 再如, 在故障检测中, 某些样本中故障的可能性是 p, 但针对某个样本并不能确定其是否有故障等.

为了解决这类问题, 本文先揭示最小累积平方误差 (ISE) 准则下的高斯核密度估计与核化的最小包含球 (MEB) 等价, 并在此基础上提出隐私团校准的 MEB 方法 (PCC-MEB), 基本思想是构造反映隐私信息的最小包含球, 并转化为一个二次规划 (QP) 问题, 其突出优势在于可以运用CVM(core vector machine) 方法有效解决大样本问题. 本文方法侧重在解决含有隐私信息的二类及多类问题, 在解决此类问题时, PCC-MEB 和 MEB 一样存在不可分的样本区域, 故引入模糊隶属度函数, 将 PCC-MEB 拓展为模糊的 PCC-MEB(FPCC-MEB).

2 算法回顾

给定训练样本 $X = \{\boldsymbol{x}_i|\boldsymbol{x}_i \in R^d\}$, $i = 1, 2, \cdots$,

收稿日期: 2010-08-26；**修回日期:** 2010-11-04.

基金项目: 国家自然科学基金项目(60903100, 60975027)；江苏省普通高校研究生科研创新计划项目(CXZZ11_0483).

作者简介: 胡文军(1977−), 男, 讲师, 博士生, 从事模式识别、人工智能等研究；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.

m, $\boldsymbol{x}_i$ 是列向量. 首先, 简要回顾两种机器学习方法.

2.1 MEB

MEB 是在样本空间中找到一个最小球体, 设 r 和 $\boldsymbol{c}$ 分别为球的半径和球心, 并使得球体将所有目标类训练样本包络. 因为样本在原始空间很难做到准确划分, 为此可以引入所谓的核技巧, 此时 MEB 优化模型为

$$\min_{r,\boldsymbol{c}}\ r^2,$$
$$\text{s.t. } \|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2 \leqslant r^2,\ 1 \leqslant i \leqslant m. \tag{1}$$

可见, 式 (1) 与一类硬划分的 SVDD 类似[5]. 构造拉格朗日函数, 并通过拉格朗日技巧可得球心

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

以及对偶形式

$$\boldsymbol{\alpha}=\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\,\boldsymbol{\alpha}^{\mathrm{T}}\operatorname{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha},$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \alpha_i \geqslant 0,\ 1 \leqslant i \leqslant m. \tag{3}$$

其中: $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_m]^{\mathrm{T}}$, $\alpha_i \geqslant 0$ 为 m 维拉格朗日乘子向量; $\mathbf{1}=[1,1,\cdots,1]^{\mathrm{T}}$ 为 m 维列向量; $\boldsymbol{K}=[k_{ij}]_{m\times m}=[k(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{m\times m}$ 为 $m\times m$ 维核矩阵. 当 $k(\boldsymbol{x}_i,\boldsymbol{x}_i)=k$($k$ 为某一常数) 时, 式 (3) 为

$$\boldsymbol{\alpha}=\underset{\boldsymbol{\alpha}}{\operatorname{argmax}}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha},$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \alpha_i \geqslant 0,\ 1 \leqslant i \leqslant m. \tag{4}$$

根据 KKT 条件可知, 对于任何满足 $\alpha_k>0$ 对应的 $\boldsymbol{x}_k$, 有

$$r^2=k(\boldsymbol{x}_k,\boldsymbol{x}_k)-2\sum_{i=1}^{m}\alpha_i k(\boldsymbol{x}_k,\boldsymbol{x}_i)+\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{5}$$

给定待测样本 $\boldsymbol{x}\in R^d$, 可通过如下决策函数:

$$f(\boldsymbol{x})=r^2-\|\varphi(\boldsymbol{x})-\boldsymbol{c}\|^2=R^2-k(\boldsymbol{x},\boldsymbol{x})+2\sum_{i=1}^{m}\alpha_i k(\boldsymbol{x},\boldsymbol{x}_i)-\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{6}$$

进行判决, 若 $f(\boldsymbol{x})\geqslant 0$, 则属于目标类; 否则属于其他类.

2.2 FKHP

2004 年 Chung 等人提出模糊核超球感知器 (FKHP) 算法[6], 其核心是通过迭代算法获得各类最小超球的半径和球心, 并引入模糊隶属度函数解决模糊样本的判决. 该方法不需要解决二次规划问题和收敛性问题, 更重要的是在解决多类问题时不需要像典型支持向量机 (SVM) 那样通过成对组合实现, 因而在速度上快于典型 SVM. 其算法如下:

Step 1: 通过迭代规则获取各个核超球的半径和球心;

Step 2: 对于非模糊样本, 通过各超球判决函数判决, 并标上相应的类标签;

Step 3: 对于模糊样本, 通过模糊隶属度大小判决, 并标上相应的类标签.

该算法的迭代规则请参考文献 [6], 而测试样本的模糊性判断以及模糊隶属度函数等内容将在第 3.4 节中详述.

3 PCC-MEB

3.1 核密度估计

若原始样本空间 $\bar{X}=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in R^d\}$ 的某个采样空间为 $X=\{\boldsymbol{x}_i|i=1,2,\cdots,m\}\subset\bar{X}$, 则其核密度估计函数[8,13]为

$$\hat{p}(\boldsymbol{x};h;\gamma)=\sum_{i=1}^{m}\gamma_i K_h(\boldsymbol{x},\boldsymbol{x}_i). \tag{7}$$

其中: $\boldsymbol{\gamma}=[\gamma_1,\gamma_2,\cdots,\gamma_m]$ 为权向量, 且 $\sum_{i=1}^{m}\gamma_i=1$, $\gamma_i\geqslant 0$; $K_h(\boldsymbol{x},\boldsymbol{x}_i)$ 为给定的核函数, h 为给定核的带宽.

给定核函数 $K_h(\boldsymbol{x},\boldsymbol{x}_i)$ 及 h 后, 为使 $\hat{p}(\boldsymbol{x};h;\boldsymbol{\gamma})$ 最优逼近真实密度函数 $p(\boldsymbol{x})$, 可通过不同准则[7]实现, 如最大近邻估计准则 (MLE), 最小累积均方误差 (IMSE), 最小累积平方误差 (ISE), 最小累积绝对误差 (IAE). 其中, ISE 是一种保证全局精度的估计准则[8], 基于此准则提出了多种性能较佳的机器学习算法[8-10], 本文也是针对于该准则.

定理 1 若 $\hat{p}(\boldsymbol{x};h;\boldsymbol{\gamma})$ 为 $p(\boldsymbol{x})$ 在 $p(\boldsymbol{x})$ 条件下的无偏估计量, 则 ISE 准则下的密度估计等价于 MEB.

证明 1) ISE 准则. 为了保证 $\hat{p}(\boldsymbol{x};h;\boldsymbol{\gamma})$ 尽可能逼近 $p(\boldsymbol{x})$, γ 应尽可能保证 ISE 最小, 即

$$\hat{\gamma}=\underset{\gamma}{\operatorname{argmin}}\,\mathrm{ISE}(\boldsymbol{\gamma})=\underset{\gamma}{\operatorname{argmin}}\int_{R^d}\|p(\boldsymbol{x})-\hat{p}(\boldsymbol{x};h;\boldsymbol{\gamma})\|^2\mathrm{d}\boldsymbol{x}=\underset{\gamma}{\operatorname{argmin}}\left\{2E_{p(\boldsymbol{x})}[\hat{p}(\boldsymbol{x};h;\boldsymbol{\gamma})]-\int_{R^d}\hat{p}^2(\boldsymbol{x};h;\boldsymbol{\gamma})\mathrm{d}\boldsymbol{x}\right\}. \tag{8}$$

2) 等价 MEB. 不失一般性, 本文选高斯函数为核函数, 即 $K_h(\boldsymbol{x},\boldsymbol{x}_i)=G_h(\boldsymbol{x},\boldsymbol{x}_i)$, 则

$$\int_{R^d}\hat{p}^2(\boldsymbol{x};h;\boldsymbol{\gamma})\mathrm{d}\boldsymbol{x}=\int_{R^d}\sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_i\gamma_j K_h(\boldsymbol{x},\boldsymbol{x}_i)K_h(\boldsymbol{x},\boldsymbol{x}_j)\mathrm{d}\boldsymbol{x}=\sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_i\gamma_j G_{2h}(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{9}$$

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

$$E_{p(\boldsymbol{x})}[\hat{p}(\boldsymbol{x};h;\boldsymbol{\gamma})]=E[p(\boldsymbol{x})]. \tag{10}$$

将式 (9) 和 (10) 代入 (8), 得

$$\hat{\gamma}=\underset{\gamma}{\operatorname{argmax}}-\sum_{i=1}^{m}\sum_{j=1}^{m}\gamma_i\gamma_jG_{2h}(\boldsymbol{x}_i,\boldsymbol{j}_j),$$
$$\text{s.t.}\ \sum_{i=1}^{m}\gamma_i=1,\ \gamma_i\geqslant 0,\ i=1,2,\cdots,m. \tag{11}$$

对比式 (11) 和 (4), 可知定理 1 成立. □

由定理 1 可知, MEB 对偶形式的乘子 $\boldsymbol{\alpha}$ 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

$$\hat{p}(\boldsymbol{x})=\sum_{i=1}^{m}\alpha_i\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x})=\boldsymbol{c}^{\mathrm{T}}\varphi(\boldsymbol{x}). \tag{12}$$

3.2 PCC-MEB 模型

假设在训练样本为 X 中包含一个隐私数据团 $S\subset X$, 而 $\bar{S}=X-S$ 中的样本属于目标类. 已知 S 属于目标类的可能性概率为 p, 根据式 (12), 可知 p 的估计值为

$$p=\frac{\sum\limits_{\boldsymbol{x}_i\in S}\hat{p}(\boldsymbol{x}_i)}{|S|}\times\frac{m}{|S|}=\frac{m\boldsymbol{c}^{\mathrm{T}}\sum\limits_{\boldsymbol{x}_i\in S}\varphi(\boldsymbol{x}_i)}{|S|^2}. \tag{13}$$

令 $\boldsymbol{u}=\sum\limits_{\boldsymbol{x}_i\in S}\varphi(\boldsymbol{x}_i)$ 和 $P=p|S|^2/m$, 则含有隐私团的MEB模型为

$$\min_{r,\boldsymbol{c}}\ r^2,$$
$$\text{s.t.}\ \|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2\leqslant r^2,\ 1\leqslant i\leqslant m;$$
$$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{u}=P. \tag{14}$$

构造式 (14) 的拉格朗日函数, 即

$$J(r,\boldsymbol{c})=r^2+\sum_{i=1}^{m}\alpha_i(\|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2-r^2)+\beta(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{u}-P). \tag{15}$$

式 (15) 对原始变量的偏导数为零, 整理后可得

$$\sum_{i=1}^{m}\alpha_i=1, \tag{16}$$

$$\boldsymbol{c}=\sum_{i=1}^{m}\alpha_i\varphi(\boldsymbol{x}_i)-\frac{\beta\boldsymbol{u}}{2}. \tag{17}$$

将式 (17) 代入 (14) 的等式约束, 整理得

$$\beta=2\Big\{\sum_{i=1}^{m}\alpha_i\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{u}-P\Big\}\Big/\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}. \tag{18}$$

将式 (16)~(18) 代入 (15), 可得式 (14) 的对偶形式, 即

$$\max_{\alpha}\ \boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\boldsymbol{\Delta}')-\boldsymbol{\alpha}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\alpha};$$
$$\text{s.t.}\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1,\ \alpha_i\geqslant 0,\ 1\leqslant i\leqslant m. \tag{19}$$

其中

$$P_{SS}=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=\sum_{\boldsymbol{x}_i\in S}\sum_{\boldsymbol{x}_j\in S}k(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

$$\boldsymbol{\Delta}'=-\frac{2P}{P_{SS}}\Big[\sum_{\boldsymbol{x}_j\in S}k(\boldsymbol{x}_i,\boldsymbol{x}_j)\Big]_{m\times 1},$$

$$\tilde{\boldsymbol{K}}=\boldsymbol{K}-\frac{1}{P_{SS}}\Big[\sum_{\boldsymbol{x}_j\in S}k(\boldsymbol{x}_i,\boldsymbol{x}_j)\Big]_{m\times 1}\Big[\sum_{\boldsymbol{x}_j\in S}k(\boldsymbol{x}_i,\boldsymbol{x}_j)\Big]_{m\times 1}^{\mathrm{T}}.$$

再令

$$P_{LXS}=\sum_{i=1}^{m}\alpha_i\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\sum_{\boldsymbol{x}_j\in S}\varphi(\boldsymbol{x}_j),$$

根据式 (17) 和 KKT 条件, 有

$$R^2=k(\boldsymbol{x}_k,\boldsymbol{x}_k)-2\sum_{i=1}^{m}\alpha_ik(\boldsymbol{x}_k,\boldsymbol{x}_i)+\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)+\frac{P^2-P_{LXS}^2}{P_{SS}}-\frac{2(P-P_{LXS})}{P_{SS}}\sum_{\boldsymbol{x}_i\in S}k(\boldsymbol{x}_k,\boldsymbol{x}_i), \tag{20}$$

其中 $\boldsymbol{x}_k$ 为任意满足 $\alpha_k>0$ 条件的样本. 类似 MEB, 其决策函数为

$$f(\boldsymbol{x})=r^2-k(\boldsymbol{x},\boldsymbol{x})+2\sum_{i=1}^{m}\alpha_ik(\boldsymbol{x},\boldsymbol{x}_i)-\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jk(\boldsymbol{x}_i,\boldsymbol{x}_j)-\frac{P^2-P_{LXS}^2}{P_{SS}}+\frac{2(P-P_{LXS})}{P_{SS}}\sum_{\boldsymbol{x}_i\in S}k(\boldsymbol{x},\boldsymbol{x}_i). \tag{21}$$

若 $f(\boldsymbol{x})\geqslant 0$, 则属于目标类.

对应比较式 (2), (5), (6) 和 (17), (20), (21) 可知, 隐私数据团的可能性概率校准了 MEB 的球心、半径和决策函数, 故将此方法称为隐私团校准的 MEB 方法 (PCC-MEB).

3.3 复杂度分析

PCC-MEB 实际上是求解式 (19) 对应的 QP 问题, 可知式 (19) 除了二次项外还存在一次项, 因此其空间复杂度为 $O(m^2+m)$, 大于 MEB 的空间复杂度 $O(m^2)$. 而 QP 求解的时间复杂度为 $O(m^3)$[11-14], 因此 PCC-MEB 很难适用于大样本的训练. 再看决策函数 (21), 对于 1 个未知样本, 其决策复杂度为 $O(m+|S|)$, 大于 MEB 的决策复杂度 $O(m)$.

为了解决对大样本的训练, 可令

$$\boldsymbol{\Delta}=[\delta_i]_{m\times 1}^{\mathrm{T}}=-\mathrm{diag}(\tilde{\boldsymbol{K}})+\eta\boldsymbol{1}+\boldsymbol{\Delta}'+\mathrm{diag}(\boldsymbol{K}),$$

其中 η 为某一常量, 且使得 $\delta_i\geqslant 0$. 式 (19) 改写为

$$\max_{\alpha}\ \boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\tilde{\boldsymbol{K}})+\boldsymbol{\Delta}-\eta\boldsymbol{1})-\boldsymbol{\alpha}^{\mathrm{T}}\tilde{\boldsymbol{K}}\boldsymbol{\alpha};$$
$$\text{s.t.}\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1,\ \alpha_i\geqslant 0,\ 1\leqslant i\leqslant m. \tag{22}$$

因为 $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1$, 故式 (22) 中一次项中 $-\boldsymbol{\alpha}^{\mathrm{T}}\eta\boldsymbol{1}$ 可以省略. 此时, 对比式 (22) 和文献 [12] 中的式 (17), 可知式 (22) 是中心约束最小球 (CC-MEB) 问题. 因此, 运用

CVM 方法可解决大样本训练. 此外, 文献 [15] 提出的支持向量预选取方法也可融合到本文算法的加速训练. 由于本文重点讨论带隐私保护的分类问题, 关于大样本问题不进行展开, 这将作为后续的研究重点.

3.4 模糊 PCC-MEB

由 3.2 节可知, PCC-MEB 实际上也是一个 MEB 问题, 因此, 在解决二类及多类问题时, 与 MEB 一样存在区域不可分问题[6]. 为此, 本文采用文献 [6] 中引入模糊隶属度函数的方法, 将 PCC-MEB 拓展为模糊隐私团校准的 MEB(FPCC-MEB) 方法. 在整个样本空间中, 可能出现多个 PCC-MEB 重叠或不被任何一个 PCC-MEB 覆盖的区域, 因此该区域中的样本 (称为模糊样本, 其他区域称为非模糊样本) 不能确定其属于哪一类. 为此, 引入模糊隶属函数

$$\mu_j(\boldsymbol{x}) = \begin{cases} \dfrac{r_j}{2d(\boldsymbol{x}, \boldsymbol{c}_j)}, \ d(\boldsymbol{x}, \boldsymbol{c}_j) > r_j; \\ 1 - \dfrac{d(\boldsymbol{x}, \boldsymbol{c}_j)}{2r_j}, \ \text{otherwise}. \end{cases} \tag{23}$$

其中: $d(\boldsymbol{x}, \boldsymbol{c}_j)$ 为 $\boldsymbol{x}$ 到第 j 个 PCC-MEB 球心的距离, r_j 为第 j 个 PCC-MEB 的半径. 此时, FPCC-MEB 的算法为:

Step 1: 求解式 (19) 或 (20) 的 QP 问题, 并根据式 (21) 求解各个 PCC-MEB 的半径.

Step 2: 对于非模糊样本, 通过如下对应 PCC-MEB 的判决函数式进行判决:

$$a^+ = \frac{\#\ \text{positive samples correctly classified}}{\#\ \text{total positive samples classified}} \times 100\%, \tag{24}$$

并标上相应的类标签.

Step 3: 对于模糊样本, 通过下式计算的模糊隶属度大小进行判决:

$$a^- = \frac{\#\ \text{negative samples correctly classified}}{\#\ \text{total negative samples classified}} \times 100\%, \tag{25}$$

并标上相应的类标签.

4 实验结果与分析

本文选取高斯函数为核函数, 带宽从 $\{s/128, s/64, s/32, s/16, s/8, s/4, s/2, s, 2s, 4s, 8s\}$ 中选择, 其中 s 是训练样本平均 2 范数的平方. 实验环境为: Pentium Core2 2.6 GHz CPU, 2 G RAM, Windows XP, Matlab2009a. 考虑到样本的不平衡, 采用几何平均精度 $G = \sqrt{a^+ \cdot a^-}$ 作为最终的精度, 其中 a^+ 和 a^- 分别采用式 (24) 和 (25) 进行计算.

4.1 p 特性实验

p 值实际上代表了数据团的隐私程度, p 越小隐私程度越大. 利用图 1 的人造数据集研究 p 对分类精度的影响, 图中“+”和“◇”形状样本为目标类, “◇”样本固定 100 个, 固定的 30 个“o”样本和“+”样本构成隐私团 S, 其中“+”个数由概率 p 决定 (即 $p/(1-p) \times 30$). 实验中 p 取值为 0.1, 0.2, $\cdots$, 0.9, 并从“◇”中随机取 30 % 作为测试样本, 剩余 70 % 和隐私团 S 构成训练样本. 图 2 给出了实验结果, 由图 2 可知, p 值越大分类精度 g 越高, 建议取 0.8 以上.

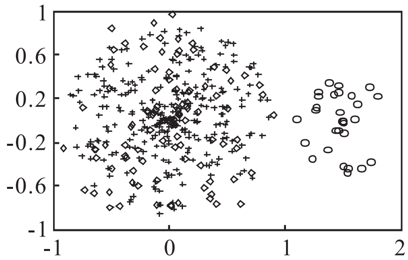

图 1 人造圆形数据集

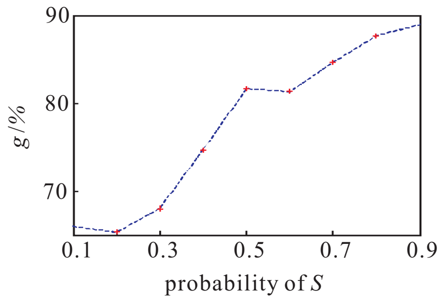

图 2 p 对分类精度的影响曲线

4.2 性能比较实验

表 1 给出了实验数据集, 可从 http://archive.ics.uci.edu/ml 和 http://prlab.tudelft.nl/users/david-tax 下载, 其中 PBRHD 为 Pen-Based Recognition of Handwritten Digits (0 和 1) 数据集, Landsat Satellite 红土和非常湿的灰土. 为了衡量 FPCC-MEB 的性能, 实验从测试精度, 训练和测试时间 (单位: s) 3 个方面与 FKHP 和 v-SVC 进行了比较, 结果以均值和标准差给出, 其中 v-SVC 算法[16-17]中的参数 $v = 0.2$, 因为 FPCC-MEB 实验中需要构造隐私团 S, 所以根据构造方法不同, FPCC-MEB 实验分为紧密隐私团和松弛隐私团两种.

紧密隐私团的实验数据构造: 1) 对于 FPCC-MEB, 先从目标类中随机抽取 30 % 用于测试, 再从剩

表 1 实验数据集

数据集	维数	样本总数	+1 样本数	−1 样本数
Arrhythmia	278	420	237	183
B.Cancer	9	699	241	458
Biomed	5	194	127	67
Connectionist Bench	60	208	111	97
PBRHD	16	1 559	779	780
Waveform	21	3 304	1 657	1 647
Landsat Satellite	36	3 041	1 533	1 508
Spectf Heart	44	267	212	55
Wine	13	178	119	59

表 2　FPCC-MEB, FKHP 和 v-SVC 三种算法的性能比较

数据集	FPCC-MEB (紧密隐私团)			FPCC-MEB (松弛隐私团)		
	g/%	训练时间	测试时间	g/%	训练时间	测试时间
Arrhythmia	**68.00±4.30**	5.534 4±0.242 7	2.500 0±0.045 6	**66.48±4.39**	5.328 1±0.343 7	1.537 5±0.366 9
B.Cancer	**96.78±1.11**	24.531 3±2.511 0	2.335 9±0.909 8	**96.73±1.09**	33.823 4±2.380 4	2.195 3±0.881 2
Biomed	**82.58±5.60**	0.918 7±0.329 7	0.203 1±0.042 8	**82.83±7.36**	0.906 3±0.226 9	0.289 1±0.134 1
Connectionist Bench	70.97±5.42	0.967 2±0.204 3	0.218 8±0.037 6	71.15±5.18	1.067 2±0.214 6	0.217 2±0.031 6
PBRHD	**99.70±0.12**	191.390 6±10.160 5	10.306 3±0.116 2	**99.91±0.19**	173.259 4±4.224 4	9.812 5±0.038 3
Waveform	**89.38±0.28**	2 341.585 9±5.292 3	44.812 5±0.044 2	**87.88±0.52**	2 500.666 7±49.084 4	45.401 0±0.023 9
Landsat Satellite	**91.80±2.89**	1 830.546 9±103.187 3	40.583 3±0.130 1	**92.25±2.09**	1 818.802 1±9.986 9	41.713 5±0.121 4
Spectf Heart	**77.65±4.54**	2.135 9±0.245 4	0.392 2±0.045 7	**76.54±3.68**	2.182 8±0.320 5	0.357 8±0.051 3
Wine	**88.72±3.63**	0.767 2±0.190 6	0.203 1±0.054 6	**92.15±4.05**	0.754 7±0.222 0	0.282 8±0.120 7

数据集	FKHP			v-SVC		
	g/%	训练时间	测试时间	g/%	训练时间	测试时间
Arrhythmia	**71.97±3.66**	3.026 6±0.250 7	0.835 9±0.008 2	**68.93±3.02**	6.385 9±0.639 4	1.093 8±0.310 3
B.Cancer	**96.86±0.78**	17.289 1±1.464 2	1.360 9±0.587 4	**97.18±0.72**	228.806 3±36.163 9	1.353 1±0.578 8
Biomed	**83.63±3.90**	0.600 0±0.201 1	0.148 4±0.039 8	**86.52±2.86**	1.167 2±0.285 1	0.157 8±0.034 9
Connectionist Bench	**77.49±6.19**	0.671 9±0.177 7	0.162 5±0.048 4	**86.15±3.66**	0.898 4±0.281 6	0.118 7±0.008 1
PBRHD	**100.00±0.00**	140.690 6±6.846 4	5.990 6±0.090 8	**99.86±0.12**	789.677 1±33.662 7	5.953 1±0.062 5
Waveform	**88.63±1.66**	1 789.026 0±71.993 8	26.822 9±0.063 1	**91.70±0.47**	12 341.929 7±2.110 3	27.781 3±0.640 8
Landsat Satellite	87.77±1.75	1 394.208 3±70.961 5	25.145 8±0.862 4	**90.90±0.84**	9 041.979 2±156.302 7	24.276 0±0.032 5
Spectf Heart	**75.14±5.31**	1.450 0±0.313 6	0.270 3±0.045 4	**75.80±6.45**	1.945 3±0.303 2	0.239 1±0.043 6
Wine	**90.56±2.80**	0.535 9±0.200 0	0.135 9±0.030 4	**90.81±4.22**	0.925 0±0.298 7	0.123 4±0.037 2

余70％中抽取50％相似度大的样本和非目标类中随机抽出的部分样本(样本数量根据隐私程度p计算)构成数据团S(本文称为紧密隐私团), 剩余20％和隐私团S一起构成训练样本; 2) 对于FKHP和v-SVC算法, 除测试样本外的70％构成训练样本. 松弛隐私团的实验数据构造类似于紧密隐私团, 只是在剩余70％中随机抽取50％的样本和非目标类中随机抽出的部分样本构成数据团S(本文称为松弛隐私团). 两种实验均取$p = 0.85$, 表2给出了实验结果.

从表2可看出: 1) 除Connectionist Bench数据集外, 不论紧密隐私团还是松弛隐私团, FPCC-MEB的几何平均精度同其他两种算法相比是可以接受的, 由此说明了本文方法的误分风险程度较低. 2) 在训练速度方面, 本文算法慢于FKHP算法, 但快于v-SVC算法. 这是因为FPCC-MEB隐私团中的非目标类样本不参与FKHP的训练; 相比于v-SVC算法, 由于其将两类训练样本一起进行QP求解实现, 而本文算法则是通过2个PCC-MEB (即2个子QP)问题实现的, 因此速度快于v-SVC, 样本较大时尤为突出, 如表中PBRHD, Waveform和Landsat Satellite数据集. 3) 在测试速度方面, FPCC-MEB需要校准决策函数, 所以测试速度慢于其他两种算法.

5　结　论

本文证明了ISE准则下的概率密度估计等价于MEB问题, 并在此基础上建立PCC-MEB数学模型用于解决含隐私数据团的分类问题. 从实验结果看, 本文算法有效, 且错分的风险程度较低. 总体而言, 本文建立了概率密度估计函数与MEB之间的联系; 提出的算法能有效解决含隐私保护数据团的分类问题. 文中虽然说明了本文方法可以拓展到CVM版本实现对大样本的学习, 但没有深入, 同时如何提高本文方法测试效率都将成为下一步研究的重点.

参考文献(References)

[1] Rüping S. SVM classifier estimation from group probabilities[C]. Proc of 27th ICML, Haifa, 2010: 911-918.

[2] Quadrianto N, Smola A J, Caetano T S, et al. Estimating labels from label proportions[C]. Proc of 25th ICML. Omnipress, 2008: 776-783.

[3] 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.

[4] Hendrik K, Nando de F. Learning about individuals from group statistics[C]. Proc of 21st Annual Conf on Uncertainty in Artificial Intelligence. Arlington: AUAI Press, 2005: 332-339.

[5] David M J T, Robert P W D. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.

[6] Chung F L, Wang S T, Deng Z H, et al. Fuzzy kernel hyperball perceptron[J]. Applied Soft Computing, 2004, 5(1): 67-74.

[7] Alan J I. Recent developments in nonparametric density estimation[J]. J of the American Statistical Association,

1991, 86(413): 205-224.

[8] Mark G, He C. Probability density estimation from optimally condensed data samples[J]. IEEE Trans on PAMI, 2003, 25(10): 1253-1264.

[9] JooSeuk K, Clayton S. Kernel classification via integrated squared error[C]. Proc of IEEE Workshop on Statistical Signal Processing. Madison, 2007: 783-787.

[10] JooSeuk K, Clayton S. Robust kernel density estimation[C]. Proc of IEEE ICASSP. Las Vegas, 2008: 3381-3384.

[11] Ivor W T, James T K, Cheung P M. Core vector machines: Fast SVM training on very large data sets[J]. J of Machine Learning Research, 2005, 6: 363-392.

[12] Ivor W T, James T K, Zurada J M. Generalized core vector machines[J]. IEEE Trans on Neural Networks, 2006, 17(5): 1126-1140.

[13] 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: 1363-1372.

[14] Chung F L, Deng Z H, Wang S T. From minimum enclosing ball to fast fuzzy inference system training on large datasets[J]. IEEE Trans on Fuzzy Systems, 2009, 17(1): 173-184.

[15] 蔡艳宁, 胡昌华, 汪洪桥, 等. 基于支持向量预选取的支持向量域故障预报[J]. 控制与决策, 2009, 24(7): 985-989.
(Cai Y N, Hu C H, Wang H Q, et al. Support vector domain fault prediction based on support vector preextracting[J]. Control and Decision, 2009, 24(7): 985-989.)

[16] Schölkopf B, Smola A J, Williamson R C, et al. New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1207-1245.

[17] Chih C C, Chih J L. Training v-support vector classifiers: Theory and algorithms[J]. Neural Computation, 2001, 13(9): 2119-2147.

(上接第210页)

[7] Collobert R, Sinz F, Weston J, et al. Large scale transductive SVMs[J]. J of Machine Learning Research, 2006, 7(8): 1687-1712.

[8] Li Y Q, Guan C, Li H, et al. A self-training semi-supervised svm algorithm and its application in an eeg-based brain computer interface speller system[J]. Pattern Recognition Letters, 2008, 29(9): 1285-1294.

[9] Adankon M M, Cheriet M. Help-training for semi-supervised discriminative classifier application to svm[C]. Proc of the 19th Int Conf on Pattern Recognition. Piscataway: IEEE, 2008.

[10] Wang X S, Tian X L, Cheng Y H. Value approximation with least squares support vector machine in reinforcement learning system[J]. J of Computational and Theoretical Nanoscience, 2007, 4(7/8): 1290-1294.

[11] Zhou Z H, Li M. Semi-supervised regression with co-training style algorithm[J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(11): 1479-1493.

[12] Suykens J A K, Vandewale J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.

(上接第215页)

[7] Dayhoff J E. Neural network architectures: An introduction[M]. New York: Van Nostrand, 1990.

[8] Proakis J G. Digital communications[M]. 3rd ed. New York: McGraw-Hill, 1995.

[9] Atiya A F. Bankruptcy prediction for credit risk using neural networks: A survey and new results[J]. IEEE Trans on Neural Networks, 2001, 12(4): 929-935.

[10] Standard & Poor's Research Insight SM COMPUSTAT. A primer for getting started[EB/OL]. http://www2.library.unr.edu/dataworks/compu_primna76.pdf.

[11] Weinberger K Q, Sha F, Saul L K. Learning a kernel matrix for nonlinear dimensionality reduction[C]. Proc of the 21st Int Conf on Machine Learning. Banff, 2004.

[12] Chang C-C, Lin C-J. LIBSVM: A library for support vector machines[DB/OL]. httpP:// www.csie.ntu.edu.tw/∼cjlin/libsvm/.

[13] Michie D, Spiegelhalter D J, Taylor C C. Machine learning[M]. Neural and Statistical Classification, London: Ellis Horwood, 1994.

[14] Hsu C, Lin C. A simple decomposition method for support vector machines[J]. Mach Learn, 2002, 46: 291-314.