**关于迁移学习的广泛调查**

**庄富贞，齐志远，段克玉，奚东波，朱红春，朱恒书，IEEE高级成员，熊辉， IEEE Fellow，何庆**

# 摘要

转移学习的目的是通过转移不同但相关的源域所包含的知识，提高目标学习者在目标域上的学习成绩。这样，在构建目标学习器时可以减少对大量目标域数据的依赖。迁移学习由于其广阔的应用前景，已成为机器学习中一个热门且有发展前景的领域。虽然已经有一些关于迁移学习的有价值和令人印象深刻的调查，但这些调查以一种相对孤立的方式介绍了迁移学习的方法，缺乏迁移学习方面的最新进展。由于迁移学习领域的迅速扩大，对相关研究进行全面的综述既必要又具有挑战性。本调查试图对已有的迁移学习研究进行联系和系统化，对迁移学习的机制和策略进行全面的总结和阐释，有助于读者更好地了解迁移学习的研究现状和思路。与以往的调查不同，本调查论文从数据和模型的角度回顾了40多种有代表性的迁移学习方法，特别是同质迁移学习方法。并简要介绍了迁移学习的应用。为了展示不同迁移学习模型的性能，我们使用了20多个有代表性的迁移学习模型进行了实验。该模型是在三个不同的数据集上进行的，即亚马逊评论、路透-21578和Office-31。实验结果表明，在实际应用中选择合适的迁移学习模型是非常重要的。

## 关键字

学习，机器学习，领域适应，解释

# 1介绍

虽然传统的机器学习技术已经取得了巨大的成功，并成功地应用于许多实际应用中，但对于某些真实场景，它仍有一定的局限性。机器学习的理想场景是有大量标记的训练实例，它们与测试数据具有相同的分布。然而，收集足够的训练数据通常是昂贵的、耗时的，甚至在许多情况下是不现实的。半监督学习可以通过放松对大量标记数据的需求，在一定程度上解决这一问题。通常，半叠加方法 只需要有限数量的标记数据，而利用大量的未标记数据来提高学习精度。但在许多情况下，未贴标签的实例也很难收集，这通常使传统模型不能令人满意。

迁移学习是解决上述问题的一种很有前途的机器学习方法。迁移学习的概念最初可能来自教育心理学。根据心理学家C.H. Judd提出的迁移泛化理论，学习迁移是**经验泛化**的结果。实现从一种情况到另一种情况的转变是可能的，只要一个人概括他的**经验**。根据这一理论，迁移的前提是两种学习活动之间要有**联系**。在实践中，一个学过小提琴的人可以比其他人更快地学习钢琴，因为小提琴和钢琴都是乐器，可以分享一些共同的知识。图1展示了迁移学习的一些直观例子。迁移学习受人跨领域转移知识能力的启发，旨在利用相关领域(称为源领域)的知识来提高学习性能或减少目标领域所需的标记示例数量。值得一提的是，知识转移并不总是会给新的任务带来积极的影响。如果领域之间没有共同点，知识转移可能是不成功的。例如，学习骑自行车并不能帮助我们更快地学习弹钢琴。此外，领域之间的相似性并不总是有利于学习，因为有时相似性可能会误导。例如，虽然西班牙语和法语关系密切，都属于罗曼语系，但学习西班牙语的人在学习法语时会遇到困难，比如使用错误的词汇或词形变化。这是因为以前在西班牙语中的成功经验会干扰学习法语中的构词、用法、发音、变化等。在心理学领域，以往的经验对学习新任务有负面影响的现象被称为**负迁移**[1]。同样，在迁移学习领域，如果目标学习者受到迁移知识的负面影响，也称为负迁移[2]、[3]。负迁移是否会发生取决于几个因素，如源域和目标域之间的关联性，学习者跨域寻找可转移有益部分的能力。在[3]中，给出了负迁移的一个正式定义和一些分析。

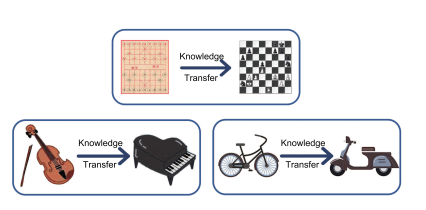


图 1迁移学习的直观例子

粗略地讲，根据领域之间的差异，迁移学习可以进一步分为两类，即同质迁移学习和异质迁移学习[4]。同质迁移学习方法的提出和发展，以处理域具有**相同特征空间**的情况。在同质迁移学习中，一些研究假设领域仅在边缘分布上不同。因此，他们通过校正**样本选择偏差**[5]或**协变量位移**[6]来适应域。然而，这种假设在很多情况下并不成立。例如，在情绪分类问题中，一个词可能在不同的领域有不同的意义倾向。这种现象也被称为**上下文特征偏差**[7]。为了解决这一问题，一些研究对条件分布进行了进一步的适应。异质迁移学习是指领域在具有不同特征空间的情况下进行的知识迁移过程。除了分布适应外，异质性迁移学习还需要特征空间适应[7]，这使得异质性迁移学习比同质迁移学习更加复杂。

调查的目的是让读者从**数据(Data)**和**模型(Model)**的角度对迁移学习有一个全面的了解。介绍迁移学习方法的机制和策略，让读者了解迁移学习方法是如何工作的。已有的许多迁移学习研究都是相互关联、系统化的。本文具体介绍了40多种有代表性的迁移学习方法。此外，我们还进行了实验来证明一个迁移学习模型在哪个数据集上表现得很好。

在这次调查中，我们更关注**同质迁移学习**。还有一些有趣的迁移学习课题没有在本次调查中提及，如强化迁移学习[8]，终身迁移学习[9]，在线迁移学习[10]。本调查的其余部分分为七个部分。第2节阐明了迁移学习和其他相关机器学习技术的区别。第3节介绍了本调查中使用的符号和迁移学习的定义。第4节和第5节分别从数据和模型的角度解释迁移学习方法。第6节介绍迁移学习的一些应用。实验的实施，以及结果在第7节提供。最后一节总结了这个调查。本调查的主要贡献概述如下。

•介绍和总结了四十多种有代表性的迁移学习方法，让读者对迁移学习有一个全面的了解。

•进行了实验来比较不同的迁移学习方法。对二十多种不同的方法的性能进行了直观的展示和分析，对读者在实践中选择合适的方法有一定的指导意义。

# 2相关工作

介绍了与迁移学习相关的一些领域。阐明了它们与迁移学习的联系和区别。

**半监督学习[11]:**半监督学习是介于监督学习(完全标记实例)和非监督学习(没有任何标记实例)之间的机器学习任务和方法。通常，半监督方法利用大量的未标记实例和有限数量的标记实例来训练学习者。半监督学习放松了对标记实例的依赖，从而降低了昂贵的标记成本。请注意，在半监督学习中，带标记的和未带标记的实例都是从相同的分布中抽取的。相反，在迁移学习中，源域和目标域的数据分布通常是不同的。许多迁移学习方法吸收了半监督学习技术。在迁移学习中还使用了半监督学习中的关键假设，即光滑性、聚类和流形假设。值得一提的是，半监督迁移学习是一个有争议的术语。原因是标签信息在迁移学习中是否可用的概念是模糊的，因为源域和目标域都可以涉及。

**多视图学习[12]**:多视图学习关注的是多视图数据的机器学习问题。视图代表一个独特的特征集。关于多个视图的一个直观的例子是，视频对象可以从两个不同的视角来描述，即图像信号和音频信号。简单地说，多视角学习从多个视角描述一个对象，从而产生丰富的信息。通过恰当地考虑各个方面的信息，学习者的学习成绩可以得到提高。在多视角学习中有几种策略，如子空间学习、多核学习、联合训练[13]、[14]等。在一些迁移学习方法中也采用了多视图技术。例如，Zhang等人提出了一个多视图迁移学习框架，它强加了多个视图之间的一致性[15]。Yang和Gao将跨不同领域的多视角信息用于[16]知识转移。Feuz和Cook在工作中引入了一种多视角迁移学习方法，在异构传感器平台[17]之间迁移活动知识。

**多任务学习[18]**:多任务学习的思想是共同学习一组相关的任务。更具体地说，多任务学习通过利用任务之间的相互联系来强化每一个任务，即同时考虑任务间的相关性和任务间的差异性。这样，增强了每个任务的泛化性。迁移学习与多任务学习的主要区别在于前者是迁移相关领域内的知识，而后者是通过同时学习一些相关任务来迁移知识。换句话说，多任务学习对每一项任务的关注是平等的，而迁移学习对目标任务的关注多于对源任务的关注。迁移学习和多任务学习之间存在一些共性和联系。二者都旨在通过知识转移来提高学习者的学习成绩。此外，它们还采用了一些相似的构建模型的策略，如特征转换和参数共享等。请注意，一些现有的研究同时利用了迁移学习和多任务学习技术。例如，Zhang等人的工作采用了生物图像分析[19]的多任务和迁移学习技术。Liu等人的工作提出了一个基于多任务学习和多源迁移学习[20]的人类动作识别框架。

# 3概述

为了方便起见，本节列出了调查中使用的符号。此外，还介绍了迁移学习的一些定义和分类，并提供了一些相关的调查。

## 3.1符号

为了方便起见，表1列出了符号及其定义。此外，我们用来表示范数和上标来表示一个向量或矩阵的转置。

## 3.2定义

在本节中，给出了迁移学习的一些定义。在给出迁移学习的定义之前，让我们回顾一下领域和任务的定义。

**定义1(域)**。(域)一个域D由两部分组成，即特征空间X和边缘分布P(X)。也就是说，D = {X, P(X)}。符号X表示一个实例集，定义为X = {X |xi∈X, i = 1，···，n}。

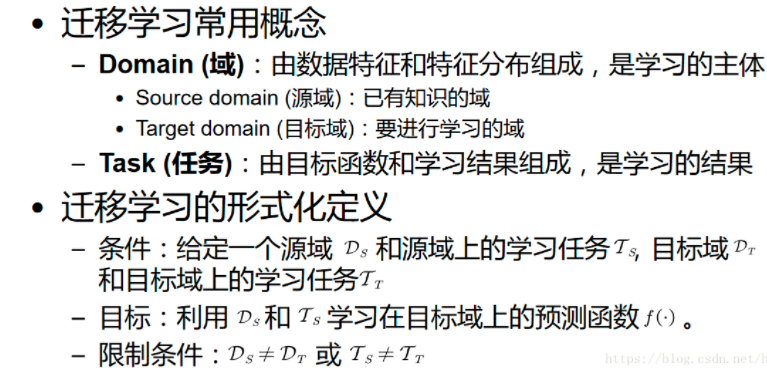
**定义2（任务Task）。**任务T由一个标签空间Y和一个决策函数f组成，即T = {Y, f}。决策函数f是隐式函数，期望从样本数据中学习。

一些机器学习模型实际上输出实例的预测条件分布。在这种情况下,.

在实践中，一个域经常被许多带有或不带有标签信息的实例观察到。例如，通常通过实例-标签对观察到的一个源任务的源域.即;对目标域的观察通常由大量未标记实例和/或有限数量的标记实例组成。

**定义3（迁移学习）.**

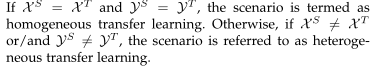
. (Transfer Learning) Given some/an observation(s)corresponding to mS ∈ N+ source domain(s) and task(s) (i.e., {(DSi, TSi)|i = 1, · · · , mS}), and some/an observation(s) about mT ∈ N+ target domain(s) and task(s) (i.e.,{(DTj, TTj)|j = 1, · · · , mT }), transfer learning utilizesthe knowledge implied in the source domain(s) to improvethe performance of the learned decision functions fTj (j =1, · · · , mT ) on the target domain(s)



上述定义涵盖了多源迁移学习的情况，是对调查[2]中所给出的定义的延伸。如果=1，该场景称为单源迁移学习。否则，它被称为多源迁移学习。此外，表示迁移学习任务的数量。少数研究集中在mT≥2[21]的设置上。现有的迁移学习研究更多地关注mT= 1的情景(尤其是mS= mT= 1的情景)。值得一提的是，对领域或任务的观察是一个广义的概念，通常将其绑定到标记/未标记的实例集或预先学习的模型中。一个常见的场景是，我们在源域上有大量的已标记实例或经过良好训练的模型，而我们只有有限的已标记的目标域实例。在这种情况下，实例和模型等资源实际上是观察，迁移学习的目标是学习目标领域上更准确的决策函数。

迁移学习领域的另一个常用术语是**领域适应**。领域适应是指通过适应一个或多个源领域来转移知识，提高目标学习器[4]学习成绩的过程。迁移学习通常依赖于领域适应过程，该过程试图减少领域之间的差异。

## 3.3迁移学习的分类

迁移学习有几个分类标准。例如，迁移学习问题可以分为三类，即传导性迁移学习、归纳迁移学习和无监督迁移学习[2]。这三个类别的完整定义在[2]中给出。这三个类别可以从**标签设置**的角度来解释。粗略地说，转导迁移学习是指标签信息只来自源域的情况。如果目标域实例的标签信息是可用的，那么该场景可以归类为归纳迁移学习。如果标签信息在源域和目标域都是未知的，这种情况称为无监督迁移学习。另一种分类是**基于源和目标特征空间以及标签空间之间的一致性。** 

Transductive learning:unlabelled data is the testing data

inductive learning:unlabelled data is not the testing data

在训练过程中，已知testing data（unlabelled data）是transductive learing

在训练过程中，并不知道testing data ，训练好模型后去解决未知的testing data 是inductive learing

链接：<https://www.zhihu.com/question/68275921/answer/354145465>。

根据调查[2]，迁移学习方法可以分为四类:基于实例的、基于特征的、基于参数的和基于关系的。基于实例的迁移学习方法主要基于实例加权策略。基于特征的方法对原始特征进行变换，生成新的特征表示;基于特征的迁移学习可以再分为基于特征的非对称迁移学习和基于特征的对称迁移学习。非对称方法转换源特性以匹配目标特性。而对称方法则试图找到一个**共同的潜在特征空间**，然后将源特征和目标特征转化为一种新的特征表示。基于参数的迁移学习方法在模型/参数层次上迁移知识。基于关系的迁移学习方法主要关注关系领域中的问题。这些方法将源域中学习到的逻辑关系或规则传输到目标域。为了更好地理解，图2展示了上述迁移学习的分类。

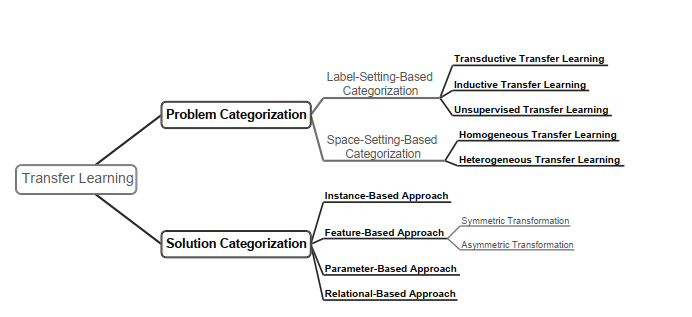


图 2迁移学习分类

一些调查提供给读者（定语从句：这些读者想有一个更全面的了解这一领域）。Pan和Yang[2]的调查是一项开创性的工作，对迁移学习进行了分类，回顾了2010年以前的研究进展。Weiss等人的调查介绍和总结了许多同质和异质迁移学习方法[4]。在Day和Khoshgoftaar[7]的调查中特别回顾了异质迁移学习。一些调查回顾了与特定主题相关的文献，如强化学习[8]、计算智能[22]和深度学习[23]、[24]。此外，还有一些针对特定应用场景的调查，包括活动识别[25]、视觉分类[26]、协同推荐[27]、计算机视觉[24]、情感分析[28]。

请注意，这次调查的组织并没有严格遵循上述分类。在接下来的两部分中，我们将从数据和模型的角度来解释迁移学习方法。粗略地说，基于数据的解释涵盖了上述基于实例的迁移学习方法和基于特征的迁移学习方法，但从更广泛的角度来看。基于模型的解释包括上述基于参数的方法。由于关于基于关系的迁移学习的研究相对较少，代表性的迁移学习方法在[2]、[4]中都有很好的介绍，所以本调查并没有关注基于关系的迁移学习方法。

# 4基于数据的解释

许多迁移学习方法，特别是基于数据库的迁移学习方法，**侧重于通过数据的调整和转换来迁移知识**。图3从数据的角度展示了这些方法的策略和目标。如图3所示，空间适应是目标之一。这一目标在异质迁移学习场景中需要得到满足。而在本次调查中，我们更多地关注同构迁移学习，这个场景中的主要目标是减少源域实例和目标域实例之间的分布差异。此外，一些高级方法可能试图在适应过程中保留数据属性。从数据的角度来看，实现目标通常有两种策略，即实例加权和特征转换。在本节中，我们将根据图3所示的策略，依次介绍一些相关的迁移学习方法。

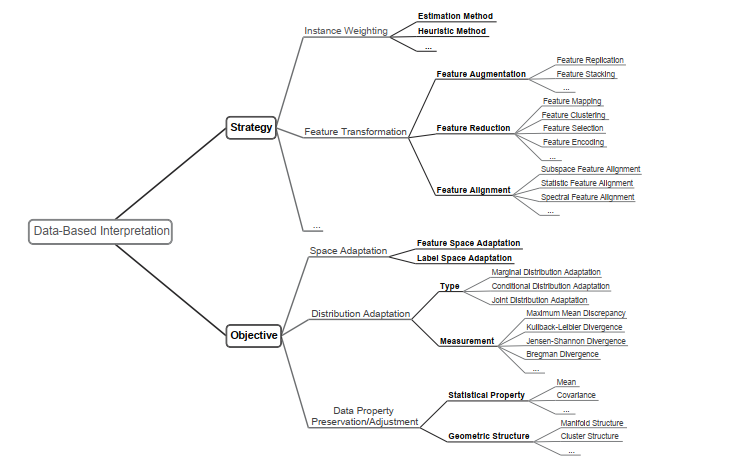
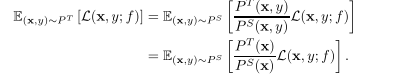


图 3从数据角度出发，迁移学习方法的策略和目标

## 4.1实例加权策略

让我们首先考虑一个简单的场景，其中有大量已标记的源域实例和有限数量的目标域实例，并且域仅在边缘分布上有所不同。（比如，）。例如，假设我们需要建立一个模型来诊断特定区域的癌症情况，这个区域老年人占大多数。给出了有限的目标域实例，并且可以从另一个以年轻人为主的区域获得相关数据。直接从其他地区转移数据可能不成功，因为存在边际分布差异，而且老年人比年轻人患癌症的风险更高。在这种情况下，考虑调整边缘分布是很自然的。**一个简单的想法是在损失函数（loss function）中为源域实例分配权重**。加权策略基于方程[5]:

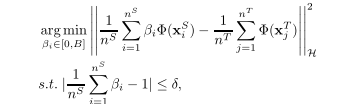


因此，学习任务的一般目标函数可以写成[5]:



其中，(i = 1,2，···，nS)是权重参数。其理论数值为。然而，这一比例通常是未知的，难以用传统方法得到。

Kernel Mean Matching (KMM)[5]是Huang等人提出的，它通过在Reproducing Kernel Hilbert Space (RKHS)中匹配源主和目标域实例的均值来解决上述未知比值的估计问题，即:



式中是个很小的参数，B是一个常量，通过对上述优化问题进行展开并利用核技巧，可将其转化为二次规划问题。这种估计分布比率的方法可以很容易地并入许多现有的算法中。获得权重后，就可以在加权的源域实例上训练学习器。

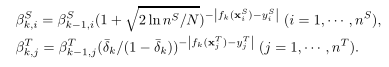
还有一些其他的研究试图估算权重。例如，Sugiyama等人提出了一种称为Kullback-Leibler重要性估计程序(KLIEP)[6]的方法。KLIEP依赖于kullbackleibler (KL)散度的最小化，并包含了一个内置的模型选择过程。在研究权值估计的基础上，提出了基于实例的迁移学习框架或算法。例如，Sun等人提出了一个多源框架，称为多源域适应的两阶段加权框架(2SW-MDA)，其中包括以下两个阶段[29]：

1.实例权重:为源域实例分配权重以减少边缘分布差异，类似于KMM。

2. 域权重:根据平滑度假设[30]，为每个源域分配权重以减小条件分布差异。

然后，根据实例权重和域权重重新对源域实例进行权重设置。这些重新加权的实例和标记的目标域实例用于训练目标分类器。

除了直接估计权重参数外，迭代调整权重也是有效的。关键是设计一种机制来减少对目标学习者有负面影响的实例的权重。一个代表性的作品是TrAdaBoost[31]，它是一个由Dai等人提出的框架。这个框架是AdaBoost[32]的扩展。AdaBoost是一种针对传统机器学习任务设计的有效的增强算法。在AdaBoost的每次迭代中，学习器都是在权重更新的实例上训练的，这导致了分类器的薄弱。实例的权重机制保证了对分类不正确的实例给予更多的关注。最后，将得到的弱分类器组合成强分类器。TrAdaBoost将AdaBoost扩展到迁移学习场景;设计了一种新的加权机构，以减少分布差异的影响。具体来说，在TrAdaBoost中，将标记的源域实例和标记的目标域实例组合为一个整体，即用一个训练集来训练弱分类器。对于源域和目标域实例，权重操作是不同的。在每次迭代中，计算一个临时变量(度量目标域实例的分类错误率)。然后，目标域实例的权重根据和单个分类结果更新，而源域实例的权重根据设计的常数和单个分类结果更新。为了更好的理解，在第k次迭代(k = 1，···，N)中用于更新权重的公式重复表示如下[31]:



请注意，每次迭代都会形成一个新的弱分类器。通过投票的方式将新生成的弱分类器的一半进行组合和集成，从而得到最终的分类器。

一些研究进一步扩展了TrAdaBoost。Yao和Doretto[33]的工作提出了多源TrAdaBoost (MsTrAdaBoost)算法，每次迭代主要有以下两步。

1. 候选分类器构造:在每个源域和目标域对的加权实例上分别训练一组候选弱分类器，即(i = 1，···，)。

2. 实例加权:选择一个在目标域实例上分类错误率最小的分类器(用j表示，在DSj∪DT上训练)，然后用于更新DSjand DT中实例的权重。

最后，将每次迭代中选择的分类器组合起来，形成最终的分类器。在工作[33]中也提出了另一种基于参数的算法TaskTrAdaBoost，在5.3节中介绍。

一些方法采用启发式的方法实现实例加权策略。例如，Jiang和Zhai提出了一个通用的权重框架[34]。框架的目标函数中有三个术语，旨在最小化三种类型实例的交叉熵损失。下面的实例类型用于构造目标分类器。

•标记的arget-domain实例:分类器应该最小化它们的交叉熵损失，这实际上是一个标准的监督学习任务。

•未标记的T arget域实例:这些实例的真实条件分布P(y|xT,U i)是未知的，应该估计。一种可能的解决方案是在已标记的源域和目标域实例上训练辅助分类器，以帮助估计条件分布或为这些实例分配伪标签。

•带标签的源域实例:作者将xS,L i的权重定义为两个部分的乘积，即:信仰和信仰。理想情况下，权重iis等于PT(xi)/PS(xi)，可以通过KMM等非参数方法估计，也可以在最坏情况下均匀设置。用于过滤出与目标域有很大差异的源域实例的权值。

可以使用一种启发式的方法来产生的值，其中包含以下三个步骤。

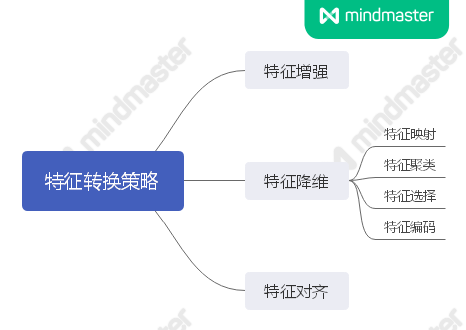
1. 辅助分类器构造:使用在标记的目标域实例上训练的辅助分类器对未标记的源域实例进行分类。

2. 实例排序:根据概率预测结果对源域实例进行排序。

3.启发式加权法(humani):预测错误的top-k源域实例的权重设置为0，其他实例的权重设置为1。

## 4.2特征转换策略

特征转换策略通常采用基于特征的方法。例如，考虑一个跨域文本分类问题。任务是通过使用来自相关的标签数据来构造目标分类器域。在这种情况下，**一个可行的解决方案是通过特征转换找到共同的潜在特征**(如潜在主题)，**并将其作为传递知识的桥梁**。基于特征的方法将每个原始特征转化为新的特征表示，用于知识转移。构造一个新的特征表示的目标包括最小化边缘和最小化条件分布差异，保持数据的性质或潜在结构，以及找到特征之间的对应关系。特征变换的操作可以分为三种类型，即特征增强、特征降维和特征对齐。此外，特征降维还可以进一步分为特征映射、特征聚类、特征选择和特征编码等几种类型。在一个算法中设计的完整的特征转换过程可能包含多个操作。



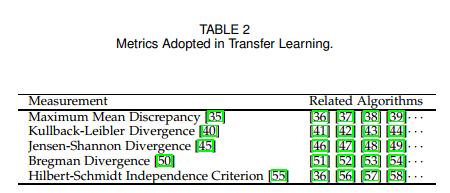
### 4.2.1 分布差异度量

特征转换的一个主要目标是减少源和目标域实例的分布差异。因此，如何有效地度量域之间的分布差异或相似性是一个重要的问题。

最大平均差异(Maximum Mean差值，MMD)在迁移学习领域得到了广泛的应用，其表达式为:



利用核技巧可以很容易地计算出MMD。简单地说，MMD通过计算RKHS中各实例均值的距离来量化分布差异。请注意，上述KMM实际上通过最小化域之间的MMD距离来生成实例的权重。表2列出了一些常用的度量标准和相关算法。除表2外，迁移学习中还采用了其他一些测量标准，包括Wasserstein distance[59]、[60]、中心矩差值[61]等。一些研究集中在对现有测量方法的优化和改进上。以MMD为例。Gretton等人提出了一种多内核版本的MMD，即MK-MMD[62]，它利用了多内核的优势。此外，Yan等人提出了一种加权版本的MMD[63]，试图解决类别权重偏差的问题。



### 4.2.2特征增强

特征增强操作在特征变换中得到了广泛的应用，特别是在对称特征变换中。具体来说，特征增强的实现方法有特征复制和特征叠加两种。为了更好地理解，我们从一个简单的迁移学习方法开始，它是建立在特征复制的基础上。

Daume的工作提出了一种简单的域自适应方法，即Feature Augmentation method (FAM)[64]。该方法通过简单的特征复制对原始特征进行变换。具体来说，在单源转移学习场景中，特征空间被扩大到原来的3倍大小。新的特性表示由通用特性、特定于源的特性和特定于目标的特性组成。注意，对于已转换的源域实例，它们的特定于目标的特性被设置为零。类似地，对于转换后的目标域实例，它们特定于源的特性被设置为零。FAM的新特征表示如下:



这里ΦS和ΦT分别表示从从源和目标域映射到新的特征空间。最终的分类器在转换后的标记实例上进行训练。值得一提的是，这种增强方法实际上是多余的。换句话说，以其他方式(使用更少的维度)增加特性空间可能能够产生合格的性能。FAM算法的优势在于其特征扩展具有优雅的形式，可以推广到多源场景中。Daume等人在[65]中提出了FAM的扩展，利用未标记的实例进一步促进知识转移过程。

然而，FAM在处理异质迁移学习任务时可能效果不佳。原因是，当源域和目标域具有不同的特征表示时，直接复制特征和填充零向量的效率较低。为了解决这个问题，Li等人提出了一种称为异构特征增强(HFA)的方法[66]，[67]。HFA的特征表示如下:



其中和具有相同的维数;和分别表示维数为和、一样的零向量。HFA将原始特征映射到一个公共特征空间，然后进行特征叠加操作。映射的特性、原始的特性和零元素按照特定的顺序堆叠，以产生新的特性表示。

### 4.2.3特征映射

在传统机器学习领域，基于映射的特征提取方法有很多可行的，如主成分分析(PCA)[68]和Kernelized-PCA (KPCA)[69]。但这些方法主要关注的是数据方差，而不是分布差异。为了求解分布差异，提出了一些用于迁移学习的特征提取方法。让我们首先考虑一个简单的场景，其中域的条件分布几乎没有差别。在这种情况下，可以使用以下简单的目标函数找到一个映射进行特征提取:



Φ是一个低维的映射函数,DIST(·)代表一个分布差异度量,Ω(Φ)是一种规范控制Φ的复杂性,VAR(·)代表实例的方差。这个目标函数的目的是找到一个映射函数Φ使得域之间的边缘分布差异最小,同时使实例间的方差尽可能大。与分母对应的目标可以通过多种方式进行优化。一种可能的方法是在方差约束下优化分子的目标。例如，映射实例的散点矩阵可以强制为单位矩阵。另一种方法是先在高维特征空间中对分子目标进行优化。然后通过PCA或KPCA等降维算法来实现分母的目标。

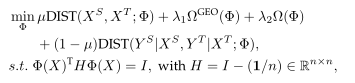
此外,发现Φ(·) 的显式是不容易的。为了解决这个问题，一些方法采用线性映射技术或转向核技巧。一般来说，有三种主要思想来处理上述优化问题。

•(映射学习+特征提取)一种可能的方法是首先找到一个高维空间，通过解决一个核矩阵学习问题或一个变换矩阵查找问题来满足目标。然后，将高维特征压缩成低维特征表示。例如，一旦学习了核矩阵，就可以提取隐式高维特征的主成分，构建基于PCA的新特征表示。

•(映射构造+映射学习)另一种方法是将原始特征映射到一个构造好的高维特征空间，然后学习低维映射来满足目标函数。例如，可以先根据选定的核函数构造核矩阵。然后，通过学习将高维特征投影到公共潜子空间的变换矩阵。

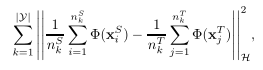
•(直接低维映射学习)通常很难直接找到想要的低维映射。然而，如果假设映射满足一定的条件，它可能是可解的。例如，如果将低维映射限制为线性映射，优化问题就可以很容易地解决。

有些方法还试图匹配条件分布并保留数据的结构。为了实现这一点，上述简单的目标函数需要包含新的术语或/和约束。例如，以下一般目标函数是一种可能的选择:

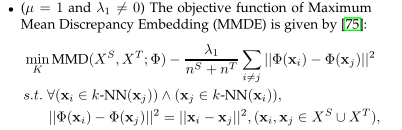


µ在哪一个参数平衡的边际条件分布差异[70],Ω地理(Φ)是一种规范控制几何结构,Φ(X)是矩阵的行是实例的源和目标域提取的新特性表示,H是定心矩阵构造散射矩阵,约束是用于最大化方差。目标函数中的最后一项表示条件分布差异的度量。

在进一步讨论上述目标函数之前，需要指出的是，目标域实例的标签信息往往是有限的，甚至是未知的。由于缺乏标签信息，很难估计分布差异。为了解决这个问题，一些方法采用**伪标签策略**，即为未标记的目标域实例分配伪标签。实现这一点的一个简单方法是训练基分类器来分配伪标签。另外，还有其他一些提供伪标签的方法，如联合训练[71]、[72]和三训练[73]、[74]。一旦伪标签信息被补全，就可以测量条件分布的差异。例如，MMD可以被修改和扩展来测量条件分布差异。具体来说，对于每个标签，收集属于同一类的源域实例和目标域实例，条件分布差异的估计表达式由[38]给出:



其中nS kand nT分别表示源域和目标域中具有相同标签Yk的实例数。这个估计实际上测量了类条件分布(即P(x|y))的差异来近似条件分布(即P(y|x))的差异。一些研究改进了上述估计。例如，Wang等人的工作采用加权方法另外解决了分类不平衡问题[70]。为了更好地理解，迁移学习方法是前一段提出的一般目标函数的特殊情况，具体如下。



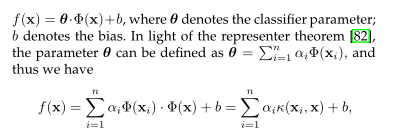
其中，k- nn (x)表示实例x的k个最近邻。作者基于最大方差展开(MVU)设计了上述目标函数[76]。与使用散射矩阵约束相反，约束和这个目标函数的第二项旨在最大化实例之间的距离以及保留局部几何。期望核矩阵K可以通过求解半定规划(SDP)[77]问题来学习。得到核矩阵后，将9个主成分分析应用于核矩阵，然后选取主导特征向量来帮助构造低维特征表示。

•(µ = 1 and λ1= 0) Pan等人的工作提出了一种称为传输分量分析(TCA)[36]的方法[78]。TCA采用MMD来测量边缘分布差异，并以散射矩阵作为约束。与MMDE先学习核矩阵再采用主成分分析不同，TCA是一种统一的方法，只需学习从经验核特征空间到低维特征空间的线性映射即可。这样避免了SDP问题的解决，使得计算量相对较低。最后的优化问题可以通过特征分解很容易地解决。还可以扩展TCA来利用标签信息。在扩展版本中，散点矩阵约束被一个新的约束取代，该约束平衡了标签依赖(由HSIC测量)和数据方差。此外，还增加了图形Laplacian regularizer[30]来保持流形的几何性质。同样，最终的优化问题也可以通过特征分解来求解。

•(µ = 0.5 and λ1= 0)Long等人提出了一种称为联合分布适应(JDA)[38]的方法。JDA试图找到一个转换矩阵，将实例映射到一个低维空间，在这个空间中边际分布和条件分布的差异都最小。为了实现这一目标，采用了多尺度度量和伪拉贝尔策略。通过特征分解求解轨迹优化问题，得到期望的变换矩阵。此外，很明显，估计的伪标签的准确性会影响JDA的性能。为了提高标记质量，采用了迭代优化操作。具体地说，在每次迭代中，执行JDA，然后使用提取的特征对实例进行分类器训练。接下来，基于训练好的分类器更新伪标签。之后，使用更新后的伪标签重复执行JDA。当收敛发生时迭代结束。需要注意的是，JDA可以利用标签和结构信息[79]、聚类信息[80]、各种统计和几何信息[81]等进行扩展。

•(0<µ <1 and λ1= 0)Wang等人的论文提出了一种称为平衡分布适应(BDA)的方法[70]，它是JDA的一种扩展。与JDA假设边际分布和条件分布在适应中具有同等重要性不同，BDA试图平衡边际分布和条件分布的适应。BDA的操作与JDA类似。此外，作者还提出了加权BDA (WBDA)。在WBDA中，通过加权的MMD版本来度量条件分布的差异，以解决类不平衡问题。

值得一提的是，有些方法将特征转换为新的特征空间(通常是高维的)，同时训练自适应分类器。为了实现这一点，需要将特征的映射函数与分类器的决策函数相关联。一种可能的方法是定义以下决策函数:



κ表示内核函数。利用核矩阵作为桥梁，为映射函数设计的正则化器可以并入到分类器的目标函数中。这样，最终的优化问题通常是关于参数(例如，再利用i)或核函数。例如，Long等人的论文提出了一个通用框架，称为基于自适应正则化的迁移学习(ARTL)[39]。ARTL的目标是学习自适应分类器，最小化结构风险，共同降低边际分布和条件分布差异，最大化数据结构和预测结构之间的流形一致性。在此框架下，基于不同的损失函数，提出了两种具体的算法。在这两种算法中，首先构造了计算多模化的系数矩阵和用于流形正则化的图拉普拉斯矩阵。然后选择一个核函数来构造核矩阵。然后将分类器学习问题转化为一个参数(即:蜗轮)求解问题，求解公式也在[39]中给出。

在ARTL中，核函数的选择将影响最终分类器的性能。为了构造鲁棒分类器，一些研究转向核学习。例如，Duan等人提出了一个统一的框架，称为域转移多核学习(DTMKL)[83]。在DTMKL中，假设核函数是一组基核的线性组合, DTMKL的目标是同时最小化分布差异、分类误差等。DTMKL的一般目标函数可以写成:



σ是单调递增函数f是相同的决策函数定义的ARTL,ΩL(βk, f)是一个一般术语代表一群regularizers等标注实例定义的最小化分类误差和控制合成模型的复杂性。作者开发了一种算法，利用降低梯度下降法同时学习核函数和决策函数[84]。在每次迭代中，确定基核的权重系数，首先更新决策函数。然后，确定决策函数，更新权重系数。注意，DTMKL可以合并许多现有的内核方法。在此框架下，作者提出了两种具体的算法。第一种是利用铰链损耗和支持向量机实现框架。第二个是第一个的扩展，带有额外的正则化器，利用伪标签信息，并且通过使用基本分类器生成未标记实例的伪标签。

### 4.2.4特征聚类

特征聚类的目的是找到对原始特征更抽象的特征表示。虽然可以看作是一种特征提取方法，但它不同于上述基于映射的提取方法。

例如，一些迁移学习方法通过使用共聚技术隐式地减少特征，即基于信息论同时聚类(或者说共聚)列联表的行和列[85]。Dai等人[41]在论文中提出了一种基于共聚的分类(CoCC)算法，用于文档分类。在文档分类问题中，迁移学习任务是利用带标记的（源文档-词）数据对目标域文档(用文档-单词矩阵表示)进行分类。CoCC将聚类技术作为知识传递的桥梁。在CoCC算法中，源和目标文档-字矩阵是共簇的。根据已知的标签信息，将源文档-词矩阵共聚生成词簇，并在目标域数据共聚过程中使用这些词簇作为约束条件。**共聚的准则是尽量减少相互信息的损失，通过迭代得到聚类结果**。每个迭代包含以下两个步骤。

1.文档集群:根据更新文档集群的目标函数，对目标文档-word矩阵的每一行重新排序。

2. 单词聚类:调整单词聚类以最小化源文档-单词矩阵和目标文档-单词矩阵的共同互信息丢失。

经过多次迭代，算法收敛，得到分类结果。注意，在CoCC中，单词聚类过程隐式提取单词特征，形成统一的单词聚类。

Dai等人还提出了一种无监督聚类方法，称为自教聚类(STC)[42]。与CoCC相似，该算法也是一种基于聚类的算法。但是，STC不需要标签信息。STC的目的是同时对源域和目标域实例进行协同聚类，假设这两个域在它们的共同特征空间中共享相同的特征集群。因此，在同一时间分别执行两个共聚任务，寻找共享特征聚类。STC的每次迭代都有以下步骤。

1. 实例聚类:更新源节点实例和目标域实例的聚类结果，使它们各自的互信息丢失最小化。

2. 特征聚类:对特征聚类进行更新，使相互信息的联合损失最小化。

当算法收敛时，得到目标域实例的聚类结果。

与上述基于聚类的方法不同，有些方法是将原始特征提取到概念(或主题)中。在文档分类问题中，概念代表了单词的高级抽象性(例如，词簇)。为了方便地介绍基于概念的迁移学习方法，让我们简要地回顾一下潜在语义分析(Latent Semantic Analysis, LSA)[86]、概率语义分析(probability LSA, PLSA)[87]和双语义分析(Dual-PLSA)[88]。

•潜在语义分析LSA: LSA是一种基于奇异值分解(SVD)技术将（文档-词）矩阵映射到低维空间(即潜在语义空间)的方法。简而言之，LSA试图找到这些单词的真正含义。为了实现这一目标，利用奇异值分解(SVD)技术进行降维，可以去除原始数据中的无关信息和滤除噪声信息。

•概率语义分析PLSA: PLSA是基于LSA的统计观点而开发的。PLSA假设存在一个潜在的类变量z，它反映了这个概念，将文档d和单词w联系在一起，并且d和w各自依赖于概念z，该图形模型的示意图如下:



其中，下标i、j和k分别表示文档、单词和概念的索引。PLSA构造一个贝叶斯网络，使用期望最大化(EM)算法估计参数[89]。

•双语义分析Dual-PLSA: Dual-PLSA是PLSA的扩展。这种方法假设有两个潜在变量zd 和 zw关联文档和单词。具体来说，变量zd和zw分别反映了文档和单词背后的概念。以下是双PLSA的图表:



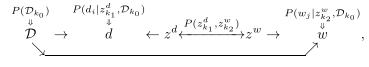
基于EM算法也可以得到双PLSA的参数。

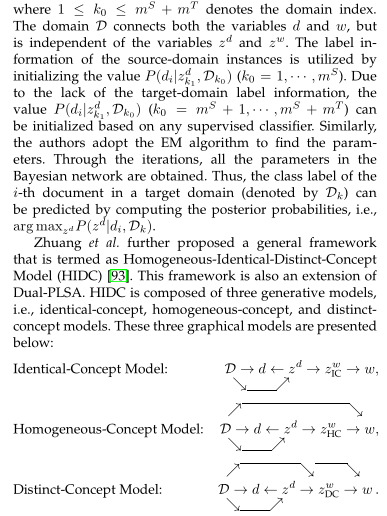
基于PLSA建立了一些基于概念的迁移学习方法。例如，Xue等人提出了一种跨域文本分类方法，称为话题桥接概率潜在语义分析(topic -桥接概率潜在语义分析，TPLSA)[90]。TPLSA是PLSA的扩展，它假设源域实例和目标域实例共享相同的单词混合概念。本文不是分别对源域和目标域执行两个PLSAs，而是利用混合概念z作为桥梁，将这两个PLSAs合并为一个整体，即每个概念都有一定的概率产生源域和目标域文档。TPLSA的示意图如下:



注意，PLSA不需要标签信息。为了更好地利用标签信息，在TPLSA的目标函数中加入了概念约束作为惩罚项，其中包括必须链接约束和不能链接约束。最后，利用EM算法迭代优化目标函数，得到分类结果(即arg maxzP(z|dT i))。

Zhuang等人的工作提出了一种用于多域文本分类(mSsource域和mTtarget域)的协作双plsa (CD-PLSA)方法11[91]和[92]。CD-PLSA是双plsa的扩展。其示意图如下:





原词概念zw分为zw IC、zw HC和zw DC三种类型。在同位概念模型中，单词分布仅依赖于单词概念，单词概念独立于域。然而，在均质概念模型中，词的分布也依赖于域。相同概念和同构概念的区别在于，zw IC是直接可转移的，而zw是特定领域的可转移概念，可能对不同领域的词分布产生不同的影响。在特殊性概念模型中，zw DC实际上是不可转移的特定于领域的DC，它可能只出现在特定的领域中。将上述三种模式结合为一个整体，即HIDC。与其他PLSA相关算法类似，HIDC也使用EM算法来获取参数。

### 4.2.5特征选择

特征选择是特征降维的另一种操作，用于提取主特征。主元特征是指在不同区域表现相同的特征。由于这些特征的稳定性，它们可以作为知识传递的桥梁。例如，Blitzer等人提出了一种称为结构对应学习 (SCL)的方法[94]。简单地说，SCL由以下步骤组成，以构造一个新的特性表示。

1. 特征选择:SCL首先进行特征选择操作，获得主元特征。

2. 映射学习:使用结构学习技术，利用枢轴特征寻找低维公共潜在特征空间[95]。3.特征叠加:通过特征增强构造一种新的特征表示，即将原始特征与得到的低维特征叠加。

以词性标注问题为例。所选的枢轴特征应该经常出现在源和目标域中。因此，限定词可以包含在主元特征中。一旦定义并选择了所有的主元特征，就可以构造出大量的二元线性分类器，其功能是预测每个主元特征的出现。在不失一般性的情况下，用于预测第i个支点特征的第i个分类器的决策函数可以表示为fi(x) = sign(nonversioni·x)，其中假设x是一个二元特征输入。第i个分类器在除第i个支点特征衍生的特征外的所有实例上训练。第i个分类器参数的估计可采用以下公式:



其中，Rowi(xj)表示第i个主元特征的未标记实例xjin项的真实值。通过叠加获得参数向量为列元素,一个矩阵˜W。接下来,基于奇异值分解),top-k左奇异向量,矩阵的主成分的˜W,来构造变换矩阵W .最后,标签上的最终的分类器训练实例在一个增广特征空间,即,(xL我;WTxL我T, yL i)。

### 4.2.6特征编码

除了特征提取和选择之外，特征编码也是一种有效的工具。例如，自编码器可以用于特征编码，这是深度学习领域经常使用的。一个编码器包括编码器和解码器。编码器试图对输入产生一个更抽象的表示，而译码器的目标是映射回那个表示，并尽量减少重构误差。自动编码器可以堆叠，以建立一个深度学习架构。一旦一个自动编码器完成训练过程，另一个自动编码器可以堆叠在它的顶部。然后，通过使用高级自动编码器的编码输出作为输入来训练新添加的自动编码器。这样就可以构建深度学习架构。

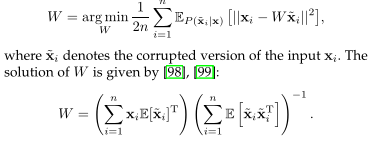
基于自编码器开发了一些迁移学习方法。例如，Glorot等人的论文提出了一种称为堆叠去噪自编码器(SDA)的方法[96]。去噪自编码器是对基础自编码器的扩展，增强了鲁棒性[97]。这种类型的自动编码器包含一个随机破坏机制，在映射之前给输入添加噪音。例如，一个输入可以通过添加掩蔽噪声或高斯噪声被破坏或部分破坏。然后对去噪自编码器进行训练，使去噪重建误差最小化。本文提出的SDA算法主要包括以下几个步骤。

1. Autoencoder T raining:源域和目标域实例被用来训练一堆自动编码器去噪贪婪的一层一层的方式。

2. 特征编码与叠加:将中间层的编码输出进行叠加，构造新的特征表示，并将实例的特征转化为得到的新表示。

3.学习者T raining:目标分类器在转换后的标记实例上进行训练。

虽然SDA算法在特征提取方面有很好的性能，但仍存在计算和参数估计代价高的缺点。为了缩短训练时间，加快传统SDA算法的速度，Chen等人提出了一种SDA的改进版本，即边缘化堆叠线性去噪自动编码器(mSLDA)[98]，[99]。该算法采用线性自编码器，以封闭形式将随机破坏步骤边缘化。似乎线性自动编码器太简单了，无法学习复杂的特性。然而，作者观察到，当遇到高维数据时，线性自动编码器通常足以达到胜任的性能。mSLDA的基本结构是一个单层线性自编码器。对应的单层映射矩阵W(为了方便用偏置列增广)应使重构的平方损失函数最小化，即:



当腐败策略确定后，上述公式可以进一步扩展简化为特定形式。注意，为了插入非线性，在我们得到一个封闭形式的矩阵W之后，一个非线性函数被用来压缩每个自动编码器的输出。然后，下一个线性自动编码器以类似于SDA的方式堆叠到当前的那个。为了处理高维数据，提出了一种可扩展的方法来进一步降低计算复杂度。

### 4.2.7特征对齐

注意，特征增加和特征减少主要关注特征空间中的显式特征。相比之下，除了明确的特征外，特征对齐还侧重于一些隐含的特征，如统计特征和光谱特征。因此，特征对齐可以在特征转换过程中发挥各种作用。例如，可以对显式特征进行对齐以生成新的特征表示，或对隐式特征进行对齐以构造满意的特征转换。

可对齐的特征包括子空间特征、谱特征和统计特征。以子空间特征对齐为例。典型的方法主要有以下步骤。

1. 子空间生成:在此步骤中，使用实例为源域和目标域生成各自的子空间。然后得到源和目标域子空间的标准正交基，分别用m表示和MT表示。这些基用于学习子空间之间的转换。

2. 子空间对齐:在第二步，一个映射，它对齐子空间的基ma和MT，被学习。并将实例的特征投影到对齐的子空间中，生成新的特征表示。

3.学习者T raining:最后，目标学习者在转换的实例上训练。

例如，Fernando等人的工作提出了一种称为子空间对齐(SA)的方法[100]。在逻辑分析中，子空间由主成分分析生成;通过选择主特征向量得到基MS和基mtn。然后学习一个变换矩阵W来对齐子空间，由[100]给出:



其中||·|| f为Frobenius规范。注意矩阵W使msmt对齐，或者说，将源子空间坐标系转换为目标子空间坐标系。分别用XSMSW和XTMT给出了转换后的低维源域实例和目标域实例。最后，学习者可以根据转换后的实例训练。

在此基础上，建立了多种迁移学习方法。例如，Sun和Saenko的论文提出了一种对齐子空间基和分布的方法[101]，称为两个子空间之间的子空间分布对齐(SDA-TS)。在SDA-TS中，变换矩阵W表示为W = MT SMTQ，其中Q是用于对齐分布差的矩阵。SA中的变换矩阵W是SDA-TS中的一种特殊情况，它将Q设为单位矩阵。请注意，SA是一种对称的基于特性的方法，而da - ts是一种不对称的方法。在SDATS中，已标记的源域实例被投影到源子空间，然后映射到目标子空间，最后映射回目标域。转换后的源域实例被表示为XSMSW MT T。

另一种具有代表性的子空间特征对齐方法是由龚等人提出的测地线流核(GFK)[102]。GFK与之前的测地线流子空间(GFS)方法密切相关[103]。在介绍GFK之前，让我们先回顾一下GFS的步骤。GFS的灵感来自于渐进式学习。直观地看，利用两个域之间的潜在路径所传达的信息可能有利于域的适应。GFS通常采取以下步骤来对齐特性。

1. 子空间生成:GFS首先通过PCA分别生成源域和目标域的两个子空间。

2. 子空间插值:得到的两个子空间可以看作是Grassmann流形上的两点[104]。基于流形的几何性质，在这两个子空间之间生成了有限数量的插值子空间。

3.特征投影叠加:将得到的所有子空间中对应的投影进行叠加，对原始特征进行变换。

尽管GFS很有用，也很优越，但是如何确定内插子空间的数量是一个问题。GFK通过将位于测地线曲线上的无穷多个子空间从源子空间积分到目标子空间来解决这个问题。GFK的关键是构造一个包含测地线流上所有子空间信息的无限维特征空间。为了计算无穷维空间的内积，定义并推导了测地流核。此外，提出了一种子空间不一致测度来选择子空间的最优维数;在多源域可用的情况下，提出了一种用于选择最优源域的秩域度量。

统计特征对齐是特征对齐的另一种。例如，Sun等人提出了一种称为关联对齐(CORAL)的方法[105]。CORAL通过对齐二阶统计特征，即协方差矩阵，构建源特征的变换矩阵。变换矩阵W由[105]给出:



式中，C为协方差矩阵。请注意，与上述基于子空间的方法相比，CORAL避免了子空间生成和投影，而且非常容易实现

建立了基于光谱特征对齐的迁移学习方法。在传统的机器学习领域，谱聚类是一种基于图论的聚类技术。该技术的关键是在聚类前利用相似矩阵的谱即特征值对特征进行降维。建立相似度矩阵来定量评估每对数据/顶点的相对相似度。在光谱聚类和特征对齐的基础上，Pan等人提出了光谱特征对齐(SFA)[106]。SFA是一种情绪分类算法。该算法尝试识别不同领域中的领域专用词和领域独立词，然后将这些领域专用词特征进行比对，以构建低维特征表示。SFA一般包含以下五个步骤。

1. 特性选择:在此步骤中，执行特性选择操作以选择域独立/枢轴特性。本文提出了三种选择领域独立特征的策略。这些策略分别基于词的出现频率、特征与标签之间的互信息[107]、特征与域之间的互信息。

2. 相似矩阵构造:一旦识别了领域特定和领域独立的特征，就构造一个二部图。这个二部图的每条边都被赋予一个权值，该权值度量领域特定词和领域独立词之间的共现关系。在二部图的基础上，构造了相似矩阵。

3.光谱特征对齐:在此步骤中，采用光谱聚类算法对特定领域的特征进行对齐[108]，[109]。具体来说，基于图Laplacian的特征向量构造特征对齐映射，将领域特有的特征映射到低维特征空间中。

4. 特征叠加:将原始特征与低维特征进行叠加，得到最终的特征表示。

5. 学习者T raining:目标学习者在被标记的实例上进行训练，最终得到特征表示。

还有一些其他的光谱转移学习方法。例如，Ling等人的工作提出了一种称为跨域光谱分类器(CDSC)的方法[110]。此方法的一般思想和步骤如下所示。

1. 相似度矩阵构建:第一步，分别构建对应于整个实例和目标域实例的两个相似度矩阵。

2. 光谱特征对齐:针对图划分指标向量设计目标函数;构造了一个约束矩阵，其中包含成对的必链接信息。不再寻求指标向量的离散解，而是将解放宽为连续解，从而解决目标函数对应的特征系统问题，构建对齐谱特征[111]。

3.学习者T raining:在转换的实例上训练传统的分类器。

更具体地说，目标函数具有广义瑞利商的形式，其目的是寻找最优的图划分，尊重裁剪小的标签信息[112]，最大限度地分离目标域实例，并拟合成对性质的约束。特征分解后，选择最后的特征向量组合成一个矩阵，然后对矩阵进行归一化。标准化矩阵的每一行表示一个转换后的实例。

# 5基于模型的解释

迁移学习方法也可以从模型的角度进行解释。图4给出了相应的策略和目标。转移学习模型的主要目的是对目标域做出准确的预测结果，如分类或聚类结果。注意，一个迁移学习模型可能由一些子模块组成，如分类器、提取器或编码器。这些子模块可以发挥不同的作用，如特征适应、伪标签生成等。在本节中，根据图4中所示的策略以适当的顺序介绍了一些相关的转移学习方法。

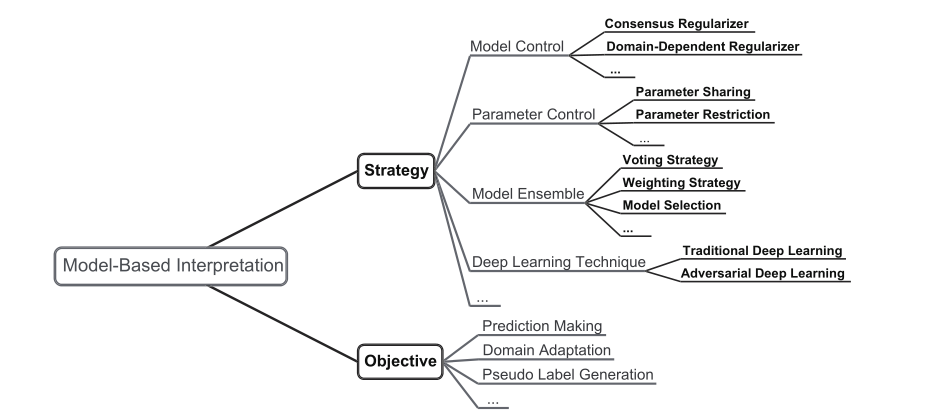


图 从模型角度来看迁移学习的策略和目标

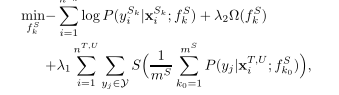
## 5.1模型控制策略

从模型的角度来看，一种自然的想法是直接将模型级的正则化添加到学习者的目标函数中。这样，在训练过程中，可以将预先获得的源模型中包含的知识转移到目标模型中。例如，Duan等人提出了一种通用框架，称为域适应机(DAM)[113]、[114]，该框架用于多源迁移学习。DAM的目标是利用在多个源域上分别训练得到的一些预先获得的基分类器，为目标域构造一个鲁棒分类器。目标函数为:



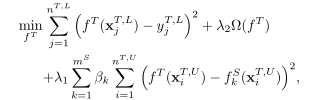
在第一项表示损失函数用于最小化分类误差的标记目标领域实例,第二项表示不同的正则化项,第三项是用于控制的复杂性最终决定函数英尺。不同类型的损失函数可以采用LT, L(英尺)如平方误差或叉损失。一些迁移学习方法在一定程度上可以看作是该框架的特例。

•(Consensus Regularizer) Luo等人的工作提出了一个框架称为共识正则化框架(CRF)[115]，[116]。CRF是专为无标记目标域实例的多源转移学习而设计的。该框架构建了对应于每个源域的msclassifier，这些分类器需要在目标域上达成一致。各源分类器的目标函数fS k (k = 1，···，mS)与DAM相似，如下图所示:



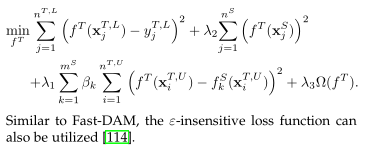
其中fS k表示第k个源域对应的决策函数，S(x) =−xlogx。第一项用于量化第k个分类器在第k个源域上的分类误差，最后一项是交叉正则化的一致性调节器。一致性调节器不仅提高了各分类器的一致性，而且降低了预测结果对目标域的不确定性。作者在逻辑回归的基础上实现了这个框架。DAM与CRF的不同之处在于，DAM明确地构建了目标分类器，而CRF则根据源分类器达成的一致意见来进行目标预测。

•(Domain-dependent Regularizer) Fast-DAM是DAM的一种特定算法[113]。根据流形假设[30]和基于图的正则化器[117]、[118]，Fast-DAM设计了一个领域相关的正则化器。目标函数为:



其中fS k(k = 1,2，··，mS)为预先得到的第k个源域的源决策函数，kkk为目标域与第k个源域的相关性所确定的可基于MMD度量的加权参数。第三个术语是领域依赖的正则化器，它传输由领域依赖激发的源分类器中包含的知识。在[113]中，作者还在上述基于对时不敏感损失函数的目标函数中引入并增加了一个新项[119]，使得所得到的模型具有较高的计算效率。

•(领域依赖的调节器+ Universum调节器)Univer-DAM是Fast-DAM的扩展[114]。它的目标函数包含一个额外的正则化器，即Universum正则化器。这个正则化器通常使用一个附加的数据集，称为Universum，其中实例既不属于正类也不属于负类[120]。将源域实例作为目标域的Universum, Univer-DAM的目标函数如下:

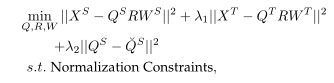


## 5.2参数控制策略

参数控制策略主要关注模型的参数。例如，在对象分类的应用中，可以通过物体的形状、颜色等属性，将来自已知源类别的知识转化为目标类别[121]。属性先验，即每个属性对应的图像特征的概率分布参数，可以从源域学习，然后用于学习目标分类器。模型的参数实际上反映了模型所学习的知识。因此，可以在参数水平上转移知识。

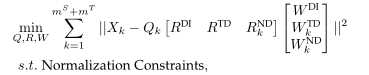
### 5.2.1参数共享

一种直观的控制参数的方法是直接将源学习者的参数共享给目标学习者。参数共享被广泛应用于基于网络的方法中。例如，如果我们有一个用于源任务的神经网络，我们可以冻结(或者说共享)它的大部分层，只调整最后几个层来生成目标网络。基于网络的方法将在5.4节中介绍。除了基于网络的参数共享外，基于矩阵分解的参数共享也是可行的。例如，Zhuang等人提出了一种文本分类方法，称为基于矩阵三因子分解的分类框架(Matrix trifactorization Based classification Framework, MTrick)[122]。作者注意到，在不同的领域，不同的词或短语有时表达相同或相似的内涵意义。因此，使用单词背后的概念比使用单词本身作为源领域知识传递的桥梁更有效。与基于plsa的迁移学习方法通过构造贝叶斯网络来利用概念不同，MTrick试图通过矩阵三因子分解来找到文档类和词簇所表达的概念之间的联系。这些联系被认为是应该被转移的稳定知识。其主要思想是将文档到word矩阵分解为三个矩阵，即文档到集群、连接和集群到单词矩阵。具体而言，分别对源文档到字矩阵和目标文档到字矩阵进行矩阵三因式分解，构造出一个联合优化问题



X表示document-to-word矩阵Q表示document-to-cluster矩阵,R代表文档集群词群的变换矩阵,W表示cluster-to-word矩阵,nddenotes文档的数量,˘QSrepresents标签矩阵。矩阵˘QSis构造基于源领域的类信息文档。如果第i个文档属于k类,˘QS (i (k) = 1。在上述目标函数中，矩阵R实际上是共享参数。第一个术语旨在对源文档到单词矩阵进行三因子分解，第二个术语分解目标文档到单词矩阵。最后一个术语包含源域标签信息。采用交替迭代法求解优化问题。一旦得到QTis的解，第k个目标域实例的类索引就是QT第k行中取值最大的类索引。

此外，Zhuang等人扩展了MTrick并提出了一种称为三形迁移学习(TriTL)的方法[123]。MTrick假设这些域在它们的词簇背后共享相似的概念。相反，TriTL假设这些域的概念可以进一步划分为三类，即独立于域的、可转移的域特定的和不可转移的域特定的概念，这与HIDC类似。这种想法是由双重转移学习(DTL)激发的，其中的概念被假定由独立于领域的概念和可转移的特定于领域的概念组成[124]。TriTL的目标函数如下:



其中，符号的定义类似于MTrick的定义，subscriptk表示域的索引，假设第一个mSdomains是源域，最后一个mTdomains是目标域。提出了一种求解优化问题的迭代算法。初始化阶段根据PLSA算法的聚类结果对WDIand WTD k进行初始化，对WUT k进行随机初始化;PLSA算法是在所有域实例的组合上执行的。

还有其他一些基于矩阵分解的方法。Wang等人提出了一种用于图像分类的迁移学习框架[125]。Wang等人提出了一种软关联方法，将两个矩阵三分解集成到一个联合框架中[126]。Do等人利用矩阵三因子分解来发现跨域推荐的隐式和显式相似性[127]。

### 5.2.2参数限制

另一个参数控制类型的策略是限制参数。与强制模型共享某些参数的参数共享策略不同，参数限制策略只要求源模型和目标模型的参数相似。

以类别学习的方法为例。分类学习问题是学习一个新的决策函数来预测一个新的类别(表示为(k + 1)第一个类别)，只有有限的目标域实例和k个预先获得的二值决策函数。这些预先获得的决策函数的作用是预测一个实例属于k个类别中的哪一个。为了解决分类学习问题，Tommasi等人提出了一种称为单模型知识转移(single model Knowledge Transfer, SMKL)的方法[128]。SMKL基于最小二乘SVM (LS-SVM)。LS-SVM的优点是将不等式约束转化为等式约束，具有较高的计算效率;它的优化等价于求解一个线性方程组问题而不是一个二次规划问题。SMKL选择一个预先获得的二元决策函数，并传递其参数中包含的知识。目标函数为



式中，分别为模型参数和第i个预得决策函数的加权参数。漏出误差也可以以一种封闭的形式得到，其中最优的值(i = 1,2，··，k)就是泛化性能最大的值。

## 5.3模型集成策略

在与产品评论相关的情绪分析应用中，来自多个产品领域的数据或模型是可用的，可以用作源领域[131]。将数据或模型直接组合到单个域可能不会成功，因为这些域的分布彼此不同。模型集成是另一个常用的策略。该策略的目的是结合大量的弱分类器来进行最终的预测。之前提到的一些迁移学习方法已经采用了这种策略。例如，TrAdaBoost和MsTrAdaBoost分别通过投票和加权来集成弱分类器。在本小节中，将介绍几种典型的基于ensembedtransferlearning方法，以帮助读者更好地理解这种策略的功能和应用。

如4.1节所述，论文[33]中提出了TaskTrAdaBoost，它是TrAdaBoost的扩展，用于处理多源场景。TaskTrAdaBoost主要有以下两个阶段。

1. 候选分类器构造:在第一阶段，通过对每个源域进行AdaBoost来构造一组候选分类器。请注意，对于每个源域，AdaBoost的每次迭代都会产生一个新的弱分类器。为了避免过拟合问题，引入了一个阈值来选择合适的分类器到候选组中。

2. 分类器选择与集成:第二阶段，在targetdomain实例上执行AdaBoost的修订版本，构建最终的分类器。在每一次迭代中，挑选出对所标记目标域实例分类误差最小的最优候选分类器，并根据分类误差分配权重。然后，根据所选分类器在目标域中的性能更新每个目标域实例的权值。在迭代过程后，选择的分类器被集成，以产生最终的预测。

原始AdaBoost与第二阶段TaskTrAdaBoost的区别在于，在每次迭代中，原始AdaBoost在加权的目标域实例上构造一个新的候选分类器，而后者在加权的目标域实例上选择一个预先获得的分类误差最小的候选分类器。

Gao等人的论文提出了另一种基于ensembleframework的框架，称为局部加权集成(LWE)[132]。LWE关注不同学习者的合奏过程;这些学习者可以构建在不同的源域上，也可以通过在单一源域上执行不同的学习算法来构建。不同于TaskTrAdaBoost每个学习者的学习全球体重,作者采用了local-weight策略,也就是说,将自适应权重分配给学习者基于局部流形结构的目标领域测试集。在LWE,学习者通常与不同的权重分配,分类不同的目标领域实例。具体地说，作者采用了一种基于图的方法来估计权重。权重的步骤概述如下。

1. 图构造:对于第i个源学习器，使用该学习器构造一个图GT Siis，对测试集中的17个目标域实例进行分类;如果两个实例被归类到同一个类中，它们在图中被连接起来。GTis还通过执行集群算法为目标域实例构造了另一个图。

2. 学习者权重:第i个学习者对第j个目标域实例xT j的权重与实例在GT Si和GT中的局部结构相似度成正比。相似度可以通过两个图中xT的共同邻域所占的百分比来衡量。

注意，该加权方案是基于聚类流形假设的，也就是说，如果两个实例在一个高密度区域内彼此接近，它们通常具有相似的标签。为了检查此假设对任务的有效性，在源域训练集上测试目标任务。具体来说，训练集的聚类质量通过使用纯度或熵等度量来量化和检查。如果聚类质量不理想，则给学习者分配均匀的权值。此外，直观地看，如果测量到的结构相似性对于每个学习者来说都特别低，那么对这些学习者进行加权和组合似乎是不明智的。为此，本文引入了一个阈值，并与平均相似度进行了比较。当相似度低于阈值时，xT j的标签由其可靠邻域之间的投票方案决定，其中可靠邻域是由组合分类器进行标签预测的那些。

上述TaskTrAdaBoost和LWE方法主要针对集成过程。与此相反，一些研究更关注弱势学习者的构建。例如锚定适配器的集成框架(Ensemble Framework of Anchor adapter, ENCHOR)[133]就是Zhuang等人提出的加权集成框架。锚是一个特定的实例。不同于TrAdaBoost通过反复调整实例的权重来训练和产生一个新的学习者，ENCHOR通过使用锚产生的实例的不同表示来构造一组弱学习者。其思想是，一个实例与锚点之间的相似性越高，该实例的特征相对锚点保持不变的可能性就越大，其中相似性可以通过余弦或高斯距离函数来度量。ENCHOR包含以下步骤。

1. 锚点选择:在此步骤中，选择一组锚点。这些锚可以根据某些规则或甚至是随机选择。为了提高ENCHOR的最终性能，作者提出了一种选择高质量锚的方法[133]。

2. 基于锚点的表示生成:对于每个锚点和每个实例，实例的特征向量直接乘以一个度量实例到锚点距离的系数。通过这种方式，每个锚都会产生一对新的适应锚的源和目标实例集。

3.学习者T raining和Ensemble:得到的对实例集可以分别用于训练学习者。然后，对所得到的学习者进行加权和组合，以做出最终的预测。

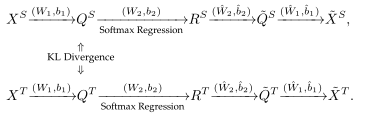
由于在每个锚点上执行的操作是独立的，因此框架ENCHOR很容易以并行方式实现。

## 5.4深度学习技术

深度学习方法在机器学习领域尤其流行。许多研究者利用深度学习技术来构建迁移学习模型。例如，第4.2.6节中提到的SDA和mSLDA方法利用了深度学习技术。在本节中，我们特别讨论与深度学习相关的迁移学习模型。本文介绍的深度学习方法分为两类，即非对抗性(或者说传统的)方法和对抗性方法。

### 5.4.1传统深度学习

如前所述，自动编码器经常用于深度学习领域。除了SDA和mSLDA之外，还有其他一些基于重构的迁移学习方法。例如，Zhuang等人的论文提出了一种名为Deep Autoencoders (TLDA)[44]迁移学习的方法[134]。TLDA对源域和目标域分别采用了两个自动编码器。这两个自动编码器共享相同的参数。编码器和解码器都具有具有激活功能的两层。两种自动编码器的示意图如下:



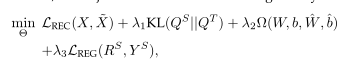
TLDA有几个目标，如下所列。

1. 重构误差最小化:译码器的输出应该非常接近编码器的输入。换句话说,之间的距离XSand˜XSas以及之间的距离XTand˜XTshould最小化。

2. 分布适应:应尽量减小qt和qt之间的分布差异。

3.回归误差最小化:编码器在标记的源域实例上的输出，即RS，应该与相应的标签信息YS一致。

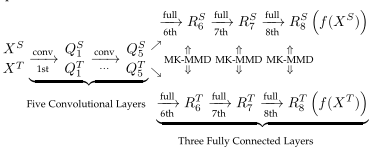
因此，TLDA的目标函数为



其中，第一项为重构误差，KL(·)为KL散度，第三项控制复杂度，最后一项为回归误差。TLDA采用梯度下降法进行训练。最终的预测有两种不同的方式。第一种方法是直接使用编码器的输出来进行预测。第二种方法是将自动编码器视为一个特征提取器，然后用编码器的第一层输出产生的特征表示在标记的实例上训练目标分类器。

除了基于重构的领域适应外，基于离散的领域适应也是一个流行的方向。在早期的研究中，浅层神经网络试图学习领域无关的特征表示[135]。研究发现，由于体系结构较浅，所得到的模型往往难以达到良好的性能。因此，许多研究转向利用深度神经网络。Tzeng等[136]在深度神经网络中加入了单一自适应层和差异损失，提高了性能。此外，Long等人进行了多层自适应并使用了多核技术，他们提出了一种称为深度自适应网络(DAN)的架构[137]。

为了更好地理解，让我们详细回顾一下DAN。DAN基于AlexNet[138]，其架构如下所示[137]。



在上述网络中，首先按照一般到特定的方式，通过5个卷积层提取特征。然后，提取的特征被输入到两个完全连接的网络中的一个由其原来的领域交换。这两个网络都由三个完全连接的层组成，它们专门用于源和目标域。DAN有以下目标。

1. 最小化分类误差:最小化被标记实例的分类误差。采用交叉熵损失函数度量标记实例的预测误差。

2. 分布适应:多个层，包括表示层和输出层，可以共同按层适应。本文采用MK-MMD代替单核MMD来测量分布差异。采用MK-MMD的线性时间无偏估计，避免了大量的内积运算[62]。

3.核参数优化:要优化MK-MMD中多个核的加权参数，使试验功率最大[62]。

DAN网络的目标函数为:



式中l为该层的索引。上述优化实际上是一个极小极大优化问题。最大化的目标函数关于核函数κ旨在测试力量最大化。在这一步之后，源域和目标域之间的细微差别将被放大。这一思路与生成式对抗网络(GAN)相似[139]。在训练过程中，DAN网络通过预先训练好的AlexNet进行初始化[138]。需要学习的参数有两类，即网络参数和多个核的加权参数。考虑到前三个卷积层输出一般特征且可转移，作者将其冻结，并对后两个卷积层和两个全连接层进行微调[140]。最后一个完全连接层(或者说，分类器层)是从头开始训练的。

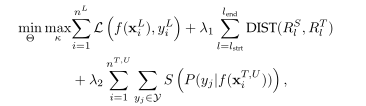
Long等人进一步扩展了上述DAN方法，提出了DAN框架[141]。新特点概括如下。

1. 添加正则化器:该框架引入了额外的正则化器来最小化未标记目标域实例的预测标签的不确定性，其动机是熵最小化准则[142]。

2. 架构泛化:DAN框架可应用于许多其他架构，如GoogLeNet[143]和ResNet[144]。

3.度量一般化:分布差异可以通过其他度量来估计。例如，除了MK-MMD之外，作者还提出了用于分布适应的平均嵌入检验[145]。

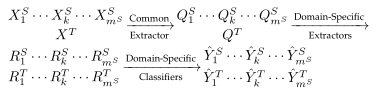
DAN框架的目标函数为:



其中，lstr和lendi表示用于适应分布的全连通层的边界指标。

还有其他一些令人印象深刻的作品。例如，Long等人构建了基于深度残差学习的区域自适应残差转移网络[146]。此外，Long等人的另一项研究提出了联合自适应网络(JAN)[147]，该网络适应了多层的联合分布差异。Sun和Saenko对珊瑚进行了深度域适应扩展，并提出了一种称为deep CORAL (d珊瑚)的方法，其中增加了珊瑚损失以最小化特征协方差[148]。Chen等人意识到具有相同标签的实例在特征空间中应该彼此靠近，他们不仅增加了珊瑚损失，还增加了基于实例的类级差异损失[149]。Pan等构建了三个原型网络(对应于DS、dtdt和DS∪DT)，并纳入了多模型共识的思想。采用伪标签策略，同时适应实例级和类级的差异[150]。Kang等人提出了对比自适应网络(CAN)，该网络基于差异度量称为对比域差异[151]。Zhu等人针对提取的多个特征表示进行适应，提出了多表示自适应网络(MRAN)[152]。

深度学习技术也可以用于多源迁移学习。例如，Zhu等人的工作提出了一个框架，称为多特征空间自适应网络(MFSAN)[153]。MFSAN的体系结构由公共特征提取器、特定于mS领域的特征提取器和特定于mS领域的19个分类器组成。对应的原理图如下图所示。



在每次迭代中，MFSAN有以下步骤。

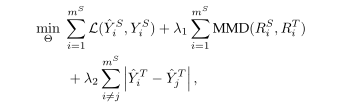
1. 共同的特征提取:每个源域(用DSkwith k = 1,···,mS), sourcedomain实例(用XS k)分别输入共同特征提取器在一个共同的潜在生产实例特征空间(用QS k)。类似的操作也在目标领域实例上执行(用XT),并产生QT。

2。特定特征提取:对于每个源域，提取的公共特征QS kis提供给第k个域特定特征提取器。同时，将QTis输入到所有领域特有的特征提取器，得到RT kwith k = 1，··，mS.

3。数据分类:将第k个领域特有特征提取器的输出输入到第k个分类器。从而以概率的形式预测分类结果的mSpairs。

4. 参数更新:更新网络参数，优化目标函数。

MFSAN有三个目标，即最小化分类误差、分布适应和一致性正则化。目标函数为:



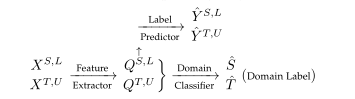
### 5.4.2对抗性深度学习

对抗性学习的思想可以整合到基于深度学习的迁移学习方法中。如前所述,在丹框架中,网络Θ和内核κ发挥极大极小的游戏,它反映了敌对的学习。然而，DAN框架在对抗性匹配方面与传统的基于gan的方法略有不同。在DAN框架中，max博弈中只有少数参数需要优化，使得优化更容易达到均衡。在介绍对抗性迁移学习方法之前，让我们简要回顾一下GAN的原始框架和相关工作。原始的GAN[139]受到了两个人博弈的启发，由生成器G和鉴别器d两个模型组成。生成器产生真实数据的伪码，目的是混淆鉴别器，使鉴别器产生错误检测。该鉴别器将真数据和假数据混合输入，用于检测数据的真伪。这两种模型实际上是在进行两方极大极小博弈，其目标函数如下:



其中˜z表示噪音实例(从一定的噪声分布采样)作为输入的发电机生产假药。整个GAN可以通过使用反向传播算法进行训练。当两方游戏达到平衡时，生成器可以产生几乎真实的实例。

在GAN的激励下，许多迁移学习方法都建立在一个良好的特征表示几乎不包含实例原始域的区分信息的假设基础上。例如，Ganin等人的工作提出了一种深度架构，称为领域对抗神经网络(DANN)，用于领域自适应[154]和[155]。DANN假设没有标记的目标域实例可以使用。它的架构由一个特征提取器、一个标签预测器和一个领域分类器组成。如图所示。

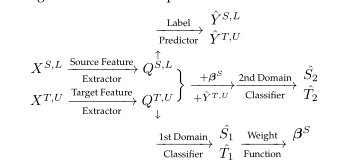


特性提取器的作用类似于生成器，其目的是生成与领域无关的特性表示，以混淆领域分类器。领域分类器扮演的角色类似于鉴别器，它尝试检测提取的特征是来自源域还是目标域。此外，标签预测器根据所提取的源域实例特征进行训练，生成实例的标签预测，即QS、L。通过插入特殊的梯度反转层(GRL)可以训练DANN。整个系统的培训后,特征提取器学习的深层特征实例,和输出ˆ欧美,ui标记目标域的标签预测实例。

还有其他一些相关的令人印象深刻的作品。Tzeng等人的工作提出了一个统一的对抗性领域适应框架[156]。Shen等的工作采用了Wasserstein distance进行领域适应[59]。Hoffman等人采用循环一致性丢失的方法来保证结构和语义的一致性[157]。Long等人提出了条件域对抗网络(CDAN)，该网络利用条件域鉴别器来辅助对抗自适应[158]。Zhang等人对源和目标分类器采用了对称设计[159]。Zhao等人利用域对抗性网络解决多源迁移学习问题[160]。Yu等人提出了一种动态对抗适应网络[161]。

有些方法是为某些特殊场景设计的。以部分迁移学习为例。部分转移学习方法是为目标域类小于sourcedomain类(即YS⊆YT)的情况而设计的。在这种情况下，具有不同标签的sourcedomain实例对域适应的重要性可能不同。更具体地说，具有相同标签的源域实例和目标域实例更可能是潜在关联的。但是，由于目标域实例是未标记的，因此如何标识和部分传输来自标记的源域实例的重要信息是一个关键问题。

Zhang等人的论文提出了一种局部域适应的方法，称为重要性加权对抗性网络域适应(IWANDA)[162]。IWANDA的建筑不同于DANN。DANN采用了一个共同的特征提取器，假设存在一个共同的特征空间，其中QS、Land QT、Uhave分布相似。但是，IWANDA分别为源域和目标域使用了两个特定于域的特征提取器。具体来说，IWANDA由两个特征提取器、两个域分类器和一个标签预测器组成。IWANDA的图表如下。



在训练之前，源特征提取器和标签预测器将在已标记的源域实例上进行预训练。这两个组件在训练过程中被冻结，这意味着只需要优化目标特征提取器和域分类器。在每次迭代中，通过以下步骤优化上述网络。

1. 实例权重:为了解决部分传递问题，根据第一个域分类器的输出给源域实例分配权重。第一个域分类器输入QS,Land QT,U，然后输出它们域的概率预测。如果一个源域实例被预测很可能属于目标域，那么这个实例很可能与目标域相关联。因此，该实例被赋予更大的权重，反之亦然。

2. 预测生成:标签预测器输出实例的标签预测。第二个分类器预测一个实例属于哪个域。

3.参数更新:优化第一个分类器，使领域分类错误最小化。第二分类器与目标特征提取器进行minmax博弈。这个分类器的目的是检测实例是否实例从目标域或加权实例从源域,并降低标签的不确定性预测ˆYT,U。目标特性提取器旨在混淆第二个分类器。这些组件可以采用与GAN类似的方式进行优化，也可以通过插入GRL进行优化。

除了IWANDA, Cao等人的工作构建了部分迁移学习的选择性对抗网络[163]。还有其他一些与迁移学习相关的研究。例如，Wang等人的工作提出了一种基于极小值的方法来选择高质量的源域数据[164]。Chen等研究了对抗性域自适应中的可转移性和可区分性，提出了一种光谱惩罚方法来增强现有对抗性迁移学习方法[165]。

# 6应用

在前面的章节中，我们介绍了一些有代表性的迁移学习方法，这些方法在他们的原始论文中已经被应用于解决各种与文本/图像相关的问题。例如，MTrick[122]和TriTL[123]利用矩阵分解技术解决跨域文本分类问题;基于深度分层的方法如DAN [137]， DCORAL [148]， DANN[154]，[155]被用于解决图像分类问题。在本节中，我们主要关注迁移学习在特定领域的应用，如医学、生物信息学、交通运输和推荐系统，而不是一般的文本相关或图像相关的应用。

## 6.1医学应用

医学影像在医学领域中起着重要的作用，是诊断的有力工具。随着机器学习等计算机技术的发展，计算机辅助诊断已成为一个热门和有前途的方向。请注意，医学图像是由特殊的医疗设备产生的，它们的标签通常依赖于有经验的医生。因此，在许多情况下，收集足够的培训数据既昂贵又困难。迁移学习技术可用于医学影像分析。一种常用的转移学习方法是在源域上对神经网络进行预训练(例如ImageNet，它是一个包含超过1400万张注释过的图像，超过2万个类别的图像数据库[166])，然后根据目标域的实例对其进行调整。

例如，Maqsood等人对AlexNet[138]进行了改进，用于检测阿尔茨海默病[167]。他们的方法有以下四个步骤。首先，对来自目标域的MRI图像进行对比度拉伸处理。其次，AlexNet架构[138]是在ImageNet[166](即源域)上预先训练的，作为学习新任务的起点。第三，修复AlexNet的卷积层，最后三个全连接层被新的softmax层、全连接层和输出层取代。最后，通过训练阿尔茨海默氏数据集[168](即目标域)对修改后的AlexNet进行微调。实验结果表明，该方法在多类分类问题(即阿尔茨海默病分期检测)中具有最高的准确率。

类似地，Shin等人改进了预训练的深度神经网络来解决计算机辅助检测问题[169]。Byra等利用转移学习技术帮助评估膝关节骨关节炎[170]。除了影像分析，迁移学习在医学领域还有其他应用。例如，Tang等人的工作将主动学习和领域适应技术结合起来对各种医疗数据进行分类[171]。Zeng等人利用迁移学习自动编码用于描述患者诊断的ICD-9代码[172]。

## 6.2生物信息学应用

生物序列分析是生物信息学领域的一项重要工作。由于对某些生物体的理解可以转移到其他生物体，迁移学习可以用于促进生物序列分析。在此应用中存在着显著的分布差异问题。例如，某些生物物质的功能可能保持不变，但两种生物之间的成分发生了变化，这可能会导致边缘分布差异。此外，如果两种生物祖先相同，但进化距离较远，则条件分布差异较大。Schweikert等以mRNA剪接位点预测问题为例分析迁移学习方法的有效性[173]。在他们的实验中，源域包含了来自一个被充分研究的模型生物体的序列实例，即秀丽隐杆线虫，而目标生物体包括另外两种线虫(即remanei和P。黄藤、黑草和植物黄藤。我们比较了许多迁移学习方法，如FAM[64]和KMM[5]的变种。实验结果表明，迁移学习有助于提高分类性能。

生物信息学领域另一个广泛遇到的任务是基因表达分析，例如，预测基因和表型之间的关联。在这个应用程序中，主要的挑战之一是数据稀疏性问题，因为已知关联的数据通常非常少。迁移学习可以通过提供额外的信息和知识来利用这一问题。例如，Petegrosso等人[174]提出了一种基于标签传播算法(LPA)分析和预测基因-表现型关联的迁移学习方法[175]。LPA利用蛋白-蛋白相互作用(PPI)网络和初始标记来预测目标关联，假设连接在PPI网络中的基因应该具有相似的标记。作者通过合并多任务和转移学习技术扩展了LPA。首先，利用人类表型本体(Human Phenotype Ontology, HPO)形成辅助任务，HPO提供了人类疾病表型特征的标准化词汇。利用表型途径以及HPO和PPI网络中的连锁知识，可以预测这些关联;PPI中相互作用的基因更有可能与相同的表型相关，而HPO中相互连接的表型更有可能与相同的基因相关。其次，以基因本体(Gene Ontology, GO)作为源域，其中包含基因功能与基因之间的关联信息。设计了更多的调控因子，并利用PPI网络和共同基因作为知识转移的桥梁。PPI网络中所有基因同时构建了基因- go术语和基因- hpo表型关联。通过传递额外的知识，预测的基因-表现型关联可能更可靠。

迁移学习也可以用于解决PPI预测问题。Xu等人[176]提出了一种将连接知识从源PPI网络转移到目标网络的方法。该方法基于集体矩阵分解技术[177]，其中一个因子矩阵是跨域共享的。

## 6.3交通运输业应用

转移学习在交通领域的一个应用是理解交通场景图像。在这个应用程序中，一个具有挑战性的问题是，从某个位置拍摄的图像经常会因为不同的天气和光照条件而发生变化。为了解决这一问题，Di等人提出了一种方法，即尝试传输在不同条件下从同一位置拍摄的图像的信息[178]。第一步，对预先训练好的网络进行微调，提取图像的特征表示。第二步，采用特征转换策略构建新的特征表示。具体来说，对提取的特征进行降维算法(偏最小二乘回归[179])，生成低维特征。然后，通过学习一个变换矩阵来最小化降维数据的域差异。其次，采用子空间对齐操作进一步减小域差值。请注意，尽管不同条件下的图像通常具有不同的外观，但它们通常具有相似的布局结构。因此，在最后一步，首先建立测试图像与被检索到的最佳匹配图像之间的跨域密集对应，然后通过马尔可夫随机场模型[180]，[181]将最佳匹配图像的标注转移到测试图像上。

转移学习也可以应用于驾驶员行为建模的任务中。在这个任务中，通常无法获得每个单独驱动程序的足够的个性化数据。在这种情况下，为新参与的驾驶员传递历史数据中包含的知识是一个很有前途的选择。例如，Lu等人提出了一种道路变化场景下驾驶员模型适应的方法[182]。源域包含描述源驱动程序行为的足够数据，而目标域有一些关于目标驱动程序的数据。第一步，对来自两个域的数据进行PCA预处理，生成低维特征。本文假设源数据和目标数据来自两个流形。因此，在第二步中，采用流形对齐方法进行域自适应。具体来说，采用动态时间翘整算法[183]来测量相似度，寻找每个目标域数据点对应的源域数据点。然后，采用局部Procrustes分析[184]，根据得到的数据点之间的对应，对两个流形进行对齐。通过这种方式，可以将源域的数据传输到目标域。在最后一步，使用随机建模方法(如高斯混合回归[185])对目标驾驶员的行为建模22。实验结果表明，在目标域数据较少的情况下，转移学习方法可以帮助目标驱动程序。此外，当目标实例数量非常小或非常大时，该方法的优越性不明显。这可能是因为在目标域实例较少的情况下无法准确地找到域间的关系，而在目标域实例充足的情况下，迁移学习的必要性降低了。

此外，迁移学习在交通领域还有其他应用。例如，Liu等人将转移学习应用到驾驶员姿态识别中[186]。Wang等人在迁移学习中采用正则化技术进行车型识别[187]。迁移学习也可以用于异常活动检测[188]、[189]、交通标志识别[190]等。

## 6.4推荐系统应用

随着信息量的快速增长，如何有效地为个人用户推荐个性化内容成为一个重要的问题。在推荐系统领域，一些传统的推荐方法，如基于因子分解的协同过滤，往往依靠对用户-物品交互矩阵的因子分解来获得预测函数。这些方法通常需要大量的训练数据才能做出准确的建议。然而，必要的训练数据，例如历史交互数据，在真实场景中往往是稀疏的。此外，对于新注册用户或新项目，传统的方法往往难以做出有效的建议，这也被称为冷启动问题。

针对这些问题，本文提出了基于实例的迁移学习方法和基于特征的迁移学习方法。这些方法试图利用来自其他推荐系统(即源域)的数据来帮助构建目标域的推荐系统。基于实例的方法主要关注于将不同类型的实例(例如，评级、反馈和检查)从源域传输到目标域。Pan等人[191]的工作利用了知识转移源域的不确定评级(表示为评级分布)。具体来说，利用源域不确定评级作为约束条件，帮助完成目标域上的评级矩阵分解任务。Hu等人[192]提出了一种transfer meeting hybrid方法，利用专注记忆网络从非结构化文本中提取知识，并有选择地传递有用信息。

基于特征的方法通常利用和转移潜在特征空间中的信息。例如，Pan等人提出了一种称为坐标系统转移(CST)的方法[193]来同时利用用户端和项目端潜在特性。源域实例来自另一个推荐系统，与目标域共享公共用户和项。基于反映用户偏好或项目因素的主坐标，描述了独立于领域的结构，并且可以跨领域转移的假设，开发了CST。CST首先在源域数据上应用稀疏矩阵三因子分解，构造了两个主要的坐标系统，这两个坐标系统实际上是用户和项目的潜在特征，然后通过设置约束将坐标系统转移到目标域。实验结果表明，CST在所有数据稀疏度水平上都显著优于非转移基线(即平均填充模型和潜在因子分解模型)[193]。

还有一些关于跨域推荐的研究[194]，[195]，[196]，[197]。例如，He等人提出了基于贝叶斯神经网络的迁移学习框架[198]。Zhu等人[199]提出了一个深度框架，该框架首先基于矩阵分解技术生成用户和项目特征表示，然后使用深度神经网络学习跨域特征映射。Yuan等人[200]提出了一种基于自编码器和改进的DANN的深度域自适应方法[154]和[155]，用于从评级矩阵中提取和转移实例。

## 6.5 其他方面应用

**通信应用:**除了WiFi定位任务[2]、[36]外，转移学习也被应用于无线网络应用中。例如，Bastug等人提出了缓存机制[201];从设备之间的相互作用中提取出的上下文信息中包含的知识被传输到目标域。此外，一些研究集中在节能问题上。Li等人的工作提出了一种利用传输学习技术的蜂窝无线接入网络的节能方案[202]。赵和Grace的工作将转移学习应用到拓扑管理中，以降低能耗[203]。

**城市计算应用**:城市计算拥有大量与城市相关的数据，在交通监测、医疗卫生、社会保障等方面具有广阔的研究前景。在许多城市计算应用中，迁移学习已被用于缓解数据稀缺问题。例如，Guo等人[204]提出了一种连锁商店网站推荐的方法，该方法利用语义相关领域的知识(如在目标城市拥有同一家商店的其他城市和其他连锁商店)向目标城市推荐。Wei等人[205]提出了一种灵活的多模式转移学习方法，将知识从拥有足够多模型数据和标签的城市转移到目标城市，以缓解数据稀疏问题。

迁移学习已被应用于一些识别任务，如手势识别[206]、人脸识别[207]、活动识别[208]和语音情绪识别[209]。此外，转移学习的专业知识也被纳入其他一些领域，如情绪分析[28]，[96]，[210]，欺诈检测[211]，社交网络[212]，高光谱图像分析[54]，[213]。

# 7实验

迁移学习技术已成功地应用于许多实际应用中。在本节中，我们进行了23个实验来评估不同类别的一些代表性迁移学习模型s1[214]在两个主流研究领域，即物体识别和文本分类上的表现。首先介绍数据集。并给出了实验结果和进一步的分析。

## 7.1数据集和预处理

实验研究了三个数据集，分别是Office31、Reuters-21578和Amazon Reviews。为了简单起见，我们主要关注分类任务。预处理数据集的统计信息如表3所示。

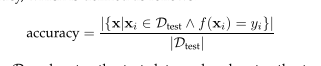
•Amazon Reviews2[107]是一个多领域的情感数据集，包含了从亚马逊网站四个领域(书籍、厨房、电子产品和dvd)获得的产品评论。这四个域中的每个评论都有一个文本和一个从0到5的评级。在实验中，小于3的比值被定义为负比值，其余的比值被定义为正比值。计算所有评论中每个单词的频率。然后，选择频率最高的5000个单词作为每个评论的属性。通过这种方式，我们最终在每个域中拥有1000个正实例、1000个负实例和大约5000个未标记实例。在实验中，四个领域中的每两个被选择产生十二个任务。

•reuters -215783是一个文本分类数据集，具有层次结构。数据集包含5个顶级类别(交换、组织、人物、地点、主题)。在我们的实验中，我们使用前三大类别Orgs, People和Places来生成三个分类任务(Orgs vs People, Orgs vs Places, People vs Places)。在每个任务中，对应两类中的子类分别被分为两部分。然后，生成的四个部分被用作组成两个域的组件。每个域大约有1000个实例，每个实例大约有4500个特性。具体来说，以任务Orgs vs People为例，一部分来自Orgs，一部分来自People，组合形成源域;类似地，其余两部分构成目标域。注意，这三个类别中的实例都被标记了。为了生成未标记的实例，将从数据集中选择标记的实例，并忽略它们的标签。

•Office-31[215]是一个对象识别数据集，包含31个类别和三个域，即Amazon, Webcam和DSLR。这三个域分别有2817、498和795个实例。亚马逊网站上的图片是在亚马逊网站上拍摄的在线电子商务图片。网络摄像头拍摄的图像是网络摄像头拍摄的低分辨率图像。DSLR中的图像是由DSLR相机拍摄的高分辨率图像。在实验中，每两个域(考虑顺序)被选择作为源域和目标域，从而产生六个任务。

## 7.2实验设置

通过实验比较了几种有代表性的迁移学习模型。具体来说，我们在数据集Office-31上执行了8种算法来解决目标识别问题。此外，我们在数据集Reuters-21578上执行和评估了14种算法来解决文本分类问题。在情绪分类问题中，对Amazon评论执行了11种算法。分类结果采用准确率进行评价，准确率的定义如下:



式中，dtest为测试数据，y为真值分类标签;f(x)表示预测的分类结果。注意，有些算法需要基分类器。在这种情况下，实验采用带线性核的SVM作为基分类器。此外，sourcedomain实例都被标记了。对于执行的算法(TrAdaBoost除外)，目标域实例没有标记。每个算法执行三次，取平均结果作为实验结果。

## 7.3实验结果

在本小节中，我们在三个数据集上比较超过二十种算法。所有算法的参数都设置为原始论文中提到的默认值或推荐值。实验结果分别对应于Amazon Reviews、Reuters-21578和Office-31的表4、表5和表6。为了让读者更直观地理解实验结果，本文提供了三幅雷达图，即图5、图6、图7，将实验结果可视化。在雷达图中，每个方向代表一个任务。一个算法的一般性能由一个多边形表示，该多边形的顶点表示该算法的精度用于处理不同的任务。

表4展示了Amazon Reviews的实验结果。基线是只在源域上训练的线性分类器(这里我们直接使用论文[107]的结果)。图5显示了结果。如图5所示，当源域是电子或厨房时，大多数算法的性能相对较好，这说明这两个域可能比其他两个域包含更多的可转移信息。此外，可以观察到，HIDC、SCL、SFA、MTrick和SDA在这12个任务中都表现良好且相对稳定。同时，其他算法，特别是mSLDA、CD-PLSA、TriTL等算法相对不稳定;他们的表现在大约20%的范围内波动。TriTL对于源域为厨房的任务具有较高的准确性，但对于其他任务的准确性相对较低。算法TCA、mSLDA和CD-PLSA在所有任务上具有相似的性能，平均准确率约为70%。在性能良好的算法中，HIDC和MTrick基于特征缩减(特征聚类)，而其他算法基于特征编码(SDA)、特征对齐(SFA)和特征选择(SCL)。这些策略是目前基于特征迁移学习的主流。

表5给出了reuet -21578的比较结果(这里我们直接使用了文献[78]和[41]中的基线和CoCC结果)。基线是一个正则化的最小二乘回归模型，只对标记的目标域实例进行训练[78]。图6与图5具有相同的结构，将性能可视化。为了清晰起见，十三种算法被分为两部分，分别对应图6中的两个子图。可以看出，对于大多数Orgs vs Places和Orgs vs People，该算法的效果相对较好，而对于People vs Places，则效果较差。 这个现象术语表示人与人之间的差异地方可能比较大。TrAdaBoost在本实验中性能比较好，因为它使用了目标域内实例的标签来减少分布差异的影响。另外，HIDC、SFA和MTrick算法在这三个任务中表现出了相对一致的性能。在之前的Amazon Reviews实验中，这些算法的表现也很好。此外，在人员vs位置方面表现最好的两种算法是CoCC和TrAdaBoost。

在第三个实验中,七个deep-learning-based转移学习模型(例如,丹,DCORAL, MRAN, CDAN,丹,简,和可以)和基线(即Alexnet [138], [140] pre-trained在ImageNet[166],然后直接对准目标域)的数据集上执行Office-31(这里我们直接使用CDAN的结果,1月,可以和基线从原始论文[137],[147],[151],[158])。ResNet-50[144]被用作这三种模型的骨干网。实验结果如表6所示，平均性能如图7所示。如图7所示，这七种算法都有很好的性能，特别是在任务D→W和W→D上，准确率非常接近100%。这一现象体现了基于深度学习的方法的优越性，也与Webcam和DSLR的区别小于Webcam/DSLR和Amazon的区别是一致的。显然，可以超过其他六种算法。在所有的6个任务中，DANN的性能与DAN相似，优于DCORAL，说明了引入对抗性学习的有效性和实用性。

值得一提的是，在上面的实验中，一些算法的性能并不理想。一个原因是我们使用了算法原始论文中提供的默认参数设置，这可能不适合我们选择的数据集。例如，GFK最初是为物体识别而设计的，在第一次实验中我们直接将其用于文本分类，但结果并不理想(平均准确率约为62%)。以上实验结果仅供参考。这些结果表明，某些算法可能不适用于特定领域的数据集。因此，在研究过程中选择合适的算法作为基线是很重要的。此外，在实际应用中，寻找合适的算法也是很有必要的。

# 8结论和未来方向

本文从数据和模型的角度对迁移学习的机制和策略进行了总结。本调查对迁移学习给出了明确的定义，并设法用一个统一的符号系统来描述大量有代表性的迁移学习方法和相关工作。我们基本介绍了基于数据解释和基于模型解释的迁移学习的目标和策略。基于数据的解释从数据的角度介绍迁移学习的目标、策略和一些迁移学习方法。同样，基于模型的解释从模型层面介绍迁移学习的机制和策略。并介绍了迁移学习的应用。最后，通过实验对有代表性的迁移学习模型在目标识别和文本分类两个主流领域的表现进行了评价。并对模型进行了比较，说明迁移学习模型的选择是一个重要的研究课题，也是一个复杂的实际应用问题。

在迁移学习领域的未来研究有几个方向可供选择。首先，迁移学习技术可以进一步探索并应用于更广泛的应用。在更复杂的情况下，需要新的方法来解决知识转移问题。例如，在现实场景中，有时与用户相关的源域数据来自另一家公司。在这种情况下，如何在保护用户隐私的同时传输源域中包含的知识是一个重要的问题。其次，如何度量域间的可转移性，避免负转移也是一个重要的问题。虽然有一些关于负迁移的研究，但[3]负迁移还需要进一步的系统分析。第三，迁移学习的可解释性也需要进一步研究[216]。最后，可以进一步进行理论研究，为迁移学习的有效性和适用性提供理论支持。在机器学习中，迁移学习是一个很有发展前景的领域，与传统机器学习相比，迁移学习具有较少的数据依赖性和标签依赖性等优点。希望我们的工作能帮助读者更好地了解研究现状和研究思路。