基于类内最大均值差异的无监督领域自适应算法 *

蔡瑞初[1]，李嘉豪[1]，郝志峰[1,2]

(1. 广东工业大学 计算机学院，广州 510006; 2. 佛山科学技术学院 数学与大数据学院，广东 佛山 528000)

摘 要：传统的无监督领域自适应算法在对齐总体分布时存在分类信息流失问题，难以保证迁移学习效果。针对这个问题，提出了一种基于类内最大均值差异的分布对齐策略。该策略首先预测所有样本的伪标签，然后借助伪标签样本信息依次对齐每个类别的领域类内分布。在深度学习框架下，所提算法能够有效保留分类信息，提高了目标领域的预测能力。与传统算法比较，实验结果表明，所提算法在多个基准数据集上获得了最优的迁移学习效果。

关键词：领域自适应；无监督学习；神经网络；最大均值差异

中图分类号：TP183 doi: 10.19734/j.issn.1001-3695.2019.03.0042

Unsupervised domain adaptation with intra-class maximum mean discrepancy

Cai Ruichu[1], Li Jiahao[1], Hao Zhifeng[1,2]

(1. *School of Computers, Guangdong University of Technology, Guangzhou* 510006, *China*; 2. *College of Mathematics & Big Data, Foshan Guangdong* 528000, *China*)

Abstract: In the unsupervised domain adaptation area, loss of label information is still an open problem during the alignment of the global distribution, and thus the effect of transfer learning is difficult to guarantee. To alleviate this problem, the proposed algorithm adopted a distribution alignment strategy based on the intra-class maximum mean discrepancy. This strategy firstly predicted the pseudo labels for all samples, then aligned the intra-class distributions of two domains with the help of the predicted labels. Under the deep learning framework, the proposed algorithm effectively avoided label information being washed away and greatly improved the prediction ability on the target domain. The experimental results show that the proposed algorithm outperforms the traditional algorithms on the benchmarks.

Key words: domain adaptation; unsupervised learning; neural network; maximum mean discrepancy

0 引言

随着大数据时代的来临，各领域已经积累了大量无标签数据，对这些无标签数据进行高效分析和学习已经成为迫切的任务。在现有的迁移学习[1]和领域自适应[2]等相关研究中，无监督领域自适应算法能够从有标注的源领域数据中提取有效的分类知识，并把这些知识迁移到无标注的目标领域数据中，避免了大量的数据标注任务，是目前主流的一种研究思路。

现有的无监督领域自适应算法主要采取分布对齐策略。Gong 等人[3]把两个领域的特征空间映射到再生核希尔伯特空间中，很好地解决了分布差异评估问题。Cortes 等人[4]对两个领域的样本进行重加权，大大减少了领域间的分布偏差。Xu 等人[5]对两个领域的样本进行挑选，有效避免了领域专用信息的干扰。然而，这些方法没有考虑样本特征之间的关系，因此，一些方法从特征子空间中对齐领域分布。Song 等人[6]对样本特征进行筛选，有效保留了领域无关的特征。Mourragui 等人[7]对样本特征进行线性加权，有效减少了领域专用特征的影响。Yan 等人[8]通过核方法寻找领域一致的子空间，有效保留了数据的重要性质。张春荣等人[9]使用核投影技术寻找高维子空间，有效减少词性标注的错误率。Liu 等人[10]为每个领域配置一个投影矩阵，从而使投影矩阵和投影结果均具备领域一致性。

考虑到神经网络拥有出色的特征提取能力，不少工作利用深度学习重新设计无监督领域自适应算法。Tzeng 等人[11]把最大均值差异技术[12]引入到神经网络中，解决了图像的自适应问题。在文献[11]的基础上，Long 等人[13~15]把多核最大均值差异技术[16]嵌入到多个领域专用层中，进一步提升了目标领域的分类精度。此外，Ghifary 等人[17]将自动编码器用于重构目标领域特征，有效对齐了源领域类别空间和目标领域重构空间。在文献[17]的基础上，Lin 等人[18]将最大均值差异技术引入到自动编码器中，有效减少了重构空间中的领域分布差异。在机器翻译中，丁亮等人[19]将 Bi-LSTM[20]用于构建自动编码器，有效提高了翻译系统的性能。最新的对抗思想[21]也被用于算法设计中。Ganin 等人[22]将梯度反转技术引入到多任务学习[23]模型中，有效抑制了特征提取器输出领域专用信息。Saito 等人[24]最大化两个分类器的评估差异，大大增强了任务决策边界的作用。此外，Liu 等人[25]固定生成器的权重，有效对齐了领域联合分布和边缘分布。在文献[25]的基础上，Tzeng 等人[26]把对抗模型或判别模型引入到生成器中，高效完成复杂的领域自适应任务。

然而，上述所有模型均存在分类信息流失问题。这是因

收稿日期：2019-03-03；**修回日期：**2019-04-27 **基金项目：**NSFC-广东联合基金资助项目（U1501254）；国家自然科学基金资助项目（61876043，61472089）；广东省自然科学基金资助项目（2014A030306004，2014A030308008）；广东省科技计划项目（2015B010108006，2015B010131015）；广东特支计划项目（2015TQ01X140）；广州市珠江科技新星项目（201610010101）；广州市科技计划项目（201604016075）

作者简介：蔡瑞初(1983-)，男，浙江温州人，教授，博士，主要研究方向为数据挖掘、机器学习等(cairuichu@gmail.com)；李嘉豪(1992-)，男(通信作者)，广东新兴人，硕士研究生，主要研究方向为深度学习、迁移学习等(jiahaoli.gdut@gmail.com)；郝志峰(1968-)，男，江苏苏州人，教授，博士，主要研究方向为机器学习、人工智能等.

为总体分布对齐策略不能保证领域共享知识就是模型分类知识。在类别不平衡情形下，总体分布对齐策略还会阻碍小样本类别信息的学习，从而导致严重的类别信息流失问题。

为此，所提算法首先预测所有样本的伪标签，然后使用类内最大均值差异(intra-class maximum mean discrepancy)技术对齐两个领域的类内分布。这种做法既能保留分类信息，又能减少领域专用信息的干扰。实验结果表明基于类内最大均值差异的无监督自适应算法能够防止分类信息的流失，获得了最优的分类效果。

接下来，本文的剩余部分首先介绍一般的无监督领域自适应模型的相关理论，然后根据这个理论提出一种基于类内最大均值差异的无监督领域自适应算法。最后通过一些对比实验验证所提算法的性能。

1　领域自适应的风险上界

为了完成领域自适应任务，目标函数的设计除了考虑经验后验分布 $q(Y|g(X))$ 和源领域后验分布 $p_S(Y|g(X))$ 的关系，还要考虑源领域后验分布 $p_S(Y|g(X))$ 和目标领域后验分布 $p_T(Y|g(X))$ 的关系。为了定量分析目标函数，首先要做的事情是定义两个分布的差异。在给定领域 $D\in\{S,T\}$ 和函数 g 后，任意两个后验分布 $p(Y|g(X))$ 和 $q(Y|g(X))$ 的差异有以下定义

$$e_g(q,p|\pi_D)=\mathbb{E}_{\pi_D(g(x))}[|q(Y|g(x))-p(Y|g(x))|] \tag{1}$$

其中: π_D 为给定领域 D 后 $g(X)$ 的分布。文献[2]的定理 1 表明，定义式(1)有以下风险上界。

定理 1　领域自适应的风险上界。若给定领域 T 后，经验后验分布 $q(Y|g(X))$ 和目标领域后验分布 $p_T(Y|g(X))$ 的差异被定义为 $e_g(q,p_T|\pi_T)$，则同理定义 $e_g(q,p_S|\pi_S)$、$e_g(q,p_S|\pi_T)$ 和 $e_g(q,p_T|\pi_S)$，并存在以下不等式

$$\begin{aligned}e_g(q,p_T|\pi_T)\le e_g(q,p_S|\pi_S)+\min\{&\\ &e_g(p_S,p_T|\pi_T)+e_g(q,p_S||\pi_T-\pi_S|),\\ &e_g(p_S,p_T|\pi_S)+e_g(q,p_T||\pi_T-\pi_S|)\}\end{aligned} \tag{2}$$

证明　$e_g(q,p_T|\pi_T)$ 加上并减掉 $e_g(q,p_S|\pi_S)$ 和 $e_g(q,p_S|\pi_T)$ 后，有

$$\begin{aligned}e_g(q,p_T|\pi_T)\le e_g(q,p_S|\pi_S)+&\left|e_g(q,p_T|\pi_T)-e_g(q,p_S|\pi_T)\right|\\ +&\left|e_g(q,p_S|\pi_T)-e_g(q,p_S|\pi_S)\right|\end{aligned} \tag{3}$$

由三角不等式知，式(3)的第一个绝对式有不等式

$$\left|e_g(q,p_T|\pi_T)-e_g(q,p_S|\pi_T)\right|\le e_g(p_S,p_T|\pi_T) \tag{4}$$

由基本不等式知，式(3)的第二个绝对式有不等式

$$\left|e_g(q,p_S|\pi_T)-e_g(q,p_S|\pi_S)\right|\le e_g(q,p_S||\pi_T-\pi_S|) \tag{5}$$

结合式(3)~(5)后，有不等式

$$e_g(q,p_T|\pi_T)\le e_g(q,p_S|\pi_S)+e_g(p_S,p_T|\pi_T)+e_g(q,p_S||\pi_T-\pi_S|) \tag{6}$$

同理，$e_g(q,p_T|\pi_T)$ 加上并减掉 $e_g(q,p_S|\pi_S)$ 和 $e_g(q,p_T|\pi_S)$ 后，有

$$e_g(q,p_T|\pi_T)\le e_g(q,p_S|\pi_S)+e_g(p_S,p_T|\pi_S)+e_g(q,p_T||\pi_T-\pi_S|) \tag{7}$$

结合式(6)(7)后，有定理 1 中的式(2)。

根据定理 1，如果 Δ 代表式(2)中的 $\min\{\cdot,\cdot\}$，那么有以下风险上界三角形，见图 1。

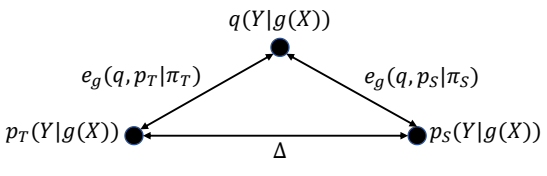


图 1　风险上界三角形

Fig. 1　The risk triangle of upper bound

由图 1 可知，为了使经验后验分布 $q(Y|g(X))$ 能近似表达目标领域后验分布 $p_T(Y|g(X))$，算法应该最小化源领域后验分布 $p_S(Y|g(X))$ 和目标领域后验分布 $p_T(Y|g(X))$ 的差异 Δ。为了使 Δ 足够小，一种有效的优化策略是对齐源领域总体分布 $\pi_S(g(X))$ 和目标领域总体分布 $\pi_T(g(X))$。此外，定理 1 还提供了另一种优化 Δ 的策略，即对齐经验后验分布 $q(Y|g(X))$ 和目标领域后验分布 $p_T(Y|g(X))$。然而，在无监督情形下，目标领域缺乏样本标签，因此这种做法并不可行。总而言之，一个无监督自适应算法可以这样实现：a)使用源领域的标注信息训练有监督模型；b)借助评估函数度量并对齐领域总体分布。

不过，根据 Δ 的组织形式，总体分布对齐策略有可能增大模型与源领域之间的后验分布差异，从而导致 Δ 不降反升。换句话说，总体分布对齐策略有可能导致模型流失分类信息。因此，算法应该对齐两个领域的类内分布。

2　类内最大均值差异

由上一章的结论可知，为了对齐类内分布，无监督自适应算法需要一个分布差异评估函数。由于最大均值差异[12]是目前主流的一种评估函数，所提算法在对齐类内分布时使用最大均值差异。为了便于后续区分，类内最大均值差异被定义为

$$\mathrm{ICMMD}[\Phi,g,\pi_S^c,\pi_T^c]:=\sup_{\phi\in\Phi}\left(\mathbb{E}_{\pi_S^c(g(x))}[\phi(g(x))]-\mathbb{E}_{\pi_T^c(g(x))}[\phi(g(x))]\right) \tag{8}$$

其中: Φ 为一类连续函数，使得特征 $g(x)$ 被投影到实数空间 $\mathbb{R}$。π_S^c 和 π_T^c 为给定类别 c 后源领域和目标领域的特征概率分布。

根据概率论[27]，若 π_S^c 和 π_T^c 都为波莱尔概率分布，那么它们相等的一个充分条件是

$$\left|\int\pi_S^c(g(x))\phi(g(x))dg(x)-\int\pi_T^c(g(x))\phi(g(x))dg(x)\right|=0 \tag{9}$$

由于连续函数 ϕ 和 $-\phi$ 同时存在于空间 Φ 中，定义式(8)是式(9)左边式子的一个上界。因此，类内最大均值差异能够评估类内分布差异。当类内最大均值差异为 0 时，领域类内分布是对齐的。由于定义式(8)要求算法遍历函数空间，算法引入再生核希尔伯特空间以便于计算类内分布差异。根据文献[3]给出的计算方式，再生核希尔伯特空间中的类内最大均值差异有以下平方形式

$$\begin{aligned}\mathrm{ICMMD}^2[\Phi,g,\pi_S^c,\pi_T^c]&:=\langle\mu_S^c-\mu_T^c,\mu_S^c-\mu_T^c\rangle_{\mathcal{H}}\\ &=\langle\mu_S^c,\mu_S^c\rangle_{\mathcal{H}}+\langle\mu_T^c,\mu_T^c\rangle_{\mathcal{H}}-2\langle\mu_S^c,\mu_T^c\rangle_{\mathcal{H}}\\ &=\mathbb{E}_{\pi_S^c(g(x))}\left[\mathbb{E}_{\pi_S^c(g(x'))}\left[\langle\psi(g(x)),\psi(g(x'))\rangle_{\mathcal{H}}\right]\right]\\ &+\mathbb{E}_{\pi_T^c(g(x))}\left[\mathbb{E}_{\pi_T^c(g(x'))}\left[\langle\psi(g(x)),\psi(g(x'))\rangle_{\mathcal{H}}\right]\right]\\ &-2\mathbb{E}_{\pi_S^c(g(x))}\left[\mathbb{E}_{\pi_T^c(g(x'))}\left[\langle\psi(g(x)),\psi(g(x'))\rangle_{\mathcal{H}}\right]\right]\end{aligned} \tag{10}$$

其中: $\mu_S^c=\mathbb{E}_{\pi_S^c(g(x))}[\psi(g(x))]$ 和 $\mu_T^c=\mathbb{E}_{\pi_T^c(g(x))}[\psi(g(x))]$ 分别为 $\psi(g(x))$ 在分布 π_S^c 和 π_T^c 上的期望。函数 ψ 把 $g(x)$ 投影到再生核希尔伯特空间 $\mathcal{H}$ 中。根据式(10)，ICMMD^2 有以下无偏估计

$$\begin{aligned}\mathrm{ICMMD}^2[\Phi,g,\mathbb{X}_S^c,\mathbb{X}_T^c]=&\frac{1}{m_c(m_c-1)}\sum_{i\ne j}^{m_c}k(g(x_i^c),g(x_j^c))\\ &+\frac{1}{n_c(n_c-1)}\sum_{i\ne j}^{n_c}k(g(\tilde{x}_i^c),g(\tilde{x}_j^c))\\ &-\frac{2}{m_cn_c}\sum_{i,j=1}^{m_c,n_c}k(g(x_i^c),g(\tilde{x}_j^c))\end{aligned} \tag{11}$$

其中: $\mathbb{X}_S^c=\{x_1^c,\cdots,x_{m_c}^c\}$ 和 $\mathbb{X}_T^c=\{\tilde{x}_1^c,\cdots,\tilde{x}_{m_c}^c\}$ 为给定类别 c 后源领域和目标领域的样本子集。m_c 和 n_c 分别代表 $\mathbb{X}_S^c$ 和 $\mathbb{X}_T^c$ 中的样本个数。k 为任意一种核函数。

由式(11)可知，算法没有必要遍历整个函数空间。算法只要把两个领域的同类样本映射到再生核希尔伯特空间中，

即可通过式(11)评估两个领域的类内分布差异。至此，只要把两个领域的同类样本分别代入到式(11)中的 $\mathbb{X}_S^c$ 和 $\mathbb{X}_T^c$，那么算法就能够对齐两个领域的类内分布。

3　具体模型及算法

综上，基于 ICMMD 的无监督自适应算法有目标函数

$$\arg\min_{f\in\mathcal{F},g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(f(g(x_i)),y_i)+\lambda\ \mathrm{ICMMD}^2[\Phi,g,\mathbb{X}_S^c,\mathbb{X}_T^c] \tag{12}$$

其中，函数 $\mathcal{L}$ 代表 softmax 损失函数，超参 λ 代表 ICMMD 正则项的权重。$f(g(x_i))$ 预测 x_i 的标签，其作用等同于

$$\arg\max_{c\in C} q(Y=c\,|\,Z=g(x_i))$$

式(12)要求算法解决一个多任务学习问题。主要任务是最小化分类误差，从而拟合源领域的后验分布。次要任务是最小化 ICMMD 正则项，从而对齐两个领域的类内分布。由于 ICMMD 正则项只对齐某一类的分布，因此若类别空间被定义为 C，那么对于每个迭代 I，算法优化以下 ICMMD 正则项。

$$\mathrm{ICMMD}^2\left[\Phi,g,\mathbb{X}_S^{(I\bmod|C|)},\mathbb{X}_T^{(I\bmod|C|)}\right]$$

考虑到无监督自适应任务中，目标领域缺失标注信息，因此这些无标签样本被模型赋予伪标签。同时考虑到领域专用信息对自适应模型的干扰，源领域样本也被模型赋予伪标签，从而减少不具备迁移能力的分类信息。因此，ICMMD 正则项只被允许访问样本的伪标签。

然而，收敛不充分的模型有可能导致伪标签样本数量不足以对齐类内分布，因此用于 ICMMD 正则项的样本是通过采样得到的。不过，在样本数量过于稀少时，采样的效果会变得很差。为此，阈值 τ 和 α 被用于调整算法的采样模式。具体来说，当源领域和目标领域都至少存在 τ 个带有伪标签 c 的样本时，算法才会对 $\mathbb{X}_S^c$ 和 $\mathbb{X}_T^c$ 执行 α 次采样操作。否则，算法直接从源领域数据集 $\mathbb{X}_S=\{x_1,\ldots,x_m\}$ 和目标领域数据集 $\mathbb{X}_T=\{\tilde{x}_1,\ldots,\tilde{x}_n\}$ 中分别采样一个数量为 α 的样本集。最终，无论选择何种采样模式，算法都会产生两个新的样本集 $\bar{\mathbb{X}}_S^c$ 和 $\bar{\mathbb{X}}_T^c$。此外，收敛不充分的模型还有可能导致 ICMMD 正则项输出十分大的评估值，并导致模型更新过于迅猛而不收敛。为此，ICMMD 正则项的评估值应做缩放处理。这要求算法对以下缩放率进行计算。

$$r=\sqrt{\frac{2\alpha(2\alpha-1)}{\sum_{x,x'\in\bar{\mathbb{X}}_S^c\cup\bar{\mathbb{X}}_T^c}\|g(x)-g(x')\|_2^2}}$$

最终，式(12)的目标函数被修改成以下形式

$$\arg\min_{f\in\mathcal{F},g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(f(g(x_i)),y_i)+\lambda\ \mathrm{ICMMD}^2\left[\Phi,rg,\bar{\mathbb{X}}_S^c,\bar{\mathbb{X}}_T^c\right] \tag{13}$$

由于式(13)的目标函数具有非凸性质，且要求 f 和 g 的表征能力足够强，使用神经网络对目标函数进行建模是一种自然的选择。基于目标函数的组织形式，网络结构按图 2 所示设计。模型的优化过程主要分为四个阶段。第一阶段评估源领域和目标领域的样本伪标签。第二阶段根据伪标签信息采样得到 $\bar{\mathbb{X}}_S^c$ 和 $\bar{\mathbb{X}}_T^c$。第三阶段计算源领域分类误差和领域类内分布差异。第四阶段使用随机梯度下降算法更新 f 和 g 的参数 θ_f 和 θ_g。其中，学习率 η 决定这些参数的优化程度。

综上所述，基于 ICMMD 的无监督自适应算法有以下流程：

对于每个迭代 I：

$c\leftarrow I \bmod |C|$。

$\mathbb{X}_S^c\leftarrow\{\}$，$m_c\leftarrow 0$。

$\mathbb{X}_T^c\leftarrow\{\}$，$n_c\leftarrow 0$。

对于每个样本 $x\in\mathbb{X}_S$：

如果 $f(g(x))$ 输出的伪标签为 c，

$\mathbb{X}_S^c\leftarrow\mathbb{X}_S^c\cup\{x\}$，$m_c\leftarrow m_c+1$。

对于每个样本 $\tilde{x}\in\mathbb{X}_T$：

如果 $f(g(\tilde{x}))$ 输出的伪标签为 c，

$\mathbb{X}_T^c\leftarrow\mathbb{X}_T^c\cup\{\tilde{x}\}$，$n_c\leftarrow n_c+1$。

如果 $m_c\geq\tau$ 且 $n_c\geq\tau$

对 $\mathbb{X}_S^c$ 采样 α 次，得到 $\bar{\mathbb{X}}_S^c$，$m_c\leftarrow\alpha$。

对 $\mathbb{X}_T^c$ 采样 α 次，得到 $\bar{\mathbb{X}}_T^c$，$n_c\leftarrow\alpha$。

否则分别从 $\mathbb{X}_S$ 和 $\mathbb{X}_T$ 中采样 α 次，得到 $\bar{\mathbb{X}}_S^c$ 和 $\bar{\mathbb{X}}_T^c$。

计算缩放率 r 并用随机梯度下降算法更新 θ_g 和 θ_f：

$$\theta_f\leftarrow\theta_f-\eta\frac{\partial\mathcal{L}}{\partial\theta_f}$$

$$\theta_g\leftarrow\theta_g-\eta\left(\frac{\partial\mathcal{L}}{\partial\theta_g}+\lambda\frac{\partial\ \mathrm{ICMMD}^2\left[\Phi,rg,\bar{\mathbb{X}}_S^c,\bar{\mathbb{X}}_T^c\right]}{\partial\theta_g}\right)$$

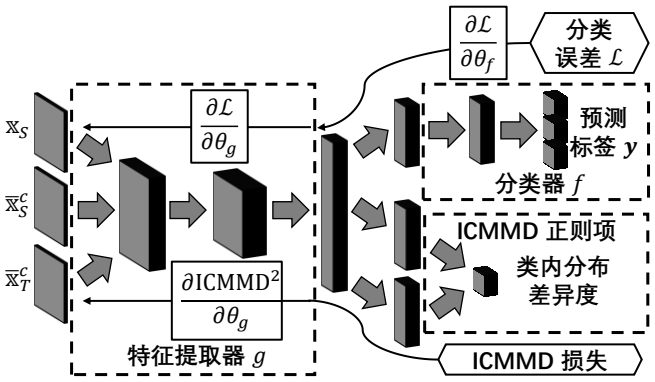


图 2　神经网络结构

Fig. 2　neural network structure

4　实验设计及结果分析

为了对比实验的公平性，所有算法被部署到 Caffe 框架[28]中，并且它们的性能指标统一采用分类精度(Accuracy)。为了发挥框架的性能，所有算法运行在一台高性能服务器上。该服务器搭载 Intel Xeon E5 中央处理器和 Nvidia Tesla K80 显卡。

4.1　OFFICE-31 实验

OFFICE-31 图片数据集 (https://pan.baidu.com/s/1o8igXT4#list/path=%2F)最先被文献[29]使用。它包含三个领域数据集，分别为 AMAZON、DSLR 和 WEBCAM。这三个领域分别包含 2817、498 和 795 张图片，且都由 31 种图片组成。

在本组实验中，DDC[11]、DAN[13]、GRL[22]和 DRCN[17]这四个主流算法被选为实验的对比算法。其中，DDC 和 DAN 都采用原始 MMD 对齐两个领域的总体分布。后者与前者的不同之处在于 MMD 被换成多核 MMD 并嵌入到多个全连接层中。GRL 通过领域混淆器寻找领域无关的特征空间，DRCN 通过自动编码器寻找类别有关的低维特征空间。

根据该数据集的组成形式，所有算法在六种领域组合下评估性能。这些领域组合分别为‘A→W’、‘W→A’、‘A→D’、‘D→A’、‘W→D’和‘D→W’。其中，‘→’两边的字母分别代表源领域和目标领域。此外，考虑到该数据集的图片分辨率较高，所有算法采用 AlexNet[30]作为网络骨架。具体网络结构如图 3 所示，它包含 3 个卷积层、3 个最大池化层和 3 个全连接层，并且每个卷积层紧跟一个最大池化层。

由于 OFFICE-31 数据集有 31 种图片，输出层的神经元数量被修改为 31。

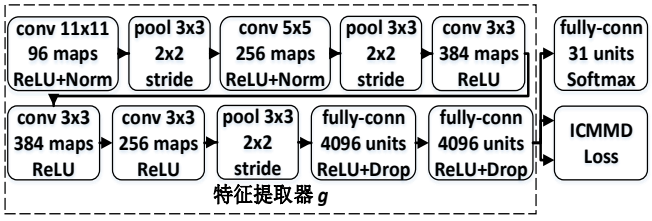


图 3　OFFICE-31 实验的具体网络结构

Fig. 3　OFFICE-31 Experiments

对比算法的超参数和正则项均使用原始论文或源代码中所使用的设计。根据之前的讨论，所提算法的正则项被设计成只包含 ICMMD 正则项的子网络。所提算法涉及的超参数有：

a)学习率 η，它控制了模型参数的优化程度。考虑到数据集的噪声较多，为了保证模型能够稳定收敛，所提算法使用较小的学习率并借助以下退火策略惩罚学习率。

$$\eta = 0.001*(1+0.001*I)^{-0.75}$$

其中 I 为算法当前迭代周期。

b)迭代周期数，它决定了算法对模型参数的优化次数。考虑到较小的学习率会导致模型收敛较慢，因此算法采用了较大的迭代周期数 50000，即 I 的最大值被设定为 50000。按照这个设定，所提算法执行 50000 次随机梯度下降算法，并随之退出。

c)正则项权重参数 λ，它控制正则项对目标函数的影响程度。考虑到 ICMMD 正则项已被设计成自动缩放形式，该参数被设置为 1.0。

d)阈值 τ 和 α，它们共同调节算法的采样模式。它们分别被设置为 5 和 64。也就是说，当两个领域的同类伪标签样本数量都达到 5 时，算法对该类伪标签样本执行 64 次采样操作。

e)核函数 k，它决定了算法对分布差异的估算方式。在本次实验中，由于高斯核函数能够估算无限维分布差异，因此 k 被设置为高斯核函数，即

$$k(x,x') = \exp\left(-\|x-x'\|_2^2\right)$$

其中: x 和 x' 为任意两个不相同的数据样本。

综上所述，将 DDC、DAN、GRL、DRCN 和所提算法进行比较，实验结果如表 1 所示。

表 1　OFFICE-31 上的算法精度比较

Table 1　Comparison of Algorithm Accuracies on OFFICE-31

算法	领域组合						Avg
	A→W	W→A	A→D	D→A	W→D	D→W	
DDC	0.618	0.522	0.644	0.521	0.985	0.950	0.707
DAN	0.685	0.531	0.670	0.540	0.990	0.960	0.729
GRL	0.726	0.527	0.671	0.545	0.992	0.964	0.738
DRCN	0.687	0.549	0.668	0.558	0.990	0.964	0.736
ICMMD	**0.731**	**0.551**	**0.733**	**0.585**	**0.998**	**0.968**	**0.761**

由表 1 知，无论在哪种领域组合中，ICMMD 的性能表现都优于主流算法。得益于 ICMMD 正则项的设计，所提算法在对齐类内分布时有效避免了分类信息的流失。受制于总体分布对齐策略的缺陷，其余算法总是出现程度不一的分类信息流失问题，从而使它们的平均精度都不同程度地低于 ICMMD。显然，总体分布对齐策略会降低模型对目标领域的分类性能。

值得注意的是，在‘W→D’和‘D→W’场景中，所有算法都因为相似的领域总体分布而获得了较高的精度。同样值得注意的是，DRCN 和 GRL 分别在‘A→W’和‘W→A’中表现不及 ICMMD，但是 DRCN 和 GRL 分别在‘W→A’和‘A→W’中表现逼近 ICMMD。这说明总体分布对齐策略未必造成严重的分类信息流失问题，而且不同算法在不同场景中流失的分类信息是不一样的。

另外，无论在哪种领域组合中，DAN 都比 DDC 有更高的精度。这说明多核 MMD 可以有效保留领域共有的分类信息。因此，ICMMD 也能改造成多核版本，进一步提升算法性能。

4.2　MNIST-USPS 实验

为了验证在其他数据集上的效果，本次实验采用 MNIST 和 USPS 这两个数据集(https://pan.baidu.com/s/1c8mwdo#list/path=%2F)。MNIST 和 USPS 数据集分别被首次使用在文献[31,32]中。这两个数据集都包含了 10 种手写数字图片，其中 MNIST 的训练集和测试集分别有 60000 和 10000 张图片。USPS 的训练集和测试集分别有 7438 和 1860 张图片。

在本组实验中，GRL[22]、DRCN[17]、CoGAN[25]和 ADDA[26]这四个主流算法被选为实验的对比算法。其中，GRL 和 DRCN 的设计原理已经在之前的文字中讨论过。CoGAN 能够通过共享生成器权重对齐两个领域的特征分布。ADDA 能够通过对抗模型或判别模型抑制生成器输出领域专用信息。

根据该数据集的组成形式，所有算法的性能评估在两种领域组合下完成。这两种领域组合分别为‘M→U’和‘U→M’。在‘M→U’领域组合下，所有算法使用 MNIST 的有标签训练集和 USPS 的无标签训练集作为模型的训练集，并使用 USPS 的有标签测试集作为模型的测试集。在‘U→M’领域组合下，所有算法使用 USPS 的有标签训练集和 MNIST 的无标签训练集作为模型的训练集，并使用 MNIST 的有标签测试集作为模型的测试集。此外，考虑到该数据集的图片分辨率较小，所有算法采用一个较小的网络骨架，具体细节如图 4 所示。它包含 2 个卷积层、2 个最大池化层和 2 个全连接层，并且每个卷积层紧跟一个最大池化层。由于 MNIST 和 USPS 数据集都有 10 个类别的手写数字图片，输出层的神经元数量被设定为 10。

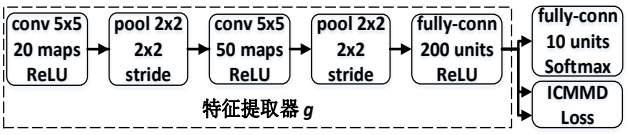


图 4　MNIST-USPS 实验的具体网络结构

Fig. 4　Concrete Network Structure for MNIST-USPS Experiments

在本次实验中，所提算法的超参数设定大致和 OFFICE-31 实验相同。不同设定的超参数有：1)学习率 η，考虑到数据集的规模比较大但分布比较简单，为了保证模型能够快速收敛，所提算法使用较大的学习率并借助以下退火策略惩罚学习率。

$$\eta = 0.01*(1+0.001*I)^{-0.8}$$

其中 I 为算法当前迭代周期。2)迭代周期数，考虑到本次实验所设定的学习率较大，因此为了节省模型的训练时间，所提算法采用了一个较小的迭代周期数 5000，即 I 的最大值被设定为 5000。按照规定，算法执行 5000 次随机梯度下降算法，并随之退出。3)阈值 τ 和 α，它们共同调节算法的采样模式。考虑到数据集的规模较大，但图片的分辨率较小，它们分别被设置为 200 和 800。也就是说，当两个领域的同类

伪标签样本数量都达到 200 时，算法对该类伪标签样本执行 800 次采样操作。

综上所述，将 GRL、DRCN、CoGAN、ADDA 和所提算法进行比较，实验结果如表 2 所示。

由表 2 可知，在 MNIST-USPS 数据集上，ICMMD 的性能优于主流算法。这个结果和 OFFICE-31 的实验结果是一致的。这一定程度上表明类内最大均值差异技术适用于所有数据集。

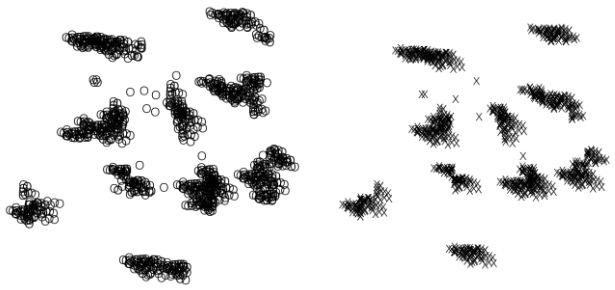
表 2　MNIST-USPS 上的算法精度比较

Table 2　Comparison of Algorithm Accuracies on MNIST-USPS

算法	领域组合		Avg
	M→U	U→M	
GRL	0.913	0.740	0.827
DRCN	0.918	0.737	0.828
CoGAN	0.912	0.891	0.902
ADDA	0.894	0.901	0.898
ICMMD	**0.975**	**0.932**	**0.954**

值得注意的是，CoGAN 和 ADDA 在‘M→U’中的性能略差于 GRL 和 DRCN，但是 CoGAN 和 ADDA 的平均性能远高于 GRL 和 DRCN。这表明对抗网络能够更好地保留源领域的分类信息。此外，所有算法在‘U→M’中的性能都不同程度地低于在‘M→U’中的性能。这是因为 USPS 数据集规模远远小于 MNIST 数据集规模，前者提供的分类信息没有后者的分类信息充足。然而，即使这样，ICMMD 依然很好地把源领域的分类信息迁移到目标领域中。

为了展示所提算法的有效性，特征提取器 g 输出的所有特征会被 t-SNE[33](t-distributed Stochastic Neighbor Embedding)算法处理成可视化图像。以‘M→U’为例，所有特征的可视化结果如图 5 所示。其中圆圈点代表 MNIST 的特征，交叉点代表 USPS 的特征。



(a) MNIST 的特征可视化　　(b) USPS 的特征可视化

图 5　特征可视化

Fig. 5　hidden feature visualization

由图 5 知，不同领域的同类样本聚集成一块，而且散落的位置也大致相同。这说明所提算法在对齐类内分布时有效保留了分类信息。

值得注意的是，在图 15(a)中，MNIST 的特征散落范围更大，而且 MNIST 的孤立点也更多。这表明 MNIST 的类内分布比 USPS 的类内分布更复杂。此外，两个子图还表明，MNIST 的孤立点分布和 USPS 的孤立点分布是不同的，这表明两个领域都存在领域专用信息。为了避免这些信息的干扰，ICMMD 正则项在评估类内分布时使用伪标签样本。

5　结束语

本文提出了一种基于类内最大均值差异的无监督领域自适应算法，解决了传统算法的分类信息流失问题。通过预测样本的伪标签，算法能够针对不同领域的同类伪标签样本进行分布对齐，有效保留源领域中具备迁移能力的分类信息。实验结果表明所提算法的性能优于传统算法，验证了本算法设计的有效性。下一步工作中，算法可借鉴 DAN 的多核思路，进一步提升所提算法的性能。

参考文献：

[1] Pan S J, Yang Qiang. A survey on transfer learning [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22 (10): 1345-1359.

[2] Ben-David S, Blitzer J, Crammer K, *et al.* A theory of learning from different domains [J]. Machine Learning, 2010, 79 (1-2): 151-175.

[3] Gong L, Jiang Shujuan, Yu Qiao, *et al.* Unsupervised deep domain adaptation for heterogeneous defect prediction [J]. IEICE Trans on Information and Systems, 2019, 102 (3): 537-549.

[4] Cortes C, Mohri M, Medina A M. Adaptation based on generalized discrepancy [J]. Journal of Machine Learning Research, 2019, 20 (1): 1-30.

[5] Xu Xinzheng, Liang Tianming, Zhu Jiong, *et al.* Review of classical dimensionality reduction and sample selection methods for large-scale data processing [J]. Neurocomputing, 2019, 328 (1): 5-15.

[6] Song Peng, Zheng Wenming. Feature selection based transfer subspace learning for speech emotion recognition [J]. IEEE Trans on Affective Computing, 2018, 2018 (1): 1.

[7] Mourragui S, Loog M, Reinders M J T, *et al.* PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors [J]. bioRxiv, 2019: 536797.

[8] Yan Ke, Kou Lu, Zhang D. Learning domain-invariant subspace using domain features and independence maximization [J]. IEEE Trans on Cybernetics, 2018, 48 (1): 288-299.

[9] 张春荣, 赵琦. 领域自适应的合成词词性标注研究 [J]. 计算机应用研究, 2018, 35 (5): 1350-1354. (Zhang Chunrong, Zhao Qi. Research on domain-adaptive POS tagging for compound words [J]. Application Research of Computers, 2018, 35 (5): 1350-1354.)

[10] Liu Ji, Zhang Lei. Optimal Projection Guided Transfer Hashing for Image Retrieval [J]. CoRR, 2019, abs/1903. 00252: 1.

[11] Tzeng E, Hoffman J, Zhang N, *et al.* Deep domain confusion: Maximizing for domain invariance [J]. CoRR, 2014, abs/1412. 3474: 1.

[12] Borgwardt K M, Gretton A, Rasch M J, *et al.* Integrating structured biological data by kernel maximum mean discrepancy [J]. Bioinformatics, 2006, 22 (14): e49-e57.

[13] Long Mingsheng, Cao Yue, Wang Jianmin, *et al.* Learning transferable features with deep adaptation networks [C]// Proc of the 32nd International Conference on Machine Learning. New York: PMLR Press, 2015: 97-105.

[14] Long Mingsheng, Zhu Han, Wang Jianmin, *et al.* Deep transfer learning with joint adaptation networks [C]// Proc of the 34th International Conference on Machine Learning. New York: PMLR Press, 2017: 2208-2217.

[15] Long Mingsheng, Cao Yue, Cao Zhangjie, *et al.* Transferable representation learning with deep adaptation networks [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2018 (9): 1.

[16] Gretton A, Sejdinovic D, Strathmann H, *et al.* Optimal kernel choice for large-scale two-sample tests [C]// Proc of the 25th Annual Conference on Neural Information Processing Systems. New York: Curran Associates, 2012: 1205-1213.

[17] Ghifary M, Kleijn W B, Zhang Mengjie, *et al.* Deep reconstruction-classification networks for unsupervised domain

adaptation [C]// Proc of the 14th European Conference on Computer Vision. Cham: Springer, 2016: 597-613.

[18] Lin Weiwei, Mak M W, Li Longxin, *et al*. Reducing domain mismatch by maximum mean discrepancy based autoencoders [C]// Proc of the 11th Odyssey Speaker and Language Recognition Workshop. [S. l.] : ISCA, 2018: 162-167.

[19] 丁亮，何彦青. 融合领域知识与深度学习的机器翻译领域自适应研究 [J]. 情报科学, 2017, 35 (10): 125-132. (Ding Liang, He Yanqing. Research on Domain Adaption in Machine Translation Combining Domain Knowledge and Deep Learning [J]. Information Science, 2017, 35 (10): 125-132.)

[20] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks [C]// Proc of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-6649.

[21] Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets [C]// Proc of the 27th Annual Conference on Neural Information Processing Systems. New York: Curran Associates, 2014: 2672-2680.

[22] Ganin Y, Lempitsky V. Unsupervised Domain Adaptation by Backpropagation [C]// Proc of the 32nd International Conference on Machine Learning. New York: PMLR Press, 2015: 1180-1189.

[23] Caruana R. Multitask learning [J]. Machine learning, 1997, 28 (1): 41-75.

[24] Saito K, Watanabe K, Ushiku Y, *et al*. Maximum classifier discrepancy for unsupervised domain adaptation [C]// Proc of the 36th IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2018: 3723-3732.

[25] Liu Mingyu, Tuzel O. Coupled generative adversarial networks [C]// Proc of the 29th Annual Conference on Neural Information Processing Systems. New York: Curran Associates, 2016: 469-477.

[26] Tzeng E, Hoffman J, Saenko K, *et al*. Adversarial discriminative domain adaptation [C]// Proc of the 35th IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 7167-7176.

[27] Dudley R M. Real analysis and probability [M]. London: Chapman and Hall/CRC, 2018: 292-293.

[28] Jia Yangqing, Shelhamer E, Donahue J, *et al*. Caffe: Convolutional architecture for fast feature embedding [C]// Proc of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 675-678.

[29] Saenko K, Kulis B, Fritz M, *et al*. Adapting visual category models to new domains [C]// Proc of the 11th European Conference on Computer Vision. Berlin: Springer, 2010: 213-226.

[30] Krzhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Proc of the 25th Annual Conference on Neural Information Processing Systems. New York: Curran Associates, 2012: 1097-1105.

[31] LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[32] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning [M]. 2nd ed. New York: Springer Series in Statistics, 2001.

[33] Maaten L, Hinton G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9 (11): 2579-26.