迁移学习问题与方法研究

（申请清华大学工学博士学位论文）

培 养 单 位：计算机科学与技术系
学　　科：计算机科学与技术
研　究　生：龙 明 盛
指 导 教 师：王 建 民 教 授

二〇一四年六月

Transfer Learning: Problems and Methods

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Long Mingsheng

Dissertation Supervisor : Professor Wang Jianmin

June, 2014

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

随着数据规模和计算资源的快速增长，机器学习在理论和实践两方面都取得了长足进展。传统机器学习方法通常依赖于数据的生成机制不随环境改变这一基本假设。然而在机器学习的各种应用领域中，如大数据分析、自然语言处理、计算机视觉、生物信息学等，上述假设往往因为过于严格而难以成立。如何分析和挖掘非平稳环境中的大规模数据是现代机器学习最具有挑战性的前沿方向之一。迁移学习放宽了传统机器学习中训练数据和测试数据必须服从独立同分布的约束，因而能够在彼此不同但又相互关联的两个领域间挖掘领域不变的本质特征和结构，使得标注数据等有监督信息可以在领域间实现迁移和复用。迁移学习是解决目标任务标注数据稀缺的基础方法，其研究仍处于富有挑战的阶段。本文面向跨领域非结构化数据的分类和预测任务，系统地研究迁移学习的问题挑战及其解决方法。

迁移学习中，过拟合、欠拟合、欠适配、负迁移等关键问题与挑战交错叠加。首先，在拟合观测数据所服从的未知概率分布时存在模型的过拟合或欠拟合问题；其次，在领域间适配不同的未知概率分布时存在模型的欠适配或负迁移问题：欠适配是指跨领域的概率分布失配问题未能充分修正，负迁移是指辅助领域任务对目标领域任务有负面效果。本文重点面向欠拟合、欠适配、负迁移等问题挑战，分析原因并设计针对性的学习方法，主要创新点包括：

1. 针对负迁移问题，提出一种图正则化联合矩阵分解模型，来构建跨领域间知识迁移的语义特征、提高特征结构的迁移能力、并避免特征结构的负面效果；该模型综合两类主流方法的优势，有效地克服了欠迁移与过迁移权衡的两难困境。

2. 针对欠适配问题，提出一种联合适配正则化学习框架，扩展最大均值差异准则用于度量联合概率分布距离；通过特征学习和监督学习使得联合概率分布在领域间适配，提出基于线性回归、支持向量机、主成份分析的三种迁移学习方法，并基于统计学习理论分析它们的泛化误差上界；提出基于核矩阵低秩近似误差的概率分布度量新准则来充分适配领域间概率分布，并从理论上分析近似误差上界。

3. 针对欠拟合、欠适配与负迁移问题，基于深度学习扩展最大均值差异准则为非线性分布差异准则，提出统一的鲁棒深度表征适配模型来协同解决上述问题；提出迁移交叉验证方法，解决目标领域无标注数据的无监督迁移学习的模型选择。

关键词：迁移学习；领域泛化；异构数据分析；概率分布适配；隐含表征学习

Abstract

The ubiquitous growth of data volumes and computing resources has accelerated machine learning with rapid advances in both theory and practice. The success of traditional machine learning algorithms often relies on the fundamental assumption that observed data is under a stationary generating mechanism. However, in a wide range of machine learning application domains, including big data analytics, natural language processing, computer vision, and bioinformatics, this assumption may be too restricted to be satisfied. Therefore, analyzing and mining the massive data under non-stationary environments is among the greatest challenges of modern machine learning. Transfer learning relaxes the assumption of traditional machine learning that the training data and testing data should be sampled independently from an identical probability distribution, thus it can be applied to discover domain-invariant intrinsic features and structures underlying two different but related domains, which establishes successful transfer and reutilization of supervised information across domains. As one of the basic tools for addressing the learning task which may fail with scarcity of labeled data, transfer learning remains an open paradigm with several unsolved challenges. To boost cross-domain classification and prediction tasks, this thesis presents a systematic study on the open issues and solutions of transfer learning.

Transfer learning involves several critical issues and challenges: overfitting, underfitting, under-adaptation, and negative-transfer. Overfitting and underfitting may happen when modeling the unknown probability distribution based on observed data; Under-adaptation and negative-transfer may happen when adapting the unknown probability distributions across domains: under-adaptation refers to the condition that the distribution mismatch cannot be corrected sufficiently; negative-transfer refers to the condition that the auxiliary task deteriorates the target task unintentionally. This thesis addresses the underfitting, under-adaptation, and negative-transfer issues, analyzes the intrinsic causes, and designs specific learning models. The novel contributions are summarized as follows.

1. For addressing the negative-transfer problem, a graph regularized collective matrix factorization model is proposed to 1) construct semantic features for cross-domain knowledge transfer, 2) enhance the transferability of semantic features, and 3) combat the negative effects of semantic features. This model synthesizes the advantages of two mainstream methods to establish effective tradeoff between under-transfer and over-transfer.

2. For addressing the under-adaptation problem, a joint adaptation regularization framework is proposed, which extends the maximum mean discrepancy to measure the divergence between different joint probability distributions; Both feature learning and classifier learning are explored to adapt the mismatched joint probability distributions across domains, from which three transfer learning methods based on linear regression, support vector machines, and principal component analysis are formulated; The generalization error bound of these methods are theoretically analyzed via the statistical learning theory. Furthermore, a novel criterion based on the low-rank approximation error of kernel matrix is proposed for comparing different probability distributions and adapting them sufficiently across domains, with theoretical analysis on the approximation error bound.

3. For addressing the underfitting, under-adaptation, and negative-transfer problems, deep learning is explored to extend the maximum mean discrepancy to nonlinear distribution discrepancy, and a unified robust deep representation adaptation model is developed to tackle the three problems collaboratively. Finally, a transfer cross-validation strategy is proposed for model selection of unsupervised transfer learning without target labels.

Key words: Transfer Learning; Domain Generalization; Heterogeneous Data Analytics; Probability Distribution Adaptation; Latent Representation Learning

目 录

第 1 章 绪论

1.1 研究背景与意义

随着数据规模和计算资源的快速增长，机器学习在理论和实践两方面都取得了长足进展，成为大数据分析的主要技术基石之一。这篇博士论文工作即来源于国家核高基科技重大专项“非结构化数据管理系统”。在进行文本、图像、视频和图谱等非结构化数据的大规模统一存储、检索、分析和挖掘过程中，一个关键性需求是对不同领域、不同来源的非结构化数据进行数据迁移和知识复用。这是因为“大数据”和“小数据”总是矛盾统一的，有大领域就有小领域，这就要求人们既要能分析大数据，也要能分析小数据，但这并不总是容易的。例如，在互联网领域存在海量高质量图像、视频数据、并且包括丰富的标注信息，容易进行监督学习和深度分析；而在众多非互联网领域（如安全监控、人物追踪）仅能获得普通规模的低质量视频、并且几乎没有标注信息，现有机器学习技术难以获得较好效果。这时候大数据的价值就体现出来了：将大数据“迁移”到小数据领域，解决小数据领域中数据稀缺、知识稀缺问题。事实上，解决小数据领域中的问题挑战对服务国计民生可能更重要，因为这些领域数据收集代价更高。而迁移学习正是利用大数据解决小数据问题的核心技术，这也是本文的研究背景和现实意义。

迁移学习（Transfer Learning）[1]，或称归纳迁移、领域适配，是机器学习中的一个重要研究问题，其目标是将某个领域或任务上学习到的知识或模式应用到不同的但相关的领域或问题中。迁移学习试图实现人通过类比进行学习的能力，例如学习走路的技能可以用来学习跑步、学习识别轿车的经验可以用来识别卡车等。

与半监督学习、主动学习等标注数据稀缺解决风范的本质不同在于，迁移学习放宽了训练数据和测试数据服从独立同分布这一假设，从而使得参与学习的领域或任务可以服从不同的边缘概率分布或条件概率分布。迁移学习的主要思想是，从相关的辅助领域中迁移标注数据或知识结构、完成或改进目标领域或任务的学习效果，其工作原理如图 1.1 所示。在很多工程实践中，为每个应用领域收集充分的标注数据代价十分昂贵、甚至是不可能的，因此从辅助领域或任务中迁移现有的知识结构从而完成或改进目标领域任务是十分必要的、是源于实践需求的重要研究问题。迁移学习可以认为是在最小人工监督代价下进行机器学习的一种崭新策略。在自然语言处理、计算机视觉、医疗健康与生物信息学等领域，目标任务的标注数据稀缺、领域分布异构等问题十分突出，迁移学习具有很强的现实需求。

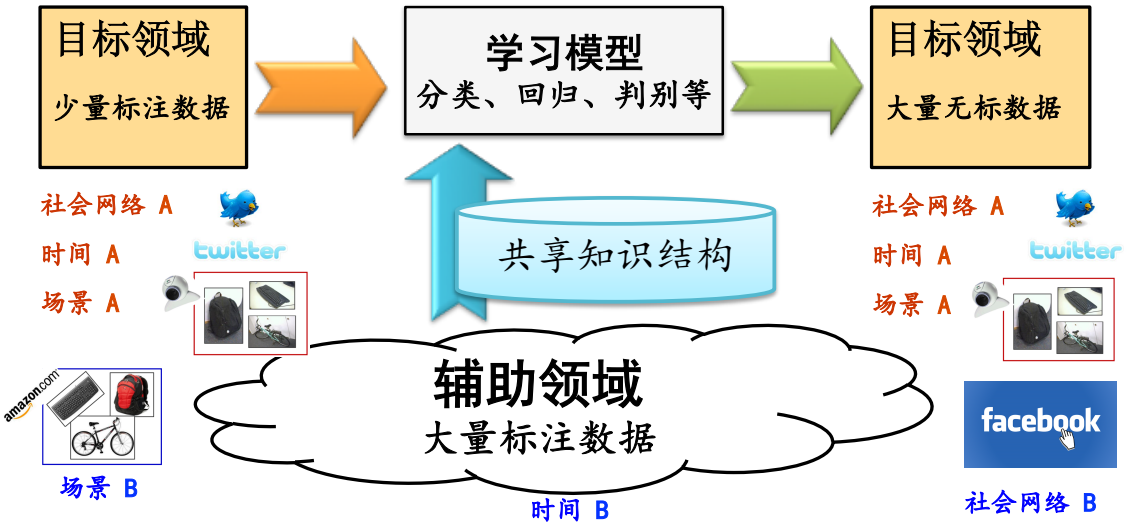


图 1.1　迁移学习原理：从辅助领域的大量标注数据中迁移知识、改进目标领域学习任务。

1.1.1　理论研究价值

迁移学习是机器学习的前沿研究方向之一，具有充分的理论和应用研究价值。

1.1.1.1　解决标注数据稀缺性

随着“大数据”时代的来临，亿万级别规模的数据导致数据的统计异构性、标注缺失性问题愈加凸显，如何在不同数据领域间进行知识迁移和模型适配将变得不可避免。标注数据稀缺性会导致经典监督学习出现严重的过拟合问题，虽然传统半监督学习、协同训练、主动学习等也可以解决数据稀缺性，但它们都要求目标领域中存在相当程度的标注数据；当标注数据十分稀缺且获取代价太大时，仍然需要从辅助领域迁移知识来提高目标领域学习效果。由于迁移学习和半监督学习、主动学习具有类似的学习目标，也可将它们结合起来解决标注稀缺性问题。

1.1.1.2　非平稳泛化误差分析

经典统计学习理论[2] 和 PAC 可学习理论[3] 给出了独立同分布条件下学习模型的泛化误差上界保证。具备理论保证是统计机器学习得以成功的关键因素之一。然而在非平稳环境中，不同数据领域不再服从独立同分布假设，使得经典学习理论不再成立，这给异构数据分析挖掘带来了理论上的风险。例如，迁移学习中存在极具挑战性的负迁移问题[1]，即难以判定迁移学习模型在什么条件下会导致性能下降而非提升。广义上看，迁移学习是经典学习在非平稳环境下的推广；换句话说，但凡经典学习不能取得很好学习效果时均可能是因为训练数据和测试数据之间存在概率分布漂移。因此，研究迁移学习是对经典学习的一个重要理论补充。

1.1.2 应用研究价值

基于上述理论方面的原因，迁移学习已经被广泛应用于很多重要实际问题中，下面就从自然语言处理、计算机视觉、医疗健康和生物信息学等方面做简要介绍。

1.1.2.1 自然语言处理

在自然语言处理、文本检索与挖掘中，迁移学习可以有很多应用，例如从 Wikipedia 长文本迁移知识到 Twitter 短文本、从 WWW 网页迁移知识到 Flickr 图像、搜索引擎中从英文文档迁移知识到中文文档，等等。迁移学习用于自然语言处理的一个根本动机是：待处理的目标领域训练数据稀缺，包括标注稀缺和内容稀缺。例如，Twitter 消息多是用户生成的无标注短文本，同时存在标注稀缺和内容稀缺，现有机器学习方法很难对这些消息进行分类管理；利用迁移学习技术，能够从 FaceBook 等长文本语料中迁移标注和内容知识助益短文本消息分类管理。

1.1.2.2 计算机视觉

计算机视觉中常常遇到用于模型训练的标注数据（辅助领域）与用于模型预测的无标数据（目标领域）具有截然不同的数据属性和统计分布，这是因为视觉场景中通常具有可变的甚至不可控的光照、朝向、遮挡、模糊等条件。典型应用包括：（1）识别手机拍摄图片中的对象（如即拍即搜式商品识别），由于收集手机图片标注数据代价很高，如何从已标注好的公共数据源（Amazon，PASCAL VOC 等）训练识别模型、并迁移到手机图片识别上？（2）对个人多媒体库（照片、视频等）进行自动分类管理、但为了保证用户体验不能过多要求用户手动标注（如 Google Picasa），如何使用 Flickr 和 YouTube 等公共数据源上的大量弱标注数据训练准确的分类模型、并迁移到个人媒体任务上？（3）如何在不同视觉场景、领域之间迁移和复用图像的形状、轮廓、图形、随机场、视觉动态等知识结构？等等。

1.1.2.3 医疗健康和生物信息学

在医疗健康和生物信息学领域，对数据进行标注代价极高（需要专业医师或生物学家给出），因而标注数据十分宝贵且十分稀缺，如何复用这些宝贵的标注数据是迁移学习的一个极佳研究问题。典型应用问题包括：（1）在如图 1.2 的医疗影像中，如何使用 CT 或 X 射线的标注图像训练模型来检测或诊断 MRI 核磁共振成像中的异常区域？如何综合利用各种多源医疗影像给出全面诊断？（2）在生物信息学中，如何利用某一种性状的 DNA 序列预测另一种性状的 DNA 序列？等等。

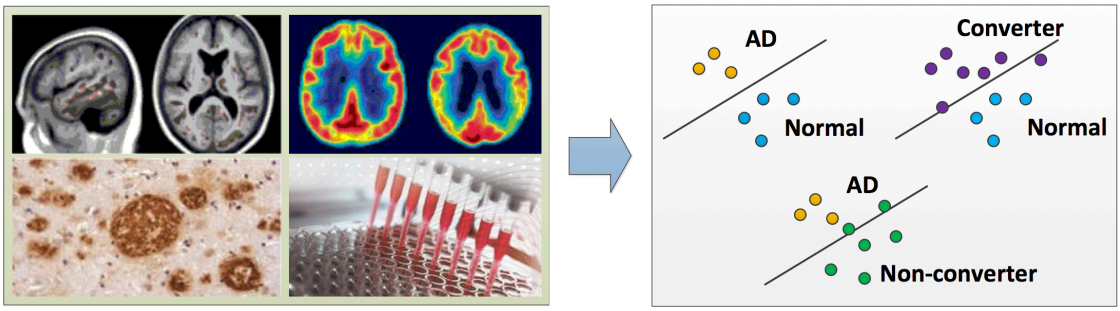


图 1.2　跨仪器设备医疗影像识别[4]：从 CT、X 射线标注图像到 MRI 核磁共振图像识别。

1.2　问题描述

迁移学习涉及领域（Domain）和任务（Task）两个重要概念，分别描述如下。

领域 $\mathcal{D}$ 定义为由 d 维特征空间 $\mathcal{X}$ 和边缘概率分布 $P(\mathbf{x})$ 组成，即 $\mathcal{D}=\{\mathcal{X}, P(\mathbf{x})\}$，$\mathbf{x}\in\mathcal{X}$。给定领域 $\mathcal{D}$，任务 $\mathcal{T}$ 定义为由类别空间 $\mathcal{Y}$ 和预测模型 $f(\mathbf{x})$ 组成，即 $\mathcal{T}=\{\mathcal{Y}, f(\mathbf{x})\}$，$y\in\mathcal{Y}$，按统计观点预测模型 $f(\mathbf{x})=P(y|\mathbf{x})$ 解释为条件概率分布。

定义 1.1 (**迁移学习**)：[1] 给定标注的辅助领域 $\mathcal{D}_s=\left\{(\mathbf{x}_1^{(s)}, y_1^{(s)}),\ldots,(\mathbf{x}_{n_s}^{(s)}, y_{n_s}^{(s)})\right\}$ 和学习任务 $\mathcal{T}_s$，无标的目标领域 $\mathcal{D}_t=\left\{\mathbf{x}_1^{(t)},\ldots,\mathbf{x}_{n_t}^{(t)}\right\}$ 和学习任务 $\mathcal{T}_t$，迁移学习的目标是在 $\mathcal{D}_s\neq\mathcal{D}_t$ 或 $\mathcal{T}_s\neq\mathcal{T}_t$ 条件下，降低目标领域预测模型 $f_t(\mathbf{x})$ 的泛化误差。

根据特征空间、类别空间、边缘概率分布、条件概率分布的异同，迁移学习可以进一步分为多个子类，在下一节详述。本文仅考察领域间特征空间与类别空间相同、边缘分布与条件分布不同的同构迁移学习，包括数据集偏移和领域适配。

为了理解迁移学习的问题场景，图 1.3 给出了迁移学习在跨领域情感分类的例子。支持向量机被证明在同领域内进行产品评论文档的情感极性（正面、负面）自动判别时能够获得很好的效果。然而，情感数据通常是领域依赖的，即不同领域的不同用户倾向于用不同的情感词来表达不同的态度。在如图 1.3 所示例子中，包含两个不同的产品领域：*books* 图书领域和 *furniture* 家具领域；在图书领域，通常用“broad”、“quality fiction”等词汇来表达正面情感，而在家具领域中却由“sharp”、“light weight”等词汇来表达正面情感。可见此任务中，不同领域的不同情感词多数不发生重叠、存在领域独享词、且词汇在不同领域出现的频率显著不同，因此会导致领域间的概率分布失配问题。为了直观地理解这一现象，用主成份分析把上万维的文本向量维度规约至二维空间、然后进行可视化。由图中可以观察到，即便在低维空间中不同领域间的概率分布差异依然很大，那么在高维空间中这种差异性一般会更大。在此情况下，辅助领域训练的情感分类器对目标领域预测效果很不理想。本文主要研究迁移学习方法来解决上述概率分布失配问题。

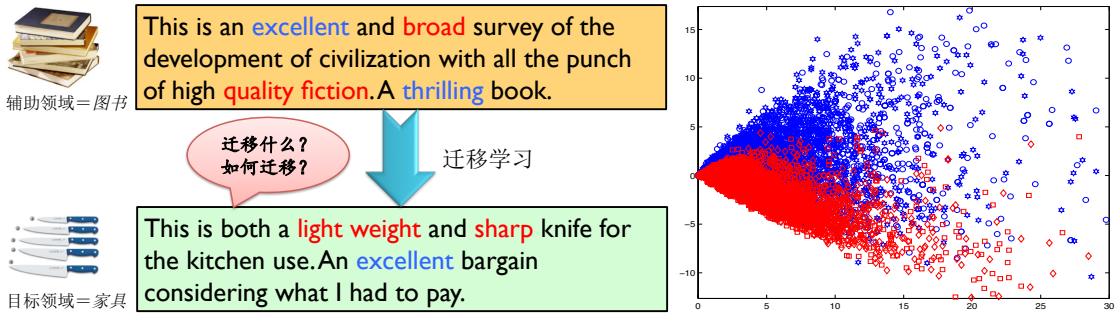


图 1.3　跨领域情感分类中的概率分布失配（不一致）问题。由于该文本数据集高到几万维，采用主成份分析将数据降至二维后进行可视化（蓝色为辅助领域、红色为目标领域）。

1.3　国内外研究现状

迁移学习是较大的研究领域，广义的迁移学习涉及多种学习框架，如多任务学习、领域适配、方差偏移、样本选择偏置、概念漂移、鲁棒学习，而狭义的迁移学习包括数据集偏移、领域适配和多任务学习[5]，本文主要研究数据集偏移和领域适配，如图 1.4 所示。迁移学习相关综述参见文献[1]，相关书籍参见文献[6]。

1.3.1　迁移学习类型

本节对迁移学习类型体系及其特点方法做一个详细回顾。按特征空间、类别空间、边缘分布、条件分布等问题因素在领域间的异同，迁移学习可大致地划分为图 1.4 所示类型体系。本文主要研究同构迁移学习中的数据集偏移和领域适配。

1.3.1.1　异构迁移学习

根据领域间特征空间和类别空间的异同，异构迁移学习包括异构特征空间和异构类别空间两种子类型，如图 1.4 所示。

异构特征空间： 首先介绍辅助领域和目标领域位于不同特征空间 $\mathcal{X}_s \neq \mathcal{X}_t$ 的异构迁移学习。典型的应用是跨语言文本分类和检索，其中训练数据和测试数据来自不同语言类型。文献[7–10] 首先采用自动翻译将辅助领域语言翻译到目标领域语言、然后再处理统一语言的概率分布失配问题，从而将异构迁移学习转化为同构迁移学习。文献[11] 同时从不同语言的训练文档（来自互译平行语料、机器翻译或其他方法）中学习语言相关的特征投影、从而将不同特征空间映射到同一个“语言无关的”抽象空间，然后利用典型相关分析得到不同语言间的关联关系。在学习到语言无关的抽象空间后，不同语言的任意文档都可以映射到公共子空间。

文献[12] 提出了翻译学习方法，可以使用一个特征空间上的训练数据建立监督模型、并对另一个特征空间上的测试数据进行预测。考察了两种任务：利用文本

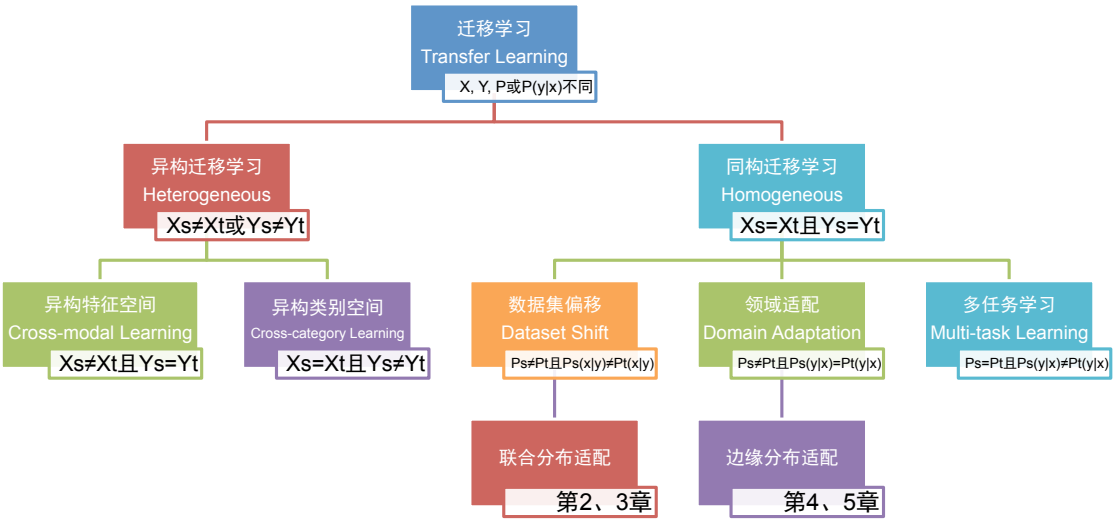


图 1.4　迁移类型分类体系。按照特征空间、类别空间、边缘概率分布、条件概率分布进行分类，比综述[1] 按照各领域是否存在标注数据的划分方法更突出不同类型的主要挑战。

帮助分类图像、跨语言文本分类。该方法依赖于不同特征空间之间的概率翻译模型，可以由双语字典或多模数据推导而得。在文献[10] 中，给定标签时词翻译条件概率由双语字典、辅助领域数据、目标领域数据的期望最大化算法协同学习得到。文献[13,14] 提出了文本到图像间的知识迁移方法。文献[15] 提出了基于特征对齐、扩充和支持向量机的通用异构迁移学习方法，在跨语言、跨媒体任务上效果良好。

异构类别空间： 辅助领域和目标领域的类别空间不一致 $\mathcal{Y}_s \neq \mathcal{Y}_t$ 场景在文本挖掘和图像理解中都受到广泛关注。文献[16] 提出了风险敏感性谱学习（Risk-sensitive Spectral Partition，RSP）算法来解决标签失配问题，文献[17] 提出了基于互信息的多任务学习来学习标签对应关系，文献[18] 提出了视觉图像的跨类别知识迁移方法，文献[19] 提出了自动辅助领域选择方法。文献[20] 提出 One-Shot Learning，仅使用非常少的训练样例就可以学习一个新类别，其基础则是旧类别到新类别的知识迁移能力，该问题已经成为计算机视觉中新类别对象识别的主要挑战之一。

在异构特征空间进行迁移学习，通常必须依赖领域特定的先验知识，包括特征空间之间的关联关系（如双语词典）、多模数据每个视图之间的对应关系（如网页中的文本和图像）、或社交关联关系（如文本和图像的情感评价来自同一个用户），等等。没有这类先验知识，则难以进行异构迁移学习。为了与不依赖于丰富先验知识的经典机器学习场景一致，本文主要研究同构迁移学习中的问题和方法。

1.3.1.2　同构迁移学习

根据边缘概率分布和条件概率分布异同，同构迁移学习可进一步分为数据集偏移、领域适配、多任务学习三种子类型，如图 1.4 所示，本文主要研究前两者。

领域间的边缘概率分布和条件概率分布都不相同即 $P_s(\mathbf{x}) \neq P_t(\mathbf{x})$ 且 $P_s(y|\mathbf{x}) \neq P_t(y|\mathbf{x})$ 的同构迁移学习称为*数据集偏移*（Dataset Shift）[6]，这是较难的迁移学习场景，已有研究工作很少且主要基于实例权重法。文献[21,22] 通过实例权重法同时修正边缘分布和条件分布差异，文献[23,24] 通过特征表示法同时减少边缘分布和条件分布差异。这些方法通常要求目标领域存在少量标注数据，这限制了迁移学习的应用范畴。本文旨在目标领域没有标注数据的场景下进一步解决数据集偏移问题。

满足领域间边缘概率分布不同即 $P_s(\mathbf{x}) \neq P_t(\mathbf{x})$ 的同构迁移学习称为*领域适配*（Domain Adaptation），包括*样本选择偏置*（Sample Selection Bias）[25,26] 和*方差偏移*（Covariate Shift）[27–29] 等，是迁移学习中研究得最为充分的问题。下文综述的迁移方法主要针对该问题，研究如何通过实例权重法或特征表示法减小边缘分布差异。

满足领域间条件概率分布不同即 $P_s(y|\mathbf{x}) \neq P_t(y|\mathbf{x})$ 的同构迁移学习称为*多任务学习*（Multi-task Learning），它通过同时学习多个任务、挖掘公共知识结构，完成知识在多个任务间的共享和迁移。多任务学习是与领域适配相对应的另一个迁移学习主流分支[17,30]，研究已近二十载。多任务学习侧重学习算法在所有领域任务上的综合性能，而领域适配则侧重目标领域任务上的学习性能。此外，多任务学习要求每个任务都存在部分标注数据，而领域适配仅要求辅助领域存在标注数据。

1.3.2 迁移学习方法

本节对统计机器学习领域与迁移学习相关的已有工作进行综述，主要侧重*无监督迁移学习*即目标领域没有标注数据的迁移学习任务。涉及两类主流方法：实例权重法和特征表示法。*实例权重法*对辅助领域中的实例进行权重调整、提升位于目标领域高密度区域的辅助领域实例权重，从而更好地与目标领域数据分布匹配；*特征表示法*试图找到原始数据的新特征表示，使得辅助领域和目标领域的数据分布更加相似、或使得领域相关的具体特征可以被领域无关的抽象特征所表示。

1.3.2.1 实例权重法

在机器学习中，通常给定目标领域和辅助领域数据集 $\mathcal{D}_t$ 和 $\mathcal{D}_s$，但生成它们的概率分布 $P_t(\mathbf{x})$ 和 $P_s(\mathbf{x})$ 却是未知的；另外，典型的机器学习问题中输入向量 $\mathbf{x}$ 维度很高，使得直接进行概率密度估计并不可行。为此，大量工作[22,25–27,31,32] 着力于如何估算目标领域和辅助领域的*概率密度比值*，即*实例权重* $w(\mathbf{x}) = \frac{P_t(\mathbf{x})}{P_s(\mathbf{x})}$。这些方法的基本假设是：$\frac{P_t(\mathbf{x})}{P_s(\mathbf{x})} < \infty$ 即 $P_t(\mathbf{x})$ 能为 $P_s(\mathbf{x})$ 所支撑、$P_s(y|\mathbf{x})$=$P_t(y|\mathbf{x})$ 即条件分布相同。文献[27] 提出*核均值匹配*（Kernel Mean Matching，KMM）方法来估计实例权重，目标是使得加权后辅助领域数据 $\mathcal{D}_s$ 和目标领域数据 $\mathcal{D}_t$ 在概率分

布上更相似；分布相似性度量为加权辅助数据与目标数据在可再生希尔伯特空间（RKHS）的均值差异，该统计量称为最大均值差异（Maximum Mean Discrepancy，MMD）[33]；由实例权重可推导得到加权支持向量机和正则化线性回归模型，并在部分实际数据上取得较好效果。类似计算实例权重的方法还有 *KL* 重要性估计过程（Kullback-Leibler Importance Estimation Procedure，KLIEP）[28]，其目标同样是估计实例权重使得加权辅助数据与目标数据的分布相似度提升，不过这里采用相对熵 $\mathrm{KL}(P_t(\mathbf{x})\|w(\mathbf{x})P_s(\mathbf{x}))$ 来度量概率分布差异；计算实例权重归结为估计一个混合基函数模型，其实验效果要好于核均值匹配[27] 以及核密度估计。文献[32] 提出基于 AdaBoost 的实例迁移学习方法，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。

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[34]，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，是本文研究的重点。

1.3.2.2 特征表示法

另一类主流迁移学习方法是特征表示法，它通过学习新的特征表示 $\phi(\mathbf{x})$，使得领域间共享特性增强而独享特性减弱。换句话说，特征表示 $\phi(\mathbf{x})$ 使得边缘概率分布 $P_s(\phi(\mathbf{x}))$ 和 $P_t(\phi(\mathbf{x}))$ 之间的差异减小，且条件分布相同 $P_s(y|\phi(\mathbf{x}))=P_t(y|\phi(\mathbf{x}))$ 的假设更容易成立，因而比实例权重法具有更好的迁移能力。多数特征表示法[8,23,35–45] 都基于如下假设：特征空间中的部分特征是领域独享的、而另一部分特征是领域共享的且可泛化的；或者存在一个领域间共享的且可泛化的隐含特征空间，该空间可以由特征学习算法在减小领域间概率分布差异的准则下抽取得到。

特征表示法可以进一步分为两个子类：隐含表征学习法和概率分布适配法。隐含表征学习法通过分析辅助领域和目标领域的大量无标样例来构建抽象特征表示，从而隐式地缩小领域间的分布差异[8,35,36,39,42]。概率分布适配法通过惩罚或移除在领域间统计可变的特征[23,41]、或通过学习子空间嵌入表示来最小化特定距离函数[40,43,45–47]，从而显式地提升辅助领域和目标领域的样本分布相似度。本文第二章属于隐含表征学习法，第三至第五章属于概率分布适配法。与已有工作不同，本文重点研究了隐含表征学习中的负迁移问题和概率分布适配中的欠适配问题。

隐含表征学习法： 在迁移学习中，常见的是目标领域中的部分特征在辅助领域中的频率为零或为不变常数，反之亦然；这些特征即为领域独享特征。例如在情感分类中，从 *books* 图书领域迁移到 *furniture* 家具领域，情感词如“sharp”（锋利的）对目标领域是重要的极性判别线索、但对辅助领域却并不含有判别信息。这时候，如果在辅助领域标注数据上直接由原始特征训练分类器，则该分类器不能对

目标领域中独享特征进行有效判别；然而，如果利用两个领域中无标数据中隐含的*共现关系*，则可以推断出新的特征表示来对领域独享特征与共享特征进行聚合，形成领域无关的特征表示。文献[42,48]采用典型相关分析、谱分析、奇异值分解等降维方法将原始特征聚合到隐含特征空间，从而提高特征表示的迁移能力。其中最著名的工作是*结构对应学习*（Structural Correspondence Learning，SCL）[8,35,36]，它基于多任务半监督学习[37]中*交互结构最小化*（Alternating Structural Minimization，ASO）算法学习隐含特征映射：ASO 算法首先利用无标数据上的多个“辅助”任务——这些任务预测由频率或互信息确定的“pivot”特征是否最终出现——来学习“判别性结构”，然后将学到的辅助任务参数向量连接成参数矩阵并执行奇异值分解，从而得到判别任务参数的隐含特征空间以及跨领域特征之间的对应关系。不过，SCL 仅对领域间存在特征关联关系（如自然语言处理）时才会有效；当特征空间维度很低、或特征之间关联关系缺失（如计算机视觉）时并无明显效果；此外，如何自动确定 SCL 的参数、进行自动的模型选择仍是一个有待解决的问题。

为了解决现有隐含表征学习方法中过度依赖于先验特征关联关系的缺点，*深度学习*[49]被成功用于抽取紧致的特征表示、强化迁移学习效果[50–53]。深度学习可以辨别隐含因式结构中反映数据分布变化的部分和不变的部分，并通过层次化结构、按照与隐含因式结构的相关性对输入特征进行聚合和抽象，这样可以降低变化部分的权重、提高不变部分的权重[50]。这显然有助于跨领域迁移学习，因为这些通用概念对领域特定的数据分布是不变的，可以作为知识迁移的桥梁。例如在 *electronics* *电子产品*领域，领域特定情感词如 “blur”、“fast”、“sharp” 需要重构领域共享词汇、或被领域共享词汇重构，这样才能建立更高层次的抽象语义（领域共享词汇是指具有相似情感的词汇如 “good” 或 “love”）。这样，辅助领域训练的分类器可以对目标领域的所有特征——甚至那些从未在辅助领域出现的词汇——赋予恰当的权重[52]。深度学习的优势是给定无标数据，通过*无监督预训练*（unsupervised pre-training）[54]即可学习数据中隐藏的抽象特征关系。然而，降低变化部分隐含因式结构的权重可能会扩大跨领域数据分布间的差异，因为在深度特征表示下辅助领域和目标领域都变得更为“紧致”从而更容易判别彼此；而扩大了的领域间分布差异会损害迁移学习效果。本文分析该问题原因并提出解决方法。

概率分布适配法：概率分布适配法[40,43,47]通过学习一个新特征表示 $\phi(\mathbf{x})$、并同时显式地减小边缘分布 $P_s(\phi(\mathbf{x}))$ 和 $P_t(\phi(\mathbf{x}))$ 之间的距离函数，使得辅助领域和目标领域的概率分布在新的特征表示 $\phi(\mathbf{x})$ 下更为相似。文献[40,43,47]都基于*最大均值差异*（Maximum Mean Discrepancy，MMD）准则来度量概率分布差异，并最小化该准则来实现领域间的概率分布适配。文献[40]提出了称为*最大均值差异嵌入*（Maximum Mean Discrepancy Embedding，MMDE）的核学习方法，在最小化

MMD 分布距离的条件下同时最大化核空间中的数据嵌入方差，然后通过核主成份分析得到数据的领域不变嵌入表征。该方法存在两个缺点：一方面它是直推式方法，仅学习了核矩阵而没有学习核函数，不具备样本外数据的泛化能力；另一方面它依赖于半正定规划，计算复杂度很高难以扩展到大规模数据上。针对这两个问题，文献[45] 提出了称为*迁移主成份分析*（Transfer Component Analysis，TCA）的降维方法，该方法能学习具备样本外数据泛化能力的核函数，并仅需要解一个本征分解问题。文献[43,47] 和 MMDE 一样通过最小化 MMD 准则实现领域间的概率分布适配，但它学习线性投影而非核矩阵，从而保证了样本外数据的泛化能力；同时它还最小化了辅助领域的结构风险泛函，从而具有目标领域的泛化误差上界。

基于特征空间中的概率分布相似度函数，可以对迁移学习的泛化误差上界进行理论分析。文献[55] 基于 $P_s(\mathbf{x}) \neq P_t(\mathbf{x})$ 假设，推导了辅助领域训练的监督模型在目标领域的泛化误差上界。该上界包括辅助领域训练误差和反映领域差异的两个度量准则——准则一度量了最佳分类器在两个领域中的综合性能，准则二基于 $\mathcal{A}$-距离[56] 度量了边缘分布 $P_s(\mathbf{x})$ 与 $P_t(\mathbf{x})$ 之间的差异。通过学习新特征表示，SCL 能够同时减小训练误差和 $\mathcal{A}$-距离，从而获得了较小的泛化误差上界。类似地，文献[34] 基于更恰当的分布距离度量函数给出了实例权重法的泛化误差上界。

概率分布适配法相对于隐含表征学习法的重要优势在于：通过距离函数显式地度量领域间概率分布的失配程度，从而自然地获得迁移学习泛化误差上界的重要组成部分，较容易进行理论分析。但现有方法仍存在*欠适配*这一关键性问题：一方面，仅适配边缘概率分布而没有适配条件概率分布；另一方面，依赖于核空间的 MMD 准则不能刻划任意非线性空间中的分布差异，因此最小化 MMD 不能充分修正跨领域概率分布失配问题。因此，欠适配问题也是本文的重点研究对象。

1.4 有待研究的问题

综上所述，迁移学习利用丰富标注的辅助领域来提高标注缺失的目标领域的泛化性能。然而，大多数迁移学习研究工作都是基于特定假设下的算法设计进行的，例如假设不同的领域间存在可以共享的隐含结构或相关实例、不同的任务间存在可以共享的子任务或稀疏表征等；虽然这些算法在很多应用领域如自然语言处理、计算机视觉等已经取得很好的实验效果，但对*何时迁移*“when to transfer”[1] 的研究仍然十分缺乏，尚没有一套机制来决策所作假设何时成立、何时不成立、何时导致*负迁移*问题。另外一些研究工作[45] 通过对概率分布相似性进行度量、并作为学习算法优化准则之一，其优点是对假设的依赖性不很严格、具有一定的理论泛化上界保证、负迁移问题发生可能性较低等；然而由于没有对联合概率分布

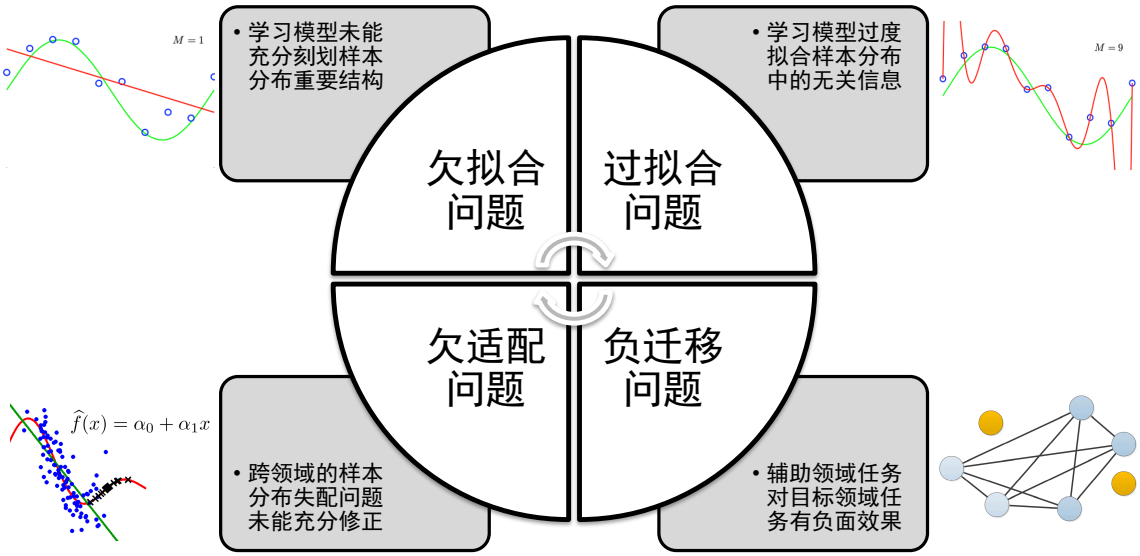


图 1.5　迁移学习的主要问题挑战：包括经典机器学习的过拟合、欠拟合问题，以及迁移学习特有的欠适配、负迁移问题。这些问题挑战交错叠加，大大增加了问题解决的难度。

进行适配、所采用的概率分布相似性度量准则过于简单、或对概率分布的抽象拟合能力不足等，这些方法存在突出的欠适配问题。由此可见，对于如何利用隐含表征学习和概率分布适配提高跨领域、跨任务迁移学习的泛化性能，在问题和方法层面都还存在较多未确定性，具体来说可以分为以下三个方面，如图 1.5 所示：

首先是负迁移问题，即指辅助领域任务对目标领域任务有负面效果。目前已有少数工作从算法设计角度对负迁移问题进行研究，主要思想是减少在领域间迁移的知识结构，例如仅在领域间共享模型的先验概率、而不共享模型参数或似然函数[57]。这些工作通过降低知识迁移来避免负迁移，因而难免又陷入“欠迁移”困境；如何权衡欠迁移和负迁移成为主要挑战之一。此外，这类算法难以应对各种假设场景，即如果假设条件变化就需要尝试其他迁移学习模型，应用代价很大。

其次是欠适配问题，即指跨领域的概率分布失配问题未能充分修正。概率分布适配方法通过最小化领域间概率分布差异，获得了具有理论保证的泛化误差上界，并在很多实际任务上取得了不错的效果，例如低维回归任务和高维分类任务等。尽管如此，现有的概率分布相似性度量函数，如最大均值差异[33]、布雷格曼散度[46] 等仍然过于简单，不能充分刻划概率分布的相似程度。例如被广泛采用的最大均值差异准则仅能匹配不同概率分布的各阶矩（如均值、方差）、而不能匹配各种更为复杂的模式（如类簇、流形）。因此，最小化这类度量函数并不能充分减小领域间的概率分布距离，导致欠适配问题。此外，当目标领域没有标注数据时，现有方法通常不考虑条件分布适配问题，但该问题对分类性能具有决定性的影响。

最后是欠拟合问题，即指学习模型未能充分刻划概率分布的重要结构。概率分布适配方法能有效工作的前提是能对概率分布自身的统计特性进行深度拟合，

表 1.1 各章节研究内容在学习问题、问题描述、学习方法三个层面的具体展现。

章节	学习问题	问题描述	学习方法
第 2 章	负迁移	欠迁移与负迁移权衡	矩阵分解
第 3 章	欠适配	条件分布适配度缺失	监督学习、特征学习
第 4 章	欠适配	矩匹配适配能力不足	谱核学习
第 5 章	欠适配、欠拟合	浅层学习抽象能力不足	深度学习

但现在基于浅层网络的学习模型显然难以挖掘概率分布中的复杂结构、特别是领域无关的抽象模式，这也是深度学习方法在部分迁移学习任务中取得优异效果的原因[50]。显然，现有工作仅通过在浅层网络中最小化领域间概率分布差异是不足以挖掘领域不变的抽象特征表示的，需要同时进行深度特征学习和概率分布适配。

1.5 研究内容与主要贡献

为了解决迁移学习的负迁移、欠适配、欠拟合等关键性基础问题，本文系统地研究了问题产生的原因及其解决方法，按照循序渐进的原则分为三大部分。第一部分：提出图正则化联合矩阵分解方法，解决负迁移问题；第二部分：提出联合分布适配方法和领域不变核学习方法，分别从条件分布适配和核适配角度解决欠适配问题；第三部分：提出深度迁移学习方法，同时解决领域内概率分布建模的欠拟合问题以及领域间概率分布匹配的欠适配问题。各个章节主要研究内容在学习问题、问题描述、学习方法三个层面的具体展现如表 1.1 所示，可以看到“概率分布适配”和“隐含表征学习”是贯穿全文研究内容的主线，具有完整的体系。

第一部分，即第 2 章，解决负迁移问题。为实现知识迁移，现有方法均假设领域间具有公共知识结构，如公共隐含语义、话题、本征谱等。然而，当领域间的概率分布差异很大时，上述假设通常难以成立，这会导致严重的负迁移问题。负迁移是迁移学习实用化的主要挑战，因为一个没有性能保证的模型难以被实践广泛采用。本章针对负迁移问题，提出了通用框架图正则化联合矩阵分解，其基本思想是：抽取领域间的公共隐含语义作为桥梁实现知识迁移，同时保持领域内几何流形结构不受领域外知识结构的破坏。在此自适应框架下，由于领域内流形结构得以完整保持，不管领域间迁移的知识结构是否能够共享，学习模型均能保证较好的目标领域学习效果。同时，本文基于矩阵分解提出了三种学习模型，并由系统性的实验证明了相对领域前沿方法的有效性。该部分主要研究成果已发表在数据挖掘顶级国际期刊 IEEE TKDE 2013[58] 和人工智能顶级国际会议 AAAI 2012[59] 上，并且成为国家核高基重大专项“非结构化数据管理系统”的代表性成果之一。

第二部分，包括第 3、4 两章，解决欠适配问题。针对领域间概率分布的适配

问题，在目标领域没有标注数据的非监督迁移学习中，已有方法通常仅适配边缘概率分布，但这会导致条件概率分布在领域间的欠适配问题，严重影响学习模型的泛化性能。第 3 章提出了联合适配正则化框架，在结构风险最小化和正则化理论的支撑下，重点解决条件概率分布的适配问题。同时，为了结合半监督学习的优势，还集成了基于流形正则化的半监督学习方法。在此基础上，提出了基于正则化线性回归、支持向量机、主成份分析等四种学习模型，并通过可再生希尔伯特空间中的表出定理给出模型的凸优化解。在文本分类和图像识别任务上的系统性实验证明了本文方法相对已有工作的优势。该部分研究成果已发表在数据挖掘顶级国际期刊 IEEE TKDE 2013[60] 和计算机视觉顶级国际会议 ICCV 2013[61] 上。

现有工作通常要在原始特征空间中进行概率分布适配，并依赖非线性核映射来实现不同概率分布之间各阶次矩匹配。然而，这类非线性核空间一般不是标准核机器（如支持向量机、核岭回归等）所依赖的最优核空间，从而导致概率分布适配与监督模型训练不能同时达到最优解，带来潜在的欠适配问题。为了解决该问题，第 4 章提出了领域不变迁移核学习，直接基于可再生希尔伯特空间学习一个领域不变核矩阵来实现辅助领域和目标领域的分布适配。具体地，首先由 Mercer 定理将目标领域本征系统外插值到辅助领域得到谱核矩阵族，其次选取与辅助领域真实核矩阵近似误差最小的插值谱核矩阵来构造领域不变核机器。在文本分类、图像识别、视频识别等大量任务上的系统性实验证明了本文方法相对已有前沿方法的优势。该部分主要研究成果已收录模式识别顶级国际会议 CVPR 2014[62] 上。

第三部分，即第 5 章，解决欠适配和欠拟合问题。前两章给出了如何在领域间适配概率分布，从而获得较低的模型泛化误差上界。但仍然存在两个根本技术挑战：（1）如何度量分布差异（2）如何学习领域不变的紧致特征表示。已有方法主要关注如何通过浅层网络进行知识巩固，但不能挖掘高度抽象、紧致的特征表示，因而无法对复杂概率分布进行深入刻划，存在欠拟合问题。由于领域内概率分布拟合是领域间概率分布适配的基础，已有方法的欠拟合问题必然同时导致欠适配问题。为同时解决欠拟合与欠适配问题，本章提出深度迁移学习框架，在深度网络架构下同时进行领域不变深度表征学习和概率分布差异修正。在本文的解决方案中，首先提出非线性分布差异，基于通用非线性特征学习来形式化领域间的概率分布失配程度；其次提出不变去噪自动编码器模型，通过数据的损失版本重构数据的原始版本来学习数据的鲁棒特征表示、并同时最小化领域间的非线性分布差异度量，进一步由多层模型堆叠形成深度网络巩固了特征表示的紧致性和不变性；最后提出迁移交叉验证策略，用于目标领域无标注数据的非监督迁移学习的模型选择。本文方法在情感极性分析、垃圾邮件过滤、视觉对象识别等跨领域任务上与前沿方法进行了比较，获得了显著的性能提升和创纪录的分类准确率。

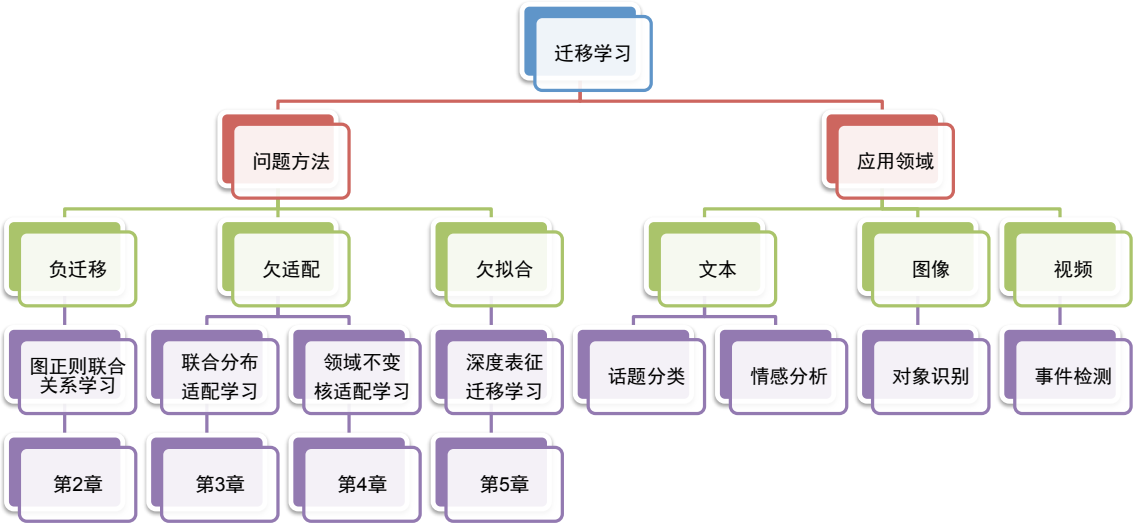


图 1.6　本文的组织结构和各章节关系。

1.6　本文的组织结构

本文组织结构和各章节关系如图 1.6 所示，按照循序渐进原则分为 6 个章节。

第 1 章为绪论部分，介绍了迁移学习的研究背景、理论和应用价值，定义了学习问题场景，综述了国内外研究现状，并概括了本文的主要研究内容和贡献。

第 2 章提出了图正则化联合矩阵分解框架，利用联合矩阵分解抽取领域间可共享的隐含语义结构，并利用图正则化对这些语义结构进行权衡和优选，从而能够选取最佳的共享知识结构，一定程度上解决了负迁移问题。

第 3 章提出了联合适配正则化框架，以及基于正则化线性回归、支持向量机、主成份分析的 4 种学习模型，实现了条件分布在领域间的适配，解决了仅适配边缘概率分布导致的欠适配问题，实验证明条件分布适配对判别问题的不可或缺性。

第 4 章提出了领域不变核学习方法，通过核矩阵的低秩近似、谱核学习等技术实现了跨领域核矩阵的匹配，解决了现有基于核空间矩匹配方法存在的欠适配问题。与第 3 章不同之处在于，本章着眼于强化边缘概率的适配程度而不考虑条件分布适配问题，因而具有更好的通用性，可以应用到分类以外的实际问题中。

第 5 章提出了深度表征适配方法，通过深度学习的非线性特征映射深入刻划了领域间的概率分布差异、并将深度特征学习与概率分布适配集成到统一的深度网络中，极大提高了领域内概率分布的抽象建模以及领域间概率分布的深度适配。

第 6 章对本文的主要研究内容和贡献进行回顾和总结，并展望了未来将继续开展的工作，特别是对迁移学习泛化误差下界这一重要的理论问题展开深入研究。

第 2 章　图正则化联合矩阵分解方法

迁移学习通过辅助领域的丰富标注数据，解决目标领域的标注稀缺问题。为实现知识迁移，现有方法均假设领域间具有公共知识结构，如公共隐含语义、话题、本征谱等。然而，当领域间的概率分布差异很大时，上述假设通常难以成立，这会导致严重的负迁移问题。负迁移是迁移学习实用化的主要挑战，因为一个没有性能保证的模型难以被实践广泛采用。本章针对负迁移问题，提出了通用框架图正则化联合矩阵分解，其基本思想是：抽取领域间的公共隐含语义作为桥梁实现知识迁移，同时保持领域内几何流形结构不受领域外知识结构的破坏。在此自适应框架下，由于领域内流形结构得以完整保持，不管领域间迁移的知识结构是否能够共享，学习模型均能保证较好的目标领域学习效果。同时，本章基于矩阵分解提出了三种学习模型，并由系统性的实验证明了相对领域前沿方法的有效性。

2.1　引言

随着多领域、多媒体大数据不断涌现，如何研究自动方法对其进行跨领域分类和组织变得愈加重要。对新领域执行机器学习常遇到标注稀缺问题，严重制约了经典监督学习方法的效果，而重新收集大量标注数据又费时费力。迁移学习[1]通过从辅助领域迁移标注数据来提升目标领域的学习效果，它放宽了经典监督学习关于训练数据和测试数据服从独立同分布这一基本假设，因而在跨领域文本分类[63]、情感分析[42]、图像识别[14]、视频摘要[64] 和协同推荐[65] 中得到广泛应用。

迁移学习的一个根本性计算问题是如何挖掘领域间的共享知识结构作为标注信息从辅助领域迁移到目标领域的桥梁。当前主要有两类主流方法：（1）通过统计似然[66] 抽取隐含语义结构，并假设这些结构可以在领域间共享和迁移[63,67–71]（2）通过谱图理论[72] 抽取数据本征谱，并假设这些结构可以在领域间共享和迁移。这些迁移学习方法的主要缺陷在于：没有同时对数据的统计和谱图信息进行协同建模，并考虑不同信息在知识迁移时的共性和特性。例如，在网络文档内容挖掘等任务中，数据中蕴含多种知识结构而非单一知识结构，这时已有方法就存在欠迁移缺点；又如，在图像内容与概念迁移等任务中，领域之间的概率分布差异如此之大以至于根本难以挖掘公共隐含语义结构作为迁移桥梁，这时已有方法就存在负迁移问题。上述讨论启发了本章的图正则化联合矩阵分解框架。

针对欠迁移问题，受文献[66] 启发，数据的统计信息和几何结构通常从不同维度刻划了数据分布、因而具有相互强化的效果。其理由是：（1）根据生成原理，

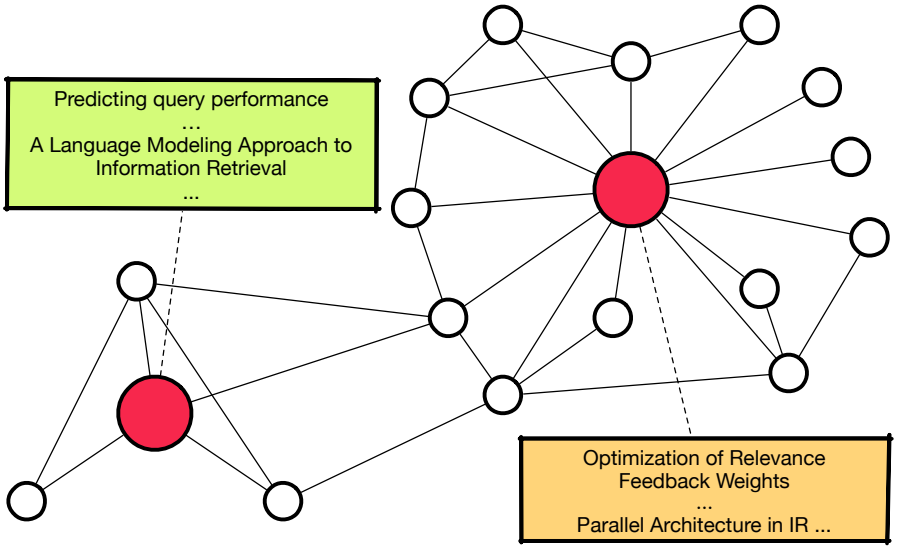


图 2.1　图正则化联合矩阵分解框架的基本原理：同时对统计信息和结构信息进行建模。图中给出的带网络（或几何流形）结构的文档集合，可以看到：由文档数据本身的词频统计信息可以挖掘隐含语义结构，由文档间的网络结构可以挖掘潜在流形结构。两者都对迁移学习至关重要，本章对它们的角色加以区分，目标是同时解决欠迁移和负迁移问题。

数据通常是由潜在语义结构混合生成的，例如文本文档通常是由隐含话题混合而成，而抽取这类隐含语义结构归结为挖掘数据的统计信息[73]；（2）根据流形理论，数据通常位于高维环绕空间中的某一潜在低维流形结构之上，抽取这些流形结构归结为挖掘数据的几何信息[72]。因此，同时对数据的统计信息和几何结构进行建模，有利于实现有效迁移学习，这可以从如图 2.1 给出的示意图中获得直观理解。

针对负迁移问题，本章给出如下的结论：当辅助领域和目标领域之间的数据分布差异足够大时，试图抽取领域间的公共语义结构会以很大概率失败；这时如果强行抽取所谓的“公共”语义结构，则必然会导致与目标领域判别结构不一致，从而引发负迁移，这也是现有方法存在的主要缺陷。基于上述原因，为了保持领域内结构信息不被破坏，本章提出在每个领域内进行图正则化、保持其几何结构不变，从而在领域间共享知识结构可能导致负迁移的情况下“力挽狂澜于既倒”。

本章针对负迁移问题，提出了通用框架图正则化联合矩阵分解，其基本思想是：抽取领域间的公共隐含语义作为桥梁实现知识迁移，同时保持领域内几何流形结构不受领域外知识结构的破坏，其原理如图 2.1 所示。主要贡献归纳如下：

- 首次提出同时对领域间统计信息和领域内几何结构进行建模，在增强正迁移的同时反制了负迁移，较为巧妙地解决了欠迁移和负迁移权衡的两难困境。
- 各种矩阵分解方法如 NMF[74]、Semi-NMF[75]、NMTF[76] 等，均可直接载入本章框架实现迁移学习任务，因而能灵活地针对具体问题采用具体方法。
- 在 222 个文本分类和图像识别任务中证明了本章方法相对已有方法的优势。

表 2.1　本章常用的符号及其描述。

符号	描述	符号	描述	符号	描述	符号	描述
$\mathcal{D}_\pi$	领域 π	p	图近邻数	$\mathbf{X}_\pi$	$\mathcal{D}_\pi$ 数据矩阵	$\mathbf{U}_\pi$	$\mathcal{D}_\pi$ 隐含语义
n_π	$\mathcal{D}_\pi$ 样例数	λ	特征图参数	$\mathbf{Y}_\pi$	$\mathcal{D}_\pi$ 标注矩阵	$\mathbf{U}$	公共隐含语义
m	特征数	γ	样例图参数	$\mathbf{L}_\pi^u$	$\mathcal{D}_\pi$ 特征图矩阵	$\mathbf{V}_\pi$	$\mathcal{D}_\pi$ 类别表示
c	类别数	σ	正交项参数	$\mathbf{L}_\pi^v$	$\mathcal{D}_\pi$ 样例图矩阵	$\mathbf{H}_\pi$	$\mathbf{U}_\pi$ 和 $\mathbf{V}_\pi$ 结构关联

2.2　图正则化联合矩阵分解

在本节中，首先给出问题定义和学习目标，其次展示图正则化框架（Graph Co-Regularized Transfer Learning，GTL），在此基础上给出两种分别基于矩阵二分解和矩阵三分解的学习算法，最后分析学习算法的复杂度、正确性和负迁移问题。

2.2.1　问题定义

本章考察直推式迁移学习，即辅助领域存在大量标注数据、而目标领域仅存在无标数据的迁移学习问题。本章形式化一个辅助领域和一个目标领域的多类分类学习模型，但该模型可直接推广到多个辅助领域和多个目标领域的学习任务上。

记 $\mathcal{D}_\pi$ 为第 π 个领域，其中 $\pi \in \Pi$ 是领域序号。对多个领域按照类型将其序号集合 Π 划分为辅助领域 Π_s 和目标领域 Π_t，即满足 $\Pi = \Pi_s \cup \Pi_t$ 且 $\Pi_s \cap \Pi_t = \emptyset$。所有领域共享特征空间 $\mathcal{X}$ 和类别空间 $\mathcal{Y}$，其中共有 $|\mathcal{X}| = m$ 个特征和 $|\mathcal{Y}| = c$ 个类别标签。记 $\mathbf{X}_\pi = [\mathbf{x}_{*1}^\pi, \ldots, \mathbf{x}_{*n_\pi}^\pi] \in \mathbb{R}^{m \times n_\pi}$ 为领域 $\mathcal{D}_\pi$ 的特征 – 样例矩阵，其中 $\mathbf{x}_{*i}^\pi$ 是领域 $\mathcal{D}_\pi$ 的第 i 个样例。记 $\mathbf{Y}_\pi \in \mathbb{R}^{n_\pi \times c}$ 为辅助领域 $\mathcal{D}_\pi, \pi \in \Pi_s$ 的标注矩阵，其中 $y_{ij}^\pi = 1$ 如果 $\mathbf{x}_{*i}^\pi$ 隶属于类别 j，否则 $y_{ij}^\pi = 0$。文中常用的符号和描述如表 2.1 所示。

问题 2.1 (**学习目标**)：　给定多个领域 $\{\mathcal{D}_\pi\}_{\pi \in \Pi}$，其中辅助领域 $\{\mathcal{D}_\pi\}_{\pi \in \Pi_s}$ 完全标注，学习在目标领域 $\{\mathcal{D}_\pi\}_{\pi \in \Pi_t}$ 上错误率最低的多类分类器 $f : \mathcal{X} \mapsto \mathcal{Y}$，且（*1*）通过保持领域间的统计属性实现知识迁移（*2*）通过保持领域内的几何结构避免负迁移。

本章提出了图正则化迁移学习（Graph Co-Regularized Transfer Learning，GTL）框架来实现上述学习目标。GTL 采用了正则化矩阵分解技术：基于输入领域间共享某些公共隐含语义的假设，通过联合矩阵分解抽取这些公共隐含语义并保持原始数据在领域间的统计属性；同时，通过图正则化对所抽取的隐含语义进行精化并保持原始数据在领域内的几何结构。这样，所得到的学习模型能够对领域间的分布差异具有鲁棒性，原因是：（1）如果统计属性与几何结构在领域间一致，则两者可以互相精化、加强知识迁移能力；（2）否则，领域内的几何结构可以支配领域内的学习任务、避免领域外知识带来的负迁移。图正则化迁移学习框架 GTL 将联合矩阵分解和图正则化两类学习目标集成为统一的优化问题，如图 2.2 所示。

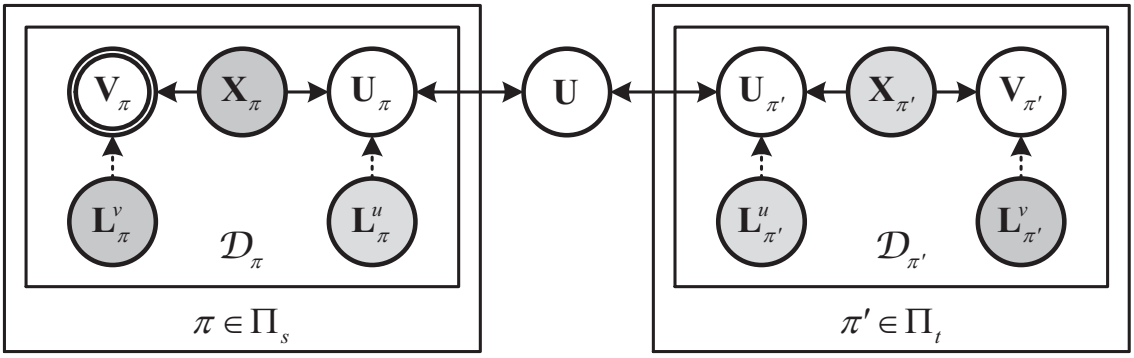


图 2.2　图正则化联合矩阵分解通用框架示意图，这里仅以共享隐含语义结构 **U** 为例。

2.2.2　联合矩阵分解

首先，通过联合矩阵分解[77] 抽取隐含语义特征，在领域间保持数据的统计信息，原始数据经过重新表征后在领域间的概率分布差异可以隐式地减小。

联合矩阵分解： 领域 $\mathcal{D}_\pi$ 的隐含语义可以通过非负矩阵分解（Nonnegative Matrix Factorization，NMF）模型[74,75] 抽取。在 NMF 中，特征 – 样例矩阵 $\mathbf{X}_\pi$ 可以分解为两个低秩非负矩阵 $\mathbf{U}_\pi$ 和 $\mathbf{V}_\pi$ 之乘积，它们对矩阵 $\mathbf{X}_\pi$ 的重构误差达到最小化、且原始数据的统计信息可以得到保持。NMF 归结为如下的优化问题：

$$\min_{\mathbf{U}_\pi, \mathbf{V}_\pi \geq \mathbf{0}} \mathcal{L}\left(\mathbf{X}_\pi, h\left(\mathbf{U}_\pi \mathbf{V}_\pi^{\mathrm{T}}\right)\right) \tag{2-1}$$

其中 h 是预测连接函数，$\mathcal{L}$ 是重构损失函数。$\mathbf{U}_\pi = [\mathbf{u}_{*1}^\pi, \ldots, \mathbf{u}_{*c}^\pi] \in \mathbb{R}^{m \times c}$ 是特征聚类矩阵，其中每列 $\mathbf{u}_{*i}^\pi$ 代表一个隐含语义；$\mathbf{V}_\pi = [\mathbf{v}_{*1}^\pi, \ldots, \mathbf{v}_{*c}^\pi] \in \mathbb{R}^{n_\pi \times c}$ 是样例类别矩阵，其中每列 $\mathbf{v}_{*i}^\pi$ 代表一个样例类别。直观上，$\mathbf{U}_\pi$ 和 $\mathbf{V}_\pi$ 是数据矩阵 $\mathbf{X}_\pi$ 在特征空间和样例空间进行协同聚类的结果。根据文献[73] 的证明，NMF 从优化问题上等价于概率隐含语义分析（PLSA），它们都归结于最大化原始数据的概率似然函数。

给定具有相互关联关系的多个领域，通过挖掘它们隐含的公共因式（如隐含语义、结构关联），就可以由标注领域迁移判别结构来提高无标领域的分类效果，这就是迁移学习有效工作的内在机制。基于上述观点，文献[77] 将非负矩阵分解模型加以扩展，从而能够同时对多个相关领域进行因式分解、并共享公共因式，得到联合矩阵分解（Collective Matrix Factorization，CMF）模型，其优化问题如下：

$$\min_{\mathbf{U}_\pi \in C_u, \mathbf{V}_\pi \in C_v} \sum\nolimits_{\pi \in \Pi} \mathcal{L}\left(\mathbf{X}_\pi, h\left(\mathbf{U}_\pi \mathbf{V}_\pi^{\mathrm{T}}\right)\right) \tag{2-2}$$

其中 C_u 和 C_v 分别是模型因式矩阵 $\mathbf{U}_\pi$ 和 $\mathbf{V}_\pi$ 的约束条件（如非负性、正交性）。CMF 的主要思想是在多个相互关联矩阵间共享因式矩阵。在文献[67,68,78] 中，通常将隐含语义矩阵 $\{\mathbf{U}_\pi\}_{\pi \in \Pi}$ 作为领域间的共享因式，从而实现知识迁移，也即设置 $C_u \triangleq \{\mathbf{U}_\pi \equiv \mathbf{U} : \forall \pi \in \Pi\}$。这等价于在语义挖掘过程中保持领域间的统计属性不变。

联合矩阵三分解： 类似地，也可以通过非负矩阵三分解（Nonnegative Matrix Tri-Factorization，NMTF）模型[76] 抽取领域间的隐含语义结构。在 NMTF 中，特征 – 样例矩阵 $\mathbf{X}_\pi$ 分解为三个低秩非负矩阵 $\mathbf{U}_\pi$、$\mathbf{H}_\pi$ 和 $\mathbf{V}_\pi$ 之乘积，优化问题如下：

$$\min_{\mathbf{U}_\pi,\mathbf{H}_\pi,\mathbf{V}_\pi\geq\mathbf{0}} \mathcal{L}\left(\mathbf{X}_\pi, h\left(\mathbf{U}_\pi\mathbf{H}_\pi\mathbf{V}_\pi^{\mathrm{T}}\right)\right) \tag{2-3}$$

其中 $\mathbf{H}_\pi \in \mathbb{R}^{c\times c}$ 是隐含语义 $\mathbf{U}_\pi$ 和样例类别 $\mathbf{V}_\pi$ 之间的关联结构，刻划了输入矩阵 $\mathbf{X}_\pi$ 的“鸟瞰视图”，即抽象“特征”与“样例”的共现关系。与 CMF 类似，可以将标准 NMTF 模型加以扩展，同时对多个相互关联的矩阵进行因式分解，得到联合矩阵三分解（Collective Matrix Tri-Factorization，CMTF）模型，优化问题如下：

$$\min_{\mathbf{U}_\pi\in C_u,\mathbf{H}_\pi\in C_h,\mathbf{V}_\pi\in C_v} \sum\nolimits_{\pi\in\Pi} \mathcal{L}\left(\mathbf{X}_\pi, h\left(\mathbf{U}_\pi\mathbf{H}_\pi\mathbf{V}_\pi^{\mathrm{T}}\right)\right) \tag{2-4}$$

由 CMTF，既可以共享隐含语义从而实现迁移学习[68,79]，即 $C_u \triangleq \{\mathbf{U}_\pi \equiv \mathbf{U} : \pi \in \Pi\}$；也可以共享结构关联从而实现迁移学习[63,69]，即 $C_h \triangleq \{\mathbf{H}_\pi \equiv \mathbf{H} : \pi \in \Pi\}$。本章同时考察了 CMF 和 CMTF 模型对迁移学习的效能，并侧重对负迁移问题的分析。

2.2.3　图正则化

其次，本章提出通过图正则化对联合矩阵分解抽取的隐含语义结构进行精化，可以使每个领域内的几何结构都分别得到保持，从而最大限度的避免负迁移问题。

样例图正则化： 从几何观点看，输入数据点可看成是由低维流形上的概率分布采样生成，而该低维流形嵌入在高维环绕空间中[72,80]。保持这种几何结构，使得学习模型可以尊重领域自身的数据分布、并实质地规避负迁移问题。根据局部不变假设[81]，如果领域 $\mathcal{D}_\pi$ 的两个样例 $\mathbf{x}_{*i}^\pi$ 和 $\mathbf{x}_{*j}^\pi$ 在数据分布的内在几何流形上彼此接近，则它们的嵌入表征 $\mathbf{v}_{i*}^\pi$ 和 $\mathbf{v}_{j*}^\pi$ 也应该相互接近。根据流形理论，上述几何结构可以通过数据集上的 p-近邻图进行有效建模[72]。考察样例图 G_π^v，其中包括 n_π 个顶点，每个顶点代表领域 $\mathcal{D}_\pi$ 的一个样例点，则图 G_π^v 的邻接矩阵可以定义为

$$(\mathbf{W}_\pi^v)_{ij} = \begin{cases} \mathrm{sim}\left(\mathbf{x}_{*i}^\pi, \mathbf{x}_{*j}^\pi\right), \text{如果 } \mathbf{x}_{*i}^\pi \in \mathcal{N}_p\left(\mathbf{x}_{*j}^\pi\right) \vee \mathbf{x}_{*j}^\pi \in \mathcal{N}_p\left(\mathbf{x}_{*i}^\pi\right) \\ 0, \text{其他} \end{cases} \tag{2-5}$$

其中 $\mathrm{sim}(\cdot,\cdot)$ 是相似性度量函数，$\mathcal{N}_p\left(\mathbf{x}_{*i}^\pi\right)$ 是位于样例 $\mathbf{x}_{*i}^\pi$ 的 p-近邻的样例集合。

注意到通过 CMF 抽取的样例 $\mathbf{x}_{*i}^\pi$ 的低维语义嵌入表征为 $\mathbf{v}_{i*}^\pi = [v_{i1}^\pi, \cdots, v_{ic}^\pi]$。采用损失函数 ℓ 来度量任意两个嵌入表征 $\mathbf{v}_{i*}^\pi$ 和 $\mathbf{v}_{j*}^\pi$ 的接近程度，即 $\ell\left(\mathbf{v}_{i*}^\pi, \mathbf{v}_{j*}^\pi\right)$。根据文献[72]，通过样例图 G_π^v 保持领域 $\mathcal{D}_\pi$ 的几何结构归结为如下的样例图正则化

$$\mathcal{R}(\mathbf{V}_\pi) = \frac{1}{2}\sum\nolimits_{i,j=1}^{n_\pi} \ell\left(\mathbf{v}_{i*}^\pi, \mathbf{v}_{j*}^\pi\right)(\mathbf{W}_\pi^v)_{ij} \tag{2-6}$$

特征图正则化： 考虑到特征样例之间存在对偶性，可认为特征也是由低维流形上的概率分布采样生成，而该低维流形嵌入在高维环绕空间中[82]。根据局部不变假设[81]，如果两个特征 $\mathbf{x}_{i*}^{\pi}$ 和 $\mathbf{x}_{j*}^{\pi}$ 在数据分布的内在几何流形上彼此接近，则它们的嵌入表征 $\mathbf{u}_{i*}^{\pi}$ 和 $\mathbf{u}_{j*}^{\pi}$ 也应该相互接近。与样例图类似，考察特征图 G_{π}^{u}，其中包括 m 个顶点，每个顶点代表领域 $\mathcal{D}_{\pi}$ 的一个特征，则图 G_{π}^{u} 的邻接矩阵可以定义为

$$(\mathbf{W}_{\pi}^{u})_{ij}=\begin{cases}\mathrm{sim}\left(\mathbf{x}_{i*}^{\pi},\mathbf{x}_{j*}^{\pi}\right), \text{如果 } \mathbf{x}_{i*}^{\pi}\in\mathcal{N}_{p}\left(\mathbf{x}_{j*}^{\pi}\right)\vee\mathbf{x}_{j*}^{\pi}\in\mathcal{N}_{p}\left(\mathbf{x}_{i*}^{\pi}\right)\\ 0, \text{其他}\end{cases} \tag{2-7}$$

其中 $\mathrm{sim}(\cdot,\cdot)$ 是相似性度量函数，$\mathcal{N}_{p}\left(\mathbf{x}_{i*}^{\pi}\right)$ 是位于特征 $\mathbf{x}_{i*}^{\pi}$ 的 p-近邻的特征集合。

注意到 CMF 为特征 $\mathbf{x}_{i*}^{\pi}$ 抽取的低维嵌入表征为 $\mathbf{u}_{i*}^{\pi}=[u_{i1}^{\pi},\cdots,u_{ic}^{\pi}]$。类似于样例图正则化，通过特征图 G_{π}^{u} 保持领域 $\mathcal{D}_{\pi}$ 的几何结构归结为如下的特征图正则化

$$\mathcal{R}(\mathbf{U}_{\pi})=\frac{1}{2}\sum\nolimits_{i,j=1}^{m}\ell\left(\mathbf{u}_{i*}^{\pi},\mathbf{u}_{j*}^{\pi}\right)(\mathbf{W}_{\pi}^{u})_{ij} \tag{2-8}$$

将公式 (2-6) 和公式 (2-8) 中的图正则项称为图协同正则项，因为它们作为协同聚类的约束出现，同时保持了样例空间和特征空间中的流形结构。由它们可以对隐含语义等知识结构进行精化、避免负迁移，这可以从下文的论述中得到证实。

2.2.4　优化框架

为了进一步提升跨领域分类的性能，联合矩阵分解和图正则化两个学习准则应作统一优化，其原因是：（1）由联合矩阵分解，可以抽取公共隐含语义结构用于知识迁移；（2）由图正则化，可以保持领域内的几何结构不被破坏从而避免负迁移。此外，联合矩阵分解和图正则化统一优化，还可以交互增强各自特有的效果：（1）联合矩阵分解可以在特征空间和样例空间中抽取满足统计似然函数最大化的嵌入表征；（2）图正则化可以为嵌入表征注入具有判别信息的几何结构提高分类效果。因此，统一优化上述两个学习准则可以得到如下的 GTL 通用迁移框架：

$$\begin{aligned}&\min_{\mathbf{U}_{\pi}\in C_{u},\mathbf{V}_{\pi}\in C_{v}}\sum_{\pi\in\Pi}\left[\mathcal{L}\left(\mathbf{X}_{\pi},h\left(\mathbf{U}_{\pi}\mathbf{V}_{\pi}^{\mathrm{T}}\right)\right)+\lambda\mathcal{R}(\mathbf{U}_{\pi})+\gamma\mathcal{R}(\mathbf{V}_{\pi})\right]\\ &\text{s.t.}\quad C_{u}\triangleq\{\mathbf{U}_{\pi}\equiv\mathbf{U}:\pi\in\Pi\},C_{v}\triangleq\{\mathbf{V}_{\pi}\equiv\mathbf{Y}_{\pi}:\pi\in\Pi_{s}\}\end{aligned} \tag{2-9}$$

其中 $\lambda>0$ 是特征图正则项参数，$\gamma>0$ 是样例图正则项参数。由于辅助领域存在标注数据，可通过约束条件 $C_{v}\triangleq\{\mathbf{V}_{\pi}\equiv\mathbf{Y}_{\pi}:\forall\pi\in\Pi_{s}\}$ 将其注入优化问题中。为了实现迁移学习，根据文献[67,77,78] 可在领域间共享隐含语义，即 $C_{u}\triangleq\{\mathbf{U}_{\pi}\equiv\mathbf{U}:\forall\pi\in\Pi\}$，通过该公共结构作为知识迁移桥梁，可以将监督信息从辅助领域传播到目标领域。

类似地，GTL 通用框架也可以通过联合矩阵三分解（CMTF）形式化如下：

$$\begin{aligned}&\min_{\mathbf{H}_\pi\in C_h,\mathbf{V}_\pi\in C_v}\sum_{\pi\in\Pi}\left[\mathcal{L}\left(\mathbf{X}_\pi,h\left(\mathbf{U}_\pi\mathbf{H}_\pi\mathbf{V}_\pi^{\mathrm{T}}\right)\right)+\lambda\mathcal{R}\left(\mathbf{U}_\pi\right)+\gamma\mathcal{R}\left(\mathbf{V}_\pi\right)\right]\\&\text{s.t.}\quad C_h\triangleq\{\mathbf{H}_\pi\equiv\mathbf{H}:\pi\in\Pi\},C_v\triangleq\{\mathbf{V}_\pi\equiv\mathbf{Y}_\pi:\pi\in\Pi_s\}\end{aligned}\tag{2-10}$$

为了实现迁移学习，根据文献[63,69]可在领域间共享关联结构，即 $C_h\triangleq\{\mathbf{H}_\pi\equiv\mathbf{H}:\forall\pi\in\Pi\}$。通过上述优化问题的最优解，可以预测目标领域 $\mathcal{D}_\pi$ 样例 $\mathbf{x}_{*i}^{\pi}$ 类别如下：

$$f\left(\mathbf{x}_{*i}^{\pi}\right)=\arg\max_j\left(\mathbf{V}_\pi\right)_{ij}\tag{2-11}$$

上述通用 GTL 框架可以采用各种不同的算法配置：预测连接函数 h、损失函数 $\mathcal{L}$ 和 ℓ、相似度量函数 sim、约束条件 C_u 和 C_v。被广泛采用的配置选项如下：

- h 可选为恒等函数或逻辑斯特函数，即 $h(\mathbf{X})=\mathbf{X}$ 或 $h\left(X_{ij}\right)=\frac{1}{1+e^{-X_{ij}}}$。
- $\mathcal{L}$ 可选为二次损失函数或矩阵散度[72]，即 $\mathcal{L}(\mathbf{X},\widehat{\mathbf{X}})=\|\mathbf{X}-\widehat{\mathbf{X}}\|_F^2$ 或 $\mathcal{L}\left(\mathbf{X},\widehat{\mathbf{X}}\right)=\sum_{ij}\left(X_{ij}\log\frac{X_{ij}}{\widehat{X}_{ij}}-X_{ij}+\widehat{X}_{ij}\right)$。
- ℓ 可选为欧式距离或广义相对熵[72]，即 $\ell(\mathbf{x},\widehat{\mathbf{x}})=\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2$ 或 $\ell\left(\mathbf{x},\widehat{\mathbf{x}}\right)=\sum_i\left(x_i\log\frac{x_i}{\widehat{x}_i}-x_i+\widehat{x}_i\right)$。
- sim 可选为余弦相似度或热核权重[72]，即 $\mathrm{sim}(\mathbf{x},\widehat{\mathbf{x}})=\cos(\mathbf{x},\widehat{\mathbf{x}})$ 或 $\mathrm{sim}\left(\mathbf{x},\widehat{\mathbf{x}}\right)=\exp\left(-\frac{\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2}{\beta^2}\right)$，其中 β 是带宽参数。
- C_u 和 C_v 可选为非负约束（NMF）、正交约束（SVD）、概率约束（PLSA）[73]、或 L_1/L_2-范数约束。

根据特定应用，可以从上述配置选项中选择最佳的算法配置以获得最佳学习性能。

2.3　学习算法与分析

本节在 GTL 框架下，扩展标准学习算法 NMF[74] 和 NMTF[76]：采用线模型即 $h(\mathbf{X})=\mathbf{X}$、$\mathcal{L}(\mathbf{X},\widehat{\mathbf{X}})=\|\mathbf{X}-\widehat{\mathbf{X}}\|_F^2$、$\ell(\mathbf{x},\widehat{\mathbf{x}})=\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2$ 以及 $\mathrm{sim}(\mathbf{x},\widehat{\mathbf{x}})=\cos(\mathbf{x},\widehat{\mathbf{x}})$；对样例类别矩阵采用近似正交约束，即 $C_v\supset\left\{\left\|\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi-\mathbf{I}\right\|_F^2\le\varepsilon:\forall\pi\in\Pi\right\}$，其中 ε 是一个充分小的正数。从后文论述中可以看到，GTL 需要通过正交约束来避免平凡解。

2.3.1　矩阵二分解

采用非负矩阵二分解 NMF[74] 为基础，框架 (2-9) 可形式化为学习模型 GTL_2：

$$\begin{aligned}O_2=&\sum\nolimits_{\pi\in\Pi}\left\|\mathbf{X}_\pi-\mathbf{U}\mathbf{V}_\pi^{\mathrm{T}}\right\|_F^2+\frac{\sigma}{2}\sum\nolimits_{\pi\in\Pi}\left\|\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi-\mathbf{I}\right\|_F^2\\&+\lambda\sum\nolimits_{\pi\in\Pi}\mathrm{tr}\left(\mathbf{U}^{\mathrm{T}}\mathbf{L}_\pi^u\mathbf{U}\right)+\gamma\sum\nolimits_{\pi\in\Pi}\mathrm{tr}\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{L}_\pi^v\mathbf{V}_\pi\right)\end{aligned}\tag{2-12}$$

其中 σ 是正交正则参数；$\mathbf{L}_\pi^u$ 和 $\mathbf{L}_\pi^v$ 是图拉普拉斯矩阵，可计算为 $\mathbf{L}_\pi^u = \mathbf{D}_\pi^u - \mathbf{W}_\pi^u$ 和 $\mathbf{L}_\pi^v = \mathbf{D}_\pi^v - \mathbf{W}_\pi^v$；$\mathbf{D}_\pi^u$ 和 $\mathbf{D}_\pi^v$ 是对角矩阵，其对角元为 $(\mathbf{D}_\pi^u)_{ii} = \sum_{j=1}^{m}(\mathbf{W}_\pi^u)_{ij}$ 和 $(\mathbf{D}_\pi^v)_{ii} = \sum_{j=1}^{n_\pi}(\mathbf{W}_\pi^v)_{ij}$。根据正则化方法，对 $\mathbf{V}_\pi, \forall \pi \in \Pi$ 的正交约束可通过正交正则项（目标函数第二项）来近似满足。这是因为根据拉格朗日乘子法，给定任意正数 ε 满足 $\left\|\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi - \mathbf{I}\right\|_F^2 \le \varepsilon$，总能找到合适的正数 σ 满足 $\left\|\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi - \mathbf{I}\right\|_F^2 \le \varepsilon$。优化问题 (2-12) 可由交互优化算法求解，如下面的定理所述，正确性证明参见第2.3.5节。

定理 2.1：　分别由公式 (2-13)~(2-14) 交互迭代式地更新 $\mathbf{U}$，$\{\mathbf{V}_\pi\}_{\pi\in\Pi}$ 可以保证目标函数 (2-12) 单调递减并收敛到局部最优解。

$$\mathbf{U} \leftarrow \mathbf{U} \odot \frac{\left[\sum_{\pi\in\Pi}\left(\mathbf{X}_\pi\mathbf{V}_\pi + \lambda\mathbf{W}_\pi^u\mathbf{U}\right)\right]}{\left[\sum_{\pi\in\Pi}\left(\mathbf{U}\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi + \lambda\mathbf{D}_\pi^u\mathbf{U}\right)\right]} \tag{2-13}$$

$$\mathbf{V}_\pi \leftarrow \mathbf{V}_\pi \odot \frac{\left[\mathbf{X}_\pi^{\mathrm{T}}\mathbf{U} + \gamma\mathbf{W}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\right]}{\left[\mathbf{V}_\pi\mathbf{U}^{\mathrm{T}}\mathbf{U} + \gamma\mathbf{D}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right]} \tag{2-14}$$

其中 $\odot$ 和 $\frac{[\cdot]}{[\cdot]}$ 分别表示点乘和点除操作（逐项乘法和除法）。

完整的训练过程总结如算法 1 所示。由于辅助领域是全标注的，因而保持 $\{\mathbf{V}_\pi \equiv \mathbf{Y}_\pi : \pi \in \Pi_s\}$ 在整个迭代过程中不变。由于上述优化问题并非凸优化，存在退化局部极值风险，因此将类别矩阵 $\{\mathbf{V}_\pi\}_{\pi\in\Pi_t}$ 初始化为目标领域预标注，即由辅助领域标注 $\{\mathbf{X}_\pi, \mathbf{Y}_\pi\}_{\pi\in\Pi_s}$ 训练的逻辑斯特回归模型（LR）在目标领域的预测结果。

算法 1: GTL_2：基于 NMF 的 GTL 迁移学习模型

输入: 输入数据 $\{\mathbf{X}_\pi\}_{\pi\in\Pi}$, $\{\mathbf{Y}_\pi\}_{\pi\in\Pi_s}$；模型参数 $p, \lambda, \gamma, \sigma$；迭代次数 T。

输出: 隐含因式 $\mathbf{U}$, $\{\mathbf{V}_\pi\}_{\pi\in\Pi}$, 分类结果 $\{\widehat{\mathbf{Y}}_\pi\}_{\pi\in\Pi_t}$。

```
开始
    分别由公式 (2-5) 和 (2-7) 构造邻接矩阵 W_π^v 和 W_π^u。
    随机初始化 U; V_π ← Y_π, ∀π ∈ Π_s; V_π ← LR(⋃_{π'∈Π_s}{X_π', Y_π'}), ∀π ∈ Π_t。
    for t ← 1 to T do
        由公式 (2-13) 更新因式 U。
        foreach π ∈ Π_t do
            由公式 (2-14) 更新因式 V_π。
        由公式 (2-12) 计算目标函数值 O_2^(t)。
    由公式 (2-11) 预测目标领域样例的类别为 ŷ(x_{*i}^π) = f(x_{*i}^π)。
```

2.3.2　矩阵三分解

以非负矩阵三分解 NMTF[76] 为基础，框架 (2-10) 可形式化为学习模型 GTL_3：

$$\begin{aligned} O_3 = & \sum\nolimits_{\pi\in\Pi} \left\|\mathbf{X}_\pi - \mathbf{U}_\pi\mathbf{H}\mathbf{V}_\pi^{\mathrm{T}}\right\|_F^2 + \frac{\sigma}{2}\sum\nolimits_{\pi\in\Pi}\left\|\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi - \mathbf{I}\right\|_F^2 \\ & + \lambda\sum\nolimits_{\pi\in\Pi}\mathrm{tr}\left(\mathbf{U}_\pi^{\mathrm{T}}\mathbf{L}_\pi^u\mathbf{U}_\pi\right) + \gamma\sum\nolimits_{\pi\in\Pi}\mathrm{tr}\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{L}_\pi^v\mathbf{V}_\pi\right) \end{aligned} \tag{2-15}$$

优化问题 (2-15) 也可由交互优化算法求解，如下面的定理所述。其正确性证明与 GTL_2 类似，限于篇幅这里从略。完整的训练过程总结如算法 2 所示。

定理 2.2：　分别由公式 (2-16)~(2-18) 交互迭代式地更新 $\{\mathbf{U}_\pi\}_{\pi\in\Pi}$, $\{\mathbf{V}_\pi\}_{\pi\in\Pi_t}$, $\mathbf{H}$ 可以保证目标函数 (2-15) 单调递减并收敛到局部最优解。

$$\mathbf{U}_\pi \leftarrow \mathbf{U}_\pi \odot \frac{\left[\mathbf{X}_\pi\mathbf{V}_\pi\mathbf{H}^{\mathrm{T}} + \lambda\mathbf{W}_\pi^u\mathbf{U}_\pi\right]}{\left[\mathbf{U}_\pi\mathbf{H}\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\mathbf{H}^{\mathrm{T}} + \lambda\mathbf{D}_\pi^u\mathbf{U}_\pi\right]} \tag{2-16}$$

$$\mathbf{V}_\pi \leftarrow \mathbf{V}_\pi \odot \frac{\left[\mathbf{X}_\pi^{\mathrm{T}}\mathbf{U}_\pi\mathbf{H} + \gamma\mathbf{W}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\right]}{\left[\mathbf{V}_\pi\mathbf{H}^{\mathrm{T}}\mathbf{U}_\pi^{\mathrm{T}}\mathbf{U}_\pi\mathbf{H} + \gamma\mathbf{D}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right]} \tag{2-17}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\left[\sum_{\pi\in\Pi}\mathbf{U}_\pi^{\mathrm{T}}\mathbf{X}_\pi\mathbf{V}_\pi\right]}{\left[\sum_{\pi\in\Pi}\mathbf{U}_\pi^{\mathrm{T}}\mathbf{U}_\pi\mathbf{H}\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right]} \tag{2-18}$$

其中 $\odot$ 和 $\frac{[\cdot]}{[\cdot]}$ 分别表示点乘和点除操作（逐项乘法和除法）。

计算复杂度：算法计算代价包括：乘法更新公式 $O\left(\sum_{\pi\in\Pi} Tc\left(mn_\pi + m^2 + n_\pi^2\right)\right)$，构造邻接图 $O\left(\sum_{\pi\in\Pi}\left(mn_\pi^2 + m^2n_\pi\right)\right)$，其余步骤 $O\left(\sum_{\pi\in\Pi} mn_\pi\right)$。总计算复杂度为 $O\left(\sum_{\pi\in\Pi} Tc\left(mn_\pi + m^2 + n_\pi^2\right) + m^2n_\pi + mn_\pi^2\right)$，针对稀疏输入数据还可大幅度下降。

2.3.3　平凡解问题

现有图正则化非负矩阵分解方法[72,80,82] 的一个重要缺陷是可能产生平凡解[83]：即当 $\gamma\to\infty$，公式 (2-12) 的第四项支配了整个目标函数，优化问题退化为：

$$O_2' = \sum_{\pi\in\Pi}\mathrm{tr}\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{L}_\pi^v\mathbf{V}_\pi\right) = \sum_{\pi\in\Pi}\sum_{k=1}^{c}\mathbf{v}_{*k}^{\pi\mathrm{T}}\mathbf{L}_\pi^v\mathbf{v}_{*k}^\pi \tag{2-19}$$

公式 (2-19) 可分解为 $c\,|\Pi|$ 个相互独立的子问题：$O_2'' = \mathbf{v}_{*k}^{\pi\mathrm{T}}\mathbf{L}_\pi^v\mathbf{v}_{*k}^\pi$。每个子问题都导致相同的解（仅差一个标量），即 $\mathbf{v}_{*1}^\pi \propto \ldots \propto \mathbf{v}_{*c}^\pi$。因此样例类别矩阵 $\mathbf{V}_\pi$ 倾向于把所有样例都分配到同一个类别中，这显然是不合理的。文献[83] 对 $\mathbf{V}_\pi$ 增加归一化割集（Normalized-Cut）风格的正交约束，然后利用拉格朗日乘子法解一个带正交

算法 2: GTL_3：基于 NMTF 的 GTL 迁移学习模型

输入: 输入数据 $\{\mathbf{X}_\pi\}_{\pi\in\Pi}$, $\{\mathbf{Y}_\pi\}_{\pi\in\Pi_s}$；模型参数 $p, \lambda, \gamma, \sigma$；迭代次数 T。

输出: 隐含因式 $\{\mathbf{U}_\pi, \mathbf{V}_\pi\}_{\pi\in\Pi}$, $\mathbf{H}$, 分类结果 $\{\widehat{\mathbf{Y}}_\pi\}_{\pi\in\Pi_t}$。

开始

　分别由公式 (2-5) 和 (2-7) 构造邻接矩阵 $\mathbf{W}_\pi^v$ 和 $\mathbf{W}_\pi^u$。

　初始化 $\{\mathbf{U}_\pi\}_{\pi\in\Pi}$, $\mathbf{H}$; $\mathbf{V}_\pi \leftarrow \mathbf{Y}_\pi$, $\forall\pi \in \Pi_s$; $\mathbf{V}_\pi \leftarrow \text{LR}\left(\bigcup_{\pi'\in\Pi_s}\{\mathbf{X}_{\pi'}, \mathbf{Y}_{\pi'}\}\right)$, $\forall\pi \in \Pi_t$。

　for $t \leftarrow 1$ **to** T **do**

　　foreach $\pi \in \Pi$ **do**

　　　由公式 (2-16) 更新因式 $\mathbf{U}_\pi$。

　　　if $\pi \in \Pi_t$ **then**

　　　　由公式 (2-17) 更新因式 $\mathbf{V}_\pi$。

　　由公式 (2-18) 更新因式 $\mathbf{H}$。

　　由公式 (2-15) 计算目标函数值 $O_3^{(t)}$。

　由公式 (2-11) 预测目标领域样例的类别为 $\widehat{y}(\mathbf{x}_{*i}^\pi) = f\left(\mathbf{x}_{*i}^\pi\right)$。

约束的非凸优化问题。但该方法迭代过程中目标函数值会发生大幅度抖动，不能获得稳定的收敛性能。本章通过正交正则化方法，将正交约束项作为目标函数中的一个正则项，从而获得了稳定的收敛性能，解决了已有方法存在的平凡解问题。通常 $\lambda \in [0, 1] \ll \infty$ 取较小值，平凡解问题不存在，因而不用为 C_u 设置正交约束。

2.3.4　负迁移问题

首先，GTL_2 处于过迁移一端。该方法在领域间共享了所有隐含语义 $\mathbf{U} \in \mathbb{R}^{m\times c}$，这相当于共享了大量参数用以迁移知识，因而可能会导致负迁移，即从辅助领域迁移过多的知识结构更可能与目标领域内在结构不一致。在此情况下，本章提出的图正则化可以最大限度保持目标领域内的几何结构，从而有效地反制了负迁移。

反之，GTL_3 则处于欠迁移一端。该方法在领域间仅共享了结构关联信息 $\mathbf{H} \in \mathbb{R}^{c\times c}$，这相当于仅共享了少量参数用以知识迁移，这在领域间公共知识很少时能具有足够的鲁棒性，但在通常情况可能会导致低效迁移，即没有从辅助领域迁移足够的知识用于提高目标领域的学习任务。在此情况下，本章提出的图正则化可以增强似然最大化与几何一致性等准则的相互提升作用，从而强化有效迁移。

2.3.5　正确性分析

2.3.5.1　优化问题解

由带约束优化方法可求解公式 (2-12) 的 GTL_2 优化问题，一般采用控制变量法，即求解某一变量的迭代公式、并固定其他变量，上述过程交互迭代直到收敛。

记 $\mathbf{\Phi}$ 和 $\mathbf{\Psi}_\pi$ 分别为非负约束 $\mathbf{U} \geq \mathbf{0}$ 和 $\mathbf{V}_\pi \geq \mathbf{0}, \forall \pi \in \Pi$ 的拉格朗日乘子，则公式 (2-12) 的带约束目标函数的拉格朗日函数形式如下：

$$L = O_2 + \mathrm{tr}\left(\mathbf{\Phi}\mathbf{U}^{\mathrm{T}}\right) + \sum\nolimits_{\pi\in\Pi} \mathrm{tr}\left(\mathbf{\Psi}_\pi \mathbf{V}_\pi^{\mathrm{T}}\right)$$

为求取局部极值，将函数 L 相对于因式 $\mathbf{U}$ 和 $\mathbf{V}_\pi$ 的偏导数矩阵推导如下：

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{U}} = & -2\sum\nolimits_{\pi\in\Pi} \mathbf{X}_\pi \mathbf{V}_\pi + 2\sum\nolimits_{\pi\in\Pi} \mathbf{U}\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi \\ & + 2\sum\nolimits_{\pi\in\Pi} \lambda \mathbf{D}_\pi^u \mathbf{U} - 2\sum\nolimits_{\pi\in\Pi} \lambda \mathbf{W}_\pi^u \mathbf{U} + \mathbf{\Phi}\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{V}_\pi} = & -2\mathbf{X}_\pi^{\mathrm{T}}\mathbf{U} + 2\mathbf{V}_\pi \mathbf{U}^{\mathrm{T}}\mathbf{U} + 2\gamma \mathbf{D}_\pi^v \mathbf{V}_\pi \\ & - 2\gamma \mathbf{W}_\pi^v \mathbf{V}_\pi + 2\sigma \mathbf{V}_\pi \mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi - 2\sigma \mathbf{V}_\pi + \mathbf{\Psi}_\pi\end{aligned}$$

由 Karush-Kuhn-Tucker (KKT) 互补性条件[84] $\mathbf{\Phi} \odot \mathbf{U} = \mathbf{0}, \mathbf{\Psi}_\pi \odot \mathbf{V}_\pi = \mathbf{0}$，可得

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{U}} \odot \mathbf{U} = & \left(\sum\nolimits_{\pi\in\Pi} \mathbf{U}\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi + \sum\nolimits_{\pi\in\Pi} \lambda \mathbf{D}_\pi^u \mathbf{U}\right) \odot \mathbf{U} \\ & - \left(\sum\nolimits_{\pi\in\Pi} \mathbf{X}_\pi \mathbf{V}_\pi + \sum\nolimits_{\pi\in\Pi} \lambda \mathbf{W}_\pi^u \mathbf{U}\right) \odot \mathbf{U} = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{V}_\pi} \odot \mathbf{V}_\pi = & \left(\mathbf{V}_\pi \mathbf{U}^{\mathrm{T}}\mathbf{U} + \gamma \mathbf{D}_\pi^v \mathbf{V}_\pi + \sigma \mathbf{V}_\pi \mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right) \odot \mathbf{V}_\pi \\ & - \left(\mathbf{X}_\pi^{\mathrm{T}}\mathbf{U} + \gamma \mathbf{W}_\pi^v \mathbf{V}_\pi + \sigma \mathbf{V}_\pi\right) \odot \mathbf{V}_\pi = \mathbf{0}\end{aligned}$$

将正号项变量作为更新值，可以推导得到如公式 (2-13)~(2-14) 所示的更新规则。

2.3.5.2　收敛性分析

与文献[72,74] 类似，本章采用*辅助函数法*来分析定理 2.1 的收敛性。简明起见，仅证明公式 (2-12) 的目标函数值 O_2 随公式 (2-14) 对因式 $\mathbf{V}_\pi$ 的更新而单调下降，其他更新公式及定理 2.2 的收敛性也可类似证明。为此，首先引入辅助函数的定义。

定义 2.1：*[74]* $A(z, \tilde{z})$ *是* $F(z)$ *的辅助函数，如果如下条件*

$$A(z, \tilde{z}) \geq F(z) \; and \; A(z, z) = F(z)$$

对任意给定的 $z, \tilde{z}$ 均满足。

引理 2.1：[74] 如果 A 是 F 的辅助函数，则 F 在如下更新规则下保持单调下降

$$z^{(t+1)} = \arg\min_z A(z, z^{(t)})$$

证明

$$F(z^{(t+1)}) \le A(z^{(t+1)}, z^{(t)}) \le A(z^{(t)}, z^{(t)}) = F(z^{(t)})$$

□

接下来，证明公式 (2-14) 恰好是引理 2.1 在特定辅助函数下的更新规则。对变量 $\mathbf{V}_\pi$（在控制变量法中，其他变量暂且视为常量）的任意元素 v_{ij}，用 F_{ij} 表示目标函数 O_2 中仅与 v_{ij} 相关的部分，则 F_{ij} 相对于 v_{ij} 的一阶、二阶导数分别推导为

$$F'_{ij} = \left(\frac{\partial O_2}{\partial \mathbf{V}_\pi}\right)_{ij}$$

$$F''_{ij} = 2\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)_{jj} + 2\gamma(\mathbf{D}_\pi^v - \mathbf{W}_\pi^v)_{ii} + 2\sigma\left(\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{jj} + v_{jj}^2 - 1\right)$$

引理 2.2：　函数

$$\begin{aligned} A\left(v, v_{ij}^{(t)}\right) = {} & F_{ij}\left(v_{ij}^{(t)}\right) + F'_{ij}\left(v_{ij}^{(t)}\right)\left(v - v_{ij}^{(t)}\right) \\ & + \frac{\left(\mathbf{V}_\pi\mathbf{U}^{\mathrm{T}}\mathbf{U} + \gamma\mathbf{D}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{ij}}{v_{ij}^{(t)}}\left(v - v_{ij}^{(t)}\right)^2 \end{aligned}$$

是 $F_{ij}(v)$ 的一个合理的辅助函数。

证明　可以直接验证 $A(v, v) = F_{ij}(v)$，因此仅需再证明 $A\left(v, v_{ij}^{(t)}\right) \ge F_{ij}(v)$，为此将 $F_{ij}(v)$ 用泰勒级数展开，得到

$$\begin{aligned} F_{ij}(v) = {} & F_{ij}\left(v_{ij}^{(t)}\right) + F'_{ij}\left(v_{ij}^{(t)}\right)\left(v - v_{ij}^{(t)}\right) + \left(v - v_{ij}^{(t)}\right)^2 \\ & \times \left(\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)_{jj} + \gamma(\mathbf{D}_\pi^v - \mathbf{W}_\pi^v)_{ii} + \sigma\left(\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{jj} + v_{jj}^2 - 1\right)\right) \end{aligned}$$

由正交性有 $1 \approx \left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{jj} = \sum_i v_{ij}^2 \gg v_{jj}^2$，根据线性代数，有如下三个不等式成立：

$$\left(\mathbf{V}_\pi\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)_{ij} = \sum\nolimits_l v_{il}^{(t)}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)_{lj} \ge v_{ij}^{(t)}\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)_{jj}$$

$$(\mathbf{D}_\pi^v\mathbf{V}_\pi)_{ij} = \sum\nolimits_l (\mathbf{D}_\pi^v)_{il}v_{lj}^{(t)} \ge v_{ij}^{(t)}(\mathbf{D}_\pi^v - \mathbf{W}_\pi^v)_{ii}$$

$$\begin{aligned}\left(\mathbf{V}_\pi\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{ij} &= \sum_l v_{il}^{(t)}\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{lj} \geq v_{ij}^{(t)}\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{jj} \\ &\geq v_{ij}^{(t)}\left(\left(\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\right)_{jj} + v_{jj}^2 - 1\right)\end{aligned}$$

综合比较上述三个不等式左右端，可得 $A\left(v, v_{ij}^{(t)}\right) \geq F_{ij}(v)$，如此引理 *2.2* 获证。□

定理 2.1 证明： 根据引理 2.1 和引理 2.2，因式 $\mathbf{V}_\pi$ 的更新规则可由最小化辅助函数 $A\left(v_{ij}^{(t+1)}, v_{ij}^{(t)}\right)$ 得到。设置 $\frac{\partial A\left(v_{ij}^{(t+1)}, v_{ij}^{(t)}\right)}{\partial v_{ij}^{(t+1)}} = 0$，可得

$$\begin{aligned}v_{ij}^{(t+1)} &= v_{ij}^{(t)} - \frac{v_{ij}^{(t)} F'_{ij}\left(v_{ij}^{(t)}\right)}{2(\mathbf{V}_\pi\mathbf{U}^{\mathrm{T}}\mathbf{U} + \gamma\mathbf{D}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi)_{ij}} \\ &= v_{ij}^{(t)} \frac{\left(\mathbf{X}_\pi^{\mathrm{T}}\mathbf{U} + \gamma\mathbf{W}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\right)_{ij}}{(\mathbf{V}_\pi\mathbf{U}^{\mathrm{T}}\mathbf{U} + \gamma\mathbf{D}_\pi^v\mathbf{V}_\pi + \sigma\mathbf{V}_\pi\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi)_{ij}}\end{aligned}$$

该更新规则与公式 (2-14) 一致，其他更新规则也可类似证明。在每次迭代里，更新因式 $\mathbf{V}_\pi$ 都满足 $O_2\left(\mathbf{V}_\pi^{(0)}\right) = A\left(\mathbf{V}_\pi^{(0)}, \mathbf{V}_\pi^{(0)}\right) \geq A\left(\mathbf{V}_\pi^{(1)}, \mathbf{V}_\pi^{(0)}\right) \geq A\left(\mathbf{V}_\pi^{(1)}, \mathbf{V}_\pi^{(1)}\right) = O_2\left(\mathbf{V}_\pi^{(1)}\right) \geq \ldots \geq O_2\left(\mathbf{V}_\pi^{(T)}\right)$。因此，目标函数值 $O_2(\mathbf{V}_\pi)$ 在迭代过程中能够保持单调递减。同时，注意到目标函数 (2-12) 有下界 0。综上所述，定理 2.1 的收敛性获证。

2.4　实验过程与结果

本节在两类实际应用（文本分类、图像识别）数据集上进行系统性的实验，证明 GTL 框架和学习模型的有效性，实验性相关的数据集和代码均可从公网下载。为论述简明性，在没有歧义情况下后文用 GTL 统一指代 GTL_2 和 GTL_3 算法。

2.4.1　实验数据

2.4.1.1　文本数据

按照迁移学习文献[29,45,47,60,63,67,85] 介绍的通用协议，本章考察被广泛采用的 *20-Newsgroups* 文本数据集，并根据其层次结构生成 216 个跨领域文本分类任务。

20-Newsgroups① 数据集包含约 20,000 个文档，4 个大类分别为 *comp*、*rec*、*sci* 和 *talk*，每个大类包含 4 个子类，详细信息如表 2.2 所示。在实验中构造了 6 组跨领域二分类任务，每组任务由 4 个大类中随机选取 2 个大类构成，一个大类记为正例，另一个大类记为负例，6 个任务组具体为 *comp vs rec*、*comp vs sci*、*comp vs talk*、*rec vs sci*、*rec vs talk* 和 *sci vs talk*。每个跨领域分类任务（包括辅助领域和目标领域）采用文献[67] 介绍的方法生成：每个任务组 *P vs Q* 的两个大类 *P* 和

① http://people.csail.mit.edu/jrennie/20newsgroups

表 2.2　文本数据集 20-Newsgroups 的层次结构和统计信息。

数据集	大类	子类	样例数	特征数
20-Newsgroups	comp	comp.graphics	970	25,804
		comp.os.ms-windows.misc	963	
		comp.sys.ibm.pc.hardware	979	
		comp.sys.mac.hardware	958	
	rec	rec.autos	987	
		rec.motorcycles	993	
		rec.sport.baseball	991	
		rec.sport.hokey	997	
	sci	sci.crypt	989	
		sci.electronics	984	
		sci.med	987	
		sci.space	985	
	talk	talk.politics.guns	909	
		talk.politics.mideast	940	
		talk.politics.misc	774	
		talk.religion.misc	627	

表 2.3　字符、人脸、对象等图像数据集的统计信息。

数据集	类型	样例数	特征数	类别数	包含子集
USPS	字符	1,800	256	10	USPS
MNIST	字符	2,000	256	10	MNIST
PIE	人脸	11,554	1,024	68	PIE1, PIE2
MSRC	对象	1,269	240	6	MSRC
VOC2007	对象	1,530	240	6	VOC

Q 分别包含 4 个子类 P_1、P_2、P_3、P_4 和 Q_1、Q_2、Q_3、Q_4；随机选取 P 的两个子类（如 P_1 、P_2）与 Q 的两个子类（如 Q_1 、Q_2）构成辅助领域，其余子类（P 的 P_3、P_4 和 Q 的 Q_3、Q_4）构成目标领域。以上构造策略既保证辅助领域和目标领域是相关的，因为它们都来自同样的大类；又保证辅助领域和目标领域是不同的，因为它们来自不同的子类。每个任务组 P *vs* Q 可以生成 $C_4^2 \cdot C_4^2 = 36$ 个分类任务，总计 6 个任务组共生成 $6 \cdot 36 = 216$ 个分类任务。数据集经过文本预处理后包含 25,804 个词项特征和 15,033 个文档，每个文档由 *tf-idf* 向量表征，如表 2.2 所示。

2.4.1.2　图像数据

本章在如下图像集上测试各种算法的效能：字符集 USPS 和 MNIST，人脸集 Multi-PIE，对象集 MSRC 和 VOC2007 ，统计信息如表 2.3，示例如图 2.3 。

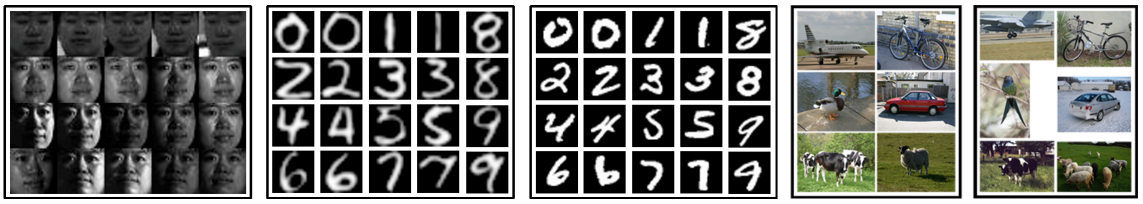


图 2.3　人脸集 Multi-PIE，字符集 USPS、MNIST，对象集 MSRC、VOC2007 的示例。

USPS① 数据集包括 7,291 张训练图片和 2,007 张测试图片，图片大小为 16×16。**MNIST**② 数据集包括 60,000 张训练图片和 10,000 张测试图片，图片大小 28×28。从图 2.3 可以直观感受到，USPS 和 MNIST 数据集分别服从显著不同的概率分布。两个数据集都包含 10 个类别，每个类别是 1–10 之间的某个字符。在实验中通过如下方式构造分类任务 *USPS vs MNIST*：在 USPS 中随机选取 1,800 张图片作为辅助数据、在 MNIST 中随机选取 2,000 张图片作为目标数据。交换辅助领域和目标领域可以得到另一个分类任务 *MNIST vs USPS*。图片预处理包括：将所有图片大小线性缩放为 16×16，每幅图片用 256 维的特征向量表征，编码了图片的像素灰度值信息。辅助领域和目标领域共享特征空间和类别空间，但数据分布显著不同。

PIE③ 代表“朝向、光照、表情”的英文单词首字母，该数据集是人脸识别的基准测试集，包括 68 个不同人物的 41,368 幅人脸照片，图片大小为 32×32，每个人物的照片由 13 个同步的相机（不同朝向）、21 个不同曝光程度拍摄。简单起见，实验中采用 PIE 的预处理集④，包括 2 个不同子集 **PIE1**[72] 和 **PIE2**[86]，是从正面朝向的人脸照片集合（C27）中按照不同的关照和曝光条件随机选出。按如下方法构造分类任务 *PIE1 vs PIE2*：将 **PIE1** 作为辅助领域、**PIE2** 作为目标领域；交换辅助领域和目标领域可以得到分类任务 *PIE2 vs PIE1*。这样，辅助领域和目标领域分别由不同光照、曝光条件的人脸照片组成，从而服从显著不同的概率分布。

MSRC⑤ 数据集包括 18 个对象类别共 4,323 张图片。**VOC2007**⑥ 数据集包括 20 个概念类别共 5,011 张图片。由于 MSRC 来自标准化图片集，而 VOC2007 来自 Flickr⑦ 网站的日常生活照片，两者所摄取的实物对象并不相同，因此 MSRC 和 VOC 服从显著不同的概率分布。两个数据集共享以下 6 个类别：“aeroplane”, “bicycle”, “bird”, “car”, “cow”, “sheep”。按如下方法构造分类任务 *MSRC vs VOC*：将 **MSRC** 中的 1,269 张图片作为辅助领域，将 **VOC2007** 中的 1,530 张图片作为目

① http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html
② http://yann.lecun.com/exdb/mnist
③ http://vasc.ri.cmu.edu/idb/html/face
④ http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
⑤ http://research.microsoft.com/en-us/projects/objectclassrecognition
⑥ http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007
⑦ http://www.flickr.com

标领域；交换辅助领域和目标领域可以得到分类任务 *VOC vs MSRC*。将所有图片均匀缩放到长边 256 像素，并以 5 像素网格抽取 128-维稠密 SIFT（DSIFT）[87] 局部特征。由 K 均值聚类得到 240-维视觉词表，并可将每张图片表征为 240-维的词袋直方图向量。这样，辅助领域和目标领域服从同一特征空间中的不同概率分布。

2.4.2　基准算法和实现细节

2.4.2.1　基准方法

本章考察 GTL 及如下 9 种迁移学习方法在文本、图像分类任务的准确率：

- 逻辑斯特回归（Logistic Regression，LR）
- 支持向量机（Support Vector Machine，SVM）
- 拉普拉斯支持向量机（Laplacian Support Vector Machine，LapSVM）[88]
- 谱特征对齐（Spectral Feature Alignment，SFA）[42]
- 迁移主成份分析（Transfer Component Analysis，TCA）[45]
- 联合矩阵分解（Collective Matrix Factorization，CMF）[77]
- 标签传播（Label Propagation，LP）[68]
- 矩阵三分解聚类（Matrix Tri-Factorization Clustering，MTrick）[69]
- 图正则化联合矩阵分解（GCMF，基于归一化约束的 GTL 方法）[59]

CMF 和 GTL_2 基于矩阵二分解，LP、MTrick、GCMF 和 GTL_3 基于矩阵三分解。

2.4.2.2　实现细节

根据文献[1,45,69] 介绍的评测协议，LR 和 SVM 在辅助领域标注数据上训练、并在目标领域无标数据上测试；SFA 和 TCA 在所有数据上学习抽象特征表示，然后基于该特征表示由辅助领域数据训练一个 LR 分类器并推广到目标领域数据；LapSVM、CMF、LP、MTrick、GCMF 和 GTL 在所有数据上直推式训练迁移模型。

经典交叉验证在目标领域没有标注数据时无法自动选择最优模型参数。本章与相关文献一样，在所有 222 个分类任务上测试 9 种基准算法，将每种算法在各种参数设置下的最佳效果用于性能对比。具体地，LR 采用 LIBLINEAR① 工具包实现，SVM 采用 LIBSVM② 工具包实现，正则参数 C（$=1/2\sigma$) 通过遍历 $C \in \{0.1, 0.5, 1, 5, 10, 50, 100\}$ 设置；LapSVM 采用文献[88] 的实现③，正则参数 γ_A 和 γ_I 分别通过遍历 $\gamma_A, \gamma_I \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ 设置；SFA、TCA、LP、MTrick 都采用文献作者的实现，子空间维度 k 通过遍历 $k \in \{4, 8, 16, 32, 64, 128, 256\}$ 设置。

① www.csie.ntu.edu.tw/~cjlin/liblinear/
② http://www.csie.ntu.edu.tw/~cjlin/libsvm
③ http://vikas.sindhwani.org/manifoldregularization.html

表 2.4　在 20-Newsgroups6 个跨领域分类任务组（各含 36 个任务）的平均准确率 (%) 。

任务组	标准学习			迁移学习					本章方法		
	LR	SVM	LapSVM	SFA	TCA	LP	CMF	MTrick	GCMF	GTL_3	GTL_2
C vs R	88.37	87.51	81.93	89.73	95.12	95.21	95.73	95.96	97.72	**98.19**	**98.05±0.00**
C vs S	77.87	75.38	68.96	78.07	77.32	85.75	87.73	86.90	88.35	86.98	**91.43±0.00**
C vs T	96.31	95.44	95.40	95.85	97.20	96.89	97.11	97.77	98.25	**98.48**	**98.35±0.00**
R vs S	75.28	73.82	74.21	79.25	82.31	85.45	86.40	87.22	93.02	95.18	**95.95±0.03**
R vs T	82.28	83.27	87.44	86.98	86.58	94.16	94.89	94.33	97.70	**98.28**	**98.07±0.00**
S vs T	76.99	76.85	80.22	79.27	79.30	86.37	88.37	90.21	**96.17**	**96.32**	95.64±0.01
平均	82.85	82.05	81.36	84.86	86.31	90.64	91.70	92.06	95.20	95.57	**96.25±0.01**

本章方法 GTL 包括 4 个可调参数：图近邻数 p，正则项参数 λ、γ 和 σ。通过参数敏感实验，GTL 可在相当大参数范围内获得稳定性能。在比较实验中，固定 $p=10$、$\gamma=10$、$\sigma=100$ 并设置：（1）文本数据 $\lambda=0$，（2）图像数据 $\lambda=0.1$ 。此外，迭代次数设置为 $T=100$，对随机初始化算法分别执行 10 次取平均准确率。

本章采用目标领域无标测试数据上的准确率（Accuracy）作为评价指标：

$$Accuracy=\frac{\left|\mathbf{x}:\mathbf{x}\in\bigcup_{\pi\in\Pi_t}\mathcal{D}_\pi\wedge f(\mathbf{x})=y(\mathbf{x})\right|}{\left|\mathbf{x}:\mathbf{x}\in\bigcup_{\pi\in\Pi_t}\mathcal{D}_\pi\right|} \tag{2-20}$$

其中 $y(\mathbf{x})$ 是测试样例 $\mathbf{x}$ 的真实标签，$f(\mathbf{x})$ 是待测学习算法为样例 $\mathbf{x}$ 预测的标签。

2.4.3　实验结果

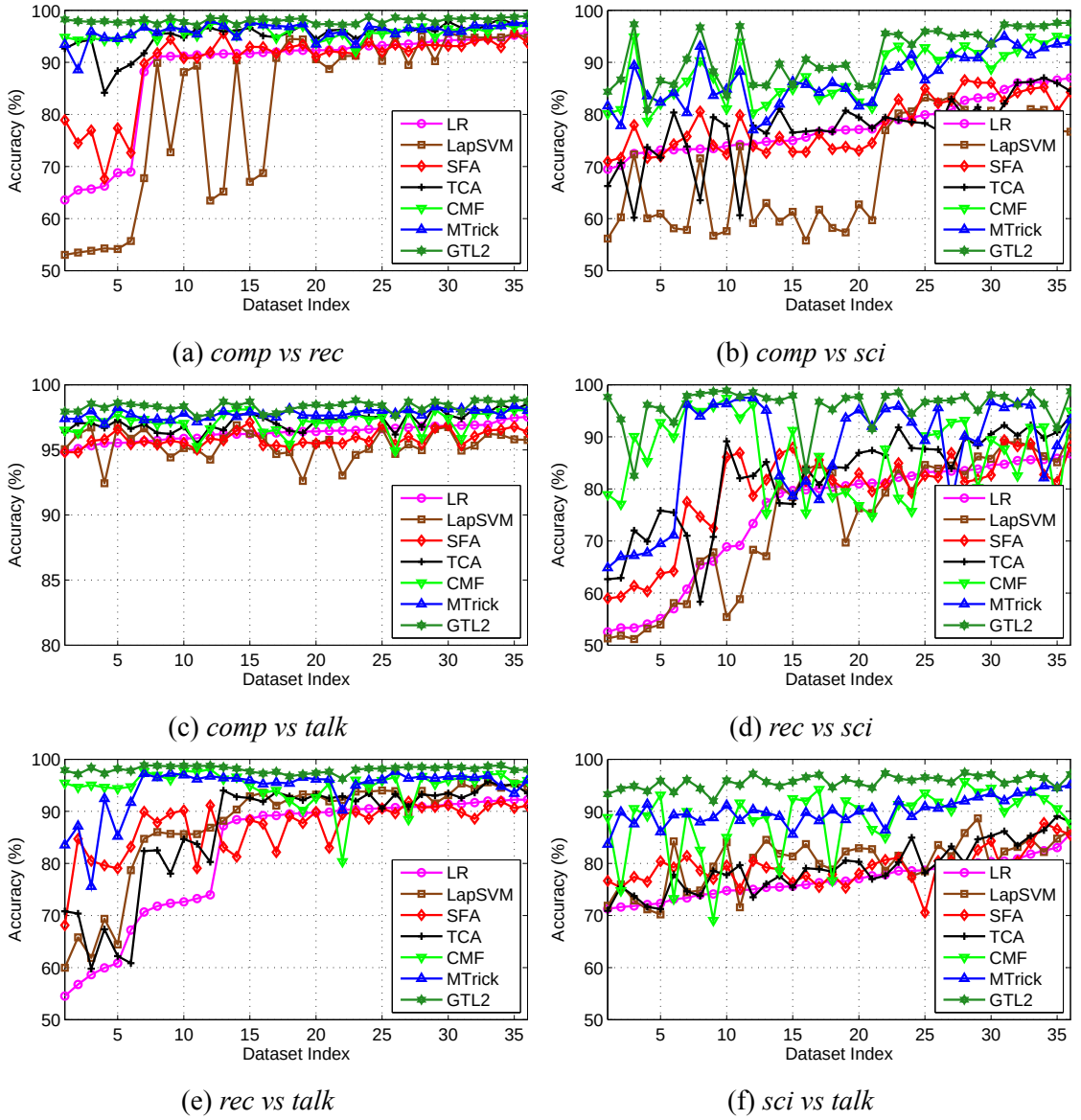
本节给出 GTL 和 9 种基准方法的对比实验结果，在各分类任务上的准确率。

2.4.3.1　文本分类

GTL 和 9 种基准方法在 6 个跨领域分类任务（共 216 个任务）上的平均准确率如表 2.4 所示，每个任务上的分类准确率如图 2.4(a)~2.4(f) 所示，其中每个子图给出每个任务组的 36 个任务的分类准确率，任务序号按照逻辑斯特回归（LR）的准确率从低到高排列，从而反映了跨领域知识迁移难度。从结果可得到如下结论。

GTL 比 8 个基准方法获得了具有统计显著性的性能提升，并且基于矩阵二分解的 GTL_2 比基于矩阵三分解的 GTL_3 性能稍好。GTL_2 在所有 216 个任务的平均准确率达到 **96.25%**，比最佳基准方法 MTrick 准确率提升了 **4.19%**、错误率下降了 **52.78%**，是当前已发表的最好成果。此外，表 2.4 给出了算法执行 10 次的平均准确率和标准差，可以看到标准差小于 0.01%，表明 GTL 对随机初始化十分稳定。

其次，观察到迁移学习方法一般比标准学习方法能获得更好的分类准确率。标准学习方法的主要局限在于将辅助领域和目标领域看成是独立同分布的，但在



(a) *comp vs rec*　(b) *comp vs sci*

(c) *comp vs talk*　(d) *rec vs sci*

(e) *rec vs talk*　(f) *sci vs talk*

图 2.4　LR、LapSVM、SFA、TCA、CMF、MTrick 和 GTL_2 在文本分类上的准确率。

实际跨领域任务中，这个独立同分布假设并不成立，因而导致了非常不理想的分类效果。需要注意的是，公认十分有效的图正则化半监督学习方法 LapSVM 在该数据集上也未能取得比标准学习方法 LR 和 SVM 更好的效果，这多少有些与经验不符。可能的原因是 LapSVM 没有抽取公共隐含语义结构来减小领域间的概率分布差异，因而直推式判别面无法与目标领域流形结构取得一致性，导致了负迁移。

再次，观察到 GTL 方法比迁移降维方法 SFA、TCA、LP、CMF 和 MTrick 获得了具有统计性的性能提高。现有迁移降维方法的一个局限在于不能同时对领域间知识结构和领域内几何结构进行权衡，从而导致领域间知识结构支配了学习模型在目标领域上的效能。在实际应用中，不同领域的不同特性（统计的、几何的）通常难以保持一致，从而导致较高的负迁移风险。GTL 方法专门针对上述风险设

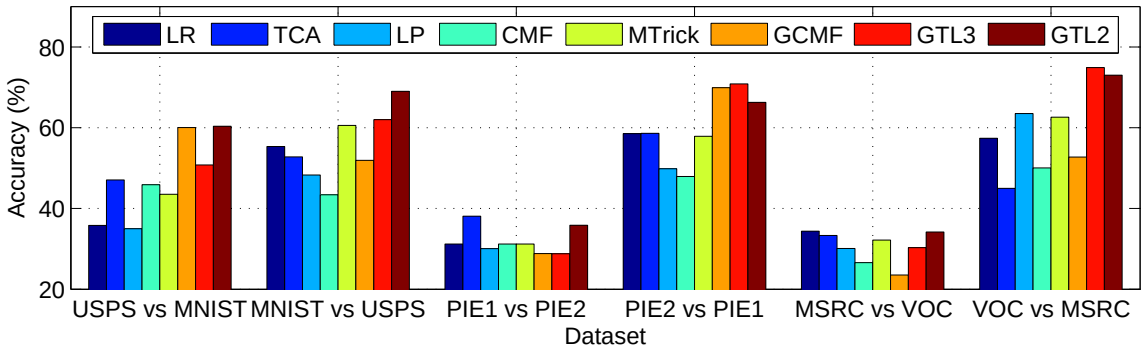


图 2.5　LR、TCA、LP、CMF、MTrick、GCMF 和 GTL 在图像识别任务上的准确率。

计了不同领域不同特性的“谈判”机制，从而有效地避免了欠迁移和负迁移问题。

最后，观察到在难以准确分类的任务上 GTL 方法取得了比基准方法更有效且更鲁棒的效果。可以从图 2.4(a)~2.4(f) 观察到以下结论：对那些 LR 仅取得极低准确率（低于 70%）的任务，GTL 方法相对于基准方法的准确率提升优势更加显著。

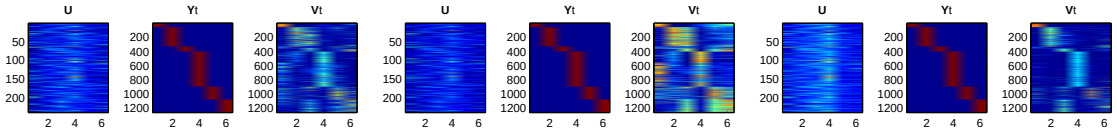
2.4.3.2　图像分类

GTL_2, GTL_3 和 6 个基准方法在 6 个图像识别任务上的准确率如图 2.5 所示。由于 SFA 不能处理非稀疏数据，LapSVM 不支持多类分类，因此两者不参与比较。

可以观察到，GTL_2 和 GTL_3 在绝大多数任务上显著地超越了所有基准方法，不管基础模型是矩阵二分解 NMF 还是矩阵三分解 NMTF，图正则化都能提升 GTL 的跨领域知识迁移效能。不过，GTL_3 的性能比 GTL_2 有明显差距，这说明在 GTL 框架下 NMF 一般会比 NMTF 更为有效，这是因为 GTL 在跨领域知识迁移更充分（甚至可能存在过迁移）的情况下效果更好（见第2.3.2节的讨论）。值得注意的是，GTL_3 比同样基于矩阵三分解 NMTF 的 GCMF 方法效果更好，这说明 GCMF 无法有效避免平凡解问题，而 GTL 通过正交正则化可有效地解决该问题。

出乎意料的是，部分迁移学习方法如 TCA、LP、CMF 和 MTrick 在部分分类任务如 *MNIST vs USPS* 和 *PIE2 vs PIE1* 上取得比标准监督学习 LR 更低的分类准确率。这就是迁移学习中十分具有挑战性的负迁移问题。在前述分类任务中，辅助领域和目标领域之间的数据分布差异如此之大，甚至不太可能抽取到所谓的公共隐含语义结构，也就无法有效构建知识迁移的桥梁。换句话说，基准方法赖以成立的公共隐含语义结构假设在这些挑战性任务上不再成立，从而导致了负迁移。

值得高兴的是，GTL 方法在这类挑战性任务上仍然能鲁棒运行，并有效地反制了负迁移。其根本原因是在 GTL 中，领域内几何结构被有效地保持下来，而不管领域间的“公共”隐含语义结构是否存在、是否能迁移过来。总之，如果领域间的语义结构与领域内的几何结构矛盾，那么 GTL 为它们建立了有效的“谈判”



(a) $\gamma = 0$　　(b) $\sigma = 0$　　(c) GTL

图 2.6　GTL_2 在分类任务 *VOC vs MSRC* 抽取的隐含因式结构 $\mathbf{U}$、$\mathbf{V}_t$ 及真实标注矩阵 $\mathbf{Y}_t$。

机制，确保领域内学习任务支配整个跨领域学习任务，从而有效地避免了负迁移。

2.4.4　负迁移分析

为了证明图正则化反制负迁移的效果，这里移除图正则化，即设置 $\lambda = \gamma = 0$。此时 GTL 抽取的隐含因式结构如图 2.6(a) 所示，其中暖色表示较大值、冷色表示较小值。注意到 $\mathbf{V}_t$ 的每一列表示一个样例类别，每一行表示样例在各个类别中的隶属概率。比较 $\mathbf{V}_t$ 和目标领域的真实标注 $\mathbf{Y}_t$，可观察到此时很多样例都被隶属到错误的类别中。其原因是移除图正则化后，相似的样例不再保证具有相似的标签，从而领域内的几何结构被领域间迁移的知识结构破坏（结构不一致性）。通过图正则化可以有效避免上述次优结果，也即避免了负迁移问题，如图 2.6(c) 所示。

其次，移除正交正则化，即设置 $\sigma = 0$，此时隐含因式结构如图 2.6(b) 所示。比较 $\mathbf{V}_t$ 和 $\mathbf{Y}_t$，可观察到很多样例都被隶属到多个（错误）类别中。其原因是移除正交正则化导致平凡解，该问题在现有图正则化矩阵分解中广泛存在。通过正交正则化可以有效避免上述平凡解，从而进一步避免负迁移问题，如图 2.6(c) 所示。

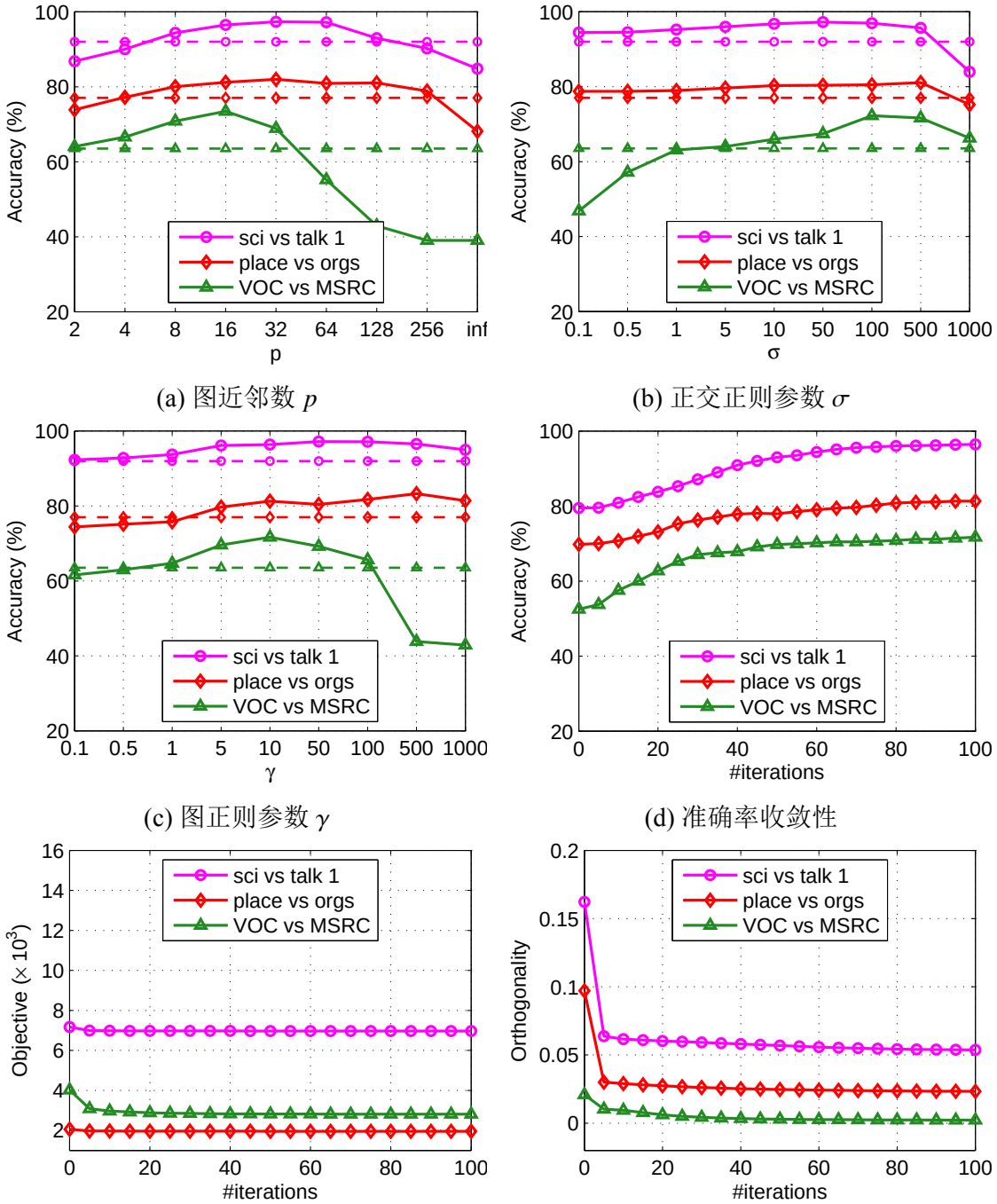
2.4.5　参数敏感性分析

本节通过参数敏感性实验，证明 GTL 在相当大参数范围内取得比基准方法更好的效果。限于空间仅在 *sci vs talk 1*、*people vs orgs* 和 *VOC vs MSRC* 任务上实验。

图近邻数 p： 以参数 p 的不同值执行 GTL_2。直观上，为了获得最佳性能 p 应取折中值，因为过大的 p 值会导致邻接图过于稠密（连接两个本不相似的样例或特征），过小的 p 值会导致邻接图过于稀疏（未连接两个十分相似的样例或特征）。分类准确率相对于 p 值的变化规律如图 2.7(a) 所示，可取 $p \in [4, 32]$ 为合理范围。

正交正则化 σ： 以参数 σ 的不同值执行 GTL_2。理论上，σ 控制了正交约束满足的程度，当 $\sigma \to 0$ 时 GTL 为非良性定义会导致平凡解；当 $\sigma \to \infty$ 时 GTL 被正则项支配而不能充分拟合训练数据。分类准确率相对于 σ 值的变化规律如图 2.7(b) 所示，可取 $\sigma \in [1, 500]$ 为合理范围。实践中可根据 γ 值来确定 σ 取值，如取 $\sigma \in [\gamma, 10\gamma]$，判据是使得正交约束误差在 10% 以内（参见后文正交约束误差）。

图正则化 γ： 以参数 γ 的不同值执行 GTL_2。理论上，γ 控制了样例图正则化



(a) 图近邻数 p　(b) 正交正则参数 σ

(c) 图正则参数 γ　(d) 准确率收敛性

(e) 目标函数收敛性　(f) 正交约束收敛性

图 2.7　GTL_2 模型的参数敏感性、收敛性分析（虚线表示最佳基准方法的准确率）。

的强度，较大的 γ 值会使几何结构保持在学习过程中更为重要。当 $\gamma \to \infty$ 时仅有几何结构保持而标注信息丢失，导致次优解。分类准确率相对 γ 的变化规律如图 2.7(c) 所示，可取 $\gamma \in [1, 100]$ 为合理范围；实践中可由辅助领域聚类准确率选取。

收敛性分析： 由于 GTL 是迭代式算法且不存在全局最优解，需要通过实验验证其收敛速度和质量。分类准确率随着迭代次数的变化规律如图 2.7(d) 所示，目

标函数值随迭代次数的变化规律如图 2.7(e) 所示。可以看到随着迭代进行，分类准确率和目标函数分别稳定地递增和递减，并在迭代 100 次后收敛到局部最优解。

此外，由于本章采用正交正则化方法来满足正交约束，通常情况下正交约束不能完全满足，而会存在合理的正交约束误差，定义为 $\sum_{\pi\in\Pi}\left\|\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi-\mathbf{I}\right\|_F^2/\sum_{\pi\in\Pi}\|\mathbf{I}\|_F^2$，一般应在 10% 以内。正交约束误差随迭代次数的变化规律如图 2.7(f) 所示，可以看到，正交约束误差随着迭代次数不断递减并在迭代 100 次后收敛到局部最优解。

2.5　小结

本章提出了通用学习框架图正则化联合矩阵分解（GTL）以及三种迁移学习模型，用于实现跨领域分类任务并重点解决负迁移问题。具体地说，GTL 抽取领域间的公共隐含语义结构实现知识迁移，并通过保持领域内的几何结构反制负迁移。GTL 的一个重要优势是同时考察了数据属性的多个维度（统计属性、几何属性），而现有方法一般仅考虑单个属性。此外，多种矩阵分解方法如 NMF 和 NMTF 等，都可以直接载入 GTL 框架以实现有效迁移学习并避免负迁移。在 222 个文本、图像分类任务上的系统性实验证明了本章方法相对于前沿方法的优越性。

第 3 章　联合分布适配方法

上一章介绍了如何通过图正则化联合矩阵分解同步解决欠迁移和负迁移问题，但未能形式化地定义辅助领域和目标领域之间的概率分布差异并最小化该准则，因而难以从理论角度分析泛化误差上界。针对领域间概率分布的适配问题，在目标领域没有标注数据的无监督迁移学习中，已有方法通常仅适配了边缘概率分布（即无标数据的聚类结构），但这会导致领域间概率分布的欠适配问题，严重影响学习模型的泛化性能。本章提出了联合适配正则化框架，在结构风险最小化和正则化理论的支撑下，重点解决条件概率分布（即标注数据的判别结构）的适配问题。同时，为了结合半监督学习的优势，还集成了基于流形正则化的半监督学习方法。在此基础上，提出了基于正则化线性回归、支持向量机、主成份分析等四种学习模型，并通过可再生希尔伯特空间中的表出定理给出了模型的凸优化解。在文本分类和图像识别任务上的系统性实验证明了本章方法相对已有工作的优势。

3.1　引言

在迁移学习中，不同领域所服从的概率分布可能会有显著的差异。因此，如何减小辅助领域和目标领域间的概率分布差异，就成为迁移学习极为重要的基本问题。相关工作可分为两类：基于表征学习的方法和基于监督学习的方法。基于表征学习的方法，直接学习跨领域间的隐含特征表示、同时显式地最小化概率分布间的距离函数[40,45,46]。在此隐含特征表示下，标准监督学习方法可以直接在辅助领域进行训练、并在目标领域进行预测[40,46]。例如文献[40]提出了最大均值差异嵌入（Maximum Mean Discrepancy Embedding，MMDE）方法，其中用于度量概率分布差异的最大均值差异距离[33]被显式地最小化；文献[46]提出了迁移子空间学习（Transfer Subspace Learning，TSL）框架，其中用于度量概率分布差异的布雷格曼散度被显式地最小化。基于监督学习的方法，直接归纳一个自适应分类器、并将分布距离函数最小化作为模型正则项[85,89–91]。但是所有这些方法仅利用辅助领域标注数据来训练判别式模型，本章提出这些标注数据还可以用来进一步减小条件分布之间的差异；此外，这些方法仅利用目标领域无标数据来适配边缘概率分布，本章提出这些无标数据还可以用来进一步挖掘目标领域的半监督学习潜力。

本章之所以要在领域间对条件分布（标注数据的判别结构）进行适配，是因为在相当多的实际应用中，仅适配边缘分布（无标数据的聚类结构）并不能获得满足应用需求的迁移学习性能，这是因为辅助领域和目标领域之间的判别分类面

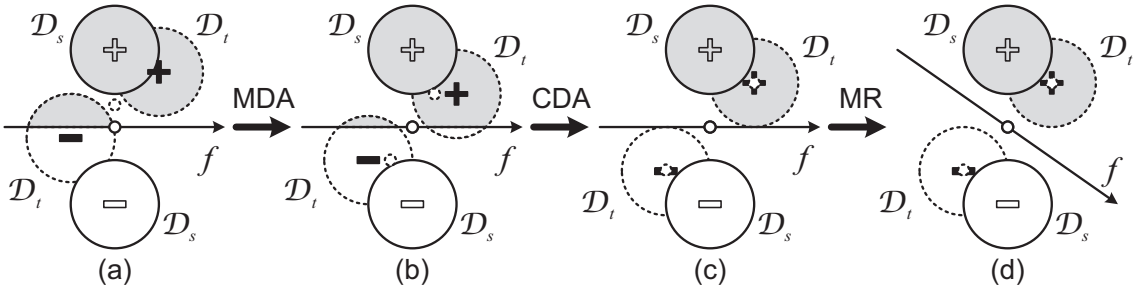


图 3.1　联合适配正则化工作原理示意图。f：判别超平面；$\mathcal{D}_s$：辅助领域；$\mathcal{D}_t$：目标领域；○：领域/类别中心；MDA：边缘分布适配；CDA：条件分布适配；MR：流形正则化。

也极有可能并不相同[24,45]。已有部分工作在此方向上进行了一些探索，主要是利用半监督学习的思想，通过目标领域无标数据揭示目标领域的判别结构[92–94]。例如文献[92] 提出了*领域适配支持向量机*（Domain Adaptation Support Vector Machine，DASVM），扩展了直推式支持向量机（TSVM）用于步进式地标注目标领域相关数据并移除辅助领域无关数据；文献[93] 提出了*隐性直推式迁移学习*（Latent Transductive Transfer Learning，LATTL），将直推式支持向量机（TSVM）和子空间学习结合起来同时进行训练。上述方法的直推式风范对样本外泛化存在困难；更重要的是，它们都没有对领域间的条件分布和判别结构差异进行形式化和适配。

根据上述讨论，总结迁移学习关键计算问题如图 3.1 所示，并启发本章的研究动机。给定标注辅助领域 $\mathcal{D}_s$ 和无标目标领域 $\mathcal{D}_t$，如子图 (a)，可见由于概率分布差异，$\mathcal{D}_s$ 上训练的判别面 f 并不能对 $\mathcal{D}_t$ 进行足够准确地预测。本章首先最小化*边缘分布*之间的距离，如子图 (b)，这样判别面 f 就可以对 $\mathcal{D}_t$ 进行更为准确地预测。值得注意的是，最小化*条件分布*差异（判别结构适配）是不可或缺的，如子图 (c)，这样领域间与标注数据相关的判别结构才能更好得对应起来。最后，基于半监督学习思想最大化了无标数据的流形一致性，如子图 (d)，这样判别面 f 需要进行适当旋转调整才能与目标领域流形结构一致，这使无标数据也可以参与预测。

本章提出了*联合适配正则化*迁移学习框架 ARTL，基于结构风险最小化准则和正则化理论，统一进行联合概率分布适配和流形结构保持，同步优化结构风险泛函、联合分布距离、流形不一致性等学习准则。本章的主要贡献总结如下：

- 首次提出联合分布适配方法，并与结构风险最小化、流形正则化等方法结合为统一的迁移学习框架。该框架仅比主流半监督学习框架[88] 增加了一个适配正则项，既简洁又同时基于可再生希尔伯特空间表出定理有全局最优解。
- 多种标准学习方法如正则化线性回归、支持向量机、主成份分析等，可直接嵌入本章框架解决迁移学习问题，满足分类预测、特征规约等各种需求。
- 在 252 个文本分类和图像识别任务上的系统性实验证明了本章方法的优势。

3.2　联合分布适配

首先给出本章的问题定义，其次提出联合适配正则化通用学习框架，基于正则化线性回归、支持向量机、主成份分析等提出四种学习模型、并利用可再生希尔伯特空间的表出定理推导凸优化解，最后分析计算复杂度和理论泛化误差上界。

3.2.1　问题定义

简明起见，本章中常用的符号及其描述总结如表3.1所示，首先给出问题定义。

定义 3.1 (**领域、任务**)：[1] 领域 $\mathcal{D}$ 由 d 维特征空间 $\mathcal{F}$ 和边缘概率分布 $P(\mathbf{x})$ 组成，即 $\mathcal{D}=\{\mathcal{F},P(\mathbf{x})\}$, $\mathbf{x}\in\mathcal{F}$。给定领域 $\mathcal{D}$，任务 $\mathcal{T}$ 由标签集合 $\mathcal{Y}$ 和分类模型 $f(\mathbf{x})$ 组成，即 $\mathcal{T}=\{\mathcal{Y},f(\mathbf{x})\}$，$y\in\mathcal{Y}$，从统计观点 $f(\mathbf{x})=Q(y|\mathbf{x})$ 可解释为条件概率分布。

一般地，如果两个领域 $\mathcal{D}_s$ 和 $\mathcal{D}_t$ 的特征空间或边缘分布不同，则认为两个领域 $\mathcal{D}_s$ 和 $\mathcal{D}_t$ 不同，即 $\mathcal{F}_s\neq\mathcal{F}_t\vee P(\mathbf{x}_s)\neq P(\mathbf{x}_t)$。类似地，如果两个任务 $\mathcal{T}_s$ 和 $\mathcal{T}_t$ 的标签集合或条件分布不同，则认为两个任务 $\mathcal{T}_s$ 和 $\mathcal{T}_t$ 不同，即 $\mathcal{Y}_s\neq\mathcal{Y}_t\vee Q(y_s|\mathbf{x}_s)\neq Q(y_t|\mathbf{x}_t)$。

问题 3.1 (**联合分布适配**)：给定有标的辅助领域 $\mathcal{D}_s=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_{n_s},y_n)\}$ 和无标的目标领域 $\mathcal{D}_t=\{\mathbf{x}_{n+1},\ldots,\mathbf{x}_{n+m}\}$ 满足 $\mathcal{F}_s=\mathcal{F}_t$, $\mathcal{Y}_s=\mathcal{Y}_t$, $P(\mathbf{x}_s)\neq P(\mathbf{x}_t)$, $Q(y_s|\mathbf{x}_s)\neq P(y_t|\mathbf{x}_t)$，学习一个特征表示 T 或分类模型 f 同时满足（*1*）边缘分布 $P_s(\mathbf{x}_s)$ 与 $P_t(\mathbf{x}_t)$ 之差异最小化且（*2*）条件分布 $Q_s(\mathbf{x}_s|y_s)$ 与 $Q_t(\mathbf{x}_t|y_t)$ 之差异最小化，从而标准分类器经领域 $\mathcal{D}_s$ 训练后可以准确地泛化到领域 $\mathcal{D}_t$（问题示意如图 *3.2*）。

通过直接估计分布密度的方法来解决领域迁移学习问题是十分有挑战性的[1]。虽然目标领域边缘分布 $P_t(\mathbf{x}_t)$ 可由核密度估计（Kernel Density Estimate，KDE）得到近似拟合[24]，但是由于目标领域没有标注数据，条件分布 $Q_t(y_t|\mathbf{x}_t)$ 无法通过核密度估计得到。因此，多数已有工作都直接假设存在一个合理的特征变换 F 满足

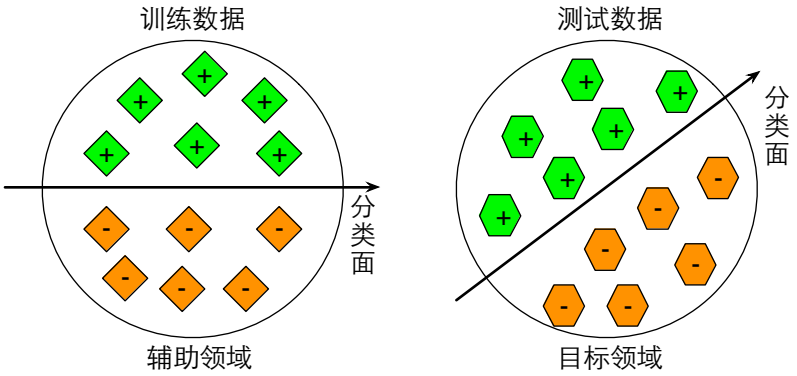


图 3.2　数据集偏移中的边缘分布偏移和条件分布偏移示意图。

表 3.1　本章常用的符号及其描述。

符号	描述	符号	描述	符号	描述	符号	描述
$\mathcal{D}_s$	辅助领域	$\mathcal{D}_t$	目标领域	$\mathbf{X}$	输入数据矩阵	$\mathbf{Y}$	输入标签矩阵
n	辅助样例数	m	目标样例数	$\mathbf{K}$	输入核矩阵	$\mathbf{E}$	标签指示矩阵
d	特征维度	C	类别数	$\mathbf{A}$	适配变换矩阵	$\mathbf{Z}$	嵌入表征矩阵
k	子空间维度	σ	模型正则项	$\mathbf{w}$	分类器参数	$\boldsymbol{\alpha}$	分类器参数
λ	适配正则项	γ	流形正则项	$\{\mathbf{M}_c\}_{c=0}^{C}$	MMD 矩阵	$\mathbf{L}$	流形邻接矩阵

$P_s(F(\mathbf{x}_s)) = P_t(F(\mathbf{x}_t))$ 且 $Q_s(F(\mathbf{x}_s)|y_s) \approx Q_t(F(\mathbf{x}_t)|y_t)$；满足条件的特征变换 F 可由最小化领域间的边缘分布距离、同时保持原始数据的关键统计属性推断得到[45]。

本章提出如下论断：

- 仅最小化边缘分布间的距离并不充分，条件分布间的距离也应显式最小化。
- 挖掘边缘分布是必要的，保持边缘分布流形一致性可获益于半监督学习[88]。

基于上述论断，主要的计算问题是如何最小化联合概率分布 J_s 和 J_t 之间的分布距离。根据概率论 $J = P \cdot Q$，因此本章通过同时最小化（1）边缘分布 P_s 和 P_t 之间的距离（2）条件分布 Q_s 和 Q_t 之间的距离，来实现联合概率分布 J_s 和 J_t 的适配。

3.2.2　边缘分布适配

要通过最小化边缘分布 $P_s(\mathbf{x}_s)$ 和 $P_t(\mathbf{x}_t)$ 之间的距离来进行边缘分布适配，首先要定义恰当的距离度量函数。由于对数据分布的概率密度进行参数化估计是一个比概率分布适配更困难的问题，根据 Vapnik[2] 统计学习理论“不要把具体问题转换成一个更一般化的问题来解决”，因此本章转而寻求概率分布的矩匹配，即匹配其各阶统计量。为了减小边缘分布 $P_s(\mathbf{x}_s)$ 和 $P_t(\mathbf{x}_t)$ 之间的分布差异，本章采用经验最大均值差异（Maximum Mean Discrepancy，MMD）[33,45,95] 来度量不同概率分布的失配程度，它定义为辅助领域和目标领域在无穷维核空间中的均值距离：

$$\mathrm{MMD}_{\mathcal{H}}^2(P_s, P_t) = \left\| \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i) - \frac{1}{m}\sum_{j=n+1}^{n+m}\phi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 \tag{3-1}$$

其中 $\phi : \mathcal{X} \mapsto \mathcal{H}$ 是核空间中的无穷阶非线性特征映射；根据泰勒展开定理，ϕ 可以展开为无穷多项式级数，因此具有匹配不同概率分布的任意阶统计量的能力[33]。

3.2.3　条件分布适配

然而，减小边缘分布之间的距离并不能保证条件分布之间的距离也能隐式地得以减小。事实上，最小化条件分布 $Q_s(\mathbf{x}_s|y_s)$ 和 $Q_t(\mathbf{x}_t|y_t)$ 之间的差异对于分布适配鲁棒性至关重要[95]。遗憾的是，由于目标领域没有标注数据，即便通过矩匹

配来进行条件分布适配也是一个非平凡任务，也就是说 $Q_t(y_t|\mathbf{x}_t)$ 不能通过已有数据拟合得到。最近的几年的一些工作开始通过核空间中的样本选择[24]、环式验证[92]、协同训练[96]、以及核密度估计[97] 来实现条件分布适配，但这些方法通常要求目标领域存在部分标注数据，因而不能解决本章提出的无监督迁移学习问题。

本章提出利用直推式学习获得目标领域的预标注从而计算条件分布距离。目标领域的预标注可以通过将辅助领域训练好的基础分类器应用到目标领域得到，该基础分类器既可以是标准学习器如支持向量机，也可以是迁移学习器如迁移主成份分析[45]。由于在分类模型未知时，类后验概率（即条件分布）$Q_s(y_s|\mathbf{x}_s)$ 和 $Q_t(y_t|\mathbf{x}_t)$ 很难拟合得到，本章转而考察类条件分布 $Q_s(\mathbf{x}_s|y_s)$ 和 $Q_t(\mathbf{x}_t|y_t)$ 的矩匹配。现在有了目标领域的预标注以及辅助领域的真标注，即可对类别空间 $c \in \{1,\ldots,C\}$ 中的每个类别 c 分别进行类条件分布 $Q_s(\mathbf{x}_s|y_s = c)$ 和 $Q_t(\mathbf{x}_t|y_t = c)$ 的矩匹配。本章扩展 MMD 准则[33] 用以度量类条件分布 $Q_s(\mathbf{x}_s|y_s = c)$ 和 $Q_t(\mathbf{x}_t|y_t = c)$ 的距离如下：

$$\mathrm{MMD}^2_{\mathcal{H}}\left(Q_s^{(c)}, Q_t^{(c)}\right) = \left\| \frac{1}{n^{(c)}} \sum_{\mathbf{x}_i \in D_s^{(c)}} \phi(\mathbf{x}_i) - \frac{1}{m^{(c)}} \sum_{\mathbf{x}_j \in D_t^{(c)}} \phi(\mathbf{x}_j) \right\|^2_{\mathcal{H}} \tag{3-2}$$

其中 $\mathcal{D}_s^{(c)} = \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{D}_s \wedge y(\mathbf{x}_i) = c\}$ 是辅助领域中属于类别 c 的样例集合，$y(\mathbf{x}_i)$ 是样例 $\mathbf{x}_i$ 的真标注，$n_s^{(c)} = |\mathcal{D}_s^{(c)}|$。相应地，$\mathcal{D}_t^{(c)} = \left\{\mathbf{x}_j : \mathbf{x}_j \in \mathcal{D}_t \wedge \widehat{y}(\mathbf{x}_j) = c\right\}$ 是目标领域中预标注属于类别 c 的样例集合，$\widehat{y}(\mathbf{x}_j)$ 是样例 $\mathbf{x}_j$ 的预标注，$n_t^{(c)} = |\mathcal{D}_t^{(c)}|$。

将公式 (3-1) 和 (3-2) 结合起来，得到本章的联合分布适配正则项，计算如下：

$$D_{\mathcal{H}}(J_s, J_t) = \mathrm{MMD}^2_{\mathcal{H}}(P_s, P_t) + \sum_{c=1}^{C} \mathrm{MMD}^2_{\mathcal{H}}\left(Q_s^{(c)}, Q_t^{(c)}\right) \tag{3-3}$$

通过最小化公式 (3-3)，边缘分布和条件分布各阶矩统计量都在无穷维 $\mathcal{H}$ 中进行了适配。注意到，如果将迁移学习器迭代式地用于获得目标领域的预标注，则该预标注一般会更为准确、并可进一步提高联合分布适配质量和目标领域分类准确率。虽然由于分布差异的存在，目标领域的预标注并不完全正确，但它是当前学习条件下所能获得的最佳分类结果，因此用它来提高联合分布适配效果是合理的。

3.3　监督学习算法与分析

3.3.1　监督学习框架

联合适配正则化 ARTL 框架基于结构风险最小化和正则化理论，有如下准则：

1. 最小化辅助领域标注数据 $\mathcal{D}_s$ 上的结构风险泛函；
2. 最小化跨领域联合概率分布 J_s 和 J_t 之间的距离函数；

3. 最大化边缘分布 P_s 和 P_t 的流形一致性，提高半监督学习效果。

记预测模型为 $f = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$，其中 $\mathbf{w}$ 是模型参数，$\phi : \mathcal{X} \mapsto \mathcal{H}$ 是从原始特征空间 $\mathcal{F}$ 到可再生希尔伯特空间空间 $\mathcal{H}$ 的特征映射函数。ARTL 学习框架形式化为：

$$f = \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \sigma \|f\|_K^2 + \lambda D_{\mathcal{H}}\left(J_s, J_t\right) + \gamma M_{\mathcal{H}}\left(P_s, P_t\right) \tag{3-4}$$

其中 K 是与 ϕ 对应的核函数，满足核技巧的内积关系 $\left\langle \phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right) \right\rangle = K\left(\mathbf{x}_i, \mathbf{x}_j\right)$；$\sigma$、$\lambda$ 和 γ 是非负正则化参数，公式 (3-4) 每一项的具体含义在下面章节详细给出。

3.3.1.1　结构风险最小化

迁移学习的最终目标是为目标领域 $\mathcal{D}_t$ 学习一个准确的自适应分类器，为此首先在辅助领域 $\mathcal{D}_s$ 归纳一个标准分类器 f 作为迁移到目标领域 $\mathcal{D}_t$ 的基础版本。该基础分类器 f 可由结构风险最小化原理[2] 最小化如下的*结构风险泛函*进行学习：

$$f = \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \sigma \|f\|_K^2 \tag{3-5}$$

其中 $\mathcal{H}_K$ 是假设空间，$\|f\|_K^2$ 是函数 f 范数，σ 是模型正则项参数，ℓ 是度量 f 对训练样本拟合精度的损失函数。两类广泛使用的损失函数是用于支持向量机的 *hinge* *损失函数* $\ell = \max\left(0, 1 - y_i f\left(\mathbf{x}_i\right)\right)$，用于线性回归的*二次损失函数* $\ell = \left(y_i - f\left(\mathbf{x}_i\right)\right)^2$。

3.3.1.2　适配正则化

公式 (3-3) 作用于非线性核映射 ϕ 而非直接作用于分类器 f，因而不能作为适配正则项且不能通过表出定理求解。为使公式 (3-3) 成为分类器 f 的正则项，引入*投影距离*，即对非线性映射 ϕ 进行投影 $\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$ 再计算联合分布距离，定义如下：

$$\begin{aligned} D_{\mathcal{H}}\left(J_s, J_t\right) &= \mathrm{MMD}_{\mathcal{H}}^2\left(P_s, P_t\right) + \sum_{c=1}^{C} \mathrm{MMD}_{\mathcal{H}}^2\left(Q_s^{(c)}, Q_t^{(c)}\right) \\ &= \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}_i\right) - \frac{1}{m} \sum_{j=n+1}^{n+m} \mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2 \\ &+ \sum_{c=1}^{C} \left\| \frac{1}{n^{(c)}} \sum_{\mathbf{x}_i \in D_s^{(c)}} \mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}_i\right) - \frac{1}{m^{(c)}} \sum_{\mathbf{x}_j \in D_t^{(c)}} \mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2 \end{aligned} \tag{3-6}$$

由于 $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$，上述正则项直接作用于分类模型的参数向量 $\mathbf{w}$，成为控制分类模型在领域间适配程度的正则项，它对迁移学习模型的效能具有决定性的作用。

3.3.1.3　流形正则化

在无监督迁移学习中同时存在标注数据和无标数据，因此可以利用半监督学习进一步提高学习效果。由于公式 (3-3) 仅可对领域间的不同概率分布进行矩匹配，为更好地拟合无标数据就需要进一步挖掘边缘分布 P_s 和 P_t 隐藏的流形信息，即由无标数据揭示目标领域的潜在知识结构。采用流形假设[88]，即如果两个数据点 $\mathbf{x}_s, \mathbf{x}_t \in \mathcal{X}$ 在边缘分布 $P_s(\mathbf{x}_s)$ 和 $P_t(\mathbf{x}_t)$ 的内在几何流形上相似，则条件分布 $Q_s(y_s|\mathbf{x}_s)$ 和 $Q_t(y_t|\mathbf{x}_t)$ 也应该相似。在流形平滑性条件下，流形正则化可形式化为：

$$M_{\mathcal{H}}\left(P_s, P_t\right) = \sum_{i,j=1}^{n+m} \left(f\left(\mathbf{x}_i\right) - f\left(\mathbf{x}_j\right)\right)^2 W_{ij} = \sum_{i,j=1}^{n+m} f\left(\mathbf{x}_i\right) L_{ij} f\left(\mathbf{x}_j\right) \tag{3-7}$$

其中 $\mathbf{W}$ 是图邻接矩阵，$\mathbf{L}$ 是归一化图拉普拉斯矩阵，$\mathbf{W}$ 可计算如下

$$W_{ij} = \begin{cases} \cos\left(\mathbf{x}_i, \mathbf{x}_j\right), & \text{如果 } \mathbf{x}_i \in \mathcal{N}_p\left(\mathbf{x}_j\right) \vee \mathbf{x}_j \in \mathcal{N}_p\left(\mathbf{x}_i\right) \\ 0, & \text{其他} \end{cases} \tag{3-8}$$

其中 $\mathcal{N}_p\left(\mathbf{x}_i\right)$ 是样例 $\mathbf{x}_i$ 的 p-近邻组成的样例集合，$\mathbf{L}$ 计算为 $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$，$\mathbf{D}$ 是对角矩阵，对角元计算为 $D_{ii} = \sum_{j=1}^{n} W_{ij}$。结构风险最小化中的流形正则化可提高模型 f 判别结构与目标领域流形结构的一致性，并提高半监督学习效果[88]。

3.3.2　监督学习算法

本节在联合适配正则化 ARTL 框架下，扩展标准学习算法（正则化线性回归、支持向量机）用以解决迁移学习问题。其主要困难是核映射 $\phi : \mathcal{X} \mapsto \mathcal{H}$ 可能导致无穷维空间，为高效求解 (3-4) 算法，需要用表出定理对 ARTL 进行重新形式化。

定理 3.1 (**表出定理**)：[88,98] 优化问题 (3-4) 的最优解具有如下展开式：

$$f\left(\mathbf{x}\right) = \sum_{i=1}^{n+m} \alpha_i K\left(\mathbf{x}_i, \mathbf{x}\right) \quad \text{和} \quad \mathbf{w} = \sum_{i=1}^{n+m} \alpha_i \phi\left(\mathbf{x}_i\right) \tag{3-9}$$

基于跨领域标注和无标数据，其中 K 是与 ϕ 对应的核函数，α_i 是分类模型参数。

在联合适配正则项 (3-3) 中带入表出定理公式 (3-9)，可得到新形式化如下：

$$\begin{gathered} D_{\mathcal{H}}\left(J_s, J_t\right) = \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\mathbf{M}_0\mathbf{K}\boldsymbol{\alpha}\right) + \sum_{c=1}^{C} \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\mathbf{M}_c\mathbf{K}\boldsymbol{\alpha}\right) = \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}\boldsymbol{\alpha}\right) \\ \text{其中 } \mathbf{M} \triangleq \sum_{c=0}^{C} \mathbf{M}_c \end{gathered} \tag{3-10}$$

其中 $\mathbf{K} \in \mathbb{R}^{(n+m)\times(n+m)}$ 是输入数据核矩阵，$K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n+m})$ 是分类模型参数向量。适配各个类别 c 的 *MMD* 指示矩阵 $\mathbf{M}_c, c \in \{0, 1, \ldots, C\}$ 计算为：

$$(M_c)_{ij} = \begin{cases} \frac{1}{n^{(c)}n^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s^{(c)} \\ \frac{1}{m^{(c)}m^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ \frac{-1}{n^{(c)}m^{(c)}}, & \begin{cases} \mathbf{x}_i \in \mathcal{D}_s^{(c)}, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ \mathbf{x}_j \in \mathcal{D}_s^{(c)}, \mathbf{x}_i \in \mathcal{D}_t^{(c)} \end{cases} \\ 0, & \text{其他} \end{cases} \tag{3-11}$$

其中 $n^{(c)}, m^{(c)}, \mathcal{D}_s^{(c)}, \mathcal{D}_t^{(c)}, c \in \{1, \ldots, C\}$ 定义如公式 (3-2) 所示。为保持公式简明，也可由公式 (3-11) 计算 $\mathbf{M}_0$，只要带入 $n^{(0)} = n, m^{(0)} = m, \mathcal{D}_s^{(0)} = \mathcal{D}_s, \mathcal{D}_t^{(0)} = \mathcal{D}_t$ 即可。

类似地，将表出定理公式 (3-9) 带入流形正则化公式 (3-7)，可得

$$M_{\mathcal{H}}(P_s, P_t) = \mathrm{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{KLK}\boldsymbol{\alpha}\right) \tag{3-12}$$

由公式 (3-10) 和 (3-12) 可实现基于正则化线性回归和支持向量机的迁移学习算法。

3.3.2.1　ARRLS：基于二次损失函数的 ARTL 模型

采用二次损失函数 $\ell(f(\mathbf{x}_i), y_i) = (y_i - f(\mathbf{x}_i))^2$，结构风险泛函可形式化如下：

$$\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \sigma\|f\|_K^2 = \sum_{i=1}^{n+m} E_{ii}(y_i - f(\mathbf{x}_i))^2 + \sigma\|f\|_K^2 \tag{3-13}$$

其中对角阵 $\mathbf{E}$ 是标注指示矩阵，$E_{ii} = 1$ 如果 $\mathbf{x}_i \in \mathcal{D}_s$，否则 $E_{ii} = 0$。将表出定理 (3-9) 带入公式 (3-13)，结构风险泛函可重新形式化为

$$\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) + \sigma\|f\|_K^2 = \left\|\left(\mathbf{Y} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\right)\mathbf{E}\right\|_F^2 + \sigma\mathrm{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha}\right) \tag{3-14}$$

其中 $\mathbf{Y} = [y_1, \ldots, y_{n+m}]$ 是标注矩阵。虽然目标领域标注未知但这并不影响上述形式化，因为标注指示矩阵 $\mathbf{E}$ 已将目标领域标注信息清零。将公式 (3-14)、(3-10) 和 (3-12) 带入 ARTL 框架 (3-4)，得适配正则化线性回归 *ARRLS* 模型，优化问题如下：

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^{n+m}}{\arg\min} \left\|\left(\mathbf{Y} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\right)\mathbf{E}\right\|_F^2 + \mathrm{tr}\left(\sigma\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\left(\lambda\mathbf{M} + \gamma\mathbf{L}\right)\mathbf{K}\boldsymbol{\alpha}\right) \tag{3-15}$$

为求解分类模型，设置上述目标函数相对于参数向量 $\boldsymbol{w}$ 的导数为 $\mathbf{0}$ 可得模型参数：

$$\boldsymbol{\alpha} = \left(\left(\mathbf{E} + \lambda\mathbf{M} + \gamma\mathbf{L}\right)\mathbf{K} + \sigma\mathbf{I}\right)^{-1}\mathbf{E}\mathbf{Y}^{\mathrm{T}} \tag{3-16}$$

当 $\lambda = \gamma = 0$ 时，联合适配正则项置零，模型退化为标准正则化线性回归（RLS）。

多分类问题： 记 $\mathbf{y} \in \mathbb{R}^C$ 为标注向量，满足 $y_c = 1$ 如果 $y(\mathbf{x}) = c$，否则 $y_c = 0$。定义“1-of-C”标注矩阵为 $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n+m}] \in \mathbb{R}^{C\times(n+m)}$，即每列仅有一个非零元，相应的分类器参数矩阵为 $\boldsymbol{\alpha} \in \mathbb{R}^{(n+m)\times C}$。如此，ARRLS 可直接扩展到多分类问题。

3.3.2.2　ARSVM：基于 hinge 损失函数的 ARTL 模型

采用 hinge 损失函数 $\ell\left(f(\mathbf{x}_i), y_i\right) = \max\left(0, 1 - y_i f\left(\mathbf{x}_i\right)\right)$，结构风险泛函转变为：

$$\sum_{i=1}^{n} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \sigma \|f\|_K^2 = \sum_{i=1}^{n} \max\left(0, 1 - y_i f\left(\mathbf{x}_i\right)\right) + \sigma \|f\|_K^2 \tag{3-17}$$

将表出定理 (3-9) 带入公式 (3-17)，并将公式 (3-17)、(3-10) 和 (3-12) 统一带入到 ARTL 框架 (3-4) 中，可得到*适配正则化支持向量机 ARSVM* 模型，优化问题如下：

$$\begin{aligned} &\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n+m}, \boldsymbol{\xi}\in\mathbb{R}^n} \sum_{i=1}^{n} \xi_i + \sigma \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{K} \boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{K} \left(\lambda \mathbf{M} + \gamma \mathbf{L}\right) \mathbf{K} \boldsymbol{\alpha} \\ &\text{s.t.} \quad y_i \left[\sum_{j=1}^{n+m} \alpha_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) + b\right] \geq 1 - \xi_i, i = 1, \ldots, n \\ &\qquad\qquad \xi_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{3-18}$$

根据文献[88] 采用拉格朗日对偶法对公式 (3-18) 重新形式化，可得到对偶优化问题

$$\begin{aligned} &\boldsymbol{\beta} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \sum_{i=1}^{n} \beta_i - \frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Q} \boldsymbol{\beta} \\ &\text{s.t.} \quad \sum_{i=1}^{n} \beta_i y_i = 0, 0 \leq \beta_i \leq \frac{1}{n}, i = 1, \ldots, n \\ &\text{其中} \quad \mathbf{Q} = \tilde{\mathbf{Y}} \tilde{\mathbf{E}} \mathbf{K} (2\sigma \mathbf{I} + 2\left(\lambda \mathbf{M} + \gamma \mathbf{L}\right) \mathbf{K})^{-1} \tilde{\mathbf{E}}^{\mathrm{T}} \tilde{\mathbf{Y}} \end{aligned} \tag{3-19}$$

其中 $\tilde{\mathbf{Y}} = \mathrm{diag}(y_1, \ldots, y_n)$ 为标注矩阵，$\tilde{\mathbf{E}} = [\mathbf{I}_n, \mathbf{0}] \in \mathbb{R}^{n\times(n+m)}$ 为指示矩阵。ARSVM 可由标准支持向量机工具包求解：首先解以矩阵 $\mathbf{Q}$ 为系数的二次规划得到对偶问题参数向量 $\boldsymbol{\beta}$，然后计算原始模型参数向量为 $\boldsymbol{\alpha} = (2\sigma \mathbf{I} + 2\left(\lambda \mathbf{M} + \gamma \mathbf{L}\right) \mathbf{K})^{-1} \tilde{\mathbf{E}}^{\mathrm{T}} \tilde{\mathbf{Y}} \boldsymbol{\beta}$。

迁移学习监督模型 ARRLS 和 ARSVM 的学习算法总结如算法 3 所示。为统一量纲、方便参数 λ 和 γ 调优，对图拉普拉斯矩阵和 MMD 指示矩阵做了归一化。

3.3.2.3　计算复杂度

记 s 为平均每个样例非零特征数，则 $s \leq d, p \ll \min(n+m, d)$。本章算法的复杂度包括以下三部分：（1）通过 LU 分解求解公式 (3-16) 或 (3-19) 的逆矩阵问题需要 $O((n+m)^3)$，可由共轭梯度法大幅降低计算量；通过 LIBSVM[99] 求解 ARSVM 模型 (3-19) 的 SVM 优化问题需要 $O((n+m)^{2.3})$。（2）构造图拉普拉斯矩阵 $\mathbf{L}$ 需要

算法 3: ARTL：联合适配正则化迁移学习监督模型 ARRLS 和 ARSVM

输入: 数据矩阵 $\mathbf{X}$、标注矩阵 $\mathbf{Y}$；图近邻数 p、模型正则项参数 σ、联合适配正则化参数 λ、流形正则项参数 γ。

输出: 迁移分类模型 $f: \mathcal{X} \mapsto \mathcal{Y}$。

开始

由公式 (3-10)、(3-11) 构造 MMD 矩阵，(3-8) 构造图拉普拉斯矩阵 $\mathbf{L}$。

选择核函数 $K(\mathbf{x}_i, \mathbf{x}_j)$ 并计算核矩阵 $\mathbf{K}$ 为 $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$。

归一化 MMD 矩阵和图拉普拉斯矩阵 $\mathbf{M} \leftarrow \frac{\mathbf{M}}{\|\mathbf{M}\|_F}$，$\mathbf{L} \leftarrow \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$。

通过公式 (3-16) 求解 ARRLS 模型的参数向量 $\boldsymbol{\alpha}$，通过公式 (3-19) 和 SVM 工具包如 LIBSVM 求解 ARSVM 模型的参数向量 $\boldsymbol{\alpha}$。

通过公式 (3-9) 返回迁移分类模型 $f: \mathcal{X} \mapsto \mathcal{Y}$。

$O\left(s(n+m)^2\right)$。（3）构造核矩阵 $\mathbf{K}$ 和 MMD 指示矩阵 $\mathbf{M}$ 需要 $O\left(C(n+m)^2\right)$。综上所述，算法 3 基于精确数值计算的复杂度为 $O\left((n+m)^3 + (s+C)(n+m)^2\right)$。如果采用共轭梯度法，计算复杂度还可以进一步大幅降低，但限于篇幅本章暂不探讨。

3.3.2.4　与已有工作的联系和区别

本章工作显著区别于迁移学习方法[14,35,42,67,71]，这些工作没有考虑概率分布适配或流形正则化。除此之外，再从技术层面讨论本章与相关工作的联系和区别。

概率分布适配：这类方法通过最小化特定距离函数如 MMD 和布雷格曼散度，来显式地减小领域间的分布差异[40,46,85,89–92]，但它们仅减小了边缘分布差异，而未减小条件分布差异。极少数文献[24,97] 试图同时匹配边缘分布和条件分布，但它们通常要求目标领域存在少量标注数据，这与本章讨论的无监督迁移学习风范不符；另外，这些方法没有考察边缘分布的流形结构，不能最大化半监督学习效能。

流形正则化：这类方法利用半监督学习，显式地最大化嵌入表征（子空间学习）[59,100–102] 或分类模型（监督学习）[88,103] 与数据内在流形结构的一致性，然而它们没有显式地最小化领域间的概率分布差异，因而会遭遇负迁移和欠适配问题。

与 ARTL 框架最相似的工作包括图正则化迁移学习（Graph co-regularized Transfer Learning，GTL）[59]、半监督迁移主成份分析（Semi-Supervised Transfer Component Analysis，SSTCA）[45]、判别性特征抽取（Discriminative Feature Extraction，DFE）[43]、隐性直推式迁移学习（Latent Transductive Transfer Learning，LATTL）[93] 以及半监督核匹配（Semi-Supervised Kernel Matching，SSKM）[94]。这些相关工作都可归类为“半监督迁移学习”，因为它们对半监督学习和迁移学习

表 3.2　本章方法与最相关工作的详细比较。

比较对象	GTL	SSTCA	DFE	LATTL	SSKM	ARTL
数据重构	√	√		√		
结构风险最小化			√	√	√	√
边缘分布适配		√	√		∠	√
条件分布适配						√
流形正则化	√	√	✓	∠	√	√
凸优化						√
通用框架			√	√		√

√ 统一优化问题；✓ 两步方法；∠ 类似方法。

进行了结合。为简明比较起见，这些方法与 ARTL 之间的联系和区别如表3.2所示。

- GTL 和 SSTCA 是维度规约方法，标注信息和流形结构都用于增强子空间学习的判别性能；ARTL 是基于结构风险最小化和正则化理论的监督学习框架。
- DFE 同时在子空间中进行分布适配和分类器训练，它虽然也考虑了流形一致性但是分开在两步过程进行，此外它没有进行领域间条件概率分布适配。
- LATTL 同时进行异构子空间学习和直推式半监督分类，它利用直推式支持向量机（TSVM）获得更好的目标领域泛化性能；但仍然存在两个局限：（1）没有显式地进行概率分布适配；（2）直推式 TSVM 难以用于外样本预测。
- SSKM 同时考察了结构风险最小化、核匹配、流形正则化，它与 ARTL 有三方面显著区别：（1）SSKM 没有显式最小化条件分布差异；（2）核匹配是非凸整数规划；（3）核匹配未作为模型正则项，无法由表出定理求凸优化解。

综上所述，本章 ARTL 框架同时考察（1）结构风险最小化（2）边缘分布适配和条件分布适配（3）流形一致性最大化。ARTL 基于可再生希尔伯特空间的正则化理论，其最优解可由通用的表出定理求取。因此，ARTL 是一个通用学习框架，可以载入各种监督学习方法。此外，ARTL 还是一个凸优化问题，具有全局最优解。

3.3.3　泛化误差分析

本节采用与文献[2,38,43]类似的推导方法，基于辅助领域的结构风险分析 ARTL 模型在目标领域的泛化误差上界。记 $f(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})\right)$ 为训练得到的预测函数，$h(\mathbf{x}) : \mathcal{X} \mapsto \{1, -1\}$ 为真实标注函数。设 $\ell(\mathbf{x})$ 是连续损失函数 $\ell(\mathbf{x}) = \left|h(\mathbf{x}) - f(\mathbf{x})\right|$，则满足 $0 \le \ell(\mathbf{x}) \le 2$。首先，定义预测函数 f 在目标领域 $\mathcal{D}_t$ 的期望风险为：

$$\epsilon_t(f) = \mathbb{E}_{\mathbf{x}\sim P_t}\left[\left|h(\mathbf{x}) - f(\mathbf{x})\right|\right] = \mathbb{E}_{\mathbf{x}\sim P_t}\left[\ell(\mathbf{x})\right]$$

类似地，定义预测函数 f 在辅助领域 $\mathcal{D}_s$ 的期望风险为：

$$\epsilon_s(f)=\mathbb{E}_{\mathbf{x}\sim P_s}\left[\left|h(\mathbf{x})-f(\mathbf{x})\right|\right]=\mathbb{E}_{\mathbf{x}\sim P_s}\left[\ell(\mathbf{x})\right]$$

下面的定理基于辅助领域风险、领域间的概率分布差异、最优假设在两个领域的误差上界给出了目标领域泛化误差上界，它是文献[38] 理论结果的一个直接推论。

定理 3.2：　设包含 f 的假设空间 VC 维是 d，则 f 在目标领域 $\mathcal{D}_t$ 的期望风险至少以概率 $1-\delta$ 满足如下上界：

$$\epsilon_t(f)\leq\hat{\epsilon}_s(f)+\sqrt{\frac{4}{n}\left(d\log\frac{2en}{d}+\log\frac{4}{\delta}\right)}+D_{\mathcal{H}}(J_s,J_t)+\Omega \tag{3-20}$$

其中 e 是自然对数底，$\hat{\epsilon}_s(f)$ 是 f 在辅助领域 $\mathcal{D}_s$ 上的经验风险，$\Omega=\inf_{f\in\mathcal{H}_K}\left[\epsilon_s(f)+\epsilon_t(f)\right]$ 是真实标注函数 h 在假设空间中的风险上界。

上述定理表明，分类模型 f 在目标领域 $\mathcal{D}_t$ 的期望风险即 $\epsilon_t(f)$ 存在上界；要最小化该上界，可以同时最小化：（1）辅助领域 $\mathcal{D}_s$ 的经验风险 $\hat{\epsilon}_s(f)$（2）辅助领域 $\mathcal{D}_s$ 和目标领域 $\mathcal{D}_t$ 在可再生希尔伯特空间 $\mathcal{H}$ 中的联合分布距离 $D_{\mathcal{H}}(J_s,J_t)$（3）真实标注函数 h 在假设空间 $\mathcal{H}_K$ 中的风险上界 Ω。

在 ARTL 框架公式 (3-4) 中，通过结构风险最小化即公式 (3-5) 可以显式地最小化 $\hat{\epsilon}_s(f)$；通过联合分布适配即公式 (3-3) 可以显式地最小化 $D_{\mathcal{H}}(J_s,J_t)$；通过流形正则化即公式 (3-7) 可以隐式地最小化 Ω。这里进一步解释为什么流形正则化即公式 (3-7) 可以隐式地最小化真实标注函数 h 在假设空间 $\mathcal{H}_K$ 的风险上界 Ω。为此首先引入下面的定理，该定理给出了流形正则化半监督学习的泛化误差上界。

定理 3.3：[104] 考察训练集合 $(\mathbf{x}_i,y_i)$，其中 $i\in\mathcal{Z}_{n+m}=\{1,\ldots,n+m\}$，假设从 $\mathcal{Z}_{n+m}$ 随机均匀地选取 n 个不同的整数 $j_1,\ldots,j_n$ 并记为 $\mathcal{Z}_n$，记 h 为真实标注函数，$\hat{\mathbf{f}}(\mathcal{Z}_n)$ 为以 $\mathcal{Z}_n$ 为标注数据、$\mathcal{Z}_{n+m}\backslash\mathcal{Z}_n$ 为无标数据训练得到的半监督学习器

$$\hat{\mathbf{f}}(\mathcal{Z}_n)=\underset{\mathbf{f}\in\mathbb{R}^{n+m}}{\arg\inf}\left[\frac{1}{n}\sum_{i\in\mathcal{Z}_n}\ell(f_i,y_i)+\gamma\mathbf{f}^{\mathrm{T}}\mathbf{L}\mathbf{f}\right]$$

如果 $|\frac{\partial}{\partial h}\ell(h,y)|\leq\tau$ 且损失函数 $\ell(h,y)$ 相对于 h 是凸函数，则半监督学习器在无标数据 $\mathcal{Z}_{n+m}\backslash\mathcal{Z}_n$ 上的泛化误差上界为

$$\mathbb{E}_{\mathcal{Z}_n}\frac{1}{m}\sum_{i\in\mathcal{Z}_{n+m}\backslash\mathcal{Z}_n}\ell\left(\hat{f}_i(\mathcal{Z}_n),y_i\right)\leq\inf_{\mathbf{f}\in\mathbb{R}^{n+m}}\left[\frac{1}{n+m}\sum_{i=1}^{n+m}\ell(f_i,y_i)+\gamma\mathbf{f}^{\mathrm{T}}\mathbf{L}\mathbf{f}+\frac{\tau^2\mathrm{tr}\left(\mathbf{L}^{-1}\right)}{2\gamma n(n+m)}\right]$$

在 ARTL 框架中，流形正则项即公式 (3-7) 和联合分布适配即公式 (3-3) 都在可再生希尔伯特空间 $\mathcal{H}$ 中执行，因此可以认为联合分布适配后的数据满足半监督

学习的条件即标注数据和无标数据独立同分布。这样，定理 3.3 揭示了在辅助领域和目标领域全集 $\mathcal{D}_s \cup \mathcal{D}_t$ 上训练的半监督分类器 $\hat{f}$ 可以在无标数据（目标领域）$\mathcal{D}_t$ 上满足上述泛化误差上界。换句话说，流形正则项 (3-7) 可以隐式地最小化 Ω。

3.4　表征学习算法与分析

3.4.1　表征学习框架

本节在联合适配正则化框架 ARTL 下，扩展表征学习算法（主成份分析、正交映射）用以解决迁移学习问题。主要思想是：通过非线性特征映射 T 使得特征向量 $\mathbf{x}$ 和类别标签 y 相对于联合概率分布 J_s 和 J_t 的期望在领域间适配（矩匹配）：

$$\begin{aligned} D_{\mathcal{H}}\left(J_s, J_t\right) &= \mathrm{MMD}_{\mathcal{H}}^2\left(P_s, P_t\right) + \sum_{c=1}^{C} \mathrm{MMD}_{\mathcal{H}}^2\left(Q_s^{(c)}, Q_t^{(c)}\right) \\ &= \left\| \frac{1}{n} \sum_{i=1}^{n} T\left(\mathbf{x}_i\right) - \frac{1}{m} \sum_{j=n+1}^{n+m} T\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2 \\ &+ \sum_{c=1}^{C} \left\| \frac{1}{n^{(c)}} \sum_{\mathbf{x}_i \in D_s^{(c)}} T\left(\mathbf{x}_i\right) - \frac{1}{m^{(c)}} \sum_{\mathbf{x}_j \in D_t^{(c)}} T\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2 \end{aligned} \tag{3-21}$$

根据文献[45]，非线性特征映射 T 可以取为核主成份分析 $T\left(\mathbf{x}\right) = \mathbf{V}^{\mathrm{T}} \phi\left(\mathbf{x}\right)$，其中矩阵 $\mathbf{V}$ 是核主成份分析的特征变换矩阵；根据文献[105]，非线性特征映射 T 可以取为正交映射 $T\left(\mathbf{x}\right) = \phi\left(\mathbf{A}^{\mathrm{T}}\mathbf{x}\right)$，其中矩阵 $\mathbf{A}$ 是正交映射矩阵；上述两种方法都满足最大均值差异 MMD 的非线性映射条件。和监督学习场景一样，上述表征学习问题也是非平凡的，因为目标领域没有标注数据且 $Q_t(y_t|\mathbf{x}_t)$ 无法准确估计。由已知数据得到的最佳估计是 $Q_t(y_t|\mathbf{x}_t) \approx Q_s(y_t|\mathbf{x}_t)$[23]，这等价于将辅助领域标注数据训练得到的分类器 f 应用到目标领域无标数据。为了获得对 $Q_t(y_t|\mathbf{x}_t)$ 的更准确估计，本章还提出了迭代式预标注精化方法，迭代式地同步改进特征变换 T 和分类器 f。

3.4.2　表征学习算法

3.4.2.1　ARPCA：基于主成份分析的 ARTL 实现

维度规约方法通过最小化输入数据重构误差来学习特征变换。为通用起见，本章采用主成份分析（Principal Component Analysis，PCA）作为数据重构方法。记 $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ 为输入数据矩阵，$\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}$ 为协方差中心化矩阵，$n \leftarrow n + m$ 是总样本数、$\mathbf{1}$ 是元素全为 1 的 $n \times n$ 矩阵，则协方差矩阵为 $\mathbf{X}\mathbf{H}\mathbf{X}^{\mathrm{T}}$。

PCA 的学习目标是找到正交变换 $\mathbf{V} \in \mathbb{R}^{d\times k}$ 使得输入数据的嵌入协方差最大化：

$$\max_{\mathbf{V}^{\mathrm{T}}\mathbf{V}=\mathbf{I}} \operatorname{tr}\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}\mathbf{H}\mathbf{X}^{\mathrm{T}}\mathbf{V}\right) \tag{3-22}$$

其中 $\operatorname{tr}(\cdot)$ 表示矩阵的迹。该优化问题可以通过本征分解 $\mathbf{X}\mathbf{H}\mathbf{X}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\boldsymbol{\Phi}$ 高效求解，其中对角矩阵 $\boldsymbol{\Phi} = \operatorname{diag}(\phi_1, \ldots, \phi_k) \in \mathbb{R}^{k\times k}$ 包含 k 个最大本征值。满足嵌入协方差最大的最优 k 维表征可由特征变换 $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] = \mathbf{V}^{\mathrm{T}}\mathbf{X}$ 得到。

KPCA： 为了在非线性的可再生希尔伯特空间中通过最大均值差异 MMD 实现概率分布适配，考察核映射 $\psi : \mathbf{x} \mapsto \psi(\mathbf{x})$ 或 $\psi(\mathbf{X}) = [\psi(\mathbf{x}_1), \ldots, \psi(\mathbf{x}_n)]$ 以及相应核矩阵 $\mathbf{K} = \psi(\mathbf{X})^{\mathrm{T}}\psi(\mathbf{X}) \in \mathbb{R}^{n\times n}$。采用表出定理 $\mathbf{V} = \phi(\mathbf{X})\,\mathbf{A}$ 对 PCA 进行核化可得

$$\max_{\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}} \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{A}\right) \tag{3-23}$$

其中 $\mathbf{A} \in \mathbb{R}^{n\times k}$ 是核主成份分析（Kernel PCA，KPCA）的变换矩阵，子空间嵌入表征为 $\mathbf{Z} = \mathbf{A}^{\mathrm{T}}\mathbf{K}$。通过非线性核映射可以在无穷维核空间中进行概率分布的矩匹配。

联合适配正则化表征学习的目标是，学习一个隐含特征表示使得领域间联合概率分布差异实现最小化。将非线性特征映射 $T(\mathbf{x}) = \mathbf{V}^{\mathrm{T}}\phi(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}\mathbf{K}(:, \mathbf{x})$ 带入联合分布适配公式 (3-21)，可得到联合适配正则化主成分分析（ARPCA）优化问题：

$$\min_{\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{A}=\mathbf{I}} \sum\nolimits_{c=0}^{C} \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{M}_c\mathbf{K}^{\mathrm{T}}\mathbf{A}\right) + \lambda \|\mathbf{A}\|_F^2 \tag{3-24}$$

其中 $\lambda > 0$ 是控制模型复杂度并保证模型适定的正则化参数。根据广义瑞利商，最小化公式 (3-21) 并最大化公式 (3-22)，等价于最小化公式 (3-21) 并使公式 (3-22) 保持不变。因此，上述优化问题与联合适配正则化表征学习的优化目标保持一致。

根据带约束最优化方法[84]，记 $\boldsymbol{\Phi} = \operatorname{diag}(\phi_1, \ldots, \phi_k) \in \mathbb{R}^{k\times k}$ 为拉格朗日乘子，则带约束优化问题 (3-24) 的拉格朗日函数可形式化为

$$L = \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\left(\mathbf{K}\sum\nolimits_{c=0}^{C}\mathbf{M}_c\mathbf{K}^{\mathrm{T}} + \lambda\mathbf{I}\right)\mathbf{A}\right) + \operatorname{tr}\left(\left(\mathbf{I} - \mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{A}\right)\boldsymbol{\Phi}\right) \tag{3-25}$$

为求全局最优解，设置上述函数相对变换矩阵 $\mathbf{A}$ 的导数 $\frac{\partial L}{\partial \mathbf{A}} = \mathbf{0}$，得广义本征分解

$$\left(\mathbf{K}\sum\nolimits_{c=0}^{C}\mathbf{M}_c\mathbf{K}^{\mathrm{T}} + \lambda\mathbf{I}\right)\mathbf{A} = \mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Phi} \tag{3-26}$$

这样，求取最优特征映射矩阵 $\mathbf{A}$ 归结为求解本征分解 (3-26) 得到 k 个最小的本征向量。整个计算过程总结如算法 4 所示。值得一提的是，通过 ARPCA 通常可以得到目标领域的一个更为准确的标注结果。因此，如果将此结果作为目标领域的预标注并迭代式地执行 ARPCA 算法，则可以不断地提高目标领域的标注效果直到算法收敛（例如目标函数值不再下降，或达到指定迭代次数）。由于目标领域预标注是根据已知数据所能获得的最好结果，因此上述迭代式标注精化过程是合理的。

算法 4: ARTL: 适配正则化迁移学习特征算法 ARPCA 和 AROM

输入: 数据矩阵 $\mathbf{X}$，标注矩阵 $\mathbf{y}$；子空间维度 k，模型正则化参数 λ。

输出: 变换矩阵 $\mathbf{A}$，特征表示 $\mathbf{Z}$，迁移分类模型 f。

1 **开始**

2　通过公式 (3-11) 构造 MMD 矩阵 $\mathbf{M}_0$，置 $\{\mathbf{M}_c := \mathbf{0}\}_{c=1}^{C}$。

3　**repeat**

4　　ARPCA：解广义本征分解 (3-26)，取 k 个最小本征向量构造 $\mathbf{A}$ 并设置 $\mathbf{Z} := \mathbf{A}^{\mathrm{T}}\mathbf{K}$；AROM：由格拉斯曼流形的共轭梯度算法求解 $\mathbf{A}$，并设置 $\mathbf{Z} := \mathbf{A}^{\mathrm{T}}\mathbf{X}$。

5　　在标注数据 $\{(\mathbf{z}_i, y_i)\}_{i=1}^{n}$ 上训练标准分类器 f，并用它更新目标领域预标注 $\{\widehat{y}_j := f(\mathbf{z}_j)\}_{j=n+1}^{n+m}$。

6　　由公式 (3-11) 构造或更新 MMD 矩阵 $\{\mathbf{M}_c\}_{c=1}^{C}$。

7　**until** 收敛

8　返回在标注数据 $\{\mathbf{A}\mathbf{x}_i, y_i\}_{i=1}^{n}$ 上训练的迁移分类模型 f。

下面用大 O 法分析算法 4 的计算复杂度。子空间维度 k 一般不大于 500，迭代次数 T 一般不大于 50，因而 $k \ll \min(d,n), T \ll \min(d,n)$。计算复杂度包括以下部分：第 4 行的广义本征分解需要 $O\left(Tkd^2\right)$，第 2、6 行计算 MMD 矩阵需要 $O\left(TCn^2\right)$，其他步骤需要 $O\left(Tdn\right)$。综上，算法 4 的总计算复杂为 $O\left(Tkn^2 + TCn^2 + Tdn\right)$。

3.4.2.2　AROM：基于正交映射的 ARTL 实现

根据文献[105]，将非线性特征映射 $T(\mathbf{x}) = \phi\left(\mathbf{A}^{\mathrm{T}}\mathbf{x}\right)$ 带入联合分布适配公式 (3-21)，可得到联合适配正则化正交映射（AROM）优化问题：

$$\begin{gathered}\min_{\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}} \sum_{c=0}^{C} \operatorname{tr}\left(\mathbf{K}_{\mathbf{A}}\mathbf{M}_c\right) \\ \text{其中 } \left(\mathbf{K}_{\mathbf{A}}\right)_{ij} = \left\langle \phi\left(\mathbf{A}^{\mathrm{T}}\mathbf{x}_i\right), \phi\left(\mathbf{A}^{\mathrm{T}}\mathbf{x}_j\right) \right\rangle = K\left(\mathbf{A}^{\mathrm{T}}\mathbf{x}_i, \mathbf{A}^{\mathrm{T}}\mathbf{x}_i\right)\end{gathered} \tag{3-27}$$

这里 K 是高斯核函数 $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}$，$\gamma$ 是带宽参数。上述优化问题的待求变量 $\mathbf{A}$ 位于非线性核函数 K 里面，因此是一个非线性优化问题，不能像 ARPCA 通过表出定理得到凸优化问题；又因为存在正交约束 $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$，也不能通过梯度下降法求取局部最优解。虽然非线性带来了求解方面的困难，但 AROM 相对 ARPCA 优势是：ARPCA 是在降维核空间中匹配概率分布的各阶矩，而 AROM 是在无穷维核空间中匹配概率分布的各阶矩，因而 AROM 的匹配程度强于 ARPCA。

本章采用格拉斯曼流形（Grassmann Manifold）上的共轭梯度下降（Conjugate

表 3.3　文本数据集 20-Newsgroups 的层次结构和统计信息。

数据集	大类	子类	样例数	特征数
20-Newsgroups	comp	comp.graphics	970	25,804
		comp.os.ms-windows.misc	963	
		comp.sys.ibm.pc.hardware	979	
		comp.sys.mac.hardware	958	
	rec	rec.autos	987	
		rec.motorcycles	993	
		rec.sport.baseball	991	
		rec.sport.hokey	997	
	sci	sci.crypt	989	
		sci.electronics	984	
		sci.med	987	
		sci.space	985	
	talk	talk.politics.guns	909	
		talk.politics.mideast	940	
		talk.politics.misc	774	
		talk.religion.misc	627	

Gradient，CG）算法[106] 来求解优化问题 (3-27)。CG 是非线性优化问题的主流求解算法，具有快速收敛速率。该算法在线性无关的共轭方向上迭代式地对目标函数进行优化。格拉斯曼流形上的 CG 方法总结如下（详细数学过程参见文献[106]）：

1. 在流形上计算目标函数 f 在当前变量值 $\mathbf{A}$ 处的梯度 $\nabla f_{\mathbf{A}} = \partial f_{\mathbf{A}} - \mathbf{A}\mathbf{A}^T \partial f_{\mathbf{A}}$；

2. 通过上次搜索方向与当前梯度 $\nabla f_{\mathbf{A}}$ 共同确定当前搜索方向 $\mathbf{H}$；

3. 在当前方向 $\mathbf{H}$ 上沿着 $\mathbf{A}$ 的最短路径（测地线）执行线性搜索。

上述步骤迭代式执行直到达到局部最优解或者最大迭代次数，如算法 4 所示。

3.5　实验过程与结果

本节在两个实际应用（文本分类、图像识别）中进行系统性实验，验证本章学习算法（ARRLS、ARSVM、ARPCA）的有效性，数据集和代码可从公网下载。

3.5.1　实验数据

3.5.1.1　文本数据

按照迁移学习文献[29,45,47,60,63,67,85] 介绍的通用协议，本章考察被广泛采用的 *20-Newsgroups* 文本数据集，并根据其层次结构生成 216 个跨领域文本分类任务。

20-Newsgroups[①] 数据集包含约 20,000 个文档，4 个大类分别为 *comp*、*rec*、*sci* 和 *talk*，每个大类包含 4 个子类，详细信息如表3.3所示。在实验中构造了 6 组跨领域二分类任务，每组任务由 4 个大类中随机选取 2 个大类构成，一个大类记为正例，另一个大类记为负例，6 个任务组具体为 *comp vs rec*、*comp vs sci*、*comp vs talk*、*rec vs sci*、*rec vs talk* 和 *sci vs talk*。每个跨领域分类任务（包括辅助领域和目标领域）采用文献[67] 介绍的方法生成：每个任务组 *P vs Q* 的两个大类 *P* 和 *Q* 分别包含 4 个子类 P_1、P_2、P_3、P_4 和 Q_1、Q_2、Q_3、Q_4；随机选取 *P* 的两个子类（如 P_1 、P_2）与 *Q* 的两个子类（如 Q_1 、Q_2）构成辅助领域，其余子类（*P* 的 P_3、P_4 和 *Q* 的 Q_3、Q_4）构成目标领域。以上构造策略既保证辅助领域和目标领域是相关的，因为它们都来自同样的大类；又保证辅助领域和目标领域是不同的，因为它们来自不同的子类。每个任务组 *P vs Q* 可以生成 $C_4^2 \cdot C_4^2 = 36$ 个分类任务，总计 6 个任务组共生成 $6 \cdot 36 = 216$ 个分类任务。数据集经过文本预处理后包含 25,804 个词项特征和 15,033 个文档，每个文档由 *tf-idf* 向量表征，如表3.3所示。

3.5.1.2　图像数据

本章在如下图像集上测试各种算法的效能：字符集 USPS 和 MNIST，人脸集 Multi-PIE，对象集 COIL20、Office 和 Caltech ，统计信息如表3.4，示例如图 3.3 。

USPS[②] 数据集包括 7,291 张训练图片和 2,007 张测试图片，图片大小为 16×16。**MNIST**[③] 数据集包括 60,000 张训练图片和 10,000 张测试图片，图片大小 28×28。从图 3.3 可以直观感受到，USPS 和 MNIST 数据集分别服从显著不同的概率分布。两个数据集都包含 10 个类别，每个类别是 1–10 之间的某个字符。在实验中通过如下方式构造分类任务 *USPS vs MNIST*：在 USPS 中随机选取 1,800 张图片作为辅助数据、在 MNIST 中随机选取 2,000 张图片作为目标数据。交换辅助领域和目标领域可以得到另一个分类任务 *MNIST vs USPS*。图片预处理包括：将所有图片大小线性缩放为 16×16，每幅图片用 256 维的特征向量表征，编码了图片的像素灰度值信息。辅助领域和目标领域共享特征空间和类别空间，但数据分布显著不同。

COIL20[④] 包含 20 个对象类别共 1,440 张图片；每个对象类别包括 72 张图片，每张图片拍摄时对象水平旋转 5 度（共 360 度）。每幅图片大小为 32×32，表征为 1,024 维的向量。实验中将该数据集划分为两个不相交的子集 **COIL1** 和 **COIL2**：COIL1 包括位于拍摄角度为 $[0°, 85°] \cup [180°, 265°]$（第一、三象限）的所有图片；

① http://people.csail.mit.edu/jrennie/20newsgroups
② http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html
③ http://yann.lecun.com/exdb/mnist
④ http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

图 3.3　字符集 USPS、MNIST，人脸集 PIE，以及对象集 COIL20、Office、Caltech-256。

表 3.4　字符、人脸、对象等图像数据集的统计信息。

数据集	类型	样例数	特征数	类别数	包括子集
USPS	字符	1,800	256	10	USPS
MNIST	字符	2,000	256	10	MNIST
PIE	人脸	11,554	1,024	68	PIE1, …, PIE5
COIL20	对象	1,440	1,024	20	COIL1, COIL2
Office	对象	1,410	800	10	A, W, D
Caltech	对象	1,123	800	10	C

COIL2 包括位于拍摄角度为 $[90°, 175°] \cup [270°, 355°]$（第二、四象限）的所有图片。这样，子集 COIL1 和 COIL2 的图片因为拍摄角度不同而服从不同的概率分布。将 COIL1 作为辅助领域、COIL2 作为目标领域，可以构造跨领域分类任务 *COIL1 vs COIL2*；交换辅助领域和目标领域，可以得到另外一个分类任务 *COIL2 vs COIL1*。

PIE① 代表“朝向、光照、表情”的英文单词首字母，该数据集是人脸识别的基准测试集，包括 68 个不同人物的 41,368 幅人脸照片，图片大小为 32×32，每个人物的照片由 13 个同步的相机（不同朝向）、21 个不同曝光程度拍摄。实验中为了充分验证本章方法的有效性和鲁棒性，特别选取了 PIE 数据集的 5 个子集，每个子集对应一个人脸朝向。具体地说，5 个子集为 **PIE1** (C05，朝左)、**PIE2** (C07，朝上)、**PIE3** (C09，朝下)、**PIE4** (C27，朝前)、**PIE5** (C29，朝右)。在每个子集（朝向）中，所有人脸照片都拍摄于不同的光照、曝光和表情条件下。随机选取任意两个子集（朝向）分别作为辅助领域和目标领域，可以构造 $5 \times 4 = 20$ 个跨领域分类任务，即 *PIE1 vs PIE2*、*PIE1 vs PIE3*、*PIE1 vs PIE4*、*PIE1 vs PIE5*、…、*PIE5 vs PIE4*。这样，辅助领域和目标领域分别由不同朝向的人脸照片组成，从而服从显著不同的概率分布。此外，不同分类任务的概率分布差异也显著不同、具有多样性，例如朝左与朝右人脸照片之间的差异要大于朝左与朝前人脸照片之间的差异。

本章采用文献[15,61,107,108] 广泛使用的基准对象数据集 *Office* 和 *Caltech-256*，根据领域先验生成 12 个跨领域视觉对象识别任务来横向评测本章算法的效果。**Office**[15,107,108] 是视觉迁移学习的主流基准数据集，包含 3 个对象领域 **Amazon** (在线电商图片)、**Webcam** (网络摄像头拍摄的低解析度图片)、**DSLR** (单反相机拍摄的高解析度图片)，共有 4,652 张图片 31 个类别标签。**Caltech-256** 是对象识别

① http://vasc.ri.cmu.edu/idb/html/face

的基准数据集，包括 1 个对象领域 **Caltech**，共有 30,607 张图片 256 个类别标签。

为便于横向比较，实验直接采用文献[108] 发布的 *Office+Caltech* 预处理数据集。对每张图片抽取 SURF 特征，并向量化为 800 维的直方图表征，所有直方图向量都进行减均值除方差的归一化处理，直方图码表由 K 均值聚类算法在 *Amazon* 子集上生成。具体共有 4 个领域 **C** (Caltech-256)、**A**（Amazon）、**W**（Webcam）和 **D**（DSLR），从中随机选取 2 个不同的领域作为辅助领域和目标领域，则可构造 $4\times3=12$ 个跨领域视觉对象识别任务，如 $C\rightarrow A$，$C\rightarrow W$，$C\rightarrow D$，…，$D\rightarrow W$。

3.5.2　基准算法与实现细节

3.5.2.1　基准算法

为了验证基于联合适配正则化框架提出的 3 种学习算法 ARRLS、ARSVM 和 ARPCA 的有效性，本节考察 10 种迁移学习基准方法，包括经典方法和前沿方法：

- 逻辑斯特回归（Logistic Regression，LR）
- 支持向量机（Support Vector Machine，SVM）
- 拉普拉斯支持向量机（Laplacian SVM，LapSVM）[88]
- 跨领域谱分类（Cross-Domain Spectral Classification，CDSC）[100]
- 谱特征对齐（Spectral Feature Alignment，SFA）[42]
- 测地流核方法（Geodesic Flow Kernel，GFK）[108]
- 迁移主成份分析（Transfer Component Analysis，TCA）[45]
- 迁移子空间学习（Transfer Subspace Learning，TSL）[46]
- 最大间隔直推式迁移学习（Large Margin Transductive TL，LMTTL）[90]
- 半监督核匹配（Semi-Supervised Kernel Matching，SSKM）[94]

LMTTL 是 ARSVM 当 $\gamma=0, C=0$ 的特例，SSKM 是 ARRLS 当 $C=0$ 的特例，TCA 是 ARPCA 当 $C=0$ 的特例，因此对比实验能科学验证 ARTL 方法的有效性。

3.5.2.2　实现细节

根据文献[1,45,94] 评测协议，LR 和 SVM 在辅助领域标注数据上训练、在目标领域无标数据上测试；CDSC、SFA、GFK、TCA、TSL 和 ARPCA 在所有数据上学习特征表示，然后基于该特征表示由辅助领域标注数据训练 LR 分类器；LapSVM、LMTTL、SSKM、ARSVM 和 ARRLS 在所有数据上直推式地训练迁移学习模型。

经典交叉验证在目标领域没有标注数据时无法自动选择最优模型参数。相关文献一样，本章在所有 252 个分类任务上测试 10 种基准算法，将每种算法在

各种参数设置下的最佳效果用于性能对比。具体地说，LR 采用 LIBLINEAR① 工具包实现，SVM 采用 LIBSVM② 工具包实现，正则项参数 C（=1/2σ) 通过遍历 $C \in \{0.1, 0.5, 1, 5, 10, 50, 100\}$ 设置；LapSVM 采用文献[88] 的实现③，正则项参数 γ_A 和 γ_I 分别通过遍历 $\gamma_A, \gamma_I \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ 设置；LMTTL、SSKM 正则项参数 λ 通过遍历 $\lambda \in \{0.01, 0.1, 1, 10, 100\}$ 设置；CDSC、SFA、TCA、TSL、GFK 子空间维度 k 通过遍历 $k \in \{10, 20, \ldots, 200\}$ 设置。对核方法，在文本数据上采用线性核 $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j$，在视觉数据上采用高斯核 $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}$；根据文献[109]，高斯核函数带宽参数设为所有训练数据间平均欧式距离 A 的倒数 $\gamma = \frac{1}{A}$。

本章方法 ARRLS 和 ARSVM 包括 4 个可调参数：模型正则项参数 σ、适配正则项参数 λ、流形正则项参数 γ、以及邻接图近邻个数 p。下文的敏感性分析表明本章方法在较大参数范围内性能稳定，特别是对参数 σ、λ 和 p。在对比实验中，参数设置如下：固定 $\sigma = 0.1$、$\lambda = 10$、$p = 10$ 且设置（1）文本数据 $\gamma = 10$（2）图像数据 $\gamma = 1$。实践中，模型选择可以采用更简单的方式进行：先确定不敏感的参数，再确定较敏感的参数。首先，ARTL 对 σ 不敏感，通常选取一个较小值即可；其次，为使边缘分布和条件分布能在领域间充分适配，通常选取较大的 λ 值；最后，基于图正则化半监督学习[88] 可选取参数 γ 和 p。本章方法 ARPCA 仅有 2 个可调参数：子空间维度 k 和适配正则项参数 λ；k 通常选取为使子空间足够准确的维度，而 λ 通常选取为使边缘分布和条件分布可以在领域间充分适配的较小值。

本章采用测试数据（目标领域无标数据）准确率（Accuracy）作为评价指标：

$$Accuracy = \frac{\left|\mathbf{x} : \mathbf{x} \in \mathcal{D}_t \wedge f(\mathbf{x}) = y(\mathbf{x})\right|}{|\mathbf{x} : \mathbf{x} \in \mathcal{D}_t|} \tag{3-28}$$

其中 $y(\mathbf{x})$ 是测试样例 $\mathbf{x}$ 的真实标签，$f(\mathbf{x})$ 是待测学习算法为样例 $\mathbf{x}$ 预测的标签。

3.5.3　实验结果

本节给出本章方法 ARRLS、ARSVM 和 ARPCA 以及 10 种基准方法在 252 个跨领域文本、图像分类任务上的准确率，并根据实验结果深入讨论某些相关问题。

3.5.3.1　文本分类

基于本章框架 ARTL 实现的 3 种迁移学习方法 ARRLS、ARSVM、ARPCA 和 10 种基准方法在 6 个跨领域文本分类任务组共 216 个分类任务中的平均准确率如表3.5，每个方法在每个任务上的详细准确率如图 3.4(a)~3.4(f) 所示，每个图展示

① www.csie.ntu.edu.tw/~cjlin/liblinear/
② http://www.csie.ntu.edu.tw/~cjlin/libsvm
③ http://vikas.sindhwani.org/manifoldregularization.html

表 3.5　文本数据集的平均分类准确率（%），包括 6 个分类任务组共计 216 个分类任务。

任务组	标准学习			迁移降维			迁移分类		适配正则化		
	LR	SVM	LapSVM	CDSC	SFA	TCA	LMTTL	SSKM	ARSVM	ARRLS	ARPCA
C vs R	88.37	87.51	81.93	87.95	89.73	95.12	92.15	96.06	95.10	96.64	**97.22**
C vs S	77.87	75.38	68.96	75.72	78.07	77.32	77.58	84.15	84.53	86.71	**89.60**
C vs T	96.31	95.44	95.40	97.33	95.85	97.20	94.93	97.40	97.53	**98.03**	**98.04**
R vs S	75.28	73.82	74.21	77.53	79.25	82.31	78.24	85.71	87.19	91.02	**95.41**
R vs T	82.28	83.27	87.44	82.14	86.98	86.58	84.55	90.15	95.99	96.82	**97.10**
S vs T	76.99	76.85	80.22	80.97	79.27	79.30	74.80	74.74	89.03	91.11	**95.02**
平均	82.85	82.05	81.36	83.62	84.86	86.31	83.71	88.03	91.56	93.40	**95.40**

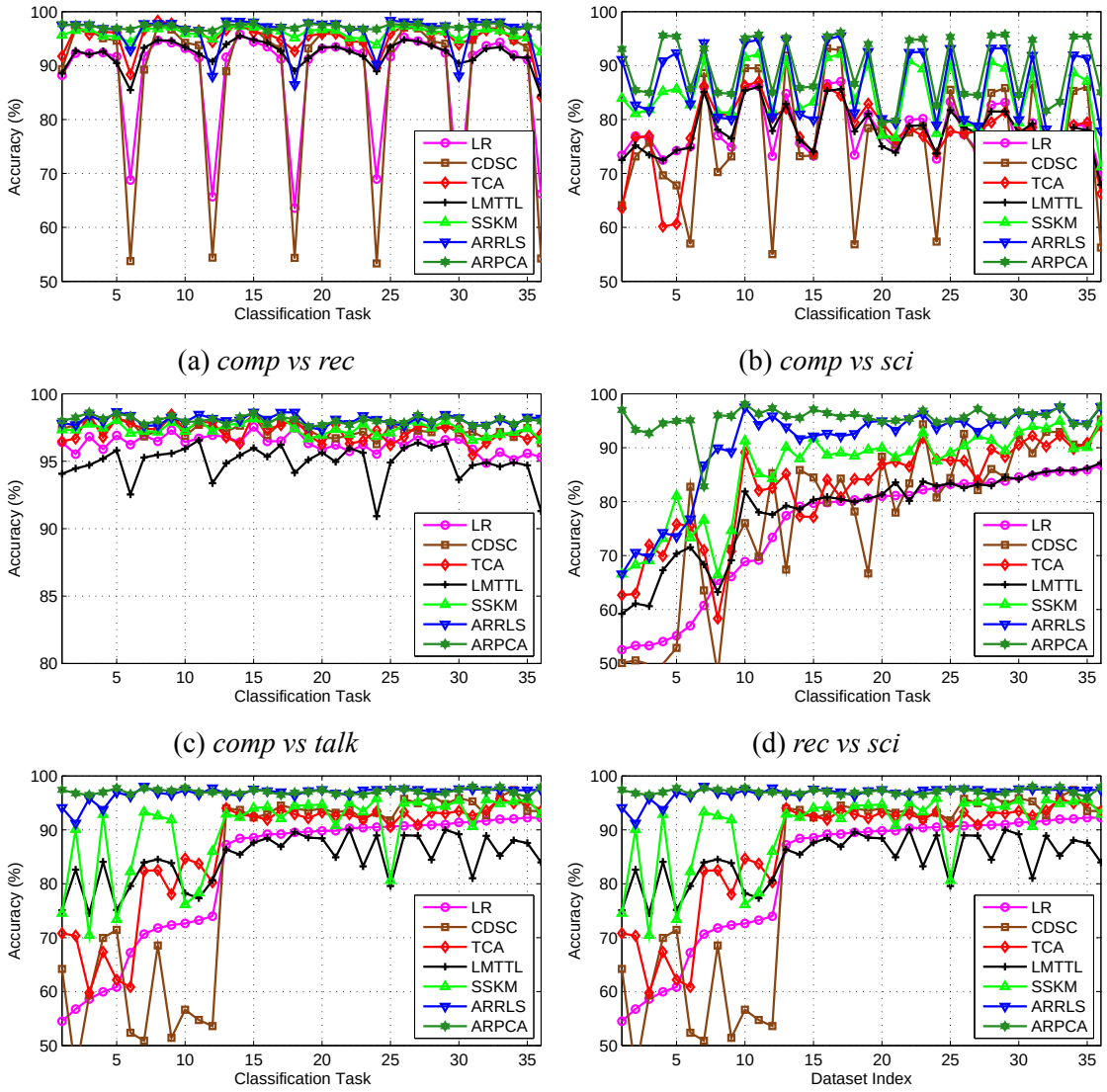
了一个任务组的 36 个任务，按照 LR 的准确率对任务序号进行了重新排序，从而反映跨领域知识迁移的难度。从图表中所呈现的实验结果，可以观察到如下结论。

ARTL 方法比基准方法获得了具有统计显著性的性能提高。ARSVM、ARRLS、ARPCA 在 216 个文本分类任务上的平均准确率分别为 **91.56%**、**93.40%**、**95.40%**，相对于最佳基准方法 SSKM 的提升幅度分别为 **3.53%**、**5.37%**、**7.37%**。上述结果是大量数据集上评测而来，因而能够有说服力地证明了 ARTL 方法可以为跨领域文档分类训练更准确的模型，同时还证明了基于表征学习 ARPCA 方法更为有效。

其次，观察到迁移学习方法在一般情况下能比标准学习方法获得更好的分类准确率。标准学习方法的主要局限在于将辅助领域和目标领域看成是独立同分布的，但在实际跨领域任务中，这个独立同分布假设通常难以不成立，因而导致了不不理想的分类效能。需要注意的是，公认效果很好的半监督学习方法 LapSVM 在该数据集上也未能取得比标准方法 LR 和 SVM 更好的效果，这多少有些与经验不符。可能的原因是 LapSVM 没有最小化领域间的概率分布差异，因而当在领域间直推式调整判别超平面时无法取得与目标领域流形结构的一致性，导致负迁移。

再次，观察到 ARTL 方法比迁移降维方法 CDSC、SFA、TCA 获得了统计显著的性能提高。现有迁移降维方法的一个局限性是不能同时最小化领域间边缘分布和条件分布的差异。虽然 SFA 被证明在跨领域情感分类任务中取得了很好的效果，但对话题分类任务效果并不理想，其原因是 SFA 仅挖掘了词特征之间的共现关系用于特征对齐学习、却没有考虑词频信息，因而仅对低频的情感数据有效、而对高频的文本数据性能下降。ARTL 解决了上述问题，从而取得了更好的效果。

最后，观察到 ARTL 方法比迁移分类方法 LMTTL 和 SSKM 获得了统计显著的性能提高。注意到 LMTTL 和 SSKM 都利用半监督学习来提高迁移分类的效果，它们在归纳监督或半监督分类器时也显式地最小化了边缘分布在领域间的差异，但条件分布失配问题却没有得到修正，因而仍然存在对目标领域的负迁移问题。



(a) *comp vs rec*　(b) *comp vs sci*

(c) *comp vs talk*　(d) *rec vs sci*

(e) *rec vs talk*　(f) *sci vs talk*

图 3.4　SVM、CDSC、TCA、LMTTL、SSKM、ARRLS 和 ARPCA 的文本分类准确率。

ARTL 方法成功避免了上述局限。此外，在难以分类的任务上 ARTL 方法取得了比基准方法更鲁棒的效果。这可以从图 3.4(a)~3.4(f) 观察到：对那些 LR 仅取得极低准确率（低于 70%）的任务，ARTL 方法相对于基准方法的提升幅度大大增加。

3.5.3.2　图像分类

ARTL 方法 ARPCA 和 5 种基准方法在 36 个跨领域图像识别任务（字符、人脸、对象）上的平均准确率图表3.6所示，每个分类任务上的详细准确率如图 3.5 所示。可观察到 ARPCA 方法比基准方法获得了统计显著的准确率提升。ARPCA 在 36 个分类任务上的平均准确率为 **57.37%**，比最好的基准方法 TSL 获得了 **7.57%** 的准确率提升，即 **15.07%** 的错误率下降。注意到 36 个分类任务的难度差别很大，

表 3.6　在字符、人脸、对象等 36 个跨领域分类任务上的平均准确率（%）。

任务组	标准学习		迁移学习			
	NN	PCA	GFK	TCA	TSL	ARPCA
字符	55.32	55.59	56.84	53.66	59.90	**63.46**
COIL	83.20	84.38	73.34	87.15	87.99	**88.89**
人脸	40.51	39.87	40.30	49.03	53.52	**62.90**
对象	31.37	39.79	42.95	43.61	42.37	**46.31**
平均	37.46	39.84	41.19	47.22	49.80	**57.37**

基准近邻分类器 NN 仅获得了 37.46% 的平均准确率，且在多个任务上效果极低。上述结果有力地证明了 ARPCA 可为跨领域图像分类构造高效且鲁棒的特征表示。

其次，可以看到 GFK 在 Office+Caltech 数据集上表现优异，但在其他数据集上表现不理想。其原因是 GFK 要求子空间维度足够小，从而使无穷个子空间可以在测地流上平滑地过渡，但子空间维度过低则无法对输入数据进行准确地表征。ARPCA 可以支持任意维度子空间、能够对输入数据进行准确地表征，从而可以学习足够精确的领域不变子空间。APRCA 也比主流特征迁移学习方法 TCA 取得了好得多的准确率，其原因是 TCA 没有对条件分布进行适配。由于基于 MMD 的矩匹配不能充分适配概率分布，因而同时进行边缘分布和条件分布适配就极为重要。

最后，ARPCA 比基于核密度估计（Kernel Density Estimate，KDE）的 TSL 取得了更好的效果。理论上，基于概率密度估计的 TSL 能比基于矩匹配的 ARPCA 和 TCA 更好地适配边缘分布，这也为实验结果所验证。然而，TSL 没有进行条件分布适配，一个潜在的困难是 TSL 无法通过 KDE 准确估计目标领域的条件分布。

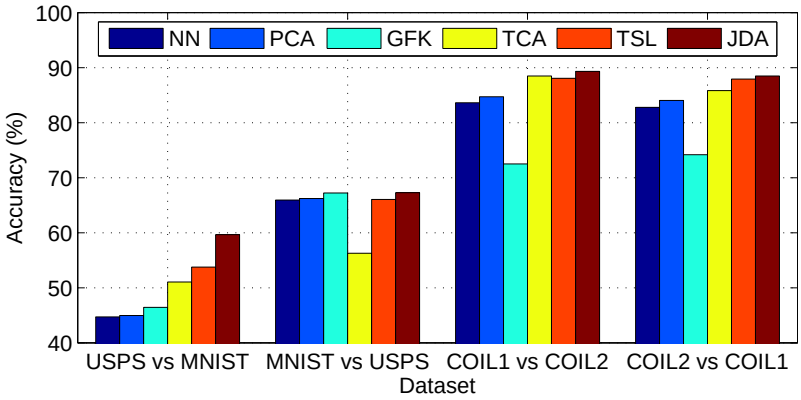
3.5.4　联合适配分析

为了深入分析联合分布适配的有效性，本节从以下四个方面进一步考察本章提出的方法：（1）分类结果（2）分布距离（3）嵌入特征表示（4）嵌入相似矩阵。

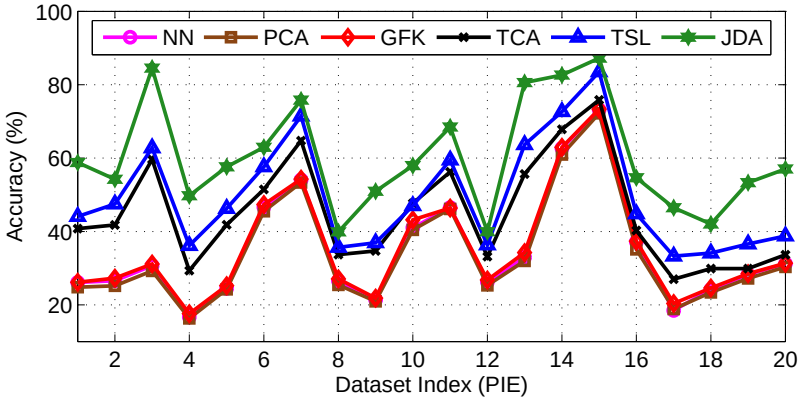
3.5.4.1　分类结果

为了检验条件适配、分布适配、流形约束等正则项的作用，随机选取分类任务 *rec vs sci 1* 并执行 ARRLS 算法，每次执行移除目标函数 (3-15) 的一个正则项。

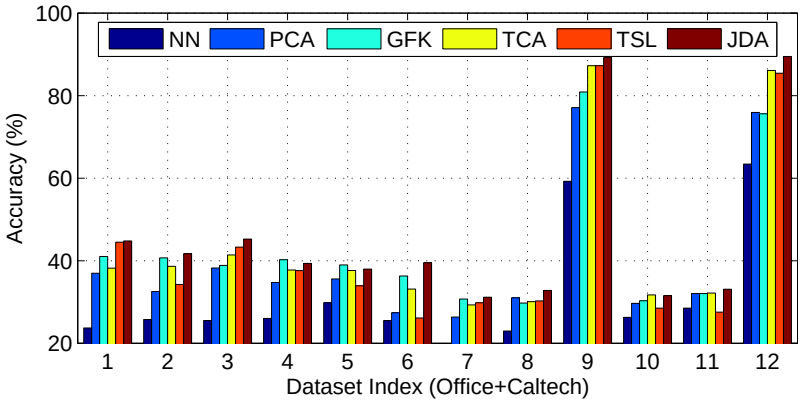
首先，移除条件分布适配正则项，即设置 $C = 0$，预测结果与分类模型参数向量如图 3.6(a) 所示。在这种情况下，无法为目标领域无标数据找到一个清晰的决策面，即目标数据并没有被区分开来。这证明了条件分布适配对迁移学习的至关重要性。类似结果还可以在图 3.6(b) 看到，该图展示了移除整个联合概率分布适配正则项后的结果，即设置 $\lambda = 0$。可以看到 $C = 0$ 和 $\lambda = 0$ 的结果非常相似，预



(a) 字符和对象识别，共 4 个分类任务



(b) 人脸识别，共 20 个分类任务，对应不同的朝向

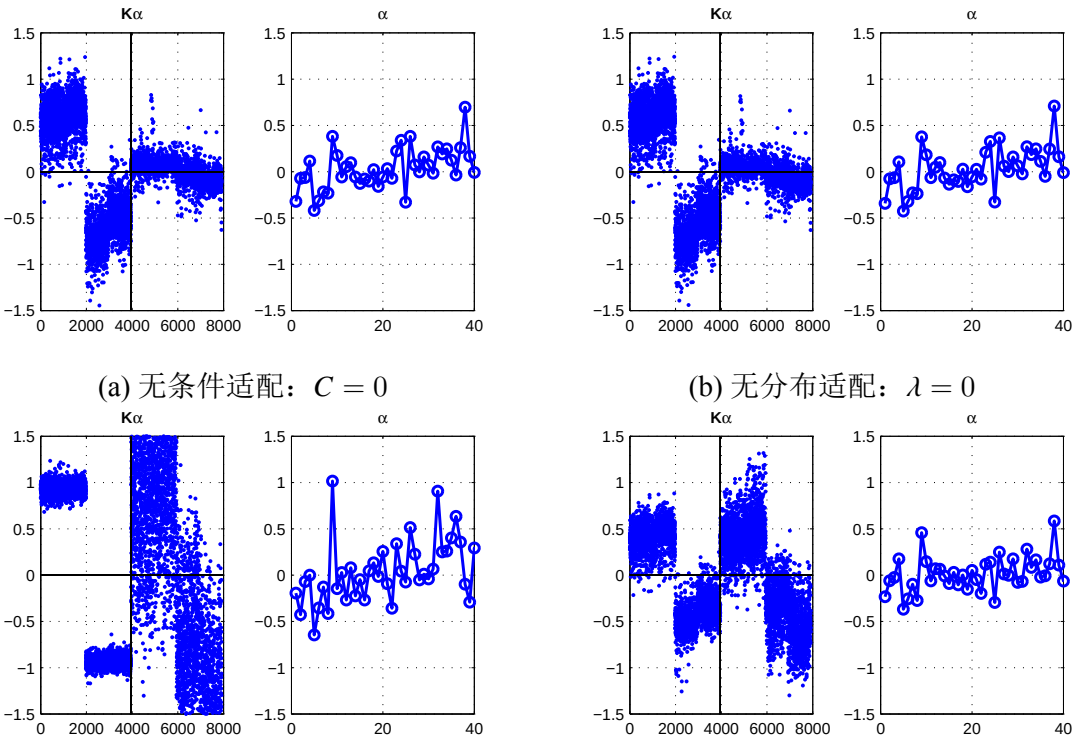


(c) 对象识别，共 12 个分类任务，对应不同的领域

图 3.5　各方法在 36 个跨领域图像分类任务上的准确率（%），每个任务的迁移难度不同。

示条件分布适配起决定性的作用，比边缘分布适配重要得多。通过条件分布适配，可以使类内差异减小、类间差异扩大，从而使数据更具有区分性，如图 3.6(d)。

其次，移除流形正则项，即设置 $\gamma = 0$，预测结果和分类模型参数向量如图 3.6(c) 所示。在这种情况下，预测值散落在真实标签区间 $[-1, 1]$ 之外的大范围内。换句话说，即便边缘分布和条件分布已经充分适配，如果移除了流形正则项，目标数据隐含的流形结构仍然会被破坏。因此，为了构建足够准确的半监督分类器，



(a) 无条件适配：$C = 0$　　(b) 无分布适配：$\lambda = 0$

(c) 无流形正则：$\gamma = 0$　　(d) 最佳参数设置

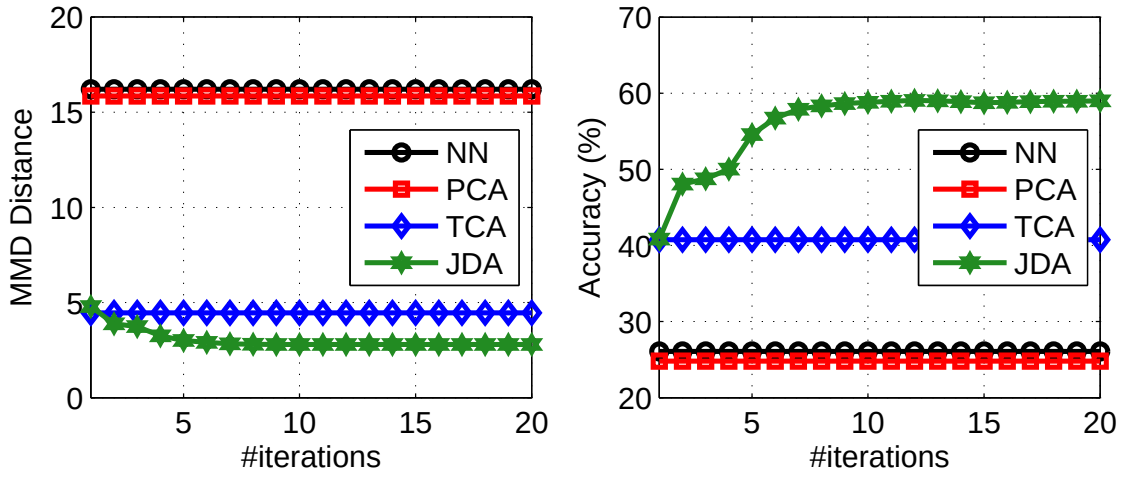
图 3.6　分类结果：ARRLS 在分类任务 *rec vs sci 1* 上的预测值 $\mathbf{K}\boldsymbol{\alpha}$ 和分类器参数向量 $\boldsymbol{\alpha}$。

保持输入数据的流形结构至关重要，这可对比图 3.6(c) 和图 3.6(d) 的结果看到。

3.5.4.2　分布距离

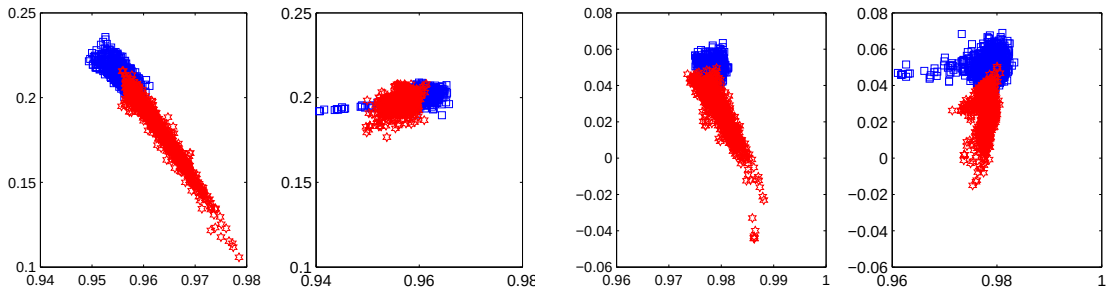
在人脸识别任务 *PIE1 vs PIE2* 上分别采用最佳参数执行 NN、PCA、TCA、ARPCA 算法，然后由公式 (3-24) 计算在各算法抽取的特征表示上的聚合 MMD 距离。为了计算不同领域间边缘分布和条件分布的真实距离，需要借用目标领域的真实标签；不过，这些真实标签仅用于这里的验证和分析，而不会用在学习过程。

每种算法执行后的跨领域分布距离如图 3.7(a) 所示，相应的分类准确率如图 3.7(b) 所示。可以得到如下结论：（1）NN 方法没有抽取新的特征表示，因此在原始空间中领域间的分布距离是最大的；（2）PCA 抽取了嵌入方差最大的特征表示，可以小幅度降低分布距离、但并不显著，因而分类准确率的提升也不显著；（3）TCA 可以显著地减小分布距离，因为它的目标函数显式地最小化了边缘分布的 MMD 距离，因此它可以获得显著的性能提升；（4）ARPCA 同时减小了边缘分布和条件分布的 MMD 距离，因此可以抽取到有效性和鲁棒性最好的特征表示，并在迭代过程中最大程度降低领域间的分布距离、提高目标领域分类准确率。



(a) MMD 距离随迭代次数变化规律　　(b) 准确率随迭代次数的变化规律

图 3.7　有效性验证：在人脸识别任务 *PIE1 vs PIE2* 上，领域间联合分布的 MMD 距离、相应的分类准确率随迭代次数的变化情况。

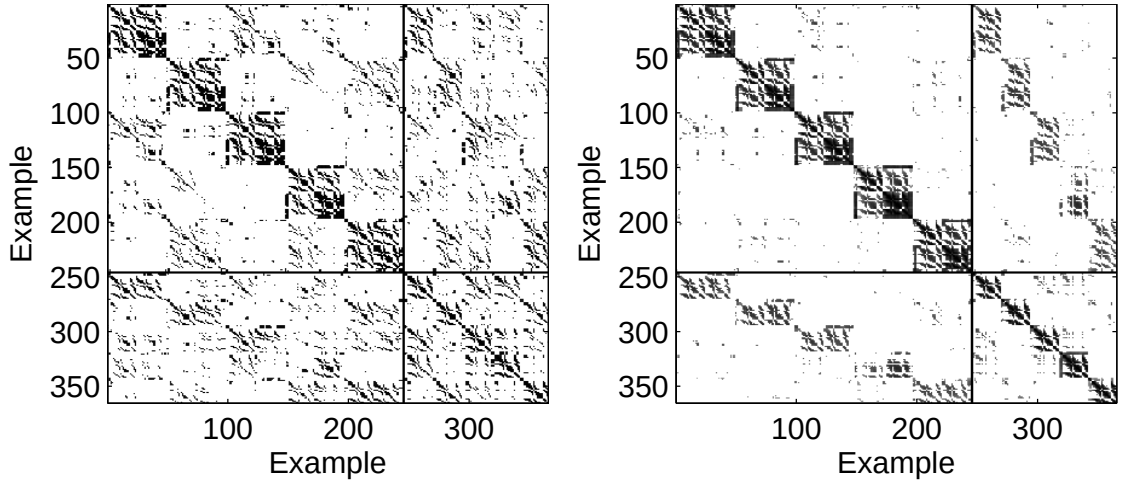


(a) TCA 嵌入特征表示　　(b) ARPCA 嵌入特征表示

图 3.8　有效性验证：ARPCA 在文本分类任务 *rec vs sci 1* 上计算得到的嵌入特征表示。

3.5.4.3　嵌入特征表示

为了从直观上判断不同迁移降维方法得到的嵌入特征表示的好坏，在文本分类任务 *rec vs sci 1* 上分别以最佳参数执行 TCA 和 ARPCA 算法，然后对嵌入表征的前 2 维（对应最大本征值的本征向量）进行可视化。图 3.8(a) 和图 3.8(b) 分别显示了在 TCA 和 ARPCA 嵌入特征表示下，辅助领域（左）和目标领域（右）的数据分布散点图。可以看到，在 TCA 嵌入特征表示下，目标领域的正反样例重叠在一起、难以区分开来，这对后续的分类任务造成极大困难；在 ARPCA 嵌入特征表示下，目标领域的正反样例相互远离，区分性非常明显，这对后续的分类任务十分有利。该任务上的分类准确率也证实了这一点：TCA 仅取得 73.50% 的准确率，而 ARPCA 取得了 95.55% 的准确率。可见，ARPCA 在领域间同时进行边缘分布适配和条件分布适配，这使 ARPCA 能够抽取比 TCA（包括其他迁移降维方法）判别性更好的嵌入特征表示，因而极大地提高了迁移模型的跨领域泛化能力。



(a) TCA 嵌入表征计算得的相似矩阵　　(b) ARPCA 嵌入表征计算得的相似矩阵

图 3.9　有效性验证：在人脸识别任务 *PIE1 vs PIE2* 上，嵌入特征表示计算得的相似矩阵。

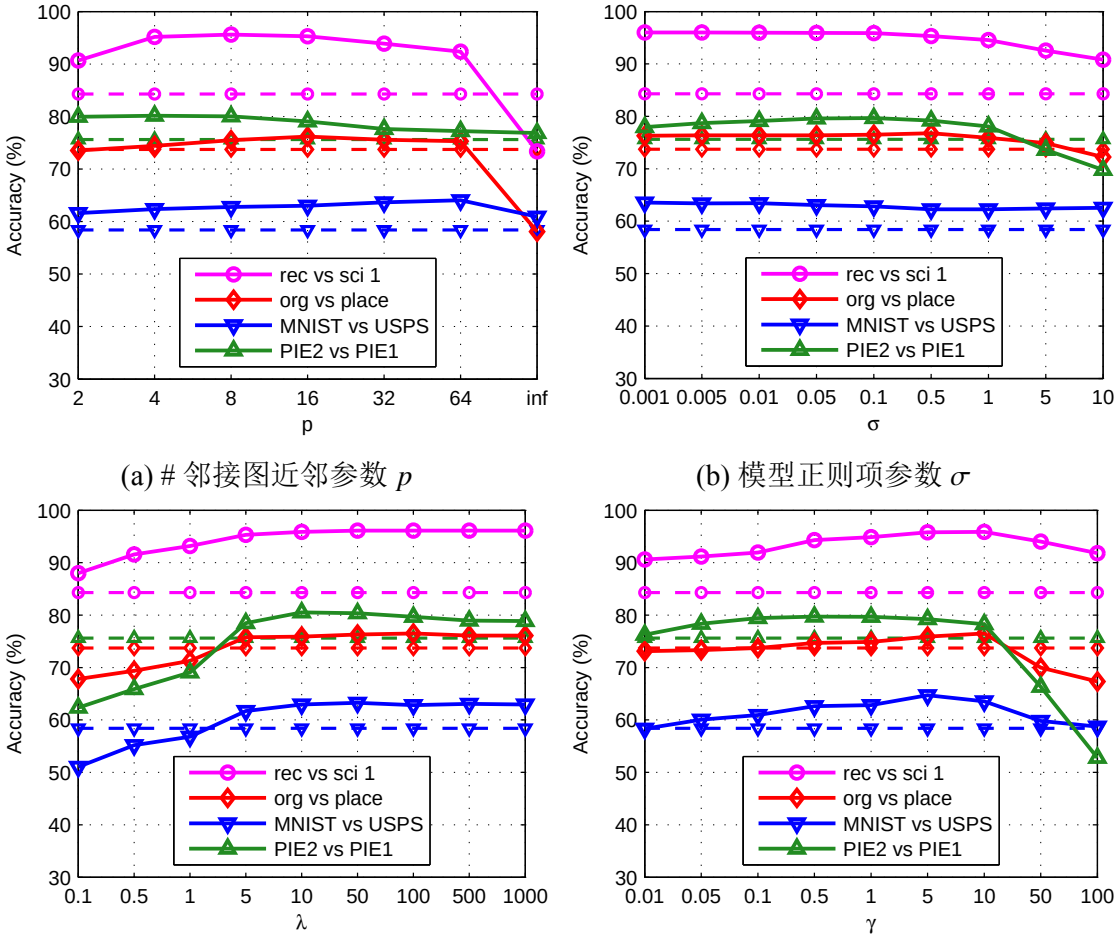
3.5.4.4　嵌入相似矩阵

在人脸识别任务 *PIE1 vs PIE2* 上分别以最优参数执行 TCA 和 ARPCA 算法，然后分别在其嵌入表征矩阵 $\mathbf{Z} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$ 上计算 20-近邻图相似矩阵。为使可视化更为直观，仅采用了前 5 个人物的 365 张人脸图片，其中 245 张来自辅助领域（某一面部朝向）、其余 120 张来自目标领域（另一面部朝向）。相应地，在相似矩阵中，左上角、右下角子矩阵表示领域内相似度，右上角、左下角子矩阵表示领域间相似度；此外，整个相似矩阵的对角块表示领域内类别内相似度，右上角、左下角子矩阵中的对角块表示领域间类别内相似度，其余矩阵块表示类别间相似度。

在 TCA 和 ARPCA 的嵌入表征上计算得到的相似矩阵分别如图 3.9(a) 和图 3.9(b) 所示。从理论上说，一个有助于跨领域分类的理想嵌入表征应该满足：（1）领域间相似度足够高，使得跨领域知识迁移得以实现；（2）类别间相似足够低，使得类别间判别更为容易。从这个意义上看，TCA 没有能够抽取足够好的嵌入表征，这表明仅做边缘分布适配不足以实现有效的迁移学习。本章方法 ARPCA 可以抽取到满足上述条件的理想嵌入表征，因此可以达到良好的跨领域的泛化能力。

3.5.5　参数敏感性分析

本节执行系统的参数敏感性实验，证明 ARRLS、ARPCA 方法可以在相当大的参数范围内稳定地获得比最佳基准方法更好分类准确率。



(a) # 邻接图近邻参数 p　　(b) 模型正则项参数 σ

(c) 适配正则项参数 λ　　(d) 流形正则项参数 γ

图 3.10　参数分析：ARRLS 在所选分类任务的性能表现（虚线表示最佳基准方法效果）。

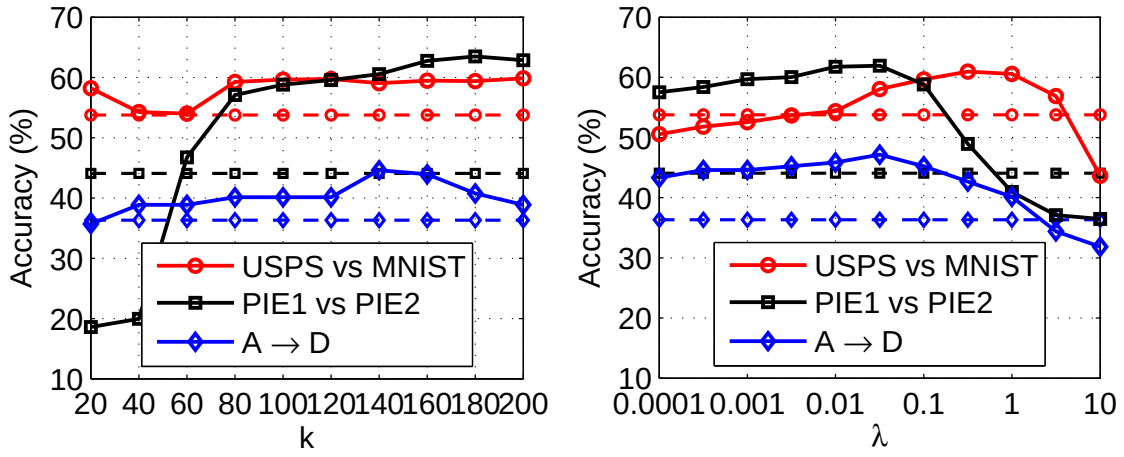
3.5.5.1　ARRLS 参数敏感性

首先考察 ARRLS 方法，这里限于空间，仅从 20-Newsgroups、Reuters-21578、USPS & MNIST 和 PIE 数据集中分别随机选取一个分类任务，观察参数的敏感性。

近邻参数 p： 选取不同的参数 p 值，执行 ARRLS 算法。理论上，p 应取折中值，因为太大的 p 值会导致稠密图（连接两个并不相似的样例），太小的 p 值会导致稀疏图（无法捕捉样例间的相似信息）。图 3.10(a) 展示了分类准确率随着参数 p 的变化规律，可以看到 $p \in [4, 64]$ 是合理的取值范围，实践中一般固定 $p = 10$。

模型正则项 σ： 选取不同的参数 σ 值，执行 ARRLS 算法。理论上，σ 控制了分类模型的复杂性，当 $\sigma \to 0$ 时分类模型退化而导致严重的过拟合问题，当 $\sigma \to \infty$ 时分类模型过简单而无法充分拟合数据的判别结构。图 3.10(b) 展示了分类准确率随着参数 σ 的变化规律，从中可以看到 $\sigma \in [0.001, 1]$ 是合理的取值范围。

适配正则项 λ： 选取不同的参数 λ 值，执行 ARRLS 算法。理论上，较大的 λ 值会使概率分布的适配程度更高，当 $\lambda \to 0$ 时分布差异未能得到减小，并导致



(a) 子空间维度 k　　(b) 适配正则项 λ

图 3.11　参数敏感性：ARPCA 在字符识别、人脸识别、对象识别等三类任务上的性能（虚线表示最佳基准方法的效果）。

负迁移问题。图 3.10(c) 展示了分类准确率随着参数 λ 的变化规律，从中可以看到 $\lambda \in [5, 1000]$ 是合理的取值范围，实践中一般取较大值 $\lambda = 10$。

流形正则项 γ： 选取不同的参数 γ 值，执行 ARRLS 算法。理论上，较大的 γ 值会使流形一致性在 ARRLS 中占据更重要的权重，当 $\gamma \to \infty$ 时仅有流形一致性得到保持，而标注等判别信息却丢失了。图 3.10(d) 展示了分类准确率随着参数 γ 的变化规律，可以看到 $\gamma \in [0.1, 10]$ 是合理的取值范围，实践中可由交叉验证确定。

3.5.5.2　ARPCA 参数敏感性

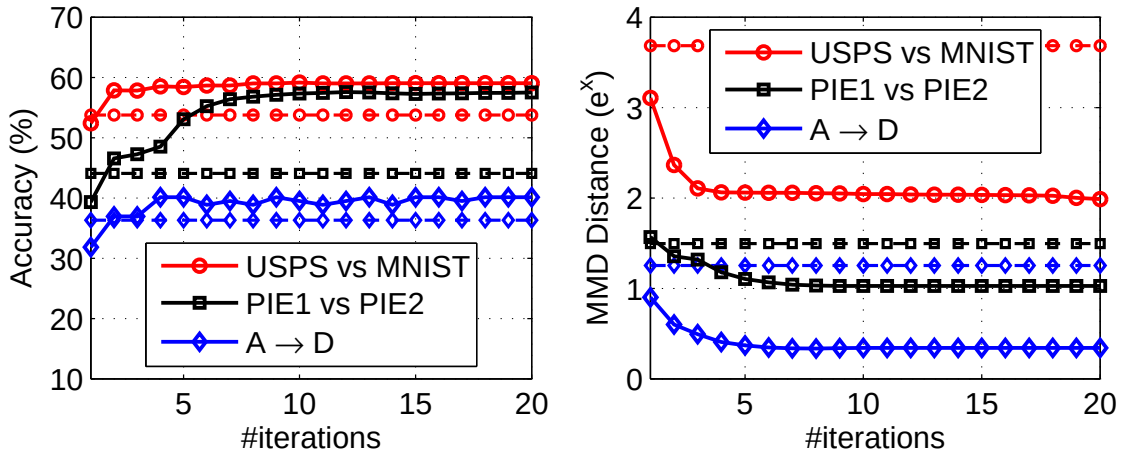
限于篇幅，对 ARPCA 方法，这里仅在字符识别任务 *USPS vs MNIST*、人脸识别任务 *PIE1 vs PIE2* 以及对象识别任务 $A \to D$ 上讨论参数敏感性结果。

在不同的 k 值下执行 ARPCA 方法，k 可选取为数据重构误差足够小的值（如 100）。图 3.11(a) 显示了分类准确率相对于参数 k 的变化规律，可选取 $k \in [60, 200]$。

在不同的 λ 值下执行 ARPCA 方法。理论上，λ 决定了适配正则项的权重，当 $\lambda \to 0$ 时优化模型退化为平凡解，当 $\lambda \to \infty$ 时分布适配程度明显不足。图 3.11(b) 显示了分类准确率随参数 λ 的变化规律，可取 $\lambda \in [0.001, 1.0]$ 为合理的取值范围。

3.5.5.3　ARPCA 收敛性分析

由于 ARPCA 是迭代式算法，这里通过实验确认其实际收敛速度和质量。图 3.12(a) 和图 3.12(b) 分别给出了分类准确率和联合分布距离随着迭代次数的变化规律。可以看到 ARPCA 随着迭代快速收敛，并在 10 次迭代以内达到局部极值。



(a) 分类准确率 k　　(b) 联合分布距离 λ

图 3.12　收敛性分析：ARPCA 在字符识别、人脸识别、对象识别等三类任务上的性能。

3.6　小结

本章提出了迁移学习通用框架，适配正则化迁移学习（ARTL），以及四种迁移学习模型，适配正则化线性回归（ARRLS）、适配正则化支持向量机（ARSVM）、适配正则化主成份分析（ARPCA）和适配正则化正交映射（AROM）。该框架同时对结构风险泛函、边缘分布适配和条件分布适配、流形一致性进行优化，从而在半监督学习风范下完成了迁移学习模型训练。该框架的优势是考察了各种必要的学习准则，并保持了模型的凸性以及简单可依赖特性。此外，多种现有监督学习方法如正则化线性回归（RLS）、支持向量机（SVM），表征学习方法如主成份分析（PCA）、独立成分分析（ICA），都可以直接载入到 ARTL 框架实现迁移学习。ARTL 对跨领域概率分布失配问题具有鲁棒性，并能极大的提高跨领域分类任务的准确率。在 252 个文本、图像分类任务上的系统性实验证明了本章方法相对于领域前沿迁移学习方法的优势。未来工作包括（1）基于统计学习理论，进一步分析本章的联合分布适配正则化的泛化性能保证（2）将本章模型推广到回归任务。

第 4 章　领域不变核学习方法

迁移学习在不同领域的服从不同概率分布的训练数据和测试数据之间推广学习模型。上一章介绍了如何通过最小化领域间边缘分布和条件分布的差异来获得更紧凑的泛化误差上界。要在样本输入空间中进行概率分布适配，由于线性映射仅能刻划概率分布之间的一阶矩差异，现有工作通常依赖于非线性核映射来实现概率分布之间的二阶矩和高阶矩匹配。然而，这类非线性核空间一般不是标准核机器（如支持向量机）所依赖的最优核空间，这导致分布适配与模型训练不能同时达到最优解，带来潜在的欠适配问题。为了解决该问题，本章提出了迁移核学习（Transfer Kernel Learning，TKL），直接在可再生希尔伯特空间中学习一个领域不变核矩阵实现辅助领域和目标领域的分布适配。具体地，首先由 Mercer 定理将目标领域本征系统外插值到辅助领域得到谱核矩阵族，其次选取与辅助领域真实核矩阵近似误差最小的谱核矩阵来构造领域不变核机器。在文本分类、图像识别、视频识别等大量任务上的系统性实验证明了本章方法相对于已有前沿方法的优势。

4.1　引言

大量研究工作表明，当标准监督学习模型在与训练数据不同的测试数据上预测时，因性能严重下降而不能满足实用需求[108,110]。这种训练数据与测试数据概率分布失配的问题在实际中广泛存在，如计算机视觉[109,111]、自然语言处理[35,42]以及其他领域。跨领域迁移学习任务包括两种不同类型的数据集，一种来自辅助领域、另一种来自目标领域。辅助领域包含大量标注数据足以训练准确的分类器，目标领域包含大量无标数据、服从与辅助领域显著不同但又潜在相关的概率分布。学习目标是修正领域间的概率分布失配，使得标准分类器可以在领域间有效迁移。

通过参数化或非参数化距离的最小化实现领域间的概率分布适配，是迁移学习的重要方法[27,33,45,46,112]。主要思想是：学习隐含特征表示或实例权重，使得辅助领域和目标领域间出现共享特性、并显式地修正分布失配。例如文献[45]提出在参数化核空间中学习一组迁移主成份，使得辅助领域和目标领域在核空间中的概率分布期望可以匹配。常用的概率分布距离度量函数包括相对熵[112]、布雷格曼散度[46]和最大均值差异（Maximum Mean Discrepancy，MMD）[27,33,45]。然而，使用 KL 和布雷格曼散度需要先进行分布密度估计，使用 MMD 距离需要先定义参数化非线性核函数，而该核函数对目标核机器如支持向量机往往并非最优核空间。

为了解决分布适配与核机器对核函数的最优参数选取不一致的问题，文

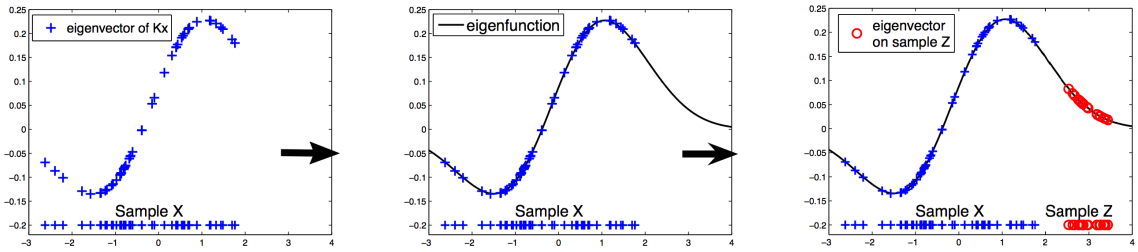


图 4.1　迁移核学习基本原理[29]。左图：计算目标领域核矩阵的离散本征向量；中图：估计该离散本征向量对应的连续本征函数；右图：在辅助领域样例上计算本征函数外插值。

献[29,40,85,108] 直接学习一个领域不变核矩阵、而不需要预设核函数参数。目标是同时最小化经验风险泛函与 MMD 准则，由半正定优化（Semi-Definite Programming，SDP）[40] 或多核学习（Multiple Kernel Learning，MKL）[85] 得到核矩阵。缺陷在于，标准 SDP 求解器的计算复杂度达 $O(n^{6.5})$，这对实际问题是计算不可行的；MKL 将复杂的概率分布差异编码到一小组核集成参数中，但这可能并不足以修正分布失配。为解决该问题，文献[29] 提出*代理核匹配*（Surrogate Kernel Matching，SKM）方法，直接在可再生希尔伯特空间中匹配辅助领域和目标领域的核矩阵。这种方法具有理论上的优势，因为领域间的分布差异直接由其核矩阵差异来形式化。该方法的一个主要缺陷在于，直接将整个辅助领域核矩阵线性地映射到目标领域本征空间中，也即用目标领域本征空间表征所有数据，会导致较大的数据近似误差。

受上述讨论的启发，本章提出*迁移核学习*（Transfer Kernel Learning，TKL）方法，在可再生希尔伯特空间中直接学习一个领域不变核矩阵，从而实现辅助领域与目标领域核矩阵的匹配。基本思想是：领域间概率分布差异可以由 Nyström 近似误差[113] 进行形式化，该误差刻划了辅助领域插值核矩阵与真实核矩阵之间的差异。图 4.1 给出如何将目标领域核矩阵的本征系统外插值到辅助领域实例空间的基本原理。TKL 方法的学习过程是：首先由 Mercer 定理[114] 将目标领域本征系统外插值到辅助领域得到谱核矩阵族，其次选取与辅助领域真实核矩阵近似误差最小的谱核矩阵来构造领域不变核机器。这样的领域不变核矩阵能够同时刻划目标领域本征系统并最小化辅助领域近似误差。本章的主要创新性贡献总结如下：

- 首次提出将 Nyström 近似误差作为领域间概率分布差异的度量准则，并在谱核设计框架下通过最小化 Nyström 近似误差来学习领域不变核矩阵。
- 所学领域不变核矩阵可直接带入现成的核机器如支持向量机（Support Vector Machine）、核岭回归（Kernel Ridge Regression）等，实现有效迁移学习。
- 在文本分类、图像识别、视频识别的大量实际任务上证明了本章方法优势。

4.2　预备知识

4.2.1　最大均值差异

概率分布在训练数据和测试数据之间的显著差异，是制约标准监督学习方法泛化性能的主要挑战。因此提高迁移学习泛化性能的关键方法之一，就是对概率分布差异进行形式化并对差异函数进行最小化。尽管参数化距离函数已被广泛应用于迁移学习并取得了良好的效果，例如相对熵[112]和布雷格曼散度[46]等，但是度量这类参数化距离函数通常要依赖非平凡的分布密度估计过程，这极大地增加了机器学习模型设计的复杂性。针对上述问题，文献[33]提出了称为最大均值差异（Maximum Mean Disrepancy，MMD）的非参数化距离函数，用来度量两个概率分布函数 P 和 Q 之间的差异。给定由 P 和 Q 采样生成的两个有限样本集 $\mathcal{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ 和 $\mathcal{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_m\}$，MMD 定义为将 $\mathcal{X}$ 和 $\mathcal{Z}$ 同时映射到一个高维的可再生希尔伯特空间 $\mathcal{H}$ 后的期望差异。MMD 及其经验估计分别定义如下[33]：

$$\begin{aligned}\mathrm{MMD}\left[P,Q\right] &\triangleq \sup_{\|f\|_{\mathcal{H}}\leq 1}\left(\mathbb{E}_{\mathbf{x}\sim P}\left[f\left(\mathbf{x}\right)\right]-\mathbb{E}_{\mathbf{z}\sim Q}\left[f\left(\mathbf{z}\right)\right]\right)\\ \mathrm{MMD}\left[\mathcal{X},\mathcal{Z}\right] &\triangleq \left\|\frac{1}{n}\sum_{i=1}^{n}\phi\left(\mathbf{x}_i\right)-\frac{1}{m}\sum_{j=1}^{m}\phi\left(\mathbf{z}_j\right)\right\|_{\mathcal{H}}\end{aligned} \tag{4-1}$$

其中 $f(\cdot)$ 是 $\mathcal{H}$ 中的任意函数，$\phi:\mathbf{x}\mapsto\mathcal{H}$ 是将原始数据映射到高维空间的非线性特征映射函数。设核函数为 $k(\mathbf{x}_i,\mathbf{x}_j)$，则有如下内积关系 $k\left(\mathbf{x}_i,\mathbf{x}_j\right)=\left\langle\phi\left(\mathbf{x}_i\right),\phi\left(\mathbf{x}_j\right)\right\rangle$。

定理 4.1 (**MMD**)：[33] 设 P，Q 为 *Borel* 概率测度，并设 $\mathcal{H}$ 为通用可再生希尔伯特空间，则 $\mathrm{MMD}[P,Q]=0$ 当且仅当 $P=Q$。

通过对非线性特征映射 $\phi(\mathbf{x})$ 进行泰勒级数展开（实际中不会显式展开此级数，而是通过核技巧间接实现），MMD 可以在高维空间中刻划概率分布的任意阶统计量，如一阶统计量均值和二阶统计量方差。换言之，最小化 MMD 可以将训练数据和测试数据的概率分布 P 和 Q 通过各阶统计量进行充分适配。MMD 被广泛应用于设计迁移学习算法，如实例权重调整[27]、特征变换[45]和分类器调整[85]。然而，这些方法无一例外地将 MMD 作为正则项加入到学习器的损失函数中，这会导致 MMD 的最小化过程无法达到最优值，原因是学习器的损失函数和 MMD 一般不存在公共的最优解。本章提出了领域不变核学习方法，以期解决上述问题。

4.2.2　Nyström 近似

本章通过学习一个领域不变核函数或核矩阵来解决基于 MMD 的概率分布适配方法的缺陷，其理论基础是 Nyström 近似[115]。Nyström 近似是核函数低秩近似的主要方法，它基于 Mercer 定理[114]用已知核的本征系统对目标核进行低秩近似。

定理 4.2 (**Mercer**)：[114] 设 $k(\mathbf{z},\mathbf{x})$ 为 $P(\mathbf{x})$ 二次可积的连续对称非负正定函数，则

$$k(\mathbf{z},\mathbf{x})=\sum_{i=1}^{\infty}\lambda_i\phi_i(\mathbf{z})\phi_i(\mathbf{x}) \tag{4-2}$$

本征值 λ_i 和标准正交本征函数 $\phi_i(\mathbf{x})$ 是下列积分方程的解

$$\int k(\mathbf{z},\mathbf{x})\phi_i(\mathbf{x})P(\mathbf{x})\mathrm{d}\mathbf{x}=\lambda_i\phi_i(\mathbf{z}) \tag{4-3}$$

该定理是可再生希尔伯特空间的理论基础，它指出任意正定核均可被其他已知核有效重构，并提供在任意数据集上生成核矩阵的机制。具体地说，给定由概率分布 $P(\mathbf{x})$ 采样而来的数据集 $\mathcal{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ 及其核矩阵 $\mathbf{K}_{\mathcal{X}}$，可通过公式 (4-3) 计算其本征值 $\{\lambda_i\}$ 和本征函数 $\{\phi_i(\mathbf{x})\}$。由此可在任意数据集 $\mathcal{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_m\}$ 上求取核函数 k，重构生成的核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 仅依赖于已知核矩阵 $\mathbf{K}_{\mathcal{X}}$ 的本征系统 $\{\lambda_i,\phi_i(\mathbf{x})\}$。

Nyström 法[115] 对公式 (4-3) 中的积分函数用如下的经验估计进行低秩近似：

$$\sum_{j=1}^{n}\frac{k(\mathbf{z},\mathbf{x}_j)\phi_i(\mathbf{x}_j)}{n}\simeq\lambda_i\phi_i(\mathbf{z}) \tag{4-4}$$

遍历和式中的变量 $\mathbf{z}\in\mathcal{X}$ 得到标准本征分解（量纲差一常数）$\mathbf{K}_{\mathcal{X}}\boldsymbol{\Phi}_{\mathcal{X}}=\boldsymbol{\Phi}_{\mathcal{X}}\boldsymbol{\Lambda}_{\mathcal{X}}$，其中 $\boldsymbol{\Phi}_{\mathcal{X}}\in\mathbb{R}^{n\times n}$ 为 n 个本征向量，$\boldsymbol{\Lambda}_{\mathcal{X}}=\mathrm{diag}(\lambda_1,\ldots,\lambda_n)\in\mathbb{R}^{n\times n}$ 为 n 个本征值。公式 (4-3) 的本征函数 $\{\phi_i(\mathbf{x})\}$ 和本征值 $\{\lambda_i\}$ 可分别由 $\boldsymbol{\Phi}_{\mathcal{X}}$ 的列向量和 $\boldsymbol{\Lambda}_{\mathcal{X}}$ 的对角元离散地近似，而本征函数 $\phi_i(\mathbf{z})$ 在任意数据点 $\mathbf{z}$ 的值可由 $\phi_i(\mathbf{z})=\sum_{j=1}^{n}\frac{k(\mathbf{z},\mathbf{x}_j)\phi_i(\mathbf{x}_j)}{n\lambda_i}$ 插值得到。在概率分布 $Q(\mathbf{z})$ 采样得到的数据集 $\mathcal{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_m\}$ 上求取所有本征函数 $\{\phi_i(\mathbf{z})\}$ 的值，则得到数据集 $\mathcal{Z}$ 的核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 对应的本征函数 $\boldsymbol{\Phi}_{\mathcal{Z}}$ 的如下离散近似：

$$\boldsymbol{\Phi}_{\mathcal{Z}}\simeq\mathbf{K}_{\mathcal{Z}\mathcal{X}}\boldsymbol{\Phi}_{\mathcal{X}}\boldsymbol{\Lambda}_{\mathcal{X}}^{-1}, \tag{4-5}$$

其中 $\mathbf{K}_{\mathcal{Z}\mathcal{X}}\in\mathbb{R}^{m\times n}$ 是跨越数据集 $\mathcal{Z}$ 和 $\mathcal{X}$ 的交叉相似度矩阵，可由核函数 k 在 $\mathcal{Z}$ 和 $\mathcal{X}$ 之间求值得到。根据 Mercer 定理，数据集 $\mathcal{Z}$ 上的核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 可低秩近似如下：

$$\mathbf{K}_{\mathcal{Z}}\simeq\boldsymbol{\Phi}_{\mathcal{Z}}\boldsymbol{\Lambda}_{\mathcal{X}}\boldsymbol{\Phi}_{\mathcal{Z}}^{\mathsf{T}}=\mathbf{K}_{\mathcal{Z}\mathcal{X}}\mathbf{K}_{\mathcal{X}}^{-1}\mathbf{K}_{\mathcal{X}\mathcal{Z}} \tag{4-6}$$

如果分别记 $\mathcal{Z}$ 和 $\mathcal{X}$ 为辅助数据和目标数据，则由于迁移学习的领域差异性，变量 $\mathbf{z}\sim Q(\mathbf{z})$ 和 $\mathbf{x}\sim P(\mathbf{x})$ 将服从不同的概率分布，即 $P\neq Q$。这种情况下，由 $\mathbf{K}_{\mathcal{X}}$ 的本征系统对 $\mathbf{K}_{\mathcal{Z}}$ 进行低秩近似会导致很大的拟合误差。文献[29] 在应用 Nyström 法设计迁移学习方法时未能充分考虑到该问题，从而可能会降低跨领域泛化性能。

4.2.3　谱核学习

核学习是度量学习的一种重要方法，包括非参数化核变换[116]、非参数化核学习[117] 等。本章采用谱核学习[118]，由已知核的本征向量对目标核进行线性重构。

定理 4.3 (**谱核学习**)：[118] 如果给定一个已知正定核矩阵 $\mathbf{K} \in \mathbb{R}^{n\times n}$ 的本征系统 $\{\gamma_i, \boldsymbol{\phi}_i\}_{i=1}^n, \gamma_1 \geq \ldots \geq \gamma_n \geq 0$，则以下矩阵族

$$\mathbf{K}_\lambda = \sum\nolimits_{i=1}^{n} \lambda_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T, \lambda_1 \geq \ldots \geq \lambda_n \geq 0 \tag{4-7}$$

是以 $\{\lambda_i\}$ 为本征值的正定核矩阵 $\mathbf{K}_\lambda$。

重构的核矩阵 $\mathbf{K}_\lambda$ 是多个本征核 $\left\{\mathbf{K}_i = \boldsymbol{\phi}_i \boldsymbol{\phi}_i^{\mathsf{T}}\right\}_{i=1}^n$ 的线性组合，但这不同于自动组合多个已知核的多核学习[119]。要完成谱核学习，算法上可以通过优化合适的准则来确定最优本征谱 $\{\lambda_i\}$，如核对齐[120] 和图对齐[121] 准则。本章将提出如何通过优化辅助核与目标核之间的 Nyström 近似误差来求领域不变核，实现迁移学习。

4.3　迁移核学习

本节首先对所研问题进行形式化，其次提出迁移核学习模型 TKL、二次优化问题及其学习算法，最后形式化地分析算法的计算复杂度和近似误差上界。

4.3.1　问题定义

简明起见，本章常用的符号及其描述总结如表 4.1 所示，首先给出问题定义。

定义 4.1 (**领域、任务**)：[1] 领域 $\mathcal{D}$ 由 d 维特征空间 $\mathcal{F}$ 和边缘概率分布 $P(\mathbf{x})$ 组成，即 $\mathcal{D} = \{\mathcal{F}, P(\mathbf{x})\}, \mathbf{x} \in \mathcal{F}$。给定领域 $\mathcal{D}$，任务 $\mathcal{T}$ 由标签集合 $\mathcal{Y}$ 和分类模型 $f(\mathbf{x})$ 组成，即 $\mathcal{T} = \{\mathcal{Y}, f(\mathbf{x})\}$，$y \in \mathcal{Y}$，从统计观点 $f(\mathbf{x}) = Q(y|\mathbf{x})$ 可解释为条件概率分布。

一般地，如果两个领域 $\mathcal{Z}$ 和 $\mathcal{X}$ 的特征空间或边缘分布不同，则认为两个领域 $\mathcal{Z}$ 和 $\mathcal{X}$ 不同，即 $\mathcal{F}_\mathcal{Z} \neq \mathcal{F}_\mathcal{X} \vee P(\mathbf{z}) \neq P(\mathbf{x})$。类似地，如果两个任务 $\mathcal{T}_\mathcal{Z}$ 和 $\mathcal{T}_\mathcal{X}$ 的标签集合或条件分布不同，则认为两个任务 $\mathcal{T}_\mathcal{Z}$ 和 $\mathcal{T}_\mathcal{X}$ 不同，即 $\mathcal{Y}_\mathcal{Z} \neq \mathcal{Y}_\mathcal{X} \vee P(y|\mathbf{z}) \neq P(y|\mathbf{x})$。

问题 4.1 (**迁移核学习**)：给定标注的辅助领域 $\mathcal{Z} = \{(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_m, y_m)\}$ 和无标的目标领域 $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 满足 $\mathcal{F}_\mathcal{Z} = \mathcal{F}_\mathcal{X}, \mathcal{Y}_\mathcal{Z} = \mathcal{Y}_\mathcal{X}, P(\mathbf{z}) \neq P(\mathbf{x}), P(y|\mathbf{z}) \simeq P(y|\mathbf{x})$，学习 $\mathcal{Z}$ 和 $\mathcal{X}$ 之间的不变核，使得标准核机器 f 经 $\mathcal{Z}$ 训练后可以准确地泛化到 $\mathcal{X}$。

给定核函数 k，如高斯核 $k(\mathbf{z}, \mathbf{x}) = e^{-\gamma\|\mathbf{z}-\mathbf{x}\|^2}$ 或拉普拉斯核 $k(\mathbf{z}, \mathbf{x}) = e^{-\gamma|\mathbf{z}-\mathbf{x}|}$，可以计算辅助核矩阵 $\mathbf{K}_\mathcal{Z}$ 和目标核矩阵 $\mathbf{K}_\mathcal{X}$，并由此构造学习领域不变核矩阵 $\overline{\mathbf{K}}_{\mathcal{Z}\cup\mathcal{X}}$。

表 4.1　符号及其描述。

符号	描述	符号	描述	符号	描述
$\mathcal{Z}$	辅助领域	$\mathcal{X}$	目标领域	$\mathbf{X}$	数据矩阵
m	辅助样例数	n	目标样例数	$\mathbf{K}$	核矩阵
d	共享特征数	c	共享类别数	$\mathbf{\Phi}$	本征向量矩阵
r	本征谱数	ζ	阻尼系数	$\mathbf{\Lambda}$	本征值矩阵

4.3.2　优化问题

在可再生希尔伯特空间中处理概率分布差异是一个非平凡过程，因为核映射 ϕ 通常不能用显式表达表出。为使得两个不同数据集（如辅助数据集 $\mathcal{Z}$ 和目标数据集 $\mathcal{X}$）服从相似的概率分布即 $P(\phi(\mathbf{z})) \simeq P(\phi(\mathbf{x}))$，其充分条件是使它们有相似的核矩阵即 $\mathbf{K}_{\mathcal{Z}} \simeq \mathbf{K}_{\mathcal{X}}$[29]。然而核矩阵是数据依赖的，不同数据集的核矩阵有不同的维度 $\mathbf{K}_{\mathcal{Z}} \in \mathbb{R}^{m\times m}, \mathbf{K}_{\mathcal{X}} \in \mathbb{R}^{n\times n}$，因而不能直接计算不同核矩阵的相似度。为解决这个问题，本章采用 *Nyström 低秩近似法*[115]，通过目标核矩阵 $\mathbf{K}_{\mathcal{X}}$ 的本征系统插值生成辅助核矩阵 $\overline{\mathbf{K}}_{\mathcal{Z}} \in \mathbb{R}^{m\times m}$；插值的辅助核矩阵 $\overline{\mathbf{K}}_{\mathcal{Z}}$ 与真实的辅助核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 具有相同维度，因而两者可通过谱核学习进行适配。图 4.2 直观描述了上述过程。

4.3.2.1　本征系统插值

本章采用 Nyström 低秩近似对本征系统进行外插值。为此目的，首先对目标核矩阵 $\mathbf{K}_{\mathcal{X}}$ 进行本征分解：

$$\mathbf{K}_{\mathcal{X}}\mathbf{\Phi}_{\mathcal{X}} = \mathbf{\Phi}_{\mathcal{X}}\mathbf{\Lambda}_{\mathcal{X}} \tag{4-8}$$

由此得到目标核矩阵 $\mathbf{K}_{\mathcal{X}}$ 的本征系统 $\{\mathbf{\Lambda}_{\mathcal{X}}, \mathbf{\Phi}_{\mathcal{X}}\}$。其次，利用 Mercer 定理计算该本征系统在辅助数据集 $\mathcal{Z}$ 的取值，则可得到辅助核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 的本征向量的插值近似：

$$\overline{\mathbf{\Phi}}_{\mathcal{Z}} \simeq \mathbf{K}_{\mathcal{Z}\mathcal{X}}\mathbf{\Phi}_{\mathcal{X}}\mathbf{\Lambda}_{\mathcal{X}}^{-1} \tag{4-9}$$

其中 $\mathbf{K}_{\mathcal{Z}\mathcal{X}} \in \mathbb{R}^{m\times n}$ 是跨越领域 $\mathcal{Z}$ 和 $\mathcal{X}$ 的交叉核矩阵，可通过核函数 k 计算得到。

在标准 Nyström 法中，插值得到的辅助领域本征向量 $\overline{\mathbf{\Phi}}_{\mathcal{Z}}$ 及目标领域本征

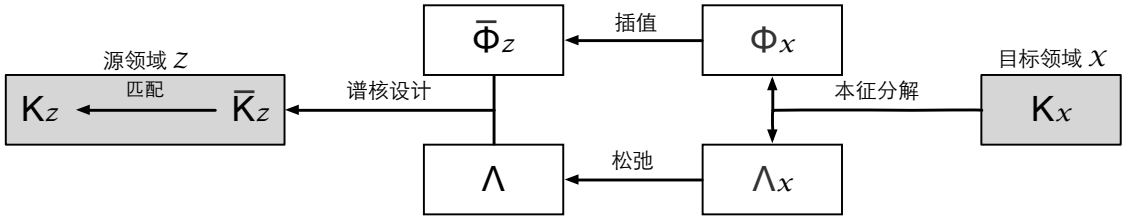


图 4.2　迁移核学习的执行过程示意图，包括本征分解、本征插值、谱核学习三部分。

值 $\mathbf{\Lambda}_{\mathcal{X}}$ 直接用于对辅助领域核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 进行近似，如公式 (4-6) 所示。本章指出，Nyström 近似法 (4-6) 当且仅当 $P(\mathbf{z}) \simeq P(\mathbf{x})$ 时有效。当辅助领域和目标领域服从不同的概率分布 $P(\mathbf{z}) \neq P(\mathbf{x})$ 时，Nyström 近似误差将会扩大并导致跨领域核近似变得不合理。因此，文献[29] 利用标准 Nyström 法设计的迁移学习算法并不有效。

上述论述给出一个重要的观点：Nyström 近似误差从本质上刻划了领域之间的分布差异，当且仅当 $P(\mathbf{z}) \simeq P(\mathbf{x})$ 时误差趋零。因此，如果可以找到一个插值核矩阵使得 Nyström 近似误差最小化，则该插值核矩阵将能挖掘领域之间潜在的不变结构并可用于提高迁移学习的泛化性能。与该结论相关的理论分析参见4.3.4节。

4.3.2.2　本征值松弛

为最小化领域间分布差异，本章采用谱核学习[118] 方法由插值得到的近似本征系统 (4-9) 重构生成原始核矩阵。生成核矩阵保持了原始核矩阵的本征向量等重要结构信息，但其本征值松弛为任意合法取值。通过最小化 Nyström 近似误差可以确定最小化分布差异的最优本征值，从而将生成核矩阵改造成领域不变核矩阵。

将标准 Nyström 法 (4-6) 的本征谱 $\mathbf{\Lambda}_{\mathcal{X}}$ 松弛为待学习参数 $\mathbf{\Lambda}$，经过谱核设计得到一族由目标领域核矩阵的本征系统外插值到辅助领域数据集上生成的核矩阵：

$$\overline{\mathbf{K}}_{\mathcal{Z}} = \overline{\mathbf{\Phi}}_{\mathcal{Z}} \mathbf{\Lambda} \overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}} \tag{4-10}$$

该矩阵族保持了目标领域的关键性结构信息即本征向量 $\overline{\mathbf{\Phi}}_{\mathcal{Z}}$，从而形成了知识迁移的桥梁。但是共享本征向量并不一定会减小领域间的分布差异，除非针对性地对其进行最小化[45]。此外还需要定义一个恰当的目标函数对自由本征谱 $\mathbf{\Lambda}$ 进行参数估计。传统谱核学习[118] 通过将待定核与已知核进行对齐来确定未知参数 $\mathbf{\Lambda}$，由于不存在同领域内的已知核，本章通过匹配不同领域的核矩阵来实现参数学习。

4.3.2.3　近似误差最小化

本章通过最小化辅助领域插值核矩阵 $\overline{\mathbf{K}}_{\mathcal{Z}}$ 和辅助领域真实核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 之间的二次误差，实现领域间分布差异的最小化，得到迁移核学习的如下优化问题：

$$\begin{aligned} \min_{\mathbf{\Lambda}} \left\| \overline{\mathbf{K}}_{\mathcal{Z}} - \mathbf{K}_{\mathcal{Z}} \right\|_F^2 &= \left\| \overline{\mathbf{\Phi}}_{\mathcal{Z}} \mathbf{\Lambda} \overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}} - \mathbf{K}_{\mathcal{Z}} \right\|_F^2 \\ \lambda_i &\geq \zeta \lambda_{i+1}, i = 1, \ldots, n-1 \\ \lambda_i &\geq 0, i = 1, \ldots, n \end{aligned} \tag{4-11}$$

其中 $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ 是 n 个非负本征谱参数。由于正定核矩阵的本征谱服从幂律分布[120,122]，一组以 $\zeta \geq 1$ 为阻尼系数的本征值阻尼约束被加入到优化问题

中。原理上这组阻尼约束对优化问题的学习效果至关重要，因为它们可以将 $\mathbf{K}_{\mathcal{Z}}$ 的本征谱先验衰减趋势注入模型，同时使较大的本征向量（对应较大的本征值）对知识迁移做贡献更大。上述优化模型学习到的最优核矩阵被确定为领域不变的。

4.3.2.4　领域不变核机器

学习到最优本征谱参数 $\mathbf{\Lambda}$ 后，可在辅助领域和目标领域全集 $\mathcal{A}=\mathcal{Z}\cup\mathcal{X}$ 上构造领域不变核矩阵 $\overline{\mathbf{K}}_{\mathcal{A}}$。基于谱核设计，$\overline{\mathbf{K}}_{\mathcal{A}}$ 可由领域不变本征系统 $\left\{\mathbf{\Lambda},\overline{\mathbf{\Phi}}_{\mathcal{A}}\right\}$ 生成：

$$\overline{\mathbf{K}}_{\mathcal{A}}=\begin{bmatrix}\overline{\mathbf{\Phi}}_{\mathcal{Z}}\mathbf{\Lambda}\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}} & \overline{\mathbf{\Phi}}_{\mathcal{Z}}\mathbf{\Lambda}\mathbf{\Phi}_{\mathcal{X}}^{\mathsf{T}}\\ \mathbf{\Phi}_{\mathcal{X}}\mathbf{\Lambda}\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}} & \mathbf{\Phi}_{\mathcal{X}}\mathbf{\Lambda}\mathbf{\Phi}_{\mathcal{X}}^{\mathsf{T}}\end{bmatrix}=\overline{\mathbf{\Phi}}_{\mathcal{A}}\mathbf{\Lambda}\overline{\mathbf{\Phi}}_{\mathcal{A}}^{\mathsf{T}} \tag{4-12}$$

其中 $\overline{\mathbf{\Phi}}_{\mathcal{A}}\triangleq\left[\overline{\mathbf{\Phi}}_{\mathcal{Z}};\mathbf{\Phi}_{\mathcal{X}}\right]$ 为所有数据集 $\mathcal{A}$ 上的插值本征向量集合。该领域不变核矩阵 $\overline{\mathbf{K}}_{\mathcal{A}}$ 可直接输入到标准核机器如支持向量机中，实现跨领域的预测和泛化任务。

具体地，在辅助领域分区 $\overline{\mathbf{K}}_{\mathcal{Z}}=\overline{\mathbf{\Phi}}_{\mathcal{Z}}\mathbf{\Lambda}\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}}$ 上训练核支持向量机分类器，并将其泛化到跨领域分区 $\overline{\mathbf{K}}_{\mathcal{X}\mathcal{Z}}=\mathbf{\Phi}_{\mathcal{X}}\mathbf{\Lambda}\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}}$ 如下：

$$\mathbf{y}_{\mathcal{X}}=\overline{\mathbf{K}}_{\mathcal{X}\mathcal{Z}}\left(\boldsymbol{\alpha}\odot\mathbf{y}_{\mathcal{Z}}\right)+b \tag{4-13}$$

其中 $\boldsymbol{\alpha}$ 是核支持向量机的拉格朗日乘子，b 是分类面的截距。常用的开源工具 LIBSVM 中实现了核支持向量机，可接受核矩阵作为输入并输出训练和预测结果。

要对样本外测试数据集 $\mathcal{X}_o$（来自目标领域 $\mathcal{X}$ 且服从同样的概率分布）进行预测，需要将目标领域的本征系统 $\mathbf{\Phi}_{\mathcal{X}}$ 由标准 Nyström 法外插值到数据集 $\mathcal{X}_o$ 上：

$$\mathbf{\Phi}_{\mathcal{X}_o}=\mathbf{K}_{\mathcal{X}_o\mathcal{X}}\mathbf{\Phi}_{\mathcal{X}}\mathbf{\Lambda}_{\mathcal{X}}^{-1} \tag{4-14}$$

其中 $\mathbf{K}_{\mathcal{X}_o\mathcal{X}}$ 是跨数据集交叉核矩阵，样本外数据集对应的跨领域分区可计算为 $\overline{\mathbf{K}}_{\mathcal{X}_o\mathcal{Z}}=\mathbf{\Phi}_{\mathcal{X}_o}\mathbf{\Lambda}\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}}$。将该矩阵代入公式 (4-13) 可得到样本外数据集 $\mathcal{X}_o$ 的预测结果。

4.3.3　学习算法

本节设计基于凸优化理论的学习算法，并基于 Nyström 法给出可扩展性实现。

算法 5: 迁移核学习（TKL）

输入: 数据矩阵 $\mathbf{X}$；核函数 k；本征谱阻尼系数 ζ。

输出: 领域不变核矩阵 $\overline{\mathbf{K}}_{\mathcal{A}}$。

1 开始

2 用核函数 k 计算辅助领域、目标领域、跨领域核矩阵 $\mathbf{K}_{\mathcal{Z}}$，$\mathbf{K}_{\mathcal{X}}$、$\mathbf{K}_{\mathcal{Z}\mathcal{X}}$。

3 对目标领域核矩阵 $\mathbf{K}_{\mathcal{X}}$ 进行本征分解 (4-8)，得到本征系统 $\{\mathbf{\Lambda}_{\mathcal{X}}, \mathbf{\Phi}_{\mathcal{X}}\}$。

4 将上述本征向量插值到辅助领域 (4-9)，得辅助领域插值本征向量 $\overline{\mathbf{\Phi}}_{\mathcal{Z}}$。

5 求解二次规划问题 (4-15)，得到本征谱参数 $\boldsymbol{\lambda}$。

6 返回领域不变核矩阵 $\overline{\mathbf{K}}_{\mathcal{A}}$ 及其训练得到的核机器，如公式 (4-12) 所示。

4.3.3.1　二次规划问题

优化问题 (4-11) 可归结为线性约束的二次规划（QP）问题。记 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ 为 n 个本征谱参数且 $\mathbf{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$。根据线性代数，公式 (4-11) 的矩阵形式如下：

$$\begin{aligned} &\min_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{Q}\boldsymbol{\lambda} - 2\mathbf{r}^{\mathsf{T}}\boldsymbol{\lambda} \\ &\mathbf{C}\boldsymbol{\lambda} \geq \mathbf{0} \\ &\boldsymbol{\lambda} \geq \mathbf{0} \end{aligned} \tag{4-15}$$

这是个线性约束二次规划问题，可由常见的凸优化工具包求解，如 MATLAB$^{\text{TM}}$ 的 *quadprog* 函数。该二次规划的系数矩阵 $\mathbf{Q}$ 和 $\mathbf{r}$ 以及约束矩阵 $\mathbf{C}$ 分别推导如下：

$$\begin{aligned} \mathbf{Q} &= \left(\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}}\overline{\mathbf{\Phi}}_{\mathcal{Z}}\right) \odot \left(\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}}\overline{\mathbf{\Phi}}_{\mathcal{Z}}\right) \\ \mathbf{r} &= \operatorname{diag}\left(\overline{\mathbf{\Phi}}_{\mathcal{Z}}^{\mathsf{T}}\mathbf{K}_{\mathcal{Z}}\overline{\mathbf{\Phi}}_{\mathcal{Z}}\right) \\ \mathbf{C} &= \mathbf{I} - \zeta\bar{\mathbf{I}} \end{aligned} \tag{4-16}$$

其中 $\zeta \geq 1$ 是本征谱阻尼系数，它是 TKL 模型中的唯一可调参数；$\mathbf{I} \in \mathbb{R}^{n\times n}$ 是单位矩阵，$\bar{\mathbf{I}} \in \mathbb{R}^{n\times n}$ 是一阶对角矩阵，它的所有非零元素为 $\bar{I}_{i,i+1} = 1, i = 1, \ldots, n-1$。

实际问题中的输入数据通常满足“本征间隔”性质，即核矩阵本征谱上最大的 r 个本征值远远大于其余本征值[122]。在这种场景下，没有必要计算目标领域核矩阵 $\mathbf{K}_{\mathcal{X}}$ 的所有本征值，保留最大的 r 个本征值和本征向量可以极大降低计算复杂度。为了在不降低模型效果的情况下加快计算效率，本章采用 $r = \min(500, n)$，相关变量维度降为 $\overline{\mathbf{\Phi}}_{\mathcal{Z}} \in \mathbb{R}^{m\times r}$，$\boldsymbol{\lambda} \in \mathbb{R}^{r\times 1}$，$\mathbf{Q} \in \mathbb{R}^{r\times r}$。算法 5 总结了上述执行过程。

4.3.3.2　可扩展性实现

标准核机器在大数据时代备受挑战，原因是其二次计算复杂度 $O(n^2)$ 难以扩展到大规模数据上。Nyström 法使大规模核学习变得行之有效[113,115]。具体地，Nyström 法通过对一个大规模核矩阵 $\mathbf{K}_{\mathcal{X}} \in \mathbb{R}^{n\times n}$ 的列向量进行随机或均匀下采样得到的一个子矩阵 $\mathbf{K}_{\mathcal{X}\widehat{\mathcal{X}}} \in \mathbb{R}^{n\times\hat{n}}, \hat{n} \ll n$ 来对原来的大矩阵进行近似。记 $\mathbf{K}_{\widehat{\mathcal{X}}}$ 为数据子集 $\widehat{\mathcal{X}}$ 上的核矩阵，其本征系统可由本征分解 $\mathbf{K}_{\widehat{\mathcal{X}}}\mathbf{\Phi}_{\widehat{\mathcal{X}}} = \mathbf{\Phi}_{\widehat{\mathcal{X}}}\mathbf{\Lambda}_{\widehat{\mathcal{X}}}$ 得到。根据 Nyström 本征向量近似公式 (4-5)，目标领域核矩阵 $\mathbf{K}_{\mathcal{X}}$ 的本征系统可由插值近似：

$$\mathbf{\Phi}_{\mathcal{X}} \simeq \mathbf{K}_{\mathcal{X}\widehat{\mathcal{X}}}\mathbf{\Phi}_{\widehat{\mathcal{X}}}\mathbf{\Lambda}_{\widehat{\mathcal{X}}}^{-1} \tag{4-17}$$

类似地，辅助领域的大规模核矩阵 $\mathbf{K}_{\mathcal{Z}} \in \mathbb{R}^{m\times m}$ 也可由对它进行下采样得到的一个子矩阵 $\mathbf{K}_{\mathcal{Z}\widehat{\mathcal{Z}}} \in \mathbb{R}^{m\times\hat{m}}, \hat{m} \ll m$ 来低秩近似。记 $\mathbf{K}_{\widehat{\mathcal{Z}}}$ 为数据子集 $\widehat{\mathcal{Z}}$ 上的核矩阵，根据 Nyström 核矩阵近似公式 (4-6)，辅助领域核矩阵 $\mathbf{K}_{\mathcal{Z}}$ 可通过如下近似得到：

$$\mathbf{K}_{\mathcal{Z}} \simeq \mathbf{K}_{\mathcal{Z}\widehat{\mathcal{Z}}}\mathbf{K}_{\widehat{\mathcal{Z}}}^{-1}\mathbf{K}_{\widehat{\mathcal{Z}}\mathcal{Z}} \tag{4-18}$$

要对二次规划问题 (4-15) 进行可扩展性实现，还需要对目标领域本征向量外插值到辅助领域的本征向量 $\overline{\mathbf{\Phi}}_{\mathcal{Z}}$ 进行小数据集近似。为此，首先通过本章提出的跨领域 Nyström 法 (4-9)，将目标领域本征向量 $\mathbf{\Phi}_{\mathcal{X}}$ 外插值到辅助领域数据子集 $\widehat{\mathcal{Z}}$：

$$\overline{\mathbf{\Phi}}_{\widehat{\mathcal{Z}}} \simeq \mathbf{K}_{\widehat{\mathcal{Z}}\mathcal{X}}\mathbf{\Phi}_{\mathcal{X}}\mathbf{\Lambda}_{\widehat{\mathcal{X}}}^{-1} \tag{4-19}$$

其次通过标准 Nyström 法 (4-6)，将子集上的本征向量 $\overline{\mathbf{\Phi}}_{\widehat{\mathcal{Z}}}$ 插值到数据全集 $\mathcal{Z}$ 上：

$$\overline{\mathbf{\Phi}}_{\mathcal{Z}} \simeq \mathbf{K}_{\mathcal{Z}\widehat{\mathcal{Z}}}\overline{\mathbf{\Phi}}_{\widehat{\mathcal{Z}}}\mathbf{\Lambda}_{\widehat{\mathcal{X}}}^{-1} \tag{4-20}$$

这样，二次规划 (4-15) 和领域不变核矩阵 (4-12) 的所有变量都转变为可扩展计算方法。由于基本过程与算法 5 类似，上述变换得到的 TKL 可扩展性实现这里从略。

4.3.3.3　计算复杂度

评估计算复杂度一般采用大 O 法。记 r 为所保留的最大本征向量的个数，取固定值 $r = \min(500, n)$。算法 5 是 TKL 的非可扩展性实现，第 2 行计算核矩阵开销为 $O(d(m+n)^2)$，第 3 行计算目标领域核矩阵的本征分解开销为 $O(rn^2)$，第 4 行本征向量插值开销为 $O(rmn)$，第 5 行求解二次优化问题开销为 $O(rn^2+r^3)$，第 6 行构造领域不变核矩阵开销为 $O(r(m+n)^2)$，总计算复杂度为二次 $O((d+r)(m+n)^2)$。

对于 TKL 的可扩展性实现（不包括领域不变核矩阵 $\mathbf{K}_{\mathcal{A}}$ 的构造），总计算复杂度降低为线性 $O((d+r)(\hat{m}+\hat{n})(m+n))$, 其中 $\hat{m} \ll m, \hat{n} \ll n$[115]，但构造 $\mathbf{K}_{\mathcal{A}}$ 仍

需 $O(r(m+n)^2)$。由于 $\mathbf{K}_{\mathcal{A}}$ 是一个低秩矩阵，实际中不会直接计算 $\mathbf{K}_{\mathcal{A}}$ 而是保留其分解形式 (4-12)，该式会带入标准核机器中并通过矩阵结合律等实现线性复杂度。

4.3.4　近似误差分析

本节依据如下定理[113]，从理论上分析分布差异与 Nyström 近似误差的关系。

定理 4.4 (近似误差)：[113] 定义 $\mathcal{E} \triangleq \left\|\mathbf{K}_{\mathcal{Z}} - \mathbf{K}_{\mathcal{Z}\mathcal{X}}\mathbf{K}_{\mathcal{X}}^{-1}\mathbf{K}_{\mathcal{X}\mathcal{Z}}\right\|_F$ 为 *Nyström* 近似误差，则该误差有上界：

$$\mathcal{E} \leq 4m\sqrt{C_k mn\epsilon} + C_k mn\epsilon\left\|\mathbf{K}_{\mathcal{X}}^{-1}\right\|_F \tag{4-21}$$

其中 C_k 为一常数，$\epsilon = \sum_{i=1}^{m}\left\|\mathbf{z}_i - \mathbf{x}_{\mathrm{NN}(i)}\right\|^2$ 为将所有样例 $\mathbf{z}_i \in \mathcal{Z}$ 分别用与其最近的样例 $\mathbf{x}_{\mathrm{NN}(i)} = \arg\min_{\mathbf{x}_j}\left\|\mathbf{z}_i - \mathbf{x}_j\right\| \in \mathcal{X}$ 进行编码的总量化误差，即一近邻量化误差。

上述定理表明，最小化 Nyström 近似误差等价于要求辅助领域数据和目标领域数据充分重叠，换言之，即要求辅助领域的每个样例都能在目标领域中至少找到一个最近邻样例使得量化误差最小。迁移核学习 TKL 的优化问题 (4-11) 是 Nyström 误差函数的谱松弛版本，因此最小化 (4-11) 从本质上实现了辅助领域和目标领域聚类结构重叠程度最大化，从而最终使概率分布在领域间得以充分适配。

4.4　实验过程与结果

这一节通过设计和执行系统性的实验来证明迁移核学习（TKL）的有效性，实验面向文本分类、视觉对象识别、视频事件识别等三类常见的跨领域实际应用。

4.4.1　实验数据

4.4.1.1　文本数据

按迁移学习文献[29,45,47,60,63,67,85] 介绍的通用协议，本节采用 *20-Newsgroups* 和 *Reuters-21578* 两个基准文本集，并根据层次结构生成 222 个跨领域文本分类任务。

20-Newsgroups① 数据集包含约 20,000 个文档，4 个大类分别为 *comp*、*rec*、*sci* 和 *talk*，每个大类包含 4 个子类，详细信息如表 4.2 所示。在实验中构造了 6 组跨领域二分类任务，每组任务由 4 个大类中随机选取 2 个大类构成，一个大类记为正例，另一个大类记为负例，6 个任务组具体为 *comp vs rec*、*comp vs sci*、*comp vs talk*、*rec vs sci*、*rec vs talk* 和 *sci vs talk*。每个跨领域分类任务（包括辅助领域

① http://people.csail.mit.edu/jrennie/20newsgroups

表 4.2　文本数据集 20-Newsgroups 和 Reuters-21578 中的层次结构和统计信息。

数据集	大类	子类	样例数	特征数
20-Newsgroups	comp	comp.graphics	970	25,804
		comp.os.ms-windows.misc	963	
		comp.sys.ibm.pc.hardware	979	
		comp.sys.mac.hardware	958	
	rec	rec.autos	987	
		rec.motorcycles	993	
		rec.sport.baseball	991	
		rec.sport.hokey	997	
	sci	sci.crypt	989	
		sci.electronics	984	
		sci.med	987	
		sci.space	985	
	talk	talk.politics.guns	909	
		talk.politics.mideast	940	
		talk.politics.misc	774	
		talk.religion.misc	627	
Reuters-21578	orgs	many subcategories	1,237	4,771
	people	many subcategories	1,208	
	place	many subcategories	1,016	

和目标领域）采用文献[67] 介绍的方法生成：每个任务组 P *vs* Q 的两个大类 P 和 Q 分别包含 4 个子类 P_1、P_2、P_3、P_4 和 Q_1、Q_2、Q_3、Q_4；随机选取 P 的两个子类（如 P_1 、P_2）与 Q 的两个子类（如 Q_1 、Q_2）构成辅助领域，其余子类（P 的 P_3、P_4 和 Q 的 Q_3、Q_4）构成目标领域。以上构造策略既保证辅助领域和目标领域是相关的，因为它们都来自同样的大类；又保证辅助领域和目标领域是不同的，因为它们来自不同的子类。每个任务组 P *vs* Q 可以生成 $C_4^2 \cdot C_4^2 = 36$ 个分类任务，总计 6 个任务组共生成 $6 \cdot 36 = 216$ 个分类任务。数据集经过文本预处理后包含 25,804 个词项特征和 15,033 个文档，每个文档由 *tf-idf* 向量表征，如表 4.2 所示。

Reuters-21578① 是一个较难的文本数据集，包含多个大类和子类。其中最大 3 个大类为 *orgs*、*people* 和 *place*，可构造 6 个跨领域文本分类任务 *orgs vs people*、*people vs orgs*、*orgs vs place*、*place vs orgs*、*people vs place* 和 *place vs people*。本章采用文献[123] 发布的 *Reuters-21578* 预处理集。值得一提的是，现有工作通常选取其中 3 个分类任务进行评测，本章在所有 6 个分类任务上进行了更为完整的评测。

① http://www.daviddlewis.com/resources/testcollections/reuters21578

表 4.3　跨领域图像和视频数据集的统计信息。

数据集	样例数	特征数	类别数	包含子集	数据集	样例数	特征数	类别数	包含子集
Office	1,410	800	10	A, W, D	Kodak	195	2,500	6	Kodak
Caltech-256	1,123	800	10	C	YouTube	906	2,500	6	YouTube

4.4.1.2　图像数据

本章采用文献[15,61,107,108]广泛使用的基准图像数据集 *Office* 和 *Caltech-256*，根据领域先验生成 12 个跨领域视觉对象识别任务来横向评测本章算法的效果。**Office**[15,107,108] 是视觉迁移学习的主流基准数据集，包含 3 个对象领域 **Amazon** (在线电商图片)、**Webcam** (网络摄像头拍摄的低解析度图片)、**DSLR** (单反相机拍摄的高解析度图片)，共有 4,652 张图片 31 个类别标签。**Caltech-256** 是对象识别的基准数据集，包括 1 个对象领域 **Caltech**，共有 30,607 张图片 256 个类别标签。

为便于横向比较，实验直接采用文献[108]发布的 *Office+Caltech* 预处理数据集。对每张图片抽取 SURF 特征，并向量化为 800 维的直方图表征，所有直方图向量都进行减均值除方差的归一化处理，直方图码表由 K 均值聚类算法在 *Amazon* 子集上生成。具体共有 4 个领域 **C** (Caltech-256)、**A**（Amazon）、**W**（Webcam）和 **D**（DSLR），从中随机选取 2 个不同的领域作为辅助领域和目标领域，则可构造 $4\times 3=12$ 个跨领域视觉对象识别任务，如 $C\to A$，$C\to W$，$C\to D$，…，$D\to W$。

4.4.1.3　视频数据

本章采用文献[109]发布的跨领域视频事件识别数据集，如图 4.3 和表 4.3 所示，由 Kodak 客户视频数据和 YouTube 视频数据组成；Kodak 客户视频数据[124] 包括 100 个真实用户一年内的相关视频和标注信息，YouTube 视频数据由关键词搜索从 YouTube 网站爬取。在该数据集上进行迁移学习十分具有挑战性：（1）YouTube 视频的解析度低于 Kodak 视频（2）YouTube 视频的标注信息来自网络准确度低。

实验中采用 6 种事件类型进行评测，具体包括“birthday”、“picnic”、“parade”、“show”、“sport”、“wedding”。实际中目标领域（Kodak）标注视频远少于辅助领域（YouTube）标注视频，因此辅助领域的所有 906 个 YouTube 视频都作为训练数据，目标领域的每个类别分别随机选取 3 个 Kodak 视频作为训练数据、其余 Kodak 视频作为测试数据。文献[109]提供了随机选取的配置信息，因此可以完整重现其实验过程和结果。该数据集进行如下预处理：按每秒 2 关键帧的频率从每个视频片段中采集关键帧，在每个关键帧的显著性区域抽取 128 维 SIFT 特征，每个视频的所有 SIFT 特征向量化为 2500 维直方图，直方图码表由 K 均值聚类得到。为更有效地对任意两个视频进行匹配，利用文献[109]提出的时空对齐金字塔

(a) Office 和 Caltech-256 数据集　　(b) Kodak 和 YouTube 数据集

图 4.3　图像对象识别与视频事件识别数据集中的示例图片或视频关键帧。

匹配（Aligned Space-Time Pyramid Matching，ASTPM）方法改进视频的特征表示。

4.4.2　基准算法和实现细节

4.4.2.1　基准算法

本章在上述过程构造的 235 个跨领域文本、图像和视频分类任务上进行系统性对比实验，下面这些基准算法基本代表了迁移学习研究的经典方法和前沿成果：

- 支持向量机（Support Vector Machine，SVM）
- 拉普拉斯支持向量机（Laplacian Support Vector Machine，LapSVM）[88]
- 跨领域谱分类（Cross-Domain Spectral Classification，CDSC）[100]
- 谱特征对齐（Spectral Feature Alignment，SFA）[42]
- 核均值匹配（Kernel Mean Matching，KMM）[27]
- 迁移成分分析（Transfer Component Analysis，TCA）[45]
- 领域迁移多核学习（Domain Transfer Multiple Kernel Learning，DTMKL）[85]
- 测地流核方法（Geodesic Flow Kernel，GFK）[108]
- 代理核匹配（Surrogate Kernel Matching，SKM）[29]

其中 DTMKL、GFK 和 SKM 是基于核学习的迁移学习方法，具体而言：DTMKL 基于多核学习框架，显式地最小化不同概率分布在每个核函数上的差异；GFK 在辅助领域和目标领域之间的测地流上平滑地抽取无穷多个子空间，用于刻划概率分布从辅助领域到目标领域的变迁规律；SKM 从目标领域核矩阵的本征系统中插值得到辅助领域的代理核矩阵，然后将辅助领域核矩阵通过线性变换与代理核矩阵进行对齐。注意 SKM 与本章 TKL 最为类似，但两者的区别仍十分明显：TKL 通过谱核学习得到一个领域不变核矩阵，而 SKM 在已知核矩阵上进行对齐操作。

4.4.2.2　实现细节

为保持对比实验的公平性，本章采用相关文献[1,29,45,85] 的测试协议。SVM 在辅助领域标注数据上训练，在目标领域无标数据上测试；LapSVM 和 DTMKL 在所有数据上归纳迁移分类器；CDSC、SFA、KMM、TCA、GFK、SKM 和 TKL 在所有数据上学习领域不变的特征空间、实例权重或相似核，并训练 SVM 分类器。

表 4.4　文本数据集上的平均分类准确率（%），包括 9 个分类任务组共 222 个分类任务。

任务组	标准学习		非核学习				核学习			
	SVM	LapSVM	CDSC	SFA	KMM	TCA	DTMKL	GFK	SKM	TKL
comp vs rec	87.51	81.93	87.95	89.73	93.64	95.12	95.08	93.74	91.31	**96.01**
comp vs sci	75.38	68.96	75.72	78.07	77.45	77.32	81.87	80.62	77.63	**88.14**
comp vs talk	95.44	95.40	97.33	95.85	96.06	97.20	97.16	96.61	**97.75**	**97.74**
rec vs sci	73.82	74.21	77.53	79.25	80.27	82.31	82.97	84.31	77.21	**91.29**
rec vs talk	83.27	87.44	82.14	86.98	85.57	86.58	88.35	92.73	86.83	**93.74**
sci vs talk	76.85	80.22	80.97	79.27	77.05	79.30	75.77	80.91	78.30	**87.52**
平均值	82.05	81.36	83.62	84.86	85.01	86.31	86.87	88.15	84.84	**92.41**
orgs vs people	78.55	82.68	80.97	77.20	80.48	81.58	81.19	81.00	78.63	**83.76**
orgs vs place	66.71	68.67	70.62	74.59	68.47	68.15	69.20	76.31	68.06	**80.85**
people vs place	59.94	60.68	64.53	67.08	57.33	57.61	57.80	58.50	59.33	**68.48**
平均值	68.40	70.68	72.04	72.96	68.76	69.11	69.40	71.94	68.67	**77.70**

在目标领域没有标注数据时，一般无法通过交叉验证来自动选择最优模型参数。与相关文献一样，本章在所有 235 个分类任务上测试 9 种基准算法，将每种算法在各种参数设置下的最佳效果用于性能对比。具体地，SVM 采用 LIBSVM① 工具包实现，正则参数 C 通过遍历 $C \in \{0.1, 0.5, 1, 5, 10, 50, 100\}$ 设置；LapSVM 采用文献[88] 的实现②，正则参数 γ_A 和 γ_I 分别通过遍历 $\gamma_A, \gamma_I \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ 设置；DTMKL 采用文献[85] 的实现，正则参数 θ 通过遍历 $\theta \in \{0.01, 0.1, 1, 10, 100\}$ 设置；CDSC、SFA、TCA、GFK 采用文献作者提供的实现，子空间参数 k 通过遍历 $k \in \{10, 20, \ldots, 100\}$ 设置。对所有核方法，在文本数据集上采用线性核 $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j$，在图像、视频数据集上采用高斯核 $k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}$；根据文献[109]，高斯核函数参数设置为所有训练数据之间平均欧式距离 A 的倒数 $\gamma = \frac{1}{A}$。

TKL 是一个简单易用的模型，因其仅有 2 个待定模型参数：本征谱阻尼系数 ζ 和 SVM 正则参数 C。后文将进行系统的参数敏感性实验，验证 TKL 能在足够大的参数范围内取得更好的效果。在对比实验中，采用统一设置：（1）222 个文本分类任务 $\zeta = 2.0$ 和 $C = 10.0$（2）13 个图像、视频分类任务 $\zeta = 1.1$ 和 $C = 10.0$。由于参数 C 比参数 ζ 更稳定，实际参数选择时可按先后顺序先确定 C 再决定 ζ。

本章采用测试数据（目标领域无标数据）准确率（Accuracy）作为评价指标：

$$Accuracy = \frac{\left|\mathbf{x} : \mathbf{x} \in \mathcal{X} \wedge f(\mathbf{x}) = y(\mathbf{x})\right|}{|\mathbf{x} : \mathbf{x} \in \mathcal{X}|} \tag{4-22}$$

其中 $y(\mathbf{x})$ 是测试样例 $\mathbf{x}$ 的真实标签，$f(\mathbf{x})$ 是待测学习算法为样例 $\mathbf{x}$ 预测的标签。

① http://www.csie.ntu.edu.tw/~cjlin/libsvm

② http://vikas.sindhwani.org/manifoldregularization.html

4.4.3　实验结果

本节给出 TKL 和 9 种基准方法在 235 个文本、图像、视频跨领域分类任务上的准确率，并根据实验结果深入讨论某些相关问题。

4.4.3.1　文本实验结果

由于 20-Newsgroups 和 Reuters-21578 层次结构各不相同，这里需要分别讨论。

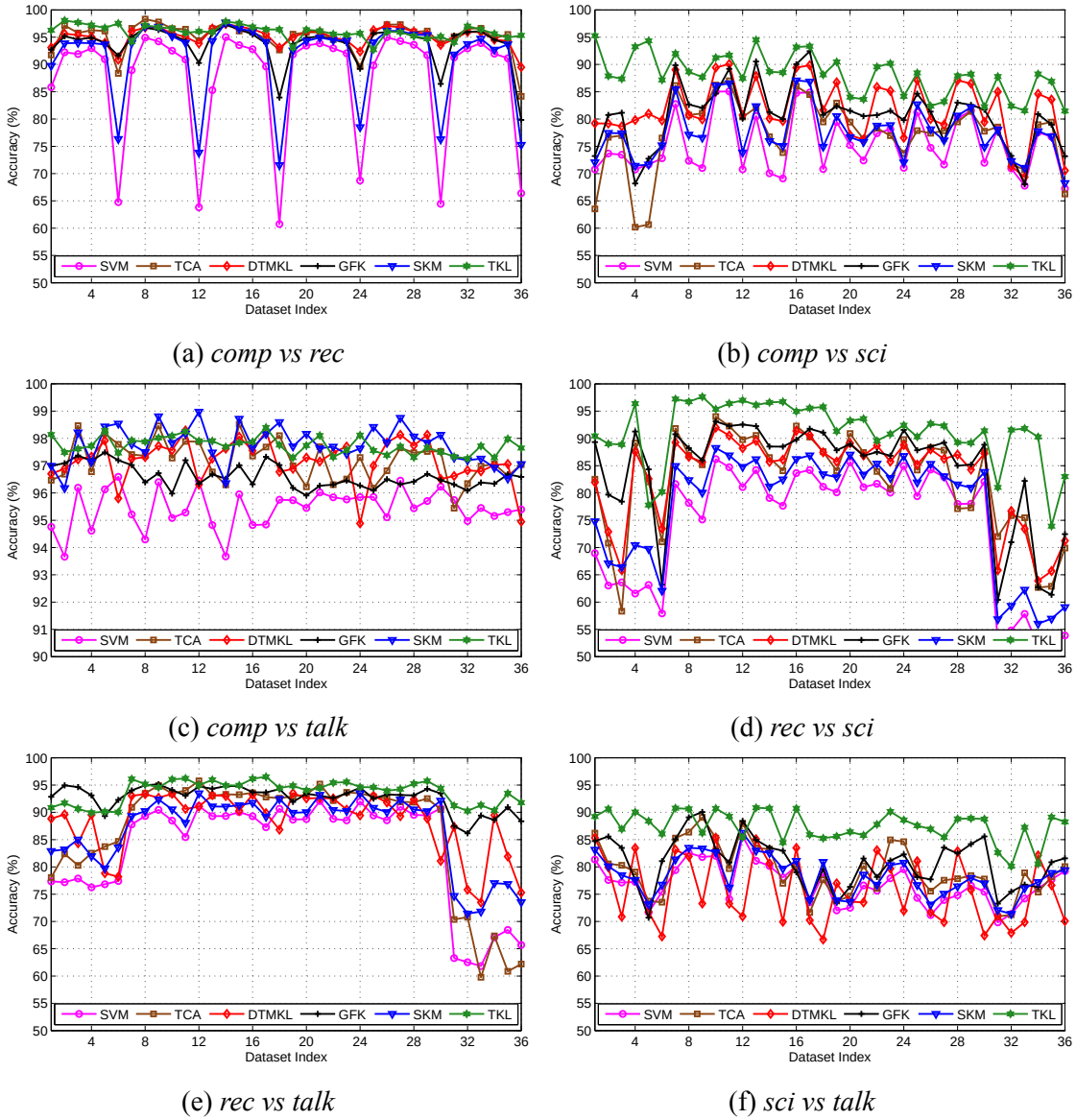
20-Newsgroups: 由该数据集构造了 6 个跨领域分类任务组 *comp vs rec*、*comp vs sci*、*comp vs talk*、*rec vs sci*、*rec vs talk* 和 *sci vs talk*，每个任务组分别由 36 个跨领域分类任务构成。实验结果包括每个分类任务的分类准确率和每个任务组的平均分类准确率，分别如图 4.4(a)~4.4(f) 和表 4.4 所示，由实验结果得出如下结论。

首先，TKL 在大多数分类任务（216 个中的 171 个）上取得了比 9 个基准方法非常显著的准确率提升：TKL 在 216 个任务上的平均准确率为 **92.41%**，较最好的基准方法 GFK 提升 **4.26%** ，即错误率下降比例达到 **35.95%**。注意到 216 个任务的迁移学习难度差别很大，基准 SVM 分类器仅能取得 82.05% 的平均准确率，且在难度较大的任务上准确率很低。尽管 TKL 未能在所有 216 个任务上都超越基准方法，但它仍不失为一种高效鲁棒方法，原因在于：（1）在基准方法表现不良时它表现尤为出色；（2）在基准方法表现良好时它也能与之匹敌。大量任务上取得一致的性能提升，有力证明了 TKL 能为跨领域文本分类任务学习领域不变核。

其次，标准监督学习（SVM）和半监督学习（LapSVM）均未能在多数任务上取得足够好的准确率，主要原因是标准学习方法基于训练数据和测试数据来自同一概率分布这一基本假设。然而，在跨领域学习问题中，这个假设并不成立，这导致了严重的欠拟合问题。有些意外的是，流形正则化半监督学习 LapSVM 仍比 SVM 性能更差，这揭示了在不同领域间适配概率分布对防止负迁移的重要性。

再次，迁移学习方法通常较标准学习方法性能更好。为便于清晰讨论各种不同类型方法的优缺点，将考察的基准迁移学习方法分为两类：非核学习和核学习。

非核学习：包括 CDSC、SFA、KMM 和 TCA 等几种方法。CDSC 通过同时保持领域内几何结构及领域外判别信息，抽取一个公共的谱嵌入空间。SFA 基于共现频率将领域相关的特征通过领域无关的特征进行对齐，学习一个公共特征空间。CDSC 和 SFA 的缺点在于没有显式地最小化领域间的概率分布差异，因此它们的泛化误差可能没有理论保证。KMM 和 TCA 通过最小化定义在可再生希尔伯特空间中的领域间最大均值差异 MMD 准则[33] 来克服前述缺点，因而可以得到有一定误差上界保证的学习模型。但是最小化 MMD 准则仅能适配不同概率分布的各阶统计量（均值和方差），适配程度较低，所得到的泛化误差上界也就比较宽松[27]。



(a) *comp vs rec*　(b) *comp vs sci*

(c) *comp vs talk*　(d) *rec vs sci*

(e) *rec vs talk*　(f) *sci vs talk*

图 4.4　SVM、TCA、DTMKL、GFK、SKM 和 TKL 在文本分类任务上的分类准确率。

核学习：包括 DTMKL、GFK 和 SKM 等几种方法。DTMKL 的目标是最小化领域间多个核上的最大均值差异 MMD，相对于 KMM 和 TCA 的优点是同时利用多个核函数实现矩匹配，可以更大程度地减小概率分布差异。不过，由于 MMD 仅能适配概率分布统计量的这一缺点，DTMKL 仍不能对不同概率分布进行充分适配。GFK 在辅助领域和目标领域之间的测地流上抽取无穷多个子空间用以平滑地刻划辅助领域到目标领域的变迁信息，这些子空间上的向量相似度可以积分为一个闭式核函数。GFK 的缺点在于，为使不同子空间可以在测地流上平滑迁移，所抽取子空间的维度通常较低，因为较高维度的子空间包含过多的次要信息，无法在领域间有效迁移，但是较低维度的子空间难以对输入数据进行准确表征。SKM 从目标领域核矩阵的本征系统插值得到辅助领域的代理核矩阵，并将辅助领域的

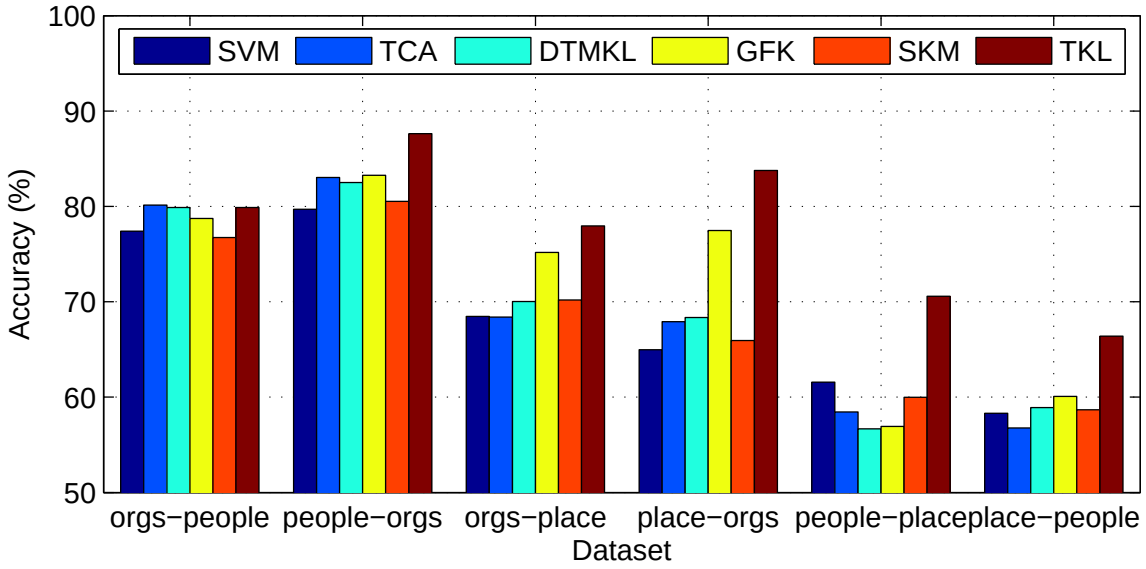


图 4.5　SVM、TCA、DTMKL、GFK、SKM 和 TKL 在 Reuters-21578 的文本分类准确率。

真实核矩阵与代理核矩阵进行对齐。SKM 的缺点在于仅对齐了已知核矩阵而不是学习一个数据依赖的领域不变核矩阵，因此它也未能很好地适配不同概率分布。

基于上述讨论，TKL 被设计为可以解决上述局限性的一种方法：1）TKL 可以显式地最小化辅助领域真实核矩阵和目标领域插值核矩阵的近似误差，这是较 MMD 能更精细地度量概率分布差异的新准则；2）TKL 探索了输入空间的整个本征谱，可以构造足够准确的核矩阵用于表征输入数据；3）TKL 是数据依赖的谱学习方法，可以更好的适配概率分布差异。这些优点共同保证了 TKL 较好的性能。

Reuters-21578: 由该数据集构造了 3 个跨领域分类任务组，每组包含 2 个文本分类任务。每个任务组的平均分类准确率如表 4.4 所示，每个任务的分类准确率如图 4.5 所示。可观察到 TKL 在所有分类任务上均显著地超越了所有基准方法。TKL 在所有 6 个分类任务的平均准确率为 **77.70%**，相较最佳基准方法 SFA 提升 **4.74%**。注意到这可能是 Reuters-21578 所有 6 个分类任务的首次完整实验结果。

值得一提的是，Reuters-21578 数据集比 20-Newsgroups 数据集要更为挑战，因为它的每个大类都包含多样化的子类。因此，CDSC、SFA、GFK 和 SKM 等方法将更难抽取领域间共享的隐含结构；KMM、TCA 和 DTMKL 也更难通过 MMD 减小分布差异，毕竟 MMD 仅能适配分布统计量而不能适配复杂结构。这解释了基准方法在该数据集上效果不好的原因。TKL 方法可以自然地探索领域间多样化结构，因为它可以将目标领域插值得到的核矩阵与辅助领域真实的核矩阵进行匹配，增大领域不变的本征谱、减小领域特定的本征谱，这保证了 TKL 优良的性能。

表 4.5　对象识别数据集的 12 个跨领域分类任务的详细准确率（10 个类别）。

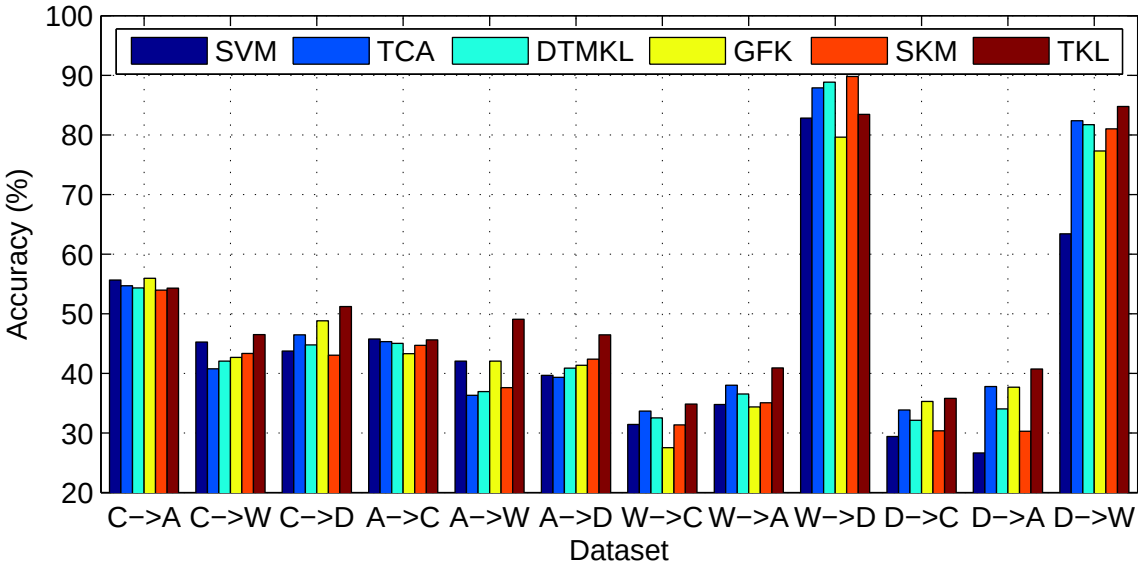
分类	标准学习		非核学习				核学习			
任务	SVM	LapSVM	CDSC	SFA	KMM	TCA	DTMKL	GFK	SKM	TKL
C→A	55.64	**56.27**	52.16	49.32	48.32	54.70	54.33	55.95	53.97	54.28
C→W	45.22	45.80	38.54	39.31	45.78	40.76	42.04	42.68	43.31	**46.50**
C→D	43.73	43.73	43.64	41.96	**53.53**	46.44	44.74	48.81	43.05	51.19
A→C	**45.77**	44.23	42.28	42.33	42.21	45.33	45.01	43.28	44.70	45.59
A→W	42.04	42.74	34.94	34.94	42.38	36.31	36.94	42.04	37.58	**49.04**
A→D	39.66	39.79	37.81	36.86	42.72	39.32	40.85	41.36	42.37	**46.44**
W→C	31.43	31.99	32.28	32.50	29.01	33.66	32.50	27.52	31.34	**34.82**
W→A	34.76	34.77	35.73	34.72	31.94	38.00	36.53	34.34	35.07	**40.92**
W→D	82.80	83.43	81.80	83.38	71.98	87.90	88.85	79.62	**89.81**	83.44
D→C	29.39	29.49	33.33	30.50	31.61	33.84	32.10	35.26	30.37	**35.80**
D→A	26.62	27.37	35.88	29.41	32.20	37.79	34.03	37.68	30.27	**40.71**
D→W	63.39	64.31	80.76	68.14	72.88	82.37	81.69	77.29	81.02	**84.75**
平均值	45.04	45.32	45.76	43.61	45.38	48.03	47.47	47.15	46.91	**51.12**

4.4.3.2　图像实验结果

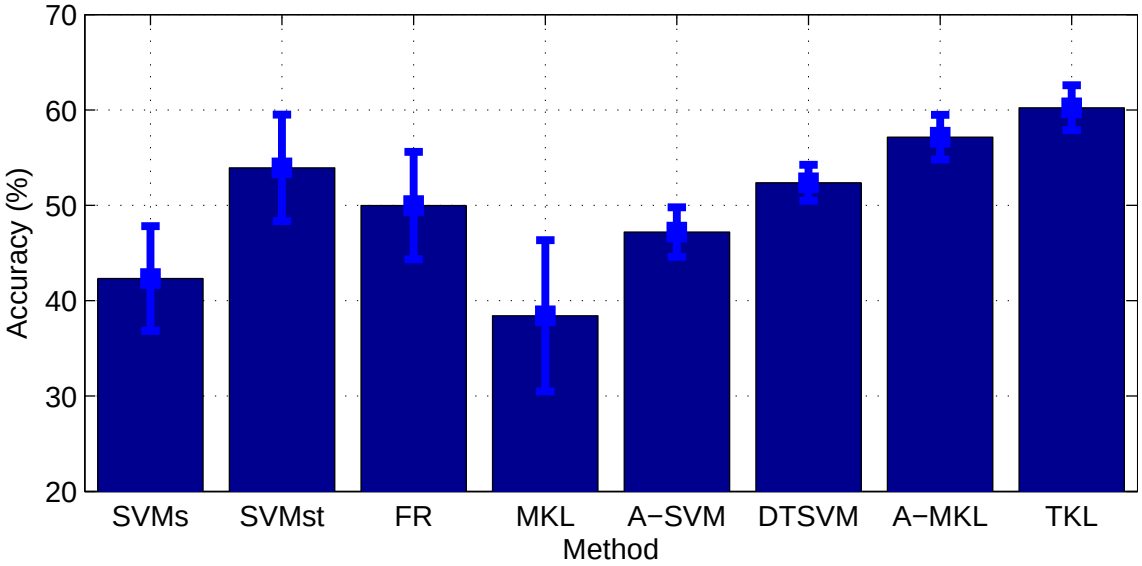
TKL 和 9 个基准方法在 12 个跨领域图像分类任务上的识别准确率如表 4.5 所示，该结果直观地绘制在图 4.6(a) 中。可以看到，TKL 在大多数任务（12 个中的 8 个）上取得了比所有基准方法显著提高的准确率：TKL 在 12 个任务上的平均分类准确率为 **51.12%**，较最好的基准方法 TCA 提升了 **3.09%**。需要注意的是，该对象识别数据集是计算机视觉中相当具有挑战的基准数据集，很多迁移学习方法在上面仅取得比标准 SVM 分类器稍好的成绩，如 TCA 取得了 2.99% 的性能提升。

视觉迁移学习相较于文本迁移学习更具挑战性，原因在于图像底层特征与高层语义之间存在的所谓*语义鸿沟*问题。该问题导致难以定义原始特征与不同领域和类别的相关性，这使得依赖特征相关性的方法 SFA 产生*负迁移*。此外，视觉迁移学习中领域间的概率分布差异通常更大，这对形式化地最小化概率分布差异提出了必然的更高要求。CDSC 并没有做到这一点，因此它的性能和标准 SVM 相仿。TCA 和 DTMKL 都显著优于 SVM，表明分布适配对视觉迁移学习的重要性。

然而，MMD 适配性较差的缺点在视觉迁移学习中更为明显。理论上，减小 MMD 等价于在可再生希尔伯特空间中对所有图像执行由辅助领域到目标领域的*平移变换*，这一过简操作显然不能十分有效地对视觉数据进行重构。虽然 GFK 和 SKM 一定程度克服这一局限，但它们又引入了新的问题：（1）GFK 受限于低维子空间，无法给出高准确度的特征表示；（2）SKM 受限于预先算好的核矩阵，无法保证该矩阵在领域间的适应性。TKL 通过探索辅助领域和目标领域的完整本征空



(a) *Office+Caltech*



(b) *YouTube+Kodak*

图 4.6　SVM、TCA、DTMKL、GFK、SKM 和 TKL 在图像、视频数据上的分类准确率。

间来学习领域不变核矩阵，因而可以有效的重构视觉数据以提高视觉迁移的效果。

4.4.3.3　视频实验结果

该实验属于崭新的研究趋势*互联网视觉*：包括图像、视频及其上下文信息（如标签、类别、标题）在内的大规模互联网数据都可用于提高计算机视觉应用的效率和效果[109]。本章采用相同的实验协议对比 TKL 和文献[109] 所考察的几种主流迁移学习方法：SVM_T（标注数据仅来自目标领域）、SVM_{ST}（标注数据同时来自辅助领域和目标领域）、FR[39]、A-SVM[89]、DTSVM[111]、MKL[111] 和 A-MKL[109]。

表 4.6　跨领域视频事件识别任务上的平均准确率（6 个类别、5 次随机实验平均）。

方法	SVM_T	SVM_{ST}	FR	MKL	A-SVM	DTSVM	A-MKL	TKL
准确率	42.32±5.50	53.93±5.58	49.98±5.63	38.42±7.93	47.19±2.59	52.36±1.88	57.14±2.34	**60.22±2.35**

TKL 和 7 种基准方法的识别准确率如表 4.6 和图 4.6(b) 所示，可以看到 TKL 显著超越了竞争者。正如文献[109] 所指出，跨领域视频事件识别十分具有挑战性：领域间的差异如此之大以至于很多迁移学习方法，如 FR、A-SVM 和 DTSVM 等，遭遇了*负迁移*即迁移学习效果竟然差于标准 SVM_{ST} 分类器。文献[109] 提出了基于多核学习的 A-MKL 方法来对不同特征进行融合、同时显式地最小化特征分布间的差异，实现了*正迁移*并取得了 +3.21% 的精度提升。但是 A-MKL 仍依赖 MMD 进行分布适配，这较大程度上限制了它处理视频在领域间复杂变化的能力。TKL 在该任务上获得了 **60.22%** 的识别准确率，相对于最好的基准算法 A-MKL 进一步提升了 **3.08%**。TKL 在数学上比 A-MKL 简单得多但却取得更好的效果，这有力地证明了学习数据依赖的领域不变核矩阵要比集成多个预算好的核矩阵效果更好。

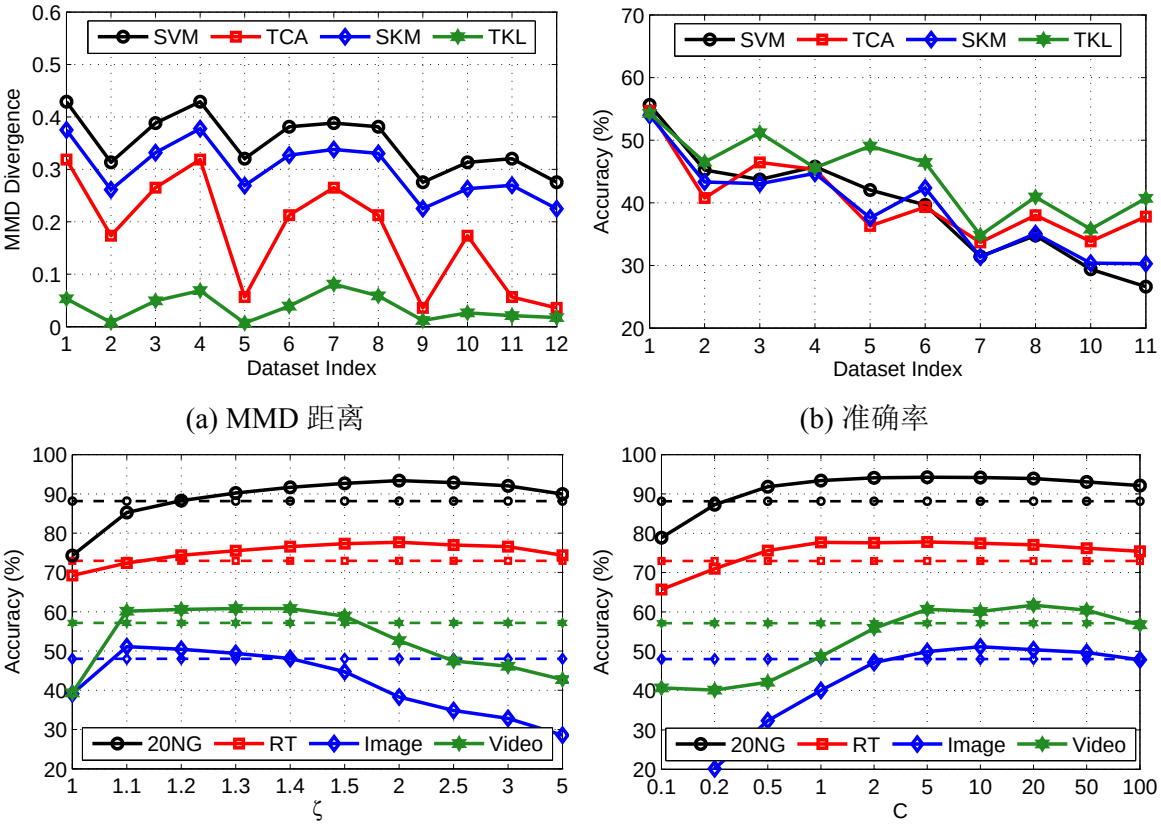
4.4.4　适配性分析

本节从领域间概率分布差异的角度来分析 TKL 的适配性能。首先在 12 个图像分类任务上执行最佳参数配置下的 SVM、TCA、SKM 和 TKL 算法，其次在抽取到的特征嵌入或核矩阵上通过公式 (4-1) 计算领域间概率分布的 MMD 值。根据文献[33] 的理论分析，较小的领域间概率分布差异预示着较好的跨领域泛化性能。

在图像数据集上的分布距离和分类准确率分别如图 4.7(a) 和图 4.7(b) 所示。可以看到，在未学习任何特征表示或核矩阵的情况下，标准 SVM 方法在原始特征空间中的分布差异是最大的，对应的分类准确率也是最低的。SKM 可以构造一个核矩阵使得分布差异一定程度减小但并不充分。TCA 通过最大均值差异 MMD 准则显式地最小化分布差异，因此它可以获得比 SKM 更高的准确率；但由于 MMD 只能适配概率分布统计量，领域间的复杂变化信息并未得到有效的处理。TKL 通过最小化辅助领域和目标领域插值核矩阵显式地减小了分布差异，因此它可以学习一个数据依赖的领域不变核矩阵以实现有效迁移学习，这保证了它的最优性能。

4.4.5　参数敏感性分析

作为一个简单易用的模型，TKL 仅包含 2 个可调参数：本征谱阻尼系数 ζ 和 SVM 正则参数 C。本节在所有文本、图像、视频数据集上进行系统性的参数实验，平均分类准确率在每个数据集上计算：*20NG*（由 20-Newsgroups 构造的 216 个文本分类任务）、*RT*（由 Reuters-21578 构造的 6 个文本分类任务）、*Image*（12 个图像



(a) MMD 距离　　(b) 准确率

(c) 阻尼系数 ζ　　(d) SVM 参数 C

图 4.7　TKL 的实验性分析：1）图像数据上的有效性验证；2）所有数据上的参数敏感性。

分类任务）和 *Video*（视频分类任务），详细实验结果如图 4.7(c) 和图 4.7(d) 所示。

首先保持 C 固定不变，在参数范围 $\zeta \in [1, 5]$ 内执行 TKL。由于正定核矩阵的本征谱按照幂律分布衰减，很自然的引入阻尼系数 $\zeta > 1$ 控制本征谱的衰减速度，其取值可由输入核矩阵的本征谱衰减趋势大致确定。平均准确率相对于 ζ 的不同取值的变化规律如图 4.7(c) 所示，可以看到，文本数据的最佳取值范围是 $\zeta \in [1.2, 5.0]$，视觉数据的最佳取值范围是 $\zeta \in [1.1, 1.4]$。注意到视觉数据的本征谱衰减速度要慢于文本数据，因此视觉数据的 ζ 最佳取值范围也就窄于文本数据的。

其次保持 ζ 不变，在参数范围 $C \in [0.1, 100]$ 内执行 TKL。平均准确率相对于 C 的不同取值的变化规律如图 4.7(d) 所示，可以看到，文本数据的最佳取值范围是 $C \in [0.5, 100]$，视觉数据的最佳取值范围是 $C \in [2, 50]$。由于 TKL 对参数 C 的性能更稳定，可先确定 C 的取值。TKL 对参数 ζ 和 C 在相当大取值范围内不敏感。

可扩展性是机器学习算法的重要属性，在包含 25,800 个特征 8,000 个文档的文本分类任务 *comp vs rec 1* 上考察所有算法的时间复杂度，得如表 4.7 所示的结果。可以看到 TKL 一般比基准方法高效得多，这显示了 TKL 对实际应用的价值。

表 4.7　可扩展性：TKL 和 9 种基准迁移学习方法的时间复杂度。

方法	时间 (秒)	方法	时间 (秒)	方法	时间 (秒)	方法	时间 (秒)	方法	时间 (秒)
SVM	6.79	LapSVM	44.20	CDSC	25.37	SFA	20.82	KMM	705.69
TCA	126.79	DTMKL	293.90	GFK	189.93	SKM	147.23	TKL	28.31

4.5　小结

本章提出了迁移核学习（TKL）方法，它通过可再生希尔伯特空间直接适配辅助领域和目标领域的核矩阵和概率分布，从而学习一个领域不变核矩阵。具体地说，TKL 从一族由目标领域本征系统插值得到的辅助领域候选谱核中，选择使得辅助领域 Nyström 近似误差最小的一个谱核用于构造领域不变核机器。该方法在大量文本、图像、视频数据集上均能大幅稳定地超越基准方法。未来工作包括探索更合理的本征谱阻尼约束，使得非幂律分布衰减的本征谱也能被准确地学习。

第 5 章　深度表征适配方法

迁移学习的目标是将学习模型泛化到与训练数据服从不同概率分布的测试数据上，被广泛用来进行异构领域挖掘和标注数据复用。前两章给出了如何在领域间适配概率分布，从而获得紧凑的模型泛化误差上界。但仍然存在两个根本性挑战：（1）如何度量分布差异（2）如何学习领域不变的紧致特征表示。已有方法主要聚焦于如何通过浅层网络进行知识巩固，但不能提取高度抽象、紧致的特征表示，因而无法对异构概率分布进行深入刻划，存在欠拟合问题。由于领域内概率分布拟合是领域间概率分布适配的基础，已有方法的欠拟合问题必然同时导致欠适配问题。为同时解决欠拟合与欠适配问题，本章提出深度迁移学习框架，在深度网络架构下同时进行领域不变深度表征学习和概率分布差异修正。在本章的解决方案中，首先提出非线性分布差异，基于通用非线性表征学习来形式化领域间的概率分布失配程度；其次提出不变去噪自动编码器模型，通过数据的损坏版本重构数据的原始版本来学习数据的鲁棒特征表示、并同时最小化领域间的非线性分布差异度量，进一步由多层模型堆叠形成的深度网络巩固特征表示的紧致性和鲁棒性；最后提出迁移交叉验证策略，用于目标领域无标注数据的无监督迁移学习的模型选择。本章方法在情感极性分析、垃圾邮件过滤、视觉对象识别等跨领域任务上与前沿方法进行了比较，获得了显著的性能提升和创纪录的分类准确率。

5.1　引言

标准监督学习在标注数据稀缺时的泛化误差极大、无法满足实际应用需求，而对新应用领域都手动标注大量数据也是不现实的。迁移学习被证明在减少数据标注代价方面十分有效，主要思想是借助相关辅助领域中存在的丰富标注数据[1]。其主要挑战是不同领域服从显著不同的概率分布，从而为预测模型在领域间泛化设置了障碍。例如，某种产品评论数据上建立的情感分类器[36] 对其他产品评论数据的极性预测可能并不准确，因为在不同产品领域中通常用不同的情感词来描述正负极性[42]：如在 *electronics* 电子产品中，用 “blur”、“fast”、“sharp” 等词来表示正面情感，而在 *books* 图书产品中，这些词并不具有明显的情感极性、甚至根本不会出现。又如，在某个手动标注图片集上训练的对象识别器对新测试集的预测性能可能并不好，因为不同场景下图片通常具有不同的朝向、遮挡、光照条件。这类应用中，亟待设计有效迁移学习算法来减小跨领域分布差异和手动标注代价。

跨领域迁移学习任务包括两种不同类型的数据集，一种来自辅助领域、另一

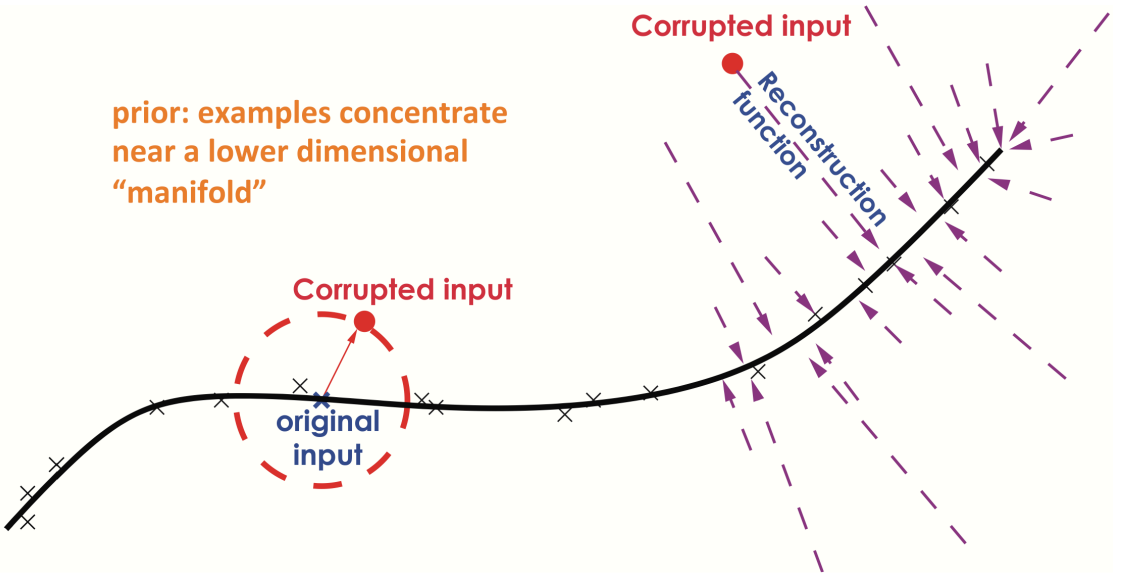


图 5.1　深度学习基本原理[49,125]：当输入数据点集中在低维流形（由黑色曲线表示）附近时，数据点的损坏版本（由原始版本经过损坏过程而得，由红色点表示）通常会与流形接近正交；深度网络的重构函数学习如何从损坏版本中恢复原始版本，等价于从低概率区域（数据的损坏版本）映射到高概率区域（数据的原始版本，位于低维流形附近）；这就建立了一个与所估计概率密度的切方向对齐的向量场，可以对概率分布进行深度拟合。

种来自目标领域。辅助领域包含大量标注数据足以准确训练分类器，目标领域包含大量无标数据、服从与辅助领域显著不同但又潜在相关的概率分布。学习目标是显式地修正领域间的概率分布失配，从而使标准分类器可以在领域间有效迁移。实现领域迁移需要解决两个关键性挑战：（1）如何度量领域间的概率分布差异（2）如何学习领域不变紧致特征表示，即如何有效实现概率分布适配和拟合问题。

为了解决第一个问题，相当一部分现有工作聚焦于如何最小化最大均值差异 Maximum Mean Discrepancy，MMD）[33] 这一非参数化统计量，它通过核空间中辅助领域和目标领域间的概率分布均值距离来度量概率分布差异。这类方法的主要目标是学习隐含特征表示或实例权重向量，使得分布差异准则 MMD 可以显式地最小化[27,45,85]。然而，基于 MMD 的核方法存在如下明显的局限性：（1）核方法并不能捕捉数据分布的通用非线性（仅能拟合局部非线性）[49] 从而不能十分有效地刻划概率分布差异；（2）预设的参数化核函数对特定分类模型可能并非最优配置[85]；（3）核方法是平方复杂度，可扩展性不好，难以应对大规模数据分析任务。

为了解决第二个问题，深度学习[49] 被成功用于抽取紧致的特征表示、强化迁移学习效果[50–53]。深度学习可以辨别隐含因式结构中反映数据分布变化的部分和不变的部分，并通过层次化结构、按照与隐含因式结构的相关性对输入特征进行分组和抽象，这样可以降低变化部分的权重、提高不变部分的权重[50]。深度学习的工作原理如图 5.1 所示。这显然有助于跨领域迁移学习，因为这些通用概念对领

域特定的数据分布是不变的，可以作为知识迁移的桥梁。例如在 *electronics* 电子产品领域，领域特定情感词如 “blur”、“fast”、“sharp” 需要重构领域共享词汇、或者被领域共享词汇重构，这样才能建立更高层次的抽象语义（领域共享词汇是指相似情感词如 “good” 或 “love” 等）。这样，辅助领域训练的分类器可以对目标领域的所有特征——甚至那些从未在辅助领域出现的词汇——赋予恰当的权重[52]。然而，降低变化部分隐含因式结构的权重可能会扩大跨领域数据分布间的差异，因为在深度特征表示下辅助领域和目标领域都变得更为“紧致”从而更容易判别彼此。本章得出如下结论：扩大了的领域间分布差异会部分损害迁移学习效果。

上述讨论提出一个有趣的问题：如何同时获得分布适配和深度学习两者的优势，建立更有效的迁移学习模型？为实现此目标，文本提出深度迁移学习框架，在深度架构中挖掘领域不变的紧致特征表示、并同时修正概率分布差异。在本章的解决方案中，首先提出非线性分布差异（Nonlinear Distribution Discrepancy，NND），基于通用非线性表征学习来形式化领域间的概率分布失配程度；其次提出不变去噪自动编码器（Invariant Denoising Autoencoder，IDA）模型，通过数据的损坏版本重构数据的原始版本来学习数据的鲁棒特征表示、并同时最小化领域间的非线性分布差异度量 NDD，进一步由多层模型堆叠形成的深度网络巩固特征表示的紧致性和鲁棒性；最后提出迁移交叉验证（Transfer Cross-Validation，TCV）策略，用于目标领域无标注数据的无监督迁移学习的模型选择。本章方法在情感极性分析、垃圾邮件过滤、视觉对象识别等跨领域任务上与前沿方法进行了比较，获得了显著的性能提升和创纪录的分类准确率。本章主要的创新性贡献总结如下：

- 提出新颖的深度迁移学习框架，可以同时抽取紧致特征表示和修正分布差异，解决欠适配与欠拟合问题，这是深度学习与分布适配方法的首次结合。
- 提出用于度量跨领域概率分布距离的非线性分布差异准则，以及自动模型选择的迁移交叉验证策略。
- 实验证据表明本章方法在情感分类任务上获得创纪录准确率，并在其他数据集上显著超越前沿方法。

5.2　非线性分布距离度量

实分析与概率论[126] 中检验两个概率分布 P 和 Q 在无穷样本集下是否相同的理论判据如下面的引理，后文提出的非线性分布差异准则即可由此推导得到。

引理 5.1 (**双样本检验**)：[33,126] 记 $(\mathcal{X}, m)$ 为可分度量空间，其中 $\mathcal{X}$ 为集合，m 为度量函数，又记 P、Q 为 $\mathcal{X}$ 上的两个 *Borel* 概率测度，则 $P = Q$ 当且仅当

$\mathbb{E}_P[f(x)] = \mathbb{E}_Q[f(x)]$ 对所有连续函数 $f \in C(\mathcal{X})$ 成立，其中 $C(\mathcal{X})$ 为 $\mathcal{X}$ 上的连续有界函数空间。

尽管连续函数空间 $C(\mathcal{X})$ 从理论上提供了唯一确定概率分布 $P = Q$ 的方法，在如此丰富的函数空间中进行有限样本集下的统计检验并非切实可行。为此，文献[33] 提出一种基于核函数的检验方法，称为最大均值差异（Maximum Mean Discrepancy，MMD），其中定义了更通用可行的统计量核均值映射，用来在一个受限的但又足够丰富的函数空间 $\mathcal{F} \subset C(\mathcal{X})$ 中度量概率分布 P 和 Q 的距离差异。

定义 5.1 (**MMD**)：[33] 记 $\mathcal{F}$ 为一个函数类 $f: \mathcal{X} \mapsto \mathbb{R}$，$\mathcal{X}_s = \{\mathbf{x}_1^s, \ldots, \mathbf{x}_{|\mathcal{X}_s|}^s\}$ 和 $\mathcal{X}_t = \{\mathbf{x}_1^t, \ldots, \mathbf{x}_{|\mathcal{X}_t|}^t\}$ 是由概率分布 P 和 Q 分别采样生成的独立同分布数据集。最大均值差异（*Maximum Mean Discrepancy*，*MMD*）及其经验估计定义如下：

$$\begin{aligned} \mathrm{MMD}\left[\mathcal{F}, P, Q\right] &= \sup_{\|f\|_{\mathcal{H}} \leqslant 1} \left(\mathbb{E}_{\mathbf{x}_i \sim p}\left[f\left(\mathbf{x}_i\right)\right] - \mathbb{E}_{\mathbf{x}_j \sim q}\left[f\left(\mathbf{x}_j\right)\right]\right) \\ \mathrm{MMD}\left[\mathcal{F}, \mathcal{X}_s, \mathcal{X}_t\right] &= \left\| \sum_{\mathbf{x}_i \in \mathcal{X}_s} \frac{\phi\left(\mathbf{x}_i\right)}{|\mathcal{X}_s|} - \sum_{\mathbf{x}_j \in \mathcal{X}_t} \frac{\phi\left(\mathbf{x}_j\right)}{|\mathcal{X}_t|} \right\|_{\mathcal{H}} \end{aligned} \tag{5-1}$$

其中 $\phi(\cdot)$ 是诱导可再生希尔伯特空间（*RKHS*）$\mathcal{H}$ 的一个非线性特征映射，满足可再生性质 $f(\mathbf{x}) = \langle \phi(\mathbf{x}), f \rangle$，$\mathcal{F}$ 可选为空间 $\mathcal{H}$ 中的一个受限单位球空间 $\mathcal{F} \subset \mathcal{H}$。

引理 5.2 (**MMD 概率判等准则**)：[33] $\mathrm{MMD}[\mathcal{F}, P, Q] = 0$ 当且仅当 $P = Q$。

根据引理5.1，上述 MMD 概率判等准则的适用性主要取决于用非线性核映射来刻划概率分布的失配程度。需要注意的是，线性核映射无法实现上述判等功能。该方法被广泛应用于解决迁移学习中的方差漂移问题[27,45,60,85]。尽管已取得广泛成功，基于 MMD 的迁移学习方法仍存在若干重要局限性。首先，核函数是一个依赖于特征空间上先验距离度量定义的固定局部响应函数（如高斯核），它要求朴素的距离函数（如欧式距离）就足以刻划数据间的相互关系。核方法仅能通过懒惰策略在训练数据近邻空间中进行局部插值得到局部泛化性[49]。从这个意义上看，核方法无法充分刻划数据中的通用非线性关系（除局部性以外的复杂关系），也就无法充分刻划分布失配程度。其次，预先设置好的核函数可能对特定学习器次优，致使人们必须选择恰当的核函数或直接从数据中学习该核函数[29,85]。再次，核方法通常不能很好地扩展到大规模数据上，这阻碍了它在大数据问题中的应用。

受此启发，本章利用非线性表征学习[49] 对 MMD 进行扩展，给出一种概率分布差异度量的新准则。具体地说，不再手动地设置核函数及其关联的非线性特征映射 $\phi(\cdot)$，而是自动地由非线性表征学习方法确定合适的特征变换 $T(\cdot)$。可采用的非线性表征学习方法包括栈式去噪自动编码器（Stacked Denoising Autoencoders，

SDA）[125] 或深度置信网络（Deep Belief Net，DBN）[54] 等。这样扩展得到的准则称为非线性分布距离（Nonlinear Distribution Discrepancy，NDD），形式化定义如下。

定义 5.2 (**NDD**)：　记 $\mathcal{F}$、P、Q、$\mathcal{X}_s$ 和 $\mathcal{X}_t$ 如前所述，记 $T:\mathcal{X}\to\mathcal{H}$ 为非线性表征学习模型 $\mathcal{M}$ 学习得到的特征变换。非线性分布距离（*Nonlinear Distribution Discrepancy*，*NDD*）的经验估计定义如下：

$$\mathrm{NDD}\left[\mathcal{F},\mathcal{X}_s,\mathcal{X}_t\right]=\left\|\sum_{\mathbf{x}_i\in\mathcal{X}_s}\frac{T\left(\mathbf{x}_i\right)}{|\mathcal{X}_s|}-\sum_{\mathbf{x}_j\in\mathcal{X}_t}\frac{T\left(\mathbf{x}_j\right)}{|\mathcal{X}_t|}\right\|_{\mathcal{H}} \tag{5-2}$$

NDD 相对于 MMD 有如下几方面优势：1）计算 NDD 所依赖的非线性映射 $T(\cdot)$ 可从数据中自动学习，不需要手动设置参数化核函数；2）NDD 复杂度线性于样本集大小和特征维度，可扩展到大规模数据集的分析处理。下面将探索如何用深度学习方法使 NDD 可以刻划领域间概率分布的失配程度。

5.3　领域不变深度表征

本节首先介绍问题定义，其次提出非线性分布距离（Nonlinear Distribution Discrepancy，NDD）度量和领域不变深度表征（Invariant Deep Representation，IDR）模型，最后讨论模型选择的迁移交叉验证（Transfer Cross-Validation，TCV）策略。

5.3.1　问题定义

本章主要解决目标领域没有任何标注数据的无监督迁移学习问题。给定标注的辅助领域 $\mathcal{X}_s=\{(\mathbf{x}_1^s,y_1^s),\ldots,(\mathbf{x}_{|\mathcal{X}_s|}^s,y_{|\mathcal{X}_s|}^s)\}$ 和无标的目标领域 $\mathcal{X}_t=\{\mathbf{x}_1^t,\ldots,\mathbf{x}_{|\mathcal{X}_t|}^t\}$，分别由不同的概率分布 P_s 和 P_t 采样生成。本章不要求领域间共享相同的特征空间，在领域间特征空间不同时通过对特征向量填零得到同一维度 d 的特征空间。本章目标是学习一个深度特征表示 $T(\mathbf{x})$ 用以刻划领域间的不变结构，从而使在辅助领域 $\mathcal{X}_s$ 训练而得的分类模型可以很好地泛化到目标领域 $\mathcal{X}_t$。上述通用迁移学习框架还可以直接推广到处理多个目标领域的情况。常用符号及其描述如表 5.1 所示。

表 5.1　本章常用的符号及其描述。

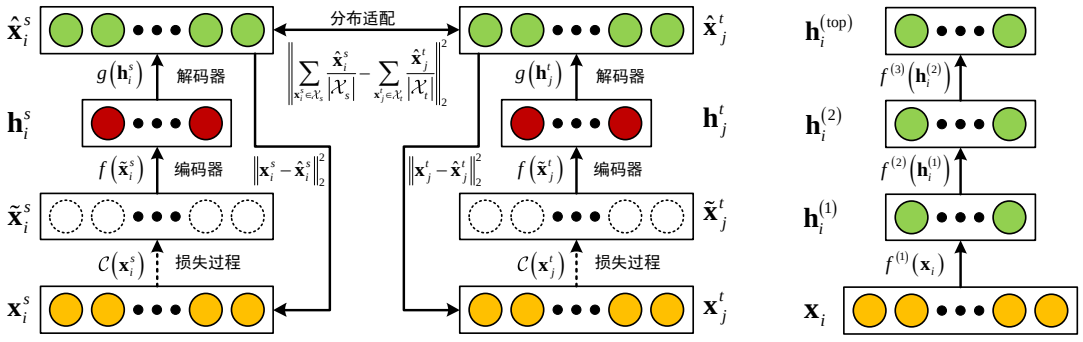
符号	描述	符号	描述	符号	描述
$\mathcal{X}_s$	辅助领域	$\mathcal{X}_t$	目标领域	$\mathbf{X}$	原始数据矩阵
d	特征维度	n	样例数目	$\widetilde{\mathbf{X}}$	损坏数据矩阵
p	损坏概率	f	编码器	$\mathbf{W}$	特征变换
λ	适配正则项	g	解码器	$\mathbf{D}$	NDD 指示矩阵

5.3.2　栈式去噪自动编码器

最近以来，深度学习[49] 由于能够从复杂数据中学习抽象的、紧致的、层次的和深度的特征表示而受到研究界的广泛关注。本章考察*栈式去噪自动编码器*（Stacked Denoising Autoencoders，SDA）[125]，其优点是可以有效地处理文本和视觉数据。SDA 的基本组件是一个称为*去噪自动编码器*（Denoising Autoencoder，DA）的单层神经网络，其目标是对人工损坏的输入特征进行去噪，即学习从数据的损坏版本重构数据的真实版本。DA 捕捉了输入数据概率分布的结构信息并以最优方式消除特征损坏的影响，其重构的数据表征往往位于原始数据附近但比损坏版本位于更高概率密度的区域，通过避开低概率密度区域来抽取更鲁棒的特征表示。具体地说，记 $\mathcal{X}$ 为输入样例集合，$\mathbf{x}_i, \widetilde{\mathbf{x}}_i, \widehat{\mathbf{x}}_i \in \mathcal{X}$ 分别为原始数据、$\mathbf{x}_i$ 的损坏版本和 $\mathbf{x}_i$ 的重构版本；记 $\mathbf{W}$ 和 $\mathbf{W}^{\mathsf{T}}$ 分别为编码器和解码器的权重矩阵，$\mathbf{b}$ 和 $\mathbf{c}$ 分别为编码器和解码器的截距向量。DA 通过如下的优化问题对损坏数据进行去噪和重构：

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{b},\mathbf{c}} \mathrm{DA}\left[\mathcal{X}\right] = \mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\sum\nolimits_{\mathbf{x}_i \in \mathcal{X}} \left\|\mathbf{x}_i - \widehat{\mathbf{x}}_i\right\|_2^2\right] \\ &\mathbf{h}_i = f\left(\widetilde{\mathbf{x}}_i\right) = a\left(\mathbf{W}\widetilde{\mathbf{x}}_i + \mathbf{b}\right), \widehat{\mathbf{x}}_i = g\left(\mathbf{h}_i\right) = a\left(\mathbf{W}^{\mathsf{T}}\mathbf{h}_i + \mathbf{c}\right) \end{aligned} \tag{5-3}$$

其中函数 $f(\cdot)$ 和 $g(\cdot)$ 分别是编码器和解码器，$a(\cdot)$ 是非线性激活函数，常用的有逻辑斯特函数 $a\left(x\right) = \frac{1}{1+e^{-x}}$ 或双曲正切函数 $a\left(x\right) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$（本章采用双曲正切函数）。$\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}[\cdot]$ 对由损坏过程 $p\left(\widetilde{\mathbf{x}}_i|\mathbf{x}_i\right)$ 产生的所有损坏版本 $\widetilde{\mathbf{x}}_i$ 求取数学期望，作为模型损坏函数的一个重要因素。由于非线性编码、解码函数 f 和 g 的限制，该数学期望无法直接计算准确值。实际中，该优化问题通过随机梯度下降求解，每次迭代的随机梯度值由样例 $\mathbf{x}_i$ 的若干损坏版本求取。本章仅考虑蒙板损坏过程 $p\left(\widetilde{\mathbf{x}}_i|\mathbf{x}_i\right)$，即每个样例 $\mathbf{x}_i$ 都独立地通过特征消除得到损坏版本——也就是 $\mathbf{x}_i$ 的每个特征都被独立地按照损坏概率 $p \in [0,1]$ 置零。这种蒙板损坏过程对稀疏文本数据尤其有效。



(a) IDA　　(b) 栈式 IDA

图 5.2　(a) 不变去噪自动编码器（IDA）与 (b) 栈式不变去噪自动编码器（栈式 IDA）。

记 $T(\cdot)$ 为 DA 学到的非线性特征变换，直观上可以定义 $T(\mathbf{x})=g(f(\widetilde{\mathbf{x}}))=\widehat{\mathbf{x}}$。由于模型中的特征损坏和网络激活都是非线性过程，上述变换具有高度的非线性，可以充分的刻划概率分布的失配程度。将该变换带入 NDD 的定义 (5-2) 得到：

$$\mathrm{NDD}[\mathcal{F},\mathcal{X}_s,\mathcal{X}_t]=\left\|\sum_{\mathbf{x}_i\in\mathcal{X}_s}\frac{\widehat{\mathbf{x}}_i}{|\mathcal{X}_s|}-\sum_{\mathbf{x}_j\in\mathcal{X}_t}\frac{\widehat{\mathbf{x}}_j}{|\mathcal{X}_t|}\right\|_2 \tag{5-4}$$

为把 DA 改造成可以在不同领域 $\mathcal{X}_s$ 和 $\mathcal{X}_t$ 之间泛化的模型，记辅助领域和目标领域的全集为 $\mathcal{X}=\mathcal{X}_s\cup\mathcal{X}_t$，将 NDD 作为适配正则项加入到 DA 的目标函数，得到如下不变去噪自动编码器（Invariant Denoising Autoencoder，IDA）联合优化问题：

$$\begin{gathered}\min_{\mathbf{W},\mathbf{b},\mathbf{c}}\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\sum_{\mathbf{x}_i\in\mathcal{X}}\left\|\mathbf{x}_i-\widehat{\mathbf{x}}_i\right\|_2^2+\lambda\sum_{s\neq t}\left\|\sum_{\mathbf{x}_i\in\mathcal{X}_s}\frac{\widehat{\mathbf{x}}_i}{|\mathcal{X}_s|}-\sum_{\mathbf{x}_j\in\mathcal{X}_t}\frac{\widehat{\mathbf{x}}_j}{|\mathcal{X}_t|}\right\|_2^2\right] \\ \mathbf{h}_i=f(\widetilde{\mathbf{x}}_i)=a(\mathbf{W}\widetilde{\mathbf{x}}_i+\mathbf{b}),\widehat{\mathbf{x}}_i=g(\mathbf{h}_i)=a\left(\mathbf{W}^{\mathsf{T}}\mathbf{h}_i+\mathbf{c}\right)\end{gathered} \tag{5-5}$$

其中 $\lambda>0$ 是适配正则项参数。通过在学习非线性特征变换的过程中同时最小化 NDD 准则，IDA 可以学习抽象的、紧致的且领域不变的特征表示以充分地刻划领域间的不变结构。IDA 的网络结构如图 5.2(a) 所示。

有了单层的领域不变自动编码器后，深度架构的构造方法如下：在首层 IDA 训练完成后，在其输出端逐层地叠加多个 IDA 模型并进行逐层地训练，即前一层 IDA 的输出作为后一层 IDA 的输入[125]，如图 5.2(b) 所示。由于上述栈式 IDA 可以学习领域不变深度表征，因此称其为（Invariant Deep Representation，IDR）。本章通过文献[125] 描述的逐层随机梯度下降算法对 IDR 模型进行训练，详细步骤从略。

5.3.3　边际化栈式去噪自动编码器

栈式 IDR 方法受限于较高的计算代价，会显著慢于主流迁移学习方法[36,96]，其主要瓶颈在于训练模型参数需要计算密集型的迭代式梯度下降算法。文献[52] 最近提出了边际化栈式去噪自动编码器（Marginalized Stacked Denoising Autoencoders，mSDA），比 SDA 取得了几个数量级的提速。受此启发，本章将边际化策略应用到 IDR 模型中，使特征损坏过程可以通过求取随机损坏变量的数学期望得到闭式解。边际化的效果等价于用原始样例的无穷多个损坏版本来训练去噪自动编码器。为使边际化操作变得可行，需要把耦合的非线性编码器和解码器简化为线性的特征变换 $\widehat{\mathbf{x}}=\mathbf{W}\widetilde{\mathbf{x}}$，并在该模型训练后注入非线性和深度结构信息。将上述线性特征变换带入 IDR 模型 (5-5) 得到边际化的 IDA 优化问题的矩阵形式：

$$\min_{\mathbf{W}}\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\left\|\mathbf{X}-\mathbf{W}\widetilde{\mathbf{X}}\right\|_F^2+\lambda\mathrm{tr}\left(\mathbf{W}\widetilde{\mathbf{X}}\mathbf{D}\widetilde{\mathbf{X}}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\right)\right] \tag{5-6}$$

其中 $\mathbf{D} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ 是 NDD 指示矩阵，定义为 $\mathbf{D} = \sum_{s\neq t} \mathbf{D}_{st}$；$\mathbf{D}_{st}$ 是领域 $\mathcal{X}_s$ 和领域 $\mathcal{X}_t$ 之间的成对 NDD 指示矩阵（代表多个领域间的两两分布距离度量），定义如下：

$$\mathbf{D}_{st} = \mathbf{d}_{st}\mathbf{d}_{st}^{\mathsf{T}} \text{ 其中 } (\mathbf{d}_{st})_i = \begin{cases} \frac{1}{|\mathcal{X}_s|}, & \mathbf{x}_i \in \mathcal{X}_s \\ \frac{-1}{|\mathcal{X}_t|}, & \mathbf{x}_i \in \mathcal{X}_t \\ 0, & \text{其他} \end{cases} \tag{5-7}$$

注意到直接计算 $\mathbf{D}$ 需要平方复杂度 $O(|\mathcal{X}|^2)$，这不满足处理大规模数据的线性复杂度要求。实际中，由于 $\mathbf{D}_{st}$ 是列向量 $\mathbf{d}_{st}$ 的外积，因此可由矩阵乘法结合律将 NDD 重写为 NDD$= \sum_{s\neq t} \mathrm{tr}\left[\left(\mathbf{W}\widetilde{\mathbf{X}}\mathbf{d}_{st}\right)\left(\mathbf{d}_{st}^{\mathsf{T}}\widetilde{\mathbf{X}}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\right)\right]$，这就将计算复杂度降低为线性 $O(|\mathcal{X}|)$。

iDA 的一个重要优势是它的优化问题公式 (5-6) 可以求得解析的闭式解：

$$\mathbf{W} = \left(\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\mathbf{X}\widetilde{\mathbf{X}}^{\mathsf{T}}\right]\right)\left(\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathsf{T}} + \lambda\widetilde{\mathbf{X}}\mathbf{D}\widetilde{\mathbf{X}}^{\mathsf{T}}\right]\right)^{-1} \tag{5-8}$$

可通过与文献[52] 类似方法对上式中的数学期望计算如下。由于 $\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathsf{T}}\right] = \sum_{\mathbf{x}_i\in\mathcal{X}} \mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathsf{T}}\right]$，因此可先计算 $\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathsf{T}}\right]$。具体地说，矩阵 $\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathsf{T}}$ 非对角元对应特征 α 和 β 没有被损坏过程置零当且仅当两个特征都没有被置零，其概率为 $(1-p)^2$；相应地，矩阵 $\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathsf{T}}$ 对角元对应特征 α 或 β 没有被损坏过程置零的概率为 $1-p$。定义“生存”概率向量为 $\mathbf{q} = [1-p, \ldots, 1-p, 1]^{\mathsf{T}} \in \mathbb{R}^{d+1}$，其中常数特征永远不会被损坏过程置零，因此其“生存”概率为 $\mathbf{q}_{d+1} = 1$。记 $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathsf{T}}$ 为原始无损数据的散度矩阵，则公式 (5-8) 中的数学期望可推导如下：

$$\mathbb{E}\left[\mathbf{X}\widetilde{\mathbf{X}}^{\mathsf{T}}\right]_{\alpha\beta} = \mathbf{S}_{\alpha\beta}\mathbf{q}_\beta,\ \mathbb{E}\left[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathsf{T}}\right]_{\alpha\beta} = \begin{cases} \mathbf{S}_{\alpha\beta}\mathbf{q}_\alpha\mathbf{q}_\beta, & \alpha \neq \beta \\ \mathbf{S}_{\alpha\beta}\mathbf{q}_\alpha, & \alpha = \beta \end{cases} \tag{5-9}$$

同理，记原始无损数据的均值散度矩阵为 $\overline{\mathbf{S}} = \mathbf{X}\mathbf{D}\mathbf{X}^{\mathsf{T}}$ 且注意到 $\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\widetilde{\mathbf{X}}\mathbf{D}\widetilde{\mathbf{X}}^{\mathsf{T}}\right] = \sum_{\mathbf{x}_i\in\mathcal{X}}\sum_{\mathbf{x}_j\in\mathcal{X}} d_{ij}\mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_j^{\mathsf{T}}\right]$，则与概率分布距离 NDD 相关的数学期望可计算如下：

$$\mathbb{E}\left[\widetilde{\mathbf{X}}\mathbf{D}\widetilde{\mathbf{X}}^{\mathsf{T}}\right]_{\alpha\beta} = \begin{cases} \overline{\mathbf{S}}_{\alpha\beta}\mathbf{q}_\alpha\mathbf{q}_\beta, & \alpha \neq \beta \\ \overline{\mathbf{S}}_{\alpha\beta}\mathbf{q}_\alpha, & \alpha = \beta \end{cases} \tag{5-10}$$

非线性和深度架构可能是深度学习得以成功的两个关键性因素[49]（虽仍存争议），因此在边际化模型 iDA 中也必须考虑这两个因素。与文献[52] 类似，在 iDA 训练得到变换矩阵 $\mathbf{W}$ 后，非线性通过双曲正切函数 $\tanh(\cdot)$ 注入到模型中。为实现深度网络，将第 $(t-1)$ 层模型的输出（注入非线性后）作为第 t 层模型的输入并将多个 iDA 模型堆叠在一起形成一个深度网络。上述由多个 iDA 经过非线性化和网络堆叠得到的迁移学习模型称为 iDR（IDR 的边际化版本），总结如算法 6 所示。

算法 6: 不变深度表征（iDR）

输入: 原始数据矩阵 $\mathbf{X}$，NDD 指示矩阵 $\mathbf{D}$；损坏概率 p，适配正则项参数 λ，网络层数 l。

输出: 领域不变深度表征 $\mathbf{R}$。

开始
  初始化 $\mathbf{X}^0 := \mathbf{X}$（注意错误的初始化方式 $\mathbf{X}^0 := \widetilde{\mathbf{X}}$）。
  for $t = 1$ **to** l **do**
    由公式 (5-8) 至 (5-10) 解得第 t 层 iDA 模型的变换矩阵 $\mathbf{W}^t$。
    计算第 t 层 iDA 模型输出的特征表示 $\mathbf{X}^t := \tanh\left(\mathbf{W}^t\mathbf{X}^{t-1}\right)$。
  返回领域不变特征表示 $\mathbf{R} := \mathbf{X}^l$ 及其辅助领域分区上训练的分类模型。

为处理超高维问题，采用文献[52]的特征分区思想，将特征集合划分为 K 个独立的子集 $\mathcal{S} = \bigcup_{k=1}^{K} \mathcal{S}_k$，并学习多个特征变换矩阵 $\mathbf{W}_k \in \mathbb{R}^{r\times|\mathcal{S}_k|}$ 实现每个特征子集 $\mathcal{S}_k$ 对高频子集 $\mathcal{S}_p$（包含 r 个最频繁特征）的有效重构。记 $\mathbf{X}_p \in \mathbb{R}^{r\times n}$ 为输入数据矩阵在高频子集 $\mathcal{S}_p$ 上的分区，$\widetilde{\mathbf{X}}_k \in \mathbb{R}^{|\mathcal{S}_k|\times n}$ 为损坏数据矩阵在第 k 个特征子集 $\mathcal{S}_k$ 上的分区。每个特征映射 $\mathbf{W}_k$ 由如下的优化问题进行学习：

$$\min_{\mathbf{W}_k} \mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\left\|\mathbf{X}_p - \mathbf{W}_k\widetilde{\mathbf{X}}_k\right\|_F^2 + \lambda\mathrm{tr}\left(\mathbf{W}_k\widetilde{\mathbf{X}}_k\mathbf{D}\widetilde{\mathbf{X}}_k^{\mathsf{T}}\mathbf{W}_k^{\mathsf{T}}\right)\right] \tag{5-11}$$

首层 iDA 网络的输出定义为 $\mathbf{X}^1 = \tanh\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{W}_k\mathbf{X}_k\right)$。首层输出维度 $r \ll d$ 的中间表征后，可以在其后堆叠多层 iDA 模型，后续学习过程如算法 6 所示。上述特征分区策略可以将 iDR 的计算复杂度由特征维度的三次方 $O(d^3)$ 降至线性 $O(dr^2)$。

负迁移问题：iDR 在不同领域 $\mathcal{X}_s$ 和 $\mathcal{X}_t$ 之间共享了深度网络的所有权重参数 $\mathbf{W}$，因此可能导致负迁移问题。为了同时增强深度表征对不同数据分布的鲁棒性和不变性，采用“参数部分共享”策略将权重参数 $\mathbf{W}$ 分解为领域不变部分 $\mathbf{W}_0$ 和领域可变部分 $\mathbf{W}_r, r \in \{s,t\}$，得到如下的鲁棒不变深度表示（Robust iDR，RiDR）：

$$\min_{\mathbf{W}_0,\{\mathbf{W}_r\}} \mathbb{E}_{p(\widetilde{\mathbf{x}}|\mathbf{x})}\left[\sum_{r\in\{s,t\}}\left\|\mathbf{X}_r - \left(\mathbf{W}_0 + \mathbf{W}_r\right)\widetilde{\mathbf{X}}_r\right\|_F^2 + \lambda\mathrm{tr}\left(\mathbf{W}_0\widetilde{\mathbf{X}}\mathbf{M}\widetilde{\mathbf{X}}^{\mathsf{T}}\mathbf{W}_0{}^{\mathsf{T}}\right)\right] \tag{5-12}$$

上述模型同时解决了迁移学习的欠拟合、欠适配和负迁移问题，通过协同解决各种问题形成强化（reinforcement）机制增强迁移能力，实现统一的迁移学习框架。

5.4　迁移交叉验证

自动化模型选择对机器学习算法的成功应用是至关重要的。对于传统单一领域问题，标准交叉验证（Standard Cross-Validation，SCV）是自动选取模型最佳参

数设置和评价模型在独立测试集上泛化性能的基本工具。在最简单的 2-折交叉验证中，标注数据被随机地划分为训练集和验证集；特定参数设置下的学习器在训练集上构造并在验证集上做模型评价，验证集上效果最好的模型被选出作为未来预测模型。对于无监督迁移学习问题目标领域没有任何标注数据，因此必须在标注好的辅助领域上执行 SCV，选取辅助领域验证集上效果最好的模型。然而，由于辅助领域和目标领域概率分布差异很大，上述策略无法选择针对目标领域性能最优的模型。换言之，上述策略随机生成的训练集和验证集将会服从相同的概率分布，这与学习器应能在服从不同概率分布的目标领域上进行泛华的基本要求矛盾。目标领域最优迁移学习器应当在与训练集服从不同概率分布的验证集上选取。

针对上述问题，本章提出*迁移交叉验证*（Transfer Cross-Validation，TCV）方法，用于目标领域没有标注数据的无监督迁移学习模型选择。关键思想是手动构造一个*异构的*训练/验证配置，用来对辅助领域/目标领域概率分布的异构性进行模拟。这种构造从理论上是合理可行的，因为迁移学习中辅助领域和目标领域具有一定程度的相关性，从而可以从辅助领域中选择一部分与目标领域概率分布一致的样例。这部分辅助领域标注数据作为“验证集”，也即表现为目标领域的代理；其余的辅助领域标注数据作为“训练集”。这样手动配置的训练/验证集合将服从*不同*的概率分布，因为根据构造方法，训练集与目标领域不同而验证集与目标领域相似。由于训练集和验证集都来自辅助领域标注数据，可用它们进行模型评价。

剩下的问题就是如何选取辅助领域中与目标领域潜在一致的样例。受到*核均值匹配方法*[27] 的启发，调整辅助领域样例权重 $\{\alpha_i : \mathbf{x}_i \in \mathcal{X}_s\}$ 使 NDD 准则最小化：

$$\min_{0\leqslant\boldsymbol{\alpha}\leqslant B}\left\|\sum\nolimits_{\mathbf{x}_i\in\mathcal{X}_s}\alpha_i\frac{T(\mathbf{x}_i)}{|\mathcal{X}_s|}-\sum\nolimits_{\mathbf{x}_j\in\mathcal{X}_t}\frac{T(\mathbf{x}_j)}{|\mathcal{X}_t|}\right\|_{\mathcal{H}}^2 \tag{5-13}$$

其中 T 是由非线性表征学习模型（本章中采用 IDR 和 iDR）抽取得到的非线性特征变换；B 是避免 $\boldsymbol{\alpha}$ 发散到无穷大的上界约束，典型取值为 $B\in[1,10]$，其取值不会影响所选取的辅助领域样例，因为仅需关心 $\{\alpha_i\}$ 之间的排序关系而不需关心它们的准确值。上述优化问题可通过如下带线性约束条件的二次规划方法求解[27]：

$$\begin{gathered}\min_{0\leqslant\boldsymbol{\alpha}\leqslant B}\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{K}\boldsymbol{\alpha}-2\boldsymbol{\kappa}^{\mathsf{T}}\boldsymbol{\alpha}\\ \text{其中 } K_{ij}=T(\mathbf{x}_i)^{\mathsf{T}}T(\mathbf{x}_j),\kappa_i=\frac{|\mathcal{X}_s|}{|\mathcal{X}_t|}\sum\nolimits_{\mathbf{x}_j\in\mathcal{X}_t}T(\mathbf{x}_i)^{\mathsf{T}}T(\mathbf{x}_j)\end{gathered} \tag{5-14}$$

最后对 $\{\alpha_i\}$ 按从大到小排序，选取 $\{\alpha_i\}$ 值最大的前 $1/k$ 比例的辅助领域样例作为代理验证集（与目标领域相似），其余 $1-1/k$ 比例的辅助领域样例作为训练集（与目标领域不同），其中 k 是验证集的比例，如 $k=2$ 表示 2-折交叉验证。值得一提的是，TCV 和方差偏移场景下的其他交叉验证方法[127,128] 有如下几方面明显

的区别：（1）TCV 执行在经过 $T(\cdot)$ 变换的数据上，考察了领域间概率分布在所谓“领域不变”深度表征下依然不能有效适配的困难情况，其他方法没有考虑这种情况；（2）TCV 通过手动配置训练/验证集合来对辅助领域/目标领域的异构性进行模拟，而其他方法只是在实例权重加权的数据上进行随机的训练/验证集合划分。

5.5　实验过程与结果

本节在情感极性分类、垃圾邮件过滤、视觉对象识别等三类实际应用问题中进行系统性实验，全面地验证本章方法 IDR 和 iDR 的准确率、有效性和可扩展性。

5.5.1　实验数据

5.5.1.1　多领域情感数据集

文献[36] 发布的多领域情感数据集① 是评测迁移学习和情感分类算法的基准数据集。该数据集来自 Amazon.com 上的产品评论，包括 4 个产品领域：*books*（**B**）、*dvds*（**D**）、*electronics*（**E**）和 *kitchen appliances*（**K**）。每个评论被赋予一个正面极性（评价分数高于 3 星）或负面极性（评分分数不高于 3 星），并由词频向量表征。每个领域包括 2,000 个标注数据和大约 4,000 个无标数据（每个领域的具体数目稍有差别），正面极性和反面极性的样例数基本平衡。绝大部分已有工作[36,42,52,96] 都给出了如下 12 个情感迁移任务的分类准确率：**D→B**、**E→B**、**K→B**、**B→D**、**E→D**、**K→D**、**B→E**、**D→E**、**K→E**、**B→K**、**D→K**、**E→K**，其中箭头前面的符号表示辅助领域，箭头后面的符号表示目标领域。该数据集详细信息如表 5.2 所示。

5.5.1.2　垃圾邮件过滤数据集

由 ECML/PKDD 2006 知识发现挑战赛② 发布的垃圾邮件过滤数据集（任务 A）包括 4 个独立的用户收件箱，可进一步分为个人收件箱 **u0**、**u1**、**u2** 和公共收件箱 **u**∗。每个个人收件箱有 1,250 个垃圾邮件和 1,250 个正常邮件；公共收件箱包括来自公共领域的 2,000 个垃圾邮件和 2,000 个正常邮件；每个邮件均由词项频率向量表征。由于组内概率分布相对比较接近，而组间概率分布差异十分显著，本实验在跨组的收件箱之间构造了 6 个迁移分类任务：**u0→u**∗、**u1→u**∗、**u2→u**∗、**u**∗**→u0**、**u**∗**→u1**、**u**∗**→u2**。例如，**u0→u**∗ 表示 **u0** 作为辅助领域、**u**∗ 作为目标领域的垃圾邮件过滤任务。该数据集的详细信息如表 5.3 所示。

① http://www.cs.jhu.edu/~mdredze/datasets/sentiment

② http://www.ecmlpkdd2006.org/challenge.html

表 5.2　多领域情感数据集的统计信息。

领域	评论数	训练样例数	测试样例数	正例比例	特征数
Books (**B**)	6,465	2,000	4,465	50%	30,000
DVD (**D**)	5,586	2,000	3,586	50%	30,000
Electronics (**E**)	7,681	2,000	5,681	50%	30,000
Kitchen (**K**)	7,945	2,000	5,945	50%	30,000

表 5.3　垃圾邮件过滤数据集的统计信息。

领域	邮件数	正例数	负例数	特征数
公共 (**u**∗)	4,000	2,000	2,000	206,908
个人 (**u0, u1, u2**)	2,500	1,250	1,250	206,908

5.5.1.3　视觉对象识别数据集

Office[107,108] 是视觉迁移学习中的主流基准数据集，包括 31 个类别 4,652 张图片，来自真实对象领域：*Amazon*（在线电商图片）、*Webcam*（网络摄像头拍摄的低解析度图片）、*DSLR*（单反相机拍摄的高解析度图片）。Caltech-256① 是对象识别的基准数据集，包括 256 个类别 30,607 张图片。在本组实验中，直接采用文献[108] 发布的 Office+Caltech 预处理集②。对每张图片抽取 SURF 特征，并向量化为 800 维的直方图表征，所有直方图向量都进行减均值除方差的归一化处理，直方图码表由 K 均值聚类算法在 *Amazon* 子集上生成。具体共有 4 个领域 **C** (Caltech-256)、**A**（Amazon）、**W**（Webcam）和 **D**（DSLR），从中随机选取 2 个不同的领域作为辅助领域和目标领域，则可构造 $4 \times 3 = 12$ 个跨领域视觉对象识别任务，如 $C \to A$，$C \to W$，$C \to D$，…，$D \to W$。该数据集的详细信息如表 5.4 所示。

5.5.2　基准算法和实现细节

5.5.2.1　基准方法

为了系统性地对本章方法 IDR 和 iDR 进行评测，考察了多种主流和前沿的迁移学习方法，具体如下。作为基线方法，在辅助领域原始特征上训练一个线性 SVM 分类器并在目标领域上进行测试。对于跨领域情感分类问题，结构对应学习（Structural Correspondence Learning，SCL）[36] 和谱特征对齐（Spectral Feature Alignment，SFA）[42] 可能是应用最广泛的两种方法，因而也考察它们在本章数据集上的效能。另外还比较了基于协同训练的迁移学习（Co-Training for Domain Adaptation，CODA）[96]，该方法是基于浅层结构的迁移学习方法中在多领域情感

① http://www.vision.caltech.edu/Image_Datasets/Caltech256
② http://www-scf.usc.edu/~boqinggo/domainadaptation.html

表 5.4　视觉对象识别数据集的统计信息。

领域	图片数	特征数	类别数	标签
Caltech (**C**)	1,123	800	10	backpack, bike, calculator,
Amazon (**A**)	958	800	10	headphones, keyboard,
Webcam (**W**)	295	800	10	laptop, monitor, mouse,
DSLR (**D**)	157	800	10	coffee-mug, video-projector

数据上效果最好的方法之一。由于 SCL、SFA 和 CODA 都仅面向文本领域，本章又考虑了两种通用迁移学习方法，*迁移成分分析*（Transfer Component Analysis、TCA）[45] 和*测地流核方法*（Geodesic Flow Kernel，GFK）[108]，它们在视觉领域取得良好的效果。最后，除了前述基于浅层结构的迁移学习方法，本章还考虑最近提出的基于*深度学习*的迁移学习方法，特别考察了*边际化栈式去噪自动编码器*（Marginalized Stacked Denoising Autoencoders，mSDA）[52]，因为它是与本章方法最相似的方法且在多领域情感数据集上取得了除本章方法外的最好效果。mSDA 由*栈式去噪自动编码器*（Stacked Denoising Autoencoders，SDA）[50] 改进而来，SDA 是最早将深度学习用于情感分类的工作。本章工作与 mSDA 有明显区别：本章方法显式地修正领域间概率分布的失配问题，从而比 mSDA 取得很大的性能提升。

5.5.2.2　实现细节

为执行公平的对比实验，本章采用与基准算法[52,96,108] 完全一致的评测协议。在无监督迁移学习中，学习算法无法获得目标领域的标注信息，因此部分基准方法难以通过标准交叉验证（SCV）进行自动化参数选择[85]。对于这类基准方法，本章要么采用它们在原文献中的调参策略（通常是启发性策略），要么为它们提供*最佳调参条件*，即在模型选择时提供目标领域的标注数据作为验证集（模型训练仍然在辅助领域标注数据上进行）。需要注意的是，在无监督迁移学习的实际应用中无法获得目标领域的标注信息，因而这种调参方法仅具有实验室场景下的意义。

IDR 基于栈式去噪自动编码器（SDA）[50]，可采用 SDA 的调参策略，考察以下参数：损坏概率 $p \in \{0.2, 0.3, 0.5, 0.6, 0.7, 0.8\}$；学习速率 $\eta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$；隐藏层神经元数 $\{1000, 2500, 5000\}$。根据文献[50]，所有这些 SDA 模型中的参数都可以通过辅助领域上的 5-折交叉验证得到。本章提出的 IDR 方法引入了一个关键性的适配正则项参数，其最优取值需要通过本章提出的迁移交叉验证（TCV）策略在范围 $\lambda \in \{0.01, 0.1, 1, 10, 100\}$ 中进行选择。IDR 的实现基于深度学习工具箱① 。

iDR 是 IDR 的边际化版本，仅包含两个超参数：损坏概率 p 和适配正则项参数

① https://github.com/rasmusbergpalm/DeepLearnToolbox

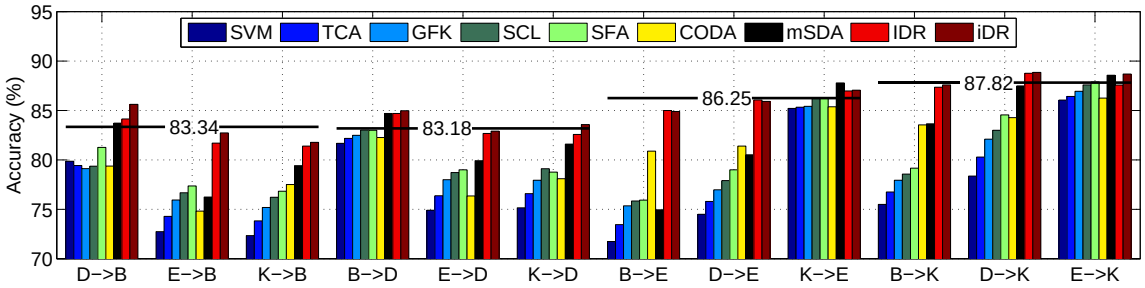


图 5.3　多领域情感数据集 12 个迁移学习任务的详细分类准确率。评论数据来自 4 个领域 *Book*、*DVD*、*Electronics* 和 *Kitchen appliances*。所有模型都在辅助领域的训练集上拟合，并在目标领域的测试集上预测。可以看到，IDR 和 iDR 在 12 个任务中的 11 个超越了基准方法，并较最好的基准方法 mSDA 准确率提升了 **3.00%**。

λ。参数 p 可类似 SDA 由辅助领域上的标准交叉验证（SCV）选择，但 SCV 无法选择最佳参数 λ（参见第5.5.5节阐述）。为此，采用本章的迁移交叉验证（TCV）自动选择参数 λ。SCV 和 TCV 考察的参数范围分别是 $p \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ 和 $\lambda \in \{0.01, 0.1, 1, 10, 100\}$。限于篇幅，本章主要关注 iDR 的实验分析；IDR 的性能与 iDR 类似，为保证实验完整性，这里给出了它在多领域情感数据集上的结果。

本章采用每个迁移任务在测试集（即目标领域无标数据）上的分类准确率（Accuracy）作为评测指标，这是因为该指标被众多文献[36,42,45,52,96,108]广泛采用：

$$Accuracy = \frac{\left|\mathbf{x} : \mathbf{x} \in \mathcal{X}_t \wedge f(\mathbf{x}) = y(\mathbf{x})\right|}{\left|\mathbf{x} : \mathbf{x} \in \mathcal{X}_t\right|} \tag{5-15}$$

其中 $y(\mathbf{x})$ 是样例 $\mathbf{x}$ 的真实类别（训练和验证阶段未知），$f(\mathbf{x})$ 是算法预测的类别。

5.5.3　实验结果

本节系统性地比较 IDR、iDR 与 7 个基准方法的分类准确率，考察范围是 30 个跨领域迁移学习任务，包括情感极性分类、垃圾邮件过滤和视觉对象识别任务。

5.5.3.1　情感分类结果

由该数据集构造了 12 个跨领域情感分类任务用于评测。图 5.3 展示了不同方法在所有任务上取得的分类准确率，每组柱条表示每个情感分类任务上的所有结果，每组颜色表示每种学习算法的所有结果。水平带数字横线表示准确率上界

表 5.5　各学习算法在所有迁移学习任务上的平均分类准确率 (%)。

数据集	SVM	TCA	GFK	SCL	SFA	CODA	mSDA	IDR	iDR
情感极性分类	77.33	78.39	79.45	80.18	80.75	80.84	*82.38*	84.91	**85.38**
垃圾邮件过滤	70.39	67.35	68.97	79.37	78.37	*82.41*	78.68	–	**89.79**
视觉对象识别	45.04	47.78	47.15	–	–	–	*49.83*	–	**51.32**

（UpperBound）[52]，即线性支持向量机（SVM）在目标领域训练集上拟合并在目标领域测试集上预测得到的准确率，代表了基于原始特征可能得到的最好效果——这里训练集和测试集的划分如表 5.2 所示，目标领域的标注数据不参与 IDR、iDR 和基准算法训练过程。表 5.5 列出所有方法在情感任务上的平均分类准确率。

本章方法 IDR 和 iDR 在 12 个分类任务中的 11 个取得了较所有基准方法更好的准确率（仅在任务 **K→E** 上比 mSDA 稍差），且在 12 个任务中的 8 个取得了大幅的准确率提升。iDR 在 12 个任务中的平均分类准确率达 **85.38%**，比最好的基准方法 mSDA 提高达 **3.00%**。尽本文作者所知，这个结果打破了多领域情感数据集上已有的性能记录，成为迄今已公开发表结果中最好的。令人印象深刻的是，IDR 和 iDR 甚至在 6 个任务中取得了比准确率上界 UpperBound 更好的效果，且在另外 4 个任务中取得了与之基本相当的效果。上述结果有力地证明 IDR 和 iDR 可以学习紧致的、高质量的、领域不变的深度表征，有效地解决迁移学习问题。

从图中还可以观察到，4 个不同产品领域大致可以进一步划分为 2 组：领域 **B** 和 **D** 彼此更相似可作为一组，领域 **K** 和 **E** 彼此也更相似可作为另一组，但组间差异很大——这也与现实生活中的感受相符：书籍和影碟都属于娱乐产品，有更多相似性；电子产品和厨具都属于生活设备，有更多相似性。因此，从领域 **K** 迁移分类器到领域 **E** 的难度，要比从领域 **B** 和 **D** 迁移分类器的难度小得多，这也与实验结果一致。十分有趣的是，随着迁移难度加大，iDR 相对于最佳基准方法的提升幅度也更大。换言之，在极其困难的迁移任务如 **E→B**、**B→K** 上，iDR 比竞争者有更好的表现。这表明对于极其困难的迁移学习任务而言，提取可以充分刻划领域不变结构的抽象的、紧致的、层次的、深度的特征表示尤为重要。iDR 同时采纳了深度学习和分布适配的思想，从而学习了符合预期的领域不变深度表征。

下面进一步分析所有基准方法的优势和不足。SVM 在标准情感分类问题中性能很好，但当训练数据和测试数据来自不同的领域并服从不同的概率分布时，SVM 的效果就差强人意了[36]。TCA 和 GFK 都是通用的迁移学习方法，都基于浅层学习架构（如主成份分析，PCA）修正概率分布在领域间的失配问题，不过它们并未取得比 SVM 明显的优势（并非所有情况，仅针对本章的情感数据如此）。值得一提的是，TCA 采用核方法和 MMD 准则来修正概率分布失配，但它的性能却远逊于本章提出的基于 NDD 准则的 iDR 方法，这有力地证明了 NDD 准则相对于 MMD 准则的优势所在。SCL 和 SFA 被广泛认为是情感极性分类问题中最先进的领域适配方法，其效能来源于对自然语言领域知识建模——确定一组领域共享的高频词项作为枢轴特征，通过普通词项与枢轴词项的共现关系抽取领域不变的特征子空间，并用于对领域间概率分布进行适配。然而，这两种方法都未能给出如何确定最佳枢轴特征的非启发式方法，这影响了它们的性能发挥。CODA 通过协

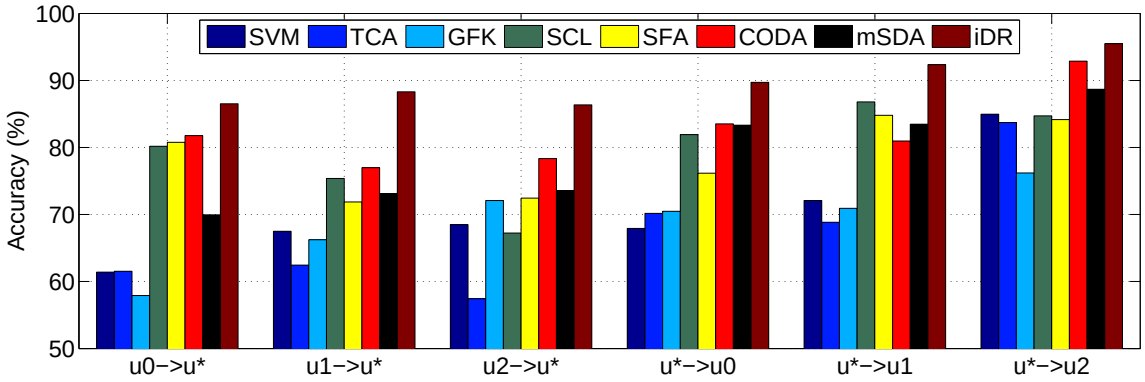


图 5.4　垃圾邮件过滤数据集 6 个迁移学习任务的分类准确率。邮件来自用户收件箱（领域）*u1*、*u2*、*u3* 和 *u∗*。iDR 在所有 6 个任务上显著超越了所有基准方法，并比最佳基准方法 CODA 准确率提升了 **7.38%**。

同训练对 SCL 和 SFA 进行改进，它根据领域相关性迭代式地执行样例和特征选择，从而得到在领域间保持不变的样例和特征。大致上，SCL、SFA 和 CODA 都局限于文本领域，不属于通用的迁移学习方法。随着深度学习的进展，浅层网络无法提取高度抽象的特征表示、不能很好地捕捉领域不变知识结构，已经成为迁移学习研究界的共识[125]。前述浅层迁移学习已被最近提出的标准深度学习方法 SDA 和 mSDA 超越，这些深度学习方法甚至没有显式地考虑领域失配问题[50,52]。基于联合的分布适配和深度学习，本章提出的 IDR 和 iDR 方法进一步大幅度超越标准深度学习方法 mSDA，这重新强调了领域失配问题的特殊性并非标准深度学习就能完全解决，强调了同时考虑深度架构和分布差异对深度迁移学习的重要性。

5.5.3.2　垃圾邮件过滤结果

由该数据集构造了 6 个跨收件箱（领域）垃圾邮件过滤任务，评测协议与上文完全一致。图 5.4 展示了所有方法的分类准确率，表 5.5 列出了所有方法在所有任务上的平均准确率。本章方法 iDR 比最佳基准方法 CODA 准确率大幅度提高了 **7.38%**。一个重要观察是，在该数据集上标准深度学习方法 mSDA 准确率较浅层迁移学习方法 CODA 下降明显。这进一步揭示了仅抽取深度特征而不进行概率分布适配不足以实现有效的迁移学习，尤其是在领域失配程度严重时。另一个观察是，CODA 取得了比 SCL 和 SFA 显著的准确率提升，这进一步强调了 SCL 和 SFA 通过启发式方法确定的枢轴特征并非最佳（选取领域不变特征是非平凡过程）。本章 iDR 是通用的学习方法，不需要上述启发式的、难调优的预处理步骤。

还可以观察到一个有趣的非对称性质：从领域 **A** 迁移到领域 **B** 与从领域 **B** 迁移到领域 **A** 是完全不同的。换言之，从公共收件箱 **u∗** 迁移到个人收件箱 **u0~u2** 的难度要远低于从反方向迁移。这可解释为，更一般性的领域通常包含更一般性

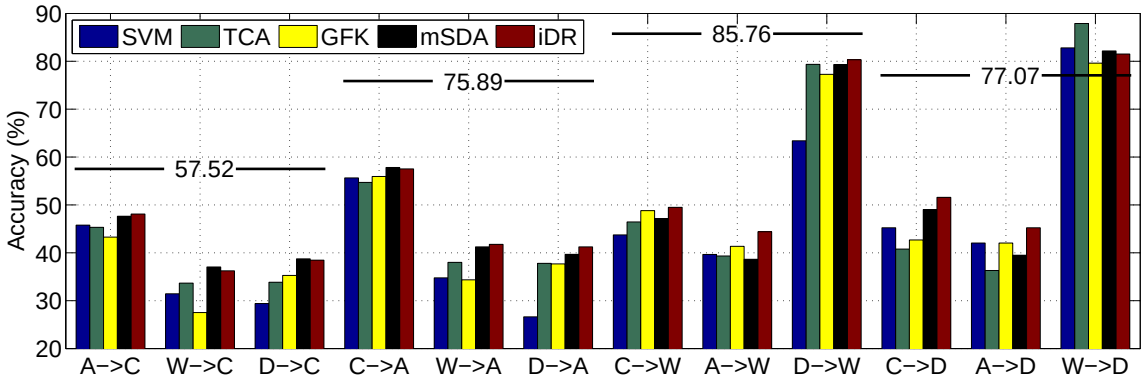


图 5.5　视觉对象识别数据集 12 个迁移学习任务的分类准确率。图像来自 4 个领域 *Caltech*、*Amazon*、*Webcam* 和 *DSLR*。在 12 个任务中的 8 个 iDR 取得了比基准方法更好的效果，并比最好的基准方法 mSDA 准确率提升了 **1.50%**。

的概念，更容易在领域间进行适配和迁移。这表明了更好的迁移学习方式是从一般性领域迁移到特殊性领域。在从特殊性领域迁移到一般性领域的困难迁移任务上，iDR 比竞争方法取得更大准确率提升，这证明 iDR 是一种鲁棒性很强的方法。

5.5.3.3　视觉对象识别结果

视觉迁移学习通常比文本迁移学习更具有挑战性，原因在于视觉底层特征与上层类别语义之间的语义鸿沟。面向文本的方法 SCL、SFA 和 CODA 无法正确地应用到视觉迁移学习，而 iDR 作为一种通用方法可以直接应用于视觉迁移学习任务。图 5.5 展示了除面向文本外的基准方法在 12 个跨领域视觉对象识别任务上的准确率。一般地，iDR 在 12 个任务中的 8 个取得了比基准方法更好的准确率。由于该数据集的适配难度极大，iDR 所取得性能提升也较小。但注意到，即便是已有的最先进迁移学习方法 GFK 也仅比标准方法 SVM 取得了有限的准确率提升。

5.5.4　深度分析

5.5.4.1　适配性逐层巩固

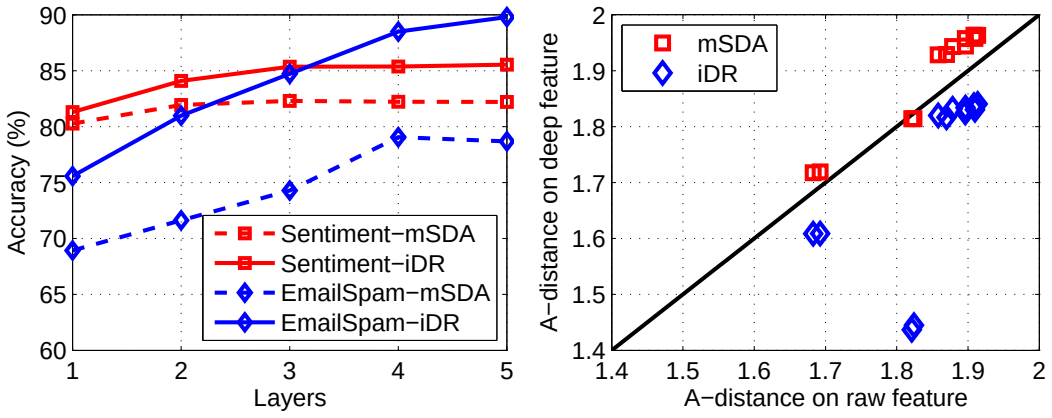
深度学习的一个最大优势是沿着网络深层方向可以抽取层次的、非线性的特征表示（通常利用神经网络或本章采用的栈式自动编码器）。因而一个值得探究的问题是：深度表征的适配性（迁移能力）是否会随着栈式自动编码器逐层加强和巩固？图 5.6(a) 展示了 iDR 分别在多领域情感数据集和垃圾邮件过滤数据集的所有迁移学习任务（12 个情感任务 6 个邮件任务）的平均分类准确率相对于网络层数的变化规律。如预期所示，mSDA 和 iDR 都随着网络层次增加取得更好的准确率——两者均随着深度增加学习到更抽象的特征表示。这证实了深度学习可以极大巩固特征表示在不同概率分布间的不变性和适配性，从而保证有效的迁移学习。

一个非常有趣的观察是，iDR 相对于 mSDA 的性能提升也随着网络层数增加而愈加显著。换言之，随着更多的自动编码器堆叠在网络上，iDR 的准确率以更大幅度超越 mSDA。这可解释为，特征不变性和适配性会在网络结构中逐层得以巩固加强。换言之，随着网络深度增加，所提取的特征表示具有更高的抽象性和非线性。根据定义 NDD 依赖于恰当的非线性变换从而有效地刻划概率分布差异，深度网络逐层增强的非线性可以增强 NDD 刻划领域间概率分布失配程度的能力。

5.5.4.2　代理 A 距离

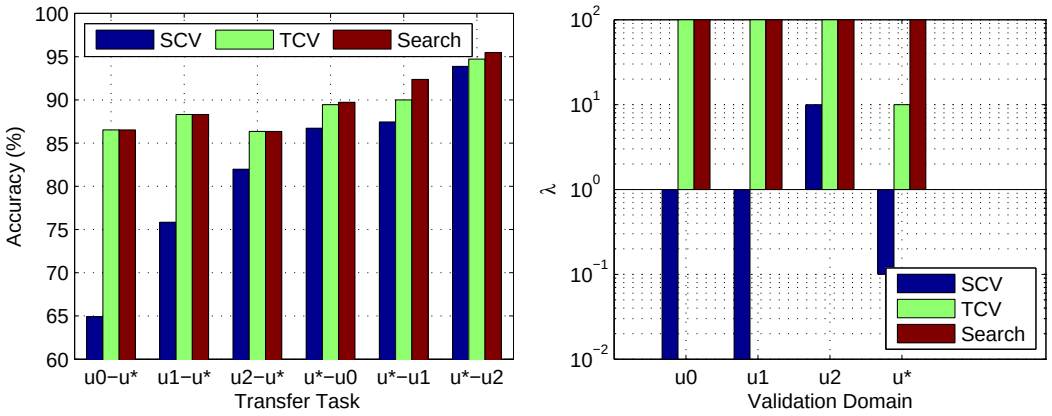
Ben-David 等人在文献[38] 中给出了迁移学习的一个理论结果，建议采用代理 A 距离（proxy-A-distance，PAD）作为两个概率分布的相似度量。他们证明了辅助领域和目标领域间的 PAD 距离是迁移学习泛化误差上界的一个重要组成部分。他们还假设为了取得良好的迁移学习效果，必须使得在新的特征表示下辅助领域和目标领域变得难以判别，因为这样意味着领域间有相似的概率分布。实际上，PAD 的准确值是无法计算的，因而需要计算它的一个近似或代理。代理 PAD 距离定义为 $\hat{d}_A = 2\,(1 - 2\epsilon)$，其中 ϵ 是分别以辅助领域和目标领域作为正例和反例训练得到的二分类器（本章采用线性 SVM）的泛化错误率或交叉验证错误率[50,52]。

图 5.6(b) 展示了分别在原始特征、mSDA 特征和 iDR 特征上计算得到的 PAD 距离。图中每个点代表多领域情感数据集的每个迁移学习任务（共 12 个任务对应 12 个点）的代理 A 距离（在辅助领域和目标领域之间），点的二维坐标定义为 $(\hat{d}_A(\text{raw}), \hat{d}_A(\text{deep}))$，也就是说，如果点在直线 $y = x$ 的上方，则表明深度表征上的 PAD 距离大于原始特征上的 PAD 距离；反之亦然。图 5.6(b) 揭示了一个令人意外的观察：在 12 个任务中的 11 任务，mSDA 特征上的 PAD 距离大于原始特征上的 PAD 距离——这意味着在 mSDA 特征表示上辅助领域和目标领域反而变得



(a) 适配逐层巩固　　(b) 代理 A 距离（PAD）

图 5.6　深度分析：(a) 表征适配性逐层巩固 (b) 作为领域泛化指标代理 A 距离（PAD）[38]。



(a) 目标领域测试准确率　　(b) 交叉验证选定的 λ 值

图 5.7　iDR 在不同的模型选择方法下的性能分析：(a) 目标领域测试准确率 (b) 不同模型选择方法选定的 λ 值。TCV 选择的模型准确率比 SCV 大幅提高且逼近最佳效果 Search。

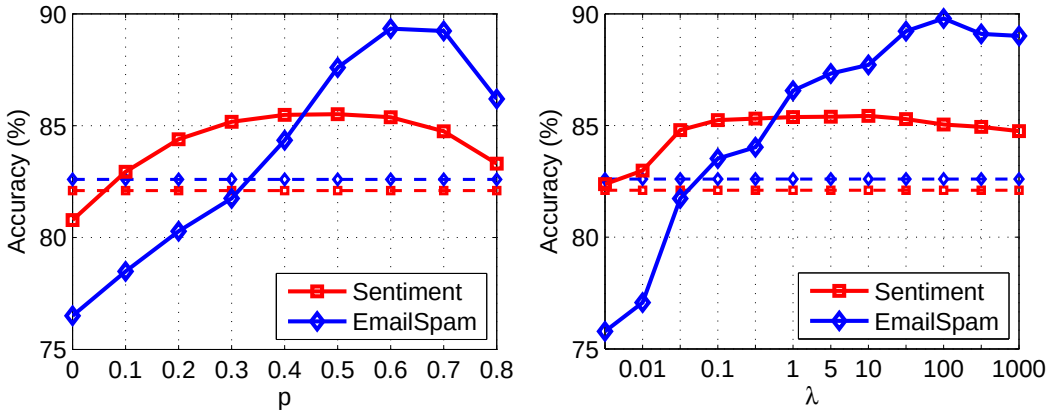
更易判别。文献[50,52] 对这个现象的解释为，深度特征抽取可以显化领域特定信息和情感特定信息，从而学习一般意义上更好的特征表示，它不但可以帮助更好地判别不同领域，还可以提高情感极性分类的效果。无论如何解释，根据 Ben-David 等人[38] 的理论，mSDA 特征表示应该会破坏迁移学习！对此本章结论为，标准深度学习抽取的特征表示不足以完全解决迁移学习问题，而这也是本章工作的动机。

如图 5.6(b) 所示，在所有 12 个任务中，iDR 表征上的 PAD 距离都显著小于原始特征上的 PAD 距离。根据 Ben-David 等人的理论[38]，更小的 PAD 距离预示着更低的领域间泛化误差上界，理论上保证了 iDR 能够取得更好的领域间泛化能力。从这个意义上看，iDR 理论上优于已有的基于深度学习的迁移学习方法[50–53]。同时，iDR 也优于已有的不基于深度学习的浅层迁移学习方法[36,42,45,96]，因为这些浅层方法无法抽取紧致的、层次的、非线性的深度表征用以刻划领域间不变结构。

5.5.5　迁移交叉验证分析

自动模型选择对于机器学习算法成功应用于实际问题至关重要。然而标准交叉验证 SCV 难以推广到目标领域没有标注数据的无监督迁移学习场景下。一个广泛采用的策略是在辅助领域标注数据上执行 SCV[50,52]，其问题是选定的参数满足在辅助领域误差最小而非在目标领域误差最小。另一个现有方法广泛采用的策略是为算法提供目标领域的标注数据以供模型选择和参数调优[45,60,85]，其问题是在无监督迁移学习中算法根本访问不到目标领域的标注数据，因而不具有实际意义。记该方法为 Search，因为它在给定目标领域标注数据作为参考下，穷举参数空间得到目标领域错误率最小的参数设置，这里把它作为算法可能达到的误差上界。

在垃圾邮件过滤数据集的 6 个迁移学习任务上评测几种模型选择方法 SCV、



(a) 损坏概率 p　　(b) 适配正则项 λ

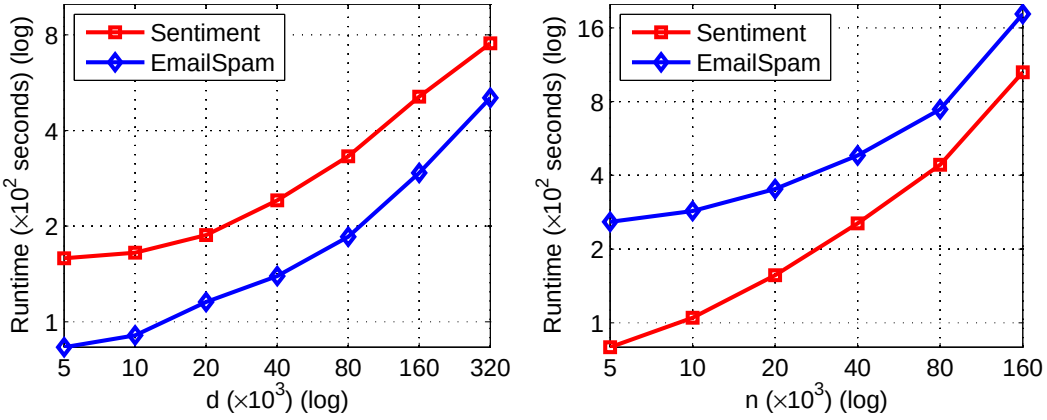
图 5.8　iDR 参数敏感性分析：(a) 损坏概率 p 和 (b) 适配正则项 λ。参数取值 $p \in [0.4, 0.8]$、$\lambda \in [0.1, 1000]$ 时，iDR 大幅超越最佳基准方法（情感数据为 mSDA，邮件数据为 CODA）。

TCV 和 Search。待选定的参数为 λ，所考察的取值范围为 $\{0.01, 0.1, 1, 10, 100\}$。图 5.7(a) 展示了 iDR 在不同模型选择策略下的目标领域测试准确率，图 5.7(b) 显示了对应选择的参数 λ 最优取值。结果显示，通过迁移交叉验证 TCV 选定的 iDR 模型显著超越了通过标准交叉验证 SCV 选定的 iDR 模型，SCV 选定的模型甚至逼近了穷举方法 Search 选定的模型的准确率。由于穷举方法 Search 在实际中不可能实现，TCV 可以认为是仅给定辅助领域标注数据条件下的最佳模型选择方法。基于上述实验，本文建议：辅助领域上的标准交叉验证 SCV 不是进行无监督迁移学习模型选择的合适方法，因为 SCV 最小化辅助领域验证误差会导致 $\lambda \to 0$，即最小化目标领域对辅助领域优化问题的副作用而非最大化领域间知识共享和迁移。

5.5.6　参数敏感性分析

基于去噪自动编码器的深度学习方法涉及一个重要的超参数，损坏概率 p。此外，本章提出的 iDR 方法还引入了一个关键性超参数，适配正则项参数 λ。注意到 iDR 可以通过本章提出的迁移交叉验证策略 TCV 自动完成参数选择，但非敏感参数性能仍是本章方法能够成功应用于实际问题的重要保证。为此，本节在多领域情感数据集和垃圾邮件过滤数据集上执行系统性的参数敏感性实验，iDR 的平均分类准确率分别在这两个数据集的所有任务（12 个情感任务 6 个邮件任务）上计算得到。特别地，当要对特定参数进行测试时，保持其余参数不变而仅变化当前参数值。例如，当测试适配正则项参数 λ 的影响时，保持损坏概率为 $p = 0.6$ 在所有任务上不变。

详细实验结果如图 5.8(a) 和图 5.8(b) 所示，其中最佳基准方法的平均准确率用虚线表示。可以看到两个参数都呈现钟形性能曲线，且在相当大取值范围内使



(a) 特征维度 d　　(b) 样本数目 n

图 5.9　iDR 的可扩展性：(a) 特征维度 d (b) 样本数目 n。iDR 相对特征维度和样本数目均保持近似线性的计算复杂度，对大规模数据具有良好的可扩展性。

iDR 性能大幅超过最佳基准方法（情感任务上为 mSDA，邮件任务上为 CODA）。一般地，iDR 在中等偏大的损坏概率即 $p \in [0.4, 0.8]$ 时获得最佳性能，这与 mSDA、SDA[50] 是一致的。这预示着输入数据损坏程度既不能过小（使学到的特征变换退化为恒等变换平凡解）也不能过大（过度破坏输入信号保真度）。类似地，iDR 在中等大小的适配正则项参数即 $\lambda \in [0.1, 1000]$ 上效果最好，这与标准正则化学习器如 SVM 等的性能曲线一致。需要注意，λ 是跨越不同概率分布的适配正则项参数，因此不能通过辅助领域上的标准 k-折交叉验证选取最优值，否则构建的训练/验证集无法反映领域间的概率分布差异。本章提出迁移交叉验证 TCV 来解决该问题。

5.5.7　可扩展性

算法的线性可扩展性对大数据和深度学习都十分重要，因为深度学习非常依赖于大规模数据抽取紧致的高质量特征表示。本节实验给出 iDR 方法相对特征维度和样本数目的可扩展性测试，评测在多领域情感数据集和垃圾邮件过滤数据集手动扩展后的“大规模”数据集进行。具体地说，当测试 iDR 相对于高维特征的可扩展性时，固定每个数据集的样本数目并生成每个特征的多个副本直到达到待考察的特征维度。类似地，当测试 iDR 相对于大规模样本的可扩展性时，固定每个数据集的特征维度并生成每个样例的多个副本直到达到待考察的样本数目。图 5.9(a) 显示了特征维度在 5,000 至 320,000 间变化时的执行时间；图 5.9(b) 显示了样本数目在 5,000 至 160,000 间变化时的执行时间。可以看到，iDR 相对于特征维度和样本数目均保持近似线性的可扩展性，因而可以高效地分析处理大规模数据。

5.6　小结

本章提出了深度迁移学习框架，通过建立领域不变的深度特征表示来实现高效的跨领域知识迁移。本章优异的实验结果证明了在统一的深度架构中，联合抽取高度紧致特征表示并修正概率分布失配的重要性。换言之，无论标准深度学习（不考虑领域失配问题）还是浅层迁移学习（不抽取深度特征表示）都可能无法充分地解决富有挑战的领域拟合和分布适配任务，特别是解决欠拟合和欠适配问题。未来工作包括扩展到统一的迁移学习框架、深度学习方法、和概率分布度量准则。

第 6 章　总结与展望

6.1　本文总结

本文研究工作主要源自国家核高基科技重大专项“非结构化数据管理系统”（2010ZX01042-002-002）以及相关项目，目标是实现多种类型、多个领域的文本、图像、视频等非结构化数据的统一分析和挖掘，并实现标注、内容等数据价值信息的有效迁移和复用。为此，本文系统地研究了迁移学习的问题挑战及解决方法，旨在提高学习模型在不同领域或任务间的泛化性能，取得的主要创新性成果如下：

第 2 章提出了通用学习框架图正则化联合矩阵分解以及三种迁移学习模型，用于抽取领域间的公共隐含语义结构实现知识迁移、并通过保持领域内的几何结构反制负迁移，从而实现了跨领域分类学习任务并避免了负迁移问题。

第 3 章提出了通用学习框架联合适配正则化以及四种迁移学习模型，同时对结构风险泛函、边缘分布和条件分布适配、流形一致性等准则进行优化，从而在半监督学习的框架下完成了迁移学习模型训练。该框架的优势是考察了各种必要的学习准则，同时保持了模型的凸性以及简单可依赖优点。学习模型对跨领域条件概率分布间的异构性具有鲁棒性，从而解决了领域间条件分布的欠适配问题。

第 4 章提出了领域不变迁移核学习方法，通过可再生希尔伯特空间直接适配辅助领域和目标领域的概率分布，从而学习一个领域不变核矩阵。首次将领域间核矩阵的 Nyström 近似误差作为概率分布差异度量准则，克服了现有核空间矩匹配方法仅能匹配概率分布统计量的缺点，解决了边缘分布在领域间的欠适配问题。

第 5 章提出了深度迁移学习方法，通过建立领域不变的深度特征表示来实现有效的跨领域知识迁移。本文分析阐明了无论标准深度学习（不考虑领域失配问题）还是浅层迁移学习（不学习高度紧致特征表示）都无法充分地解决富有挑战性的领域迁移和分布适配问题；只有在统一深度架构中同步进行深度特征表示学习和非线性概率分布适配，才能同时解决跨领域概率分布的欠适配与欠拟合问题。

6.2　未来工作展望

本文系统地研究了迁移学习问题及方法，主要从概率分布适配和隐含特征学习两个技术层面展开，深入探讨了负迁移、欠适配、欠拟合等问题挑战。虽然在

理论、方法和应用方面取得了阶段性成果，但迁移学习作为机器学习的前沿方向之一，仍然存在很多挑战性问题有待进一步探索。具体可以概括为以下三方面：

1. 基于统计学习理论，分析迁移学习的推广误差下界。

在迁移学习中，训练数据（辅助领域）和测试数据（目标领域）不再服从独立同分布假设，这导致传统的统计学习理论、PAC 可学习理论等不能直接应用于分析迁移学习模型的泛化误差上界。现在已有少数工作研究了特定条件下的泛化误差上界，如假设领域间边缘分布可变、而条件分布不变[55]，或假设所有学习任务来自一致性的任务环境等，取得了部分前瞻性理论成果。然而，迁移学习的泛化误差下界作为一个重要的理论问题一直为整个学术界所忽略。所谓泛化误差下界是指，由于领域间概率分布差异的固有存在，且任何学习算法的分布适配能力都存在上限，这决定了迁移学习模型将具有泛化性能的“天花板”，难以达到独立同分布条件下的最优泛化性能，即存在泛化误差下界。因此，研究在领域间固有差异下学习算法所能达到的最好效果，有助于深入理解迁移学习对不同问题的有效性上界，在理论和实践上都具有十分重要的意义，是本文后续工作的重中之重。

2. 基于深度学习方法，设计和实现迁移学习高效算法。

近年来，深度学习在学术界和工业界都获得了广泛成功，引发了机器学习、计算机视觉、自然语言处理等领域的新一轮革命浪潮，被认为是进行大数据深度分析的最有潜力的技术之一。最近两年的研究表明，深度学习能够抽取紧致的、层次的、抽象的数据表示，具备在领域间迁移和复用的能力。因此，深度学习是实现迁移学习的重要技术之一，具有重要的研究价值。本文后续工作的重点之二，就是研究基于深度学习方法和迁移学习理论的新算法，提高迁移学习的实践效果。

3. 将迁移学习算法应用到非平稳环境大数据分析平台。

在非平稳环境下的大数据分析平台中，存在从多个数据源、多个时间段采集到的多模态数据，这些数据具有不同的标注程度且服从不同的概率分布，要对它们进行统一分析和管理，就需要借助迁移学习技术。本文提出的迁移学习算法有希望应用于大规模数据标注与分类问题，如人脸识别、对象跟踪等。有必要研究更高效的学习算法，以及如何在多核并行或分布式内存集群上进行并行模型训练。将迁移学习与大数据系统结合起来、更好地服务国计民生，是本文的长期目标。

参考文献

[1] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22:1345–1359.

[2] Vapnik V. Statistical Learning Theory. John Wiley, 1998.

[3] Valiant L. A theory of the learnable. Communications of the ACM, 1984, 27(11).

[4] Yuan L, Wang Y, Thompson P M, et al. Multi-Source Learning for Joint Analysis of Incomplete Multi-Modality Neuroimaging Data. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012.

[5] Margolis A. A Literature Review of Domain Adaptation with Unlabeled Data. Technical report, 2011.

[6] Quiñonero-Candela J, Sugiyama M, Schwaighofer A, et al. Dataset Shift in Machine Learning. The MIT Press, 2009.

[7] Wei B, Pal C. Cross lingual adaptation: an experiment on sentiment classifications. Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics, 2010.

[8] Prettenhofer P, Stein B. Cross-Language Text Classification using Structural Correspondence Learning. Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics, 2010.

[9] Ling X, Xue G R, Dai W, et al. Can Chinese web pages be classified with English data source? Proceedings of the 17th international conference on World Wide Web, 2008.

[10] Shi L, Mihalcea R, Tian M. Cross language text classification by model translation and Semi-Supervised learning. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.

[11] Platt J, Toutanova K, Yih W T. Translingual document representations from discriminative projections. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.

[12] Dai W, Chen Y, Xue G R, et al. Translated learning: Transfer learning across different feature spaces. Neural Information Processing Systems, 2008.

[13] Qi G J, Aggarwal C, Huang T. Towards semantic knowledge propagation from text corpus to web images. Proceedings of the 20th international conference on World wide web, 2011.

[14] Zhu Y, Chen Y, Lu Z, et al. Heterogeneous Transfer Learning for Image Classification. Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011.

[15] Li W, Duan L, Xu D, et al. Learning with Augmented Features for Supervised and Semi-supervised Heterogeneous Domain Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 99(PrePrints).

[16] Shi X, Fan W, Yang Q, et al. Relaxed transfer of different classes via spectral partition. Proceedings of the European Conference on Machine Learning and Knowledge Discoverty in Databases, 2009.

[17] Quadrianto N, Smola A J, Caetano T S, et al. Multitask learning without label correspondences. Neural Information Processing Systems, 2010.

[18] Qi G J, Aggarwal C, Rui Y, et al. Towards cross-category knowledge propagation for learning visual concepts. IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[19] Xiang E W, Pan S J, Pan W, et al. Source-free transfer learning. International Joint Conference on Artificial Intelligence, 2011.

[20] Fei-Fei L, Fergus R, Perona P. One-Shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4).

[21] Bickel S, Bruckner M, Scheffer T. Discriminative learning for differing training and test distributions. Proceedings of the 24th international conference on Machine learning, 2007.

[22] Jiang J, Zhai C. Instance weighting for domain adaptation in nlp. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007.

[23] Arnold A, Nallapati R, Cohen W W. A Comparative Study of Methods for Transductive Transfer Learning. Proceedings of the IEEE International Conference on Data Mining Workshop, 2007.

[24] Zhong E, Fan W, Peng J, et al. Cross Domain Distribution Adaptation via Kernel Mapping. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.

[25] Zadrozny B. Learning and evaluating classifiers under sample selection bias. Proceedings of the twenty-first international conference on Machine learning, 2004.

[26] Cortes C, Mohri M, Riley M, et al. Sample selection bias correction theory. Algorithmic Learning Theory, 2008.

[27] Huang J, Smola A J, Gretton A, et al. Correcting Sample Selection Bias by Unlabeled Data. Neural Information Processing Systems, 2006.

[28] Sugiyama M, Nakajima S, H Kashima P v B, et al. Direct importance estimation with model selection and its application to covariate shift adaptation. Neural Information Processing Systems, 2007.

[29] Zhang K, Zheng V W, Wang Q, et al. Covariate Shift in Hilbert Space: A Solution via Surrogate Kernels. Proceedings of the 30th International Conference on Machine Learning, 2013.

[30] Argyriou A, Evgeniou T. Multi-Task Feature Learning. Neural Information Processing Systems, 2006.

[31] Rosset S, Zhu J, Zou H, et al. A method for inferring label sampling mechanisms in semi-supervised learning. Neural Information Processing Systems, 2005.

[32] Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning. Proceedings of the 24th International Conference on Machine Learning, 2007.

[33] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel method for the two-sample problem. Neural Information Processing Systems, 2006.

[34] Cortes C, Mansour Y, Mohri M. Learning bounds for importance weighting. Neural Information Processing Systems, 2010.

[35] Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006.

[36] Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007.

[37] Ando R K, Zhang T, Bartlett P. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 2005, 6.

[38] Ben-David S, Blitzer J, Crammer K, et al. Analysis of Representations for Domain Adaptation. Advances in Neural Information Processing Systems, 2006.

[39] Daumé III H. Frustratingly Easy Domain Adaptation. Proceedings of the Annual Meeting of Association for Computational Linguistics, 2007.

[40] Pan S J, Kwok J T, Yang Q. Transfer Learning via Dimensionality Reduction. Proceedings of the 22nd AAAI Conference on Artificial Intelligence, 2008.

[41] Satpal S, Sarawagi S. Domain Adaptation of Conditional Probability Models Via Feature Subsetting. Proceedings of the European Conference on Machine Learning and Knowledge Discoverty in Databases, 2007.

[42] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment. Proceedings of the 19th International Conference on World Wide Web, 2010.

[43] Chen B, Lam W, Tsang I, et al. Extracting Discriminative Concepts for Domain Adaptation in Text Mining. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.

[44] Raina R, Battle A, Lee H, et al. Self-taught learning: transfer learning from unlabeled data. Proceedings of the 24th international conference on Machine learning, 2007.

[45] Pan S J, Tsang I W, Kwok J T, et al. Domain Adaptation via Transfer Component Analysis. IEEE Transactions on Neural Networks, 2011, 22(2):199–210.

[46] Si S, Tao D, Geng B. Bregman Divergence-Based Regularization for Transfer Subspace Learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(7).

[47] Chen B, Lam W, Tsang I W, et al. Discovering Low-Rank Shared Concept Space for Adapting Text Mining Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(6).

[48] Huang F, Yates A. Distributional representations for handling sparsity in supervised Sequence-Labeling. Proceedings of the 47th Annual Meeting of the Association of Computational Linguistics, 2009.

[49] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798–1828.

[50] Glorot X, Bordes A, Bengio Y. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. Proceedings of the 28th International Conference on Machine Learning, 2011.

[51] Ngiam J, Khosla A, Kim M, et al. Multimodal Deep Learning. Proceedings of the 28th International Conference on Machine Learning, 2011.

[52] Chen M, Xu Z E, Weinberger K Q, et al. Marginalized Denoising Autoencoders for Domain Adaptation. Proceedings of the 29th International Conference on Machine Learning, 2012.

[53] Ge L, Gao J, Li X, et al. Multi-Source Deep Learning for Information Trustworthiness Estimation. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013.

[54] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief net. Neural Computation, 2006, 18:1527–1554.

[55] Blitzer J, Crammer K, Kulesza A, et al. Learning Bounds for Domain Adaptation. Neural Information Processing Systems, 2008.

[56] Kifer D, Ben-David S, Gehrke J. Detecting change in data streams. Proceedings of the Thirtieth international conference on Very large data bases, 2004.

[57] Cao B, Pan S J, Zhang Y, et al. Adaptive Transfer Learning. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.

[58] Long M, Wang J, Ding G, et al. Transfer Learning with Graph Co-Regularization. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7).

[59] Long M, Wang J, Ding G, et al. Transfer Learning with Graph Co-Regularization. Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012.

[60] Long M, Wang J, Ding G, et al. Adaptation Regularization: A General Framework for Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(5).

[61] Long M, Wang J, Ding G, et al. Transfer Feature Learning with Joint Distribution Adaptation. IEEE International Conference on Computer Vision, 2013.

[62] Long M, Wang J, Ding G, et al. Transfer Joint Matching for Unsupervised Domain Adaptation. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[63] Zhuang F, Luo P, Shen Z, et al. Mining Distinction and Commonality across Multiple Domains using Generative Model for Text Classification. IEEE Transactions on Knowledge and Data Engineering, 2011, 24(11).

[64] Li L, Zhou K, Xue G R, et al. Video Summarization via Transferrable Structured Learning. Proceedings of International Conference on World Wide Web, 2011.

[65] Li B, Yang Q, Xue X. Transfer Learning for Collaborative Filtering via a Rating-Matrix Generative Model. Proceedings of the 26th International Conference on Machine Learning, 2009.

[66] Zhu X, Lafferty J. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 2005.

[67] Dai W, Xue G R, Yang Q, et al. Co-clustering based classification for out-of-domain documents. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.

[68] Wang Z, Song Y, Zhang C. Knowledge Transfer on Hybrid Graph. Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009.

[69] Zhuang F, Luo P, Xiong H, et al. Exploiting associations between word clusters and document classes for cross-domain text categorization. Proceedings of the 10th SIAM International Conference on Data Mining, 2010.

[70] Wang H, Huang H, Nie F, et al. Cross-Language Web Page Classification via Dual Knowledge Transfer Using Nonnegative Matrix Tri-Factorization. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011.

[71] Long M, Wang J, Ding G, et al. Dual Transfer Learning. Proceedings of the 12th SIAM International Conference on Data Mining, 2012.

[72] Cai D, He X, Han J, et al. Graph Regularized Nonnegative Matrix Factorization for Data Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8).

[73] Ding C, Li T, Peng W. Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence, Chi-square Statistic, and a Hybrid Method. Proceedings of the 21st AAAI Conference on Artificial Intelligence, 2006.

[74] Lee D D, Seung H S. Algorithms for Non-negative Matrix Factorization. Neural Information Processing Systems, 2000.

[75] Ding C, Li T, Jordan M I. Convex and Semi-Nonnegative Matrix Factorizations. IEEE Transactions on Pattern Analysis Machine Intelligence, 2010, 32(1).

[76] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix tri-factorizations for clustering. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

[77] Singh A P, Gordon G J. Relational Learning via Collective Matrix Factorization. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[78] Xue G R, Dai W, Yang Q, et al. Topic-bridged PLSA for cross-domain text classification. Proceedings of the 31st ACM SIGIR conference on Research and development in information retrieval, 2008.

[79] Li T, Sindhwani V, Ding C, et al. Bridging Domains with Words: Opinion Analysis with Matrix Tri-factorizations. Proceedings of the 10th SIAM International Conference on Data Mining, 2010.

[80] Cai D, He X, Wang X, et al. Locality Preserving Nonnegative Matrix Factorization. Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009.

[81] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. Neural Information Processing Systems, 2001.

[82] Gu Q, Zhou J. Co-Clustering on Manifolds. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.

[83] Gu Q, Ding C, Han J. On Trivial Solution and Scale Transfer Problems in Graph Regularized NMF. Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011.

[84] Boyd S, Vandenberghe L. Convex Optimization. Cambridge University Press, 2004.

[85] Duan L, Tsang I W, Xu D. Domain Transfer Multiple Kernel Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3):465–479.

[86] Cai D, He X, Han J. Spectral Regression: A Unified Approach for Sparse Subspace Learning. Proceedings of the IEEE International Conference on Data Mining, 2007.

[87] Vedaldi A, Fulkerson B. VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/, 2008.

[88] Belkin M, Niyogi P, Sindhwani V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. Journal of Machine Learning Research, 2006, 7:2399–2434.

[89] Yang J, Yan R, Hauptmann A G. Cross-domain video concept detection using adaptive svms. Proceedings of the 15th international conference on Multimedia, 2007.

[90] Quanz B, Huan J. Large margin transductive transfer learning. Proceedings of the 18th ACM conference on Information and knowledge management, 2009.

[91] Tao J, Chung F L, Wang S. On minimum distribution discrepancy support vector machine for domain adaptation. Pattern Recognition, 2012, 45(11).

[92] Bruzzone L, Marconcini M. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(5).

[93] Bahadori M T, Liu Y, Zhang D. Learning with Minimum Supervision: A General Framework for Transductive Transfer Learning. Proceedings of the 11th IEEE International Conference on Data Mining, 2011.

[94] Xiao M, Guo Y. Semi-Supervised Kernel Matching for Domain Adaptation. Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012.

[95] Sun Q, Chattopadhyay R, Panchanathan S, et al. A Two-Stage Weighting Framework for Multi-Source Domain Adaptation. Neural Information Processing Systems, 2011.

[96] Chen M, Weinberger K Q, Blitzer J C. Co-Training for Domain Adaptation. Neural Information Processing Systems, 2011.

[97] Quanz B, Huan J, Mishra M. Knowledge transfer with low-quality data: A feature extraction issue. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10).

[98] Schölkopf B, Herbrich R, Smola A J. A Generalized Representer Theorem. Proceedings of the 14th Annual Conference on Computational Learning Theory, 2001.

[99] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[100] Ling X, Dai W, Xue G R, et al. Spectral domain-transfer learning. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[101] Wang C, Mahadevan S. Heterogeneous Domain Adaptation using Manifold Alignment. Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011.

[102] Shi X, Liu Q, Fan W, et al. Transfer across Completely Different Feature Spaces via Spectral Embedding. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(4).

[103] Liu Q, Liao X, Carin L. Semi-Supervised Multitask Learning. Advances in Neural Information Processing Systems, 2007.

[104] Johnson R, Zhang T. Graph-Based Semi-Supervised Learning and Spectral Kernel Design. IEEE Transactions on Information Theory, 2008, 54(1).

[105] Baktashmotlagh M, Harandi M T, Lovell B C, et al. Unsupervised Domain Adaptation by Domain Invariant Projection. IEEE International Conference on Computer Vision, 2013.

[106] Edelman A, Arias T, Smith S. The geometry of algorithms with orthogonality constraints. SIAM, 1998, 20(2).

[107] Saenko K, Kulis B, Fritz M, et al. Adapting Visual Category Models to New Domains. European Conference on Computer Vision, 2010.

[108] Gong B, Shi Y, Sha F, et al. Geodesic Flow Kernel for Unsupervised Domain Adaptation. IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[109] Duan L, Xu D, Tsang I W H, et al. Visual Event Recognition in Videos by Learning from Web Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(9).

[110] Perronnin F, Sánchez J, Liu Y. Large-scale image categorization with explicit data embedding. IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[111] Duan L, Tsang I, Xu D, et al. Domain Transfer SVM for Video Concept Detection. IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[112] Cao X, Wipf D, Wen F, et al. A Practical Transfer Learning Algorithm for Face Verification. IEEE International Conference on Computer Vision, 2013.

[113] Zhang K, Tsang I W, Kwok J T. Improved Nyström Low-Rank Approximation and Error Analysis. Proceedings of the 25th International Conference on Machine Learning, 2008.

[114] Schölkopf B, Smola A J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, 2001.

[115] Williams C, Seeger M. Using the Nyström method to speed up kernel machines. Neural Information Processing Systems, 2001.

[116] Zhu X, Kandola J, Ghahramani Z, et al. Nonparametric Transforms of Graph Kernels for Semi-Supervised Learning. Neural Information Processing Systems, 2004.

[117] Zhuang J, Tsang I W, Hoi S C. A Family of Simple Non-Parametric Kernel Learning Algorithms. Journal of Machine Learning Research, 2011, 12:1313–1347.

[118] Zhang T, Ando R K. Analysis of spectral kernel design based semi-supervised learning. Neural Information Processing Systems, 2006.

[119] Rakotomamonjy A, Bach F R, Canu S, et al. SimpleMKL. Journal of Machine Learning Research, 2008, 9:2491–2521.

[120] Hoi S C H, Lyu M R, Chang E Y. Learning the Unified Kernel Machines for Classification. Proceedings of the 12nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

[121] Mao Q, Tsang I W. Parameter-Free Spectral Kernel Learning. Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, 2010.

[122] Jin R, Yang T, Mahdavi M, et al. Improved Bounds for the Nyström Method with Application to Kernel Classification. IEEE Transactions on Information Theory, 2013, 59(10).

[123] Gao J, Fan W, Jiang J, et al. Knowledge transfer via multiple model local structure mapping. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[124] Loui A, Luo J, Chang S F, et al. Kodak: Consumer Video Benchmark Data Set: Concept Definition and Annotation. Proceedings of the International Workshop on Multimedia Information Retrieval, 2007.

[125] Vincent P, Larochelle H, Lajoie I, et al. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. Journal of Machine Learning Research, 2010, 11:3371–3408.

[126] Dudley R M. Real analysis and probability. Cambridge University Press, 2002.

[127] Sugiyama M, Krauledat M, Müller K R. Covariate Shift Adaptation by Importance Weighted Cross Validation. Journal of Machine Learning Research, 2007, 8:985–1005.

[128] Zhong E, Fan W, Yang Q, et al. Cross Validation Framework to Choose Amongst Models and Datasets for Transfer Learning. Proceedings of the European Conference on Machine Learning and Knowledge Discoverty in Databases, 2010.

致 谢

转身回眸之间，我在清华园已经度过了整整十年的青春岁月，博士生涯临近尾声之际，我要诚挚地感谢我的导师王建民教授。这篇学位论文的每一个环节，从选题开题、资料搜集、方法论证、写作技巧到最终的成文定稿，王老师都倾注了极大的心血与劳动。几年的时光中，王老师既是我学业的导师，又是我人生的导师，无论是在学业上、生活上还是在未来的职业规划上，都给予了我悉心的指导和关怀，为我提供了优越的科研学习环境。同时他谦和的人格修养、渊博的专业知识、严谨的治学态度以及高效的科研管理风格，值得我终身学习和铭记。师恩重如高山，我只有在今后的工作中，加倍努力，用优异的成绩报答师恩。

感谢孙家广院士对我殷切的鼓励和悉心的教导，使我树立了超越自我的坚定信念。在与孙院士的每一次交流中，我都不断的审视自己的科研学习状态，反思自己的不足，并在审视和反思之中获得进步，这一切都令我如沐春风，受益匪浅。

感谢伊利诺伊大学芝加哥分校 Philip S. Yu 教授，作为数据库和数据挖掘领域的国际权威学者，他针对我的研究方向给出很多建设性意见，帮助我更好的理清了研究思路，通过和俞教授的合作，我更加顺利地完成了本文的第三至第五章的工作。感谢香港科技大学 Qiang Yang 教授，作为迁移学习方向的国际权威学者，他引导我更好地理解了该研究前沿方向中的关键性问题，帮助我做到有的放矢。感谢加州大学洛杉矶分校 Wei Wang 教授在论文合作中给予我的指导和帮助。

在大数据与知识工程实验室，从热烈的学术讨论，到忙碌的项目开发，伴随着我这些年来的成长，记录着我每一次的进步。实验室的丁贵广、叶晓俊、张力、闻立杰、王朝坤、宋韶旭等老师对我论文工作给予了热忱的关心和指导，黄向东、卓安、衣国垒、石江枫、向武、曹越、刘璋、余志伟等同学在项目研发上给予了我很多的支持和帮助。在此，对这些良师益友一并致以我深深的感谢。

感谢答辩评审会的各位老师对我的学位论文提出了宝贵的意见和建议，他们是伊利诺伊大学芝加哥分校 Philip S. Yu 教授，中国人民大学杜小勇教授，北京大学苏开乐教授，清华大学计算机系胡事民教授，清华大学软件学院孙家广院士、顾明教授。谨向在百忙中抽出宝贵时间评审本文的专家、学者致以由衷的谢意。

最后，衷心地感谢我的妻子田琳及家人，博士生涯漫长而又艰辛，你们对我的爱、理解和支持是我不断前进的动力源泉。

本文的研究工作受到国家核高基重大专项、国家 863 计划项目、国家 973 计划项目和国家自然科学基金项目的资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1985 年 11 月 20 日出生于广西壮族自治区河池市。

2004 年 9 月考入清华大学电机工程与应用电子技术系电气工程及其自动化专业，2008 年 7 月本科毕业并获得工学学士学位。

2008 年 9 月免试进入清华大学计算机科学与技术系攻读博士学位至今。

发表的学术论文

[1] **Mingsheng Long**, Yue Cao, Jianmin Wang, Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks. *International Conference on Machine Learning (**ICML**)*, 2015. (中国计算机学会推荐 *A* 类会议)

[2] **Mingsheng Long**, Jianmin Wang, Jiaguang Sun, and Philip S. Yu. Domain Invariant Transfer Kernel Learning. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 27(6):1519-1532, 2015. (中国计算机学会推荐 *A* 类期刊)

[3] **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and Philip S. Yu. Adaptation Regularization: A General Framework for Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 26(5):1076-1089, 2014. (*SCI* 检索号*: 14299513,* 中国计算机学会推荐 *A* 类期刊)

[4] **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer Learning with Graph Co-Regularization. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 26(7):1805-1818, 2014. (*SCI* 检索号*: 14430037,* 中国计算机学会推荐 *A* 类期刊)

[5] **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer Joint Matching for Unsupervised Domain Adaptation. *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**)*, 2014. (中国计算机学会推荐 *A* 类会议)

[6] **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer Feature Learning with Joint Distribution Adaptation. *Proceedings of the 14th IEEE International Conference on Computer Vision (**ICCV**)*, 2013. (*EI* 检索号*: 20141717632010,* 中国计算机学会推荐 *A* 类会议)

[7] **Mingsheng Long**, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, and Philip S. Yu. Transfer Sparse Coding for Robust Image Representation. *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**)*, 2013. (*EI 检索号: 20134616982520,* 中国计算机学会推荐 *A* 类会议)

[8] **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer Learning with Graph Co-Regularization. *Proceedings of the 26th AAAI Conference on Artificial Intelligence (**AAAI**)*, 2012. (*EI 检索号: 20124515646446,* 中国计算机学会推荐 *A* 类会议)

[9] **Mingsheng Long**, Jianmin Wang, Guiguang Ding, Wei Cheng, Xiang Zhang, and Wei Wang. Dual Transfer Learning. *Proceedings of the 12th SIAM International Conference on Data Mining (**SDM**)*, 2012. (*EI 检索号: 20133016523090,* 中国计算机学会推荐 *B* 类会议) (**最佳论文奖提名**)

[10] **Mingsheng Long**, Wei Cheng, Xiaoming Jin, Jianmin Wang, and Dou Shen. Transfer Learning via Cluster Correspondence Inference. *Proceedings of the 10th IEEE International Conference on Data Mining (**ICDM**)*, 2010. (*EI 检索号: 20110813686985,* 中国计算机学会推荐 *B* 类会议)

[11] Yue Cao, **Mingsheng Long***, Jianmin Wang, Han Zhu, Qingfu Wen. Deep Quantization Network for Efficient Image Retrieval. *AAAI Conference on Artificial Intelligence (**AAAI**)*, 2015. (通讯作者, 中国计算机学会推荐 *A* 类会议)

[12] Han Zhu, **Mingsheng Long***, Jianmin Wang, Yue Cao. Deep Hashing Network for Efficient Similarity Retrieval. *AAAI Conference on Artificial Intelligence (**AAAI**)*, 2015. (通讯作者, 中国计算机学会推荐 *A* 类会议)

[13] Wu Xiang, Jianmin Wang, and **Mingsheng Long**. Local Hybrid Coding for Image Classification. *Proceedings of the 22nd International Conference on Pattern Recognition (**ICPR**)*, 2014.

[14] Xiangdong Huang, Jianmin Wang, Jian Bai, Guiguang Ding, and **Mingsheng Long**. Inherent Replica Inconsistency in Cassandra. *Proceedings of the 3rd International Congress on Big Data (**BigData**)*, 2014.

[15] Jiangfeng Shi, **Mingsheng Long**, Qiang Liu, Guiguang Ding, and Jianmin Wang. Twin Bridge Transfer Learning for Sparse Collaborative Filtering. *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (**PAKDD**)*, 2013.

[16] Lianghao Li, Xiaoming Jin, and **Mingsheng Long**. Topic Correlation Analysis for Cross-Domain Text Classification. *Proceedings of the 26th AAAI Conference on*

*Artificial Intelligence (**AAAI**)*, 2012. (*EI 检索号: 20124515646441,* 中国计算机学会推荐 *A* 类会议)

待发表的学术论文

[1] **Mingsheng Long**, Jianmin Wang, Yue Cao, Jiaguang Sun, and Philip S. Yu. Deep Learning of Transferable Representation for Safe Domain Adaptation. *IEEE Transactions on Knowledge and Data Engineering (**TKDE**)*, 2016. (小改后接收) (中国计算机学会推荐 *A* 类期刊)

[2] **Mingsheng Long**, Han Zhu, Jianmin Wang, Zhangjie Cao, Michael I. Jordan. Learning Transferable Visual Features with Very Deep Adaptation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**)*, 2016. (会议论文扩展，在审) (中国计算机学会推荐 *A* 类期刊)

[3] **Mingsheng Long**, Jianmin Wang. Learning Multiple Tasks with Deep Relationship Networks. arXiv preprint arXiv:1506.02117.

[4] **Mingsheng Long**, Jianmin Wang, Philip S. Yu. Compositional Correlation Quantization for Large-Scale Multimodal Search. arXiv preprint arXiv:1504.04818.

已授权国家发明专利

[1] 王建民, 丁贵广, 龙明盛, 姜晓伟。一种用于推荐系统的计算机数据挖掘方法。专利号：CN2012101932292。

主持或参与的科研项目

[1] 国家自然科学基金青年科学基金项目, 面向大数据的安全迁移学习方法, 项目编号: 61502265, 2016.01.01–2018.12.31。(主持)

[2] 中国博士后基金特别资助, 可扩展迁移学习理论与方法研究, 项目编号: 2015T80088, 2014.07.15–2016.07.15。(主持)

[3] 清华信息科学与技术国家实验室大数据科学与技术专项, 面向领域的大数据应用系统开发与运行平台, 2014.07.01–2016.06.30。(技术负责人)

[4] 国家杰出青年科学基金, 大规模过程数据管理与挖掘, 项目编号: 613250154, 2014.01.01–2017.12.31。(参与)

[5] **国家核高基科技重大专项**, 非结构化数据管理系统, 项目编号: 2010ZX01042-002-002, 2010.01.01–2013.06.30。(骨干)

[6] 国家发改委高技术服务业研发及产业化专项, 面向医疗卫生行业的大数据分

析平台与服务创新, 2012.01.01–2014.12.31。(参与)

[7] 国家自然科学基金项目, 非结构化数据管理若干关键技术研究, 项目编号: 61073005, 2011.01.01–2013.06.30。(骨干)

[8] 国家自然科学基金项目, 自然科学基金项目辅助评审关键技术与系统研发, 项目编号: 61050010, 2010.07.01–2011.06.30。(参与)

攻读博士学位期间的获奖情况

[1] 清华大学优秀博士学位论文，清华大学，2014 年。

[2] 清华大学优秀博士毕业生，清华大学，2014 年。

[3] 北京市优秀博士毕业生，北京市，2014 年。

[4] 软件学院学术新秀，清华大学，2014 年。

[5] 钟士模奖学金，清华大学，2013 年。

[6] 英特尔学者，英特尔公司 & 清华大学，2012 年。

[7] **最佳论文奖提名**，SDM 数据挖掘国际会议，2012 年。