广东工业大学硕士学位论文

（工学硕士）

无领域信息的跨领域适应算法研究

李嘉豪

二〇一九年五月二十七日

分类号：　　学校代号：11845

UDC：　　密级：　　学　　号：2111605020

广东工业大学硕士学位论文

（工学硕士）

无领域信息的跨领域适应算法研究

李嘉豪

指导教师姓名、职称：　蔡瑞初　教授

专 业 或 领 域 名 称 ：　计算机科学与技术

学 生 所 属 学 院 ：　计算机学院

论 文 答 辩 日 期 ：　二〇一九年五月

A Dissertation Submitted to Guangdong University of Technology

for the Degree of Master

（Master of Engineering Science）

Cross-domain Adaptive Algorithm without Domain Information

Candidate: Li Jiahao

Supervisor: Prof. Cai Ruichu

May 2019

School of Computer Science and Technology
Guangdong University of Technology
Guangzhou, Guangdong, P. R. China, 510006

摘　　要

随着大数据时代的来临，人类的各种活动产生了大量的数据。这很大程度归功于入网设备和网络社交的普及。由于数据处理往往滞后于数据生成，大量的无标签数据无法得到及时处理。这些数据不应该通过人工标注获得类别信息，因为人工标注会阻碍数据处理。因此，常用做法是利用已训练好的模型把标注信息迁移到新增数据上。不过，由于这种模型往往具备领域专用性，这种做法的效果往往不尽人意。为此，无监督领域自适应算法被提了出来。这种算法试图从有标注的源领域数据中挖掘具备迁移性的类别信息，并把这些信息迁移到无标注的目标领域数据中，是当前最热门的一种研究方向。为了将这种算法应用到多个源领域中，这种算法又发展出了无监督多源领域自适应算法。不过，几乎所有的多源领域自适应算法都受限于领域可分假设，无法适用于混合源场景中。少量没有这种假设的多源自适应算法因为其他假设而适用于特定的应用场景中。

本文基于类别同态假设提出了一种无领域信息的跨领域自适应算法。在这种假设下，不同的源领域有不同的类别分布和数据噪声模式，但所有源领域的同一类别有相同的数据产生机制。也就是说，存在一个特征转换函数，使得所有源领域的类别决策空间是一致的。如果这个空间和目标领域的隐空间是相似的，那么多个源领域的类别信息就能够迁移到目标领域上。为了得到这个特征转换函数，所提算法不仅约束输出特征具备类别信息，还要对齐源领域期望分布和目标领域分布。考虑到类别信息的流失问题，所提算法选择对齐领域的类内分布。为了对齐类内分布，本文采用了两种策略来评估分布差异。第一种策略是最小化总体均方离差，能够集中同类样本到一致的类簇中心。第二种策略是最小化最大均值差异，能够对齐领域之间的类内分布。在无监督情形下，目标领域的类别空间是不可访问的。因此，目标领域的训练样本事先通过模型获得伪标签。此外，考虑到特征转换函数可能比较复杂，所提算法采用多任务学习的方式训练卷积神经网络。

对比实验表明，在没有源领域信息的帮助下，基于类别同态假设的跨领域适应算法总体上依然优于其他方法。此外，基于最大均值差异的策略要稍好于基于总体均方离差的策略。隐层特征的可视化结果表明所提算法能够很好地保留数据的类别信息，从而获得了鲁棒的性能。

关键字：无监督领域自适应；总体均方离差；最大均值差异；多任务学习；卷积神经网络

ABSTRACT

With the advent of the big data era, human activities have generated a large amount of data. This is largely due to the popularity of networking devices and social networking. Since data processing is less efficient than data generation, a large number of unlabeled data cannot be processed in time. These data should not be labeled manually to obtain category information, because manual labeling can hinder data processing. Therefore, it is common practice to transfer annotation information to new data using well-trained models. However, the result of this approach is often unsatisfactory because of the existing model with domain-specificity. Thus, unsupervised domain adaptive algorithm is proposed. As one of the most popular research, this algorithm extracts transferable category information from labeled data in the source domain and transfers this information to unlabeled data in the target domain. In order to apply this algorithm to multiple source domains, unsupervised multi-source domain adaptive algorithm is developed. However, almost all multi-source domain adaptive algorithms are limited by domain-separable assumptions and cannot be applied to mixed-source scenarios. A small number of multi-source adaptative algorithms without this assumption are only suitable for specific application scenarios because of other assumptions.

In this paper，a cross-domain adaptive algorithm without domain information is proposed based on the category homomorphism assumption. Under this assumption, different source domains have different category distributions and data noise patterns, but the same category of all source domains has the same data generation mechanism. That is to say, there exists a feature transformation function, which makes all source domains have a consistent decision space. If this space is similar to the hidden space of the target domain, the category information of multiple source domains can be migrated to the target domain. In order to obtain this feature transformation function, the proposed algorithm not only constrains the output feature to preserve category information but also aligns the expected distribution of source domain and target distribution. Considering the loss of category information, the proposed algorithm chooses to align the intra-class feature distribution. In order to align intra-class distribution, this paper provides two strategies to evaluate distribution discrepancy. The first strategy is to minimize the total mean square deviation, which can centralize intra-class instances to the cluster center. The second strategy is to minimize the maximum mean discrepancy, which can align the intra-class distributions of domains. Under the unsupervised scenario, the category space of the target domain is inaccessible. Therefore, all the target instances are given pseudo labels from the model. In addition, considering that the feature transformation function may

be complex, the proposed algorithm uses multi-task learning to train convolution neural networks.

The comparative experiments show that, without the help of source domain information, the category-homomorphism-based cross-domain adaptive algorithm is still better than other methods in general. In addition, the strategy based on maximum mean discrepancy is slightly better than that based on the total mean square deviation. The visualization of hidden features shows that the proposed algorithms can well retain the category information of data and thus achieve robust performance.

Key words: unsupervised domain adaptation; total mean square deviation; maximum mean discrepancy; multi-task learning; convolution neural network

目　　录

CONTENTS

第一章 绪论

1.1 研究背景及意义

数据是机器学习的主要处理对象。根据数据的标注程度，学界已经发展出三种经典的机器学习算法来处理数据。有监督学习算法被用于处理有标签数据，从而能够使模型按照人类的行为准则做出相应的决策。无监督学习算法被用于处理无标签数据，从而能够使模型根据数据的统计特性归纳内在的规律。半监督学习算法被用于处理混合数据，从而能够集中有监督学习和无监督学习的优势得到更优质的模型。

然而，这些经典算法都假设训练数据和测试数据来自同一个分布。在大数据时代背景下，大量来自各个领域的数据明显不符合这种假设。为了对这些包含多个领域的数据进行高效分析和学习，学界在这些经典算法的基础上引入了迁移学习[1]和领域自适应[2]等建模思想。基于这些思想发展起来的算法能够从一个或多个源领域中提取出有效的知识，并把这些知识迁移到其他目标领域中。这不仅避免了频繁的模型重建工作，而且还能借助已有模型降低人工干预的程度。

早期，为了能够对目标领域进行快速建模，从源领域训练出来的模型会基于目标领域数据进行再训练。通过引入源领域先验知识，这种串行式做法一定程度上弥补了目标领域数据不足的缺点。不过在源领域和目标领域差异较大时，这种做法并不比经典方法有更好的性能表现。其主要原因在于源领域先验知识没有在后续训练中得到很好的保留。为此，不少研究工作[3,4,5]基于多任务学习[6,7,8]的思路重新设计了目标函数。这个目标函数包含两个任务。主要任务往往与业务数据相关，可能是一个分类任务[9]或者聚类任务[10]。次要任务用于挖掘源领域和目标领域之间的公共知识，通常是一个领域分布对齐任务。这个任务不仅可以对齐领域总体分布[11]，也可以对齐领域条件分布。不过后者需要人工给定真实标签或使用模型预测伪标签[12]。

目前，这种跨领域适应算法拥有广阔的应用前景。在生物信息行业中，基因测序的开销已经降至几百美元，从而使很多医疗平台能够开放基因测序及分析服务。将这些来自不同平台的基因序列数据整合分析，能够更好地刻画基因疾病的发生机理，并制定新的基因疗法。在公共安全行业中，成千上万的监控摄像头每小时都会产生大量的图像数据。这要求有一种通用的处理技术对可疑人物或物品进行实时报警。不过，考虑到主流算法只能应用在单个或多个独立源领域到单个目标领域的场景中，因此在混淆多个源领域的场景下，对无领域信息的跨领域适应算法展开研究是很有必要的。

1.2 国内外研究现状

由于数据生成和数据处理在效率上存在巨大的鸿沟，大量堆积的数据都没有备注相应的标签。为了高效地学习和处理这些无标签数据，无监督领域自适应算法成为了迁移学习中新的研究方向。这种算法主要解决了分类知识的迁移问题，从而使源领域的分类知识能够移植到目标领域中，避免了低效的人工标注工作。

当前，主流的研究工作都是在分布对齐的基础上设计并完善无监督领域自适应算法。一些工作假设源领域和目标领域之间存在一个低维映射函数，使得与领域弱相关的局部特征得以保留，从而在低维空间中对齐两个领域的总体分布[13,14,15]。另外一些工作假设源领域和目标领域之间存在一个分布包含关系，使得目标领域的总体分布能够通过挑选或重加权源领域样本来得到[11,16,17]。这两种方法能够互补地对齐领域分布。最有效的一种做法是为源领域和目标领域分别配置一个投影矩阵，接着借助相应的约束或策略使得加权结果和投影矩阵均具备领域一致性[18,19]。为了寻得能够对齐条件分布或总体分布的投影矩阵，算法一般将两个领域的原始样本映射到再生核希尔伯特空间中，并执行类内对齐策略[20,21,22]和类间分离策略[23,24,25]。其中，最大均值差异是一种基于概率统计的分布对齐技术。它能够配合支持向量机[26]对齐源领域和目标领域的总体分布[27,28]。此外，投影矩阵的分布对齐效果也可以通过一阶或二阶度量方法来量化。一般来说，基于一阶度量的方法采用欧氏距离或马氏距离作为相似性度量[29,30,31]，而基于二阶度量的方法采用协方差作为两个领域的相似性度量[32,33,34]。然而，这些面向浅层模型设计的算法只能解决简单的领域自适应场景。由于浅层模型的参数空间不足以让算法对齐两个领域的总体分布或条件分布，算法无法在复杂的领域自适应场景下挖掘足够的分类知识。

为了避免浅层模型的缺点，不少研究工作把无监督领域自适应算法移植到深度学习模型上。这些工作首先把神经网络的一部分视作特征提取器，接着利用最大均值差异对齐不同领域间的特征分布 [3,35,36,37]。在图像分类任务中，为了保证特征提取器能够输出具备分类能力的特征，神经网络的剩余部分会充当分类器，并且被交叉熵等损失函数优化[3,35]。这种做法不仅可以对齐两个领域的边缘分布 $\pi_S(X)$ 和 $\pi_T(X)$，也可以对齐两个领域的后验分布 $p_S(Y|X)$ 和 $p_T(Y|X)$，从而统一分类决策空间[38,39,40]。此外，一些研究结果表明，通过在端对端神经网络中额外插入一个分支网络用于最小化领域重构误差[4,41,42,43]或领域预测准确率[5,44,45]，两个领域的样本能够被同时投影到一个不受领域影

响的隐空间中，从而使两个领域的分布尽可能地相似。无监督领域自适应算法也可以基于对抗学习思想[46,47,48]进行设计。只要把传递到特征提取器的梯度进行反转，领域判别器就能与特征提取器建立起对立关系，从而使特征提取器能够过滤掉领域专用的信息[5,49,50,51]。为了提高模型的对抗力度，领域判别器可以额外乘上一个权重，从而改进领域判别器的混淆效率[52]。此外，通过共享局部权重或判别器，两个耦合的对抗生成网络能够对齐生成器的输出特征分布[53,54]。然而，虽然深度领域自适应模型有效避免了浅层模型所带来的各种潜在缺点，但是自由度过大的模型参数空间往往导致分类误差和分布差异无法被算法同时优化，尤其是在类别不平衡程度受领域因素强烈影响的情况下。

归根结底，上述问题的产生原因是这些方法都针对源领域和目标领域的所有数据来评估和优化分布差异。这种方法未必能够从源领域中迁移分类知识到目标领域中。因为对齐的领域总体分布只能表明源领域确实具备某种可迁移的知识，所以通过对齐分布而间接得到的知识有可能是一种对模型分类无益的知识。总而言之，真正有意义的分类知识往往因对齐领域总体分布而流失。为此，一些研究工作事先通过模型预测所有数据的伪标签，然后根据伪标签对齐不同领域中的同类特征分布[55,56,57]。这种做法不仅有效避免类别信息的流失，还能筛减掉那些只包含领域专用信息的样本。此外，在某种程度上，伪标签的使用也能够减轻源领域对分类器的主导作用，避免了分类器对源领域的过拟合。

至此，适用于两个领域的自适应算法基本发展成熟。不过，这些算法只考虑单个源领域的情况，因此把这些算法扩展到多源领域是很有必要的。目前，已有部分算法能够适用于多源领域的情况。以 Just DIAL[58]为例，这个算法能够对不同领域执行不同的归一化策略，从而使不同领域被映射到同一个决策空间中。在多源领域可以被区分的情况下，这种做法可以推广到任意多的源领域中[59]。受核方法的启发，各种基于核的度量也能够同时对多个领域的分布差异进行评估[60,61]。

然而，在现实世界中，人类各种活动数据可能来自多个不明确的来源，并统一由某些学术或商业机构所收集保存。在这种情形下，领域可区分的前提假设被破坏，从而导致大部分的多源领域自适应算法失效。为此，无领域信息的跨领域适应算法被提出，避免了因领域不可分而失效的问题。目前，唯一可行的方法是使用半监督聚类驱使概率图模型挖掘数据中的隐含领域[62]。不过，这种方法也面临其他假设不成立的窘境。

1.3 本文主要工作

本文主要针对混合源场景下的无监督多源领域自适应问题展开研究，并提出了一种基于类别同态假设的无领域信息的跨领域适应算法。在类别同态假设下，本文所提算法只要对齐类内分布就能够从有标签的混合源数据中提取类别信息，并迁移到无标签的目标数据中。由于目标领域的样本缺乏相应的标签，所提算法会通过模型临时对这些样本赋予伪标签，从而使类内分布能够顺利对齐。在类别同态假设下，基于类内分布的对齐策略能够减少不同领域之间的噪声影响，从而使模型更具泛化性和鲁棒性。本文的主要贡献包含以下两个方面：

（1）跨领域自适应的风险上界及类别同态假设。在单源领域自适应的风险上界基础上，本文提供了三个跨领域自适应的风险上界。借助这个上界，本文能够从理论上解释类别同态假设在混合源场景中的作用。

（2）两种类内分布对齐策略。总体均方离差和最大均值差异是常用的两种分布差异度量。然而，在无监督场景下，这两种度量很难评估和对齐类内分布。主要原因是这两种度量都需要收集某些特定信息。总体均方离差需要收集类内均值，而最大均值差异需要收集两份同类样本。本文针对这些信息分别提供了对应的解决方案。

1.4 论文组织架构

本文主要包含以下四个章节：

第一章为绪论部分，主要介绍本文研究的背景及意义，并概述无监督领域自适应的相关工作及其发展历程。此外，绪论部分还对本文主要工作及组织架构进行了说明。

第二章为相关理论及技术，主要介绍了霍夫丁不等式、VC 泛化上界、最大均值差异、多任务学习、卷积神经网络和 Caffe 深度学习框架。前三个主要涉及到本文的理论分析部分，其余三个则涉及到模型的设计。

第三章为无领域信息的跨领域自适应，主要介绍了跨领域自适应中的三个风险上界和类别同态假设，并详细描述了两种类内分布对齐策略。其中的风险上界能够分析类别同态假设在混合源场景中的作用。

第四章为实验结果及分析，主要介绍两种类内分布对齐策略在基因数据和图像数据上的性能表现，并分析了结果背后的原因。

第二章 相关理论及技术

2.1 霍夫丁不等式

霍夫丁不等式（Hoeffding's inequality）于 1963 年被 Wassily Hoeffding 提出并且证明[63]。这个不等式能够描述独立随机变量序列的经验平均值及其理论期望值之间的偏差概率收敛情况。正式来说，假定$\xi_1,\xi_2,\ldots,\xi_n$为一个独立随机变量序列，且每一个随机变量ξ_i有取值范围$[a_i,b_i]$，那么定义这些变量的经验均值变量为

$$\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$$

并且对于任意实数$\varsigma>0$，有不等式

$$\Pr\left(\bar{\xi}-\mathbb{E}\left[\bar{\xi}\right]>\varsigma\right)\le\exp\left(-\frac{2n^2\varsigma^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) \tag{2.1}$$

和

$$\Pr\left(\left|\bar{\xi}-\mathbb{E}\left[\bar{\xi}\right]\right|>\varsigma\right)\le 2\exp\left(-\frac{2n^2\varsigma^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) \tag{2.2}$$

其中，$\mathbb{E}[\bar{\xi}]$为经验均值变量的理论期望值。

由公式（2.1）和（2.2）知，随着独立随机变量的数量n增大，经验平均值会愈加逼近于理论期望值。这个理论结果可以说明现实世界中的很多事实。举个例子，若一个罐子里存放着红球和绿球，其中红球占所有小球的比例为μ。由于事先无法得知这个比例μ，因此有必要对罐中小球进行有放回抽样，并统计被选中小球中红色球的比例ν。由公式（2.2）知，只要抽样次数n足够大，那么ν很有可能接近于μ，因为

$$\Pr\left(\left|\nu-\mu\right|>\varsigma\right)\le 2\exp\left(-2n\varsigma^2\right)$$

此外，霍夫丁不等式还可以推广到一般的二分类问题上。假设输入空间和输出空间分别为$\mathcal{X}$和$\mathcal{Y}=\{+1,-1\}$，并且$\mathcal{X}\times\mathcal{Y}$上存在一个未知分布$P_{XY}$。若给定一个假设空间$\mathcal{F}\subset\mathcal{Y}^{\mathcal{X}}$和一个二分损失函数$\ell:\mathcal{F}\times\mathcal{X}\times\mathcal{Y}\to\{0,1\}$，那么对于任意假设函数$f\in\mathcal{F}$，有期望误差

$$E(f)=\mathbb{E}_{P_{XY}}\left[\ell\left(f,X,Y\right)\right] \tag{2.3}$$

和经验误差

$$E_n(f)=\frac{1}{n}\sum_{i=1}^{n}\ell\left(f,x_i,y_i\right) \tag{2.4}$$

其中，(x_i,y_i)为第i次采样所得到的一个实例，n代表从$\mathcal{X}\times\mathcal{Y}$中采样的次数。

一般情况下，假设空间$\mathcal{F}$中至少存在一个函数f^*使得公式（2.3）中的E最小。然而，由于概率分布P_{XY}无法事先被得知，这个最优函数f^*无法通过遍历整个假设空间$\mathcal{F}$来得到。不过，如果把公式（2.1）中的每个随机变量ξ_i替换为$\ell(f,x_i,y_i)$，那么存在以下不等式

$$\Pr\left(E_n(f)-E(f)>\varsigma\right)\le\exp\left(-2n\varsigma^2\right)$$

也就是说，只要采样的次数n足够大，经验误差$E_n(f)$可以逼近期望误差$E(f)$。因此，即使概率分布P_{XY}未知，最优函数f^*依然可以通过公式（2.4）来寻得。此外，如果把公式（2.1）中的每个随机变量ξ_i替换为$-\ell(f,x_i,y_i)$，那么还有以下式子

$$\Pr\left(E(f)-E_n(f)>\varsigma\right)\le\exp\left(-2n\varsigma^2\right) \tag{2.5}$$

如果把$E(f)-E_n(f)>\varsigma$看作一个事件，那么其对立事件$E(f)-E_n(f)\le\varsigma$的发生概率为$1-\exp(-2n\varsigma^2)$。令$\exp(-2n\varsigma^2)=\delta$，那么公式（2.5）可以写成

$$\Pr\left(E(f)-E_n(f)\le\sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right)\ge1-\delta$$

其中，δ是一个取值于$(0,1)$的实数。这个式子表明至少有概率$1-\delta$保证以下上界成立。

$$E(f)\le E_n(f)+\sqrt{\frac{\log\frac{1}{\delta}}{2n}} \tag{2.6}$$

因此，只要保证训练数据足够多，$E(f)$可以间接通过最小化$E_n(f)$来优化。

值得注意的是，公式（2.6）只能用于假设空间$\mathcal{F}$有限的情形。由公式（2.6）知，所有假设函数$f\in\mathcal{F}$都满足公式（2.6）的概率至少为$(1-\delta)^{|\mathcal{F}|}$。为了让这个概率足够大，一般情况下，$\delta$会取得更小，并且为了让上界保持紧凑，训练数据的规模n也必须相应取得更大。不过，当假设空间$\mathcal{F}$为无限时，无论怎么调整δ的数值，所有假设函数$f\in\mathcal{F}$都满足公式（2.6）的概率都接近于 0，从而导致公式（2.6）失效。为了解决这个问题，一般需要引入 VC 维[64]对假设空间$\mathcal{F}$进行分析，从而避免考虑假设空间$\mathcal{F}$的所有情形。相关分析请参考下一节和文献[64,65]。

2.2 VC 泛化上界

在上一节中，式（2.6）已经表明，在给定假设函数 $f \in \mathcal{F}$ 后，经验误差和期望误差之间的差异是依概率收敛于 0 的。不过，这个结论只能验证单个假设函数，并不能让算法在假设空间 $\mathcal{F}$ 中构成学习。因此，式（2.6）仍需进一步被推广到所有的假设函数 $f \in \mathcal{F}$ 上。然而，这样的推广是无意义的。由于算法只能通过损失函数 ℓ 来辨别假设函数 f，算法往往无法遍历整个假设空间 $\mathcal{F}$。具体来说，在给定损失函数 ℓ 后，存在一个集合 $\Gamma \coloneqq \{\{\ell(f,x,y)\}_{x\times y} \mid f \in \mathcal{F}\}$，使得其中的元素和假设空间 $\mathcal{F}$ 的多个元素对应。也就是说，假设空间 $\mathcal{F}$ 可能存在 $f_1 \neq f_2$，使得 $\{\ell(f_1,x,y)\}_{x\times y} = \{\ell(f_2,x,y)\}_{x\times y}$。显然，受限于损失函数的形式，算法可能无法对 f_1 和 f_2 进行区分，而只能区分假设空间 $\mathcal{F}$ 中不同的等价类。这些等价类中的假设函数都会使损失函数 ℓ 输出相同的像。若把拥有相同像的假设函数 f 记为等价类 $\hat{f}$，并将其组成类空间 $\widehat{\mathcal{F}}$，那么集合 Γ 和类空间 $\widehat{\mathcal{F}}$ 存在双射关系，亦即

$$\widehat{\mathcal{F}} \underset{\ell^{-1}}{\overset{\ell}{\rightleftarrows}} \Gamma$$

由于集合 Γ 的元素和类空间 $\widehat{\mathcal{F}}$ 的元素一一对应，算法在类空间 $\widehat{\mathcal{F}}$ 上的学习相当于在集合 Γ 上的学习。为了更好地描述这个学习过程，不妨定义 $\{(x_i,y_i)\}_{i=1}^n$ 在 Γ 上的投影为

$$\hat{\Gamma}\left(\{(x_i,y_i)\}_{i=1}^n\right) = \left\{\{\gamma(x_1,y_1),\gamma(x_2,y_2),\ldots\gamma(x_n,y_n)\} \,\middle|\, \gamma \in \Gamma\right\}$$

显然，如果损失函数 ℓ 是一个二分损失函数，那么 $\hat{\Gamma}$ 是一个 n 维二分向量集合，且无论集合 Γ 是否有限，$\hat{\Gamma}$ 的元素个数都不超过 2^n。此外，任意 $\hat{f} \in \widehat{\mathcal{F}}$ 的经验误差完全由 $\hat{\Gamma}$ 决定，这是因为集合 Γ 的元素和类空间 $\widehat{\mathcal{F}}$ 的元素一一对应。

其实，正如上一节所说，霍夫丁不等式可以简单地推广到一致成立的情况。不过，这个推广只适用于有限的假设空间 $\mathcal{F}$ 中。在无限的假设空间中，霍夫丁不等式要么几乎不成立，要么上界被无限推高。然而，刚才的分析表明，任意 $\hat{f} \in \widehat{\mathcal{F}}$ 的经验误差可以由有限样本集完全决定，而与类空间 $\widehat{\mathcal{F}}$ 是否无限无关。因此，很有可能存在一种分析方法可以通过有限样本集来描述无限的情况。唯一的问题是，期望误差 E 不能像经验误差 E_n 那样只依赖有限样本集，而需要遍历整个空间 $\mathcal{X}\times\mathcal{Y}$。为此，一种叫做 Symmetrization 的技术[65,66]起到了关键作用。这个技术假设除了样本集 $\{(x_i,y_i)\}_{i=1}^n$，还存在另一个假想样本集 $\{(x_i',y_i')\}_{i=1}^n$。这个假想样本集也是从 $\mathcal{X}\times\mathcal{Y}$ 中采样得到的。值得注意的是，这个假想样本集并不要求实际存在，它仅限于理论分析。这个 Symmetrization 技术表明，对于任意实数 $\varsigma > 0$，且 $n\varsigma^2 \geq 2$，有以下不等式

$$\Pr\left(\sup_{\hat{\gamma}\in\hat{\Gamma}}\left(E(\hat{\gamma})-E_n(\hat{\gamma})\right)>\varsigma\right)\le 2\Pr\left(\sup_{\hat{\gamma}\in\hat{\Gamma}}\left(E'_n(\hat{\gamma})-E_n(\hat{\gamma})\right)>\frac{\varsigma}{2}\right) \tag{2.7}$$

其中，E'_n被定义为假想样本集$\{(x'_i,y'_i)\}_{i=1}^n$的经验误差。

由式（2.7）知，经验误差和期望误差之间的一致收敛上界能够被两个有限集合所刻画，从而避免了对无限集的遍历。这个结论正是得益于假想样本集的引入。不难发现，根据霍夫丁不等式（2.1），式（2.7）不等号右边式子有以下不等式

$$\begin{aligned}&\Pr\left(\sup_{\hat{\gamma}\in\hat{\Gamma}}\left(E'_n(\hat{\gamma})-E_n(\hat{\gamma})\right)>\frac{\varsigma}{2}\right)\le\sum_{\hat{\gamma}\in\hat{\Gamma}}\Pr\left(E'_n(\hat{\gamma})-E_n(\hat{\gamma})>\frac{\varsigma}{2}\right)\\&=\sum_{\hat{\gamma}\in\hat{\Gamma}}\Pr\left(\frac{1}{n}\sum_{i=1}^n\left(\hat{\gamma}\left(x'_i,y'_i\right)-\hat{\gamma}(x_i,y_i)\right)-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(\hat{\gamma}\left(x'_i,y'_i\right)-\hat{\gamma}(x_i,y_i)\right)\right]>\frac{\varsigma}{2}\right)\\&\le\left|\hat{\Gamma}\right|\exp\left(-\varsigma^2\frac{n}{2}\right)=2^{2n}\exp\left(-\varsigma^2\frac{n}{2}\right)=\exp\left(\left(2\log 2-\frac{\varsigma^2}{2}\right)n\right)\end{aligned} \tag{2.8}$$

该式子所给出的上界并不是一个好的上界。由于$2\log 2\approx 1.39$，几乎任何一个合理的ς都会使得上界的指数部分为正数，从而使得上界随着n的增大而迅速增长。也就是说，随着样本量的增大，上界反而越差。此外，这个上界还产生一个荒谬且无意义的结论。当$n=0$时，这个上界达到最好，且上界成立的概率小于等于 1。

之所以出现上述情况，是因为上界包含$\hat{\Gamma}$的最大元素个数2^{2n}。其实，由于$\hat{\Gamma}$依赖于Γ、$\{(x_i,y_i)\}_{i=1}^n$和$\{(x'_i,y'_i)\}_{i=1}^n$，$\hat{\Gamma}$的复杂度往往达不到2^{2n}。比如，若$\hat{\Gamma}$所对应的空间$\mathcal{F}$是一类与坐标纵轴平行的函数，那么任意两个样本点的二分输出情况最多有三种，即$\{(0,0),(1,1),(0,1)\}$。根据这个现象，n个样本点在空间$\hat{\Gamma}$中所产生的二分输出情况最多为

$$S_{\hat{\Gamma}}(n)=\sup_{\{(x_i,y_i)\}_{i=1}^n}\left|\hat{\Gamma}\left(\left\{\left(x_i,y_i\right)\right\}_{i=1}^n\right)\right|$$

此外，Vapnik 和 Chervonenkis 还指出[64]，如果$\hat{\Gamma}$的 VC 维$d_{\hat{\Gamma}}$被定义为满足$S_{\hat{\Gamma}}(m)=2^m$的最大整数m，那么对于任意正整数n，有以下不等式

$$S_{\hat{\Gamma}}(n)\le\sum_{i=0}^{d_{\hat{\Gamma}}}\binom{n}{i} \tag{2.9}$$

于是，将式（2.9）的$S_{\hat{\Gamma}}(2n)$代入到式（2.8）中的$|\hat{\Gamma}|$后，有概率$1-\delta$保证以下上界成立

$$E(\hat{\gamma})\le E_n(\hat{\gamma})+\sqrt{\frac{2}{n}\log 2+\frac{2d_{\hat{\Gamma}}}{n}\log\left(\frac{2ne}{d_{\hat{\Gamma}}}\right)+\frac{2}{n}\log\frac{1}{\delta}} \tag{2.10}$$

其中，δ被定义如下

$$\delta=2\left(\frac{2ne}{d_{\hat{\Gamma}}}\right)^{d_{\hat{\Gamma}}}\exp\left(-\varsigma^2\frac{n}{2}\right)$$

2.3 最大均值差异

迁移学习任务往往需要度量两个分布之间的相似性。根据概率论，对于任意的连续函数ϕ，相等的波莱尔概率分布p和q均满足以下等式

$$\left|\int_{z\in\mathcal{Z}} p(z)\phi(z)dz-\int_{z\in\mathcal{Z}} q(z)\phi(z)dz\right|=0$$

如果函数空间Φ是由所有的连续函数$\phi:\mathcal{Z}\to\mathbb{R}$构成的，那么由于连续函数$\phi$和$-\phi$是成对出现在$\Phi$中的，上式可以省略掉绝对值符号并得到以下不等式

$$\left|\int_{z\in\mathcal{Z}} p(z)\phi(z)dz-\int_{z\in\mathcal{Z}} q(z)\phi(z)dz\right|\le\sup_{\phi\in\Phi}\left(\int_{z\in\mathcal{Z}} p(z)\phi(z)dz-\int_{z\in\mathcal{Z}} q(z)\phi(z)dz\right)$$

其中小于号右边式子即为最大均值差异的定义式。它可以简写成以下形式

$$\mathrm{MMD}\left[\Phi,p,q\right]:=\sup_{\phi\in\Phi}\left(\mathbb{E}_{p(z)}\left[\phi(z)\right]-\mathbb{E}_{q(z)}\left[\phi(z)\right]\right) \tag{2.11}$$

不难发现，当最大均值差异的值为0时，两个波莱尔分布相互对齐。因此，为了使两个分布的差异最小化，算法可以优化最大均值差异。然而，式（2.11）的形式不利于实际算法的设计。首先，由于连续函数Φ可能是无限大的，算法不能在有限时间内遍历所有的连续函数$\phi\in\Phi$。其次，由于事先无从得知分布p和q，算法在评估分布差异时只能依赖于采样的数据。

为了避免遍历函数空间，算法将所有数据映射到再生核希尔伯特空间中。如果$\mathcal{H}$是一个完备内积空间，使得其中的函数ϕ能够把非空紧集$\mathcal{Z}$映射到实数空间$\mathbb{R}$中，并且对于所有的$z\in\mathcal{Z}$，空间$\mathcal{H}$存在连续的线性点估计函数映射$\phi\to\phi(z)$，那么$\mathcal{H}$是一个再生核希尔伯特空间，并且其中的$\phi(z)$存在内积形式

$$\phi(z)=\langle\phi,\psi(z)\rangle_{\mathcal{H}} \tag{2.12}$$

其中，函数ψ把z投影到空间$\mathcal{H}$中的一点。同时，两个特征的正定核有以下内积形式

$$k(z,z'):=\langle\psi(z),\psi(z')\rangle_{\mathcal{H}} \tag{2.13}$$

由公式（2.12）知，当万有再生核希尔伯特空间$\mathcal{H}$的内积范数收敛于一个单位球Φ时，式（2.11）所定义的最大均值差异有以下展开

$$\begin{aligned}\mathrm{MMD}\left[\Phi,p,q\right]&=\sup_{\|\phi\|_{\mathcal{H}}\le1}\left(\mathbb{E}_{p(z)}\left[\phi(z)\right]-\mathbb{E}_{q(z)}\left[\phi(z)\right]\right)\\&=\sup_{\|\phi\|_{\mathcal{H}}\le1}\left(\mathbb{E}_{p(z)}\left[\langle\phi,\psi(z)\rangle_{\mathcal{H}}\right]-\mathbb{E}_{q(z)}\left[\langle\phi,\psi(z)\rangle_{\mathcal{H}}\right]\right)\\&=\sup_{\|\phi\|_{\mathcal{H}}\le1}\langle\phi,\mu_p-\mu_q\rangle_{\mathcal{H}}=\left\|\mu_p-\mu_q\right\|_{\mathcal{H}}\end{aligned} \tag{2.14}$$

其中，$\mu_p = \mathbb{E}_{p(z)}[\psi(z)]$和$\mu_q = \mathbb{E}_{q(z)}[\psi(z)]$分别代表$\psi(z)$关于分布$p$和$q$的期望值。这个式子表明，在映射函数$\psi$的帮助下，即使没有遍历所有的连续函数$\phi \in \Phi$，两个波莱尔分布之间的差异依然可以求得。从向量空间的角度看，由于函数空间Φ被定义为单位球，任意连续函数$\phi \in \Phi$都可以看作模为 1 的单位向量。虽然这些向量能够通过内积操作扭曲函数ψ的输出图像，但在任意两个分布上计算函数ψ的均值差异时，这些扭曲所造成的影响会互相抵消。这是因为在再生核希尔伯特空间中，函数ψ与单位向量ϕ是线性相关的，根据柯西-施瓦兹不等式（Cauchy-Schwarz inequality），单位向量ϕ可以单独求范数，从而使得均值差异的系数为 1。

不过，公式（2.14）对分布差异的经验评估是有偏的，因此一般转而考虑以下式子

$$\begin{aligned}\mathrm{MMD}^2\left[\Phi,p,q\right] &= \langle \mu_p-\mu_q, \mu_p-\mu_q\rangle_{\mathcal{H}} = \langle \mu_p,\mu_q\rangle_{\mathcal{H}} + \langle \mu_p,\mu_q\rangle_{\mathcal{H}} - 2\langle \mu_p,\mu_q\rangle_{\mathcal{H}} \\ &= \mathbb{E}_{p(z)}\left[\mathbb{E}_{p(z')}\left[\langle \psi(z),\psi(z')\rangle_{\mathcal{H}}\right]\right] + \mathbb{E}_{q(z)}\left[\mathbb{E}_{q(z')}\left[\langle \psi(z),\psi(z')\rangle_{\mathcal{H}}\right]\right] \\ &\quad -2\mathbb{E}_{p(z)}\left[\mathbb{E}_{q(z')}\left[\langle \psi(z),\psi(z')\rangle_{\mathcal{H}}\right]\right]\end{aligned}$$

把公式（2.13）代入到上式中，上式有以下无偏经验评估

$$\begin{aligned}\mathrm{MMD}^2\left[\Phi,\mathbb{Z}_p,\mathbb{Z}_q\right] = &\frac{1}{n_p(n_p-1)}\sum_{i\neq j}^{n_p} k(z_i,z_j) + \frac{1}{n_q(n_q-1)}\sum_{i\neq j}^{n_q} k(z'_i,z'_j) \\ &-\frac{2}{n_p n_q}\sum_{i,j=1}^{n_p,n_q} k(z_i,z'_j)\end{aligned} \tag{2.15}$$

其中，$\mathbb{Z}_p = \{z_1, z_2, ...\}$和$\mathbb{Z}_q = \{z'_1, z'_2, ...\}$分别为服从分布$p$和$q$的数据集，$n_p$和$n_q$分别为数据集$\mathbb{Z}_p$和$\mathbb{Z}_q$的样本数量。这个式子表明，即使无法得知真实分布$p$和$q$，两个分布之间的差异依然可以从它们各自的数据集中评估出来。

由公式（2.15）知，不同的核函数所估计的分布差异是不同的，因此核函数的选择是很有必要的。主流的核函数有线性核函数、多项式核函数和高斯核函数。其中，线性核函数能够对两个输入特征执行内积操作，并得到一个简单的线性超平面，但是仅适用于线性可分的数据集。多项式核函数能够拟合出复杂的分隔超平面，但是较多的超参数容易造成模型求解困难。高斯核函数能够通过泰勒展开将特征映射到无限维空间中，而且参数选择和模型求解均比多项式核函数容易，因此，它被广泛用于各种核方法中，尤其是最大均值差异。

最近，一种基于多核的最大均值差异被提出[67]。在保持单核假设[11]的基础上，它还假设多个核函数的线性组合能够得到最优的核函数。这个思想被后续大量工作所继承，其中最著名的是 DAN[35]。

2.4 多任务学习

早期，机器学习模型都是基于单个任务来设计的。在服从原有任务的假设前提下，这些模型都表现良好。然而，这些模型并不能很好地适配到其他目标任务中。其主要原因在于这种单任务模型可能会忽略一些能够提升目标任务的潜在信息。也就是说，单任务模型往往会过度地适配原有任务，从而失去对其他目标任务的推广和泛化能力。

为了让模型更具泛化性，目标函数的设计可以加入一些约束，从而使模型的参数空间适当缩小，避免了模型朝过拟合方向发展。比如，L_1 正则项可以约束模型参数倾向于稀疏性，从而达到特征选择的效果。L_2 正则项可以约束模型参数保持较小的值，从而达到抗干扰的效果。考虑到 L_1 和 L_2 正则项的先验分布分别为拉普拉斯分布和高斯分布，因此这两个正则项可以保持模型的参数空间具备对应的先验知识。总的来说，通过权衡主任务和辅助任务中的信息，模型能够具备更好的泛化性。

传统的多任务学习主要关注两点。第一点是通过范数正则项使模型在任务之间具备稀疏性。第二点是对多任务之间的关系进行建模。

基于第一点发展起来的方法是块稀疏正则化（Block-Sparsity Regularization）。这个方法首先假设所有任务的参数空间是相同的，并为每个任务定义一个表征参数的列向量。然后，所有任务的列向量组成一个参数矩阵，其中每一行代表不同任务的同一个特征，每一列代表不同任务的参数。此时，如果假设所有任务只共享部分参数[68]，即参数矩阵中只有少数几行为非零，那么为了找到这少数几行的非零参数，一般会引入 L_1 正则项使参数矩阵有尽可能多的全零行。具体流程是先对参数矩阵的每一行计算 L_q 范数并得到一个范数向量，接着对这个范数向量计算 L_1 正则项，从而强制向量中的大部分元素为 0。这种使用多个范数约束的方法一般被称为混合 L_1/L_q 正则化，其中 L_q 正则化可根据实际需要设置不同的约束[69,70]。

然而，这种基于特征的方法只适用于多个联系紧密的任务中。为此，不少研究工作基于第二点发展了各种聚类方法，从而能够使用某种先验知识来表明不同任务之间的相关性。这些聚类方法可以惩罚参数矩阵中的列向量及其方差[71]，也可以驱使支持向量机逼近多个任务的均值模型[72]。此外，一些多任务场景虽然没有聚类假设，但却可能共享一个结构。为了找到这种结构，最简单的方法是将块稀疏正则化扩展到树结构和图结构上[73,74]。不过，最常用的方法是在贝叶斯网络中引入模型参数先验，从而使不同任务的参数互相逼近[75]。

目前，多任务学习的思想已经深刻融入到深度学习模型的训练中。这体现在神经网络参数的硬共享和软共享上。

硬共享参数的神经网络允许多个任务共用部分参数。在机器视觉中，这种神经网络的卷积层会被所有任务所共享，而全连接层则是任务专用的。通过对所有全连接层添加矩阵先验[76]或最大均值差异约束[3,35]，神经网络可以学习任务之间的关系。然而，对于非机器视觉问题，这个方法过于依赖预先设定的共享结构。因此，一种称作完全自适应共享方法应运而生[77]。这个方法会事先准备一个网络骨架，然后自底向上对相似任务进行分组并增加网络分支。不过，这种方法不允许模型学到高层次的任务关系。考虑到硬共享参数容易造成两种极端，因此不少工作转而研究软共享参数。

基于软共享参数的方法一般会为每个任务分配一个神经网络。每个神经网络的参数都是任务专用的。为了在多个独立任务之间建立关系，常用的方法是对网络中的所有参数使用相似性约束，如L_2距离正则化[78]和迹正则化[79]。在机器视觉中，所有神经网络的卷积层会被嵌入相应的正则项，从而约束参数的相似性。除此之外，通过对隐层的输出结果进行线性组合，任务之间的信息可以互相流通，从而有助于建立任务关系。这方面最具代表性的工作是十字绣网络（Cross-Stitch Networks）[80]。这种网络会保留中间部分用于各个任务，其余部分则使用十字绣单元对所有隐层特征进行线性加权并重新分配到各个任务所对应的网络中。后来，这种网络被扩展成了水闸网络[81]，并能够学习每一层的共享子结构和序列表征。

多任务学习的理论观点比较多，主要集中在这几个层面：（1）隐式数据增加机制。多任务学习一般会接收更多来源的训练样本。这些样本都带有不同的噪声模式。当这些噪声被带入模型时，那些与噪声相关的表征将被忽略，从而保证了模型的泛化性。（2）注意力集中机制。当训练样本具备小规模和高维度的特点时，模型通常很难判断哪些特征是专用的还是通用的。不过，多任务学习可以使模型关注最通用的特征上。（3）特征窃听机制。由于不同任务的特征交互方式是不同的，每个任务对同一种特征的学习难度是不同的。如果某种特征的提取对于任务 A 是容易的，而对于任务 B 是困难的，那么多任务学习允许任务 A 窃听任务 B 所习得的特征。（4）表示偏置机制。多任务学习倾向于把所有任务看作同一类任务。因此，基于多任务学习的模型一般都是从泛化程度较高的假设空间中筛选出来的。（5）正则化机制。多任务学习往往自带归纳偏置功能，避免了模型对随机噪声的过拟合问题。

2.5 卷积神经网络

深度学习的诞生离不开对人脑认知机理的研究，尤其是生物视觉原理的研究。1959年，Hubel 和 Wiesel 发现视觉系统中的信息是由可视皮层分级处理的[82]。原始的像素信号首先会通过瞳孔并在视网膜上被转化为一系列的神经冲动。这些神经冲动信号会被大脑皮层中的某些视觉细胞所捕获并进行边缘检测。然后这些检测结果会进一步上传至大脑皮层中的低级区域进行形状检测。最后这些形状检测结果会被大脑的高级区域判定并赋予某种已知的物体概念。

在卷积神经网络诞生之前，人们苦于寻找一种鲁棒的算法用于银行支票金额的识别上。当时，不少算法工程师编写了一系列的规则用于判断支票上的手写数字。然而，由于不同客户写出来的数字风格是不一样的，基于规则的做法有很高的识别错误率。后来受生物视觉原理的启发，LeCun[83]等人设计了 LeNet-5 卷积神经网络用于手写数字识别任务上，使得识别错误率一下子降了下来。从此，人们从大量的规则编写工作中被解放出来，转而对卷积神经网络展开研究。和传统神经网络一样，LeNet-5 也能使用反向传播算法[84]训练。它包含两个卷积层、两个池化层和三个全连接层。其中，每个卷积层都会紧跟一个池化层，并且最后一个全连接层会输出每个分类的评估值。由于卷积神经网络能够逐层提取出原始图像的有效表征，LeNet-5 能够以极少的预处理来归纳视觉信息中的规律。

然而，受制于当时小规模的训练数据和落后的计算能力，LeNet-5 在复杂问题上的表现并不理想。后来，人们设计了很多方法，试图克服卷积神经网络的训练难题。其中，最著名的一种网络结构是 AlexNet[85]。与 LeNet-5 类似，这个网络结构也是使用卷积层和池化层来搭建的。不过，AlexNet 的层次结构更深，而且使用非线性激活函数 ReLU[86]和 Dropout[87]方法来缓解梯度消失和过拟合问题。由于 AlexNet 取得了卓越的效果，各界都掀起了一股卷积神经网络的研究热潮。各种改进版本的卷积神经网络层出不穷，其中最著名的是 VGGNet[88]、GoogleNet[89]和 ResNet[90]。从网络的整体结构来看，卷积神经网络的一个发展趋势是设计更深的网络结构。这是因为更深的网络可以利用更充足的非线性单元逼近目标函数，同时得出更好的特征表示。不过，太深的网络会使缓解下来的梯度消失问题再次变得严峻起来，于是人们又设计了各种基于旁路结构的网络，如 ResNet。此外，太深的网络还会导致训练难度大大增加，于是人们发明了局部响应归一化（Local Response Normalization）[85]和批归一化（Batch Normalization）[91]。

卷积神经网络的成功离不开有效特征的提取和降维，而有效特征的提取和降维分别依赖于卷积层和池化层的优良设计。

卷积层会预先设定滑动步幅大小和滑动窗口大小，并对窗口中的所有权重初始化。然后滑动窗口会遍历整个输入特征图，并计算权重和每个特征子图的内积。最后，这些内积结果会按遍历顺序整合成一张特征图。若卷积层的输入特征是多通道的，那么卷积层会对所有通道的输出特征图执行一次逐元素加和操作。因此，从计算方式的角度看，卷积层可被视为一个关于通道的全连接层，只不过所有连接所对应的特征和权重都以矩阵的形式存在。特别地，当滑动步幅大小和滑动窗口大小均被设定为 1 时，卷积层直接对所有通道的输入特征图加权求和。此外，卷积操作有三种模式。‘valid’模式强制滑动窗口处于整个特征图的里面，‘same’模式强制滑动窗口的中心遍历整个特征图，‘full’模式允许滑动窗口在保持与特征图相交的前提下有最大的滑动空间。前两种模式多用于前向过程的特征提取中，后一种多用于后向过程的梯度传播中。

池化层的计算步骤与卷积层大致相同。不过，在遍历整个输入特征图时，池化层会通过某种操作把所有特征子图都转化成一个表征值。根据操作的不同，池化层可大致分为三种，分别为最大池化层、平均池化层和随机池化层。最大池化层能够从特征子图中选择最大值作为输出，平均池化层能够输出特征子图中所有像素的平均值，随机池化层首先根据数值大小为特征子图中的像素点赋予概率，然后依概率随机选择一个像素点作为输出。这些不同类型的池化层都能够保持平移操作等不变性，但保持不变性的主要原因是不一样的。最大池化层能够减少因卷积层参数所造成的均值偏移，有效保留特征图中的纹理信息。平均池化层能够减少因邻域大小受限所造成的方差，有效保留特征图中的背景信息。随机池化层的不变性解释介于最大池化层和平均池化层。

然而，传统的卷积网络只能用于欧氏数据，如语音数据和图像数据。因此，图卷积网络[92]被设计用于具备图谱结构的非欧数据。图卷积网络有三种矩阵。第一种是邻接矩阵，定义了每个节点与图中其余节点的关系。第二种是特征矩阵，定义了每个节点所持有的特征。第三种是参数矩阵，定义了上一层特征到下一层特征的变换操作。这三个矩阵的相乘即可得到一次图卷积的结果。这个结果会作为下一层的特征矩阵输入到新一轮的图卷积中，直到网络的输出层。从计算流程看，一次图卷积可看作两次全连接。第一次会根据图结构对邻居节点的特征进行加和，第二次则对加和后的特征执行多个不同的线性加权操作。

2.6 Caffe 深度学习框架

Caffe（Convolutional Architecture for Fast Feature Embedding）是一个兼顾表达、速度和模块化的深度学习框架[93]。这个框架最早由作者贾扬清所创建，后由伯克利人工智能研究所（BAIR）和社区贡献者陆续完善。

Caffe 拥有一个符号表达系统，鼓励用户在其上进行开发和创新。在 Caffe 中，模型和优化都被定义为配置文件，从而大大减少了用户的代码编写工作。由于这个符号表达系统能够有效解耦前后端的工作，Caffe 提供一个设备标记即可决定框架运行在 CPU 还是 GPU 环境中，从而方便用户部署到集群或移动设备中。Caffe 拥有可扩展性的代码组织形式，能够促使开发者为其添砖加瓦。由于这些开发者的贡献，Caffe 框架能够跟踪代码和模型的最新进展。Caffe 拥有极高的处理速度，是研究实验和工业部署的完美选择。Caffe 能够借助 NVIDIA K40 GPU 在一天内处理 6 千万张图片。这意味着学习一张图片仅需 4 毫秒，而推断一张图片仅需 1 毫秒。随着新运行库和新硬件的推出，Caffe 能够做得更快。此外，Caffe 还拥有良好的社区。在这个社区中，有各种学术研究团队、初创企业和各个领域的工业开发团队为 Caffe 注入新的活力。

Caffe 有四个核心模块，分别为 Blob、Layer、Net 和 Solver。Blob 负责数据的存储、交互的处理，并提供统一的数据访问接口。Layer 负责对数据进行转换，是神经网络的核心。Net 负责统筹整个计算图，是一系列 Layer 的集合。Solver 负责控制计算图的计算过程、训练周期和测试周期，并且必要时保存计算图的各类参数。

Blob 是一个存有前向特征数据和后向梯度数据的张量。对于二维图像数据来说，Blob 通常是一个格式为（N,C,H,W）的 4 维张量，其中 N 代表一个 Batch 的大小，C 代表图像的通道数，H 和 W 分别代表图像的高度和宽度。需要说明的是，一个 Batch 所能设定的大小与内存显存容量有关。一般情况下，一个 Batch 越大，网络的数据吞吐量也会越大。不过，更大的 Batch 虽然有助于模型的收敛，但可能会影响模型的泛化性。

Layer 是一个数据转换节点。根据转换的类型，Layer 可以是卷积层、池化层或损失层。将这些特化的 Layer 任意组合可以得到任意的网络计算图。Layer 主要包含三个操作。创建操作首先根据输入 Blob 的形状（shape）推断参数 Blob 和输出 Blob 的形状，然后根据预设的超参数对参数 Blob 初始化。前向计算操作能够利用特定算法从输入 Blob 中提取特征，并把特征整合到输出 Blob 中。后向计算操作能够根据前向算法逆向考虑输出 Blob 到输入 Blob 的关系，把输出 Blob 的梯度信息分配到输入 Blob 中。

Net 是一个包含多个 Layer 的计算图。这个计算图定义了输入输出和多个 Layer，并按照某种拓扑结构把多个 Layer 连接起来。在 Caffe 中，Net 的输入端可以接受图像数据或序列数据，并且允许 lmdb、h5py 等数据格式。Net 的输出端允许用户根据不同的任务选择不同的损失函数。比如，分类任务会选择 softmax 损失函数，回归任务会选择均方误差函数。此外，Caffe 允许 Net 有多个输出端，从而运行多任务学习。

Solver 在 Caffe 中充当指挥官角色。Net 的初始化、前向后向计算、权重参数优化和模型导入导出都由 Solver 控制，保证了模型训练和测试流程的准确无误。由于权重参数的优化方式比较多，Solver 提供了 6 种优化模式以供用户选择。这些模式分别是 Stochastic Gradient Descent（SGD）、AdaDelta、Adaptive Gradient（AdaGrad）、Adam、Nesterov's Accelerated Gradient（Nesterov）和 RMSprop。

2.7 本章小结

本章首先介绍了霍夫丁不等式的数学形式，并通过相关例子来阐释该不等式对于机器学习问题的重要性。根据这个不等式，机器学习算法的误差收敛性可以从大数定律中得到分析，并通过 VC 泛化上界推广到一致收敛性。因此，跨领域适应算法的收敛性分析也会用到霍夫丁不等式和 VC 泛化上界。然后，引入最大均值差异，并分析跨领域适应算法的分布对齐问题。在跨领域适应算法中，一个好的分布对齐方案往往意味着一个好的迁移学习策略。引入多任务学习的原因是跨领域适应算法一般要优化两个目标。这种多目标的优化任务属于多任务学习的范畴。接着，由于跨领域适应算法一般需要提取足够多的高层特征来对齐分布，卷积神经网络的引入是一件很自然的事情。最后，考虑到算法实现的高效性和简便性，Caffe 框架将作为后续算法的底层开发平台。

第三章　无领域信息的跨领域自适应

3.1 跨领域自适应的有限制风险界

在跨领域自适应场景中，每一个源领域S都有自己的总体分布$\pi_S(X)$和后验分布$p_S(Y|X)$，其中X和Y分别为源领域S的样本空间和标签空间。这些源领域的产生机制可以描述为$\sigma \to S \to (X,Y)$，其中参数σ控制了源领域的产生概率$\pi_\sigma(S)$。与此同时，目标领域的产生机制可以描述为$T \to (X,Y)$,其中参数T控制了目标领域的总体分布$\pi_T(X)$和后验分布$p_T(Y|X)$。为了让多个源领域后验分布都对齐到同一个目标领域上，模型需要借助一个风险上界来构造目标函数。这个风险上界由两部分组成。第一部分为模型经验后验分布$q(Y|X)$与源领域期望后验分布$\mathbb{E}_{\pi_\sigma(S)}[p_S(Y|X)]$的差异。第二部分为源领域期望后验分布$\mathbb{E}_{\pi_\sigma(S)}[p_S(Y|X)]$和目标领域后验分布$p_T(Y|X)$的差异。如果给定任意领域$D \in \{S,T\}$后，任意两个后验分布$q(Y|X)$和$p(Y|X)$的差异有定义

$$e(q,p|\pi_D)=\mathbb{E}_{\pi_D(x)}\left[\left|q(Y|X=x)-p(Y|X=x)\right|\right] \tag{3.1}$$

并且有如下定理

定理 1（跨领域自适应的有限制风险界） 如果给定目标领域T后，模型经验后验分布$q(Y|X)$与目标领域后验分布$p_T(Y|X)$之间的差异可以定义为$e(q,p_T|\pi_T)$，那么同理得$e(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]|\mathbb{E}_{\pi_\sigma(S)}[\pi_S])$、$e(q,p_T|\mathbb{E}_{\pi_\sigma(S)}[\pi_S])$和$e(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]|\pi_T)$，并有不等式

$$\begin{aligned} e\left(q,p_T|\pi_T\right) \le e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]|\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)+\min\Big\{ & \\ e\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T|\pi_T\right)+e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]|\left|\pi_T-\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right|\right), & \\ e\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T|\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)+e\left(q,p_T|\left|\pi_T-\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right|\right)\Big\} & \end{aligned} \tag{3.2}$$

证明：根据文献[2]中定理 1 的证明过程，该定理的详细证明见附录 A。

根据定理 1，若把式（3.2）的$\min\{\cdot,\cdot\}$简写为Δ，则有风险上界三角形，见图 3-1。

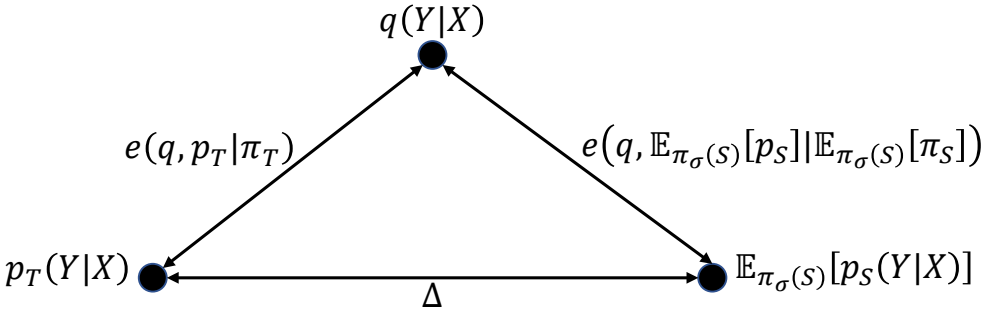


图 3-1 有限制的风险上界三角形

Figure 3-1 The limited triangle of upper risk bound

由图 3-1 知，若源领域期望后验分布 $\mathbb{E}_{\pi_{\sigma(S)}}[p_S(Y|X)]$ 与目标领域后验分布 $p_T(Y|X)$ 的差异 Δ 足够小，那么通过最小化模型经验后验分布 $q(Y|X)$ 和源领域期望后验分布 $\mathbb{E}_{\pi_{\sigma(S)}}[p_S(Y|X)]$ 的差异，$q(Y|X)$ 可以近似表达 $p_T(Y|X)$。然而，在无监督的情形下，算法对分布差异 Δ 的优化是有限的。这是因为 Δ 的形式只允许模型经验后验分布逼近源领域期望后验分布，从而无法用上目标领域中的无标签数据。

3.2 跨领域自适应的无限制风险界

其实，通过对齐源领域期望总体分布 $\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S(X)]$ 和目标领域总体分布 $\pi_T(X)$，分布差异 Δ 也可以得到缩小。然而受限于式（3.1）所定义的后验分布差异，分布差异 Δ 没有得到足够的参数空间来优化。因此，后验分布差异需要重新定义。如果给定任意领域 $D \in \{S,T\}$ 和任意转换函数 g 后，任意两个后验分布 $q(Y|g(X))$ 和 $p(Y|g(X))$ 的差异有定义

$$e_g(q,p|\pi_D)=\mathbb{E}_{\pi_{D(g(x))}}\left[\left|q(Y|g(x))-p(Y|g(x))\right|\right]$$

根据定理 1，有以下推论。

推论 2（跨领域自适应的无限制风险界） 若给定目标领域 T 和转换函数 g 后，模型经验后验分布 $q(Y|g(X))$ 与目标领域后验分布 $p_T(Y|g(X))$ 的差异定义为 $e_g(q,p_T|\pi_T)$，那么同理得 $e_g(q,\mathbb{E}_{\pi_{\sigma(S)}}[p_S]|\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S])$、$e_g(q,p_T|\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S])$ 和 $e_g(q,\mathbb{E}_{\pi_{\sigma(S)}}[p_S]|\pi_T)$，并有不等式

$$\begin{aligned} e_g\left(q,p_T|\pi_T\right) \le e_g\left(q,\mathbb{E}_{\pi_{\sigma(S)}}[p_S]|\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S]\right)+\min\Big\{ & \\ e_g\left(\mathbb{E}_{\pi_{\sigma(S)}}[p_S],p_T|\pi_T\right)+e_g\left(q,\mathbb{E}_{\pi_{\sigma(S)}}[p_S]|\left|\pi_T-\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S]\right|\right), & \\ e_g\left(\mathbb{E}_{\pi_{\sigma(S)}}[p_S],p_T|\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S]\right)+e_g\left(q,p_T|\left|\pi_T-\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S]\right|\right)\Big\} & \end{aligned} \tag{3.3}$$

根据推论 2，若把式（3.3）的 $\min\{\cdot,\cdot\}$ 简写为 Δ，则有风险上界三角形，见图 3-2。

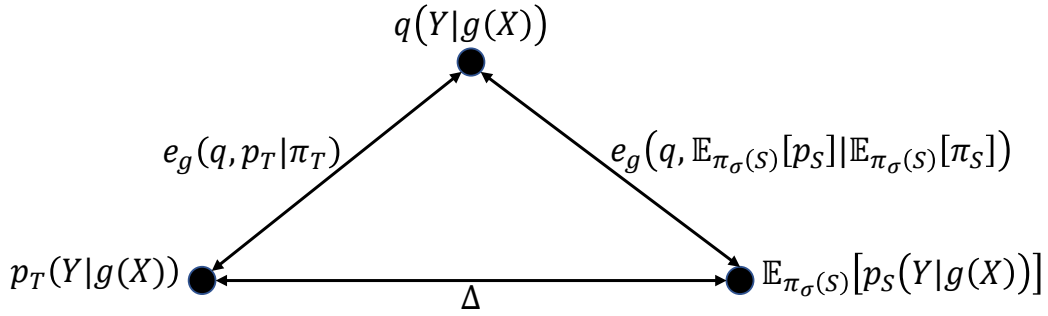


图 3-2 无限制的风险上界三角形

Figure 3-2 The unlimited triangle of upper risk bound

由图 3-2 可知，源领域与目标领域之间的后验分布差异 Δ 可以通过对齐源领域期望总体分布 $\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S(g(X))]$ 和目标领域总体分布 $\pi_T(g(X))$ 来优化。于是，无监督跨领域适应算法有两个主要任务：1）对齐模型经验后验分布 $q(Y|g(X))$ 和源领域期望后验分布

$\mathbb{E}_{\pi_{\sigma(S)}}[p_S(Y|g(X))]$；2）对齐源领域期望总体分布$\mathbb{E}_{\pi_{\sigma(S)}}[\pi_S(g(X))]$和目标领域总体分布$\pi_T(g(X))$。需要说明的是，推论 2 还额外表明算法可以通过对齐模型经验后验分布$q(Y|g(X))$和目标领域后验分布$p_T(Y|g(X))$来进行优化。然而，在无监督的情形下，目标领域的数据标签是不可访问的，因此这种做法是不可行的。总之，一个无监督跨领域自适应算法可以这样实现：a）根据任务 1），使用一个已有的有监督模型如神经网络来拟合源领域的有标签数据；b）根据任务 2），使用一个函数来评估源领域和目标领域之间的总体分布差异，并优化这个分布差异评估函数以对齐两个领域的总体分布。

3.3 跨领域自适应的源领域收敛界

根据上一节的结论，跨领域自适应的其中一个任务是拟合源领域的期望分布。由于源领域数据中包含若干个领域，算法只能拟合源领域的经验分布。亦即，算法所评估的分布是有误差的。这个误差可以通过 VC 泛化上界得到分析，因此有以下定理

定理 3（跨领域自适应的源领域收敛界）令$f\circ g$是一个定义在空间$\mathcal{F}\circ\mathcal{G}$中的复合函数，其中空间$\mathcal{F}\circ\mathcal{G}$有 VC 维$d_{\mathcal{F}\circ\mathcal{G}}$。令$E(f,g)\coloneqq\mathbb{E}_{P(g(X),Y,S)}[\mathbb{I}(f\circ g(x)\neq y)]$为源领域的二分期望损失，$\hat{E}(f,g)\coloneqq 1/mn\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{I}(f\circ g(x_i^j)\neq y_i^j)$为$m$个源领域的二分经验损失。如果每个源领域中的所有样本实例服从独立同分布，那么对于任意的$\delta\in(0,1)$，至少有概率$(1-\delta)^2$保证以下上界成立

$$E(f,g)-\hat{E}(f,g)\leq\sqrt{\frac{\log\frac{1}{\delta}}{2m}}+\sqrt{\frac{2}{n}\log 2+\frac{2d_{\mathcal{F}\circ\mathcal{G}}}{n}\log\left(\frac{2ne}{d_{\mathcal{F}\circ\mathcal{G}}}\right)+\frac{2}{n}\log\frac{1}{\delta}} \tag{3.4}$$

证明：该定理的详细证明见附录 B。

这个风险上界受到三个因素的影响。第一个因素为源领域的数量，第二个因素为单个源领域的样本数量，第三个因素为模型的 VC 维。显然，随着前两个数量的增大，源领域的经验损失会逼近期望损失。最后一个因素的情况与传统 VC 理论相关，具体分析请参考文献[64,65]。

3.4 跨领域自适应的类别同态假设

在前三节中，跨领域自适应的基本假设及理论上界已经讨论过。简单地说，源领域的产生机制可以描述为$\sigma\to S\to(X,Y)$，而目标领域的产生机制描述为$T\to(X,Y)$。目前大部分的领域自适应算法只考虑单个源领域到单个目标领域的情形。如果把隐含多个源领域的数据看作一个源领域的数据，并将其用于模型的训练上，那么算法必须假设源领域期望总体分布和目标领域总体分布是相似的。

然而，这种假设在现实世界中几乎不可能成立。不过，现实世界更多地服从另一种假设，即同一类的物体或行为往往有相同的局部特点。比如，自行车和汽车都属于车，而且都有轮子和方向盘。人类和鸟类都属于生物链上的消费者，而且都以有机物为食物。盗窃和盗版都属于违法犯罪，而且都侵犯了别人的合法权益。物品交换和钱物交换都属于贸易，而且买卖双方都得到了各自想要的东西。这些例子表明，不同形式的同类物体或行为隐含了相同的子结构或特点。从领域自适应的角度看，不同领域的同类样本有相同的局部特征，即不同领域之间存在类别同态假设。不同领域对同类样本的影响仅限于额外特征或噪声。总之，类别同态假设可以描述为以下产生机制。

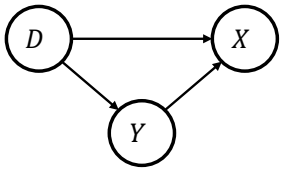


图 3-3 类别同态假设下的产生机制

Figure 3-3 The generation mechanism under category homomorphism assumption

由图 3-3 知，领域 D 影响类别 Y 的分布，类别 Y 影响样本 X 的分布，而领域 D 以额外特征或噪声的形式影响 X 的分布。显然，无论样本 X 属于哪个领域，必定包含类别 Y 的信息，因为类别 Y 主导了样本 X 的产生机制。因此，如果源领域和目标领域均服从上述假设，那么根据图 3-3，在给定类别 c 后，源领域 S 对 X 有以下影响

$$\mathbb{E}_{\pi_\sigma(S)}\left[\pi_S^c(X)\right]=\mu^c(X)+\mathbb{E}_{\pi_\sigma(S)}\left[\xi_S(X)\right] \tag{3.5}$$

其中，$\pi_S^c(X)$ 为给定源领域 S 和类别 c 后 X 的分布，$\mu^c(X)$ 为给定类别 c 后 X 的分布，$\xi_S(X)$ 为源领域 S 的噪声分布。同理，在给定类别 c 后，目标领域 T 对 X 有以下影响

$$\pi_T^c(X)=\nu^c(X)+\xi_T(X) \tag{3.6}$$

其中，$\pi_T^c(X)$ 为给定目标领域 T 和类别 c 后 X 的分布，$\nu^c(X)$ 为给定类别 c 后 X 的另一个分布，$\xi_T(X)$ 为目标领域 T 的噪声分布。这里需要注意的是，$\nu^c(X)$ 与 $\mu^c(X)$ 不一定相同，因为目标领域的超参数可能和源领域的超参数 σ 不同。

为了有效对齐式（3.5）和式（3.6）中的两个分布，根据式（3.3），任意一种类别 c 的无限制风险上界都可修改为以下形式

$$\begin{aligned} e_g\left(q,p_T\mid\pi_T^c\right)\le e_g\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S^c]\right)+\min\Big\{ & \\ e_g\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T\mid\pi_T^c\right)+e_g\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\left|\pi_T^c-\mathbb{E}_{\pi_\sigma(S)}[\pi_S^c]\right|\right), & \\ e_g\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S^c]\right)+e_g\left(q,p_T\mid\left|\pi_T^c-\mathbb{E}_{\pi_\sigma(S)}[\pi_S^c]\right|\right)\Big\} & \end{aligned} \tag{3.7}$$

上式表明，为了将源领域中类别 c 的信息迁移到目标领域中，算法不仅对齐模型后验分

布 $q(Y|g(X))$ 和源领域期望后验分布 $\mathbb{E}_{\pi_\sigma(S)}[p_S(Y|g(X))]$，还要最小化类内分布差异，即

$$\pi_T^c\left(g\left(X\right)\right)-\mathbb{E}_{\pi_\sigma(S)}\left[\pi_S^c\left(g\left(X\right)\right)\right]$$

根据式（3.5）和式（3.6），最小化上式相当于最小化以下式子

$$\nu^c\left(g\left(X\right)\right)-\mu^c\left(g\left(X\right)\right)+\xi_T\left(g\left(X\right)\right)-\mathbb{E}_{\pi_\sigma(S)}\left[\xi_S\left(g\left(X\right)\right)\right] \tag{3.8}$$

显然，当上式等于 0 时，多个源领域和目标领域的类内分布是相互对齐的，并且来自多个领域的噪声被相互抵消。考虑到算法所使用的源领域数量有限，因此式（3.8）中的 $\mathbb{E}_{\pi_\sigma(S)}[\xi_S(g(X))]$ 只能替换为其对应的经验评估形式。在对齐领域类内分布时，这势必导致式（3.8）对齐的效果出现偏差。不过根据源领域收敛上界（3.4），随着源领域数据所包含的领域越多，这个偏差将趋于 0。亦即，随着源领域数量的增加，领域类内分布的对齐效果越好，类别信息的迁移能力就越好。这一点在实验中将得到验证。

总的来说，如果 C 被定义为类别相关的随机变量，那么根据式（3.7），基于类内分布对齐策略的跨领域自适应可以描述为以下示意图。

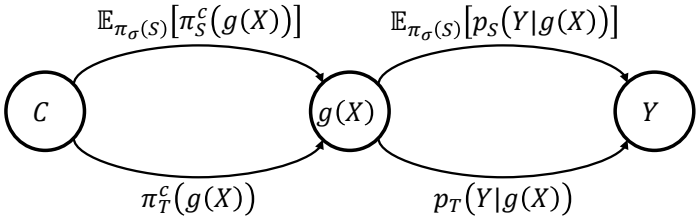


图 3-4 基于类内分布对齐的跨领域自适应

Figure 3-4 Cross-domain adaptation by intra-class distributed alignment

由图 3-4 知，类内分布对齐策略相当于构造了一条可逆通路，使得类别信息被编码到特征中后，这些信息又能够从特征中被解码出来，从而有效保留了类别信息。

3.5 基于类内总体均方离差的策略

根据之前的讨论，如果算法采用总体均方离差对齐类内分布，那么领域自适应的任务是优化特征提取器 $g\in\mathcal{G}$ 和分类器 $f\in\mathcal{F}$，使得以下目标函数最小化。

$$\arg\min_{f\in\mathcal{F},g\in\mathcal{G}}\frac{1}{n_S}\sum_{i=1}^{n_S}\mathcal{L}(f(g(x_i)),y_i)+\alpha\ \text{ICMSD}$$

其中，损失函数 $\mathcal{L}$ 一般选用 softmax 函数或均方误差，n_S 和 n_T 分别为源领域和目标领域的样本数。α 用于控制 ICMSD 对目标函数的影响。ICMSD 正则项用于约束特征提取器 g 的类别同态特性。它有以下定义

$$\text{ICMSD}=\frac{1}{2(n_S+n_T)}\sum_{i=1}^{n_S+n_T}\left\|g(x_i)-u(y_i)\right\|_2^2$$

其中，带有标签c的样本均值被定义为

$$u(c)=\frac{1}{n(c)}\sum_{i=1}^{n_S+n_T}\mathbb{I}(y_i=c)g(x_i)$$

其中，$\mathbb{I}(\cdot)$被定义为指示函数，即当括号内的条件成立时输出 1，否则输出 0。$n(c)$为带有标签c的样本总数，并有以下定义

$$n(c)=\sum_{i=1}^{n_S+n_T}\mathbb{I}(y_i=c)$$

这里需要说明的是，在无监督自适应场景下，目标领域训练样本的标签是由模型临时给定的。此外，算法每次迭代都要重新计算所有类别的$n(c)$和$u(c)$，从而满足目标函数对计算的先决要求。

由于未收敛的模型有可能导致所有类别的$u(c)$过于相似，目标函数额外增加了一个正则项。这个正则项能够将不同类别的$u(c)$分散开，其具体形式为

$$\mathcal{R}=\frac{-1}{2(n_S+n_T)\,|Y|}\sum_{i=1}^{n_S+n_T}\sum_{y\in Y}\mathbb{I}(y_i\neq y)\left\|g(x_i)-u(y)\right\|_2^2$$

因此，算法最终所优化的目标函数为

$$\arg\min_{f\in\mathcal{F},g\in\mathcal{G}}\frac{1}{n_S}\sum_{i=1}^{n_S}\mathcal{L}(f(g(x_i)),y_i)+\alpha\ \mathrm{ICMSD}+\beta\mathcal{R} \tag{3.9}$$

其中，β用于控制$\mathcal{R}$对目标函数的误差贡献。这个目标函数的第一项用于最小化源领域样本的分类误差。第二项用于最小化类内总体均方离差，并约束特征提取器g的类别同态特性。第三项用于最大化类间总体均方离差，保证特征提取器g输出类别明确的特征，减少分类器f的判别难度。

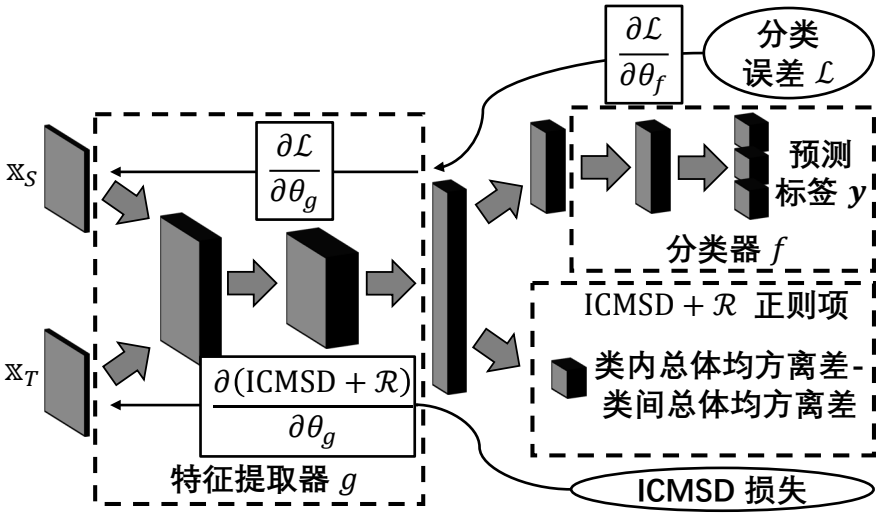


图 3-5 基于类内总体均方离差的神经网络

Figure 3-5 The neural network based on intra-class total mean square deviation

考虑到特征提取器 g 和分类器 f 的形式比较多，一种可行的方案是构建一个包含三个子网络的深度学习模型，如图 3-5 所示。左边的子网络负责承载特征提取器 g 的参数。这些参数会被类内总体均方离差（ICMSD）约束，从而使特征提取器 g 保持类别同态特性。右上角的子网络负责承载分类器 f 的参数。这些参数会被 softmax 损失函数约束，从而使分类器 f 能够对特征提取器 g 的同态特征进行分类。右下角的子网络用于评估特征提取器 g 的同态程度。整个网络将依据式（3.9）的目标函数进行训练。详细的训练过程如下。

Algorithm 1 Cross-domain adaptation with intra-class total mean square deviation

Input：samples $\{x_1, x_2, \ldots, x_{n_S+n_T}\}$ and labels $\{y_1, y_2, \ldots, y_{n_S}\}$

Output：the model parameters θ_f and θ_g

for each iteration I **do**

$Y \leftarrow \{\}$

for each index $i \in \{1, 2, \ldots n_S + n_T\}$ **do**

if $i > n_S$ **then**

$y_i \leftarrow f(g(x_i))$

end if

if $y_i \notin Y$ **then**

$Y \leftarrow Y \cup \{y_i\}$

$u(y_i) \leftarrow 0$

$n(y_i) \leftarrow 0$

end if

$u(y_i) \leftarrow (n(y_i)/(n(y_i)+1))u(y_i) + (1/(n(y_i)+1))g(x_i)$

$n(y_i) \leftarrow n(y_i)+1$

end for

Update θ_f and θ_g using Stochastic Gradient Descent:

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

$$\theta_g \leftarrow \theta_g - \eta \left(\frac{\partial \mathcal{L}}{\partial \theta_g} + \alpha \frac{\partial \mathrm{ICMSD}}{\partial \theta_g} + \beta \frac{\mathcal{R}}{\partial \theta_g} \right)$$

end for

由上述算法可知，前向过程预测目标领域样本的伪标签，并统计各类样本的个数和均值。后向过程使用随机梯度下降算法更新分类器 f 的参数 θ_f 和特征提取器 g 的参数 θ_g。整个优化过程受学习率 η 控制。在更新阶段，θ_f 只受分类器 f 的分类误差更新，而 θ_g 同时受分类误差和两个正则项更新。这是因为特征提取器 g 同时影响分类误差和两个正则项的评估，而分类器 f 只影响分类误差。需要说明的是，由于未收敛的模型预测的伪标签带有噪声，正则项的梯度需要根据分类器的梯度进行放缩，即

$$\frac{\partial \mathrm{ICMSD}}{\partial \theta_g} \leftarrow \frac{\partial \mathrm{ICMSD}}{\partial \theta_g} \cdot \left(\left\| \frac{\partial \mathcal{L}}{\partial \theta_g} \right\|_2 / \left\| \frac{\partial \mathrm{ICMSD}}{\partial \theta_g} \right\|_2 \right), \quad \frac{\mathcal{R}}{\partial \theta_g} \leftarrow \frac{\mathcal{R}}{\partial \theta_g} \cdot \left(\left\| \frac{\partial \mathcal{L}}{\partial \theta_g} \right\|_2 / \left\| \frac{\mathcal{R}}{\partial \theta_g} \right\|_2 \right)$$

3.6 基于类内最大均值差异的策略

与之前的策略相似，如果算法采用最大均值差异对齐类内分布，那么领域自适应的任务是优化特征提取器 $g \in \mathcal{G}$ 和分类器 $f \in \mathcal{F}$，使得以下目标函数最小化。

$$\arg\min_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{1}{n_S} \sum_{i=1}^{n_S} \mathcal{L}(f(g(x_i)), y_i) + \lambda \ \mathrm{ICMMD}^2 \left[\Phi, g, \mathbb{X}_S^c, \mathbb{X}_T^c \right] \tag{3.10}$$

其中，损失函数 $\mathcal{L}$ 一般选用 softmax 函数或均方误差，变量 n_S 和 n_T 分别为源领域和目标领域的样本数。$\mathbb{X}_S^c$ 和 $\mathbb{X}_T^c$ 分别为源领域和目标领域中带有类别 c 的样本子集，变量 n_S^c 和 n_T^c 为给定类别 c 后源领域和目标领域的样本数。超参 λ 用于控制 ICMMD 对目标函数的误差贡献。复合函数 $f(g(x_i))$ 用于评估 x_i 的潜在标签，其功能等价于 $\arg\max_{c \in C} q(Y = c \mid g(x_i))$。

与式（3.9）类似，式（3.10）的目标函数也是用于训练多任务学习模型的。这个函数能够优化特征提取器 g 和分类器 f，并最小化分类误差。与此同时，这个函数也会借助 ICMMD 正则项使多个源领域的类内分布对齐到目标领域上。需要说明的是，如果对多个类别同时计算 ICMMD，那么算法有可能出现内存不足问题。为此，算法会使用轮转法迭代计算每个类别的 ICMMD 正则项。也就是说，如果两个领域都具有相同的类别空间 $\mathcal{C}$，那么对于当前迭代 I，算法评估并优化以下 ICMMD 正则项。

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

最终，所得模型能够把所有样本投影到领域一致的决策空间中，并且能够在决策空间中预测样本特征的类别。此外，考虑到目标领域在无监督场景下缺失样本标签，因此模型为目标领域训练样本预测伪标签。同时，考虑到源领域存在少量领域专用样本，模型也为源领域训练样本预测伪标签。因此，所有领域的训练样本都附带伪标签。

和 ICMSD 一样，ICMMD 也会受到未收敛模型的负面影响。这是因为未收敛模型的预测能力一般很弱，容易导致某些类的伪标签样本过于稀疏。一种可行的方案是对

这些样本进行过采样。不过在样本极度稀疏的情况下，过采样操作就显得没有意义了。因此，算法设计了一套流程用于过采样。这套流程受到阈值τ_1和τ_2控制。详细地说，当目标领域和源领域中附带伪标签c的样本个数均不少于τ_1时，算法分别对$\mathbb{X}_S^c$和$\mathbb{X}_T^c$采样τ_2次。其余情况下，算法直接对$\mathbb{X}_S=\{x_1,x_2,\ldots,x_m\}$和$\mathbb{X}_T=\{\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_n\}$采样$\tau_2$次。此时的ICMMD退化成原始MMD。最终，这套流程都会为ICMMD正则项提供两个新的样本集$\bar{\mathbb{X}}_S^c$和$\bar{\mathbb{X}}_T^c$。此外，未收敛模型还带来另一个负面影响。这个影响来自于模型所产生的病态特征。这些特征会导致 ICMMD 的评估值很大，从而导致梯度爆炸问题。一种简单的处理方法是对ICMMD的评估值进行缩放处理，亦即计算以下缩放率。

$$r=\sqrt{\frac{2\tau_2(2\tau_2-1)}{\sum_{x,x'\in\bar{\mathbb{X}}_S^c\cup\bar{\mathbb{X}}_T^c}\left\|g(x)-g(x')\right\|_2^2}}$$

综上所述，算法最终所优化的目标函数为

$$\arg\min_{f\in\mathcal{F},g\in\mathcal{G}}\frac{1}{n_S}\sum_{i=1}^{n_S}\mathcal{L}(f(g(x_i)),y_i)+\lambda\ \mathrm{ICMMD}^2\left[\Phi,rg,\bar{\mathbb{X}}_S^c,\bar{\mathbb{X}}_T^c\right] \tag{3.11}$$

与式（3.9）类似，式（3.11）的目标函数也具备非凸特性。这要求模型对f和g的建模能力足够强。考虑到神经网络有足够强的函数拟合能力，算法采用神经网络对目标函数进行建模。根据目标函数的组成形式，神经网络结构按照图 3-6 设计。算法主要有四个阶段。第一阶段是通过模型获得所有领域样本的伪标签。第二阶段是根据过采样条件采样得到$\bar{\mathbb{X}}_S^c$和$\bar{\mathbb{X}}_T^c$。第三阶段执行一次前向传播，并评估类内分布差异度和分类误差。第四阶段执行一次反向传播算法得到参数θ_f和θ_g的梯度，并使用随机梯度下降算法更新参数。其中，参数θ_f和θ_g分别为f和g的参数，η负责调控参数的更新过程。

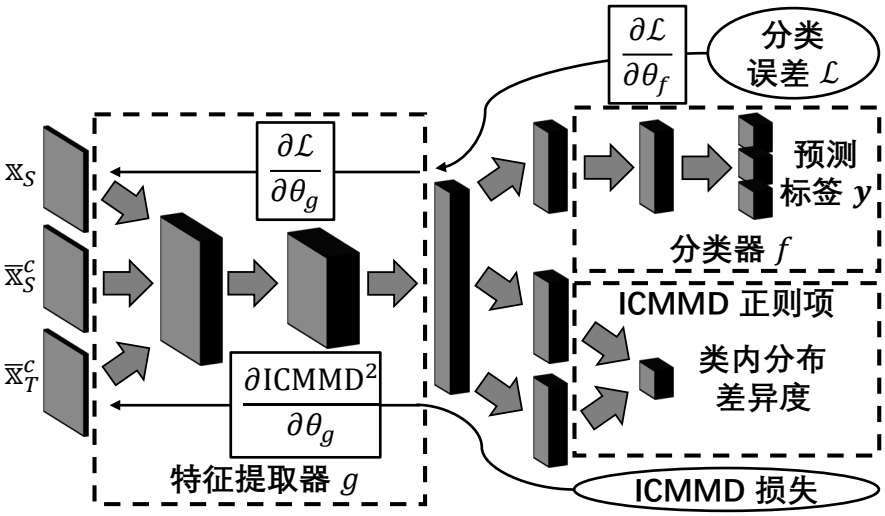


图 3-6 基于类内最大均值差异的神经网络

Figure 3-6 The neural network based on intra-class maximum mean discrepancy

综上所述，基于类内最大均值差异的策略有以下训练过程。

Algorithm 2 Cross-domain adaptation with intra-class maximum mean discrepancy

Input：samples $\{x_1, x_2, \ldots, x_{n_S}\} \cup \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{n_T}\}$ and labels $\{y_1, y_2, \ldots, y_{n_S}\}$

Output：the model parameters θ_f and θ_g

for each iteration I **do**

$c \leftarrow I \bmod |C|$

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

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

for each index $i \in \{1, 2, \ldots n_S\}$ **do**

if $f(g(x_i)) = c$ **then**

$\mathbb{X}_S^c \leftarrow \mathbb{X}_S^c \cup \{x_i\}$, $n_S^c \leftarrow n_S^c + 1$

end if

end for

for each index $j \in \{1, 2, \ldots n_T\}$ **do**

if $f(g(\tilde{x}_j)) = c$ **then**

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

end if

end for

if $n_S^c \geq \tau_1$ and $n_T^c \geq \tau_1$ **then**

Sample τ_2 times from $\mathbb{X}_S^c$ and obtain $\overline{\mathbb{X}}_S^c$, $n_S^c \leftarrow \tau_2$

Sample τ_2 times from $\mathbb{X}_T^c$ and obtain $\overline{\mathbb{X}}_T^c$, $n_T^c \leftarrow \tau_2$

else then

Sample τ_2 times from $\mathbb{X}_S$ and obtain $\overline{\mathbb{X}}_S^c$, $n_S^c \leftarrow \tau_2$

Sample τ_2 times from $\mathbb{X}_T$ and obtain $\overline{\mathbb{X}}_T^c$, $n_T^c \leftarrow \tau_2$

end if

Calculate the scaling rate r and update θ_f and θ_g using Stochastic Gradient Descent:

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

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

end for

3.7 本章小结

本章首先引入跨领域自适应的三个风险上界。由于没有引入可优化的参数空间，第一个风险上界无法使用目标领域中的无标签数据。第二个风险上界在第一个上界的基础上引入了特征转换函数。这个特征转换函数允许算法对其优化，从而对齐领域总体分布。第三个风险上界分析了算法在评估源领域期望分布时所产生的误差。这个风险上界表明源领域数量和样本数量能够同时影响算法对分布的评估。

接着，引入类别同态假设，并在这个假设下提出了给定具体类别后的无限制风险上界。根据这个风险上界，算法能够基于类内分布对齐策略进行类别信息的迁移。

最后，提出了两种类内分布对齐策略。这两种策略分别使用了总体均方离差和最大均值差异来评估类内分布的对齐程度。在深度学习的帮助下，这两个策略所对应的正则项能够被随机梯度下降算法优化，从而为算法寻得类别同态算子。

第四章　实验配置及结果分析

本章通过两个数据集验证上述两种策略的性能。一些共享配置会在此列出。所有算法的性能表示为在目标域测试集上的分类精度，即 $\mathrm{Acc}=\sum_{i=1}^{n_T'}\mathbb{I}(\mathrm{softmax}(f(g(x_i))),y_i)/n_T'$。其中，$n_T'$ 为目标域测试集的样本数量。值得注意的是，测试集上的标签均是 ground true 的，不是模型给定的伪标签。为了方便阅读，最优和次优性能分别用粗体和下划线表示。为了达到最佳性能，所有算法部署在 Caffe 框架[93]中并运行在一台高性能 CentOS 服务器上。该服务器装有一张 Nvidia Tesla K80 显卡和一颗 Xeon(R) E5 中央处理器。

基准算法包含五个最高水准的算法，分别为 LatentDA[59]、GRL[5]、DRCN[4]、DAN[35] 和 DDC[3]。其中，LatentDA 对不同领域使用不同的归一化。GRL 和 DRCN 专注于寻找一个领域不变的特征空间。DAN 和 DDC 使用最大均值差异技术对齐领域总体分布。值得注意的是，一些方法被修改为适用于多领域场景下。具体地说，GRL 使用 softmax 损失函数来替代交叉熵函数，并最大化领域预测错误率。DAN 和 DDC 会在任意两个领域上执行最大均值差异。这些基准算法均使用原始论文或代码所给出的超参数。

在基于类内总体均方离差的策略（ICMSD）中，权重 α 和 β 被分别设定为 0.2 和 0.1。在基于类内最大均值差异的策略（ICMMD）中，权重 λ 被设定为 1.0。此外，在基因实验上，ICMMD 的阈值 τ_1 和 τ_2 被分别设定为 1000 和 2000。在图像实验上，阈值 τ_1 和 τ_2 被分别设定为 128 和 64。这两种策略在相同数据集上使用相同的迭代周期和学习率退火方案，即基因实验上的迭代周期 $I_{\max}$ 和初始学习率 η_0 被分别设定为 500 和 0.01，而图像实验上的 $I_{\max}$ 和 η_0 被分别设定为 5000 和 0.001。两种策略都基于初始学习率 η_0 和 Caffe 自带的退火策略‘inv’计算每个迭代 $I\le I_{\max}$ 的学习率 η，即 $\eta=\eta_0(1+0.001\,I)^{-0.75}$。

4.1 基因数据实验

4.1.1 基因数据描述

在本次实验中，Gene Expression Omnibus (GEO) 数据集用于测试所有算法的性别分类性能。GEO 是一个公共存储库，用于存档和分发由科学界提交的完整基因表达数据。GEO 目前存储了超过三十万个样本。每个样本均包含基因表达水平及其相应的样本信息。这些数据是从 300 多个不同平台中收集而来的，并且大部分样本缺失平台信息。在数据预处理阶段，数据平台被视为一个领域，并且从 GEO 数据集中任选七个领域作为实验数据集。这七个领域共包含 36033 个样本。具体细节见表 4-1。此外，由于不同领域的特征数值范围比较悬殊，所有样本成分 x_v 都被转换为 x_v'，即

$$x'_v = \mathrm{sign}(x_v)\log(1+|x_v|) \tag{4.1}$$

表 4-1 GEO 数据集的统计信息

Table 4-1 Statistics of GEO dataset

领域	男性	女性	总数
GPL570	10273	9098	19371
GPL4133	1119	882	2001
GPL6102	383	464	847
GPL6480	1203	980	2183
GPL6884	1008	928	1936
GPL6947	3046	2533	5579
GPL10558	1898	2218	4116

4.1.2 网络结构说明

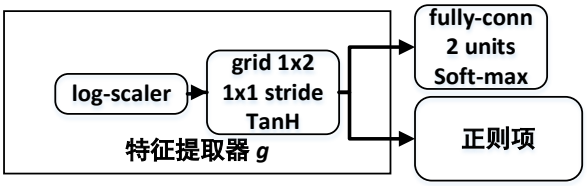


图 4-1 基因数据实验所采用的神经网络

Figure 4-1 The neural network for gene data

ICMSD 和 ICMMD 及对比算法均使用上述网络结构。这个网络有四个组成部分。详细地说，LogScaler 层和 Grid 层用于同态特征的提取，不同算法所使用的正则项被插入到全连接层的前面，全连接层和 softmax 损失函数用于评估分类误差。根据 ICMSD 和 ICMMD 的输入形式，指向正则项的箭头在 ICMSD 中代表源领域和目标领域的一个批量数据，而在 ICMMD 中代表两个领域中的同类样本子集。值得说明的是，LogScaler 层的计算形式如式（4.1），而 Grid 层的计算过程如下图所示。

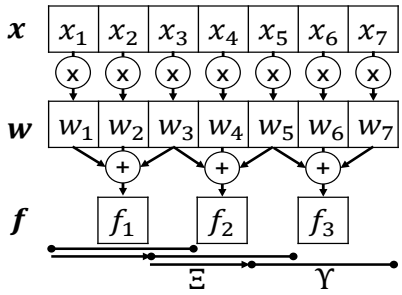


图 4-2 Grid 层的计算过程

Figure 4-2 The calculation of grid layer

由图 4-2 知，与 Conv 层不同的是，Grid 层中相邻感受野共享局部参数，而 Conv 层中相邻感受野共享所有参数。这是因为 Grid 层的参数与输入特征相互耦合，而 Conv 层的参数与滑动窗口相互耦合的。由于 Conv 层和 Grid 层拥有不同的设计，它们的计算形式也是不同的。具体地说，在 Conv 层中，滑动窗口在遍历整个输入特征图时，每个特征子图会与滑动窗口中的参数计算内积结果。在 Grid 层中，滑动窗口会同时遍历输入特征图和参数矩阵，每个特征子图会与对应的参数子矩阵计算内积结果。除此之外，由于 Grid 层的参数与输入特征耦合，Grid 层的参数个数和输入特征个数相同。因此，Grid 层比 Conv 层拥有更多的参数，减少了过多特征对共享参数的学习干扰。需要说明的是，在后续实验中，滑动窗口的大小 Υ 和移动步幅 Ξ 分别被设定为 2 和 1。

4.1.3 源域数量实验

本节实验首先围绕源领域数量展开。在图 4-1 网络结构下，表 4-2 和表 4-3 分别展示了 ICMSD 和 ICMMD 策略在各种源领域数量下的性能。当 Grid 层被替换为 Conv 层后，ICMSD 和 ICMMD 策略在各种源领域数量下的性能由表 4-4 和表 4-5 给出。

表 4-2 ICMSD-GRID 的源域数量实验结果

Table 4-2 The results of ICMSD-GRID varying the source number

GPL{源领域组合}→GPL10558	分类精度	5 个最重要的特征块
570	0.904±0.025	6519 4413 9050 4412 6518
6947 570	0.949±0.004	6519 6518 4412 4413 9051
6480 6947 570	0.951±0.003	6519 6518 4413 4412 9050
4133 6480 6947 570	0.950±0.003	6519 6518 4413 4412 9050
6884 4133 6480 4947 570	<u>0.953</u>±0.002	6519 6518 4413 4412 6520
6102 6884 4133 6480 4947 570	**0.959**±0.002	6519 6518 4413 9050 9051

表 4-3 ICMMD-GRID 的源域数量实验结果

Table 4-3 The results of ICMMD-GRID varying the source number

GPL{源领域组合}→GPL10558	分类精度	5 个最重要的特征块
570	0.913±0.004	4413 4412 6519 6518 9050
6947 570	0.961±0.002	6519 6518 9051 9050 6517
6480 6947 570	0.962±0.001	6519 6518 4412 4413 4414
4133 6480 6947 570	<u>0.963</u>±0.002	6519 6518 4413 4412 9050

6884 4133 6480 4947 570	0.963±0.002	6519 6518 9051 4412 4413
6102 6884 4133 6480 4947 570	**0.964**±0.001	6519 6518 4413 4412 6517

表 4-4 ICMSD-CONV 的源域数量实验结果

Table 4-4 The results of ICMSD-CONV varying the source number

GPL{源领域组合}→GPL10558	分类精度	5 个最重要的特征块
570	0.512±0.026	9865 6865 13783 9035 13874
6947 570	0.497±0.031	8965 8350 9871 12736 13513
6480 6947 570	0.505±0.028	7432 9050 6518 4412 6679
4133 6480 6947 570	0.507±0.030	9050 9051 5644 7433 4413
6884 4133 6480 4947 570	0.497±0.031	9050 7041 9051 5644 6221
6102 6884 4133 6480 4947 570	**0.520**±0.027	9050 7432 7041 6221 9051

表 4-5 ICMMD-CONV 的源域数量实验结果

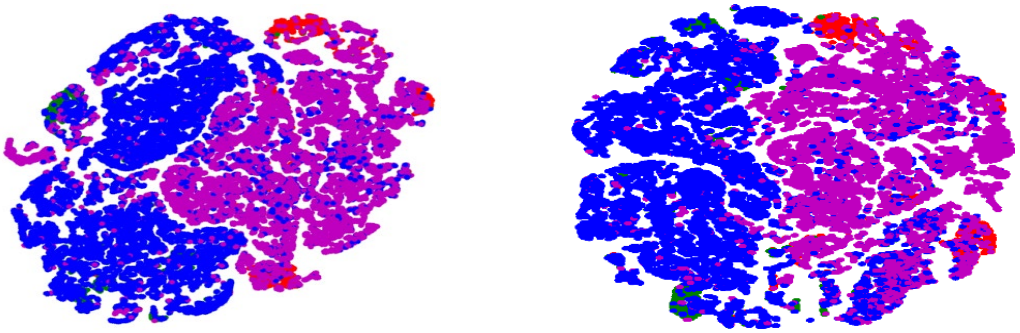
Table 4-5 The results of ICMMD-CONV varying the source number

GPL{源领域组合}→GPL10558	分类精度	5 个最重要的特征块
570	0.554±0.017	9050 13509 9770 5644 13436
6947 570	0.574±0.025	9050 7040 4412 6221 6679
6480 6947 570	0.591±0.033	9050 6221 7432 6679 9051
4133 6480 6947 570	**0.596**±0.025	9050 5644 9051 6221 7041
6884 4133 6480 4947 570	0.586±0.021	9050 7041 9051 5644 6222
6102 6884 4133 6480 4947 570	0.585±0.021	9050 7041 9051 6221 6222

从以上表格中可以知道，Grid 层所提取的特征比 Conv 层的要好。因为无论算法采用哪种类内分布对齐策略，Gird 层所提供的精度都比 Conv 层的好。此外，表 4-2 和表 4-3 表明，随着源领域数量的增加，算法的精度越来越高。在上一章的讨论中，源领域的风险上界（3.4）就表明，参与训练的源领域越多，源领域期望类内分布的估计偏差越小，从而使得类内分布的对齐效果越好。值得注意的是，ICMMD 的性能比 ICMSD 的性能更好。这是因为 ICMSD 仅仅对齐分布均值，没有全面考虑分布的其他性质。而 ICMMD 能够从所有线性转换中放大分布之间的微小差异，并对这些差异进行最小化。

除此之外，相比于 Conv 层，Gird 层所选的特征块更一致。从表 4-2 和表 4-3 可知，6519、6518、4412、4413、9050 和 9051 是最通用的六个重要特征块。因此，在包含六

个源领域的场景下，实验使用 t-SNE 算法[94]把这些特征块处理成如下可视化图像。



a) ICMSD 的可视化图像　　b) ICMMD 的可视化图像

a) The visual image on ICMSD　　b) The visual image on ICMMD

图 4-3 两种策略的可视化图像

Figure 4-3 Two visual images on two strategies

在图 4-3 中，蓝色点和品红点分别代表源领域中的男性和女性，绿色点和红色点分别代表目标领域中的男性和女性。不难看出，男性和女性被中轴线分割。这表明存在一个统一类别判别空间，使得所有领域能够在其上进行类别预测。此外，ICMMD 的中轴对称性比 ICMSD 的好，因此 ICMMD 的精度比 ICMSD 的精度更高。

4.1.4 类别比例实验

接下来，实验主要研究类别和领域之间的耦合影响。为了制造不同的耦合关系，实验会在两个源领域中以相反比例采样数据。具体来说，GPL6947 和 GPL10558 是任选的两个源领域，而 GPL570 是任选的目标领域。在数据准备阶段，GPL6947 和 GPL10558 分别按比例 $a:(10-a)$ 和 $(10-a):a$ 采样男性和女性样本，GPL570 按比例 5:5 采样两性样本。在图 4-1 网络结构下，将 $a=\{6,7,8,9\}$ 的实验结果整理后得到表 4-6 和表 4-7。

表 4-6 X-GRID 的类别比例实验结果

Table 4-6 The results of X-GRID varying the class ratio

	a=6	a=7	a=8	a=9
DDC-GRID	0.905±0.010	0.879±0.014	0.868±0.016	0.826±0.023
DAN-GRID	0.874±0.039	0.852±0.029	0.838±0.038	0.828±0.047
DRCN-GRID	0.819±0.037	0.799±0.043	0.762±0.038	0.752±0.038
GRL-GRID	0.910±0.003	0.807±0.019	0.677±0.025	0.580±0.054
LatentDA-GRID	0.872±0.003	0.822±0.002	0.785±0.003	0.742±0.004
ICMSD-GRID	0.923±0.001	0.920±0.002	0.918±0.002	0.917±0.001

ICMMD-GRID	**0.930**±0.003	**0.928**±0.002	**0.921**±0.003	**0.919**±0.001

表 4-7 X-CONV 的类别比例实验结果

Table 4-7 The results of X-CONV varying the class ratio

	a=6	a=7	a=8	a=9
DDC-CONV	0.507±0.001	0.503±0.004	0.503±0.001	0.504±0.003
DAN-CONV	0.506±0.005	0.504±0.007	0.503±0.004	0.502±0.004
DRCN-CONV	0.501±0.004	0.500±0.005	0.501±0.005	0.500±0.005
GRL-CONV	0.500±0.001	0.500±0.002	0.499±0.004	0.502±0.003
LatentDA-CONV	**0.526**±0.004	**0.532**±0.006	**0.509**±0.002	0.509±0.001
ICMSD-CONV	0.500±0.002	0.500±0.003	0.500±0.003	0.501±0.004
ICMMD-CONV	0.515±0.032	0.525±0.026	0.496±0.029	**0.524**±0.030

表 4-6 和表 4-7 表明，相比于 CONV 层，GRID 层的优势更大。此外，从表 4-6 中可以发现，随着类别和领域之间的耦合程度的增加，所有算法都出现了不同程度的性能损失，但是 ICMSD 和 ICMMD 依然稳定保持在最佳性能。这是因为 ICMSD 和 ICMMD 是基于类别同态假设设计的，从而保持对耦合影响的不变性。其余算法都出现大幅度的性能下降，这是因为这些算法在设计之初就没有考虑到类别问题。这些算法忽略了类别产生机制，并盲目追求特征或分布的对齐。需要说明一点的是，LatentDA 能够针对不同领域拟合不同的分布，但在类别不平衡场景下依然造成类别信息的迁移困难。

4.1.5 常规对比实验

最后，为了比较所有算法在默认数据配置下的性能表现，实验选择 GPL6102、GPL6884、GPL4133、GPL6480、GPL6947 和 GPL570 作为源领域，GPL10558 作为目标领域，并使用所有算法完成领域自适应任务。所有算法的自适应效果见下表

表 4-8 X-GRID 在默认配置下的结果

Table 4-8 The results of X-GRID in default configuration

算法	精度
DDC-GRID	0.857±0.001
DAN-GRID	0.838±0.001
DRCN-GRID	0.841±0.001
GRL-GRID	0.921±0.001

LatentDA-GRID	0.917±0.001
ICMSD-GRID	0.959±0.001
ICMMD-GRID	**0.964**±0.001
Train on Target	0.965±0.002

在表 4-8 中，最有一行的‘Train on Target’只使用目标领域训练模型，代表领域自适应的最高精度上限。从表中可知，ICMSD 和 ICMMD 均获得了接近上限的精度，其余算法则因为类别信息丢失问题而出现了性能下滑问题。

4.2 图像数据实验

4.2.1 图像数据描述

在本次实验中，OFFICE-31 数据集用于测试所有算法的日常物品分类性能。该数据集是评估自适应算法性能的一个常用数据集。它包含 3 个领域的图像数据，分别为 AMAZON、DSLR 和 WEBCAM，并且每个领域均包含 31 个类。具体细节见表 4-9。

表 4-9 OFFICE-31 数据集的统计信息

Table 4-9 Statistics of OFFICE dataset

领域	图片数	类别数
AMAZON	2817	31
DSLR	498	31
WEBCAM	795	31

4.2.2 网络结构说明

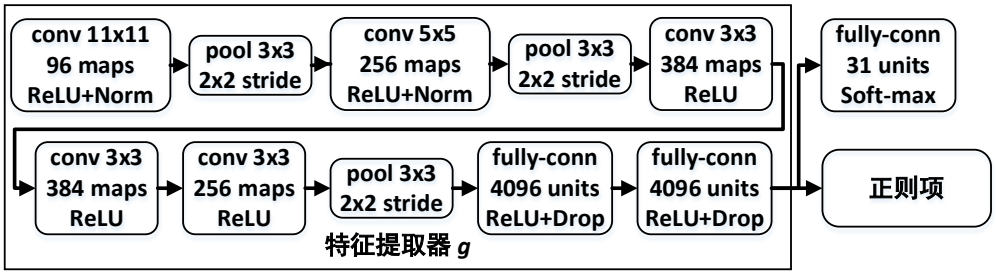


图 4-4 图像数据实验所采用的神经网络

Figure 4-4 The neural network for image data

上述网络结构扩展于 AlexNet[83]。具体来说，多个特征提取层组成同态特征提取器，正则项被插在最后一个全连接层前面。由于使用的数据包含 31 个分类，最后一个全连接层有 31 个神经元，分别输出每个类别的评估值。与主流算法类似，所有图片数据都经过减均值预处理，即每张图片均减去 ILSVRC12 所提供的均值。

4.2.3 类别比例实验

由于图像数据给出的领域比较少，图像数据上的源域数量实验被直接跳过。接下来的类别比例实验保持基因数据实验上的设计。不过，由于图像数据有 31 个类别，在任选的两个源领域中，第一类和最后一类的采样比例是相反的，其余的 29 个类别的采样比例按顺序以等差数列形式递减或递增。以‘AD2W’为例，AMAZON 的第一类和最后一类的采样比例为$a:(10-a)$，DSLR 的第一类和最后一类的采样比例为$(10-a):a$。AMAZON 中第二类到倒数第二类的采样数量逐渐递减，DSLR 中第二类到倒数第二类的采样数量逐渐递增。从另一个角度看，任意两个源领域的同一类样本数是互补的，亦即所有类别的样本总数大致相同。

表 4-10 在 AD2W 场景中的类别比例实验结果

Table 4-10 The results on AD2W varying the class ratio

	a=6	a=7	a=8	a=9
DDC	0.911±0.008	0.901±0.009	0.901±0.009	0.903±0.012
DAN	0.742±0.009	0.722±0.009	0.700±0.009	0.702±0.013
DRCN	0.934±0.003	0.923±0.005	0.916±0.007	0.915±0.010
GRL	0.925±0.003	0.787±0.013	0.754±0.024	0.731±0.012
LatentDA	0.820±0.001	0.814±0.001	0.814±0.001	0.801±0.001
ICMSD	0.944±0.001	0.941±0.002	**0.941**±0.001	**0.941**±0.002
ICMMD	**0.946**±0.001	**0.945**±0.002	**0.941**±0.003	**0.941**±0.002

表 4-11 在 AW2D 场景中的类别比例实验结果

Table 4-11 The results on AW2D varying the class ratio

	a=6	a=7	a=8	a=9
DDC	0.959±0.009	0.957±0.004	0.953±0.005	0.942±0.004
DAN	0.864±0.008	0.833±0.009	0.814±0.008	0.811±0.011
DRCN	0.975±0.003	0.973±0.007	0.972±0.008	0.972±0.005
GRL	0.970±0.004	0.812±0.015	0.761±0.021	0.743±0.016
LatentDA	0.841±0.001	0.835±0.001	0.831±0.001	0.832±0.001
ICMSD	0.981±0.002	0.982±0.002	0.982±0.002	0.978±0.002
ICMMD	**0.987**±0.002	**0.986**±0.001	**0.986**±0.002	**0.983**±0.001

表 4-12 在 DW2A 场景中的类别比例实验结果

Table 4-12 The results on DW2A varying the class ratio

	a=6	a=7	a=8	a=9
DDC	0.397±0.009	0.382±0.016	0.375±0.006	0.373±0.033
DAN	0.328±0.007	0.318±0.008	0.297±0.008	0.268±0.011
DRCN	-	-	-	-
GRL	0.534±0.006	0.518±0.008	0.498±0.005	0.479±0.005
LatentDA	0.622±0.002	**0.621**±0.002	0.618±0.002	0.617±0.003
ICMSD	0.619±0.003	0.619±0.002	0.618±0.002	**0.618**±0.004
ICMMD	**0.623**±0.003	**0.621**±0.003	**0.619**±0.002	0.616±0.003

表 4-10、4-11 和 4-12 均表明 ICMSD 和 ICMMD 对类别比例不敏感，并保持优于其他算法的性能。从算法假设的角度看，ICMSD 和 ICMMD 能够帮助模型寻得一个一致的类别决策空间。LatentDA 虽然为每个源领域单独提取类别信息并迁移到目标领域中，但是在极端的类别不平衡场景中依然面临性能下降的窘境。由于假设类别信息是独立于领域，GRL 有可能在类别不平衡场景中消除有用的特征。DRCN 虽然假设有用的特征不仅有益于类别预测也有益于特征重构，但是依然导致类别相关特征是有偏的，并倾向于预测主要类别。DAN 和 DDC 虽然使用了 MMD 做为正则项，但是针对领域总体分布进行对齐，导致了类别信息的流失。需要说明的是，DRCN 在 DW2A 场景下出现了崩溃现象。这是因为相比于 DSLR 和 WEBCAM，AMAZON 的分布更复杂。

4.2.4 常规对比实验

表 4-13 各种算法在默认配置下的结果

Table 4-13 The results of various algorithms in default configuration

	AD2W	AW2D	DW2A
DDC	0.851±0.004	0.873±0.003	0.455±0.004
DAN	0.885±0.011	0.943±0.011	0.486±0.003
DRCN	0.834±0.004	0.846±0.004	0.479±0.003
GRL	0.934±0.002	0.967±0.001	0.538±0.002
LatentDA	0.931±0.003	0.943±0.002	**0.603**±0.003
ICMSD	0.940±0.003	0.981±0.002	0.585±0.003

ICMMD	**0.946**±0.003	**0.983**±0.002	<u>0.591</u>±0.002
Train on Target	0.976±0.002	0.985±0.004	0.840±0.003

最后，在默认配置下，将本文提到的两种策略与其他主流算法进行比较，并得到表4-13 的实验结果。与基因数据实验类似，‘Train on Target’被视作无监督自适应的性能上界。从表 4-13 可知，ICMSD 和 ICMMD 几乎在所有自适应场景中获得了优于其他主流算法的性能。具体来说，ICMSD 和 ICMMD 在 AD2W 和 AW2D 场景中均获得了接近上界的性能。值得注意的是，所有算法在 DW2A 场景中的性能都不高，因为 AMAZON 的分布比 DSLR 和 WEBCAM 更复杂。然而，ICMSD 和 ICMMD 依然在 DW2A 场景中获得了次优性能，这表明这两种策略能够在复杂环境中捕获到领域一致的类别信息。

4.3 本章小结

本章主要针对类内总体均方离差和类内最大均值差异这两种策略展开实验，并比较了这两种策略与其他主流算法的性能表现。为了避免实验的偶然性，所有实验配置都在基因数据和图像数据上得到相应的结果。这些结果表明基于类内分布对齐的策略能够获得更优的性能表现。此外，从特征可视化的效果看，ICMMD 的性能表现比 ICMSD 更优越。这是因为 ICMSD 能够评估所有造成分布差异的成分，而 ICMSD 只考虑分布之间的均值偏差。

总结与展望

无监督领域自适应算法能够在有标签的源领域和无标签的目标领域之间建立一座信息桥梁。在这个桥梁上，源领域和目标领域互相交换彼此的信息。为了从这些信息中找到有用的类别信息，算法会通过分布对齐的手段把控这座桥梁上的信息进出情况，从而把类别无关的信息过滤掉。

然而，大部分主流的无监督领域自适应算法只考虑单个源领域的情况。为此，某些研究工作在已有算法基础上发展出了多源领域自适应算法。这些多源领域自适应算法能够为每个源领域搭建通往目标领域的信息桥梁。这些信息桥梁能够根据相应源领域的特点实行不同的把控手段。不过，这种做法默认服从源领域可分假设，因此无法应用到源领域不明确的混合场景中。少量不依赖这种假设的算法也因为其他假设而应用受限。为此，本文基于类别同态假设提出了一种基于类内分布对齐准则的跨领域适应算法，并且设计了两个分布对齐策略以验证所提算法的有效性。总的来说，本文主要涉及以下两个贡献：

（1）推广了单源领域自适应的风险上界，并且将其应用到无领域信息的跨领域自适应场景中。这个场景下的风险上界不建议算法对齐领域的总体分布，而是指导算法服从关于类别的同态产生机制并制定相应的类内分布对齐策略。

（2）设计了两套类内分布对齐策略，并且解决了算法部署中的多个问题。基于类内总体均方离差的策略最小化类内距离，并且必要时最大化类间距离。基于类内最大均值差异的策略使用轮转法对齐领域同类经验分布。

本文所提算法虽然能够基于类内分布对齐准则完成了无监督领域自适应任务，但是只服从了一种类别同态假设，即给定类别后，领域给样本施加了一种噪声影响。然而，样本还有可能受到领域的另一种影响，即领域专用特征的影响。这些特征有可能阻碍所提算法对齐类内分布，因为过多的领域相关信息会增大算法出错的可能性。一种可能的解决方案是在卷积层或池化层中执行类内分布对齐策略，从而实现注意力机制。具体思路是寻找一个特征窗口使得窗口内外的分布差异最大化，并使窗口内的特征具备分类能力。

参考文献

[1] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE TKDE, 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] Tzeng E, Hoffman J, Zhang N, et al. Deep domain confusion: Maximizing for domain invariance[J]. arXiv preprint arXiv:1412.3474, 2014

[4] Ghifary M, Kleijn W B, Zhang M, et al. Deep reconstruction-classification networks for unsupervised domain adaptation[C]//ECCV. Springer, Cham, 2016: 597-613.

[5] Ganin Y, Lempitsky V. Unsupervised Domain Adaptation by Backpropagation[C]//ICML. 2015: 1180-1189

[6] Baxter J. A model of inductive bias learning[J]. JAIR, 2000, 12: 149-198.

[7] Thrun S. Is learning the n-th thing any easier than learning the first?[C]//NIPS. 1996: 640-646.

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

[9] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning[M]. MIT press, 2018.

[10] Duda R O, Hart P E, Stork D G. Unsupervised learning and clustering[J]. Pattern classification, 2001: 517-601.

[11] 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.

[12] Xie S, Zheng Z, Chen L, et al. Learning semantic representations for unsupervised domain adaptation[C]//ICML. 2018: 5419-5428.

[13] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis[J]. IEEE TNN, 2011, 22(2): 199-210.

[14] Gopalan R, Li R, Chellappa R. Domain adaptation for object recognition: An unsupervised approach[C]//ICCV. 2011: 999-1006.

[15] Baktashmotlagh M, Harandi M T, Lovell B C, et al. Unsupervised domain adaptation by domain invariant projection[C]//ICCV. 2013: 769-776.

[16] Huang J, Gretton A, Borgwardt K M, et al. Correcting sample selection bias by unlabeled data[C]//NIPS. 2007: 601-608.

[17] Gong B, Grauman K, Sha F. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation[C]//ICML. 2013: 222-230.

[18] Liu J, Zhang L. Optimal Projection Guided Transfer Hashing for Image Retrieval[J]. arXiv preprint arXiv:1903.00252, 2019.

[19] Fu J, Zhang L, Zhang B, et al. Guided Learning: A New Paradigm for Multi-task Classification[C]//CCBR. Springer, Cham, 2018: 239-246.

[20] Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature representation[C]//ICML. 2013: 10-18.

[21] Ghifary M, Balduzzi D, Kleijn W B, et al. Scatter component analysis: A unified framework for domain adaptation and domain generalization[J]. IEEE TPAMI, 2017 (1): 1-1.

[22] Li S, Song S, Huang G. Prediction reweighting for domain adaptation[J]. TNNLS, 2017, 28(7): 1682-1695.

[23] Zhang L, Liu Y, Deng P. Odor recognition in multiple E-nose systems with cross-domain discriminative subspace learning[J]. IEEE TIM, 2017, 66(7): 1679-1692.

[24] Lu H, Shen C, Cao Z, et al. An embarrassingly simple approach to visual domain adaptation[J]. IEEE TIP, 2018, 27(7): 3403-3417.

[25] Li S, Song S, Huang G, et al. Domain invariant and class discriminative feature learning for visual domain adaptation[J]. IEEE TIP, 2018, 27(9): 4260-4273.

[26] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[27] Sun Z, Wang C, Wang H, et al. Learn multiple-kernel SVMs for domain adaptation in hyperspectral data[J]. IEEE GRSL, 2013, 10(5): 1224-1228.

[28] Li W, Xu Z, Xu D, et al. Domain generalization and adaptation using low rank exemplar SVMs[J]. IEEE TPAMI, 2018, 40(5): 1114-1127.

[29] Hu J, Lu J, Tan Y P. Deep transfer metric learning[C]//CVPR. 2015: 325-333.

[30] Ding Z, Fu Y. Robust transfer metric learning for image classification[J]. IEEE TIP, 2017, 26(2): 660-670.

[31] Xu Y, Pan S J, Xiong H, et al. A unified framework for metric transfer learning[J]. IEEE TKDE, 2017, 29(6): 1158-1171.

[32] Herath S, Harandi M, Porikli F. Learning an invariant hilbert space for domain adaptation[C]//CVPR. 2017: 3845-3854.

[33] Zhang Z, Wang M, Huang Y, et al. Aligning infinite-dimensional covariance matrices in reproducing kernel hilbert spaces for domain adaptation[C]//CVPR. 2018: 3437-3445.

[34] Li L, Zhang Z. Semi-supervised Domain Adaptation by Covariance Matching[J]. IEEE TPAMI, 2018 (1): 1-1.

[35] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[J]. arXiv preprint arXiv:1502.02791, 2015.

[36] Liu L, Lin W, Wu L, et al. Unsupervised deep domain adaptation for pedestrian detection[C]//ECCV. 2016: 676-691.

[37] Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks[C]//ICML. 2017: 2208-2217.

[38] Zhang X, Yu F X, Chang S F, et al. Deep transfer network: Unsupervised domain adaptation[J]. arXiv preprint arXiv:1503.00591, 2015.

[39] Motiian S, Piccirilli M, Adjeroh D A, et al. Unified deep supervised domain adaptation and generalization[C]//ICCV. 2017: 5715-5725.

[40] Long M, Cao Y, Cao Z, et al. Transferable representation learning with deep adaptation networks[J]. IEEE TPAMI, 2018: 1-1.

[41] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach[C]//ICML. 2011: 513-520.

[42] Chen M, Xu Z, Weinberger K, et al. Marginalized denoising autoencoders for domain adaptation[J]. arXiv preprint arXiv:1206.4683, 2012.

[43] Wen L, Gao L, Li X. A new deep transfer learning based on sparse auto-encoder for fault diagnosis[J]. IEEE TSMCS, 2017 (99): 1-9.

[44] Pinheiro P O. Unsupervised domain adaptation with similarity learning[C]//CVPR. 2018: 8004-8013.

[45] Chen Y, Li W, Sakaridis C, et al. Domain adaptive faster r-cnn for object detection in the wild[C]//CVPR. 2018: 3339-3348.

[46] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//NIPS. 2014: 2672-2680.

[47] Zhao J, Mathieu M, LeCun Y. Energy-based generative adversarial network[J]. arXiv preprint arXiv:1609.03126, 2016.

[48] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[J]. arXiv preprint arXiv:1701.07875, 2017.

[49] Zhang J, Ding Z, Li W, et al. Importance weighted adversarial nets for partial domain adaptation[C]//CVPR. 2018: 8156-8164.

[50] Pei Z, Cao Z, Long M, et al. Multi-adversarial domain adaptation[C]//AAAI. 2018: 3934-3941.

[51] Zhang W, Ouyang W, Li W, et al. Collaborative and Adversarial Network for Unsupervised domain adaptation[C]//CVPR. 2018: 3801-3809.

[52] Chen Q, Liu Y, Wang Z, et al. Re-weighted adversarial adaptation network for unsupervised domain adaptation[C]//CVPR. 2018: 7976-7985.

[53] Liu M Y, Tuzel O. Coupled generative adversarial networks[C]//NIPS. 2016: 469-477.

[54] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation[C]//CVPR. 2017: 7167-7176.

[55] Saito K, Ushiku Y, Harada T. Asymmetric tri-training for unsupervised domain adaptation[C]//ICML. 2017: 2988-2997.

[56] Zhang W, Ouyang W, Li W, et al. Collaborative and Adversarial Network for Unsupervised domain adaptation[C]//CVPR. 2018: 3801-3809.

[57] Xie S, Zheng Z, Chen L, et al. Learning semantic representations for unsupervised domain adaptation[C]//ICML. 2018: 5419-5428.

[58] Carlucci F M, Porzi L, Caputo B, et al. Just dial: Domain alignment layers for unsupervised domain adaptation[C]//ICIAP. 2017: 357-369.

[59] Mancini M, Porzi L, Rota Bulò S, et al. Boosting domain adaptation by discovering latent domains[C]//CVPR. 2018: 3771-3780.

[60] Gong B, Grauman K, Sha F. Reshaping visual datasets for domain adaptation[C]//NIPS. 2013: 1286-1294.

[61] Xu Z, Li W, Niu L, et al. Exploiting low-rank structure from latent domains for domain generalization[C]//ECCV. Springer, Cham, 2014: 628-643.

[62] Hoffman J, Kulis B, Darrell T, et al. Discovering latent domains for multisource domain adaptation[C]//ECCV. 2012: 702-715.

[63] Hoeffding W. Probability inequalities for sums of bounded random variables[M]//The Collected Works of Wassily Hoeffding. Springer, New York, NY, 1994: 409-426.

[64] Vapnik V N, Chervonenkis A Y. On the uniform convergence of relative frequencies of events to their probabilities[M]//Measures of complexity.

Springer, Cham, 2015: 11-30.

[65] Vapnik V. The nature of statistical learning theory[M]. Springer science & business media, 2013.

[66] Kearns M J, Vazirani U V, Vazirani U. An introduction to computational learning theory[M]. MIT press, 1994.

[67] Gretton A, Sejdinovic D, Strathmann H, et al. Optimal kernel choice for large-scale two-sample tests[C]//NIPS. 2012: 1205-1213.

[68] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning[C]//NIPS. 2007: 41-48.

[69] Zhang C H, Huang J. The sparsity and bias of the lasso selection in high-dimensional linear regression[J]. The Annals of Statistics, 2008, 36(4): 1567-1594.

[70] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. JRSS, 2006, 68(1):49-67.

[71] Evgeniou T, Micchelli C A, Pontil M. Learning multiple tasks with kernel methods[J]. JMLR, 2005, 6(Apr): 615-637.

[72] Evgeniou T, Pontil M. Regularized multi-task learning[C]//ACM SIGKDD. 2004: 109-117.

[73] Kim S, Xing E P. Tree-guided group lasso for multi-task regression with structured sparsity[C]//ICML. 2010, 2: 1.

[74] Chen X, Kim S, Lin Q, et al. Graph-structured multi-task regression and an efficient optimization method for general fused lasso[J]. arXiv preprint arXiv:1005.3579, 2010.

[75] Heskes T M. Empirical Bayes for learning to learn[C]//ICML. 2000: 367-374.

[76] Long M, Wang J. Learning multiple tasks with deep relationship networks[J]. arXiv preprint arXiv:1506.02117, 2015, 2.

[77] Lu Y, Kumar A, Zhai S, et al. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification[C]//CVPR. 2017: 5334-5343.

[78] Duong L, Cohn T, Bird S, et al. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser[C]//IJCNLP. 2015, 2: 845-850.

[79] Yang Y, Hospedales T M. Trace norm regularised deep multi-task learning[J]. arXiv preprint arXiv:1606.04038, 2016.

[80] Misra I, Shrivastava A, Gupta A, et al. Cross-stitch networks for multi-task learning[C]//CVPR. 2016: 3994-4003.

[81] Ruder S, Bingel J, Augenstein I, et al. Learning what to share between loosely related tasks[J]. arXiv preprint arXiv:1705.08142, 2017.

[82] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of physiology, 1962, 160(1): 106-154.

[83] 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.

[84] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Cognitive modeling, 1988, 5(3): 1

[85] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//NIPS. 2012: 1097-1105.

[86] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models[C]//ICML. 2013, 30(1): 3.

[87] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.

[88] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[89] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//CVPR. 2015: 1-9.

[90] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//CVPR. 2016: 770-778.

[91] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

[92] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[93] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]// ACM Multimedia. 2014: 675-678.

[94] Maaten L, Hinton G. Visualizing data using t-SNE[J]. JMLR, 2008, 9(Nov): 2579-2605.

攻读学位期间发表论文

1. 蔡瑞初，李嘉豪，郝志峰. 基于类内最大均值差异的无监督领域自适应算法[J]. 计算机应用研究（已录用）.
2. Ruichu Cai, Jiahao Li, Zhenjie Zhang, Xiaoyan Yang, Zhifeng Hao. DACH: Domain Adaptation without Domain Information[J]. IEEE TNNLS.
3. 发明专利：一种基于同态神经网络的跨领域图片分类方法（专利号：201710584948.X）.

学位论文独创性声明

本人郑重声明：所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明，并表示了谢意。本人依法享有和承担由此论文所产生的权利和责任。

论文作者签名：李嘉豪 日期：2019年5月29日

学位论文版权使用授权声明

本学位论文作者完全了解学校有关保存、使用学位论文的规定："研究生在广东工业大学学习和工作期间参与广东工业大学研究项目或承担广东工业大学安排的任务所完成的发明创造及其他技术成果，除另有协议外，归广东工业大学享有或特有"。同意授权广东工业大学保留并向国家有关部门或机构送交该论文的印刷本和电子版本，允许该论文被查阅和借阅。同意授权广东工业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、扫描或数字化等其他复制手段保存和汇编本学位论文。保密论文在解密后遵守此规定。

论文作者签名：李嘉豪 日期：2019年5月29日

指导教师签名： 日期：2019年5月29日

致谢

时光荏苒，白驹过隙。不知不觉，三年的研究生涯也将迎来结束。回首当初，自己怀着对人工智能的热爱义无反顾地选择了读研之路。在这条道路上，我遇到了很多人和事，也碰到了很多喜与悲。正是这些人生片段造就了现在的我。在此，我想感谢那些给予我支持和关心的人。

首先，我想对我的导师蔡瑞初教授说一声谢谢。没有他的指点，我可能不会了解到人工智能的前沿研究及进展。没有他的引导，我可能不会接触到迁移学习的最新工作及应用。没有他的教诲，我可能不会学习到文章写作的各种技巧及工具。蔡老师一直以来治学严谨、严于律己，深深影响了我对学术研究的思考，是我学术道路上的一面旗帜。在我遇到瓶颈的时候，是蔡老师不厌其烦地指引我走出误区并获得灵感。从另一方面说，没有蔡老师的正确带领，我可能无法走进这个纷繁复杂但妙趣横生的学术圈。

其次，我要感谢 DMIR 实验室的全体老师。正是有了他/她们开设的大学课程和学术研讨会，我才能了解到其他研究方向的理论发展及应用场景，避免陷入自身的思维误区。正是有了他/她们邀请的学术大牛和领域巨匠，我才能感受到其他研究人员的学术素养和治学作风。正是有了他/她们传递的热情问候和嘘寒问暖，我才能在失落的时候收获一份小小的自我安慰。

接着，我还要感激那些陪伴我健康成长的伙伴们，包括甄启琪、侯永杰、陈子彬、陈培辉、赵坤垚、林泽钿、乔杰、申策、唐钟洋、钟椿荣、黄礼泊、李可爱、曾艳等。感谢他/她们在研究生阶段给予我学习和生活上的帮助和勉励，让我的研究生涯留下了很多美好的回忆。感谢室友李嘉兴、梁有懿和赖琪多年的无怨无悔。

然后，我感谢家人多年来对我的理解。他们的默默支持是我生活学习两不误的重要保证。他们对我的养育之恩，我永世难忘。

最后，十分感谢各位专家和教授在百忙之中担任本论文的评审。他们的宝贵意见和建议确保了本论文的质量！

2019 年于广州

李嘉豪

附录 A

证明：对 $e(q,p_T\mid\pi_T)$ 同时加减 $e(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S])$ 和 $e(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\pi_T)$ 后，有

$$\begin{aligned}e\left(q,p_T\mid\pi_T\right)=&e\left(q,p_T\mid\pi_T\right)\\&+e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\\&+e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\pi_T\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\pi_T\right)\end{aligned}\tag{A.1}$$

调整式（A.1）等号右边各项的顺序，并对两项相减部分取绝对值，有不等式

$$\begin{aligned}e\left(q,p_T\mid\pi_T\right)=&e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\\&+\left|e\left(q,p_T\mid\pi_T\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\pi_T\right)\right|\\&+\left|e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\pi_T\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\right|\end{aligned}\tag{A.2}$$

根据三角不等式，式（A.2）中的第一个绝对式有不等式

$$\left|e\left(q,p_T\mid\pi_T\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\pi_T\right)\right|\le e\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T\mid\pi_T\right)\tag{A.3}$$

根据分布差异的定义式（3.1），式（A.2）中的第二个绝对式有不等式

$$\left|e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\pi_T\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\right|\le e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\left|\pi_T-\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right|\right)\tag{A.4}$$

把式（A.3）和（A.4）代入到（A.2）后，有以下不等式

$$\begin{aligned}e\left(q,p_T\mid\pi_T\right)\le&e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\\&+e\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T\mid\pi_T\right)+e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\left|\pi_T-\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right|\right)\end{aligned}\tag{A.5}$$

同理，对 $e(q,p_T\mid\pi_T)$ 加减 $e(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S])$ 和 $e(q,p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S])$ 后，有

$$\begin{aligned}e\left(q,p_T\mid\pi_T\right)=&e\left(q,p_T\mid\pi_T\right)\\&+e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\\&+e\left(q,p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)-e\left(q,p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\end{aligned}\tag{A.6}$$

调整式（A.6）等号右边各项的顺序，并对两项相减部分取绝对值，有不等式

$$\begin{aligned}e\left(q,p_T\mid\pi_T\right)=&e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\\&+\left|e\left(q,p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\right|\\&+\left|e\left(q,p_T\mid\pi_T\right)-e\left(q,p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\right|\end{aligned}\tag{A.7}$$

根据三角不等式，式（A.7）中的第一个绝对式有不等式

$$\left|e\left(q,p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)-e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\right|\le e\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\tag{A.8}$$

根据分布差异的定义式（3.1），式（A.7）中的第二个绝对式有不等式

$$\left|e\left(q,p_T\mid\pi_T\right)-e\left(q,p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\right|\le e\left(q,p_T\mid\left|\pi_T-\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right|\right) \tag{A.9}$$

把式（A.8）和（A.9）代入到（A.7）后，有以下不等式

$$\begin{aligned} e\left(q,p_T\mid\pi_T\right)=&e\left(q,\mathbb{E}_{\pi_\sigma(S)}[p_S]\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)\\ &+e\left(\mathbb{E}_{\pi_\sigma(S)}[p_S],p_T\mid\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right)+e\left(q,p_T\mid\left|\pi_T-\mathbb{E}_{\pi_\sigma(S)}[\pi_S]\right|\right)\end{aligned} \tag{A.10}$$

结合式（A.5）和（A.10），即得定理中的式（3.2）。

附录 B

证明：根据 $E(f,g)$ 和 $\hat{E}(f,g)$ 的定义，有以下等式

$$\begin{aligned}
&E(f,g)-\hat{E}(f,g)\\
&=\mathbb{E}_{p(g(x),y,S)}\left[\mathbb{I}(f\circ g(x)\neq y)\right]-\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{p(g(x),y|S_j)}\left[\mathbb{I}(f\circ g(x)\neq y)\right]\right]\\
&+\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{p(g(x),y|S_j)}\left[\mathbb{I}(f\circ g(x)\neq y)\right]\right]-\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{p(g(x),y|S_j)}\left[\mathbb{I}(f\circ g(x)\neq y)\right] \qquad \text{(B.1)}\\
&+\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{p(g(x),y|S_j)}\left[\mathbb{I}(f\circ g(x)\neq y)\right]-\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{I}(f\circ g(x_i^j)\neq y_i^j)\\
&=B_1+B_2+B_3
\end{aligned}$$

其中，B_1、B_2 和 B_3 分别代表第二个等式中的三个相减项。显然，根据源领域的产生机制，B_1 有以下等式

$$B_1=\mathbb{E}_{p(g(x),y,S)}\left[\mathbb{I}(f\circ g(x)\neq y)\right]-\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\mathbb{E}\left[p(g(x),y,S_j)\right]}\left[\mathbb{I}(f\circ g(x)\neq y)\right]=0 \qquad \text{(B.2)}$$

根据霍夫丁不等式（2.1），B_2 至少有概率 $1-\delta$ 满足以下不等式

$$B_2\leq\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \qquad \text{(B.3)}$$

不难发现，B_3 有以下等式

$$\begin{aligned}
B_3&=\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{p(g(x),y|S_j)}\left[\mathbb{I}(f\circ g(x)\neq y)\right]-\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{I}(f\circ g(x_i^j)\neq y_i^j)\\
&=\mathbb{E}_{\frac{1}{m}\sum_{j=1}^{m}p(g(x),y|S_j)}\left[\mathbb{I}(f\circ g(x)\neq y)\right]-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{I}(f\circ g(x_i^j)\neq y_i^j)\right]
\end{aligned}$$

由于 $1/m\sum_{j=1}^{m}\mathbb{I}(f\circ g(x_i^j)\neq y_i^j)$ 服从独立同分布，根据 VC 泛化上界（2.10），B_2 至少有概率 $1-\delta$ 满足以下不等式

$$B_3\leq\sqrt{\frac{2}{n}\log 2+\frac{2d_{\mathcal{F}\circ\mathcal{G}}}{n}\log\left(\frac{2ne}{d_{\mathcal{F}\circ\mathcal{G}}}\right)+\frac{2}{n}\log\frac{1}{\delta}} \qquad \text{(B.4)}$$

把式（B.2）、（B.3）和（B.4）代入到式（B.1）中，定理得证。