基于最小包含球的领域迁移学习新方法

顾 鑫 王士同
(江南大学数字媒体学院 无锡 214122)

摘 要 传统机器学习方法认为不同的学习任务彼此无关,但事实上不同的学习任务常常相互关联。迁移学习试图利用任务之间的联系以及过去的学习经验加速对于新任务的学习。将最小包含球(Minimum Enclosing Ball,MEB)算法与 Parzen Windows 概率估计公式相结合,提出了一种新的迁移学习算法 MEBTL((Minimum Enclosing Ball Transfer Learning)。该算法同时结合 CVM(Core Vector Machines)理论提出了 CCMEBTL(Center Constrained Minimum Enclosing Ball Transfer Learning)算法,其可以在不同领域之间完成大样本的迁移学习。作为验证,将其应用在 WIFI 数据的室内定位、人脸识别检测上,并取得了较好的效果。

关键词 中心约束型最小包含球,数据校正,迁移学习,领域自适应
中图法分类号 TP391 **文献标识码** A

Novel Domain Transfer Learning Approach Using Minimum Enclosing Ball

GU Xin WANG Shi-tong
(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract Traditional machine learning methods assume that different learning tasks have nothing with each other, but in fact there are some links between them. Transfer learning attempts to use these links and even past learning experiences between different tasks to accelerate the learning for new tasks. This paper integrated the MEB (Minimum enclosing ball algorithm together with Parzen windows probability estimation to develop a new transfer learning method named MEBTL (Minimum enclosing ball Transfer learning). We also used CVM (Core Vector Machines) theory to develop its fast version of the proposed algorithm CCMEBTL for large domain adaptation. The experimental results about "WIFI indoor positioning" and "face detection" indicate the effectiveness of the proposed algorithm.

Keywords CCMEB, Data correction, Transfer learning, Domain adaptation

1 引言

传统的知识学习一般假定训练数据和测试数据来自同样的数据分布。但在实际情况下,由于多种原因,这种假设并不成立,即训练数据和测试数据往往有不同的分布,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是昂贵的,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。由此,迁移学习方法(Transfer Learning)[1]被提出来解决这一问题,只要不同域之间有一部分相似,这种学习便可获得成功。而领域自适应可以看作是一种特殊的迁移学习,其任务是传递和共享不同任务和域之间的知识。

目前,迁移学习存在多种形式。文献[2]将其总结为3种形式:①归纳式迁移学习,其目标领域标签数据较少;②直推式迁移学习,原始领域中存在大量标签数据,而目标领域中没有标签数据;③无人监督式迁移学习,原始和目标领域中都无标签数据。本文算法有两种应用背景:一种是直推式迁移学习,原始领域与目标领域的数据分布不同,原始领域与目标领域中大量的标签数据可用,而目标领域中没有给出标记数据;还有一种情况是原始领域与目标领域之间存在干扰或扰动,而领域自适应算法能够有效对数据进行校正。

在直推式迁移学习中,大部分是基于特征表示的迁移学习方法,都是在无监督学习框架下的。Blitzer 等人[3]提出一种结构化相似学习(SCL)方法,该方法利用目标领域中未标记数据来提取一些特征,从而降低领域之间的差异。虽然 SCL 算法能降低不同领域之间的差异,但选择枢纽特征是很困难的,并且具有领域依赖性。

在 NLP 领域中的迁移学习有时也被称为领域适应。在这个领域,Daumé III[4]为 NLP 问题提出了一种简单的内核映射函数,它同时映射原始和目标领域中的数据到一个高维的特征空间中,标准判别算法被用来训练分类器。然而,构建内核映射函数是领域知识驱动的,推广内核映射到其他领域或应用程序中是不容易的。

在文献[5]中,一个基于联合集群的算法用来跨不同的领域传输标记信息。Xing 等人在文献[6]中提出了一种新的叫做"Bridged refinement"的算法来纠正由未注意变换的分类器所预测的目标分布上的标记,将训练和测试数据的混合分布

到稿日期:2012-09-19 返修日期:2013-01-14 本文受国家自然科学基金项目(60903100,60975027)资助。
顾 鑫(1979—),男,博士生,工程师,主要研究方向为模式识别、人工智能等,E-mail:guxinbest@sina.com;王士同(1964—),男,教授,博士生导师,主要研究方向为模式识别、人工智能、数据挖掘、模糊系统等。

作为一个桥梁来更好地从训练数据到测试数据进行迁移。

本文主要致力于解决直推式迁移学习问题，原始领域与目标领域的数据分布不同，原始领域中大量的标记数据可用，而目标领域中没有给出标记数据，其目的是为目标领域训练并找出与原始领域之间的相似度。为了完成这一目标，提出了基于 MEB[7] 的迁移学习(Minimum enclosing ball Transfer learning，简称 MEBTL)方法，其核心思想是分别计算出原始领域、目标领域相对于最小包含球球心的概率估计值，然后通过它们之间的比值判断出原始领域与目标领域数据分布的差异度。为了在不同领域之间完成大样本的迁移学习，我们参考 CCMEB[8] 理论提出了 CCMEBTL(Center Constrained Minimum Enclosing Ball Transfer Learning)算法。试验结果表明该算法具有较高的效率和准确性。

2 最小包含球理论

2.1 传统最小包含球(MEB)

其主要思想是找到包含一类数据的超球，并且使球的半径尽量小。其算法表述如下[8]：

$$\min R^2$$

$$\text{s.t. } \| c-\phi(x_i) \|^2 \leqslant R^2, i=1,\cdots,N \tag{1}$$

式中，c 为最小包含球的球心，R 为最小包含球的球半径。我们同时可以将式(1)进行 QP 化处理、核化处理。其对偶问题如下，其中 $\partial=[\partial_1,\cdots,\partial_N]^T$ 为拉格朗日系数，$k(x_i,x_j)$、$k(x_i,x_i)$ 是核矩阵。

$$\arg\max -\sum_{i,j=1}^{N}\partial_i k(x_i,x_j)\partial_j+\sum_{i=1}^{N}\partial_i diag(k(x_i,x_i))$$

$$\text{s.t. } \sum_{i=1}^{N}\partial_i=1, 0\leqslant\partial_i\leqslant c \tag{2}$$

通过式(2)计算可求出半径和球心：

$$R=\sqrt{\partial^T diag(K)-\partial^T K\partial}, c=\sum_{i=1}^{N}\partial_i\phi(x_i) \tag{3}$$

且核矩阵对角线恒为一常数 κ，即

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

2.2 CCMEB 理论

在文献[9]中 I Tsang 指出同时满足式(2)和式(4)的二次规划问题等价于一最小包含球的问题。在此基础上提出了利用最小包含球理论的核心集(Core-set)技术[5,6]，开发了核心向量机(CoreVector Machines，CVM)算法[9]。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。

2008 年，I Tsang，JWORK 和 JZURADA 进一步在文献[8]中探讨了更多不能同时满足式(2)和式(4)的核方法与 MEB 问题之间的关系。他们从 MEB 延伸出了中心约束型 MEB(Center Constrained-Minimum Enclosing Ball，CCMEB)。CCMEB 中，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，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 求解可表示为如下约束优化问题：

$$\arg\max -\sum_{i,j=1}^{N}\partial_i k(x_i,x_j)\partial_j+\sum_{i=1}^{N}\partial_i diag(k(x_i,x_i)+\Delta)$$

$$\text{s.t. } \sum_{i=1}^{N}\partial_i=1, 0\leqslant\partial_i\leqslant c \tag{5}$$

式中，

$$\Delta=[\delta_1^2,\cdots,\delta_N^2] \tag{6}$$

由式(5)的最优解可得该最小包含球的中心点 c 和半径 r：

$$r=\sqrt{\partial^T(diag(k)+\Delta)-\partial^T K\partial}, c=\sum_{i=1}^{N}\partial_i\phi(x_i) \tag{7}$$

由于 $\partial^T 1=1$，因此在式(5)的目标函数中增一项：$\eta\partial^T 1$，$\eta\in R$ 将不会影响最优解的值，于是得式(8)：

$$\arg\max -\sum_{i,j=1}^{N}\partial_i k(x_i,x_j)\partial_j+\sum_{i=1}^{N}\partial_i diag(k(x_i,x_i)+\Delta-\eta 1) \tag{8}$$

$$\text{s.t. } \sum_{i=1}^{N}\partial_i=1, 0\leqslant\partial_i$$

此外，任意点 $\begin{bmatrix}\phi(x_l)\\ \delta_l\end{bmatrix}$ 和中心点 $\begin{bmatrix}c\\ o\end{bmatrix}$ 的距离可表示成：

$$\| c-\phi(x_l) \|^2+\delta_l^2=\sum\partial_i\partial_j k(x_i,x_j)-2\sum\partial_i k(x_i,x_l)+k(x_l,x_l)+\delta_l^2 \tag{9}$$

通过不断选择样本点，迭代比较式(7)与式(9)的值，我们先求出样本空间的核心点(Core-set)，继而可求出最小包含球球心 c，具体过程参见文献[8]。

3 MEBTL 设计方法

3.1 MEBTL 算法

我们先推导该算法：

假设存在两个样本空间 $D1$、$D2$，其中 $D1$ 为训练样本空间(train)，含有 N 个样本点 x_i；$D2$ 为测试样本空间(test)，含有 N 个样本点 x_j。我们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。根据 Parzen Windows 理论可知，利用有限的采样样本可以计算出对应点的概率估计。这里我们设 x_j 相对于 $D1$ 样本空间的概率估计为 $P_{D1}(x_j)$，设 x_j 相对于 $D2$ 样本空间的概率估计为 $P_{D2}(x_j)$。我们认为如果两类样本为近似空间，则满足两类样本的概率估计比应尽量靠近 1，即如下所示。

$D1$ 样本空间相对其最小包含球的球心 c 的概率估计为：

$$\frac{P_{D2}(x_j)}{P_{D1}(x_j)}=r(x_j) \tag{10}$$

如果 $r(x_j)\to 1$，test 与 train 同类；

如果 $r(x_j)\to 0$，test 与 train 非同类。

根据最小包含球公式可以求出 $D1$ 空间的最小包含球球心 c 和对应半径 R，具体公式表述如下：

$$\min R^2$$

$$\text{s.t. } \| \phi(c)-\phi(x_i) \|^2 \leqslant R^2, i=1,\cdots,N \tag{11}$$

$$\phi(c)=\sum_{i=1}^{N}B_i\phi(x_i), R=\sqrt{B^T diag(K)-B^T KB} \tag{12}$$

同时根据 Parzen Windows 概率估计公式可知：

$$P_{D1}(x_j)=\phi^T(x_j)\phi(c)=\sum_{i,j=1}^{N}B_i k(x_i,x_j) \tag{13}$$

同理可知：

$$\phi(\bar{c})=\sum_{j=1}^{N}A_j\phi(x_j) \tag{14}$$

$$P_{D2}(x_j)=\phi^T(x_j)\phi(\bar{c})=\sum_{j=1}^{N}A_j k(x_j,x_j) \tag{15}$$

令 $\phi(x_i)$ 为 train 空间样本；

令 $\phi(x_j)$ 为 test 空间样本；

令 $\phi(x_k)$ 为 test 加上 train 的总空间样本；

令 $\phi(c)$为 train 空间样本的最小包含球球心；

令 $\phi(\bar{c})$为 test 空间样本的最小包含球球心；

令 B_i 为 train 空间样本的最小包含球的拉格朗日系数；

令 A_i 为 test 空间样本的最小包含球的拉格朗日系数。

将式(13)、式(15)代入式(10)中得：

$$P_{D2}(x_j)=\phi^{\mathrm{T}}(x_j)\phi(\bar{c})=P_{D1}(x_j)r(x_j)$$
$$=\phi^{\mathrm{T}}(x_j)\phi(c)r(x_j) \tag{16}$$

将 $D2$ 样本空间的最小包含球的求解公式展开如下：

$$(\phi(x_j)-\phi(\bar{c}))^2=\phi^{\mathrm{T}}(x_j)\phi(x_j)-2\phi(x_j)\phi(\bar{c})+$$
$$\phi^{\mathrm{T}}(\bar{c})\phi(\bar{c}) \tag{17}$$

取高斯核后得到下式：

$$(\phi(x_j)-\phi(\bar{c}))^2=2-2\phi^{\mathrm{T}}(x_j)\phi(\bar{c}) \tag{18}$$

将式(16)代入式(18)后得：

$$(\phi(x_j)-\phi(\bar{c}))^2=2-2\phi^{\mathrm{T}}(x_j)\phi(c)r(x_j)$$

$$(\phi(x_j)-\phi(\bar{c}))^2=2-2r(x_j)\sum_{i=1}^{N}B_ik(x_i,x_j) \tag{19}$$

同时 $r(x)$可用奇函数的线性表示如下：

$$r(x_j)=\sum_{k=1}^{N}\partial_k k(x_j,x_k) \tag{20}$$

将式(20)代入式(19)中得以下公式：

$$(\phi(x_i)-\phi(c))^2=2-2\sum_{k=1}^{N}\partial_k k(x_k,x_j)\sum_{i=1}^{N}B_ik(x_i,x_j) \tag{21}$$

我们可通过构造下列优化模型求解：

$$\min(\partial_k^{\mathrm{T}}\partial_k)/2$$

$$\text{s. t. } 2-2\sum_{k=1}^{N}\partial_k k(x_k,x_j)\sum_{i=1}^{N}B_ik(x_i,x_j)\leqslant R^2 \tag{22}$$

同时将其 QP 化：

$$\max\sum_{i=1}^{N}a_i diag(2-R^2)-2\sum_{i=1}^{N}(a_i)^2(k(x_k,x_j)\sum_{i=1}^{N}B_ik(x_i,x_j))^2$$

$$\text{s. t }\sum_{i=1}^{N}a_i=1,0\leqslant a_i \tag{23}$$

a_i 为拉格朗日系数，R 为 $D1$ 样本空间的最小包含球半径。求解可参考式(12)，求解后得：

$$\partial_k=2\sum_{i=1}^{n}a_i\sum_{i=1}^{N}B_ik(x_i,x_j)(k(x_k,x_j)) \tag{24}$$

将式(24)代入式(20)得：

$$r(x_j)=2\sum_{i=1}^{n}a_i\sum_{i=1}^{N}B_ik(x_i,x_j)(k(x_k,x_j)k(x_k,x_j)^{\mathrm{T}}) \tag{25}$$

最后通过 $r(x_j)$的值比较不同域之间的相关性，完成不同域之间的迁移学习。

3.2 CCMEBTL

MEBTL 在小样本条件下有着较好的运算速度，但对大样本数据的处理就显得力不从心。在此我们提出了 CCMEBTL 算法，实验发现其有着较好的运行效率。首先可以通过 CVM 算法求出训练集 $D1$ 样本空间的核心点(Core-set-1)，具体算法见文献[9]。通过训练样本的 Core-set-1 可以快速求出 $D1$ 的球心 c 和半径 R。再将 c、R 代入到式(23)中，利用 CCMEB 快速求出训练集的核心点(Core-set-2)，具体算法如下。

在 MEBTL 算法结论的式(23)中，令 $2-R^2$ 为矩阵 $\boldsymbol{D}$，令 $2\sum_{i=1}^{N}B_ik(x_i,x_j)k(x_k,x_j)$为矩阵 $\boldsymbol{L}$。则式(23)可简化为下式：

$$\max a^{\mathrm{T}}diag(\boldsymbol{D})-a^{\mathrm{T}}\boldsymbol{L}a$$

$$\text{s. t }\sum_{i=1}^{N}a_i=1,0\leqslant a_i \tag{26}$$

在式(26)的基础上参考 2.2 节的 CCMEB 算法，取

$$\Delta=-diag(\boldsymbol{L})+diag(\boldsymbol{D})+\eta 1 \tag{27}$$

其中 $1=[1,\cdots,1]^{\mathrm{T}}$。

此时只要选择足够大的 η，就能使 $\Delta\geqslant 0$。式(28)即是一个标准的 CCMEB 问题，于是结合 Core-set 技术就得到了本文的 CCMEBTL 算法。其 QP 公式如下：

$$\max a^{\mathrm{T}}(diag(\boldsymbol{L})+\Delta-\eta 1)-a^{\mathrm{T}}\boldsymbol{L}a$$

$$\text{s. t. }\sum_{i=1}^{N}a_i=1,0\leqslant a_i \tag{28}$$

该式为一最小包含球问题，其解如下：

$$R1=\sqrt{a^{\mathrm{T}}(diag(\boldsymbol{L})+\Delta-a^{\mathrm{T}}\boldsymbol{L}a} \tag{29}$$

$$C1=\sum_{i=1}^{m}a_i\phi(x_i) \tag{30}$$

任意点$\begin{bmatrix}\phi(x_D)\\ \delta_l\end{bmatrix}$和中心点$\begin{bmatrix}C1\\ o\end{bmatrix}$的距离可表示成：

$$\|C1-\phi(x_l)\|^2+\delta_l^2=a^{\mathrm{T}}K(x_i,x_i)a-2aK(x_i,x_l)+$$
$$K(x_l,x_l) \tag{31}$$

注：x_i 为原有集合中的样本点，x_l 为新添比较的样本点，K 取高斯核。

其求解步骤为：

1. 在现有样本点 x_i 中求出中心点 $C1$ 和半径 $R1$。

2. 判断有无点在半径 $R1$ 外，没有则停止，有则继续。

3. 将距离 $C1$ 点最远的点 x_l 添加到原有样本点中构成(Core-set-2)。

4. 在新样本点中求出 $C1$，$R1$。

5. 循环至第 2 步。

最后将核心点(Core-set-2)代入式(23)中，求出 $r(x_j)$，通过比较可以判断出两类样本点之间的相似度。

3.3 MMEBTL 解题步骤

综合 3.2 节内容，将 MMEBTL 算法解题步骤归纳为表 1。

表 1

MEBTL 解题步骤
输入：D1、D2，其中 D1 为源样本空间，含有 N 个样本点 x_i，D2 为目标样本空间，含有 N 个样本点 x_j
输出：两类样本空间相对于最小包含球球心 c 的概率估计比 r(x)
步骤 1　利用 CVM 算法求出 D1 空间核心集的 CORE-SET-1
步骤 2　利用 CCMEBTL 算法求出 D2 空间核心集的 CORE-SET-2
步骤 3　利用核心集 CORE-SET-1 求出 D1 空间的最小包含球的球心 c 和半径 R、拉格朗日系数 B_i
$\phi(c)=\sum_{i=1}^{N}B_i\phi(x_i)$　$R=\sqrt{\partial^{\mathrm{T}}\mathrm{diag}(k)-\partial^{\mathrm{T}}K\partial}$
步骤 4　利用核心集 CORE-SET-2 求出公式 D2 空间的最小包含球的球心 C1 和半径 R1，求出式(21)对应的拉格朗日系数∂_k 和概率比 $r(x_j)$
$C1=\sum_{i=1}^{m}a_i\phi(x_i)$　$R1=\sqrt{a^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{L})+\Delta-a^{\mathrm{T}}\boldsymbol{L}a}$
$\partial_k=2\sum_{i=1}^{N}a_i\sum_{j=1}^{N}B_ik(x_i,x_j)(k(x_k,x_j))$　$r(x_j)=\sum_{k=1}^{N}\partial_k k(x_j,x_k)$
(2)测试
如果 $r(x_j)\to 1$，test 与 train 同类
如果 $r(x_j)\to 0$，test 与 train 非同类

4 实验结果及其分析

4.1 自定义数据测试

此处我们选择 11 组二维向量数据，每一组数据均含 5000 个样本点，11 组数据均为指定条件下的随机正态分布。其分布情况如表 2 所列。

表 2　人工数据集设定

数据集	正态分布均值	正态分布标准差
Train	0	4
Test1	0	4
Test2	0	4
Test3	20	4
Test4	40	4
Test5	60	4
Test6	80	4
Test7	−20	4
Test8	−40	4
Test9	−60	4
Test10	−80	4

表中的 Train、Test1、Test2 为同样的数据分布，其余则不同。为了满足测试要求，我们人为地对 Test1、Test2 加以扰动，最后生成的所有数据集分布如图 1 所示。

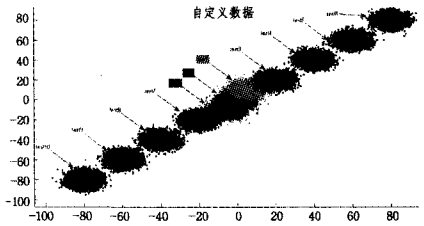

图 1　人工数据集分布图

其中 Train 为训练数据，Test1、Test2 为经过扰动后的与训练集相同的正态分布随机数据，其余 Test3－Test10 部分则为不同的正态分布随机数据。我们首先算出训练球的球心坐标，然后采用 CCMEBTL 算法计算出不同测试集相对于训练集的概率估计比，我们将数据生成为表 3。

表 3　不同数据设定下的球心坐标和概率比(人工数据)

训练集/测试集	测试集球心坐标	训练集与测试集之间概率比值
Train	(0.0440,−0.0165)	0.9977
Test1	(5.0440,9.9835)	0.0693
Test2	(−4.7570,−9.8795)	0.5656
Test3	(19.9918,19.9641)	7.0244 e−4
Test4	(39.9601,39.7514)	3.2992 e−35
Test5	(60.0302,60.1870)	6.2293e−92
Test6	(80.0658,80.0943)	3.0612e−219
Test7	(−19.9714,−20.0552)	3.6565 e−4
Test8	(−40.2375,−40.1465)	9.4482 e−40
Test9	(−60.2330,−60.2676)	1.0259e−108
Test10	(−80.3364,−79.9662)	2.0899 e−228

通过我们的算法可以计算出不同测试集相对于训练集的概率估计比值。通过表 3 比较我们可以发现其概率估计比值与数据集生成函数相一致，较好地反应了不同向量集之间的相似性。

4.2　WIFI 定位算法验证

1)算法验证

通过 WIFI 数据获取位置信息(标签)是一种常用的定位方式[10,11]，但标签信息不一定存在于每一组 WIFI 数据当中，此时我们需要采取对应算法分析不同数据组中的信息，对位置信息进行校正。本文采用的 WIFI 数据来自 http://www.cse.ust.hk/～qyang/ICDMDMC07/。该数据组收集的是 145.5m 长、37.5m 宽的建筑内的 WIFI 信号量，建筑内部位置被人为定义为 247 个 1.5m×1.5m 的方格，也就是 240 个定位点。具体如图 2 所示。

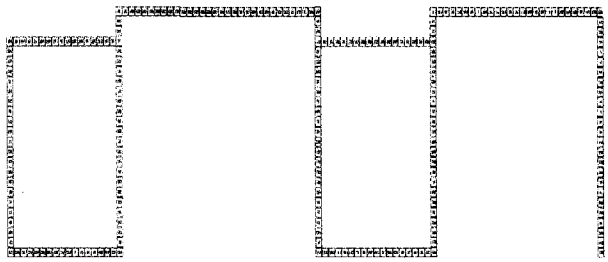

图 2　WIFI 定位示意图

本文选择了 20 组 WIFI 定位数据，每组 WIFI 数据格式如下：

Trace_1

〈Record Time〉〈\t〉〈Location Label〉〈\t〉〈AP's Index_1〉〈:〉〈RSS value〉〈\t〉〈AP's Index_2〉〈:〉〈RSS value〉〈\t〉…〈AP's Index_n〉〈:〉〈RSS value〉〈\n〉

118629xxx63.484　163　61:−82　67:−79　71:−76　73:−82　75:−66　76:−93　79:−88　80:−84　81:−94

118629xxx64.484　−1　67:−79　71:−76　73:−82　75:−66　76:−93　79:−88　80:−84　81:−94

118629xxx70.484　−1　61:−88　67:−73　71:−76　73:−82　75:−55　76:−87　79:−84　80:−91

118629xxx70.984　−1　61:−88　67:−72　71:−70　73:−73　75:−48　76:−83　79:−86　80:−89　81:−92

……

……

其中

Record Time 为 WIFI 信息记录时间。

Location Label 为位置标签信息(0－247)，当值为－1 时为标签信息缺失。

AP's Index 为 WIFI 信息数据库对应的记录位置。

RSS value 为 AP's Index 的接受信号强度(received signal strength)，RSS 值越大则 AP's Index 越有效。

在实验中我们选择 Trace_1－Trace_20 这 20 组实验数据，将大部分位置标签信息隐去，以判断迁移算法的领域自适应性。

在实验中选取 20 组 WIFI 数据 1—20。每组数据样本点个数如表 4 所列。

表 4　WIFI 数据大小设定

1 组(训练集)	320
2 组(测试集)	278
3 组(测试集)	244
4 组(测试集)	399
5 组(测试集)	499
6 组(测试集)	178
7 组(测试集)	267
8 组(测试集)	224
9 组(测试集)	285
10 组(测试集)	510
11 组(测试集)	280
12 组(测试集)	432
13 组(测试集)	353
14 组(测试集)	342
15 组(测试集)	435
16 组(测试集)	155
17 组(测试集)	297
18 组(测试集)	243
19 组(测试集)	474
20 组(测试集)	377

首先根据每组 WIFI 数据中的标签信息(从后台得知)绘制 WIFI 定位点示意图,此示意图为真实定位点坐标图。具体作法是将图 2 进行坐标化,再将 20 组位置信息依次在坐标中标记出,以显示 20 组 WIFI 数据的相对定位位置,并将此图作为衡量算法正确性的依据。具体图示如图 3 所示。

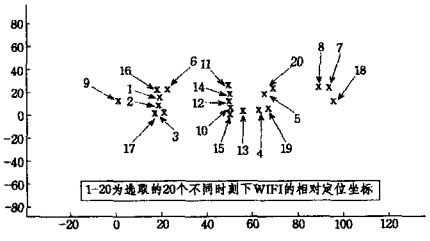

图 3 WIFI 定位点坐标示意图

然后我们选取第一组数据为训练样本,其余数据为测试样本。根据 CCMEBTL 算法计算出测试样本与训练样本的概率估计比 $r(x_j)$(见式(8)),并取其均值 $r(x)$,如表 5 所列。

表 5 不同数据设定下概率比(WIFI 数据)

训练集/测试集	Train1
Train1	r(x):0.9698
Test2	r(x):0.7145
Test3	r(x):0.7103
Test4	r(x):0.1296
Test5	r(x):2.5060 e−5
Test6	r(x):0.2449
Test7	r(x):1.4343 e−7
Test8	r(x):3.6560 e−7
Test9	r(x):0.6183
Test10	r(x):0.4321
Test11	r(x):0.4871
Test12	r(x):0.4424
Test13	r(x):0.2527
Test14	r(x):0.4204
Test15	r(x)0.3379
Test16	r(x):0.7354
Test17	r(x):0.6913
Test18	r(x):2.4436 e−10
Test19	r(x):0.0136
Test20	r(x):2.8234 e−5

通过表 5 与图 3 比较可知,真实训练点与测试点之间的距离趋势与概率估计比趋势相一致,即与第 1 位置点越接近,点的概率估计比越接近 1,相反则越接近 0。说明 CCMEBTL 算法通过迁移学习能尽量多地利用原有信息,实现了领域自适应[3,12]。该算法有很好的位置校正功能,可通过对 $r(x)$ 设定阈值来去除干扰信息,从而判断位置点。

2)算法对比分析

在 WIFI 定位实验中,在 3.3 节表 1 步骤 4 当中我们可以通过 CCMEBTL 算法求得目标域 $D2$ 空间的最小包含球球心,在源域标签信息已知的情况下比较源域与目标域最小包含球球心坐标,能够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]、CCMEBTL 3 种领域自适应算法进行了比较。首先定义平均误差距离 AED (average error distance),$AED=(\sum_{(x_i,y_i)\in D2}|f(x_i)-y_i|)/N$,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,此处选择 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 4 所示。

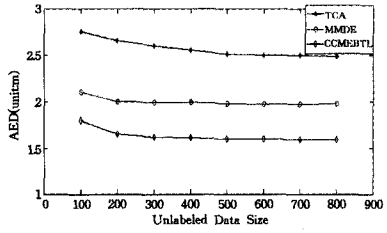

图 4 WIFI 不同算法下的 AED 对比

观察上图发现,CCMEBTL 算法的 AED 值相比其他算法更小,更加接近真实的 WIFI 定位,该算法有着较好的位置校正和预测功能。

4.3 人脸识别验证

人脸识别[15]数据检验

我们采用 ORL 人脸库,下载地址:http://download.csdn.net/source/1583590。在这里选择 10 幅人脸图像进行比较,其中前 5 幅为同一个人,只是面部表情存在差异,后 5 幅为另外一人,也是面部表情有差异,如图 5 所示。我们可以将面部表情的差异看作是一种数据扰动,而不同人图像的差异看作是一种分类。通过 CCMEBTL 算法能够消除相似领域数据之间的扰动,同时能够显著区分满足不同分布的领域数据,做到领域自适应。10 幅人脸图像为 92×112 的灰度图像,每幅图像含有 10304 个灰度数据。轮留选取其中一幅为训练图像,其余为测试图像,通过算法可计算出不同图像的概率比均值 $r(x)$ 参数,如表 6 所列。比较 $r(x)$ 参数值可评判图像之间的相似性(见式(10))。

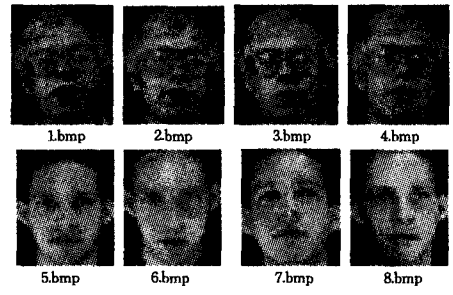

图 5 ORL 数据集

表 6 不同人脸图像的概率比

训练集/测试集	1.bmp	2.bmp	3.bmp	4.bmp	5.bmp	6.bmp	7.bmp	8.bmp
1.bmp	0.9982	0.8130	0.7883	0.8114	0.9795	0.5271 e−2	0.7189 e−2	0.8904 e−2
2.bmp	0.9574	0.9740	0.8281	0.8382	0.9743	0.8715 e−2	0.6739 e−2	0.7803 e−2
3.bmp	0.9686	0.7685	0.9955	0.7860	0.9443	0.5111 e−2	0.6970 e−2	0.8713 e−2
4.bmp	0.9497	0.7988	0.7850	0.9789	0.9881	0.6260 e−2	0.6040 e−2	0.8206 e−2
5.bmp	0.9110	0.8567	0.8324	0.8569	0.9998	0.7195 e−2	0.6801 e−2	0.8857 e−2
6.bmp	0.7747 e−2	0.6579 e−2	0.5626 e−2	0.6024 e−2	0.6583 e−2	0.9889	0.9006	0.8832
7.bmp	0.5730 e−2	0.6694 e−2	0.4502 e−2	0.4944 e−2	0.5361 e−2	0.7871	0.9248	0.7465
8.bmp	0.5847 e−2	0.6359 e−2	0.4380 e−2	0.4949 e−2	0.5368 e−2	0.8728	0.8656	0.9238

通过表 6 我们可以发现,同一人的人脸数据的 $r(x)$ 一般

(下转第 210 页)

[15] Wu Jun, Lin Zheng-kui, Lu Ming-yu. Asymmetric Semi-Supervised Boosting for SVM Active Learning in CBIR[C]// Proceedings of the ACM International Conference on Image and Video Retrieval. Xi'an, China, 2010

[16] Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning[J]. Advances in Neural Information Processing Systems, 1995(7): 231-238

[17] Liu Ying, Zhang Bai, Huang Li-hua, et al. A novel optimization parameters of support vector machines model for the land use/cover classification[J]. International Journal of Food, Agriculture & Environment, 2012, 10(2): 132-138

[18] Maulik U, Chakraborty D. A self-trained ensemble with semisupervised SVM: An application to pixel classification of remote sensing imagery[J]. Pattern Recognition, 2011, 44: 615-623

[19] 黄金杰，李士勇，蔡云泽. 一种建立粗糙数据模型的监督模糊聚类方法[J]. 软件学报，2005，16(6)：744-753

[20] 单丹丹，杜培军，夏俊士. 基于多分类器集成的"北京一号"小卫星遥感影像分类研究[J]. 遥感应用，2011，2：69-78

(上接第 191 页)

趋向于 1(一般大于 0.9)，不同人的人脸数据的 $r(x)$ 都较小并趋向于 0。CCMEBTL 通过迁移学习尽量多地利用原有信息实现了领域自适应，我们可以做到对人脸的有效识别。

同时通过 CCMEBTL 算法，在 3.3 节表 1 中可求得训练集(源域空间 D1)的最小包含球球心 c 的坐标和测试集(目标域域空间 D2)最小包含球球心 C1 的坐标，继而可求出不同人脸图像的球间距离，如表 7 所列。球间距离越小，表示源域与目标域相似度越高。

表 7 不同人脸图像的球间距离

训练集/测试集	1.bmp	2.bmp	3.bmp	4.bmp	5.bmp	6.bmp	7.bmp	8.bmp
1.bmp	9.53 e−14	4.57	3.44	4.72	11.10	13.79	12.86	14.87
2.bmp	4.5705	9.74 e−14	2.24	1.26	10.92	12.42	13.20	16.52
3.bmp	3.44	2.24	9.55 e−14	2.51	10.83	12.70	12.94	15.68
4.bmp	4.72	1.26	2.51	9.71 e−14	11.59	13.11	13.94	16.93
5.bmp	11.10	10.92	10.83	11.59	9.23 e−14	5.16	5.11	7.74
6.bmp	13.79	12.42	12.70	13.11	5.16	9.00 e−14	6.38	8.70
7.bmp	12.86	13.20	12.94	13.94	4.70	6.38	8.87 e−14	8.32
8.bmp	14.87	16.52	15.68	16.93	7.74	8.70	8.32	0.10e−14

通过表 7 观察可以发现，两大类数据内部子集的球心间距明显小于不同类子集之间的球心间距，即同一人的人脸图像球间距明显小于不同人的人脸图像球间距。结果显示算法能较好地体现不同领域之间的相关性，具有较好的领域自适应性。

结束语 本文将 MEB、CCMEB 理论应用在迁移学习研究上，提出了 MEBTL 算法和 CCMEBTL 算法。在求解目标域球心位置时尽可能多地利用到源域数据完成知识传递，并发现不同域之间的内部联系。最后通过比较不同域的概率统计比可实现数据的修正和校正。为了满足大样本数据集运算要求，引入了 CVM、CCMEB 理论。大量的实验内容验证了本文算法的有效性和快速性。应当指出本文算法仍有可深入研究之处，如何将其应用于数据分类和数据回归将是我们下一步的研究重点。

参 考 文 献

[1] Dai W, Yang Q, Xue G, et al. Boostimg for transerfer learning [C]// Proceedings of the 24th International Conference on Machine Learning. USA Corvasllis: ACM, 2007: 193-200

[2] Pan S J, Yang Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[3] Blitzer J, McDonald R, Percira F. Domain adaptation with structural correspondence learning [C]// Processings of the 2006 Conference on Empirical Methods in Natural Language Processing. PA USA: SIAM, 2006: 120-128

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

[5] Blitzer J, Crammer K, Kulesza A, et al. Learning bounds for domain adaptation[C]// Proceedings of the 21st Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT, 2008: 129-136

[6] Dai W, Xue G, Yang Q, et al. Co-clustering based classification for out-of-domain documents[C]// Proceedings of 13th ACM SIGKDD. New York: ACM, 2007: 210-219

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

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

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

[10] Fang S-H, Lin T-N. Indoor location system based on discriminant-adaptive neural network in IEEE 802.11 environments [J]. IEEE Transactions on Neural Networks, 2008, 19(11): 1973-1978

[11] Yang Q, Pan S J, Zheng V W. Estimating location using Wi-Fi [J]. Intelligent Systems, IEEE, 2008, 23(1): 8-13

[12] Satpal S, Sarawagi S. Domain Adaptation of Conditional Probability Models via Feature Subsetting [C]// Proceedings of PKDD. Heidelberg: Springer-Verlag Press, 2007, 4702: 224-235

[13] Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction[C]// Proc. 23rd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence. Chicago, IL, July 2008: 677-682

[14] Pan S J, Tsang I W, Kwok J T, et al. Domain Adaptation via Transfer Component Analysis [J]. IEEE Transactions on Neural Networks, 2011, 22(2): 199-210

[15] Osuna E, Freund R, Giroal F. Training support vector machines: an application to face diction[C]// Prceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Juan, 1997: 130-136