基于分类误差一致性准则的自适应知识迁移

梁 爽[①] 杭文龙*[②] 冯 伟[②] 刘学军[②]

[①](南京邮电大学地理与生物信息学院 南京 210023)

[②](南京工业大学计算机科学与技术学院 南京 211816)

摘 要：目前大多数迁移学习方法在利用源域数据辅助目标域数据建模时，通常假设源域中的数据均与目标域数据相关。然而在实际应用中，源域中的数据并非都与目标域数据的相关程度一致，若基于上述假设往往会导致负迁移效应。为此，该文首先提出分类误差一致性准则(CCR)，对源域与目标域分类误差的概率分布积分平方误差进行最小化度量。此外，该文提出一种基于CCR的自适应知识迁移学习方法(CATL)，该方法可以快速地从源域中自动确定出与目标域相关的数据及其权重，以辅助目标域模型的构建，使其能在提高知识迁移效率的同时缓解负迁移学习效应。在真实图像以及文本数据集上的实验结果验证了CATL方法的优势。

关键词：迁移学习；负迁移；概率分布；分类误差一致性规则

中图分类号：TP181 **文献标识码**：A **文章编号**：1009-5896(2019)11-2736-08

DOI: 10.11999/JEIT181054

Adaptive Knowledge Transfer Based on Classification-error Consensus Regularization

LIANG Shuang[①] HANG Wenlong[②] FENG Wei[②] LIU Xuejun[②]

[①](*School of Geographic and Biologic Information*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210023, *China*)

[②](*Department of Computer and Technology*, *Nanjing Tech University*, *Nanjing* 211816, *China*)

Abstract: Most current transfer learning methods are modeled by utilizing the source data with the assumption that all data in the source domain are equally related to the target domain. In many practical applications, however, this assumption may induce negative learning effect when it becomes invalid. To tackle this issue, by minimizing the integrated squared error of the probability distribution of the source and target domain classification errors, the Classification-error Consensus Regularization (CCR) is proposed. Furthermore, CCR-based Adaptive knowledge Transfer Learning (CATL) method is developed to quickly determine the correlative source data and the corresponding weights. The proposed method can alleviate the negative transfer learning effect while improving the efficiency of knowledge transfer. The experimental results on the real image and text datasets validate the advantages of the CATL method.

Key words: Transfer learning; Negative transfer; Probability distribution; Classification-error Consensus Regularization (CCR)

1 引言

迁移学习旨在通过大量已标注源域数据辅助目标域分类模型的建立[1]。主要研究内容有：(1)迁移什么：源域中何种类型的知识能迁移至目标域，通常采用实例[2]，特征[3–6]以及模型参数[7]3种类型；(2)如何迁移：源域中的知识能通过何种途径迁移至目标域，主流方法包括boosting[2]，支持向量机(Support Vector Machine, SVM)[8]以及模糊系统[9]等；(3)何时迁移：何时适合将源域中的知识迁移至目标域，确定是否对源域知识进行迁移。

传统迁移学习方法通常假设源域中的所有数据均与目标域数据具有相似的特征空间或数据分布，然而在实际应用中此假设未必成立。若从源域中强行迁移与目标域不相关的数据知识，可能会导致负

收稿日期：2018-11-20；改回日期：2019-04-30；网络出版：2019-05-16
*通信作者： 杭文龙 wlhang@njtech.edu.cn
基金项目：国家自然科学基金(61802177)，江苏省高校自然科学研究面上项目(18KJB520020)，南京邮电大学引进人才科研启动基金(NY219034)，江苏省重点研发计划(BE2015697)

Foundation Items: The National Nature Science Foundation of China (61802177), The Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (18KJB520020), NUPTSF (NY219034), Key Research and Development Program of Jiangsu Province (BE2015697)

迁移效应[2]。目前，相关学者针对传统迁移方法进行改进，通过自动识别出仅与目标域相关的源域数据知识来辅助目标域分类模型的构建。其中，自适应支持向量机(Adaptive Support Vector Machine, ASVM)[8]通过基于损失最小化原则的采样策略，从源域中选择与目标域相关的数据。然而，ASVM算法仅考虑源域数据与目标域数据是否相关，却忽略了其相关程度可能存在差异。选择性迁移学习机(Selective Transfer Machine, STM)[10]通过最小化由核均值匹配方法[11]所度量的源域数据和目标域数据分布误差，能够同时确定出与目标域相关的源域数据及其权重。然而，在目标域数据有限时，STM算法无法保证源域中相关数据权重估计的可靠性，且该方法求解权重的计算复杂度较高。此外，文献[12]提出了一种多源迁移学习方法，通过协同训练的思想利用多个源域数据和目标域已标注数据训练的弱分类器分别对目标域测试样本进行标注，并选择在多个弱分类器下具有相同标签的测试样本作为高可信度样本加入目标域训练集，最终利用目标域训练集获取最终目标域分类器。可以看出，文献[12]中必须使用多个源域以得到多个弱分类器对目标域测试样本的预测标签一致性进行判断，限制了其在单源域场景下的应用。文献[13]提出了一种基于部分相关实例特征知识的迁移学习方法，该方法利用协同聚类方法对源域和目标域数据特征进行聚类，找出并修正源域数据特征，使其与目标域接近，此时源域数据能更好地辅助目标域数据迁移。

近年来，由于深度学习方法具有强大的特征学习能力，被广泛应用于迁移学习领域[14–17]。目前，基于深度神经网络的迁移学习算法研究尚处于起步阶段，且几乎均属于特征迁移学习方法，即利用深度神经网络学习源域和目标域高度抽象的可迁移特征并用于辅助目标域模型的构建[14–17]。可以看出，目前基于深度神经网络的迁移学习算法大多是假设源域中的数据均与目标域数据相关，在特征提取过程中对深层特征进行分布一致性约束来学习可迁移特征，然而这种假设在真实应用场景并非总是成立，强行迁移容易对目标域深度模型构建产生影响甚至导致负迁移效应。

由上述方法看出，研究能快速确定出源域中相关数据及其权重的方法，实现源域知识的有效迁移很有必要。本文首先提出分类误差一致性准则(Classification-error Consensus Regularization, CCR)，在此基础上，提出一种基于CCR的自适应知识迁移学习方法(CCR-based Adaptive Transfer Learning, CATL)，自适应地选择出与目标域相关的源域数据及其权重，实现迁移同时能够有效缓解负迁移学习效应。

2　基本定义及一致性准则

2.1　基本数学符号和框架

首先介绍本文使用的数学符号：$\boldsymbol{Q}\in\mathbb{R}^{M\times N}$表示$(i,j)$位置为元素$\boldsymbol{Q}_{ij}$的$M\times N$矩阵，$\boldsymbol{\eta}=[\eta_1,\eta_2,\cdots,\eta_N]^{\mathrm{T}}\in\mathbb{R}^N$为列向量，$\|\boldsymbol{\eta}\|_p=\left(\sum_{i=1}^{N}|\eta_i|^p\right)^{1/p}$表示$\boldsymbol{\eta}\in\mathbb{R}^N$的$p$范数。$D_S=\{(\boldsymbol{x}_{S_1},y_{S_1}),(\boldsymbol{x}_{S_2},y_{S_2}),\cdots,(\boldsymbol{x}_{S_{\mathrm{NS}}},y_{S_{\mathrm{NS}}})\}$是有NS个样本的源域，$\boldsymbol{x}_{S_i}\in\boldsymbol{X}_S$为源域数据，$y_{S_i}\in\boldsymbol{Y}_S$为数据标签。$D_T=\{(\boldsymbol{x}_{T_1},y_{T_1}),(\boldsymbol{x}_{T_2},y_{T_2}),\cdots,(\boldsymbol{x}_{T_{\mathrm{NT}}},y_{T_{\mathrm{NT}}})\}$是目标域，$\boldsymbol{x}_{T_i}\in\boldsymbol{X}_T$为输入数据，$y_{T_i}\in\boldsymbol{Y}_T$为数据标签。不失一般性，考虑二分类情况，即$\boldsymbol{Y}_S,\boldsymbol{Y}_T\in\{-1,1\}$。

基于最小二乘损失函数的标准二分类学习系统可表示为

$$\left.\begin{aligned}&\min_{f}\quad \Phi(f)+C\sum_{i=1}^{N}\xi_i^2\\&\text{s.t.}\quad y_i=f(\boldsymbol{x}_i)+\xi_i,\quad \forall i=1,2,\cdots,N\end{aligned}\right\}\tag{1}$$

其中，数据$\boldsymbol{x}_i$在决策函数f下的分类误差为ξ_i，$\Phi(f)$为规则化项，平衡参数$C>0$。

由于所有的线性分类模型均可以表示为

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b\tag{2}$$

其中，$\phi(\boldsymbol{x})$为特征映射函数。$\boldsymbol{x}_i$在决策函数f下的分类误差表示为

$$\begin{aligned}\xi_i&=y_i-f(\boldsymbol{x}_i)=y_i-\left(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b\right)\\&=y_i-\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)-b\end{aligned}\tag{3}$$

2.2　分类误差一致性准则

基于式(3)，本节提出了分类误差一致性准则，通过最小化源域分类器与目标域分类器之间分类误差的不一致性，来提高目标域分类性能。

定义 1　(分类误差一致性准则，CCR) 根据式(3)，假设源域和目标域的分类误差分别为$\xi_{S_1},\xi_{S_2},\cdots,\xi_{S_{\mathrm{NS}}}$与$\xi_{T_1},\xi_{T_2},\cdots,\xi_{T_{\mathrm{NT}}}$，则分类误差一致性准则可表示为

$$\Delta=\sum_{i=1}^{\mathrm{NT}}\sum_{j=1}^{\mathrm{NS}}\left(\xi_{T_i}-\xi_{S_j}\right)^2\tag{4}$$

利用Parzen窗密度估计[18]对源域和目标域的分类误差进行估计，可以得到

$$P_S(\xi)=\frac{1}{\mathrm{NS}}\left(2\pi\sigma_S^2\right)^{-d/2}\sum_{i=1}^{\mathrm{NS}}\exp\left(-\|\xi-\xi_{S_i}\|^2/2\sigma_S^2\right)\tag{5}$$

$$P_T(\xi)=\frac{1}{\mathrm{NT}}\left(2\pi\sigma_T^2\right)^{-d/2}\sum_{j=1}^{\mathrm{NT}}\exp\left(-\left\|\xi-\xi_{T_j}\right\|^2/2\sigma_T^2\right) \tag{6}$$

其中，σ为高斯核宽。由式(5)以及式(6)看出，若使源域和目标域间知识共享的越多，则应使源域与目标域分类误差的概率分布积分平方误差最小化，即最小化$\int_{D_S,D_T}(P_S(\xi)-P_T(\xi))^2\mathrm{d}\xi$，有

$$\begin{aligned}&\int_{D_S,D_T}(P_S(\xi)-P_T(\xi))^2\mathrm{d}\xi\\&=\int_{D_S}(P_S(\xi))^2\mathrm{d}\xi-2\int_{D_S,D_T}P_S(\xi)\,P_T(\xi)\,\mathrm{d}\xi\\&\quad+\int_{D_T}(P_T(\xi))^2\mathrm{d}\xi\end{aligned} \tag{7}$$

无任何先验知识的情况下，假设$\int_{D_S}P_S(\xi)\,\mathrm{d}\xi=1$，则有

$$\begin{aligned}\int_{D_S}(P_S(\xi))^2\mathrm{d}\xi&=\int_{D_S}P_S(\xi)\,P_S(\xi)\,\mathrm{d}\xi\\&=\mathrm{E}[P_S(\xi)]\approx 1/\mathrm{NS}\end{aligned} \tag{8}$$

类似地，$\int_{D_T}(P_T(\xi))^2\mathrm{d}\xi\approx 1/\mathrm{NT}$，则式(7)表示为

$$\begin{aligned}&\mathrm{E}\left[P_S(\xi)\right]-2\int_{D_S,D_T}P_S(\xi)\,P_T(\xi)\,\mathrm{d}\xi+\mathrm{E}[P_T(\xi)]\\&\approx\frac{1}{\mathrm{NS}}-2\int_{D_S,D_T}P_S(\xi)\,P_T(\xi)\,\mathrm{d}\xi+\frac{1}{\mathrm{NT}}\end{aligned} \tag{9}$$

此时$\Delta_1=\int_{D_S,D_T}P_S(\xi)\,P_T(\xi)\,\mathrm{d}\xi$越大越好。根据文献[19]，有式(10)成立

$$\begin{aligned}&\int G\left(\xi,\xi_{T_j},\sigma_T\right)\,G\left(\xi,\xi_{S_i},\sigma_S\right)\mathrm{d}\xi\\&=G\left(\xi_{T_j}-\xi_{S_i},\sigma_T+\sigma_S\right)\end{aligned} \tag{10}$$

这里的高斯分布函数为$G\left(\xi,\xi_i,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-d/2}\cdot\exp\left(-\left\|\xi-\xi_i\right\|^2/2\sigma^2\right)$，则$\Delta_1$等价于

$$\begin{aligned}&\frac{1}{\mathrm{NS}\cdot\mathrm{NT}}\sum_{i=1}^{\mathrm{NT}}\sum_{j=1}^{\mathrm{NS}}\left(2\pi\left(\sigma_T^2+\sigma_S^2\right)\right)^{-d/2}\\&\cdot\exp\left(-\left\|\xi_{T_i}-\xi_{S_j}\right\|^2/2\left(\sigma_T^2+\sigma_S^2\right)\right)\end{aligned} \tag{11}$$

由于最大化e^{-x^2}等价于最小化x^2，若最小化$\int_{D_S,D_T}(P_S(\xi)-P_T(\xi))^2\mathrm{d}\xi$(即最大化$\Delta_1$)，应使$\sum_{i=1}^{\mathrm{NT}}\sum_{j=1}^{\mathrm{NS}}\left(\left\|\xi_{T_i}-\xi_{S_j}\right\|^2/2\left(\sigma_T^2+\sigma_S^2\right)\right)$最小化，即最小化$\sum_{i=1}^{\mathrm{NT}}\sum_{j=1}^{\mathrm{NS}}\left(\xi_{T_i}-\xi_{S_j}\right)^2$，即定义1。

3 基于CCR的自适应迁移学习

传统的迁移学习方法利用源域数据辅助目标域数据建模时，通常假设源域中的数据均与目标域数据相关。然而，现实场景中的源域数据并非都与目标域数据的相关程度一致，强行迁移则会影响目标域模型的识别精度甚至会导致负迁移效应。针对上述问题，本文利用所提CCR，最小化源域和目标域数据的分类误差，自适应选择出与目标域相关的源域数据及其相关程度，避免由于迁移不相关的源域数据而导致的负迁移效应。在此基础上，根据模型参数迁移方法，借鉴源域的模型参数知识辅助构建目标域数据分类模型。

本节提出一种基于CCR的自适应迁移学习方法(CATL)，该方法能够自适应识别出相关的源域数据，使得它们与目标域数据的分类误差的差异最小；此外，考虑到这些源域数据的相关性程度存在差异，本文将权重参数η_j引入分类误差ξ_{S_j}，并采用快速留一交叉验证法求解上述权重参数；最终，利用识别出的相关源域数据及其权重来辅助目标域分类模型的构建。

3.1 自适应迁移学习模型构建

由于引入权重参数，源域分类误差表示为线性组合$\sum_{j=1}^{\mathrm{NS}}\eta_j\xi_{S_j}$，则得到本文的自适应迁移学习模型

$$\begin{aligned}&\min_{\boldsymbol{w},b}\frac{1}{2}\left\|\boldsymbol{w}_T-\boldsymbol{w}_S\right\|^2+\frac{C}{2}\sum_{i=1}^{\mathrm{NT}}\sum_{j=1}^{\mathrm{NS}}\left(\xi_{T_i}-\eta_j\xi_{S_j}\right)^2\\&\text{s.t.}\quad y_{T_i}=\boldsymbol{w}_T^{\mathrm{T}}\phi\left(\boldsymbol{x}_{T_i}\right)+b_T+\xi_{T_i},\forall i\\&\qquad=1,2,\cdots,\mathrm{NT},\forall j=1,2,\cdots,\mathrm{NS}\end{aligned} \tag{12}$$

这里的权重参数η_j用于确定源域第j个数据是否被迁移，权重η_j越大说明对应的源域数据与目标域数据相关性越大。由式(12)看出，目标函数可以通过最小化源域和目标域之间的分类误差来确定相关的源域数据及其权重。式(12)的拉格朗日形式表示为

$$\begin{aligned}L=&\frac{1}{2}\left\|\boldsymbol{w}_T-\boldsymbol{w}_S\right\|^2+\frac{C}{2}\sum_{i=1}^{\mathrm{NT}}\sum_{j=1}^{\mathrm{NS}}\left(\xi_{T_i}-\eta_j\xi_{S_j}\right)^2\\&-\sum_{i=1}^{\mathrm{NT}}\alpha_i\left(\boldsymbol{w}_T^{\mathrm{T}}\phi\left(\boldsymbol{x}_{T_i}\right)+b_T+\xi_{T_i}-y_{T_i}\right)\end{aligned} \tag{13}$$

其中，α_i为拉格朗日乘子，通过求解L关于$\boldsymbol{w}_T$,b_T,ξ_{T_i}和α_i的导数，并将结果代入目标表达式(12)的等式约束中，可以得到

$$\sum_{i=1}^{\mathrm{NT}}\alpha_i\phi\left(\boldsymbol{x}_{T_i}\right)^{\mathrm{T}}\phi\left(\boldsymbol{x}_{T_k}\right)+b_T+\frac{\alpha_k}{C\cdot\mathrm{NS}}=y_{T_k}-\boldsymbol{w}_S^{\mathrm{T}}\phi\left(\boldsymbol{x}_{T_k}\right)-\frac{1}{\mathrm{NS}}\sum_{j=1}^{\mathrm{NS}}\eta_j\xi_{S_j}\tag{14}$$

利用核函数，即$K\left(\boldsymbol{x},\boldsymbol{x}'\right)=\phi\left(\boldsymbol{x}\right)^{\mathrm{T}}\phi\left(\boldsymbol{x}'\right)$，将式(14)中的线性等式组合表示为矩阵形式

$$\begin{bmatrix}\boldsymbol{K}+\dfrac{1}{C}\boldsymbol{\Lambda} & 1\\ 1^{\mathrm{T}} & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ b\end{bmatrix}=\begin{bmatrix}\boldsymbol{y}-\hat{\boldsymbol{y}}-\dfrac{1}{\mathrm{NS}}\displaystyle\sum_{j=1}^{\mathrm{NS}}\eta_j\xi_{S_j}1\\ 0\end{bmatrix}\tag{15}$$

这里的$\boldsymbol{\Lambda}=\mathrm{diag}\left\{\mathrm{NS}^{-1},\mathrm{NS}^{-1},\cdots,\mathrm{NS}^{-1}\right\}$，$\boldsymbol{y}$是目标域已知标签样本的标签向量，即$\boldsymbol{y}$=$[y_{T_1},y_{T_2},\cdots,y_{T_{\mathrm{NT}}}]^{\mathrm{T}}$，$\hat{\boldsymbol{y}}$=$\left[\boldsymbol{w}_S^{\mathrm{T}}\phi\left(\boldsymbol{x}_{T_1}\right),\boldsymbol{w}_S^{\mathrm{T}}\phi\left(\boldsymbol{x}_{T_2}\right),\cdots,\boldsymbol{w}_S^{\mathrm{T}}\phi\left(\boldsymbol{x}_{T_{\mathrm{NT}}}\right)\right]^{\mathrm{T}}$，则式(15)的解析解为

$$\begin{bmatrix}\boldsymbol{\alpha}\\ b\end{bmatrix}=\boldsymbol{Q}\begin{bmatrix}\boldsymbol{y}-\hat{\boldsymbol{y}}-\dfrac{1}{\mathrm{NS}}\displaystyle\sum_{j=1}^{\mathrm{NS}}\eta_j\xi_{S_j}1\\ 0\end{bmatrix}\tag{16}$$

这里矩阵$\boldsymbol{H}$表示等式(15)等号左侧的矩阵，$\boldsymbol{Q}$是矩阵$\boldsymbol{H}$的逆矩阵。

3.2 相关源域数据及其权重的快速决策

求解权重向量$\boldsymbol{\eta}$是建立学习模型的关键步骤，本文借鉴快速留一交叉验证法，求解源域中相关数据及其权重的方法。通过最小化留一误差求解权重，是对泛化误差的几乎无偏估计[20]，可以有效缓解目标数据有限情况下权重不可靠的问题。

定理 1　将式(16)重新表示为$\left[\boldsymbol{\alpha}'^{\mathrm{T}},b'\right]^{\mathrm{T}}=\boldsymbol{Q}\left[\boldsymbol{y}^{\mathrm{T}}-\hat{\boldsymbol{y}}^{\mathrm{T}},0\right]^{\mathrm{T}}$，$\left[\boldsymbol{\alpha}''_j{}^{\mathrm{T}},b''_j\right]^{\mathrm{T}}=\boldsymbol{Q}\left[\xi_{S_j}1^{\mathrm{T}},0\right]^{\mathrm{T}}$以及$\boldsymbol{\alpha}=\boldsymbol{\alpha}'-1/\mathrm{NS}\cdot\sum_{j=1}^{\mathrm{NS}}\eta_j\boldsymbol{\alpha}''_j$，则样本$\boldsymbol{x}_{T_i}$的预测标签$\tilde{y}_i,i=1,2,\cdots,\mathrm{NT}$可以表示为

$$\tilde{y}_i=y_i-\alpha'_i/\boldsymbol{Q}_{ii}+\boldsymbol{\eta}^{\mathrm{T}}\xi_S/\boldsymbol{Q}_{ii}\tag{17}$$

这里$\boldsymbol{\xi}_S$=$[\xi_{S_1},\xi_{S_2},\cdots,\xi_{S_{\mathrm{NS}}}]^{\mathrm{T}}$，即为源域分类误差的向量形式。

证明过程略。

从式(17)看出，模型参数$\boldsymbol{\alpha}$与权重向量$\boldsymbol{\eta}$线性相关。只要确定出权重$\eta_j,\ j=1,2,\cdots,\mathrm{NS}$，便能建立目标域分类模型。显而易见，最优$\eta_j$可以通过样本$\boldsymbol{x}_{T_i}$的输出，即$\tilde{y}_iy_i$的值来确定。然而直接最大化$\mathrm{sign}(\tilde{y}_iy_i)$求和函数是非凸优化问题，求解较为困难。因此，本文利用式(18)来最小化预测误差

$$l\left(\tilde{y}_i,y_i\right)=\left|1-\tilde{y}_iy_i\right|_+=\left|y_i\frac{\alpha'_i-\boldsymbol{h}^{\mathrm{T}}\boldsymbol{A}''_i}{\boldsymbol{Q}_{ii}}\right|_+\tag{18}$$

这里$|x|_+=\max\left\{0,x\right\}$。综上，最优化权重向量$\boldsymbol{\eta}$转化为求解式(19)的目标函数

$$\min_{\eta}\sum_{i=1}^{N}l\left(\tilde{y}_i,y_i\right),\ \ \mathrm{s.t.}\left\|\boldsymbol{\eta}\right\|_p\le1,\eta_j\ge0\tag{19}$$

这里p范数用于向量$\boldsymbol{\eta}$的规则化约束，能缓解由样本量较少引起的过拟合。由于L_2型正则化约束能保证权重的稳定性，通过投影子梯度下降法优化算法[21]能方便求解式(19)。本文取$p=2$，即L_2型约束。采用的损失函数为关于变量$\boldsymbol{\eta}$的利普希茨连续(Lipschitz-continuous)凸损失函数。根据文献[21]可知，在第t次迭代过程中，$\boldsymbol{\eta}^{(t)}$施加的权重为$(t+1)$，采用投影子梯度下降法优化算法的收敛速度可以达到$O\left(1/t\right)$。

3.3 决策函数以及时间复杂度分析

根据式(16)，在已求解权重向量$\boldsymbol{\eta}$的基础上得到$\boldsymbol{\alpha}$，此时目标域模型参数为$\boldsymbol{w}=\sum_{i=1}^{\mathrm{NT}}\alpha_i\phi\left(\boldsymbol{x}_i\right)$，最终得到目标域上的分类模型

$$f\left(\boldsymbol{z}\right)=\boldsymbol{w}^{\mathrm{T}}\phi\left(\boldsymbol{z}\right)=\sum_{i=1}^{\mathrm{NT}}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi\left(\boldsymbol{z}\right)\tag{20}$$

在训练阶段，CATL算法的时间复杂度为$O\left(\mathrm{NT}^3+t_{\max}\cdot\mathrm{NS}\cdot\mathrm{NT}\right)$，NT表示目标域训练样本个数，NS表示源域样本个数，$t_{\max}$表示求解向量$\boldsymbol{\eta}$所需的迭代次数。第1项是求解矩阵$\boldsymbol{O}$的时间复杂度，第2项是求解权重向量$\boldsymbol{\eta}$的时间复杂度。在预测阶段，与传统迁移学习方法不同，CATL算法仅仅利用相关的源域数据辅助目标域模型建立，因此CATL算法所需的预测时间相对较少。

4　实验结果与分析

4.1 实验数据

为验证本文算法的有效性，本节将在常用图像以及文本数据集上对CATL算法进行分析与评估。

实验中使用两个图像数据集USPS和MNIST中的部分数字。USPS和MNIST均是手写字体数据集，其中数字“0”，“4”和“6”来自USPS数据集，“7”和“9”来自MNIST数据集。尽管USPS和MNIST均是数字图像，然而由于手写字体比较潦草，给自动识别带来困难，特别是部分相近的数字，如数字“7”和“9”，数字“4”和“9”以及数字“0”和“6”，因此本文选择此3组数据作为测试数据集。参考文献[22]，本文实验从USPS中随机抽取1800幅图像作为源域数据，从MNIST中随机抽取2000幅图像作为目标域数据。

在所有实验中，统一将图像缩放至16×16像素并转换为灰度图，实验数据集详细设置如表1所示。

表 1 图像数据集USPS及MNIST中源域数据与目标域数据的详细设置

任务	源域数据		目标域数据	
	正类	负类	正类	负类
1	USPS7	USPS9	MNIST7	MNIST9
2	USPS4	USPS9	MNIST4	MNIST9
3	USPS0	USPS6	MNIST0	MNIST6

此外，利用文本数据集—新闻数据集(20-Newsgroups)[22]来评估所提算法的分类效果。所采用数据集包含4个大类，即comp, rec, sci以及talk，每个大类中分别包含4个子类。在实验中，通过使用任意两类中的两个子类作为源域，选择其它两个子类作为目标域。本文利用词频特征提取法[22]对数据进行预处理。实验数据集详细设置如表2所示。

4.2 实验设置

本文实验共采用如下7种对比算法：(1) 最小二乘支持向量机(Least Square Support Vector Machine, LSSVM)[23]; (2) 自适应支持向量机(Adaptive SVM, ASVM)[8]; (3) 跨领域支持向量机(Cross-Domain SVM, CDSVM)[24]; (4) Boosting迁移学习(Boosting for Transfer Learning, TrAdaBoost)[2]; (5) 选择性迁移学习机(Selective Transfer Machine, STM)[10]; (6) 部分相关实例特征迁移学习(Partial Related Instance-Feature based transfer learning, PRIF)[13]; (7) 本文提出的基于L_2型约束的自适应迁移算法(CCR-based Adaptive Transfer Learning, CATL_2)。其中，算法(2)—(5)，(7)迁移学习算法均属于归纳式迁移学习，即目标域中有少量已标注的数据可用于目标域分类模型的构建。本文提出的CATL_2算法不仅能从源域中选择相关数据，而且同时能确定出数据对应的权重参数。此外，为了验证各迁移算法的有效性，LSSVM算法将已标注的目标域数据作为训练样本，STM算法将已标注的目标域数据的权重设置为1，并与相关的源域数据一起以构成新的训练样本。对于PRIF算法，原文采用直推式迁移学习方法对协同聚类后的源域数据进行迁移，为了公平对比，本文采用文献[20]中的自适应迁移学习方法。

在实验中，所有对比方法均执行10次并给出均值，源域与目标域所使用的核函数均为高斯核 $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\gamma \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2\right)$，平衡参数$C$和高斯核宽参数$\gamma$均采用5折交叉验证方法从集合{1e − 5, 1e − 4, 1e − 3, 1e − 2, 1e − 1, 1e0, 1e1, 1e2, 1e3, 1e4, 1e5}中选取。在通过投影子梯度下降法优化算法求解最优化权重向量η时，迭代次数设为100。根据参考文献[10]，STM参数设置为$\text{B} = 1000$, $\varepsilon = \sqrt{\text{NT} - 1}/\sqrt{\text{NT}}$。本文采用准确度(ACC)[25]结果作为算法的评价指标。

表 2 文本数据集20-Newsgroups中源域数据与目标域数据的详细设置

任务	源域数据		目标域数据	
	正类	负类	正类	负类
1	comp.graphics	rec.autos	comp.os.ms-windows.misc	rec.motorcycles
2	comp.sys.ibm.pc.hardware	rec.sport.baseball	comp.sys.mac.hardware	rec.sport.hokey
3	sci.crypt	talk.politics.guns	sci.electronics	talk.politics.mideast
4	sci.med	talk.politics.misc	sci.space	talk.religion.misc
5	rec.autos	talk.politics.guns	rec.motorcycles	talk.politics.mideast
6	rec.sport.baseball	talk.politics.misc	rec.sport.hokey	talk.religion.misc

4.3 实验结果

4.3.1 图像数据集分类结果

对上述7种对比算法在图像数据集上的实验结果进行验证与分析，得到如下结论：

(1) 表3表示不同算法在手写字体图像数据集上的平均分类精度。在多数情况下，CATL_2表现出较好的分类效果。主要原因在于，CDSVM算法和ASVM算法利用了所有的源域数据，或者仅仅利用了部分源域数据而忽略了它们的权重；TrAdaBoot算法对数据分布(由KL距离度量)较为敏感，其分类性能在目标域已标注数据较少时不够理想；STM算法需要较多的目标域数据以保证核均值匹配方法对权重估计的准确性，因而较少的目标域数据会导致源域数据权重分配不精确；PRIF算法通过自适应学习源域特征，能够纠正源域中不相关的实例特

表 3　各种算法在图像任务上的分类精度

任务	已标注样本	LSSVM	CDSVM	ASVM	TrAdaBoost	STM	PRIF	$CATL_2$
1	4	0.5287	0.5611	0.5913	0.5799	0.6018	0.6245	**0.6359**
	6	0.5520	0.5800	0.6094	0.6133	0.6298	0.6384	**0.6477**
	8	0.5897	0.6112	0.6266	0.6007	0.6319	0.6421	**0.6528**
	10	0.6030	0.6392	0.6502	0.6213	0.6487	0.6539	**0.6672**
	12	0.6381	0.6461	0.6383	0.6588	0.6643	0.6753	**0.6791**
	14	0.6541	0.6587	0.6754	0.6682	0.6901	0.6982	**0.7014**
2	4	0.5354	0.5743	0.5998	0.5887	0.5983	**0.6223**	0.6133
	6	0.5897	0.5992	0.6293	0.5903	0.6426	0.6478	**0.6520**
	8	0.6276	0.6387	0.6492	0.6690	0.6803	0.6893	**0.6927**
	10	0.6508	0.6641	0.6843	0.6905	0.7067	0.7029	**0.7168**
	12	0.6892	0.6698	0.6988	0.7123	0.7234	0.7326	**0.7387**
	14	0.7098	0.7156	0.7207	0.7076	0.7266	0.7391	**0.7421**
3	4	0.6578	0.6903	0.7026	0.6873	0.7235	0.7472	**0.7492**
	6	0.7013	0.7445	0.7529	0.7354	0.7541	0.7632	**0.7726**
	8	0.7452	0.7695	0.7721	0.7455	0.7618	07726	**0.7829**
	10	0.7762	0.7803	0.7789	0.7836	0.7928	0.7918	**0.8193**
	12	0.7923	0.7944	0.8034	0.7994	0.8288	0.8172	**0.8301**
	14	0.8234	0.8213	0.8178	0.8145	0.8397	0.8263	**0.8452**

征，能够达到较好的识别精度，但聚类个数对该方法影响较大。本文提出的$CATL_2$算法，能够快速且精确地从源域中确定出与目标域相关的数据及其权重，从而提高知识迁移的效率同时缓解负迁移学习效应。

(2) 特别是在目标域上的已标注训练数据较少时，迁移学习算法的分类性能几乎都优于传统的LSSVM算法，其中$CATL_2$表现出更好的分类性能，而随着已标注训练数据的不断增加，迁移学习方法的分类性能

与LSSVM算法逐渐接近。主要原因在于，利用目标域上丰富的已标注数据能帮助LSSVM算法构建较为稳健的学习模型。

4.3.2　文本数据集分类结果

基于Boosting的迁移学习是对Adaboost学习框架的拓展。它通过组合源域和目标域样本，然后在迭代过程中降低源域数据的权重，以减少对学习过程的影响。根据文献[2]，对于文本数据集，TrAdaBoost所需的迭代次数至少为50，这需要花费较多的计算时间。因此，在数据规模偏大的文本数据集部分不采用TrAdaBoost方法。具有不同数目的已标注目标域样本的分类结果如图1所示。图1(a)—1(f)展示了6种算法在6个文本数据集分类任务上的平均分类准确度。根据这些结果，得到以下观察结果：

(1) 在大多数情况下，$CATL_2$在分类准确性方面优于其它方法。这主要归因于基于CCR准则的自适应源域数据及其权重选择策略。

(2) 在目标域已标注数据较少时，CDSVM比LSSVM具有更好的分类精度，随着标注目标样本的增加，这一优势迅速减弱。在大多数情况下，ASVM和STM可以获得比CDSVM更好的结果，这主要是因为ASVM和STM均可以选择性利用源域数据，而PRIF通过纠正源域实例不相关特征达到了与所提算法较为接近的分类性能。由于$CATL_2$能够准确地选择相关的源域数据及其相应的权重并构建目标分类器，因此其分类性能更加突出。

5　结束语

本文首先提出了一种基于分类误差的一致性准则(CCR)，在此基础上提出了一种自适应知识迁移学习方法(CATL)，使用快速留一交叉验证法从源域中确定出与目标域相关的数据及其权重，该方法可以自适应地选择出与目标域相关的源域数据及其权重用于迁移的同时能够有效缓解负迁移学习效应，实验结果验证了本文所提自适应迁移方法的有效性。虽然本文方法能明显提高目标域分类性能，但仍有待改进。例如，在建立目标域分类模型时需要预先确定出源域的模型参数及误差，如何更优地选择源域参数进一步提高目标域分类性能，并拓展到多源迁移是将来研究的工作。

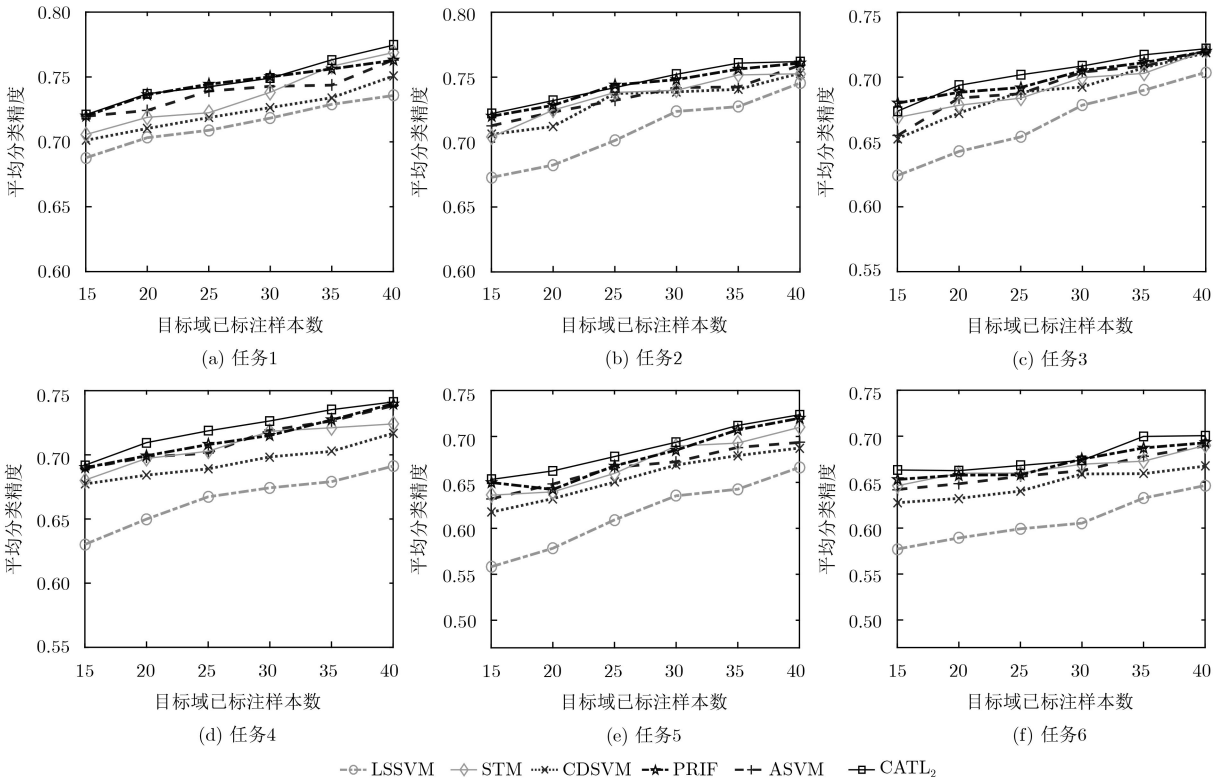

图 1 6种对比算法在文本数据集上的分类精度

参 考 文 献

[1] DENG Zhaohong, JIANG Yizhang, CHOI K S, *et al*. Knowledge-leverage-based TSK fuzzy system modeling[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, 24(8): 1200–1212. doi: 10.1109/TNNLS.2013.2253617.

[2] DAI Wenyuan, YANG Qiang, XUE Guirong, *et al*. Boosting for transfer learning[C]. The 24th International Conference on Machine Learning, Corvalis, USA, 2007: 193–200. doi: 10.1145/1273496.1273521.

[3] JIANG Yizhang, DENG Zhaohong, CHUNG F L, *et al*. Recognition of epileptic EEG signals using a novel multiview TSK fuzzy system[J]. *IEEE Transactions on Fuzzy Systems*, 2017, 25(1): 3–20. doi: 10.1109/TFUZZ.2016.2637405.

[4] ZHUANG Fuzhen, LUO Ping, DU Changying, *et al*. Triplex transfer learning: Exploiting both shared and distinct concepts for text classification[J]. *IEEE Transactions on Cybernetics*, 2014, 44(7): 1191–1203. doi: 10.1109/TCYB.2013.2281451.

[5] PAN S J, NI Xiaochuan, SUN Jiantao, *et al*. Cross-domain sentiment classification via spectral feature alignment[C]. Proceedings of the 19th International Conference on World Wide Web, Raleigh, USA, 2010: 751–760. doi: 10.1145/1772690.1772767.

[6] ZANG Shaofei, CHENG Yuhu, WANG Xuesong, *et al*. Semi-supervised transfer discriminant analysis based on cross-domain mean constraint[J]. *Artificial Intelligence Review*, 2018, 49(4): 581–595. doi: 10.1007/s10462-016-9533-3.

[7] WANG Guanjin, ZHANG Guangquan, CHOI K S, *et al*. Deep additive least squares support vector machines for classification with model transfer[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, 49(7): 1527–1540. doi: 10.1109/TSMC.2017.2759090.

[8] YANG Jun, YAN Rong, and HAUPTMANN A G. Adapting SVM classifiers to data with shifted distributions[C]. The Seventh IEEE International Conference on Data Mining Workshops, Omaha, USA, 2007: 69–76. doi: 10.1109/ICDMW.2007.37.

[9] JIANG Yizhang, DENG Zhaohong, CHUNG F L, *et al*. Realizing two-view TSK fuzzy classification system by using collaborative learning[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017, 47(1): 145–160. doi: 10.1109/TSMC.2016.2577558.

[10] CHU Wensheng, DE LA TORRE F, and COHN J F. Selective transfer machine for personalized facial action unit detection[C]. 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 3515–3522. doi: 10.1109/CVPR.2013.451.

[11] GRETTON A, SMOLA A, HUANG Jiayuan, *et al*. Covariate Shift by Kernel Mean Matching[M]. QUIÑONERO-CANDELA J, SUGIYAMA M, SCHWAIGHOFER A, *et al*. Dataset Shift in Machine Learning. Cambridge, USA: MIT

Press, 2009: 131–160. doi: 10.7551/mitpress/9780262170055.003.0008.

[12] CHENG Yuhu, WANG Xuesong, and CAO Ge. Multi-source tri-training transfer learning[J]. *IEICE Transactions on Information and Systems*, 2014, E97-D(6): 1668–1672. doi: 10.1587/transinf.e97.d.1668.

[13] WANG Yunyun, ZHAI Jie, LI Yun, *et al*. Transfer learning with partial related "instance-feature" knowledge[J]. *Neurocomputing*, 2018, 310: 115–124. doi: 10.1016/j.neucom.2018.05.029.

[14] CHEN Minmin, XU Zhixiang, WEINBERGER K Q, *et al*. Marginalized denoising autoencoders for domain adaptation[C]. The 29th International Conference on Machine Learning, Edinburgh, Scotland, 2012: 1627–1634.

[15] ZHOU J T, PAN S J, TSANG I W, *et al*. Hybrid heterogeneous transfer learning through deep learning[C]. The 28th AAAI Conference on Artificial Intelligence, Québec City, Canada, 2014: 2213–2219.

[16] GLOROT X, BORDES A, and BENGIO Y. Domain adaptation for large-scale sentiment classification: A deep learning approach[C]. The 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011: 513–520.

[17] LONG Mingsheng, WANG Jianmin, CAO Yue, *et al*. Deep learning of transferable representation for scalable domain adaptation[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(8): 2027–2040. doi: 10.1109/TKDE.2016.2554549.

[18] PARZEN E. On estimation of a probability density function and mode[J]. *The Annals of Mathematical Statistics*, 1962, 33(3): 1065–1076. doi: 10.1214/aoms/1177704472.

[19] DENG Zhaohong, CHUNG F L, and WANG Shitong. FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation[J]. *Pattern Recognition*, 2008, 41(4): 1363–1372. doi: 10.1016/j.patcog.2007.09.013.

[20] TOMMASI T, ORABONA F, and CAPUTO B. Learning categories from few examples with multi model knowledge transfer[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(5): 928–941. doi: 10.1109/TPAMI.2013.197.

[21] LACOSTE-JULIEN S, SCHMIDT M, and BACH F. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method[J]. arXiv:1212.2002, 2012.

[22] LONG Mingsheng, WANG Jianmin, DING Guiguang, *et al*. Transfer learning with graph co-regularization[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(7): 1805–1818. doi: 10.1109/TKDE.2013.97.

[23] SUYKENS J A K and VANDEWALLE J. Least squares support vector machine classifiers[J]. *Neural Processing Letters*, 1999, 9(3): 293–300. doi: 10.1023/a:1018628609742.

[24] BART E and ULLMAN S. Cross-generalization: Learning novel classes from a single example by feature replacement[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005: 672–679. doi: 10.1109/CVPR.2005.117.

[25] GU Xiaoqing, CHUNG F L, and WANG Shitong. Bayesian Takagi-Sugeno-Kang fuzzy classifier[J]. *IEEE Transactions on Fuzzy Systems*, 2017, 25(6): 1655–1671. doi: 10.1109/TFUZZ.2016.2617377.

梁　爽：女，1987年生，讲师，研究方向为机器学习、信号处理.
杭文龙：男，1988年生，讲师，研究方向为机器学习、模式识别.
冯　伟：男，1995年生，硕士生，研究方向机器学习、模式识别.
刘学军：男，1970年生，教授，硕士生导师，研究方向为数据挖掘、大数据分布式处理.