文章编号: 1001-0920 (2013) 02-0177-06

领域自适应的最小包含球设计方法

顾 鑫[1,2], 王士同[1], 许 敏[1,3]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 江苏北方湖光光电有限责任公司，
江苏 无锡 214035；3. 无锡职业技术学院 计算机系，江苏 无锡 214000)

摘 要: 支持向量域描述 (SVDD) 算法适用于异常点检测, 但对于不同领域样本组的整体快速识别则力不从心. 为此, 基于 SVDD 算法提出一种基于最小包含球的领域自适应算法 (MEB-DA), 并将其发展为基于中心约束型最小包含球的领域自适应法 (CCMEB-DA), 以满足大样本数据的快速计算. 该算法通过计算各自数据组的包含球球心对不同领域数据进行整体校正和相似度识别, 具有较好的便捷性和自适应性. 将所提出的算法应用于无限保真 (WIFI) 数据的室内定位和人脸识别检测, 均取得了较好的效果, 从而验证了所提出算法的有效性和快速性.

关键词: 中心约束型最小包含球；领域；最小包含球；数据校正

中图分类号: TP391 **文献标志码:** A

Minimum enclosing ball for domain adaptation

GU Xin[1,2], *WANG Shi-tong*[1], *XU Min*[1,3]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. Jiangsu North Huguang Opto-Electronics Co Ltd，Wuxi 214035，China；3. Department of Computer，Wuxi Institute of Technology，Wuxi 214000，China. Correspondent：GU Xin，E-mail：guxinbest@sina.com)

Abstract: Support vector domain description(SVDD) is very suitable for testing a single anomaly point and is inadequate for testing the whole testing dataset. Based on SVDD, the algorithm of minimum enclosing ball for domain adaptation(MEB-DA) is proposed. In order to achieve the rapid calculation for large datasets, an algorithm named center constrained minimum enclosing ball for domain adaptation(CCMEB-DA) is proposed. By calculating the center of each dataset, the dataset is corrected and the similarity of data is identified between different domains, which shows a good adaptability. The proposed method is applied to the fields of wireless fidelity(WIFI) indoor positioning and face detection, and the obtained experimental results show the effectiveness of the proposed algorithm.

Key words: center constrained minimum enclosing ball(CCMEB)；domain；minimum enclosing ball；data correction

0 引 言

传统的知识学习方法一般都假定训练数据和测试数据均来自同样的数据分布. 然而, 在实际情况下, 由于多种原因这种假设并不成立. 训练数据和测试数据往往有着不同的分布, 当分布发生变化时, 传统的机器学习方法必须从头开始, 需要用户重新收集大量的训练数据. 重新收集训练数据和再次训练模型的代价很昂贵, 因此希望能够运用先前任务中所学到的知识来帮助学习新的任务, 以减少对新的训练数据的需求. 为解决这一问题, 人们提出了迁移学习方法[1], 只要不同域之间有一部分相似, 这种学习便能够获得成功. 领域自适应[2-4]可以被看作一种特殊的迁移学习, 其任务是传递和共享不同任务与域之间的知识.

本文提出一种基于最小包含球领域自适应算法 (MEB-DA). 该算法以 SVDD[5]为基础, 首先求出源域的最小包含球球心, 并将不同领域的球心位置、球半径信息作为约束条件, 在此条件下求出目标域的最小包含球球心; 然后与源域进行比较, 以判定其相似度. 该算法能够在不采用数据标签属性的情况下对不同域或近似域进行整体比较, 以判断他们的相似度, 对不同类型的领域数据具有较好的泛化能力. 为了满足大样本数据的快速计算, 将 MEB-DA 发展为基于中心约束型最小包含球的领域自适应算法 (CCMEB-DA), 该算法可应用于大样本数据的比较. 最后通过实验验

收稿日期: 2011-07-17；**修回日期:** 2011-12-16.

基金项目: 国家自然科学基金项目(60903100, 60975027).

作者简介: 顾鑫(1979−), 男, 工程师, 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、数据挖掘等研究.

证了所提出算法的有效性和快速性.

1 最小包含球理论

1.1 传统最小包含球(MEB)理论

MEB 理论的主要思想是找到包含一类数据的超球, 并使球的半径尽量小. 算法表述如下[6]:

$$\max_{R}\ R^2+C'\sum_{i=1}^{N}\xi_i;$$
$$\text{s.t. } \|c-\phi(x_i)\|^2\leqslant R^2+\xi_i^2,\ \xi_i\geqslant 0, i=1,2,\cdots,N. \tag{1}$$

其中: c 为最小包含球的球心, R 为最小包含球的球半径, C' 为惩罚系数, ξ_i 为松弛因子. 对式 (1) 进行 QP 化、核化后的对偶问题为

$$\max_{\boldsymbol{\alpha}}\ \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha};$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ 0\leqslant\boldsymbol{\alpha}\leqslant C'. \tag{2}$$

其中: $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$ 为 Lagrange 乘子, $\boldsymbol{K}=[k(x_i,x_j)]=[\phi(x_i^{\mathrm{T}})\phi(x_j)]$ 为核矩阵, $\mathbf{1}=[1,\cdots,1]^{\mathrm{T}}$ 为 N 维列向量. 由式 (2) 可以计算出半径和球心

$$r=\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}},\ c=\sum_{i=1}^{N}\alpha_i\phi(x_i), \tag{3}$$

且核矩阵 $\boldsymbol{K}$ 对角线恒为一常数 κ, 即

$$k(x_i,x_i)=\kappa. \tag{4}$$

1.2 CCMEB 理论

Tsang 等[6-7]指出, 同时满足式 (2) 和 (4) 的二次规划问题等价于一最小包含球问题, 并在此基础上利用核心集 (Core-set) 技术[6-7]开发了 CVM 算法. CVM 改进了最小包含球算法, 在大样本数据集处理上有着较快的速度. Tsang 进一步探讨了不能同时满足式 (2) 和 (4) 的核方法与 MEB 的关系, 从 MEB 延伸出 CCMEB. CCMEB 给核空间中任意样本点 $\phi(x_i)$ 附加一维新特征 $\delta_i\in R$, 形成新样本 $\begin{bmatrix}\phi(x_i)\\ \delta_i\end{bmatrix}$, 然后寻求新的样本集对应的最小包含球, 但对该最小包含球增加了一个约束条件, 即最小包含球中增加的特征维对应的中心固定在原点, 亦即 CCMEB 的中心为 $\begin{bmatrix}c\\ o\end{bmatrix}$, 这里 c 对应未扩展的核空间中相应的球中心特征向量.

CCMEB 求解可表示为如下的约束优化问题[8]:

$$\max_{\boldsymbol{\alpha}}\ \boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\Delta)-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha};$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ 0\leqslant\boldsymbol{\alpha}\leqslant C'. \tag{5}$$

其中 $\Delta=[\delta_1^2,\cdots,\delta_N^2]$. 由式 (5) 的最优解可得到该最小包含球的半径 r 和球心点 c, 即

$$r=\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\Delta)-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}},$$
$$c=\sum_{i=1}^{N}\alpha_i\phi(x_i). \tag{6}$$

此外, 任意点 $\begin{bmatrix}\phi(x_l)\\ \delta_l\end{bmatrix}$ 和球心点 $\begin{bmatrix}c\\ o\end{bmatrix}$ 的距离可表示为

$$\|c-\phi(x_l)\|^2+\delta_l^2=\sum\alpha_i\alpha_j k(x_i,x_j)-2\sum\alpha_i k(x_i,x_l)+k(x_l,x_l)+\delta_l^2. \tag{7}$$

因为 $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1$, 所以在式 (5) 的目标函数中增加一项 $-\eta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1},\eta\in R$, 不会影响最优解的值, 于是有

$$\max_{\boldsymbol{\alpha}}\ \boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\Delta-\eta\mathbf{1})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha};$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ 0\leqslant\boldsymbol{\alpha}\leqslant C'. \tag{8}$$

式 (8) 与 (5) 等价, 可将不满足式 (4) 的核方法转换成 CVM 快速求解. 通过不断选择样本点, 迭代比较式 (6) 中的 r 与式 (7) 的值, 可求出样本空间的核心集, 继而求出最小包含球球心 c.

2 领域自适应设计方法

2.1 MEB-DA 算法

首先推导该算法. 假设存在两个样本空间 D_1 和 D_2. 其中: D_1 为源域空间含有 N 个样本点 x_j, D_2 为目标域空间含有 N 个样本点 x_i. 需要判断的是, 两个样本域之间是否相似或存在某种关联性. 如果两类样本为近似空间, 则满足两类样本的最小包含球的球心 $(c,\bar{c})$ 与对应的半径 $(R,\bar{R})$ 之间应该无限接近, 可表示如下:

$$\min\ (R-\bar{R})^2+\mu\|\phi(c)-\phi(\bar{c})\|^2,$$
$$\text{s.t. } \|\phi(x_i)-\phi(\bar{c})\|^2\leqslant R^2. \tag{9}$$

其中: $\phi(c)$ 和 $\phi(\bar{c})$ 分别为 D_1 和 D_2 空间球心的高维映射函数; $\phi(x_j)$ 和 $\phi(x_i)$ 分别为 D_1 和 D_2 空间的高维映射函数; μ 为领域依赖系数, 该值越大, 表示目标域对源域的依赖度越高. 式 (9) 看似合理, 但在求解过程中因未知量过多而难以求解. 后来经过证明发现, 两类样本空间相似性与半径 $(R,\bar{R})$ 无关, 只与球心位置 $(c,\bar{c})$ 有关. 具体证明如下:

由 Parzen Windows 理论[9]可知, 利用有限的采样样本可以计算出对应点的概率密度估计. 这里假设 x_i 相对于 D_1 样本空间概率密度估计为 $P_{D_1}(x_i)$, 相对于 D_2 样本空间概率密度估计为 $P_{D_2}(x_i)$. 利用最小累积平方误差, 使得 $P_{D_2}(x_i)$ 最优逼近源域概率密度估计 $P_{D_1}(x_i)$ 可表示如下:

$$\min\int(P_{D_2}(x_i)-P_{D_1}(x_i))^2\mathrm{d}z. \tag{10}$$

由 Parzen Windows 概率公式可知

$$P_{D_1}(x_i)=\phi^{\mathrm{T}}(x_i)\phi(c),\ P_{D_2}(x_i)=\phi^{\mathrm{T}}(x_i)\phi(\bar{c}).$$

代入式 (10), 得

$$\min(\phi(c)-\phi(\bar{c}))^2\int\phi^{\mathrm{T}}(x_i)\phi(x_i)\mathrm{d}x. \tag{11}$$

将式 (11) 核化, $k(x_i,x_i)=\phi^{\mathrm{T}}(x_i)\phi(x_i)$, 采用高斯核 $k(x,y)=\mathrm{e}^{-\|x-y\|^2/\delta^2}$, 则通过观察可以发现

$$\int k(x_i,x_i)\mathrm{d}x=1,$$

所以两样本是否相似与各自球的半径无关而与球心相关. 总结后的公式如下:

$$\begin{aligned}&\min\ R^2+\mu\|\phi(c)-\phi(\bar{c})\|^2;\\&\text{s.t. }\|\phi(x_i)-\phi(\bar{c})\|^2\leqslant R^2.\end{aligned}\tag{12}$$

其对应的 QP 公式为

$$\begin{aligned}&\max_{\alpha_i}\sum_{i=1}^{N}\alpha_i(k(x_i,x_i)-2\phi(x_i)\mu c/(1+\mu))-\\&\qquad\sum_{i=1}^{N}2(k(x_i,x_i)/(1+\mu))\alpha_i^2;\\&\text{s.t. }\sum_{i=1}^{N}\alpha_i=1,\ \alpha_i\geqslant 0.\end{aligned}\tag{13}$$

其中: c 为源域空间的最小包含球球心, 可通过 CVM 算法求出 $c=\sum\limits_{j=1}^{N}\alpha_j\phi(x_j)$. 代入式 (13), 得

$$\begin{aligned}&\max_{\alpha_i}\sum_{i=1}^{N}\alpha_i(k(x_i,x_i)-2k(x_i,x_j)\alpha_j\mu/(1+\mu))-\\&\qquad\sum_{i=1}^{N}2(k(x_i,x_i)/(1+\mu))\alpha_i^2;\\&\text{s.t. }\sum_{i=1}^{N}\alpha_i=1,\ \alpha_i\geqslant 0.\end{aligned}\tag{14}$$

通过计算可求出映射在二维空间的目标域球心

$$\bar{c}=\sum_{i=1}^{N}(\alpha_i x_i+\mu c)/(1+\mu),\ c=\sum_{j=1}^{N}\alpha_j x_j.\tag{15}$$

通过比较 c 与 $\bar{c}$ 并计算二者之间的距离, 可以判断出两类样本点之间的相似性.

2.2　CCMEB-DA 算法

2.2.1　算法描述

在小样本条件下 MEB-DA 具有较好的运算速度, 但对于大样本数据的处理则显得力不从心. 为解决该问题, 本文提出如下 CCMEB-DA 算法.

在式 (14) 的基础上, 令矩阵

$$\begin{aligned}\boldsymbol{K}_1&=[2k(x_i,x_i)/(1+\mu)],\\\boldsymbol{K}_2&=[-2k(x_i,x_i)/(1+\mu)+k(x_i,x_i)-\\&\qquad 2k(x_i,x_j)\alpha_j\mu/(1+\mu)].\end{aligned}$$

参考 CCMEB 算法, 取 $\Delta=\mathrm{diag}(\boldsymbol{K}_2)+\eta\boldsymbol{1}$. 此时, 只要选择足够大的 $\eta\boldsymbol{1}$ 使 $\Delta\geqslant 0$, 就可以将式 (14) 改造成式 (16). 对比式 (8) 可发现, 式 (16) 是一个标准的 CCMEB 问题, 于是, 结合 Core-set 技术便可以得到本文所提出的 CCMEB-DA 算法. 其 QP 公式为

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K}_1)+\Delta-\eta\boldsymbol{1})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{\alpha};\\&\text{s.t. }\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1,\ \boldsymbol{\alpha}\geqslant 0.\end{aligned}\tag{16}$$

由式 (16) 的最优解 $\boldsymbol{\alpha}$ 可求得最小包含球半径和球心分别为

$$\begin{aligned}R_t&=\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K}_1)+\Delta)-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_1\boldsymbol{\alpha}};\\C_t&=\sum_{i=1}^{N}(\alpha_i\phi(x_i)+\mu c)/(1+\mu).\end{aligned}\tag{17}$$

此时, 任意点 $\begin{bmatrix}\phi(x_l)\\ \delta_l\end{bmatrix}$ 与中心 $\begin{bmatrix}C_1\\ 0\end{bmatrix}$ 的距离可表示为

$$\begin{aligned}&\|C_t-\phi(x_l)\|^2+\delta_l^2=\\&\quad k(x_l,x_l)-\sum_{j=1}^{N}\frac{2\mu}{(1+\mu)}\alpha_j k(x_l,x_j)+\\&\quad\sum_{i=1}^{M}\frac{1}{(1+\mu)^2}\alpha_i^2k(x_i,x_i)+\\&\quad\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{2\mu}{(1+\mu)^2}\alpha_i\alpha_j k(x_i,x_j)+\\&\quad\sum_{j=1}^{M}\frac{\mu^2}{(1+\mu)^2}\alpha_j^2k(x_j,x_j)+\delta_l^2.\end{aligned}\tag{18}$$

2.2.2　求解步骤

CCMEB-DA 求解步骤如下.

输入: 源域数据集 D_1, 目标域数据集 D_2, 源域球心 c, 领域依赖系数 μ, Core-set 逼近精度 ε;

输出: 目标域样本空间核心集 Q, 最小包含球球心 $\bar{c}$.

Step 1: 初始化目标域核心集 Q_0,R_0,C_0, 设 t 为迭代计数器, 且初值为 0;

Step 2: 在扩展的特征空间中, 若所有点都被球体 $B(C_t,(1+\varepsilon)R_t)$ 包围, 则转 Step 6;

Step 3: 选取核心集之外的点 x_l, 通过式 (18) 找到与中心点 C_t 最远点 x_l, 将该点加入核心集 $Q_{t+1}=Q_t\bigcup\{x_l\}$;

Step 4: 通过式 (16) 和 (17) 求解新的核心集 Q_{t+1} 下的 C_{t+1} 和 R_{t+1};

Step 5: 令 $t=t+1$, 并返回 Step 2;

Step 6: 终止训练, 将核心集 Q 代入式 (14), 继而可求出 c 和 $\bar{c}$, 通过 c 与 $\bar{c}$ 之间的距离比较便可以判断出两类样本点之间的相似性.

2.3　CCMEB-DA 算法复杂度分析

标准 SVDD 算法的时间复杂度为 $O(m^2)$, 空间复杂度为 $O(m^2)$, 其中 m 为样本的规模. CCMEB 利用基于 MEB 的核心集快速求解, 时间复杂度与样本规模 m 呈线性关系, 空间复杂度不依赖于样本规模, 大大提高了运算速度. 第 2.2.2 节 Step 3 求最远点时, 每

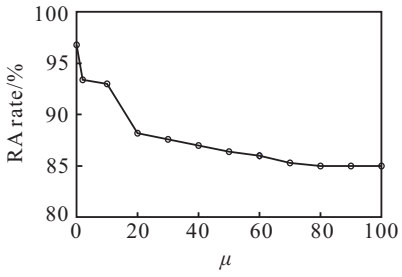
进行一次迭代运算的时间复杂度为$O(|Q_t|^2+|D|\cdot|Q_t|)$, 当样本D规模较大时会非常耗时. 本文参考文献[10]提出的加速方法, 在样本D中随机抽取一子集D_1作为替代寻找距离中心最远的点. 文献[13]已经证明, 当子集大小为59时, 最远点在D_1中的概率为95%, 时间复杂度则下降为$O(|Q_t|^2+|D_1|\cdot|Q_t|)$, 其中$|D_1|$为59. 在Step 4中, QP求解的数据为核心集, 其规模远远小于样本总体规模, 时间复杂度也远远小于求解所有样本的时间复杂度. CCMEB-DA算法将CCMEB理论作用于大数据集快速运算, 因此算法复杂度可参考CVM, 其时间复杂度上界为$O(m/\varepsilon^2+1/\varepsilon^4)$, 与样本规模$m$呈线性关系; 其空间复杂度上界为$O(1/\varepsilon^2)$, 可以使用存储核心集代替所有样本而不依赖于样本规模.

2.4 领域依赖系数参数选择

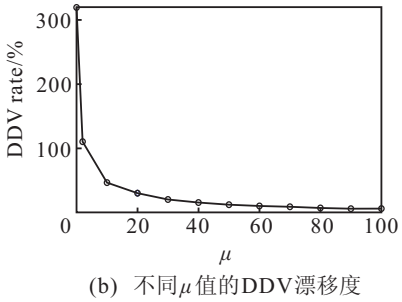
式(9)中定义了领域依赖系数μ. 该系数越大, 源域球心与目标域球心越接近, 领域校正能力越强, 但求得的目标域最小包含球识别精度越低(球外点越多). 为能够尽可能多地将测试点用最小包含球包围, 同时又尽可能地靠近源域球心, 需对μ进行选择. 首先定义如下两个概念:

1) 最小包含球识别精度, 简称为RA. RA=(SPI/TSP)×100%, SPI为最小包含球内部样本点个数, TSP为总样本点个数. RA的值越大, 算法识别率越高.

2) 领域漂移度, 简称为DDV. DDV=(Distance/R)×100%, R为目标域最小包含球半径, Distance为目标域与源域球心之间的距离. DDV的值越小, 说明领域相似度越高. 通常将DS设定为一阈值, 以判别领域相似性, 如DDV $\leqslant$ 50%表示领域相似, DDV $>$ 50%表示领域无关.



(a) 不同μ值的RA精度



(b) 不同μ值的DDV漂移度

图1 领域依赖系数选择

在对μ进行取值时, 应尽可能地提高RA, 同时应尽量减小DDV. 只有这样才能实现最大限度地利用已有数据学习, 同时又不丢失现有数据, 从而达到领域自适应性. 选择不同的μ值, 在相似的领域数据集中通过反复实验比较得到的结果如图1所示.

由图1可以发现, 当μ取值大于10时, RA显著下降; 当μ小于10时, DDV明显增强算法的校正功能减弱. 显然, 理想的情况应该是RA大于90%, DDV小于50%. 基于这样的标准, μ的参数应设定为10.

3 实验结果及分析

本实验分为3部分, 第1部分为人工数据, 主要从抗扰动、大样本测试两方面验证算法的合理性、自适应性; 第2部分为WIFI真实数据集, 从数据校正和定位预测角度验证算法的领域自适应性, 并与其他算法进行对比分析; 第3部分为从图像处理的角度以人脸数据为例进行算法验证. 实验环境为Intel Core 2 GHz CPU, 1 G RAM Windows XP, Matlab 7.5.0.

3.1 人工生成数据测试

3.1.1 算法验证

选择11组二维向量数据, 每一组数据均含5 000个样本点, 11组数据均为指定条件下的随机正态分布. 其中train, test 1和test 2为同样的数据分布, 其余则不同. 为满足测试要求, 人为对test 1和test 2加以扰动, 最后生成的所有数据集分布如图2所示. 图2中: train为训练数据, test 1和test 2为经过扰动后的与训练集相同的正态分布随机数据, 其余test 3～test 10部分则为不同的正态分布随机数据.

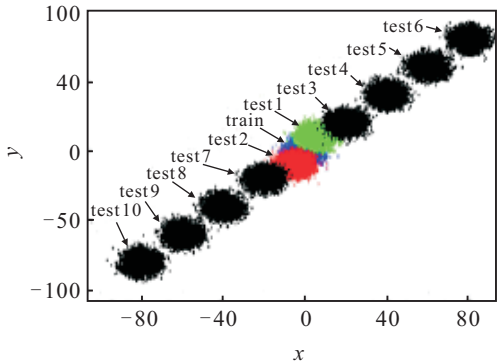


图2 人工数据集分布图

分别采用CCMEB算法和CCMEB-DA算法计算出的结果如表1所示. 通过表1的对比观察可以发现: 当DDV阈值设为50%时, CCMEB-DA算法相对于CCMEB算法具有较好的领域自适应性, 能够校正test 1和test 2中的扰动, 判断出train, test 1, test 2之间的相似性; 对于数据分布明显不同的test 3～test 10也能够通过球心之间的距离反应出不同领域之间的差异大小.

表 1　不同算法实验结果与对比

dataset	CCMEB		CCMEB-DA	
	球间距	DDV/%	球间距	DDV/%
train,train	0	0	0.713 5	17.83
train,test 1	9.080 5	227.01	2.100 4	49.51
train,test 2	9.028 8	225.72	1.963 7	49.09
train,test 3	25.218 8	630.47	3.015 1	75.37
train,test 4	51.136 6	1 278.41	5.197 0	129.92
train,test 5	77.254 9	1 931.37	7.732 2	193.30
train,test 6	102.982 0	2 574.55	10.237 2	255.93
train,test 7	25.346 2	633.65	2.976 3	74.40
train,test 8	51.664 0	1 291.60	5.195 6	129.89
train,test 9	77.432 7	1 935.81	7.793 5	194.83
train,test 10	103.112 8	2 577.82	10.258 4	256.46

3.1.2　大样本数据验证

为了验证 CCMEB-DA 算法在大样本下的运算速度, 将人工生成样本大小从 1 000 递增到 50 000, 其训练时间与 MEB-DA 算法的对比结果如表 2 所示.

表 2　大样本运算时间对比

样本规模	MEB-DA 训练时间/s	CCMEB-DA 训练时间/s
1 000	2.2	6.33
2 000	4.80	8.38
3 000	9.51	10.53
4 000	18.27	12.50
10 000	–	20.53
20 000	–	33.18
30 000	–	46.21
40 000	–	65.76
50 000	–	81.42

由表 2 可见: MEB-DA 在样本数据超过 4 000 之后, 系统发生了内存溢出错误, 使得算法无法继续运行; 样本容量较小时, CCMEB-DA 由于存在核心集的迭代外扩过程, 运行速度尚不及 MEB-DA, 但随着样本容量的增大, 其快速的优势便能得以体现, 而且会越来越明显. 图 3 表明了 CCMEB-DA 算法的时间复杂度与样本规模呈线性关系.

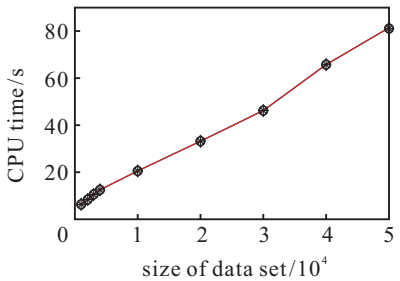


图 3　样本规模与运算时间

3.2　WIFI 定位实验

3.2.1　算法验证

通过 WIFI 数据获取位置信息 (标签) 是一种较为常用的定位方式[11-12], 但是标签信息不一定存在于每一组 WIFI 数据中, 此时需采用对应算法分析不同数据组中的信息, 对位置信息进行校正或预测. 本文采用的 WIFI 数据来自 http://www.cse.ust.hk/~qyang/ICDMDMC07/. 实验选择 20 组 WIFI 数据并将大部分位置标签信息隐去, 以判断算法的领域自适应性. 首先根据每组 WIFI 数据中的标签信息 (从后台得知) 模拟绘制 20 组 WIFI 定位分布示意图, 并将此图作为衡量算法正确性的依据, 其具体分布如图 4 所示; 然后将第 1 组 WIFI 数据作为源域数据, 其余组作为目标域数据, 通过 CCMEB-DA 算法计算出目标域各自最小包含球的球心, 绘制成图 5. 比较图 4 和图 5 可以发现, 20 组 WIFI 数据的相对位置分布基本一致, 这表明了 CCMEB-DA 算法通过迁移学习尽量多地利用原有信息实现了领域自适应, 该算法具有很好的位置校正功能. 同时, 通过比较可知, 真实定位点之间的距离趋势与各球心之间的距离趋势相一致.

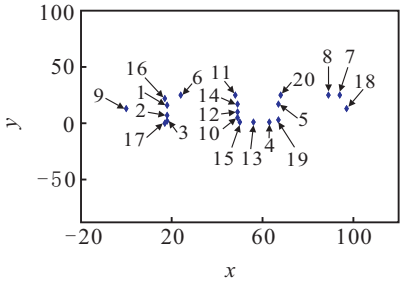


图 4　WIFI 定位点坐标示意图

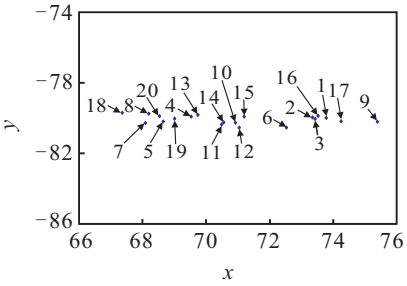


图 5　WIFI 最小包含球球心坐标示意图

3.2.2　算法对比分析

在 WIFI 定位实验中, CCMEB-DA 在源域标签信息已知的情况下, 通过比较源域与目标域最小包含球球心坐标, 能在数据标签信息缺失的情况下对目标域位置进行预测. 本文对 MMDE[13], TCA[14]和 CCMEB-DA 这 3 种领域自适应算法进行比较. 首先定义平均误差距离

$$\mathrm{AED}=\frac{\sum_{(x_i,y_i)\in D_2}|f(x_i)-y_i|}{N}.$$

其中: x_i 为每一组 WIFI 数据, $f(x_i)$ 为预测位置的算法函数, y_i 为每一组 WIFI 数据的真实定位位置, N 为测试样本组数目, 选择 20 组 WIFI 数据. 在标签信息缺失的情况下重复 10 次计算 AED 后取其平均值, 具体结果如图 6 所示.

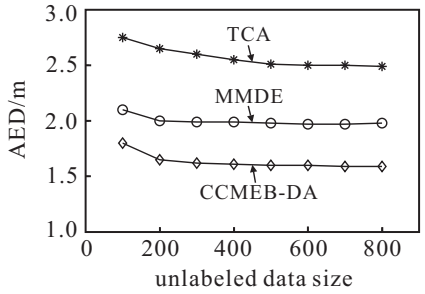


图 6 WIFI 不同算法下的 AED 对比

观察图 6 发现, CCMEB-DA 算法的 AED 值相比其他算法更小, 更加接近真实的 WIFI 定位, 表明该算法具有较好的位置校正和预测功能.

3.3 人脸识别检验[15]

本文采用 ORL 人脸库, 下载地址: http://download.csdn.net/source/1583590. 这里选择 8 幅人脸图像进行比较, 详见图 7.

本文将面部表情的差异看作一种数据扰动, 不同人脸的差异看作一种分类. 采用 CCMEB-DA 算法能够消除相似领域数据之间的扰动, 同时能够显著区分

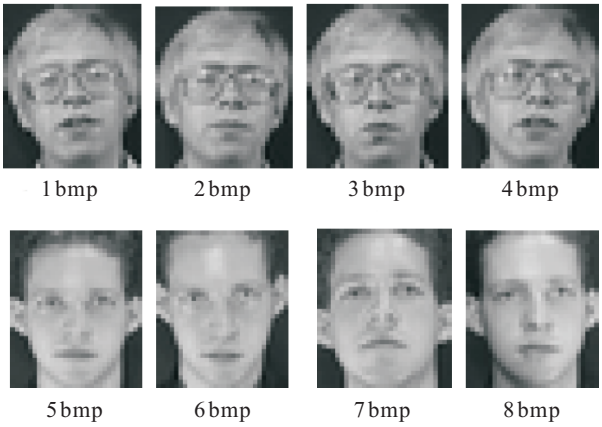


图 7 ORL 数据集

满足不同分布的领域数据, 做到领域自适应. 通过算法计算出图像球心之间的距离参数如表 3 所示.

比较距离值可以判断图像之间的相似性. 由表 3 可以看出, 同一人的人脸数据球心间距都较小 (一般小于 10), 不同人的人脸数据球心间距都较大 (一般都大于 10). 因此, 通过 CCMEB-DA 算法可以做到对人脸的有效识别.

表 3 不同人脸图像的球间距离

源域	目标域							
	1 bmp	2 bmp	3 bmp	4 bmp	5 bmp	6 bmp	7 bmp	8 bmp
1 bmp	9.53e-14	4.57	3.44	4.72	11.10	13.79	12.86	14.87
2 bmp	4.570 5	9.74e-14	2.24	1.26	10.92	12.42	13.20	16.52
3 bmp	3.44	2.24	9.55e-14	2.51	10.83	12.70	12.94	15.68
4 bmp	4.72	1.26	2.51	9.71e-14	11.59	13.11	13.94	16.93
5 bmp	11.10	10.92	10.83	11.59	9.23e-14	5.16	5.11	7.74
6 bmp	13.79	12.42	12.70	13.11	5.16	9.00e-14	6.38	8.70
7 bmp	12.86	13.20	12.94	13.94	4.70	6.38	8.87e-14	8.32
8 bmp	14.87	16.52	15.68	16.93	7.74	8.70	8.32	0.10e-14

4 结 论

本文将 MEB 和 CCMEB 理论应用于领域自适应的研究, 提出了 MEB-DA 和 CCMEB-DA 算法. 在求解目标域球心位置时应尽可能多地利用源域数据, 通过比较不同域球心位置判断领域相似度, 完成数据的修正和校正. 实验结果验证了所提出算法的有效性和快速性, 但如何将其应用于数据分类和数据回归尚有待进一步研究.

参考文献(References)

[1] Dai W, Yang Q, Xue G, et al. Boostimg for transerfer learning[C]. Proc of the 24th Int Conf on Machine Learning. New York: ACM Press, 2007: 20-24.

[2] Hal Daum'e I I I, Daniel Mareu. Domain adaptation for statistical classifiers[J]. J of Artificial Intelligence Research, 2006, 26(4): 101-126.

[3] John Blitzer, Ryan McDonald, Fernando Percira. Domain adaptation with structural correspondence learning[C]. Proc of the 2006 Conf on Empirical Methods in Natural Language Processing. Sydney, 2006: 120-128.

[4] Sandeepkumar Satpal, Sunita Sarawagi. Domain adaptation of conditional probability models via feature subsetting[C]. Proc of PKDD. Heidelberg: Springer-Verlag Press, 2007: 224-235.

[5] Tax D M J, Duin R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11): 1191-1199.

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

[7] 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.

[8] Fu-lai Chung, Zhaohong Deng, Shitong Wang. From minimum enclosing ball to fast fuzzy inference system training on large datasets[J]. IEEE Trans on Fuzzy Systems, 2009, 17(1): 173-184.

(下转第187页)