

# 迁移学习简明手册

一点心得体会

版本号： v1.0

王晋东

中国科学院计算技术研究所

[tutorial.transferlearning.xyz](http://tutorial.transferlearning.xyz)

2018 年 4 月

## 摘要

迁移学习作为机器学习的一大分支，已经取得了长足的进步。本手册简明地介绍迁移学习的概念与基本方法，并对其中的领域自适应问题中的若干代表性方法进行讲述。最后简要探讨迁移学习未来可能的方向。

本手册编写的目的是帮助迁移学习领域的初学者快速入门并掌握基本方法，为自己的研究和应用工作打下良好基础。

本手册的编写逻辑很简单：是什么——介绍迁移学习；为什么——为什么要用迁移学习、为什么能用；怎么办——如何进行迁移（迁移学习方法）。其中，是什么和为什么解决概念问题，这是一切的前提；怎么办是我们的重点，也占据了最多的篇幅。为了最大限度地方便初学者，我们还特别编写了一章上手实践，直接分享实现代码和心得体会。

## 推荐语

看了王晋东同学的“迁移学习小册子”，点三个赞！迁移学习被认为是机器学习的下一个爆点，但介绍迁移学习的文章却很有限。这个册子深入浅出，既回顾了迁移学习的发展历史，又囊括了迁移学习的最新进展。语言流畅，简明通透。应该对机器学习的入门和提高都有很大帮助！

——杨强 (迁移学习权威学者,香港科技大学教授,IJCAI president, AAAI/ACM fellow)

<b>目录</b>		
<b>写在前面</b>	<b>I</b>	<b>5 迁移学习的基本方法</b> <span style="float: right;">19</span>
<b>致谢</b>	<b>II</b>	<b>5.1 基于样本迁移</b> . . . . . <span style="float: right;">19</span>
<b>手册说明</b>	<b>III</b>	<b>5.2 基于特征迁移</b> . . . . . <span style="float: right;">20</span>
<b>1 迁移学习基本概念</b>	<b>1</b>	<b>5.3 基于模型迁移</b> . . . . . <span style="float: right;">20</span>
<b>1.1 引子</b>		<b>5.4 基于关系迁移</b> . . . . . <span style="float: right;">21</span>
<b>1.2 迁移学习的概念</b>		
<b>1.3 为什么需要迁移学习?</b>		
<b>1.4 与已有概念的区别和联系</b>		
<b>1.5 负迁移</b>		
<b>2 迁移学习的研究领域</b>	<b>7</b>	<b>6 第一类方法：数据分布自适应</b> <span style="float: right;">23</span>
<b>2.1 按目标域标签分</b>		<b>6.1 边缘分布自适应</b> . . . . . <span style="float: right;">23</span>
<b>2.2 按学习方法分类</b>		<b>6.1.1 基本思路</b> . . . . . <span style="float: right;">23</span>
<b>2.3 按特征分类</b>		<b>6.1.2 核心方法</b> . . . . . <span style="float: right;">23</span>
<b>2.4 按离线与在线形式分</b>		<b>6.1.3 扩展</b> . . . . . <span style="float: right;">25</span>
<b>3 迁移学习的应用</b>	<b>9</b>	<b>6.2 条件分布自适应</b> . . . . . <span style="float: right;">26</span>
<b>3.1 计算机视觉</b>		<b>6.3 联合分布自适应</b> . . . . . <span style="float: right;">27</span>
<b>3.2 文本分类</b>		<b>6.3.1 基本思路</b> . . . . . <span style="float: right;">27</span>
<b>3.3 时间序列</b>		<b>6.3.2 核心方法</b> . . . . . <span style="float: right;">27</span>
<b>3.4 医疗健康</b>		<b>6.3.3 扩展</b> . . . . . <span style="float: right;">29</span>
<b>4 基础知识</b>	<b>12</b>	<b>6.4 小结</b> . . . . . <span style="float: right;">30</span>
<b>4.1 迁移学习的问题形式化</b>		<b>7 第二类方法：特征选择</b> <span style="float: right;">31</span>
<b>4.1.1 领域</b>		<b>7.1 核心方法</b> . . . . . <span style="float: right;">32</span>
<b>4.1.2 任务</b>		<b>7.2 扩展</b> . . . . . <span style="float: right;">32</span>
<b>4.1.3 迁移学习</b>		<b>7.3 小结</b> . . . . . <span style="float: right;">32</span>
<b>4.2 总体思路</b>		<b>8 第三类方法：子空间学习</b> <span style="float: right;">33</span>
<b>4.3 度量准则</b>		<b>8.1 统计特征对齐</b> . . . . . <span style="float: right;">33</span>
<b>4.3.1 常见的几种距离</b>		<b>8.2 流形学习</b> . . . . . <span style="float: right;">35</span>
<b>4.3.2 相似度</b>		<b>8.3 扩展与小结</b> . . . . . <span style="float: right;">37</span>
<b>4.3.3 KL 散度与 JS 距离</b>		<b>9 深度迁移学习</b> <span style="float: right;">38</span>
<b>4.3.4 最大均值差异 MMD</b>		<b>9.1 深度网络的可迁移性</b> . . . . . <span style="float: right;">38</span>
<b>4.3.5 Principal Angle</b>		<b>9.2 最简单的深度迁移：finetune</b> . . . . . <span style="float: right;">42</span>
<b>4.3.6 A-distance</b>		<b>9.3 深度网络自适应</b> . . . . . <span style="float: right;">43</span>
<b>4.3.7 Hilbert-Schmidt Independence Criterion</b>		<b>9.3.1 基本思路</b> . . . . . <span style="float: right;">43</span>
<b>4.3.8 Wasserstein Distance</b>		<b>9.3.2 核心方法</b> . . . . . <span style="float: right;">44</span>
<b>4.4 迁移学习的理论保证 *</b>		<b>9.3.3 小结</b> . . . . . <span style="float: right;">49</span>
		<b>9.4 深度对抗网络迁移</b> . . . . . <span style="float: right;">49</span>
		<b>9.4.1 基本思路</b> . . . . . <span style="float: right;">49</span>
		<b>9.4.2 核心方法</b> . . . . . <span style="float: right;">49</span>
		<b>9.4.3 小结</b> . . . . . <span style="float: right;">52</span>
		<b>10 上手实践</b> <span style="float: right;">53</span>

<b>11 迁移学习前沿</b>	<b>59</b>	<b>12 总结语</b>	<b>63</b>
11.1 机器智能与人类经验结合迁移	59		
11.2 传递式迁移学习 . . . . .	59	<b>13 附录</b>	<b>64</b>
11.3 终身迁移学习 . . . . .	60	13.1 迁移学习相关的期刊和会议 . . . . .	64
11.4 在线迁移学习 . . . . .	61	13.2 迁移学习研究学者 . . . . .	64
11.5 迁移强化学习 . . . . .	62	13.3 迁移学习资源汇总 . . . . .	67
11.6 迁移学习的可解释性 . . . . .	62	13.4 迁移学习常用算法及数据资源	68

## 写在前面

一直以来都有这样的愿望：无论学习什么知识，总是希望可以快速准确地找到对应的有价值资源进行学习。我相信我们每个人都梦寐以求。然而，越来越多的学科，尤其是我目前从事的计算机科学、人工智能领域，当下正在飞速地发展着。太多的新知识都难以事半功倍地找到快速入手的教程。庄子曰：“吾生也有涯，而知也无涯。以有涯随无涯，殆已。”

我只是迁移学习领域一个很普通的博士生，也同样经历了由“一问三不知”到“稍稍理解”的艰难过程。我在 2016 年初入门迁移学习之时，迁移学习这个概念还未曾像今天一样炙手可热。当时所能找到的学习资源只有两种：别人已发表的论文和已做过的演讲。这些还是不够简单、不够直观。我需要从如此众多的材料中不断归纳，才能站在博士研究的那个圈子的边缘，以便将来可以做出一点点贡献，往圆圈外突破一点点。

相信不只是我，任何一个刚刚入门的学习者都会经历此过程。

“沉舟侧畔千帆过，病树前头万木春。”

已所不欲，勿施于人。正是因为我初学之时也经历过如此沮丧的时期，我才在 Github 上对迁移学习进行了整理归纳，在知乎网上以“王晋东不在家”为名分享自己对于迁移学习和机器学习的理解和教训、在线上线下与大家讨论相关的问题。很欣慰的是，这些免费开放的资源或多或少地，帮助到了一些初学者，使他们更快速地步入迁移学习之门。

但这些还是不太够。Github 上的资源模式已经固定，目前主要是进行日常更新，不断加入新的论文和代码。目前还是缺乏一个人人都能上手的初学者教程。也只一次，有读者提问有没有相关的入门教程，能真正从 0 到 1 帮助初学者进行入门。

最近，南京大学博士（现任旷视科技南京研究院负责人）魏秀参学长写了一本《解析卷积神经网络—深度学习实践手册》，给很多深度学习的初学者提供了帮助。受他的启发，我也决定将自己在迁移学习领域的一些学习心得体会整理成一本手册，免费进行分享。希望能借此方式，帮助更多的初学者。我们不谈风月，只谈干货。

我不是大佬，我也是迁移学习路上的一名小学生。迁移学习领域比我做的好的同龄人太多了。因此，不敢谈什么指导。所有的目的都仅为分享。

本手册在互联网上免费开放。随着作者理解的深入（以及其他有意者的增补），本手册肯定会不断修改、越来越好。因此，我打算效仿软件的开发、采取版本更新的方式进行管理。

希望未来可以有更多的有志之士加入，让我们的教程日渐丰富。

## 致谢

本手册编写过程中得到了许多人的帮助。在此对他们表示感谢。

感谢我的导师、中国科学院计算技术研究所的陈益强研究员。是他一直以来保持着对我的信心，相信我能做出好的研究成果，不断鼓励我，经常与我讨论以明确问题，才有了今天的我。陈老师给我提供了优良的实验环境。我一定会更加努力地科研，做出更多更好的研究成果。

感谢香港科技大学计算机系的杨强教授。杨教授作为迁移学习领域国际泰斗，经常不厌其烦地回答我一些研究上的问题。能够得到杨教授的指导，是我的幸运。希望我能在杨教授带领下，做出更踏实的研究成果。

感谢新加坡南洋理工大学的于涵老师。作为我论文的共同作者，于老师认真的写作态度、对论文的把控能力是我一直学习的榜样。于老师还经常鼓励我，希望可以和于老师有着更多合作，发表更好的文章。

感谢清华大学龙明盛助理教授。龙老师在迁移学习领域发表了众多高质量的研究成果，是我入门时学习的榜样。龙老师还经常对我的研究给予指导。希望有机会可以真正和龙老师合作。

感谢美国伊利诺伊大学芝加哥分校的 Philip S. Yu 教授对我的指导和鼓励。

感谢新加坡 A\*STAR 的郝书吉老师。我博士生涯的发表的第一篇论文是和郝老师合作完成的。正是有了第一篇论文被发表，才增强了我的自信，在接下来的研究中放平心态。

感谢我的好基友、西安电子科技大学博士生段然同学和我的同病相怜，让我们可以一起吐槽读博生活。

感谢我的室友沈建飞、以及实验室同学的支持。

感谢我的知乎粉丝和所有交流过迁移学习的学者对我的支持。

最后感谢我的女友和父母对我的支持。

本手册中出现的插图，绝大多数来源于相应论文的配图。感谢这些作者做出的优秀的研究成果。希望我能早日作出可以比肩的研究。

## 说明

本手册的编写目的是帮助迁移学习领域的初学者快速进行入门。我们尽可能绕开那些非常理论的概念，只讲经验方法。我们还配有多方面的代码、数据、论文资料，最大限度地方便初学者。

本手册的方法部分，关注点是近年来持续走热的领域自适应 (Domain Adaptation) 问题。迁移学习还有其他众多的研究领域。由于作者研究兴趣所在和能力所限，对其他部分的研究只是粗略介绍。非常欢迎从事其他领域研究的读者提供内容。

本手册的每一章节都是自包含的，因此，初学者不必从头开始阅读每一部分。直接阅读自己需要的或者自己感兴趣的部分即可。本手册每一章节的信息如下：

第 1 章介绍了迁移学习的概念，重点解决什么是迁移学习、为什么要进行迁移学习这两个问题。

第 2 章介绍了迁移学习的研究领域。

第 3 章介绍了迁移学习的应用领域。

第 4 章是迁移学习领域的一些基本知识，包括问题定义，域和任务的表示，以及迁移学习的总体思路。特别地，我们提供了较为全面的度量准则介绍。度量准则是迁移学习领域重要的工具。

第 5 章简要介绍了迁移学习的四种基本方法，即基于样本迁移、基于特征迁移、基于模型迁移、基于关系迁移。

第 6 章到第 8 章，介绍了领域自适应的 3 大类基本的方法，分别是：数据分布自适应法、特征选择法、子空间学习法。

第 9 章重点介绍了目前持续最火的深度迁移学习方法。

第 10 章提供了简单的上手实践教程。

第 11 章对迁移学习进行了展望，提出了未来几个可能的研究方向。

第 12 章是对全手册的总结。

第 13 章是附录，提供了迁移学习领域相关的学习资源，以供读者参考。

由于作者水平有限，不足和错误之处，敬请不吝批评指正。

**手册的相关资源：**

网站 (内含勘误表): <http://t.cn/RmasEFe>

开发维护地址: <http://github.com/jindongwang/transferlearning-tutorial>

作者的联系方式:

邮箱: [jindongwang@outlook.com](mailto:jindongwang@outlook.com), 知乎: 王晋东不在家。

微博: 秦汉日记，个人网站:<http://jd92.wang>。

# 1 迁移学习基本概念

## 1.1 引子

冬末春初，北京的天气渐渐暖了起来。这是一句再平常不过的气候描述。对于我们在北半球生活的人来说，这似乎是一个司空见惯的现象。北京如此，纽约如此，东京如此，巴黎也如此。然而此刻，假如我问你，阿根廷的首都布宜诺斯艾利斯，天气如何？稍稍有点地理常识的人就该知道，阿根廷位于南半球，天气恰恰相反：正是夏末秋初的时候，天气渐渐凉了起来。

我们何以根据北京的天气来推测出纽约、东京和巴黎的天气？我们又何以不能用相同的方式来推测阿根廷的天气？

答案显而易见：因为它们的地理位置不同。除去阿根廷在南半球之外，其他几个城市均位于北半球，故而天气变化相似。

我们可以利用这些地点地理位置的相似性和差异性，很容易地推测出其他地点的天气。这样一个简单的事实，就引出了我们要介绍的主题：迁移学习。

## 1.2 迁移学习的概念

迁移学习，顾名思义，就是要进行迁移。放到我们人工智能和机器学习的学科里讲，迁移学习是一种学习的思想和模式。

我们都对机器学习有了基本的了解。机器学习是人工智能的一大类重要方法，也是目前发展最迅速、效果最显著的方法。机器学习解决的是让机器自主地从数据中获取知识，从而应用于新的问题中。迁移学习作为机器学习的一个重要分支，侧重于将已经学习过的知识迁移应用于新的问题中。

迁移学习的核心问题是，找到新问题和原问题之间的相似性，才可以顺利地实现知识的迁移。比如在我们一开始说的天气问题中，那些北半球的天气之所以相似，是因为它们的地理位置相似；而南北半球的天气之所以有差异，也是因为地理位置有根本不同。

其实我们人类对于迁移学习这种能力，是与生俱来的。比如，我们如果已经会打乒乓球，就可以类比着学习打网球。再比如，我们如果已经会下中国象棋，就可以类比着下国际象棋。因为这些活动之间，往往有着极高的相似性。生活中常用的“举一反三”、“照猫画虎”就很好地体现了迁移学习的思想。

回到我们的问题中来。我们用更加学术更加机器学习的语言来对迁移学习下一个定义。  
迁移学习，是指利用数据、任务、或模型之间的相似性，将在旧领域学习过的模型，应用于新领域的一种学习过程。

迁移学习最权威的综述文章是香港科技大学杨强教授团队的 A survey on transfer learning [Pan and Yang, 2010]。

图 1 简要表示了一个迁移学习过程。图 2 给出了生活中常见的迁移学习的例子。



图 1: 迁移学习示意图



图 2: 迁移学习的例子

值得一提的是，新华社报道指出，迁移学习是中国领先于世界的少数几个人工智能领域之一 [xinhua, 2016]。中国的人工智能赶超的机会来了！

### 1.3 为什么需要迁移学习？

了解了迁移学习的概念之后，紧接着还有一个非常重要的问题：迁移学习的目的是什么？或者说，为什么要用迁移学习？

我们把原因概括为以下四个方面：

#### 1. 大数据与少标注之间的矛盾。



图 3: 多种多样的数据来源

我们正处在一个大数据时代，每天每时，社交网络、智能交通、视频监控、行业物流等，都产生着海量的图像、文本、语音等各类数据。数据的增多，使得机器学习和深度学习模型可以依赖于如此海量的数据，持续不断地训练和更新相应的模型，使得模型的性能越来越好，越来越适合特定场景的应用。然而，这些大数据带来了严重的问题：总是缺乏完善的数  
据标注。

众所周知，机器学习模型的训练和更新，均依赖于数据的标注。然而，尽管我们可以获取到海量的数据，这些数据往往是很初级的原始形态，很少有数据被加以正确的人工标注。数据的标注是一个耗时且昂贵的操作，目前为止，尚未有行之有效的方式来解决这一问题。这给机器学习和深度学习的模型训练和更新带来了挑战。反过来说，特定的领域，因为没有足够的标定数据用来学习，使得这些领域一直不能很好的发展。

#### 2. 大数据与弱计算之间的矛盾。

大数据，就需要大设备、强计算能力的设备来进行存储和计算。然而，大数据的大计算能力，是“有钱人”才能玩得起的游戏。比如 Google, Facebook, Microsoft，这些巨无霸公司有着雄厚的计算能力去利用这些数据训练模型。例如，ResNet 需要很长的时间进行训练。Google TPU 也都是有钱人的才可以用得起的。

绝大多数普通用户是不可能具有这些强计算能力的。这就引发了大数据和弱计算之间的矛盾。在这种情况下，普通人想要利用这些海量的大数据去训练模型完成自己的任务，基本上不太可能。那么如何让普通人也能利用这些数据和模型？

#### 3. 普适化模型与个性化需求之间的矛盾。

机器学习的目标是构建一个尽可能通用的模型，使得这个模型对于不同用户、不同设备、不同环境、不同需求，都可以很好地进行满足。这是我们的美好愿景。这就是要尽可能



图 4: 大数据与强计算能力

地提高机器学习模型的泛化能力，使之适应不同的数据情形。基于这样的愿望，我们构建了多种多样的普适化模型，来服务于现实应用。然而，这只能是我们竭尽全力想要做的，目前却始终无法彻底解决的问题。人们的个性化需求五花八门，短期内根本无法用一个通用的模型去满足。比如导航模型，可以定位及导航所有的路线。但是不同的人有不同的需求。比如有的人喜欢走高速，有的人喜欢走偏僻小路，这就是个性化需求。并且，不同的用户，通常都有不同的隐私需求。这也是构建应用需要着重考虑的。

所以目前的情况是，我们对于每一个通用的任务都构建了一个通用的模型。这个模型可以解决绝大多数的公共问题。但是具体到每个个体、每个需求，都存在其唯一性和特异性，一个普适化的通用模型根本无法满足。那么，能否将这个通用的模型加以改造和适配，使其更好地服务于人们的个性化需求？

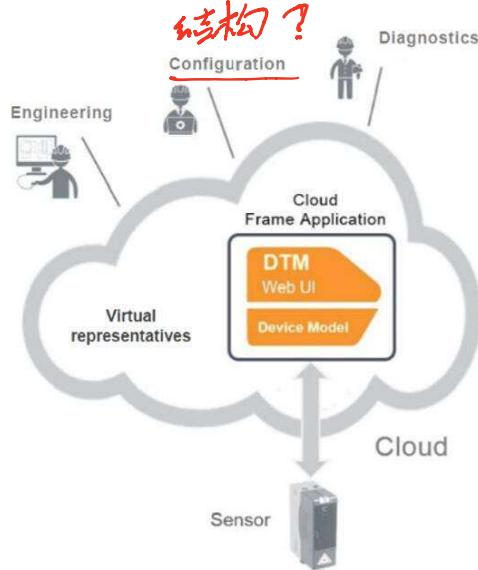


图 5: 普适化模型与个性化需求

#### 4. 特定应用的需求。

机器学习已经被广泛应用于现实生活中。在这些应用中，也存在着一些特定的应用，它们面临着一些现实存在的问题。比如推荐系统的冷启动问题。一个新的推荐系统，没有足够的用户数据，如何进行精准的推荐？一个崭新的图片标注系统，没有足够的标签，如何进行精准的服务？现实世界中的应用驱动着我们去开发更加便捷更加高效的机器学习方法来加以解决。

上述存在的几个重要问题，使得传统的机器学习方法疲于应对。迁移学习则可以很好



图 6: 特定应用需求: 冷启动

地进行解决。那么, **迁移学习是如何进行解决的呢?**

#### 1. 大数据与少标注: 迁移数据标注

单纯地凭借少量的标注数据, 无法准确地训练高可用度的模型。为了解决这个问题, 我们直观的想法是: 多增加一些标注数据不就行了? 但是不依赖于人工, 如何增加标注数据?

利用迁移学习的思想, 我们可以寻找一些与目标数据相近的有标注的数据, 从而利用这些数据来构建模型, 增加我们目标数据的标注。

#### 2. 大数据与弱计算: 模型迁移

不可能所有人都有能力利用大数据快速进行模型的训练。利用迁移学习的思想, 我们可以将那些大公司在大数据上训练好的模型, 迁移到我们的任务中。针对于我们的任务进行微调, 从而我们也可以拥有在大数据上训练好的模型。更进一步, 我们可以将这些模型针对我们的任务进行自适应更新, 从而取得更好的效果。

#### 3. 普适化模型与个性化需求: 自适应学习

为了解决个性化需求的挑战, 我们利用迁移学习的思想, 进行自适应的学习。考虑到不同用户之间的相似性和差异性, 我们对普适化模型进行灵活的调整, 以便完成我们的任务。

#### 4. 特定应用的需求: 相似领域知识迁移

为了满足特定领域应用的需求, 我们可以利用上述介绍过的手段, 从数据和模型方法上进行迁移学习。

表1概括地描述了迁移学习的必要性。

表 1: 迁移学习的必要性

矛盾	传统机器学习	迁移学习
大数据与少标注	增加人工标注, 但是昂贵且耗时	数据的迁移标注
大数据与弱计算	只能依赖强大计算能力, 但是受众少	模型迁移
普适化模型与个性化需求	通用模型无法满足个性化需求	模型自适应调整
特定应用	冷启动问题无法解决	数据迁移

## 1.4 与已有概念的区别和联系

迁移学习并不是一个横空出世的概念, 它与许多已有的概念都有些联系, 但是也有着一些区别。我们在这里汇总一些与迁移学习非常接近的概念, 并简述迁移学习与它们的区别和联系。

#### 1. 迁移学习 VS 传统机器学习:

迁移学习属于机器学习的一类, 但它在如下几个方面有别于传统的机器学习 (表 2):

#### 2. 迁移学习 VS 多任务学习:

表 2: 传统机器学习与迁移学习的区别

比较项目	传统机器学习	迁移学习
数据分布	训练和测试数据服从相同的分布	训练和测试数据服从不同的分布
数据标注	需要足够的数据标注来训练模型	不需要足够的数据标注
模型	每个任务分别建模	模型可以在不同任务之间迁移

**多任务学习**指多个相关的任务一起协同学习；迁移学习则强调知识由一个领域迁移到另一个领域的过程。**迁移是思想，多任务是其中的一个具体形式。**

### 3. 迁移学习 VS 终身学习：

**终身学习**可以认为是**序列化的多任务学习**，在已经学习好若干个任务之后，面对新的任务可以继续学习而不遗忘之前学习的任务。**迁移学习则侧重于模型的迁移和共同学习。**

### 4. 迁移学习 VS 领域自适应：

**领域自适应问题**是迁移学习的研究内容之一，它侧重于解决**特征空间一致、类别空间一致，仅特征分布不一致的问题**。而迁移学习也可以解决上述内容不一致的情况。

### 5. 迁移学习 VS 增量学习：

**增量学习**侧重解决**数据不断到来，模型不断更新的问题**。迁移学习显然和其有着不同之处。

### 6. 迁移学习 VS 自我学习：

**自我学习**指的是**模型不断地从自身处进行更新**，而迁移学习强调知识在不同的领域间进行迁移。

### 7. 迁移学习 VS 协方差漂移

协方差漂移指数据的边缘概率分布发生变化。领域自适应研究问题解决的就是协方差漂移现象。

## 1.5 负迁移

我们都希望迁移学习能够比较顺利地进行，我们得到的结果也是满足我们要求的，皆大欢喜。然而，事情却并不总是那么顺利。这就引入了迁移学习中的一个负面现象，也就是所谓的**负迁移**。

用我们熟悉的成语来描述：如果说成功的迁移学习是“举一反三”、“照猫画虎”，那么负迁移则是“东施效颦”。东施已经模仿西施捂着胸口皱着眉头，为什么她还是那么丑？

要理解负迁移，首先要理解什么是迁移学习。**迁移学习**指的是，利用**数据和领域之间存在的相似性关系，把之前学习到的知识，应用于新的未知领域**。**迁移学习的核心问题是，找到两个领域的相似性**。找到了这个相似性，就可以合理地利用，从而很好地完成迁移学习任务。比如，之前会骑自行车，要学习骑摩托车，这种相似性指的就是自行车和摩托车之间的相似性以及骑车体验的相似性。这种相似性在我们人类看来是可以接受的。

所以，如果这个相似性找的不合理，也就是说，两个领域之间不存在相似性，或者基本不相似，那么，就会大大损害迁移学习的效果。还是拿骑自行车来说，你要拿骑自行车的经验来学习开汽车，这显然是不太可能的。因为自行车和汽车之间基本不存在什么相似性。所以，这个任务基本上完不成。这时候，我们可以说出现了**负迁移 (Negative Transfer)**。

所以，为什么东施和西施做了一样的动作，反而变得更丑了？因为东施和西施之间压根就不存在相似性。

迁移学习领域权威学者、香港科技大学杨强教授发表的迁移学习的综述文章 A survey on transfer learning [Pan and Yang, 2010] 给出了负迁移的一个定义：

负迁移指的是，在源域上学习到的知识，对于目标域上的学习产生负面作用。

文章也引用了一些经典的解决负迁移问题的文献。但是普遍较老，这里就不说了。

所以，产生负迁移的原因主要有：

- 数据问题：源域和目标域压根不相似，谈何迁移？
- 方法问题：源域和目标域是相似的，但是，迁移学习方法不够好，没找到可迁移的成分。

负迁移给迁移学习的研究和应用带来了负面影响。在实际应用中，找到合理的相似性，并且选择或开发合理的迁移学习方法，能够避免负迁移现象。

### 最新的研究成果

随着研究的深入，已经有新的研究成果在逐渐克服负迁移的影响。杨强教授团队 2015 在数据挖掘领域顶级会议 KDD 上发表了传递迁移学习文章 Transitive transfer learning [Tan et al., 2015]，提出了传递迁移学习的思想。传统迁移学习就好比是踩着一块石头过河，传递迁移学习就好比是踩着连续的两块石头。

更进一步，杨强教授团队在 2017 年人工智能领域顶级会议 AAAI 上发表了远领域迁移学习的文章 Distant domain transfer learning [Tan et al., 2017]，可以用人脸识别飞机！这就好比是踩着一连串石头过河。这些研究的意义在于，传统迁移学习只有两个领域足够相似才可以完成，而当两个领域不相似时，传递迁移学习却可以利用处于这两个领域之间的若干领域，将知识传递式的完成迁移。这个是很有意义的工作，可以视为解决负迁移的有效思想和方法。可以预见在未来会有更多的应用前景。

图 7 对传递迁移学习给出了简明的示意。

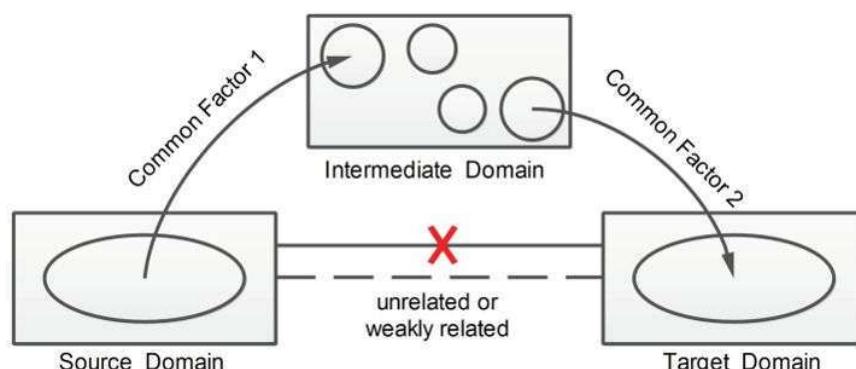


图 7：传递式迁移学习示意图

## 2 迁移学习的研究领域

依据目前较流行的机器学习分类方法，机器学习主要可以分为有监督、半监督和无监督机器学习三大类。同理，迁移学习也可以进行这样的分类。需要注意的是，依据的分类准则不同，分类结果也不同。在这一点上，并没有一个统一的说法。我们在这里仅根据目前较流行的方法，对迁移学习的研究领域进行一个大致的划分。

图 8 给出了迁移学习的常用分类方法总结。

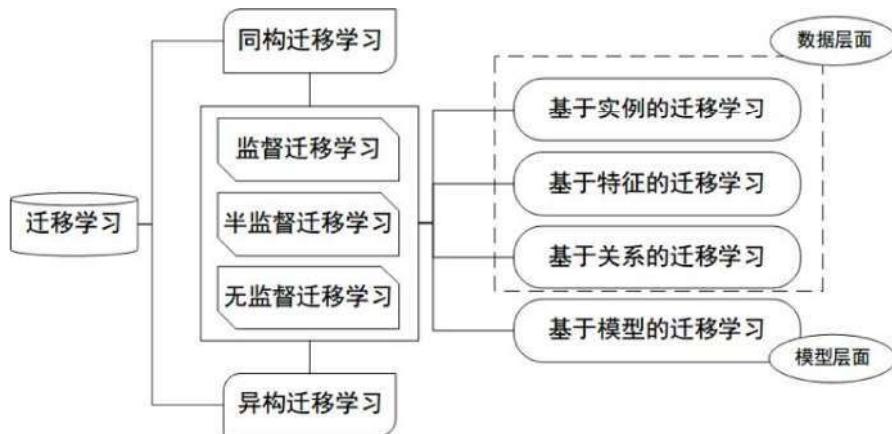


图 8: 迁移学习的研究领域与研究方法分类

大体上讲，迁移学习的分类可以按照四个准则进行：按目标域有无标签分、按学习方法分、按特征分、按离线与在线形式分。不同的分类方式对应着不同的专业名词。当然，即使是一个分类下的研究领域，也可能同时处于另一个分类下。下面我们将对这些分类方法及相应的领域作简单描述。

### 2.1 按目标域标签分

这种分类方式最为直观。类比机器学习，按照目标领域有无标签，迁移学习可以分为以下三个大类：

1. 监督迁移学习 (Supervised Transfer Learning)
2. 半监督迁移学习 (Semi-Supervised Transfer Learning)
3. 无监督迁移学习 (Unsupervised Transfer Learning)

显然，少标签或无标签的问题（半监督和无监督迁移学习），是研究的热点和难点。这也是本手册重点关注的领域。

### 2.2 按学习方法分类

按学习方法的分类形式，最早在迁移学习领域的权威综述文章 [Pan and Yang, 2010] 给出定义。它将迁移学习方法分为以下四个大类：

1. 基于样本的迁移学习方法 (Instance based Transfer Learning)
2. 基于特征的迁移学习方法 (Feature based Transfer Learning)

3. 基于模型的迁移学习方法 (Model based Transfer Learning)
4. 基于关系的迁移学习方法 (Relation based Transfer Learning)

这是一个很直观的分类方式，按照数据、特征、模型的机器学习逻辑进行区分，再加上不属于这三者中的关系模式。

**基于实例的迁移**，简单来说就是通过权重重用，对源域和目标域的样例进行迁移。就是说直接对不同的样本赋予不同权重，比如说相似的样本，我就给它高权重，这样我就完成了迁移，非常简单非常直接。

**基于特征的迁移**，就是更进一步对特征进行变换。意思是说，假设源域和目标域的特征原来不在一个空间，或者说它们在原来那个空间上不相似，那我们就想办法把它们变换到一个空间里面，那这些特征不就相似了？这个思路也非常直接。这个方法是用得非常多的，一直在研究，目前是感觉是研究最热的。

**基于模型的迁移**，就是说构建参数共享的模型。这个主要就是在神经网络里面用的特别多，因为神经网络的结构可以直接进行迁移。比如说神经网络最经典的 finetune 就是模型参数迁移的很好的体现。

**基于关系的迁移**，这个方法用的比较少，这个主要就是说挖掘和利用关系进行类比迁移。比如老师上课、学生听课就可以类比为公司开会的场景。这个就是一种关系的迁移。

目前最热的就是基于特征还有模型的迁移，然后基于实例的迁移方法和他们结合起来使用。

迁移学习方法是本手册的重点。我们在后续的篇幅中介绍。

### 2.3 按特征分类

按照特征的属性进行分类，也是一种常用的分类方法。这在最近的迁移学习综述 [Weiss et al., 2016] 中给出。按照特征属性，迁移学习可以分为两个大类：

1. 同构迁移学习 (Homogeneous Transfer Learning)
2. 异构迁移学习 (Heterogeneous Transfer Learning)

这也是一种很直观的方式：如果特征语义和维度都相同，那么就是同构；反之，如果特征完全不相同，那么就是异构。举个例子来说，不同图片的迁移，就可以认为是同构；而图片到文本的迁移，则是异构的。

### 2.4 按离线与在线形式分

按照离线学习与在线学习的方式，迁移学习还可以被分为：

1. 离线迁移学习 (Offline Transfer Learning)
2. 在线迁移学习 (Online Transfer Learning)

目前，绝大多数的迁移学习方法，都采用了离线方式。即，源域和目标域均是给定的，迁移一次即可。这种方式的缺点是显而易见的：算法无法对新加入的数据进行学习，模型也无法得到更新。与之相对的，是在线的方式。即随着数据的动态加入，迁移学习算法也可以不断地更新。

### 3 迁移学习的应用

迁移学习是机器学习领域的一个重要分支。因此，其应用并不局限于特定的领域。凡是满足迁移学习问题情景的应用，迁移学习都可以发挥作用。这些领域包括但不限于计算机视觉、文本分类、行为识别、自然语言处理、室内定位、视频监控、舆情分析、人机交互等。图 9 展示了迁移学习可能的应用领域。

下面我们选择几个研究热点，对迁移学习在这些领域的应用场景作一简单介绍。



图 9: 迁移学习的应用领域概览

#### 3.1 计算机视觉

迁移学习已被广泛地应用于计算机视觉的研究中。特别地，在计算机视觉中，迁移学习方法被称为 Domain Adaptation。Domain adaptation 的应用场景有很多，比如图片分类、图片哈希等。

图 10 展示了不同的迁移学习图片分类任务示意。同一类图片，不同的拍摄角度、不同光照、不同背景，都会造成特征分布发生改变。因此，使用迁移学习构建跨领域的鲁棒分类器是十分重要的。



图 10: 迁移学习图片分类任务

计算机视觉三大顶会 (CVPR、ICCV、ECCV) 每年都会发表大量的文章对迁移学习在视觉领域的应用进行介绍。

#### 3.2 文本分类

由于文本数据有其领域特殊性，因此，在一个领域上训练的分类器，不能直接拿来作用到另一个领域上。这就需要用到迁移学习。例如，在电影评论文本数据集上训练好的分类

器，不能直接用于图书评论的预测。这就需要进行迁移学习。图 11 是一个由电子产品评论迁移到 DVD 评论的迁移学习任务。

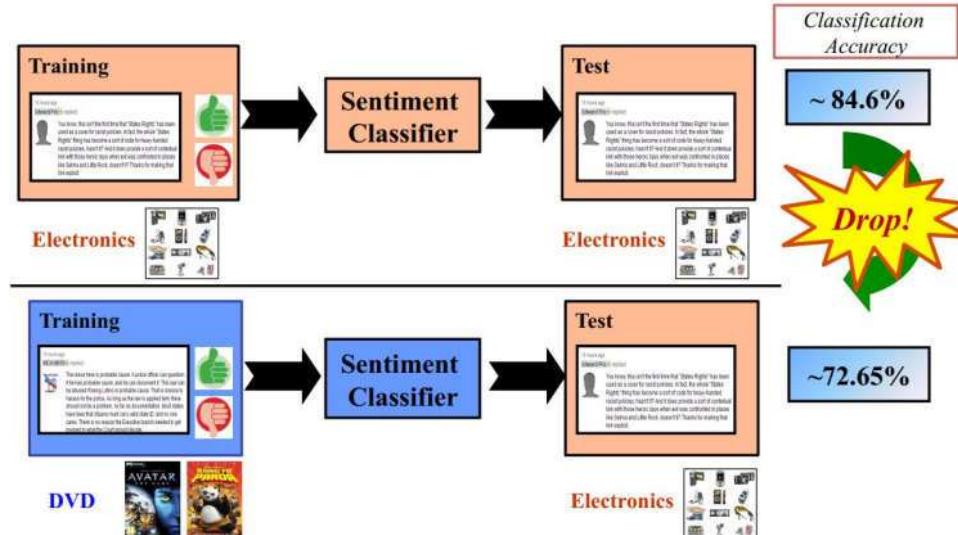


图 11: 迁移学习文本分类任务

文本和网络领域顶级会议 WWW 和 CIKM 每年有大量的文章对迁移学习在文本领域的应用作介绍。

### 3.3 时间序列

行为识别 (Activity Recognition) 主要通过佩戴在用户身体上的传感器，研究用户的行为。行为数据是一种时间序列数据。不同用户、不同环境、不同位置、不同设备，都会导致时间序列数据的分布发生变化。此时，也需要进行迁移学习。图 12 展示了同一用户不同位置的信号差异性。在这个领域，华盛顿州立大学的 Diane Cook 等人在 2013 年发表的关于迁移学习在行为识别领域的综述文章 [Cook et al., 2013] 是很好的参考资料。

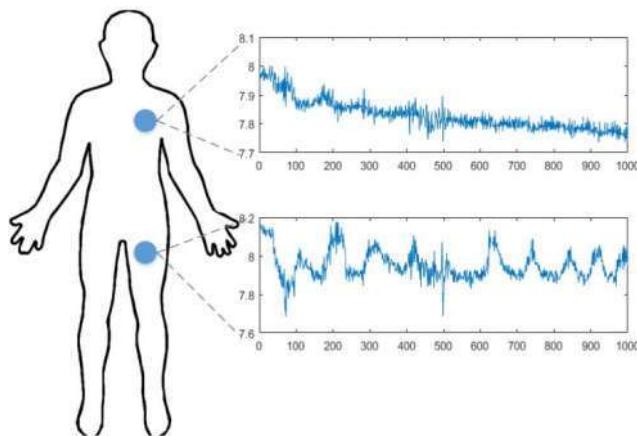


图 12: 不同位置的传感器信号差异示意图

室内定位 (Indoor Location) 与传统的室外用 GPS 定位不同，它通过 WiFi、蓝牙等设备研究人在室内的位置。不同用户、不同环境、不同时刻也会使得采集的信号分布发生变化。图 13 展示了不同时间、不同设备的 WiFi 信号变化。

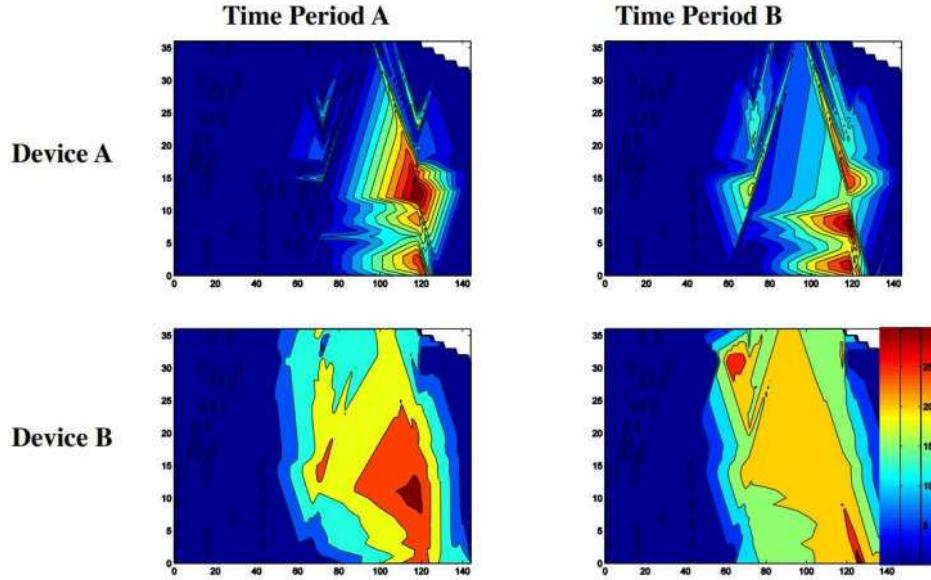


图 13: 室内定位由于时间和设备的变化导致的信号变化

### 3.4 医疗健康

医疗健康领域的研究正变得越来越重要。不同于其他领域，医疗领域研究的难点问题是，无法获取足够有效的医疗数据。在这一领域，迁移学习同样也变得越来越重要。

最近，顶级生物期刊细胞杂志报道了由张康教授领导的广州妇女儿童医疗中心和加州大学圣迭戈分校团队的重磅研究成果：基于深度学习开发出一个能诊断眼病和肺炎两大类疾病的 AI 系统 [Kermany et al., 2018]，准确性匹敌顶尖医生。这不仅是中国研究团队首次在顶级生物医学杂志发表有关医学人工智能的研究成果；也是世界范围内首次使用如此庞大的标注好的高质量数据进行迁移学习，并取得高度精确的诊断结果，达到匹敌甚至超越人类医生的准确性；还是全世界首次实现用 AI 精确推荐治疗手段。细胞杂志封面报道了该研究成果。

我们可以预见到的是，迁移学习对于那些不易获取标注数据的领域，将会发挥越来越重要的作用。

## 4 基础知识

本部分介绍迁移学习领域的一些基本知识。我们对迁移学习的问题进行简单形式化，给出迁移学习的总体思路，并且介绍目前常用的一些度量准则。本部分中出现的所有符号和表示形式，是以后章节的基础。已有相关知识的读者可以直接跳过。

### 4.1 迁移学习的问题形式化

迁移学习的问题形式化，是进行一切研究的前提。在迁移学习中，有两个基本的概念：领域 (Domain) 和 任务 (Task)。它们是最基础的概念。定义如下：

#### 4.1.1 领域

**领域 (Domain):** 是进行学习的主体。领域主要由两部分构成：数据和生成这些数据的概率分布。通常我们用花体  $\mathcal{D}$  来表示一个 domain，用大写斜体  $P$  来表示一个 概率分布。

特别地，因为涉及到迁移，所以对应于两个基本的领域：源领域 (Source Domain) 和 目标领域 (Target Domain)。这两个概念很好理解。源领域就是有知识、有大量数据标注的领域，是我们要迁移的对象；目标领域就是我们最终要赋予知识、赋予标注的对象。知识从源领域传递到目标领域，就完成了迁移。

领域上的数据，我们通常用小写粗体  $\mathbf{x}$  来表示，它也是向量的表示形式。例如， $\mathbf{x}_i$  就表示第  $i$  个样本或特征。用大写的黑体  $\mathbf{X}$  表示一个领域的数据，这是一种矩阵形式。我们用大写花体  $\mathcal{X}$  来表示数据的特征空间。

通常我们用小写下标  $s$  和  $t$  来分别指代两个领域。结合领域的表示方式，则： $\mathcal{D}_s$  表示源领域， $\mathcal{D}_t$  表示目标领域。

值得注意的是，概率分布  $P$  通常只是一个逻辑上的概念，即我们认为不同领域有不同的概率分布，却一般不给出（也难以给出） $P$  的具体形式。

#### 4.1.2 任务

**任务 (Task):** 是学习的目标。任务主要由两部分组成：标签和标签对应的函数。通常我们用花体  $\mathcal{Y}$  来表示一个标签空间，用  $f(\cdot)$  来表示一个学习函数。

相应地，源领域和目标领域的类别空间就可以分别表示为  $\mathcal{Y}_s$  和  $\mathcal{Y}_t$ 。我们用小写  $y_s$  和  $y_t$  分别表示源领域和目标领域的实际类别。

#### 4.1.3 迁移学习

有了上面领域和任务的定义，我们就可以对迁移学习进行形式化。

**迁移学习 (Transfer Learning):** 给定一个有标记的源域  $\mathcal{D}_s = \{\mathbf{x}_i, y_i\}_{i=1}^n$  和一个无标记的目标域  $\mathcal{D}_t = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$ 。这两个领域的数据分布  $P(\mathbf{x}_s)$  和  $P(\mathbf{x}_t)$  不同，即  $P(\mathbf{x}_s) \neq P(\mathbf{x}_t)$ 。迁移学习的目的就是要借助  $\mathcal{D}_s$  的知识，来学习目标域  $\mathcal{D}_t$  的知识（标签）。

更进一步，结合我们前面说过的迁移学习研究领域，迁移学习的定义需要进行如下的考虑：

- # } (1) 特征空间的异同，即  $\mathcal{X}_s$  和  $\mathcal{X}_t$  是否相等。
- (2) 类别空间的异同：即  $\mathcal{Y}_s$  和  $\mathcal{Y}_t$  是否相等。
- (3) 条件概率分布的异同：即  $Q_s(y_s|\mathbf{x}_s)$  和  $Q_t(y_t|\mathbf{x}_t)$  是否相等。

结合上述形式化，我们给出领域自适应 (Domain Adaptation)这一热门研究方向的定义：

**领域自适应 (Domain Adaptation):** 给定一个有标记的源域  $\mathcal{D}_s = \{\mathbf{x}_i, y_i\}_{i=1}^n$  和一个无标记的目标域  $\mathcal{D}_t = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$ ，假定它们的特征空间相同，即  $\mathcal{X}_s = \mathcal{X}_t$ ，并且它们的类别空间也相同，即  $\mathcal{Y}_s = \mathcal{Y}_t$  以及条件概率分布也相同，即  $Q_s(y_s|\mathbf{x}_s) = Q_t(y_t|\mathbf{x}_t)$ 。但是这两个域的边缘分布不同，即  $P_s(\mathbf{x}_s) \neq P_t(\mathbf{x}_t)$ 。迁移学习的目标就是，利用有标记的数据  $\mathcal{D}_s$  去学习一个分类器  $f : \mathbf{x}_t \mapsto y_t$  来预测目标域  $\mathcal{D}_t$  的标签  $y_t \in \mathcal{Y}_t$ 。

在实际的研究和应用中，读者可以针对自己的不同任务，结合上述表述，灵活地给出相关的形式化定义。

### 符号小结

我们已经基本介绍了迁移学习中常用的符号。表 3 是一个符号表：

表 3: 迁移学习形式化表示常用符号

符号	含义
下标 $s / t$	指示源域 / 目标域
$\mathcal{D}_s / \mathcal{D}_t$	源域数据 / 目标域数据
$\mathbf{x} / \mathbf{X} / \mathcal{X}$	向量 / 矩阵 / 特征空间
$\mathbf{y} / \mathcal{Y}$	类别向量 / 类别空间
$(n, m)$ [或 $(n_1, n_2)$ 或 $(n_s, n_t)$ ]	(源域样本数, 目标域样本数)
$P(\mathbf{x}_s) / P(\mathbf{x}_t)$	源域数据 / 目标域数据的边缘分布
$Q(\mathbf{y}_s \mathbf{x}_s) / Q(\mathbf{y}_t \mathbf{x}_t)$	源域数据 / 目标域数据的条件分布
$f(\cdot)$	要学习的目标函数

## 4.2 总体思路

形式化之后，我们可以进行迁移学习的研究。迁移学习的总体思路可以概括为：开发算法来最大限度地利用有标注的领域的知识，来辅助目标领域的知识获取和学习。

迁移学习的核心是，找到源领域和目标领域之间的相似性，并加以合理利用。这种相似性非常普遍。比如，不同人的身体构造是相似的；自行车和摩托车的骑行方式是相似的；国际象棋和中国象棋是相似的；羽毛球和网球的打球方式是相似的。这种相似性也可以理解为不变量。以不变应万变，才能立于不败之地。

举一个杨强教授经常举的例子来说明：我们都应该知道在中国大陆开车时，驾驶员坐在左边，靠马路右侧行驶。这是基本的规则。然而，如果在英国、香港等地区开车，驾驶员是坐在右边，需要靠马路左侧行驶。那么，如果我们从中国大陆到了香港，应该如何快速地适应他们的开车方式呢？诀窍就是找到这里的不变量：不论在哪个地区，驾驶员都是紧靠马路中间。这就是我们这个开车问题中的不变量。

找到相似性（不变量），是进行迁移学习的核心。

有了这种相似性后，下一步工作就是，如何度量和利用这种相似性。度量工作的目标有两点：一是很好地度量两个领域的相似性，不仅定性地告诉我们它们是否相似，更定量地给出相似程度。二是以度量为准则，通过我们所要采用的学习手段，增大两个领域之间的相似性，从而完成迁移学习。

一句话总结：相似性是核心，度量准则是重要手段。



### 4.3 度量准则

度量不仅是机器学习和统计学等学科中使用的基础手段，也是迁移学习中的重要工具。它的核心就是衡量两个数据域的差异。计算两个向量（点、矩阵）的距离和相似度是许多机器学习算法的基础，有时候一个好的距离度量就能决定算法最后的结果好坏。比如 KNN 分类算法就对距离非常敏感。本质上就是找一个变换使得源域和目标域的距离最小（相似度最大）。所以，相似度和距离度量在机器学习中非常重要。

这里给出常用的度量手段，它们都是迁移学习研究中非常常见的度量准则。对这些准则有很好的理解，可以帮助我们设计出更加好用的算法。用一个简单的式子来表示，度量就是描述源域和目标域这两个领域的距离：

$$DISTANCE(\mathcal{D}_s, \mathcal{D}_t) = \text{DistanceMeasure}(\cdot, \cdot) \quad (4.1)$$

下面我们从距离和相似度度量准则几个方面进行简要介绍。

#### 4.3.1 常见的几种距离

##### 1. 欧氏距离

定义在两个向量（空间中的两个点）上：点  $\mathbf{x}$  和点  $\mathbf{y}$  的欧氏距离为：

$$d_{Euclidean} = \sqrt{(\mathbf{x} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y})} \quad (4.2)$$

##### 2. 闵可夫斯基距离

Minkowski distance，两个向量（点）的  $p$  阶距离：

$$d_{Minkowski} = (||\mathbf{x} - \mathbf{y}||^p)^{1/p} \quad (4.3)$$

当  $p = 1$  时就是曼哈顿距离，当  $p = 2$  时就是欧氏距离。

##### 3. 马氏距离

定义在两个向量（两个点）上，这两个数据在同一个分布里。点  $\mathbf{x}$  和点  $\mathbf{y}$  的马氏距离为：

$$d_{Mahalanobis} = \sqrt{(\mathbf{x} - \mathbf{y})^\top \Sigma^{-1} (\mathbf{x} - \mathbf{y})} \quad (4.4)$$

其中， $\Sigma$  是这个分布的协方差。

当  $\Sigma = \mathbf{I}$  时，马氏距离退化为欧氏距离。

#### 4.3.2 相似度

##### 1. 余弦相似度

衡量两个向量的相关性（夹角的余弦）。向量  $\mathbf{x}, \mathbf{y}$  的余弦相似度为：

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|} \quad (4.5)$$

##### 2. 互信息

定义在两个概率分布  $X, Y$  上， $x \in X, y \in Y$ 。它们的互信息为：

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (4.6)$$

### 3. 皮尔逊相关系数

衡量两个随机变量的相关性。随机变量  $X, Y$  的 Pearson 相关系数为：

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad (4.7)$$

理解：协方差矩阵除以标准差之积。

范围： $[-1, 1]$ ，绝对值越大表示（正/负）相关性越大。

### 4. Jaccard 相关系数

对两个集合  $X, Y$ ，判断他们的相关性，借用集合的手段：

$$J = \frac{|X \cap Y|}{|X \cup Y|} \quad (4.8)$$

理解：两个集合的交集除以并集。

扩展：Jaccard 距离  $= 1 - J$ 。

#### 4.3.3 KL 散度与 JS 距离

KL 散度和 JS 距离是迁移学习中被广泛应用的度量手段。

##### 1. KL 散度

Kullback-Leibler divergence，又叫做相对熵，衡量两个概率分布  $P(x), Q(x)$  的距离：

$$D_{KL}(P||Q) = \sum_{i=1} P(x) \log \frac{P(x)}{Q(x)} \quad (4.9)$$

这是一个非对称距离： $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ .

##### 2. JS 距离

Jensen-Shannon divergence，基于 KL 散度发展而来，是对称度量：

$$JSD(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M) \quad (4.10)$$

其中  $M = \frac{1}{2}(P + Q)$ 。

#### 4.3.4 最大均值差异 MMD

最大均值差异是迁移学习中使用频率最高的度量。Maximum mean discrepancy，它度量在再生希尔伯特空间中两个分布的距离，是一种核学习方法。两个随机变量的 MMD 平方距离为

$$MMD^2(X, Y) = \left\| \sum_{i=1}^{n_1} \phi(\mathbf{x}_i) - \sum_{j=1}^{n_2} \phi(\mathbf{y}_j) \right\|_{\mathcal{H}}^2 \quad (4.11)$$

其中  $\phi(\cdot)$  是映射，用于把原变量映射到再生核希尔伯特空间 (Reproducing Kernel Hilbert Space, RKHS) [Borgwardt et al., 2006] 中。什么是 RKHS？形式化定义太复杂，简单来说希尔伯特空间是对于函数的内积完备的，而再生核希尔伯特空间是具有再生性  $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}} = K(x, y)$  的希尔伯特空间。就是比欧几里得空间更高端的。将平方展开后，RKHS 空间中的内积就可以转换成核函数，所以最终 MMD 可以直接通过核函数进行计算。

理解：就是求两堆数据在 RKHS 中的均值的距离。

*Multiple-kernel MMD:* 多核的 MMD，简称 MK-MMD。现有的 MMD 方法是基于单一核变换的，多核的 MMD 假设最优的核可以由多个核线性组合得到。多核 MMD 的提出和计算方法在文献 [Gretton et al., 2012] 中形式化给出。MK-MMD 在许多后来的方法中被大量使用，最著名的方法是 DAN [Long et al., 2015a]。我们将在后续单独介绍此工作。

### 4.3.5 Principal Angle

也是将两个分布映射到高维空间 (格拉斯曼流形) 中，在流形中两堆数据就可以看成两个点。Principal angle 是求这两堆数据的对应维度的夹角之和。

对于两个矩阵  $\mathbf{X}, \mathbf{Y}$ ，计算方法：首先正交化 (用 PCA) 两个矩阵，然后：

$$PA(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{\min(m,n)} \sin \theta_i \quad (4.12)$$

其中  $m, n$  分别是两个矩阵的维度， $\theta_i$  是两个矩阵第  $i$  个维度的夹角， $\Theta = \{\theta_1, \theta_2, \dots, \theta_t\}$  是两个矩阵 SVD 后的角度：

$$\mathbf{X}^\top \mathbf{Y} = \mathbf{U}(\cos \Theta) \mathbf{V}^\top \quad (4.13)$$

### 4.3.6 A-distance

$A$ -distance 是一个很简单却很有用的度量。文献 [Ben-David et al., 2007] 介绍了此距离，它可以用来估计不同分布之间的差异性。 $A$ -distance 被定义为建立一个线性分类器来区分两个数据领域的 hinge 损失 (也就是进行二类分类的 hinge 损失)。它的计算方式是，我们首先在源域和目标域上训练一个二分类器  $h$ ，使得这个分类器可以区分样本是来自于哪一个领域。我们用  $err(h)$  来表示分类器的损失，则  $A$ -distance 定义为：

$$\mathcal{A}(\mathcal{D}_s, \mathcal{D}_t) = 2(1 - 2err(h)) \quad (4.14)$$

$A$ -distance 通常被用来计算两个领域数据的相似性程度，以便与实验结果进行验证对比。

### 4.3.7 Hilbert-Schmidt Independence Criterion

希尔伯特-施密特独立性系数，Hilbert-Schmidt Independence Criterion，用来检验两组数据的独立性：

$$HSIC(X, Y) = \text{trace}(HXHY) \quad (4.15)$$

其中  $X, Y$  是两堆数据的 kernel 形式。

### 4.3.8 Wasserstein Distance

Wasserstein Distance 是一套用来衡量两个概率分部之间距离的度量方法。该距离在一个度量空间  $(M, \rho)$  上定义，其中  $\rho(x, y)$  表示集合  $M$  中两个实例  $x$  和  $y$  的距离函数，比如欧几里得距离。两个概率分布  $\mathbb{P}$  和  $\mathbb{Q}$  之间的  $p$ -th Wasserstein distance 可以被定义为

$$W_p(\mathbb{P}, \mathbb{Q}) = \left( \inf_{\mu \in \Gamma(\mathbb{P}, \mathbb{Q})} \int \rho(x, y)^p d\mu(x, y) \right)^{1/p}, \quad (4.16)$$

其中  $\Gamma(\mathbb{P}, \mathbb{Q})$  是在集合  $M \times M$  内所有的以  $\mathbb{P}$  和  $\mathbb{Q}$  为边缘分布的联合分布。著名的 Kantorovich-Rubinstein 定理表示当  $M$  是可分离的时候，第一 Wasserstein distance 可以等价地表示成一个积分概率度量 (integral probability metric) 的形式

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}}[f(x)] - \mathbb{E}_{x \sim \mathbb{Q}}[f(x)], \quad (4.17)$$

其中  $\|f\|_L = \sup |f(x) - f(y)|/\rho(x, y)$  并且  $\|f\|_L \leq 1$  称为 1- 利普希茨条件。

#### 4.4 迁移学习的理论保证 \*

本部分的标题中带有 \* 号，有一些难度，为可看可不看的内容。此部分最常见的形式是当自己提出的算法需要理论证明时，可以借鉴。

在第一章里我们介绍了两个重要的概念：迁移学习是什么，以及为什么需要迁移学习。但是，还有一个重要的问题没有得到解答：为什么可以进行迁移？也就是说，迁移学习的可行性还没有探讨。

值得注意的是，就目前的研究成果来说，迁移学习领域的理论工作非常匮乏。我们在这里仅回答一个问题：为什么数据分布不同的两个领域之间，知识可以进行迁移？或者说，到底达到什么样的误差范围，我们才认为知识可以进行迁移？

加拿大滑铁卢大学的 Ben-David 等人从 2007 年开始，连续发表了三篇文章 [Ben-David et al., 2007, Blitzer et al., 2008, Ben-David et al., 2010] 对迁移学习的理论进行探讨。在文中，作者将此称之为 “Learning from different domains”。在三篇文章也成为了迁移学习理论方面的经典文章。文章主要回答的问题就是：在怎样的误差范围内，从不同领域进行学习是可行的？

**学习误差：** 给定两个领域  $\mathcal{D}_s, \mathcal{D}_t$ ,  $X$  是定义在它们之上的数据，一个假设类  $\mathcal{H}$ 。则两个领域  $\mathcal{D}_s, \mathcal{D}_t$  之间的  $\mathcal{H}$ -divergence 被定义为

$$\hat{d}_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) = 2 \sup_{\eta \in \mathcal{H}} \left| \mathbb{P}_{\mathbf{x} \in \mathcal{D}_s} [\eta(\mathbf{x}) = 1] - \mathbb{P}_{\mathbf{x} \in \mathcal{D}_t} [\eta(\mathbf{x}) = 1] \right| \quad (4.18)$$

因此，这个  $\mathcal{H}$ -divergence 依赖于假设  $\mathcal{H}$  来判别数据是来自于  $\mathcal{D}_s$  还是  $\mathcal{D}_t$ 。作者证明了，对于一个对称的  $\mathcal{H}$ ，我们可以通过如下的方式进行计算

$$d_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) = 2 \left( 1 - \min_{\eta \in \mathcal{H}} \left[ \frac{1}{n_1} \sum_{i=1}^{n_1} I[\eta(\mathbf{x}_i) = 0] + \frac{1}{n_2} \sum_{i=1}^{n_2} I[\eta(\mathbf{x}_i) = 1] \right] \right) \quad (4.19)$$

其中  $I[a]$  为指示函数：当  $a$  成立时其值为 1，否则其值为 0。

在目标领域的泛化界：

假设  $\mathcal{H}$  为一个具有  $d$  个 VC 维的假设类，则对于任意的  $\eta \in \mathcal{H}$ ，下面的不等式有  $1 - \delta$  的概率成立：

$$R_{\mathcal{D}_t}(\eta) \leq R_s(\eta) + \sqrt{\frac{4}{n} (d \log \frac{2en}{d} + \log \frac{4}{\delta})} + \hat{d}_{\mathcal{H}}(\mathcal{D}_s, \mathcal{D}_t) + 4 \sqrt{\frac{4}{n} (d \log \frac{2n}{d} + \log \frac{4}{\delta})} + \beta \quad (4.20)$$

其中

$$\beta \geq \inf_{\eta^* \in \mathcal{H}} [R_{\mathcal{D}_s}(\eta^*) + R_{\mathcal{D}_t}(\eta^*)] \quad (4.21)$$

并且

$$R_s(\eta) = \frac{1}{n} \sum_{i=1}^n I[\eta(\mathbf{x}_i) \neq y_i] \quad (4.22)$$

具体的理论证明细节，请参照上述提到的三篇文章。

在自己的研究中，如果需要进行相关的证明，可以参考一些已经发表的文章的写法，例如 [Long et al., 2014a] 等。

另外，英国的 Gretton 等人也在进行一些学习理论方面的研究，有兴趣的读者可以关注他的个人主页：<http://www.gatsby.ucl.ac.uk/~gretton/>。

## 5 迁移学习的基本方法

按照迁移学习领域权威综述文章 A survey on transfer learning [Pan and Yang, 2010], 迁移学习的基本方法可以分为四种。这四种基本的方法分别是：基于样本的迁移，基于模型的迁移，基于特征的迁移，及基于关系的迁移。

本部分简要叙述各种方法的基本原理和代表性相关工作。基于特征和模型的迁移方法是我们的重点。因此，在后续的章节中，将会更加深入地讨论和分析。

### 5.1 基于样本迁移

基于样本的迁移学习方法 (Instance based Transfer Learning) 根据一定的权重生成规则，对数据样本进行重用，来进行迁移学习。图 14形象地表示了基于样本迁移方法的思想。源域中存在不同种类的动物，如狗、鸟、猫等，目标域只有狗这一种类别。在迁移时，为了最大限度地和目标域相似，我们可以人为地提高源域中属于狗这个类别的样本权重。



图 14: 基于样本的迁移学习方法示意图

在迁移学习中，对于源域  $\mathcal{D}_s$  和目标域  $\mathcal{D}_t$ ，通常假定产生它们的概率分布是不同且未知的 ( $P(\mathbf{x}_s) \neq P(\mathbf{x}_t)$ )。另外，由于实例的维度和数量通常都非常大，因此，直接对  $P(\mathbf{x}_s)$  和  $P(\mathbf{x}_t)$  进行估计是不可行的。因而，大量的研究工作 [Khan and Heisterkamp, 2016, Zadrozny, 2004, Cortes et al., 2008, Dai et al., 2007, Tan et al., 2015, Tan et al., 2017] 着眼于对源域和目标域的分布比值进行估计 ( $P(\mathbf{x}_t)/P(\mathbf{x}_s)$ )。所估计得到的比值即为样本的权重。这些方法通常都假设  $\frac{P(\mathbf{x}_t)}{P(\mathbf{x}_s)} < \infty$  并且源域和目标域的条件概率分布相同 ( $P(y|\mathbf{x}_s) = P(y|\mathbf{x}_t)$ )。特别地，上海交通大学 Dai 等人 [Dai et al., 2007] 提出了 TrAdaBoost 方法，将 AdaBoost 的思想应用于迁移学习中，提高有利于目标分类任务的实例权重、降低不利于目标分类任务的实例权重，并基于 PAC 理论推导了模型的泛化误差上界。TrAdaBoost 方法是此方面的经典研究之一。文献 [Huang et al., 2007] 提出核均值匹配方法 (Kernel Mean Matching, KMM) 对于概率分布进行估计，目标是使得加权后的源域和目标域的概率分布尽可能相近。在最新的研究成果中，香港科技大学的 Tan 等人扩展了实例迁移学习方法的应用场景，提出了传递迁移学习方法 (Transitive Transfer Learning, TTL) [Tan et al., 2015] 和远域迁移学习 (Distant Domain Transfer Learning, DDTL) [Tan et al., 2017]，利用联合矩阵分解和深度神经网络，将迁移学习应用于多个不相似的领域之间的知识共享，取得了良好的效果。

虽然实例权重法具有较好的理论支撑、容易推导泛化误差上界，但这类方法通常只在领域间分布差异较小时有效，因此对自然语言处理、计算机视觉等任务效果并不理想。而基于

特征表示的迁移学习方法效果更好，是我们研究的重点。

## 5.2 基于特征迁移

基于特征的迁移方法 (Feature based Transfer Learning) 是指将通过特征变换的方式互相迁移 [Liu et al., 2011, Zheng et al., 2008, Hu and Yang, 2011]，来减少源域和目标域之间的差距；或者将源域和目标域的数据特征变换到统一特征空间中 [Pan et al., 2011, Long et al., 2014b, Duan et al., 2012]，然后利用传统的机器学习方法进行分类识别。根据特征的同构和异构性，又可以分为同构和异构迁移学习。图 15 很形象地表示了两种基于特征的迁移学习方法。



图 15: 基于特征的迁移学习方法示意图

基于特征的迁移学习方法是迁移学习领域中最热门的研究方法，这类方法通常假设源域和目标域间有一些交叉的特征。香港科技大学的 Pan 等人 [Pan et al., 2011] 提出的迁移成分分析方法 (Transfer Component Analysis, TCA) 是其中较为典型的一个方法。该方法的核心内容是以最大均值差异 (Maximum Mean Discrepancy, MMD) [Borgwardt et al., 2006] 作为度量准则，将不同数据领域中的分布差异最小化。加州大学伯克利分校的 Blitzer 等人 [Blitzer et al., 2006] 提出了一种基于结构对应的学习方法 (Structural Corresponding Learning, SCL)，该算法可以通过映射将一个空间中独有的一些特征变换到其他所有空间中的轴特征上，然后在该特征上使用机器学习的算法进行分类预测。清华大学龙明盛等人 [Long et al., 2014b] 提出在最小化分布距离的同时，加入实例选择的迁移联合匹配 (Transfer Joint Matching, TJM) 方法，将实例和特征迁移学习方法进行了有机的结合。澳大利亚卧龙岗大学的 Jing Zhang 等人 [Zhang et al., 2017a] 提出对于源域和目标域各自训练不同的变换矩阵，从而达到迁移学习的目标。

近年来，基于特征的迁移学习方法大多与神经网络进行结合 [Long et al., 2015a, Long et al., 2016, Long et al., 2017, Sener et al., 2016]，在神经网络的训练中进行学习特征和模型的迁移。由于本文的研究重点即是基于特征的迁移学习方法，因此，我们在本小节对这类方法不作过多介绍。在下一小节中，我们将从不同的研究层面，系统地介绍这类工作。

## 5.3 基于模型迁移

基于模型的迁移方法 (Parameter/Model based Transfer Learning) 是指从源域和目标域中找到他们之间共享的参数信息，以实现迁移的方法。这种迁移方式要求的假设条件是：源域中的数据与目标域中的数据可以共享一些模型的参数。其中的代表性工作主要

有 [Zhao et al., 2010, Zhao et al., 2011, Pan et al., 2008b, Pan et al., 2008a]。图 16 形象地表示了基于模型的迁移学习方法的基本思想。

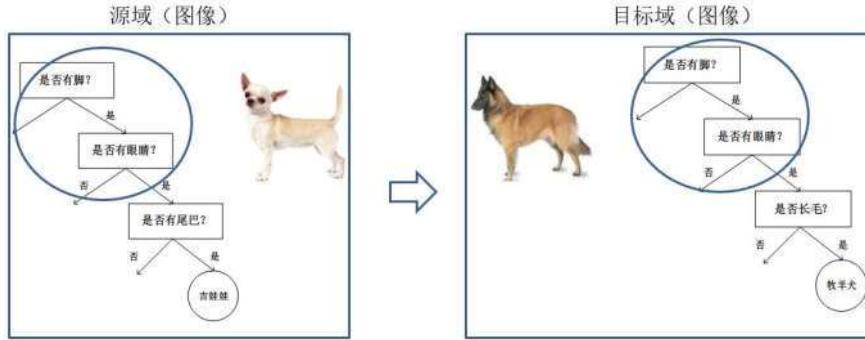


图 16: 基于模型的迁移学习方法示意图

其中, 中科院计算所的 Zhao 等人 [Zhao et al., 2011] 提出了 TransEMDT 方法。该方法首先针对已有标记的数据, 利用决策树构建鲁棒性的行为识别模型, 然后针对无标定数据, 利用 K-Means 聚类方法寻找最优化的标定参数。西安邮电大学的 Deng 等人 [Deng et al., 2014] 也用超限学习机做了类似的工作。香港科技大学的 Pan 等人 [Pan et al., 2008a] 利用 HMM, 针对 WiFi 室内定位在不同设备、不同时间和不同空间下动态变化的特点, 进行不同分布下的室内定位研究。另一部分研究人员对支持向量机 SVM 进行了改进研究 [Nater et al., 2011, Li et al., 2012]。这些方法假定 SVM 中的权重向量  $\mathbf{w}$  可以分成两个部分:  $\mathbf{w} = \mathbf{w}_0 + \mathbf{v}$ , 其中  $\mathbf{w}_0$  代表源域和目标域的共享部分,  $\mathbf{v}$  代表了对于不同领域的特定处理。在最新的研究成果中, 香港科技大学的 Wei 等人 [Wei et al., 2016b] 将社交信息加入迁移学习方法的正则项中, 对方法进行了改进。清华大学龙明盛等人 [Long et al., 2015a, Long et al., 2016, Long et al., 2017] 改进了深度网络结构, 通过在网络中加入概率分布适配层, 进一步提高了深度迁移学习网络对于大数据的泛化能力。

通过对现有工作的调研可以发现, 目前绝大多数基于模型的迁移学习方法都与深度神经网络进行结合 [Long et al., 2015a, Long et al., 2016, Long et al., 2017, Tzeng et al., 2015, Long et al., 2016]。这些方法对现有的一些神经网络结构进行修改, 在网络中加入领域适配层, 然后联合进行训练。因此, 这些方法也可以看作是基于模型、特征的方法的结合。

#### 5.4 基于关系迁移

基于关系的迁移学习方法 (Relation Based Transfer Learning) 与上述三种方法具有截然不同的思路。这种方法比较关注源域和目标域的样本之间的关系。图 17 形象地表示了不同领域之间相似的关系。

就目前来说, 基于关系的迁移学习方法的相关研究工作非常少, 仅有几篇连贯式的文章讨论: [Mihalkova et al., 2007, Mihalkova and Mooney, 2008, Davis and Domingos, 2009]。这些文章都借助于马尔科夫逻辑网络 (Markov Logic Net) 来挖掘不同领域之间的关系相似性。

我们将重点讨论基于特征和基于模型的迁移学习方法, 这也是目前绝大多数研究工作的热点。

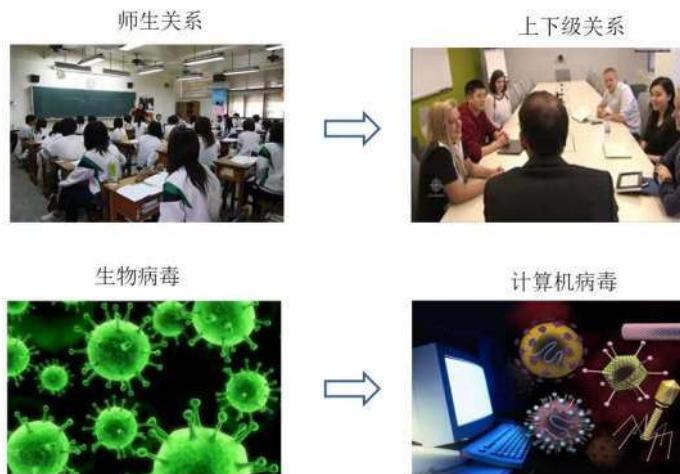


图 17: 基于关系的迁移学习方法示意图

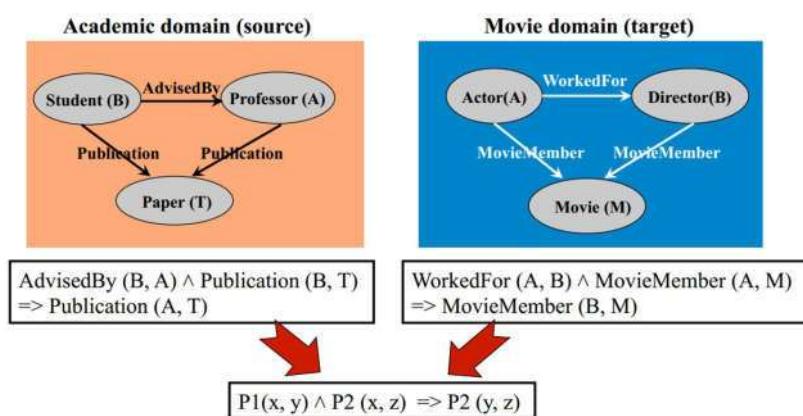


图 18: 基于马尔科夫逻辑网的关系迁移

## 6 第一类方法：数据分布自适应

数据分布自适应 (Distribution Adaptation) 是一类最常用的迁移学习方法。这种方法的基本思想是，由于源域和目标域的数据概率分布不同，那么最直接的方式就是通过一些变换，将不同的数据分布的距离拉近。

图 19形象地表示了几种数据分布的情况。简单来说，数据的边缘分布不同，就是数据整体不相似。数据的条件分布不同，就是数据整体相似，但是具体到每个类里，都不太相似。

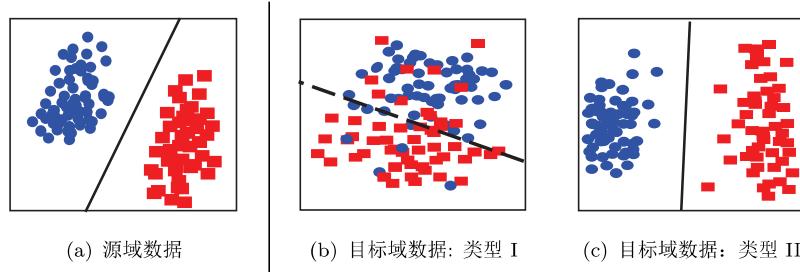


图 19: 不同数据分布的目标域数据

根据数据分布的性质，这类方法又可以分为边缘分布自适应、条件分布自适应、以及联合分布自适应。下面我们分别介绍每类方法的基本原理和代表性研究工作。介绍每类研究工作时，我们首先给出基本思路，然后介绍该类方法的核心，最后结合最近的相关工作介绍该类方法的扩展。

### 6.1 边缘分布自适应

#### 6.1.1 基本思路

边缘分布自适应方法 (Marginal Distribution Adaptation) 的目标是减小源域和目标域的边缘概率分布的距离，从而完成迁移学习。从形式上来说，边缘分布自适应方法是用  $P(\mathbf{x}_s)$  和  $P(\mathbf{x}_t)$  之间的距离来近似两个领域之间的差异。即：

$$DISTANCE(\mathcal{D}_s, \mathcal{D}_t) \approx ||P(\mathbf{x}_s) - P(\mathbf{x}_t)|| \quad (6.1)$$

边缘分布自适应对应于图 19中由图 19(a)迁移到图 19(b)的情形。

#### 6.1.2 核心方法

边缘分布自适应的方法最早由香港科技大学杨强教授团队提出 [Pan et al., 2011]，方法名称为迁移成分分析 (Transfer Component Analysis)。由于  $P(\mathbf{x}_s) \neq P(\mathbf{x}_t)$ ，因此，直接减小二者之间的距离是不可行的。TCA 假设存在一个特征映射  $\phi$ ，使得映射后数据的分布  $P(\phi(\mathbf{x}_s)) \approx P(\phi(\mathbf{x}_t))$ 。TCA 假设如果边缘分布接近，那么两个领域的条件分布也会接近，即条件分布  $P(y_s|\phi(\mathbf{x}_s)) \approx P(y_t|\phi(\mathbf{x}_t))$ 。这就是 TCA 的全部思想。因此，我们现在的目标是，找到这个合适的  $\phi$ 。

但是世界上有无穷个这样的  $\phi$ ，也许终我们一生也无法找到合适的那一个。庄子说过，吾生也有涯，而知也无涯，以有涯随无涯，殆已！我们肯定不能通过穷举的方法来找  $\phi$  的。那么怎么办呢？

回到迁移学习的本质上来：最小化源域和目标域的距离。好了，我们能不能先假设这个  $\phi$  是已知的，然后去求距离，看看能推出什么呢？

更进一步，这个距离怎么算？机器学习中有很多种形式的距离，从欧氏距离到马氏距离，从曼哈顿距离到余弦相似度，我们需要什么距离呢？TCA 利用了一个经典的也算是比较“高端”的距离叫做最大均值差异 (MMD, maximum mean discrepancy)。我们令  $n_1, n_2$  分别表示源域和目标域的样本个数，那么它们之间的 MMD 距离可以计算为：

$$DISTANCE(\mathbf{x}_s, \mathbf{x}_t) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(\mathbf{x}_i) - \frac{1}{n_2} \sum_{j=1}^{n_2} \phi(\mathbf{x}_j) \right\|_{\mathcal{H}} \quad (6.2)$$

MMD 是做了一件什么事呢？简单，就是求映射后源域和目标域的均值之差。

事情到这里似乎也没什么进展：我们想求的  $\phi$  仍然没法求。

TCA 是怎么做的呢，这里就要感谢矩阵了！我们发现，上面这个 MMD 距离平方展开后，有二次项乘积的部分！那么，联系在 SVM 中学过的核函数，把一个难求的映射以核函数的形式来求，不就可以了？于是，TCA 引入了一个核矩阵  $\mathbf{K}$ ：

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{s,s} & \mathbf{K}_{s,t} \\ \mathbf{K}_{t,s} & \mathbf{K}_{t,t} \end{bmatrix} \quad (6.3)$$

以及一个 MMD 矩阵  $\mathbf{L}$ ，它的每个元素的计算方式为：

$$l_{ij} = \begin{cases} \frac{1}{n_1^2} & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s, \\ \frac{1}{n_2^2} & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t, \\ -\frac{1}{n_1 n_2} & \text{otherwise} \end{cases} \quad (6.4)$$

这样的好处是，直接把那个难求的距离，变换成了下面的形式：

$$\text{tr}(\mathbf{KL}) - \lambda \text{tr}(\mathbf{K}) \quad (6.5)$$

其中， $\text{tr}(\cdot)$  操作表示求矩阵的迹，用人话来说就是一个矩阵对角线元素的和。这样是不是感觉离目标又进了一步呢？

其实这个问题到这里就已经是可解的了，也就是说，属于计算机的部分已经做完了。只不过它是一个数学中的半定规划 (SDP, semi-definite programming) 的问题，解决起来非常耗费时间。由于 TCA 的第一作者 Sinno Jialin Pan 以前是中山大学的数学硕士，他想用更简单的方法来解决。他是怎么做的呢？

他想出了用降维的方法去构造结果。用一个更低维度的矩阵  $\mathbf{W}$ ：

$$\tilde{\mathbf{K}} = (\mathbf{KK}^{-1/2}\widetilde{\mathbf{W}})(\widetilde{\mathbf{W}}^\top \mathbf{K}^{-1/2}\mathbf{K}) = \mathbf{KWW}^\top \mathbf{K} \quad (6.6)$$

这里的  $\mathbf{W}$  矩阵是比  $\mathbf{K}$  更低维度的矩阵。最后的  $\mathbf{W}$  就是问题的解答了！

好了，问题到这里，整理一下，TCA 最后的优化目标是：

$$\begin{aligned} \min_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^\top \mathbf{KLK}\mathbf{W}) + \mu \text{tr}(\mathbf{W}^\top \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^\top \mathbf{KHKW} = \mathbf{I}_m \end{aligned} \quad (6.7)$$

这里的  $\mathbf{H}$  是一个中心矩阵， $\mathbf{H} = \mathbf{I}_{n_1+n_2} - 1/(n_1+n_2)\mathbf{1}\mathbf{1}^\top$ 。

这个式子下面的条件是什么意思呢？那个  $\min$  的目标我们大概理解，就是要最小化源域和目标域的距离，加上  $\mathbf{W}$  的约束让它不能太复杂。那么下面的条件是什么呢？下面的条件就是要实现第二个目标：维持各自的数据特征。

TCA 要维持的是什么特征呢？文章中说是 variance，但是实际是 scatter matrix，就是数据的散度。就是说，一个矩阵散度怎么计算？对于一个矩阵  $\mathbf{A}$ ，它的 scatter matrix 就是  $\mathbf{A}\mathbf{H}\mathbf{A}^\top$ 。这个  $\mathbf{H}$  就是上面的中心矩阵啦。

解决上面的优化问题时，作者又求了它的拉格朗日对偶。最后得出结论， $\mathbf{W}$  的解就是它的前  $m$  个特征值！简单不？数学美不美？

好了，我们现在总结一下 TCA 方法的步骤。输入是两个特征矩阵，我们首先计算  $\mathbf{L}$  和  $\mathbf{H}$  矩阵，然后选择一些常用的核函数进行映射（比如线性核、高斯核）计算  $\mathbf{K}$ ，接着求  $(\mathbf{KLK} + \mu\mathbf{I})^{-1}\mathbf{KHK}$  的前  $m$  个特征值。仅此而已。然后，得到的就是源域和目标域的降维后的数据，我们就可以在上面用传统机器学习方法了。

为了形象地展示 TCA 方法的优势，我们借用 [Pan et al., 2011] 中提供的可视化效果，在图中展示了对于源域和目标域数据（红色和蓝色），分别由 PCA（主成分分析）和 TCA 得到的分布结果。从图 20 中可以很明显地看出，对于概率分布不同的两部分数据，在经过 TCA 处理后，概率分布更加接近。这说明了 TCA 在拉近数据分布距离上的优势。

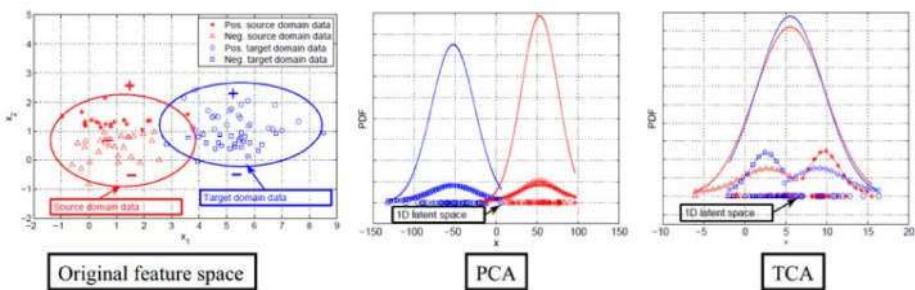


图 20: TCA 和 PCA 的效果对比

### 6.1.3 扩展

TCA 方法是迁移学习领域一个经典的方法，之后的许多研究工作都以 TCA 为基础。我们列举部分如下：

- ACA (Adapting Component Analysis) [Dorri and Ghodsi, 2012]: 在 TCA 中加入了 HSIC
- DTMKL (Domain Transfer Multiple Kernel Learning) [Duan et al., 2012]: 在 TCA 中加入了 MK-MMD，用了新的求解方式
- TJM (Transfer Joint Matching) [Long et al., 2014b]: 在优化目标中同时进行边缘分布自适应和源域样本选择
- DDC (Deep Domain Confusion) [Tzeng et al., 2014]: 将 MMD 度量加入了深度网络特征层的 loss 中（我们将会在深度迁移学习中介绍此工作）
- DAN (Deep Adaptation Network) [Long et al., 2015a]: 扩展了 DDC 的工作，将 MMD 换成了 MK-MMD，并且进行多层 loss 计算（我们将会在深度迁移学习中介绍此工作）

- DME (Distribution Matching Embedding): 先计算变换矩阵，再进行特征映射（与 TCA 顺序相反）
- CMD (Central Moment Matching) [Zellinger et al., 2017]: MMD 着眼于一阶，此工作将 MMD 推广到了多阶

## 6.2 条件分布自适应

条件分布自适应方法 (Conditional Distribution Adaptation) 的目标是减小源域和目标域的条件概率分布的距离，从而完成迁移学习。从形式上来说，条件分布自适应方法是用  $P(y_s|\mathbf{x}_s)$  和  $P(y_t|\mathbf{x}_t)$  之间的距离来近似两个领域之间的差异。即：

$$DISTANCE(\mathcal{D}_s, \mathcal{D}_t) \approx \|P(y_s|\mathbf{x}_s) - P(y_t|\mathbf{x}_t)\| \quad (6.8)$$

条件分布自适应对应于图 19 中由图 19(a) 迁移到图 19(c) 的情形。

目前单独利用条件分布自适应的工作较少，这些工作主要可以在 [Saito et al., 2017] 中找到。最近，中科院计算所的 Wang 等人提出了 STL 方法 (Stratified Transfer Learning) [Wang et al., 2018]。作者提出了类内迁移 (Intra-class Transfer) 的思想。指出现有的绝大多数方法都只是学习一个全局的特征变换 (Global Domain Shift)，而忽略了类内的相似性。类内迁移可以利用类内特征，实现更好的迁移效果。

STL 方法的基本思路如图 21 所示。首先利用大多数投票的思想，对无标定的位置行为生成伪标签；然后在再生核希尔伯特空间中，利用类内相关性进行自适应地空间降维，使得不同情境中的行为数据之间的相关性增大；最后，通过二次标注，实现对未知标定数据的精准标定。

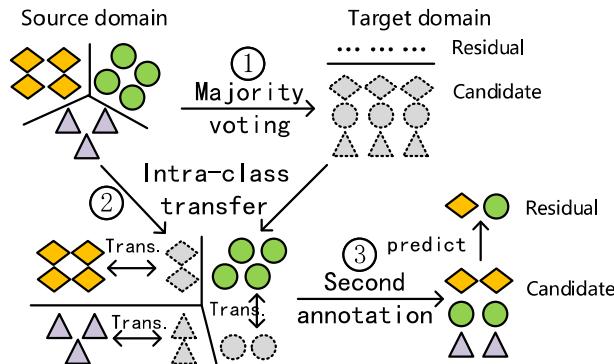


图 21: STL 方法的示意图

为了实现类内迁移，我们需要计算每一类别的 MMD 距离。由于目标域没有标记，作者使用来自大多数投票结果中的伪标记。更加准确地说，用  $c \in \{1, 2, \dots, C\}$  来表示类别标记，则类内迁移可以按如下方式计算：

$$D(\mathcal{D}_s, \mathcal{D}_t) = \sum_{c=1}^C \left\| \frac{1}{n_1^{(c)}} \sum_{\mathbf{x}_i \in \mathcal{D}_s^{(c)}} \phi(\mathbf{x}_i) - \frac{1}{n_2^{(c)}} \sum_{\mathbf{x}_j \in \mathcal{D}_t^{(c)}} \phi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 \quad (6.9)$$

其中， $\mathcal{D}_s^{(c)}$  和  $\mathcal{D}_t^{(c)}$  分别表示源域和目标域中属于类别  $c$  的样本。 $n_1^{(c)} = |\mathcal{D}_s^{(c)}|$ ，且  $n_2^{(c)} = |\mathcal{D}_t|$ 。