# 无监督域适应中不同分布度量下的理论保证

Yuntao Du

September 2020

## 1 Introduction

机器学习，尤其是深度学习方法在很多领域中多取得了令人瞩目的成果，包括计算机视觉，自然语言处理，语音识别等 [1]。但是，这些方法都需要大量的有标注的数据，才能训练出一个比较好的机器学习模型。为了克服这一问题，研究人员们提出了迁移学习这一研究范式 [2]。迁移学习旨在于从大量有标注数据的源领域领域 (辅助领域) 中迁移知识到有少量标注数据，甚至无标记数据的目标领域中，从而提高模型在目标领域上的效果。这一过程与人类学习知识中的" 举一反三" 相类似。

在传统的机器学习中，我们通常采用数据 "独立同分布" 这一假设，即假设训练数据和测试数据是在同一数据分布相互独立地采用出来的，并基于此构建了诸如 PAC 可学习理论 [3] 的机器学习理论。这些理论表明模型的泛化误差可以由模型的训练误差所以及训练样本的数目所界定，并且会随着训练样本的增加而减小。在迁移学习中，源域和目标域的数据通常来自不同的数据分布，使得在源域上训练好的模型很难泛化到目标领域中，因此如何衡量并降低两个领域之间的分布差异从而使得源域上的模型可以更好的泛化到目标领域成为了迁移学习领域的难点 [2]。

在本文中我们关注于迁移学习中的一个子领域–无监督域适应。在无监督域适应中，我们假设源域存在大量的有标注数据，在目标域中只有大量的无标注数据，并且源域和目标域的特征空间相同。这种场景也是迁移学习中最具挑战的一种场景 [2]。在过去的二十多年内，很多相关的理论和算法被提出来用于解决上述问题。在理论层面，研究人员们提出了 $H$-distance [4], $H\Delta H$-distance [5], Margin Disparity Discrepancy [6] 等距离度量，并基于此构建了相应的学习理论。在算法层面，受上述理论的启发，人们提

出了基于样本加权 [7, 8]，基于分布匹配 [9, 10, 11] 和基于对抗的学习算法 [6, 12, 13]，显著提升了模型的泛化效果。本文主要关注于无监督域适应学习理论，尤其是在不同分布度量下的目标域上泛化误差界限。

本文主要内容安排如下: 在第二章首先介绍相关概念和相关工作。在之后的章节中,本文将依次介绍 $H$-distance [4], $H\Delta H$-distance [5],Discrepancy Distance[14], 标签函数差异 [15] 和 Margin Disparity Discrepancy [6] 等距离度量, 以及基于此度量推导出的误差界限，最后对这几种误差界限进行了对比和分析。

# 2 相关概念及相关工作

## 2.1 相关概念

在迁移学习中, 源域样本和目标域样本分别来自两个不同的数据分布, 分别记作 $P$ 和 $Q$。这两个分布是在样本和内积空间 $\mathcal{X} \times \mathcal{Y}$ 上的联合分布，其中 $\mathcal{X} \in R^d$, 对于二分类问题，$\mathcal{Y} = \{0, 1\}$, 对于多分类问题,$\mathcal{Y} = \{0, 1, 2, ..., K\}$。对于分布 $D$, 我们将在分布 $D$ 上采样出的样本集合记为 $\hat{D}$。在无监督域适应中，存在一个在源域分布 $P$ 中采样的有标注数据集合 $\hat{\mathcal{P}} = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ 和在目标领分布 $Q$ 中采样的无标记数据集合 $\hat{Q} = \{x_i^t\}_{i=1}^{n_t}$。

在二分类的场景下，定义分布 $D$ 上真实的标签函数为 $f : \mathcal{X} \to [0, 1]$。对于任意一个分类器 $h : \mathcal{X} \to [0, 1]$, 分类器的误差被定义为:

$$\epsilon(h, f) = E_{x\sim\mathcal{D}}[h(x) \neq f(x)] = E_{x\sim\mathcal{D}}[|h(x) - f(x)|] \tag{1}$$

因此，分类器 $h$ 在源域和目标域上的分类误差可以表示为

$\epsilon_s(h) = \epsilon_s(h, f_s)$, $\epsilon_t(h) = \epsilon_s(h, f_t)$。分类器在源域和目标域样本集合上的经验误差被记作 $\hat{\epsilon}_s(h)$ 和 $\hat{\epsilon}_t(h)$。在多分类的场景下，误差的定义会在下文进行相应的修改。

## 2.2 相关工作

根据文献 [16] 的分类，可以将现有的域适应理论分为以下几类: 基于差异的误差界限 [4, 5, 14, 15, 6]，基于积分概率矩阵的误差界限 [17, 18, 19] 和基于 PAC-Bayesian 的误差界限 [20, 21]。最早的域适应理论是基于误差的泛化界限，该理论 [4] 针对于二分类问题，基于 0-1 损失函数和 $H$-distance,

作者提了第一个域适应的理论框架。根据该理论可知，分类器在目标域上的泛化误差由分类器在源域上的经验误差，两个域之间的分布差异和一些常数项所界定。该框架也成为后续的算法设计的指导框架。Mansour 等人将该理论扩展到对于任意满足三角不等式的损失函数 [14]。Yuchen Zhang 等人考虑在多分类下的场景，并基于 margin 距离推导出了相应的误差界限 [6]。在基于积分概率矩阵的这类理论中，主要包括优化传输 [17, 18, 19] 和最大均值差异两类。前者通常采用 Wasserstein 距离进行域差异度量，后者采用最大均值差异进行度量。基于这两种度量，作者也提出例如相应的理论界限。在基于 PAC-Bayesian 的这些理论中 [20, 21]，模型需要对一组分类器进行多数投票，根据其不一致性进行泛化误差的界定。在本文中，我们关注于基于差异的域适应理论。

## 3 基于 $H$-distance 和 $H\Delta H$-distance 的误差界

该理论 [4] 最早于 2006 年在 NIPS(NeurIPS) 上提出，后续的工作扩展于 2010 年的 Machine Learning 期刊上 [5]。在这个理论中，作者考虑二分类的情形，并基于 0-1 损失函数推导出了相应的理解界限。

为了衡量两个分布之间的差异，一个很自然的想法是采用 $L_1$ 距离。但 $L_1$ 距离存在两个缺点：1) 对于任意分布，$L_1$ 距离不能从有限的样本中被准确地估计。2) $L_1$ 散度是一个过分严格的度量，因为它涉及所有可测子集的最大值。因此造成误差界限较松。为了解决这两个问题，作者在本文中提出差异度量 $\mathcal{H}$-divergence。

**Definition 1** *给定两个分布 $\mathcal{P}$ 和 $\mathcal{Q}$，令 $\mathcal{H}$ 为假设类，$I(h)$ 为特性函数，其中 $h \in \mathcal{H}$, 即 $x \in I(h) \Leftrightarrow h(x) = 1$。$\mathcal{H}$-divergence 被定义为:*

$$d_{\mathcal{H}}(\mathcal{P}, \mathcal{Q}) = 2 \sup_{h \in \mathcal{H}} |Pr_{\mathcal{P}}[I(h)] - Pr_{\mathcal{Q}}[I(h)]| \tag{2}$$

在有限的样本集上，通常采用经验 $\mathcal{H}$-divergence 来进行度量。对于一个对称的假设类 $\mathcal{H}$ 和两个样本数为 $m$ 的样本集 $\hat{P}, \hat{Q}$, 经验经验 $\mathcal{H}$-divergence 可以表示为:

$$\hat{d}_{\mathcal{H}}(\hat{\mathcal{P}}, \hat{\mathcal{Q}}') = 2(1 - \min_{h \in \mathcal{H}} [\frac{1}{m} \sum_{x:h(x)=0} I[x \in \hat{\mathcal{P}}] + \frac{1}{m} \sum_{x:h(x)=1} I[x \in \hat{\mathcal{Q}}]) \tag{3}$$

其中 $I[\cdot]$ 为指示函数，基于 $\mathcal{H}$-divergence, 作者提出了相应的学习理论。

**Theorem 1** 令 $H$ 表示一个 $VC$ 维为 $d$ 的假设空间，给定从源域上以 $iid$ 方式采样的大小为 $m$ 的样本集，则至少以 $1-\delta$ 的概率，对于任意一个 $h \in \mathcal{H}$ 有:

$$\epsilon_t(h) \leq \hat{\epsilon}_s(h) + d_{\mathcal{H}}(\hat{D}_s, \hat{D}_t) + \lambda^* + \sqrt{\frac{4}{m}(d\log\frac{2em}{d} + \log\frac{4}{\delta})} \tag{4}$$

其中，$e$ 是自然底数，$\lambda^* = \epsilon_s(h^*) + \epsilon_t(h^*)$ 是理想联合误差，$h^* = \arg\min_{h\in\mathcal{H}} \epsilon_s(h) + \epsilon_t(h)$ 是在源域和目标域上的最优分类器。

基于理论1，可以发现，目标域上的泛化误差由四项所界定:1) 源域上的经验误差，2）源域和目标域之间的分布差异，3) 理想联合误差，4) 与样本数，VC 维等相关的常数项。理想联合误差 $\lambda$ 无法进行准确计算，因为其需要目标域上的真实标签。在很多情况下，我们都假设 $\lambda^*$ 是一个很小的值，即存在一个分类器使得其在源域和目标域上的分类误差都比较小，从而使得我们可以进行知识迁移。在此假设下，影响目标域泛化误差的就只有前两项：源域泛化误差和两个域之间的分布差异。受理论1的启发，Ganin 等人在 2016 年提出了 DANN 算法 [12]，基于域判别器来衡量两个域上的差异。

基于 $H$-distance, 作者更进一步定义了 $H\Delta H$ 空间和 $H\Delta H$-distance[5]。

**Definition 2** 对于一个假设空间 $\mathcal{H}$, 对称差假设空间 $\mathcal{H}\Delta\mathcal{H}$ 是这种空间的集合

$$g \in \mathcal{H}\Delta\mathcal{H} \Leftrightarrow g(x) = h(x) \oplus h'(x), \quad h, h' \in \mathcal{H} \tag{5}$$

**Definition 3** 对于任意的 $h, h' \in \mathcal{H}$,

$$d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) = 2 \sup_{h,h'\in\mathcal{H}} |Pr_{x\sim P}[h(x) \neq h'(x)] - Pr_{x\sim Q}[h(x) \neq h'(x)]| \tag{6}$$

基于 $H\Delta H$-distance，作者又进一步给出了新的误差界限。

**Theorem 2** 令 $\mathcal{H}$ 是一个 $VC$ 维为 $d$ 的假设空间。$\hat{P}, \hat{Q}$ 是从分布 $P$ 和 $Q$ 中采样出的大小为 $m$ 的样本集。则对于任意的 $\delta \in (0, 1)$ 和任意的 $h \in \mathcal{H}$, 至少有 $1-\delta$ 的概率有

$$\epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}, \hat{Q}) + 4\sqrt{\frac{2d\log(2m) + \log(\frac{2}{\delta})}{m}} + \lambda \tag{7}$$

为了便于理解，下面附上本理论的证明过程

$$
\begin{aligned}
\epsilon_t(h) &= \epsilon_t(h, f_t) \\
&\leq \epsilon_t(h*) + \epsilon_t(h, h^*) \\
&\leq \epsilon_t(h^*) + \epsilon_s(h, h^*) + \epsilon_t(h, h^*) - \epsilon_s(h, h^*) \\
&\leq \epsilon_t(h^*) + \epsilon_s(h, h^*) + |\epsilon_t(h, h^*) - \epsilon_s(h, h^*)| \\
&\leq \epsilon_t(h^*) + \epsilon_s(h, h^*) + \frac{1}{2}\hat{d}_{H\Delta H}(\hat{P}, \hat{Q}) \\
&\leq \epsilon_t(h^*) + \epsilon_s(h^*) + \epsilon_s(h) + \frac{1}{2}\hat{d}_{H\Delta H}(\hat{P}, \hat{Q}) \\
&\leq \epsilon_s(h) + \frac{1}{2}\hat{d}_{H\Delta H}(\hat{P}, \hat{Q}) + \lambda \\
&\leq \epsilon_s(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}, \hat{Q}) + 4\sqrt{\frac{2d\log(2m) + \log(\frac{2}{\delta})}{m}} + \lambda
\end{aligned}
\tag{8}
$$

由此推导过程可以看出，推导过程第 4 到第 5 行的过程中，实际上是在给 $|\epsilon_t(h, h^*) - \epsilon_s(h, h^*)|$ 寻找一个上界，因此作者提出的 $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}, \hat{Q})$ 距离实际上是定义出的上界。通过比较 $\mathcal{H}$-distance 和 $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}, \hat{Q})$。我们可以发现 $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\hat{P}, \hat{Q})$ 是在假设空间取 $\mathcal{H}\Delta\mathcal{H}$ 时的特例。基于理论2，Saito 等人提出了 MCD 算法 [22]，通过设计两个分类器的差异来近似 $H\Delta H$-distance，进而降低两个域之间的差异。

# 4 基于差异距离 (Discrepancy Distance) 的误差界限

$H$-distance 或 $H\Delta H$-distance 只考虑损失函数为 0-1 损失函数的情景，更进一步，Mansour 等人将其扩展到任意满足三角不等式的损失函数 [14]。作者首先定义了差异距离 (Discrepancy Distance)。

**Definition 4** *令 $H$ 表示一类假设空间，$L : \mathcal{Y} \times \mathcal{Y} \to R$ 表示在 $\mathcal{Y}$ 上的损失函数，两个分布 $P$ 和 $Q$ 之间的差异距离 $disc_L$ 被定义为:*

$$
disc_L(P, Q) = \max_{h, h' \in \mathcal{H}} |L_P(h, h') - L_Q(h, h')| \tag{9}
$$

可以看出，差异距离实际上是将 $H\Delta H$ 距离从 0-1 损失函数向任意损失函数的扩展。为了方便进行误差界限的推导，作者约束损失函数需满足三角不等式，即 $disc_L(P, Q) \leq disc_L(P, M) + disc_L(M, Q)$。

定义 $h_Q^* \in \arg\min_{h\in\mathcal{H}} L_Q(h, f_Q)$, 其中 $f_Q$ 是在从分布 $Q$ 上的标签函数。相似地，定义 $h_P^*$ 是 $L_P(h, f_P)$ 的最优分类器。为了能够进行适应 (迁移)，作者假设这两个最优分类器之间的平均损失 $L_Q(h_Q^*, h_P^*)$ 很小。和理论1,2中的假设不同，在理论1,2中，作者假设是在源域和目标域上存在一个最优分类器，而在此理论中，作者是假设源域和目标域各自存在一个最优分类器，并且这两个分类器之间差异很小。

**Theorem 3** 假设损失函数 $L$ 是对称的并且满足三角不等式，则对于任意 $h \in \mathcal{H}$，都有

$$L_Q(h, f_Q) \leq L_Q(h_Q^*, f_Q) + L_P(h, h_P^*) + disc_L(P, Q) + L_P(h_P^*, h_Q^*) \quad (10)$$

对比理论2, 作者也进行了一些简单的分析。如假定 $h_Q^* = h_P^*$, 则有 $h^* = h_P^* = h_Q^*$, 在此时，理论3变成了 $L_Q(h, f_Q) \leq L_Q(h^*, f_Q) + L_P(h, h^*) + disc(P, Q)$。则理论2变为了 $L_Q(h, f_Q) \leq L_Q(h^*, f_Q) + L_P(h, f_P) + L_P(h^*, f_P) + disc(P, Q)$。根据三角不等式可以有 $L_P(h, h^*) \leq L_P(h, f_P) + L_P(h^*, f_P)$, 因此在此条件下，理论3是比理论2更紧的一个误差界限。

# 5 结合标签函数差异的误差界限

理论1和2已经被提出和使用了很多年。基于这个理论的启发，许多算法在进行设计时，其目标通常是在最小化源域分类损失的同时学习一个领域无关的特征。然而这类算法在某些情况下可能会失效。在文献 [15] 中，HanZhao 构造了一个反例来对这个现象进行说明。如图1所示，左图表示的是在原空间下，源域和目标域的分布情况，右图是经过某一个特性变换后，源域和目标域数据的分布情况，可以看出，此时两个域之间的差异为 0，但对于任意一个分类器，其在源域和目标域上的分类误差之和始终为 1。在这种极端条件下，最小化源域上的分类误差，反而会使目标域上的误差变大。

针对于这个问题，作者提出了一个新的理论。

**Theorem 4** 令 $f_s$,$f_t$ 表示源域和目标域上的标签函数，$\hat{P}$,$\hat{Q}$ 表示从两个域中采用出的样本，每个样本集的大小都为 $m$, $Rads(\mathcal{H})$ 表示 *Redemacher* 复杂度。那么，对于任何一个 $\mathcal{H} \in [0,1]^{\mathcal{X}}$ 和 $h \in \mathcal{H}$，都有

$$\begin{aligned} \epsilon_t(h) \leq & \hat{\epsilon}_s(h) + d_{\mathcal{H}}(\hat{P}, \hat{Q}) + 2Rads(\mathcal{H}) + 4Rads(\hat{\mathcal{H}}) \\ & + min\{\mathbb{E}_P[|f_s - f_t|], \mathbb{E}_Q[|f_s - f_t|]\} + O(\sqrt{\log(1/\delta)/m}) \end{aligned} \quad (11)$$

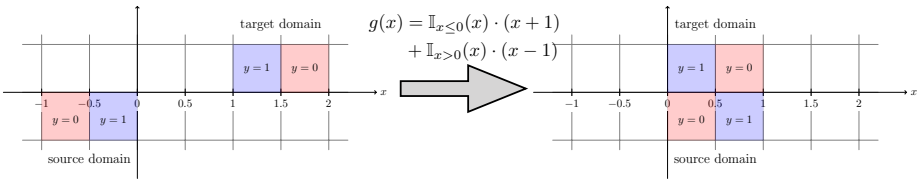


图 1: 理论2的反例示意图

其中，$\hat{\mathcal{H}} = \{sgn(|h(x) - h'(x)| - t)|h, h' \in \mathcal{H}, t \in [0, 1]\}$。

该泛化界限可以分为 3 部分，第一部分为域适应部分，包括源域经验误差，经验 $\mathcal{H}$-distance 和标签函数差异。第二部分对应着对假设空间 $\mathcal{H}$ 和 $\hat{\mathcal{H}}$ 的复杂度测量，第三部分描述有限样本造成的误差。对比理论4和理论2，最大的不同是理论4中的 $min\{\mathbb{E}_P[|f_s - f_t|], \mathbb{E}_Q[|f_s - f_t|]\}$ 项和理论2中的 $\lambda^*$ 项，后者依赖对假设空间 $\mathcal{H}$ 的选择，而前者则不需要。并且在理论4中，揭示了条件偏差的问题，可以很好的解释上面的反例。

# 6 基于 margin 距离的多类域适应理论

理论1,2,3,4都是考虑在二分类场景下的误差界限。在实际的应用场景中，我们更多的是面临多分类的问题。Yuchen Zhang 等人基于 Margin 距离首先提出了在多分类问题下的误差界限 [6]。

实际上，$\mathcal{H}\Delta\mathcal{H}$ 距离是两个域上分布距离差异的一个上界，但是在计算该距离时，要分别对 $h$ 和 $h'$ 在所有的假设空间中进行遍历。作者发现，实际上在计算分布差异距离上界的时候，没有必要对两个分类器都进行遍历，可以固定其中的一个分类器，对于另外一个分类器在假设空间中就足够。基于此，作者定义了一个和差异距离 (Discrepancy Distance) 相似的距离度量–Disparity Discrepancy。

**Definition 5** *给定一个假设空间 $\mathcal{H}$ 和一个假设空间中的分类器 $h \in \mathcal{H}$, Disparity Discrepancy 被定义为:*

$$d_{h,\mathcal{H}} = \sup_{h' \in \mathcal{H}} (disp_Q(h', h) - disp_P(h', h)) = \sup_{h' \in \mathcal{H}} (\mathbb{E}_Q \mathbb{1}(h', h) - \mathbb{E}_P \mathbb{1}(h', h)) \tag{12}$$

对于二分类问题，分类器只有两个输出，而在多分类问题中，分类器有多个输出。因此不能直接简单地将二分类问题中的损失函数用到多分类问

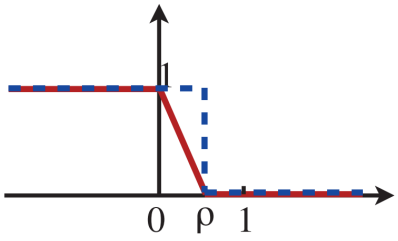


图 2: margin 损失示意图

题中。作者采用在传统机器学习中常用的 margin loss 来应对多分类问题。

下面对 margin loss 相关的符号和定义进行一个简要的说明。定义 $\mathcal{F}$ 为打分函数 (scoring function)$f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|} = \mathbb{R}^K$ 的假设空间，其中 $f$ 输出的每一维都表示预测的置信度。为了表示方便，用 $f(x, y)$ 表示 $f(x)$ 对于标签 $y$ 的预测值。使得样本 $x$ 预测分数最大的是其预测标签，因此，存在一个包含 $h_f$ 的标签函数空间 $\mathcal{H}$:

$$h_f : x \longmapsto \arg\max_{y \in \mathcal{Y}} f(x, y) \tag{13}$$

对于一个有标记的样本 $(x, y)$, 一个假设 $f$ 的间隔被定义为:

$$\rho_f(x, y) = \frac{1}{2}(f(x, y) - \max_{y' \neq y} f(x, y')) \tag{14}$$

相应的间隔损失和经验间隔损失被定义为:

$$\begin{aligned} err_D^{(\rho)}(f) &= \mathbb{E}_{x \sim D} \Phi_\rho \circ \rho_f(x, y) \\ err_{\hat{D}}^{(\rho)}(f) &= \mathbb{E}_{x \sim \hat{D}} \Phi_\rho \circ \rho_f(x, y) = \frac{1}{n} \sum_{i=1}^{n} \Phi_\rho(\rho_f(x_i, y_i)) \end{aligned} \tag{15}$$

其中 $\circ$ 表示函数复合，$\Phi_\rho$ 定义为

$$\Phi_\rho = \begin{cases} 0 & \rho \leq x \\ 1 - \frac{x}{\rho} & 0 \leq x \leq \rho \\ 1 & x \leq 0 \end{cases} \tag{16}$$

关于 margin 损失的示意图如图2所示。通过用 margin 损失替换 0-1 损

失，作者定义了 margin disparity 和它的经验距离:

$$
\begin{aligned}
&disp_D^{(\rho)}(f', f) = \mathbb{E}_D \Phi_\rho \circ \rho_{f'}(\cdot, h_f) \\
&disp_{\hat{D}}^{(\rho)}(f', f) = \mathbb{E}_{\hat{D}} \Phi_\rho \circ \rho_{f'}(\cdot, h_f) = \frac{1}{n} \sum_{i=1}^{n} \Phi_\rho \circ \rho_{f'}(x_i, h_f(x_i))
\end{aligned} \tag{17}
$$

基于 margin disparity，作者定义了 Margin Disparity Discrepancy (MDD)。

**Definition 6** 基于 *margin disparity*，定义 *Margin Disparity Discrepancy (MDD)* 和其经验版本为:

$$
\begin{aligned}
d_{f,\mathcal{F}}^{(\rho)}(P, Q) &= \sup_{f' \in \mathcal{F}} (disp_Q^{(\rho)}(f', f) - disp_P^{\rho}(f', f)) \\
d_{f,\mathcal{F}}^{(\rho)}(\hat{P}, \hat{Q}) &= \sup_{f' \in \mathcal{F}} (disp_{\hat{Q}}^{(\rho)}(f', f) - disp_{\hat{P}}^{\rho}(f', f))
\end{aligned} \tag{18}
$$

基于 MDD, 作者推导了一个新的泛化界限。

**Theorem 5** 令 $\mathcal{F}$ 是一簇从 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 到 $[a, b]$ 的映射函数，$\mathcal{R}_{n,D}(F)$ 表示 $F$ 在分布 $D$ 上 *Redemacher* 复杂度，$n$ 为从分布 $D$ 上采样的样本数，对于任意的 $\delta > 0$, 至少有 $1 - 3\delta$ 的概率，对于任意的打分函数 $f$ 都有，

$$
\begin{aligned}
err_Q(f) \leq & err_{\hat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\hat{P}, \hat{Q}) + \lambda + \frac{2k^2}{\rho} \mathcal{R}_{n,p}(\Pi_1 \mathcal{F}) + \\
& \frac{k}{\rho} \mathcal{R}_{n,p}(\Pi_{\mathcal{H}} \mathcal{F}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{k}{\rho} \mathcal{R}_{m,Q}(\Pi_{\mathcal{H}} \mathcal{F}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}
\end{aligned} \tag{19}
$$

其中 $\Pi_1 \mathcal{F} = \{x \longmapsto f(x, y) | y \in \mathcal{Y}, f \in \mathcal{F}\}$,$\Pi_{\mathcal{H}} \mathcal{F} = \{x \longmapsto f(x, h(x)) | h \in \mathcal{H}, f \in \mathcal{F}\}$, $\lambda = \lambda(\rho, \mathcal{F}, P, Q)$ 是和 $f$ 无关的常数。

该理论与基于 0-1 损失和 $H\Delta H$-distance 的理论2相比，通过选择一个更好的间隔 $\rho$, 可以实现在目标域上更好的泛化性。该理论指出，在泛化性和间隔 $\rho$ 之间存在一个权衡。对于小的 $\rho$ 和很大的假设空间，前两项没有很大的差异，因此随着 $\rho$ 的增加，右边的项会逐渐减小。对于比较大的 $\rho$, 这些项无法优化以达到可接受的较小值。基于此理论，作者提出了 MDD 算法，通过采用两个分类器的结构，近似并最小化 MDD 距离。

# 7 比较与分析

在表1中, 我们对这几种误差界限进行了对比。可以发现，之前的学习理论大部分基于二分类问题，只有少部分理论考虑多分类问题。并且，这些理论可以直接指导并归纳出相应的算法设计。

| 文献 | 类别个数 | 框架 | 距离度量 | 基于理论归纳的算法 |
|---|---|---|---|---|
| [4] | 2 | VC | $H$-distance | DANN[12] |
| [5] | 2 | VC | $H\Delta H$-distance | MCD[22] |
| [14] | 2 | Rademacher | Discrepancy Distance | - |
| [15] | 2 | Rademacher | $H$-distance& 标签函数差异 | - |
| [6] | K | Rademacher | Margin Disparity Discrepancy | MDD[6] |

表 1: 几种误差界限对比

# 8 总结

在本文中，我们对无监督域适应问题的误差界限进行了总结和梳理。本文主要关注于基于差异的学习理论，介绍了五种比较有代表性的学习理论。针对于无监督域适应问题的学习理论还有待进一步探索的空间，相信学习理论的突破也会给这个领域带来更大的进展。

# 参考文献


[1] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael S. Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14:1–37, 2007.

[2] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge & Data Engineering*, 22(10):1345–1359, 2010.

[3] L. Valiant. A theory of the learnable. *Commun. ACM*, 27:1134–1142, 1984.

[4] Shai Ben-David, John Blitzer, K. Crammer, and Fernando C Pereira. Analysis of representations for domain adaptation. In *NIPS*, 2006.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando C Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2009.

[6] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I. Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, 2019.

[7] W. Dai, Qiang Yang, G. Xue, and Y. Yu. Boosting for transfer learning. In *ICML '07*, 2007.

[8] Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1855–1862, 2010.

[9] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jia-Guang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. *2013 IEEE International Conference on Computer Vision*, pages 2200–2207, 2013.

[10] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[11] Sinno Jialin Pan, Ivor Wai-Hung Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22:199–210, 2011.

[12] Yaroslav Ganin, E. Ustinova, Hana Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

[13] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018.

[14] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *ArXiv*, abs/0902.3430, 2009.

[15] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representation for domain adaptation. *ArXiv*, abs/1901.09453, 2019.

[16] I. Redko, Emilie Morvant, Amaury Habrard, M. Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. 2020.

[17] I. Redko, Amaury Habrard, and M. Sebban. Theoretical analysis of domain adaptation with optimal transport. In *ECML/PKDD*, 2017.

[18] N. Courty, Rémi Flamary, Amaury Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *ArXiv*, abs/1705.08848, 2017.

[19] Sofien Dhouib, I. Redko, and C. Lartizien. Margin-aware adversarial domain adaptation with optimal transport. 2020.

[20] P. Germain, Amaury Habrard, F. Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML*, 2013.

[21] P. Germain, Amaury Habrard, F. Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In *ICML 2016*, 2015.

[22] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.