智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于分类差异与信息熵对抗的无监督域适应算法

李庆勇， 何军， 张春晓

引用本文:
李庆勇, 何军, 张春晓. 基于分类差异与信息熵对抗的无监督域适应算法[J]. 智能系统学报, 2021, 16(6): 999–1006.
LI Qingyong, HE Jun, ZHANG Chunxiao. Unsupervised domain adaptation algorithm based on classification discrepancy and information entropy[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(6): 999–1006.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202010020

DOI: 10.11992/tis.202010020
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210831.1640.008.html

基于分类差异与信息熵对抗的无监督域适应算法

李庆勇[1], 何军[1,2], 张春晓[1]
（1. 南京信息工程大学 电子与信息工程学院, 江苏 南京 210044; 2. 南京信息工程大学 人工智能学院, 江苏 南京 210044）

摘　要: 采用对抗训练的方式成为域适应算法的主流, 通过域分类器将源域和目标域的特征分布对齐, 减小不同域之间的特征分布差异。但是, 现有的域适应方法仅将不同域数据之间的距离缩小, 而没有考虑目标域数据分布与决策边界之间的关系, 这会降低目标域内不同类别的特征的域内可区分性。针对现有方法的缺点, 提出一种基于分类差异与信息熵对抗的无监督域适应算法 (adversarial training on classification discrepancy and information entropy for unsupervised domain adaptation, ACDIE)。该算法利用两个分类器之间的不一致性对齐域间差异, 同时利用最小化信息熵的方式降低不确定性, 使目标域特征远离决策边界, 提高了不同类别的可区分性。在数字标识数据集和 Office-31 数据集上的实验结果表明, ACDIE 算法可以学习到更优的特征表示, 域适应分类准确率有明显提高。

关键词: 域适应; 对抗训练; 神经网络; 无监督学习; 迁移学习; 分类差异; 信息熵; 决策边界
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2021)06−0999−08

中文引用格式: 李庆勇, 何军, 张春晓. 基于分类差异与信息熵对抗的无监督域适应算法 [J]. 智能系统学报, 2021, 16(6): 999–1006.
英文引用格式: LI Qingyong, HE Jun, ZHANG Chunxiao. Unsupervised domain adaptation algorithm based on classification discrepancy and information entropy[J]. CAAI transactions on intelligent systems, 2021, 16(6): 999–1006.

Unsupervised domain adaptation algorithm based on classification discrepancy and information entropy

LI Qingyong[1], HE Jun[1,2], ZHANG Chunxiao[1]
(1. School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: The adversarial training method has become the mainstream of the domain adaptation algorithm. The feature distributions of the source and target domains are aligned by a domain classifier to reduce the feature distribution discrepancy among different domains. However, existing domain adaptation methods only reduce the distance between different domain data without considering the relationship between the data distribution of the target domain and decision boundaries, thus decreasing the intradomain distinguishability of different categories in the target domain. Considering the shortcomings of the existing methods, an unsupervised domain adaptation algorithm based on classification discrepancy and information entropy confrontation (ACDIE) is proposed in this study. The algorithm uses the discrepancy and the domain aligning discrepancy between two classifiers and minimizes the information entropy to reduce uncertainty. Consequently, the proposed method makes the target domain feature far away from the decision boundaries and improves the distinguishability of different categories. The experimental results of the digital identification and Office-31 datasets show that the ACDIE algorithm can learn better feature representation. Moreover, the accuracy of the domain adaptation classification is considerably improved.

Keywords: domain adaptation; confrontation training; neural network; unsupervised learning; transfer learning; classification discrepancy; information entropy; decision boundary

在大数据时代的背景下, 数据呈爆炸式增长, 但大部分数据缺失有效的标注信息。由于数据标注任务的成本较高, 通过无监督学习的方式进行

收稿日期: 2020−10−19.　**网络出版日期**: 2021−09−01.
基金项目: 国家自然科学基金项目 (61601230).
通信作者: 何军. E-mail: jhe@nuist.edu.cn.

模型训练可以大大减少投入的人力、物力和时间成本，所以无监督学习成为机器学习领域一个重要的研究方向[1-2]。其次，传统机器学习算法中存在用训练集数据进行训练得到的模型无法适应现实场景的问题，这是由训练集数据与实际测试数据的特征分布不同导致的[3]。

针对以上问题，迁移学习 (transfer learning, TL) 方法被提出[4]，域适应学习 (domain adaptation learning, DAL) 作为一种同构迁移学习方法[5]，在源域与目标域样本特征分布不同但相似的前提下，将源域样本分类模型迁移到目标域，使模型适应目标域数据。无监督域适应模型通过带标签源域数据和无标签目标域数据进行训练，即使训练过程中不包含目标域标注信息，也可以在目标域数据中实现很好的识别效果。

Ghifary 等[6] 利用传统 DAL 思想，使用自编码器学习共享编码以获得域不变特征，实现在特征向量空间中，不同域样本特征之间的距离减小的目的，从而使无标签目标域样本得到正确分类。Sener 等[7] 提出利用聚类和伪标签的方法来获取分类特征，从而实现在无标签目标域上的分类。卷积神经网络中间特征的分布匹配被认为是实现域适应的有效方法[8]。最大均值差异 (maximum mean discrepancy, MMD)[9] 使用核函数映射特征来度量两不同分布之间的距离，通过最小化源域与目标域之间的距离得到域共享特征。Tzeng 等[10] 在分类损失的基础上加了一层适配层，通过在适配层上引入 MMD 距离来度量最小化两个领域的分布差异。Long 等[11-12] 在 MMD 方法的基础上改进，采用多层适配和多核 MMD 使域差异最小化，实现源域和目标域特征具有相似的特征分布。借鉴生成对抗网络 (generative adversarial network, GAN)[13] 独特的对抗训练方式，Ganin 等[14] 提出包含特征生成器和域分类器结构的模型 DANN，利用特征生成器生成欺骗域分类器的特征，从而将源域和目标域数据映射到相似的概率分布上。王格格等[15] 通过联合使用生成对抗网络和多核最大均值差异度量准则优化域间差异，以学习源域分布和目标域分布之间的共享特征。Sankaranarayanan 等[16] 提出了一个能够直接学习联合特征空间的对抗图像生成的无监督域适应方法 GTA，利用图像生成的对抗过程学习一个源域和目标域特征分布最小化的特征空间。但由于上述使用 GAN 或 MMD 的分布对齐方法仅将不同域之间的距离拉近，没有考虑目标样本与决策边界之间的关系，因此无法优化域内类间差异，从而影响域适应分类效果。Saito 等[17] 通过训练两个分类器以最大化分类差异，但其方法只是减少源域和目标域之间的距离，而未增大目标域不同类之间的距离，这会使目标域样本靠近决策边界，使分类不确定性增加。

为此，本文提出一种基于分类差异和信息熵对抗的无监督域适应模型。利用两个分类器之间的不一致性对齐域间差异，使源域和目标域数据之间的距离最小，同时利用最小化熵的方式降低不确定性，使目标域特征远离决策边界，提高了目标域样本的类间差异。

1　分类差异和信息熵对抗

假设给定带标签的源域数据集 $D_s=\{X_s,Y_s\}$，源域图像 x_s 对应标签为 y_s，同时给定无标签目标域数据集 $D_t=\{X_t\}$，目标域图像为 x_t。本文模型包括特征生成网络 G 和分类器网络 F_1、F_2，G 网络接收图像 x_s 或 x_t 的输入，经过特征提取输出特征向量 $\boldsymbol{f}$，分类器 F_1 和 F_2 将特征向量分为 K 类，即输出 K 维向量，对向量应用 Softmax 函数得到类别概率。本文使用符号 $p_1(y|x)$、$p_2(y|x)$ 来分别表示由 F_1 和 F_2 获得的输入图像 x 的 K 维概率输出。

相比于其他域适应算法，本文算法在最小化域间差异的同时，可以使目标域内不同类别样本之间的差异最大化。如图 1 所示，对于目标域数据，其他方法因为仅对齐域间差异，缩小源域和目标域数据之间的距离，所以特征生成器会在分类边界附近生成模糊特征。本文模型方法利用对抗训练思想，最小化源域与目标域数据之间的距离，同时使目标域不同类别远离分类边界，获得更加具有区分性的特征，从而提高域适应分类的准确率。

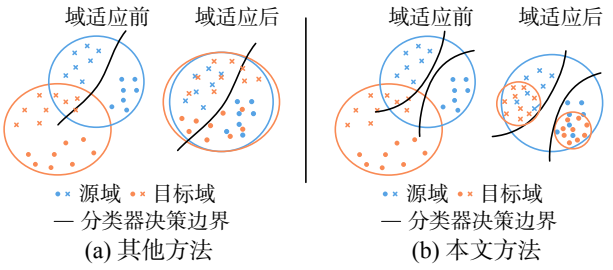


图 1　不同方法特征分布对比

Fig. 1　Comparison of the feature distribution of different methods

1.1　信息熵对抗

分类器的输出为经过 Softmax 函数得到的不同类别概率，根据信息熵的定义，可以得到该分类器结果的信息熵大小，信息熵越大表示不同类

别的概率值越接近, 表明分类边界越模糊, 反之, 信息熵越小, 表明分类边界越清晰。如图 2 所示, 借鉴对抗训练思想、特征生成器最小化信息熵、分类器最大化信息熵, 实现使生成的特征向量 $\boldsymbol{f}$ 远离分类边界的目的, 其中不同形状的标志点代表不同类别的样本。

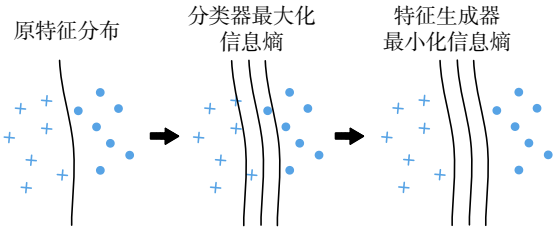


图 2　信息熵对抗过程

Fig. 2　Information entropy confrontation process

1.2　算法分析

本文算法的目标是利用特定任务的分类器作为判别器来减小源域和目标域特征的距离, 以考虑类边界和目标样本之间的关系。为实现这个目标, 必须检测到靠近分类边界的目标域样本, 本文算法利用了两种分类器在目标样本预测上的不一致性。由于源域数据带标签, 所以分类器可以对源域样本正确分类, 两分类器 F_1 和 F_2 的初始化不同必然使决策边界不同。如图 3 所示, 处于阴影处的目标域样本会被错误分类, 如果能够测量两个分类器分类结果之间的不一致, 并训练生成器使之最小化, 则生成器将避免生成错误分类的目标域特征。同时分类器输出结果 $p_1(y|x)$ 和 $p_2(y|x)$ 的信息熵越小, 表示预测结果越具有确定性, 所以训练生成器使分类结果信息熵最小化, 则特征生成器将生成远离分类器决策边界的更加具有区分性的特征。

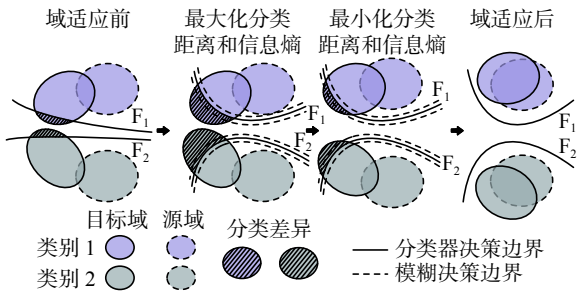


图 3　本文算法特征分布对齐过程

Fig. 3　Alignment process of the feature distribution is presented in this paper

使用距离 $d(p_1(y|x_t), p_2(y|x_t))$ 度量分类器 F_1 和 F_2 之间的差异, 其中 d 表示计算两概率分布散度的函数。根据 Ben-David 等[18] 提出的目标域样本误差限的计算理论, 目标域样本的误差限 $R_T(h)$ 与 3 个因素有关, 包括源域样本误差限 $R_S(h)$、度量分类器差异的 $\mathcal{H}$ 距离和常数 λ, 其中 $\mathcal{H}$ 距离用来度量区分不同域分类器的差异, λ 表示理想假设的共享误差, 通常被认为是一个极小的值。使用 H 表示分类器假设空间, 对于给定的源域 S 和目标域 T, 则:

$$\forall h \in H, R_T(h) \leqslant R_S(h) + \frac{1}{2} d_{\mathcal{H}}(S,T) + \lambda \tag{1}$$

$$d_{\mathcal{H}}(S,T) = 2 \sup_{(h,h') \in H^2} \left| \underset{x \sim S}{E} I[h(x) \neq h'(x)] - \underset{x \sim T}{E} I[h(x) \neq h'(x)] \right| \tag{2}$$

$$\lambda = \min[R_S(h) + R_T(h)] \tag{3}$$

式中: $I[a]$ 是一个二值函数, 当预测 a 正确时函数值为 1, 否则为 0。对于 $d_{\mathcal{H}}(S,T)$, 通过对带标签的源域数据的监督学习, 可以认为预测函数 h 和 h' 可以对源域数据实现很好地分类, 所以 $\underset{x \sim S}{E} I[h(x) \neq h'(x)]$ 部分值极小, 因此可以近似认为:

$$d_{\mathcal{H}}(S,T) = \sup_{(h,h') \in H^2} \underset{x \sim T}{E} I[h(x) \neq h'(x)] \tag{4}$$

式 (4) 表示两个分类器对目标域样本预测差异的极限值。将 h 用特征提取器 G 的函数 $G(x)$ 和分类器 F_1 的函数 F_1 表示, h' 用特征提取器 G 的函数 $G(x)$ 和分类器 F_2 的函数 F_2 表示, 用符号“∘”表示不同网络结构之间输入输出的连接, 则可以得到

$$\sup_{F_1, F_2} \underset{x \sim T}{E} I[F_1 \circ G(x) \neq F_2 \circ G(x)] \tag{5}$$

引入对抗训练的方式, 实现对特征提取器 G 的优化:

$$\min_{G} \max_{F_1, F_2} \underset{x \sim T}{E} I[F_1 \circ G(x) \neq F_2 \circ G(x)] \tag{6}$$

本文算法的目标是获得一个特征生成器, 这个特征生成器可以将目标样本的分类不确定性最小化, 并且可以使目标域样本与源域样本的距离最小化。

1.3　Softmax 交叉熵损失

本文使用 Softmax 交叉熵损失来优化有标注源域数据集上的监督学习分类任务, 通过对源域数据的监督学习可以保证特征生成器在先验特征空间上有合理的构造。Softmax 交叉熵损失定义为

$$L_{\mathrm{cl}}(X_s, Y_s) = -\frac{1}{K} \sum_{i=1}^{K} I(i = y_s^{(i)}) \log p_s(x_s^{(i)}) \tag{7}$$

式中: $I(i = y_s^{(i)})$ 是一个二值函数, 当 i 与 $y_s^{(i)}$ 相等时, 其值为 1, 否则为 0; p_s 是经过映射函数得到的分类概率输出, $p_s = \mathrm{Softmax} \circ F \circ G$。

1.4　分类差异损失

将两个分类器的概率输出之差的绝对值之和定义为分类距离损失:

$$L_d(X_t) = d(p_1(y|x_t), p_2(y|x_t)) = \frac{1}{K} \sum_{k=1}^{K} |p_{1k} - p_{2k}| \tag{8}$$

式中 p_{1k} 和 p_{2k} 分别表示第 k 类 p_1 和 p_2 的概率输出。

1.5　信息熵损失

在目标域中，一个理想的特征向量 $\boldsymbol{f}$ 输入分类器得到的概率输出应该集中于某一类上。由于目标域数据没有标注信息，无法知道样本的类别，因此本文通过最小化信息熵的方法来促使目标域样本分类概率集中于某一类上，使得到的分类结果更加具有确定性。定义熵损失如下：

$$L_{\text{ent}}(X_t)=H(X_t)=\frac{1}{K}\sum_{i=1}^{K}-F(G(x_t^{(i)}))\log F(G(x_t^{(i)})) \tag{9}$$

源域由于有标注信息，其样本的分类概率往往集中在所标注的类别上；而目标域由于存在域间差异，其在分类概率上往往不够集中。训练特征提取器最小化信息熵可以在特征向量层减小源域和目标域的域间差异，即使特征提取器具有更强的泛化能力。

1.6　算法流程

L_{cl1} 和 L_{cl2} 分别表示分类器 F_1 和 F_2 的 Softmax 交叉熵损失，L_{ent1} 和 L_{ent2} 分别表示分类器 F_1 和 F_2 的信息熵损失。输入源域数据集 $D_s=\{X_s,Y_s\}$，目标域数据集 $D_t=\{X_t\}$，批次大小为 m，特征提取器训练次数为 n。ACDIE 模型训练的整体算法流程为：

1) 从 D_s 中采样 m 个有标注数据 $\{x_{si},y_{si}\}_{i=1}^{m}$，记为 $\{X_{sm},Y_{sm}\}$；从 D_t 中采样 m 个无标注数据 $\{x_{ti}\}_{i=1}^{m}$，记为 $\{X_{tm}\}$；

2) 通过有标注数据进行监督训练；

3) 计算损失函数 $L_1=L_{\text{cl1}}+L_{\text{cl2}}$；

4) 反向传播梯度信号，更新 G、F_1 和 F_2 中的参数；

5) 通过无标注数据进行域适应训练；

6) 计算损失函数 $L_2=L_{\text{cl1}}+L_{\text{cl2}}-L_d(X_{tm})-L_{\text{ent1}}(X_{tm})-L_{\text{ent2}}(X_{tm})$；

7) 计算损失函数 $L_3=L_d(X_{tm})+L_{\text{ent1}}(X_{tm})+L_{\text{ent2}}(X_{tm})$；

8) 反向传播梯度信号，更新 G 中的参数；

9) 重复训练步骤 7)~8) n 次。

2　训练步骤

分类器 F_1 和 F_2 接收特征生成器 G 生成的特征向量作为输入，F_1 和 F_2 需要最大化分类距离差异 $d(p_1(y|x_t),p_2(y|x_t))$ 和信息熵 $H(x_t)$，而特征生成器最小化分类距离和信息熵。由此形成特征生成器 G 与分类器 F 的关于分类距离和信息熵的对抗训练。ACDIE 模型训练流程如图 4 所示，ACDIE 模型的训练可以分为以下 3 步。

1) 模型预训练

为了使特征生成器获得特定任务的区分特征，首先通过监督学习的方式训练特征生成器和分类器以正确地对源域样本进行分类。训练网络 G、F_1 和 F_2，以最小化 Softmax 交叉熵优化目标，如式 (10) 所示：

$$\min_{G,F_1,F_2} L_{\text{cl}}(X_s,Y_s) \tag{10}$$

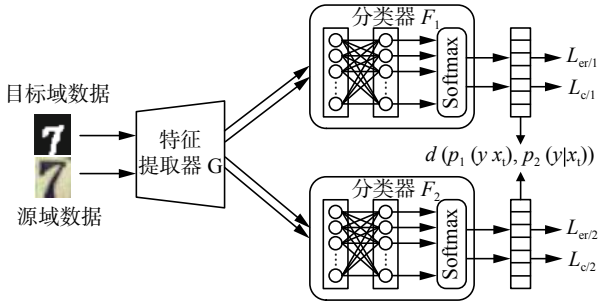


图 4　ACDIE 模型流程

Fig. 4　ACDIE model flow

2) 训练分类器

固定特征生成器 G 的参数，利用目标域数据训练分类器 F_1 和 F_2，使分类概率输出的差异增大，同时最大化分类输出的信息熵，优化目标，如式 (11) 所示：

$$\min_{F_1,F_2} L_{\text{cl}}(X_s,Y_s)-L_d(X_t)-L_{\text{ent1}}(X_t)-L_{\text{ent2}}(X_t) \tag{11}$$

3) 训练特征生成器

固定分类器 F_1 和 F_2 的参数，利用目标域数据训练特征生成器 G，最小化分类差异和分类概率信息熵，使目标域特征靠近相似类别的源域特征，同时远离决策边界，使特征更加具有区分性。优化目标如式 (12) 所示：

$$\min_{G} L_d(X_t)+L_{\text{ent1}}(X_t)+L_{\text{ent2}}(X_t) \tag{12}$$

在训练过程中，将不断重复上述 3 个步骤，以实现特征生成器和分类器关于分类距离和信息熵的对抗训练。

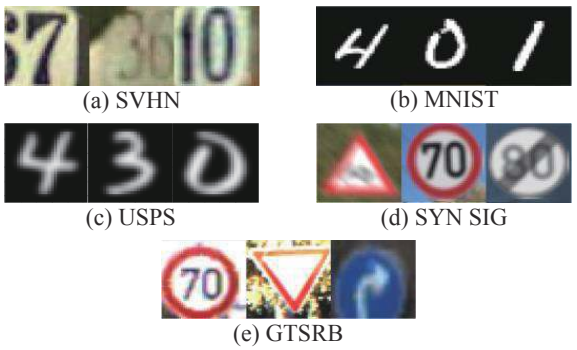
3　实验设计与结果分析

为了评价 ACDIE 算法的性能和效果，本文设计了 4 种实验：数字标识域适应实验、实物域适应实验、t-SNE 图可视化实验、信息熵损失对比实验。特征生成器 G 采用包括卷积层、池化层的卷积神经网络进行特征提取，分类器 F_1 和 F_2 采用具有相同网络结构的全连接神经网络进行分类。在 G、F_1、F_2 网络中加入批次归一化 (batch normalization, BN) 层来提高网络的训练和收敛的速度，防止梯度爆炸和梯度消失的发生，同时通过 Dropout 层来防止模型过拟合。本文实验基于 pytorch 深度学习框架，Ubuntu16.04 操作系统，采用 E5-2 670 处理器，GPU 为 GeForce GTX1080Ti，内存 32 GB。

3.1　数字标识域适应实验

3.1.1　数据集

选择机器学习领域常用数据集进行域适应实验, 包括 MNIST[19]、USPS[20]、SVHN[21]、SYN SIG[22] 和 GTSRB[23], 示例图片如图 5 所示。SVHN 是现实生活中的街道门牌号数字数据集, 包含 99289 张 32 像素×32 像素的彩色图片; MNIST 为手写数字识别数据集, 包含 65 000 张 32 像素×32 像素的灰度图片; USPS 为美国邮政服务手写数字识别数据集, 包含 6 562 张 28 像素×28 像素的灰度图像, 这些数据集共计 10 个类别的图像; SYN SIG 是合成的交通标志数据集; GTSRB 是真实世界的标志数据集, 共计 43 个类别的图像。



(a) SVHN　(b) MNIST　(c) USPS　(d) SYN SIG　(e) GTSRB

图 5　数字标识数据集示例

Fig. 5　Digital ID dataset example

对于这 5 个域的数据样本, 设置 5 种不同的域适应情况: SVHN → MNIST、SYN SIG → GTSRB、MNIST → USPS、MNIST → USPS*和 USPS → MNIST。在本文实验中, USPS 表示使用 1 800 张 USPS 数据集样本, USPS*表示使用全部的 USPS 数据集样本来训练模型, 数据集样本数量设置与文献 [17] 相同。

3.1.2　实验超参数

使用 mini-batch 随机梯度下降的优化器算法, batch size 设置为 128, 随机种子值设置为 1, Learning rate 设置为 0.000 2, 通过 Adam 优化器实现网络参数更新, weight decay 设置为 0.000 5。

3.1.3　对比实验结果

将本文算法与其他在域适应领域有代表性的方法进行比较, 包括 MMD[9]、DANN[14]、分离域共享特征和域独有特征的 DSN[24]、基于域鉴别器对抗训练的 ADDA[25]、学习多域联合分布的 CoGAN[26]、利用图像生成的对抗过程学习源域和目标域特征分布差异最小化的 GTA[16], 以及最大化决策分类器差异的 MCD[17]。表 1 展示了不同方法在 5 种实验设置情况下的域适应准确率, 其中: Source Only 表示只使用源域数据进行训练而不进行域适应; 分类精度最高的值用粗体表示。根据实验结果, 对于 5 种不同的域适应情况, ACDIE 算法的准确率都为最高值。特别是, 在 MNIST→USPS 的实验中, ACDIE 模型的域适应分类准确率可以达到 97.4%, 相较于 MCD 的分类准确率提高了 3.2%。另外, 在其他 4 种域适应情况下, 相较于其他最好的域适应算法, ACDIE 模型的分类准确率也提高了 2.1%～2.6%。对比 MNIST→USPS 和 MNIST→USPS*的准确率结果, 可以发现通过更多的目标域数据可以进一步提高域适应效果。

表 1　数字标识数据集域适应准确率对比

Table 1　Comparison of the domain adaptation accuracies of digital ID datasets　%

算法	SVHN→ MNIST	SYNSIG→ GTSRB	MNIST→ USPS	MNIST→ USPS*	USPS→ MNIST
Source Only	67.1	85.1	76.7	79.4	63.4
MMD	71.1	91.1	—	81.1	—
DANN	76.0	88.7	77.1	85.1	73.2
DSN	82.7	93.1	91.3	—	—
ADDA	76.0	—	89.4	—	90.1
CoGAN	—	—	91.2	—	89.1
GTA	92.4	—	92.8	95.3	90.8
MCD	96.2	94.4	94.2	96.5	94.1
ACDIE	**98.8**	**96.7**	**97.4**	**98.6**	**96.2**

3.2　实物域适应实验

3.2.1　Office-31 数据集

为了测试模型对于实际物体图片的域适应效果, 设计在 Ofiice-31 数据集的域适应实验。Ofiice-31 数据集含有 31 类不同物品的图片, 共计 4 652 张, 是测试域适应算法的通用数据集。该数据集的图片分别来自 3 种不同的数据域, 包括在亚马逊网站收集的样本数据 Amazon(A)、通过电脑摄像头拍摄得到的样本数据 Webcam(W)、利用单反相机拍摄得到的样本数据 DSLR(D)。图 6 分别为 A、D、W 这 3 个不同域的图片数据。对于这 3 个域的数据样本, 设置 6 种不同的域适应情况: A→D、A→W、D→A、D→W、W→A、W→D。

(a) A 域　(b) D 域　(c) W 域

图 6　Office-31 数据集示例

Fig. 6　Office-31 dataset example

3.2.2　实验超参数

使用 mini-batch 随机梯度下降的优化器算法，batch size 设置为 32，随机种子值设置为 2 020。特征提取器 G 采用预训练的 ResNet-50 网络，使用 SGD 优化器进行梯度更新，学习率设置为 0.001，权重衰减参数为 0.000 5。分类器 F 采用两层全连接的网络结构，使用 SGD 优化器进行梯度更新，学习率设置为 0.001，权重衰减参数为 0.000 5，momentum 值设置为 0.9。

3.2.3　对比实验结果

为了对比实验的合理性，所有方法在同等条件下进行对比实验，选取 ResNet-50 网络作为特征提取网络，对比方法包括 DANN[14]、GTA[16] 和使用条件对抗域适应的 CDAN[27]。表 2 展示了不同方法在 6 种实验设置情况下的域适应准确率，其中 ResNet-50 表示使用 ResNet-50 作为特征提取器对源域数据进行训练而不进行域适应。

表 2　Office-31 数据集域适应准确率对比

Table 2　Comparison of the domain adaptation accuracies of Office-31 dataset　%

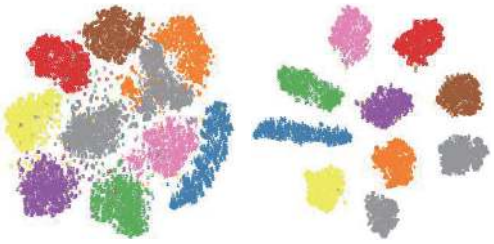
算法	A→D	A→W	D→A	D→W	W→A	W→D	平均值
ResNet-50	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DANN	79.7	82.0	68.2	96.9	67.4	99.1	82.2
GTA	87.7	89.5	72.8	97.9	71.4	99.8	86.5
CDAN	**89.8**	93.1	70.1	98.2	68.0	99.9	86.5
ACDIE	89.2	**93.1**	**73.0**	**98.6**	**71.7**	**100.0**	**87.6**

从实验结果可以看出，相较于现有的算法模型，本文所提出的 ACDIE 模型在不同域适应情况下的分类准确率都有不同程度的提高。在 D→W 和 W→D 的情况下的域适应结果分别达到 98.6% 和 100%，因为 D 与 W 两个域之间的图片差异较小，所以可以达到一个很高的分类准确率。在 A→D 和 A→W 的情况下准确率较 GTA 算法分别提高了 1.5% 和 3.6%，说明 ACDIE模型在两个域之间的差异较大的情况下仍能达到较好的域适应效果。ACDIE 模型在 Office-31 数据集上的平均域适应准确率达到 87.6%。

3.3　t-SNE 图可视化实验

为了更加直观地看到经过域适应后特征向量的变化，本文采用 t-SNE[28] 方法将高维特征向量映射到适合观察的二维向量，进而实现数据的可视化。

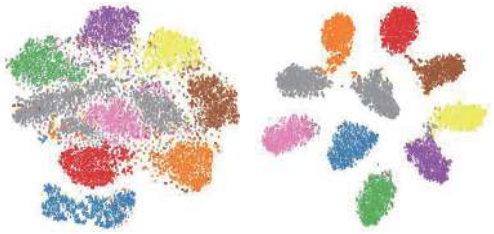
图 7 和图 8 分别是在 SVHN→MNIST 和 USPS→MNIST 两种域适应情况下，目标域样本特征分布的变化情况。每种颜色代表一个类别，左边为进行域适应前不同类别样本的可视化，右边为进行域适应后不同样本的可视化。通过 t-SNE 图发现，在域适应前目标域数据不同类别之间的距离较小，且决策边界较为模糊。通过 ACDIE 模型的域适应后，目标域相同种类的数据更加集中，不同种类的数据之间的距离增大，这使得分类器更加容易实现对目标域数据的分类。



(a) 域适应前　　(b) 域适应后

图 7　SVHN→MNIST 的 t-SNE 图

Fig. 7　t-SNE diagram of SVHN→MNIST



(a) 域适应前　　(b) 域适应后

图 8　USPS→MNIST 的 t-SNE 图

Fig. 8　t-SNE diagram of USPS→MNIST

3.4　信息熵损失对比实验

为了验证将信息熵损失加入对抗训练的有效性，以基于分类差异的域适应模型为基础，设置 4 组对比实验：1) 不加入信息熵损失；2) 仅在优化 F 时加入信息熵损失；3) 仅在优化 G 时加入信息熵损失；4) 信息熵损失对抗训练，即 ACDIE 模型。

从表 3 的对比实验结果可以看出，在实验 3 的情况下，通过在优化特征生成器 G 时加入信息熵损失，使信息熵损失减小，可以使生成的特征远离决策边界，从而达到更高的域适应准确率，证明引入信息熵损失的有效性。在实验 2 的情况下，通过在优化分类器 F 时加入信息熵损失，使信息熵损失增大，实验结果与实验 1 大致相同，在 MNIST→USPS(p) 和 USPS→MNIST 下准确率有所下降，因为分类器 F 信息熵增加，决策边界更加模糊，一部分靠近边界的样本数据会被错误分类。在实验 4 中，即 ACDIE 模型，通过对抗训练的方式实现特征生成器 G 的信息熵损失最小化，域适应准确率相较于实验 3 进一步提高，证明了将信息熵损失加入对抗训练的有效性。

表 3　信息熵损失对比实验

Table 3　Comparative experiment of information entropy loss　%

实验设置	SVHN→MNIST	MNIST→USPS	MNIST→USPS(p)	USPS→MNIST
1	96.3	94.5	96.6	94.3
2	96.7	94.8	95.7	92.4
3	97.9	96.6	98.2	95.5
4	98.8	97.4	98.6	96.2

4　结束语

现有无监督域适应算法仅将不同域之间的距离拉近, 没有考虑目标样本与决策边界之间的关系, 没有扩大目标域内不同类别样本之间的距离。针对上述问题, 本文提出利用两个分类器之间的不一致性对齐域间差异, 减小源域和目标域之间的距离, 同时通过最小化信息熵来降低分类不确定性的 ACDIE 模型。最小化信息熵能使相同类别的数据更加聚集, 不同类别数据之间的距离更大, 而且可以使目标域样本与源域样本在语义空间上分布更加对齐。大量的实验表明, 本文提出的的模型相比于领域内其他模型取得了更优的性能, 验证了所提改进算法的有效性。

尽管 ACDIE 模型在多个数据集中都有不错的表现, 但它仍存在一些提升空间。在今后的工作中, 将进一步从信息论的角度思考, 考虑互信息等因素对模型的影响, 以提升模型的准确率和鲁棒性。同时将进一步探究不同距离分布度量对域适应结果的影响。

参考文献:

[1] WANG Xiaolong, GUPTA A. Unsupervised learning of visual representations using videos[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 2794−2802.

[2] MAHJOURIAN R, WICKE M, ANGELOVA A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 5667−5675.

[3] 刘建伟, 孙正康, 罗雄麟. 域自适应学习研究进展 [J]. 自动化学报, 2014, 40(8): 1576–1600.
LIU Jianwei, SUN Zhengkang, LUO Xionglin. Review and research development on domain adaptation learning[J]. Acta automatica sinica, 2014, 40(8): 1576–1600.

[4] PAN S J, YANG Qiang. A survey on transfer learning[J]. IEEE transactions on knowledge and data engineering, 2010, 22(10): 1345–1359.

[5] ROZANTSEV A, SALZMANN M, FUA P. Beyond sharing weights for deep domain adaptation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(4): 801–814.

[6] GHIFARY M, KLEIJN W B, ZHANG Mengjie, et al. Deep reconstruction-classification networks for unsupervised domain adaptation[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016: 597−613.

[7] SENER O, SONG H O, SAXENA A, et al. Learning transferrable representations for unsupervised domain adaptation[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016: 2110−2118.

[8] SUN Baochen, FENG Jiashi, SAENKO K. Return of frustratingly easy domain adaptation[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona: AAAI Press, 2016: 2058−2065.

[9] GRETTON A, BORGWARDT K M, RASCH M J, et al. A kernel two-sample test[J]. The journal of machine learning research, 2012, 13: 723–773.

[10] TZENG E, HOFFMAN J, ZHANG Ning, et al. Deep domain confusion: maximizing for domain invariance[J]. Computer science, 2014.

[11] LONG Mingsheng, CAO Yue, WANG Jianmin, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR, 2015: 97−105.

[12] LONG Mingsheng, ZHU Han, WANG Jianmin, et al. Unsupervised domain adaptation with residual transfer networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016: 136−144.

[13] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014: 2672−2680.

[14] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The journal of machine learning research, 2016, 17(1): 2096–2030.

[15] 王格格, 郭涛, 余游, 等. 基于生成对抗网络的无监督域适应分类模型 [J]. 电子学报, 2020, 48(6): 1190–1197.
WANG Gege, GUO Tao, YU You, et al. Unsupervised

domain adaptation classification model based on generative adversarial network[J]. Acta electronica sinica, 2020, 48(6): 1190–1197.

[16] SANKARANARAYANAN S, BALAJI Y, CASTILLO C D, et al. Generate to adapt: aligning domains using generative adversarial networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 8503−8512.

[17] SAITO K, WATANABE K, USHIKU Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 3723−3732.

[18] 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.

[19] 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.

[20] HULL J J. A database for handwritten text recognition research[J]. IEEE transactions on pattern analysis and machine intelligence, 1994, 16(5): 550–554.

[21] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning [C]//Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain, 2011: 5−16.

[22] MOISEEV B, KONEV A, CHIGORIN A, et al. Evaluation of traffic sign recognition methods trained on synthetically generated data[C]//Proceedings of the 15th International Conference on Advanced Concepts for Intelligent Vision Systems. Poznań, Poland: Springer, 2013: 576−583.

[23] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. The German traffic sign recognition benchmark: a multi-class classification competition[C]//Proceedings of 2011 International Joint Conference on Neural Networks. San Jose, USA: IEEE, 2011: 1453−1460.

[24] BOUSMALIS K, TRIGEORGIS G, SILBERMAN N, et al. Domain separation networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016: 343−351.

[25] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 7167−7176.

[26] LIU Mingyu, TUZEL O. Coupled generative adversarial networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016: 469−477.

[27] LONG Mingsheng, CAO Zhangjie, WANG Jianmin, et al. Conditional adversarial domain adaptation[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc., 2018: 1647−1657.

[28] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(2605): 2579–2605.

作者简介：

李庆勇，硕士研究生，主要研究方向为无监督学习和计算机视觉。

何军，副教授，主要研究方向为机器学习、计算机视觉、最优化方法。获发明专利授权 4 项，发表学术论文 30 余篇。

张春晓，硕士研究生，主要研究方向为无监督学习和计算机视觉。