分类号 TP3 密级

UDC 编号

中国科学院研究生院
博士学位论文

迁移学习中文本分类算法研究

庄 福 振

指导教师 何 清 研究员

中国科学院计算技术研究所

申请学位级别 工学博士 学科专业名称 计算机软件与理论

论文提交日期 2011 年 4 月 论文答辩日期 2011 年 5 月

培养单位 中国科学院计算技术研究所

学位授予单位 中国科学院研究生院

答辩委员会主席

Research on Text Classification Algorithms in Transfer Learning

by

Zhuang Fuzhen

Dissertation submitted to
Graduate School of the Chinese Academy of Sciences
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Dissertation Supervisor: Prof. He Qing

Institute of Computing Technology

Chinese Academy of Sciences

April 2011

声　　明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：　　日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：　　导师签名：　　日期：

本文受

1、 国家自然科学基金面上项目“分布式计算环境下的并行数据挖掘算法与理论研究”(项目编号：No. 60975039)

2、 国家自然科学基金重点项目“WEB 搜索与挖掘的新理论与方法”(项目编号：No. 60933004)

3、 国家 973 项目子课题“非结构化(图像)信息的内容理解与语义表征”(项目编号：No.2007CB311004)

4、 国家自然科学基金面上项目“基于超曲面的覆盖分类算法与理论研究”(项目编号：No. 60675010)

5、 北京市自然科学基金“海量高维、多类数据分类法研究及其应用”(项目编号：No. 4052025)

资助

摘 要

在传统机器学习中，为了保证训练得到的分类模型具有高准确性和可靠性，都有两个基本的假设：(1) 用于学习的训练样本与新的测试样本满足独立同分布条件；(2) 必须有足够可利用的训练样本才能学习得到一个好的分类模型。但是，在实际应用中我们发现这两个条件往往无法满足。首先，随着时间的推移，原先可利用的有标签样本数据可能变得不可用，与新来的测试样本的分布产生语义、分布上的缺口。另外，有标签的样本数据往往很缺乏，很难获得而且人工标记费时耗力。为了解决这两个问题，迁移学习研究成为近年来十分重要和具有挑战性的课题。迁移学习是运用已有的知识对不同但相关领域问题进行求解的新的一种机器学习方法。它放宽了传统机器学习中的两个基本假设，目的是迁移已有的知识来解决目标领域中仅有少量有标签样本数据甚至没有的学习问题。

本文对迁移学习中分类算法进行了研究，由浅入深，由简单到复杂，提出了几种迁移学习分类算法。首先研究了从单个源领域(也叫训练集)到单个目标领域(也叫测试集)的学习算法，然后对多个源领域到单个目标领域的算法进行了研究，继而到更加一般的情况，研究同时处理多个源领域与多个目标领域的学习算法。最后，对本文提出的几个多源跨领域学习算法进行比较。取得的主要成果如下：

1. 提出基于混合正则化的无标签领域归纳迁移学习方法。首先，分析了直推式迁移学习(Transductive Transfer Learning)方法桥接精化(Bridged Refinement)中存在的类别比例漂移问题，然后提出归一化的方法使得预测的类别比例接近于实际样本类别比例。更进一步，提出了一种基于混合正则化框架的归纳迁移学习算法。其中包括目标领域分布结构的流形正则化，预测概率的熵正则化，以及类别比例的期望正则化。实验结果表明，1) 加入类别先验可以提高桥接精化算法的分类性能；2) 本文提出的归纳迁移学习算法优越于其他现有常用算法，同时最后得到的模型可以直接对新来的目标数据进行预测。

2. 提出一种有效挖掘词特征聚类与文档类别关联关系的迁移学习算法。虽然文本分类问题中源领域与目标领域数据在原始词特征上分布不一样，但不同数据领域可能共享词特征聚类与文档类别之间的关联关系。非负矩阵分解算法已经广泛运用于文本分类，聚类，模式识别等机器学习领域，而且非负矩阵分解算法可以很好的利用该关联关系，因此本文把非负矩阵分解算法引入到迁移学习领域，提出一种有效挖掘词特征聚类与文档类别关联关系的联合优化框架。为了求解该优化问题，设计了一个迭代算法并从理论上分析了该迭代算法的收敛性。大量实验表明，所提出的算法可以有效解决迁移学习分类问题，并且在知识迁移比较困难的情况下表现更加优异。

3. 提出基于一致性正则化的多源跨领域学习框架。在实际应用中，有标签样本往往来自于多个源领域，而且多个源领域之间分布不同但语义相关。因此，如何开发利用多

个源领域之间的分布差异性来进行知识迁移，使得在目标领域数据上的分类性能尽可能好？这是一个比从单个源领域学习更有挑战性的研究问题。本文提出了基于一致性正则化从多源领域到目标领域的跨领域分类学习框架。在这个框架下，局部的子分类器不仅考虑在源领域上可利用的局部数据，而且考虑了这些由源领域标签数据得到的子分类器在目标领域上的预测的一致性。更进一步，从理论上分析了一致性正则化框架的有效性。最后，为了处理各个源领域数据在地理上分布的情况，提出一致性正则化框架的分布式实现，可避免收集各个领域数据到中心节点，而只是传递一些统计变量，这在一定程度上减轻了数据信息的隐私性担忧。

4. 研究基于生成模型的挖掘多领域之间共性与特性的跨领域分类方法。该项工作对有效挖掘词特征聚类与文档类别关联关系进行了深入研究。基于非负矩阵的迁移学习方法缺乏完整的概率解释，而且很难用到多个源领域和多个目标领域数据。本文进一步提出基于生成模型的有效挖掘多领域之间共性与特性的跨领域分类方法。共性是指与领域数据独立的词特征聚类与文档类别之间的关联关系，而特性是指不同的领域数据用不同的关键词特征来表示同一词概念。因此可以有效开发利用不同领域之间的共性作为知识迁移的桥梁。为了求解该算法，设计了一个 EM 算法。实验结果表明该算法可以同时处理多个源领域和多个目标领域，而且可以有效解决分布不同性程度较高的迁移学习问题。

5. 对多源跨领域学习算法进行比较。该项工作首先扩展基于非负矩阵分解的跨领域学习方法到处理多源领域的情况，然后对基于生成模型跨领域学习算法进行改进。最后进行了系统的实验比较。实验表明，该算法优于传统的监督学习算法，也比以往的跨领域学习方法优越，而且能够处理迁移学习比较难的分类问题，具有较强的迁移学习能力。

关键词：跨领域分类学习；迁移学习；混合正则化；非负矩阵分解；一致性正则化；主题模型；概率潜在语义分析；EM 算法；半监督学习

Research on Text Classification Algorithms in Transfer Learning

Zhuang Fuzhen (Computer Software & Theory)
Directed By He Qing

Abstract

To ensure that classification algorithms can obtain high accuracy and reliability in traditional machine learning, two basic assumptions must be supposed: (1) the training (also referred as source domain) and test data (also referred target domain) follow the *independent* and *identical distributed* (*i.i.d*.) condition; (2) there are enough labeled samples to learn a good classification model. However, these two assumptions always do not hold in real-world applications due to the following two facts. First, the new test data coming from fast evolving information sources usually generate distribution gap, which causes the unavailability of existing labeled data. Second, it is always very hard to obtain the labeled data, and also the manual labeling is expensive. To solve these two problems, transfer learning[1] becomes an important and challenging research topic in recent years. Transfer learning is a new machine learning method that applies the knowledge from related but different domains to target domains. It relaxes the two basic assumptions in traditional machine learning, and aims to solve the problems that there are few or even not any labeled data in target domains.

This dissertation focuses on the research of text classification algorithms in transfer learning. We proceed with the study deeper and deeper along the line from simple problems to hard ones, and propose several classification algorithms for transfer learning. First, the classification problems from one single source domain to one single target domain are studied. Then, the ones with multiple source domains are investigated. Third, the classification algorithm, which can simultaneously handle multiple source domains and multiple target domains, is proposed. Finally, we make a systemic comparison of the cross-domain text classification algorithms learning from multiple source domains. The main contributions of this dissertation are summarized as follows:

1. An inductive transfer learning algorithm for unlabeled target domain via hybrid regularization is proposed. Firstly, the problem of class ratio drift in the previous work bridged refinement of transductive transfer learning is discussed, and then a normalization method to move towards the desired class ratio is proposed. Furthermore, a hybrid regularization framework for inductive transfer learning is proposed. It considers three factors, including the geometry distribution of the target domain by manifold regularization, the entropy value of prediction probability by entropy regularization, and the class prior by expectation

[1] In text classification or natural language processing, transfer learning is also referred as cross-domain learning or domain adaptation.

regularization. Experimental results show that: (1) the classification performance of bridged refinement is improved when incorporating the class prior; (2) the proposed inductive method outperms all compared approaches and the output model can directly deal with unseen test points.

2. A cross-domain text categorization algorithm is proposed to effectively exploit the association between word clusters and document classes. Though the distributions of source and target domains are different in raw word features, the associations between word clusters and document classes may remain stable across different domains. The non-negative matrix factorization algorithm is widely used in text classification, text clustering, and pattern recognition and so on. Since the associations can be easily incorporated in non-negative matrix factorization, thus a joint optimization framework is formulated based on non-negative matrix tri-factorization to exploit the stable associations. Furthermore, an iterative algorithm is given to solve this optimization and is theoretically proved to be converged. Extensive experiments validate the effectiveness of our algorithm on cross-domain classification problems. In particular, the proposed method can take advantage of handling more difficult transfer learning scenarios.

3. A consensus regularization framework for learning from multiple source domains is proposed. In real-world applications, the labeled data are always from multiple information sources, and these domains can be semantically related but have different distributions. How to exploit the distribution differences among multiple source domains to boost the learning performance in a target domain is a challenging research topic. In this paper, a consensus regularization framework is proposed, in which a local classifier is trained by considering both local data available in one source domain and the prediction consensus with the classifiers learned from other source domains. Moreover, a theoretical analysis is provided as well as an empirical study of the proposed consensus regularization framework. Finally, to deal with the situation that the multiple source domains are geographically distributed, the distributed version of the proposed algorithm is also developed, which avoids the need to upload all the data to a centralized location and helps to mitigate privacy concerns.

4. To deal with the transfer learning scenario with multiple source domains and multiple target domains, a cross-domain algorithm based on generative model is proposed to mine the distinction and commonality among multiple domains. This work is further to provide the probabilistic explanation for the previous work non-negative matrix tri-factorization based method, and tackle the classification problems with multiple source domains and multiple target domains. The two issues of different domains using different key words to express the same concept and different domains sharing the same associations between word clusters and document classes are actually the distinction and commonality. An EM algorithm is developed

to solve the proposed model, and also the experiments show its effectiveness to handle multiple source and multiple target domains.

5. The cross-domain text classification algorithms learning from multiple source domains are compared. First, the non-negative tri-factorization matrix based cross-domain method is further extended to deal with multiple source domains. Then, the generative model based cross-domain approach is improved, in which a more reasonable assumption is captured. Finally, these cross-domain algorithms, dealing with multiple source domains, are expirically compared. Experimental results show that the proposed methods are better than the supervised-learning algorithm and previous cross-domain approaches. Also they are more tolerant of distribution differences, which indicate that they can handle much more difficult transfer learning problems.

Keywords: Cross-domain Classification Learning; Transfer Learning; Hybrid Regularization; Non-negative Matrix Factorization; Consensus Regularization; Topic Modeling; Probabilistic Latent Semantic Analysis; EM Algorithms; Semi-supervised Learning

目　录

图目录

表目录

第一章　绪论

1.1. 研究背景与意义

随着社会发展的信息化和网络化，人们在日常生活和工作中无时无刻不在获取信息，分析信息，并以此作为决策的依据。在一定程度上，信息的拥有量已经成为决定和制约人类社会发展的重要因素。

近年来，互联网正处于快速发展的时期。根据统计，互联网上在线发布的网页数量高达亿级，并以每天百万页的速度增长。其中包含的内容极为丰富，几乎囊括了人类社会从政治、经济、军事到生活、娱乐、体育的各个方面，信息量极为丰富且完全开放。从发展趋势来看，互联网将成为人们获取信息的主要来源。可惜互联网并不是组织严密、条理清晰的数字信息库，而是一个杂乱无章的信息仓库。人们要想从其中获取自己感兴趣的信息已经变得越来越困难，往往花费很多时间却所获寥寥。

如何从互联网中获取信息？想要高效准确地寻找到所需的信息，信息分类是必不可少的第一步。通过分类，信息可以得到有效的组织管理，有利于快速准确的定位信息。分类学习问题，是机器学习中一种重要的学习方法，目前已经得到广泛的研究与发展。它根据带有标签的数据样本(也称为“源领域数据”或者“训练样本”)训练分类模型，然后运用分类模型对新数据样本(也称为“目标领域数据”)进行预测类标。近年来，各种分类学习算法[Cover, 1967; Friedman, 1996; Breiman, 1984; Quinlan, 1993; Joachims, 1999; Joachims, 1999a; He, 2002; He, 2003; Joachims, 2003; Hastie, 2001; Vapnik, 1998; He, 2008]不断被提出，并被广泛地应用到信息分类中，使得人们可以高效准确地定位所需要的信息。

在传统分类学习中，为了保证训练得到的分类模型具有准确性和高可靠性，都有两个基本的假设：(1) 用于学习的训练样本与新的测试样本满足独立同分布的条件；(2) 必须有足够可利用的训练样本才能学习得到一个好的分类模型。但是，在实际应用中我们发现这两个条件往往无法满足。首先，随着时间的推移，原先可利用的有标签的样本数据可能变得不可用，与新来的测试样本的分布产生语义、分布上的缺口。比如，股票数据就是很有时效性的数据，利用上月份的训练样本学习得到的模型并不能很好的预测本月份的新样本。另外，有标签的样本数据往往很缺乏，而且很难获得。在 Web 数据挖掘领域，新数据不断涌现，已有的训练样本已经不足以训练得到一个可靠的分类模型，而标注大量的样本又非常地费时费力，而且由于人的主观因素容易出错。这就引起了机器学习中另外一个重要问题，如何利用少量的有标签训练样本或者源领域数据，建立一个可靠的模型对目标领域数据进行预测(源领域数据和目标领域数据可以不具有相同的数据分布)。He 等人[He, 2008]一文指出数据分类首先要解决训练集样本抽样问题，如何抽

到具有代表性的样本集作为训练集是一个值得研究的重要问题。该文提出极小样本集抽样方法用于基于超曲面分类算法，该方法可感知非结构化数据的分布，并以极小样本集作为代表子集。该文还指出了极小样本集有多少种表达方式。给出了样本缺失情况下准确率的精确估计。这篇文章表明在实际中保证训练得到的分类模型具有高准确性和可靠性的两个基本的假设并不是每个算法都能做到的，因此研究迁移学习变得非常重要。

近年来，迁移学习已经引起了广泛的关注和研究[Ben-David, 2007; Blitzer, 2006; Dai, 2007; Dai, 2007a; Liao, 2005; Xing, 2007; Mahmud, 2007a; Samarth, 2006; Bel, 2003; Zhai, 2004; Dai, 2007b; Pan, 2010; Luo, 2008; Duan, 2009; Dai, 2008a; Zhuang, 2009; Zhuang, 2010]。根据维基百科的定义[2]，迁移学习是运用已存有的知识对不同但相关领域问题进行求解的新的一种机器学习方法。它放宽了传统机器学习中的两个基本假设，目的是迁移已有的知识来解决目标领域中仅有少量有标签样本数据甚至没有的学习问题。迁移学习广泛存在于人类的活动中，两个不同的领域共享的因素越多，迁移学习就越容易，否则就越困难，甚至出现“负迁移”[Rosenstein, 2005；Dai, 2009]，产生副作用。比如：一个人要是学会了自行车，那他就很容易学会开摩托车；一个人要是熟悉五子棋，也可以轻松地将知识迁移到学习围棋中。但是有时候看起来很相似的事情，却有可能产生“负迁移”，比如，学会自行车的人来学习三轮车反而不适应，因为它们的重心位置不同[戴 08; 施 09]。近几年来，已经有相当多的研究者投入到迁移学习领域中，每年在机器学习和数据挖掘的顶级会议中都有关于迁移学习的文章发表，比如，ICML，SIGKDD，NIPS，ICDM 以及 CIKM 等。目前对迁移学习的研究主要集中在单个源领域到单个目标领域的迁移分类学习[Dai, 2007a; Xing, 2007, Ling, 2008; Wu, 2004]，而对多个源领域的知识进行迁移到单个或多个目标领域还很缺乏。从一个源领域到目标领域的迁移学习往往不够，容易产生学习偏见，或者知识不够，导致性能不好。例如：要想学习好泛函分析，仅有代数学的基础是不够的，可能还得有几何学、微积分学以及函数论等相关学科的基础。文献[Gao, 2008; Gao, 2009]中提出的方法虽然可以处理多个源领域到单个目标领域的迁移学习，但只是采取了优化的策略对各个源领域训练得到的子分类器进行优化集成，从而提高算法的分类性能。如何深入地挖掘、开发各个源领域数据的内部结构或者数据分布，最终提高算法的性能是本文研究的主要内容之一。

由于在实际应用中，源领域与目标领域的数据往往服从不同分布，而且有标签数据很难获得，因此对于迁移学习算法的研究已经变得越来越重要，而且具有挑战性。此外迁移学习算法的应用前景还非常广泛，包括文本、图像分类，情感分类，强化学习，排序学习，度量学习，人工智能规划，文本、图像聚类，协同过滤以及基于传感器定位估计等。本文主要对迁移学习中的文本分类算法进行研究。

[2] http://en.wikipedia.org/wiki/Transfer_learning

1.2. 问题描述

迁移学习研究的是如何利用相关源领域中的知识，应用到目标领域，从而提高机器学习算法在目标领域上的性能。分类算法研究如何利用源领域中大量有标签的样本，来提高目标领域数据的分类效果。给出训练集 $D_s = \{(\mathbf{x}_i, y_i)\}|_{i=1}^{n_s}$ 来自于源领域，服从于某种数据分布 $P_s(X,Y)$；而测试集 $D_t = \{(\mathbf{x}_i, y_i)\}|_{i=1}^{n_t}$ 来自于目标领域，服从于某种数据分布 $P_t(X,Y)$，其中 n_s 和 n_t 分别为训练集和测试集的样本个数，$\mathbf{x}$ 为样本点，y 为对应样本类别。如图 1.2.1 所示的二维数据集，源领域数据和目标领域数据服从于不同的数据分布，即 $P_s(X,Y) \neq P_t(X,Y)$，且 h_s 和 h_t 分别为源领域和目标领域的理想分类器。如果从源领域数据训练得到一个分类器 h_s，可以看到虽然 h_s 可以在源领域中表现很好，但却不能对目标领域数据进行很好的预测。

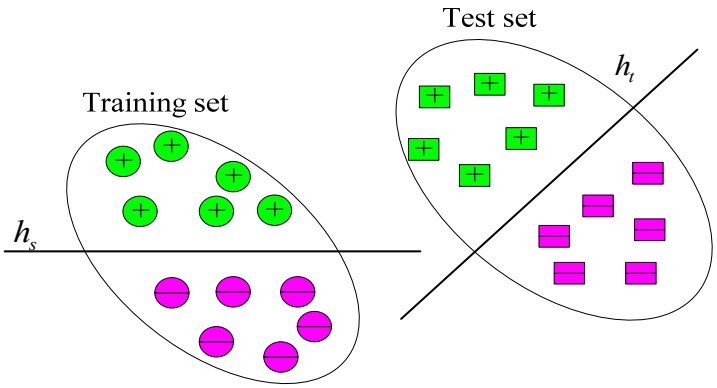


图 1.2.1 源领域与目标领域数据分布不一致示意图

表 1.2.1 主要符号说明表

符号	说明	符号	说明
$\mathbb{R}$	实数空间	$\mathbb{R}_+$	非负实数空间
$\mathbf{x}$	样本空间的一个样本点，列向量	y	对应样本 $\mathbf{x}$ 的类别输出
X	表示 $\mathbf{x}$ 的随机变量	Y	表示 y 的随机变量
h	称为分类模型，也叫假设	f	称为分类模型，也叫假设
$\mathbf{T}$	表示矩阵	$\mathbf{T}^{\mathrm{T}}$	表示矩阵的转置
$P(X,Y)$	表示随机变量 X 和 Y 的概率分布		

我们还进一步假设源领域与目标领域数据的特征空间一样，而且目标领域中无任何有标签的数据。本文致力于利用源领域中有标签的样本数据，把知识迁移到目标领域中，

从而在目标领域中得到理想的分类性能。本文提出的跨领域分类算法主要对文本进行处理，但也可以扩展到其他领域。表 1.2.1 给出了本文的主要符号说明表。值得强调的是，本文对文本数据进行处理，且采用 $tf \cdot idf$ 作为样本点的特征，因此特征空间的取值为非负，即 $\mathbf{x} \in \mathbb{R}_{+}^{m}$，其中 m 为样本的特征维度。

1.3. 本文的主要贡献

本文主要对迁移学习中的文本分类算法进行研究。传统机器学习中的两个基本假设条件，(1) 用于学习的训练样本与新的测试样本满足独立同分布的条件；(2) 必须有足够可利用的训练样本才能学习得到一个好的分类模型。本文研究的跨领域学习算法放宽了这两个基本假设条件，旨在解决目标领域与源领域数据服从不同数据分布，且目标领域中无任何标签数据的分类问题。针对研究的问题，从简单到复杂，由浅到深，对文本分类算法进行了研究。主要完成的工作包括从单个源领域到单个目标领域，单个源领域到多个目标领域，多个源领域到单个目标领域以及多个源领域到多个目标领域的迁移学习等。首先针对从单个源领域到单个目标领域的学习问题，提出了基于混合正则化的无标签目标领域的迁移学习算法以及一种有效挖掘词特征聚类与文档类别关联关系的迁移学习算法；第二，针对从多源领域学习问题，研究了一致性正则化框架；第三，区别于判别模型，研究基于生成模型的迁移学习分类算法，该方法可以同时处理多个源领域多个目标领域的学习问题；最后对本文提出的几种多源领域学习算法以及以往的迁移学习算法进行了比较。

(1) 提出基于混合正则化的无标签领域归纳迁移学习方法

为了解决目标领域无标签数据以及源领域数据是不同分布的分类问题，并建立一个归纳分类模型对新来的目标数据进行预测，首先分析了直推式迁移学习(Transductive Transfer Learning)中存在的类别比例漂移问题，然后提出归一化的方法使得预测的类别比例接近于实际样本类别比例。更进一步，提出了一种基于混合正则化框架的归纳迁移学习算法。其中包括目标领域分布结构的流形正则化，预测概率的熵正则化，以及类别比例的期望正则化。这个框架被用于从源领域到目标领域学习的归纳模型中。

(2) 提出一种有效挖掘词特征聚类与文档类别关联关系的迁移学习算法

跨领域分类学习的目标是在源领域数据与目标领域数据具有不同数据分布的情况下，把从有标签源领域学习到的知识适应到无标签目标领域中。虽然在原始词特征上，源领域与目标领域的数据分布不同，但是不同领域词特征聚类(词概念)与文档类别之间的关联关系可能一样。因此，开发可以这种与领域独立的关联关系，并且作为源领域与目标领域之间知识迁移的桥梁。由此提出了同时分解源领域与目标领域数据矩阵的联合优化框架，其中共享词特征聚类(词概念)与文档类别之间的关联关系。为了求解该优化框架，提出了一

个迭代算法，并从理论上分析了其收敛性。

(3) 提出基于一致性正则化的多源跨领域学习框架

提出了基于一致性正则化从多源领域到目标领域的跨领域分类框架。在这个框架下，局部的子分类器不仅考虑了在源领域上的可利用的局部数据，而且考虑了这些由源领域知识得到的子分类器在目标领域上的预测的一致性。更进一步，在理论上分析了一致性正则化的有效性。最后，为了处理各个源领域数据在地理上分布的情况，提出了一致性正则化的分布式实现，可避免收集各个领域数据到中心节点，而只是传递一些统计变量，一定程度上减轻了数据信息的隐私性担忧。

(4) 给出基于生成模型的挖掘多领域共性与特性的跨领域分类方法

从生成模型的角度研究了多领域学习，有效挖掘多领域间的共性与特性。区别于概率隐性语义分析模型(PLSA)，只有一个隐性变量，这里提出的 CD-PLSA 模型有两个隐性变量 y 和 z，分别表示词概念和文档类别。不同领域间的共性把它们的特性联系起来，并且作为知识迁移的桥梁。提出一个 EM 算法来求解 CD-PLSA 模型，并实现了处理领域数据分布在不同节点的分布式算法。

(5) 对多源领域跨领域学习算法进行比较

对本文提出的几种多源领域跨领域学习算法进行了比较。首先扩展基于非负矩阵的跨领域方法，使之能同时处理多源领域。然后，对基于生成模型的算法 CD-PLSA 进行改进。实验表明这几种多源跨领域算法各有优缺点，但都比传统监督学习算法性能优越。本文提出的多源领域学习算法也比以往的跨领域学习算法 CoCC，LWE 表现得好，且能处理迁移学习问题比较难的情况，具有较强的迁移学习能力。

1.4. 论文的组织

本文共分为八章，除第一章之外的各章节组织如下：

在第二章，介绍了迁移学习的研究现状，主要讨论了当前迁移学习方法采用的技术以及模型等。

在第三章，研究分析了几种半监督学习方法，并把它们应用到迁移学习中，提出了基于混合正则化的迁移学习方法，其中包括熵正则化，流形正则化以及期望正则化。

在第四章，提出了一种有效挖掘词特征聚类与文档类别关联关系的迁移学习算法。通过观测发现虽然不同领域数据分布不一致，但是词特征聚类与文档类别之间的关联关系可能是一样的。因此本章运用非负矩阵算法，提出一个联合优化框架来有效挖掘不用领域之间这种关联关系。

在第五章，提出了基于一致性正则化的多元跨领域学习算法，并且从理论上分析了该算法的有效性。

在第六章，对第四章提出的算法进行了分析和完善，基于矩阵分解的方法缺乏概率解释，比如不同领域之间词特征聚类与文档类别之间的关联关系矩阵就不满足概率条件。因此对生成模型进行了研究，并且提出了有效挖掘多领域之间共性与特性的跨领域分类方法。该方法还能同时处理多个源领域和多个目标领域。

在第七章，对多源领域跨领域学习算法进行了实验比较。

在第八章，对本文做出的工作进行总结，并且指出了一些可能的研究方向。

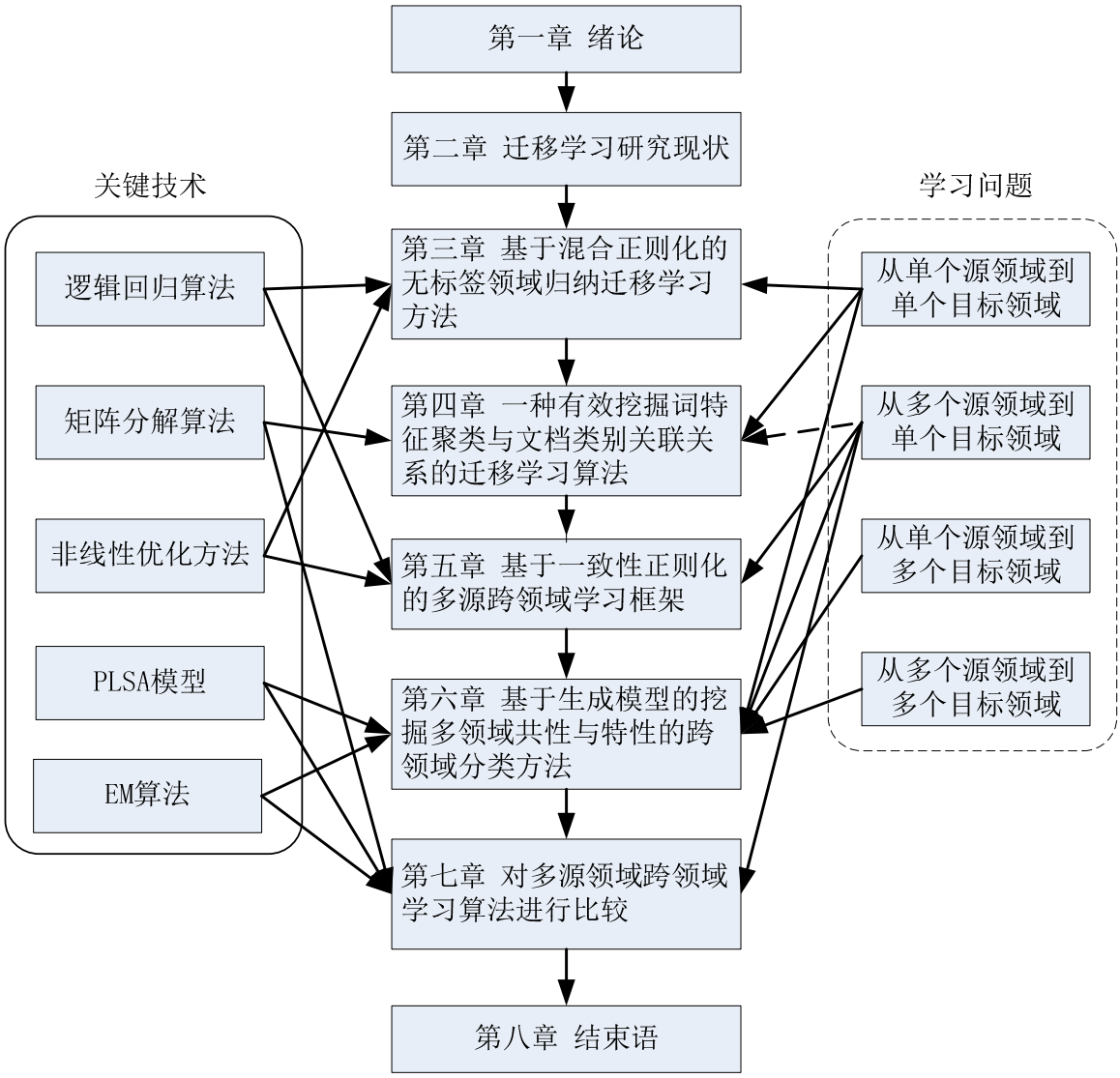


图 1.4.1 本文的内容组织结构图

这些章节的内容组织结构如图 1.4.1 所示。根据研究的迁移学习问题的逐层深入，第三章和第四章属于从单个源领域到单个目标领域的工作，而第五章和第六章属于多源领域学习问题。第六章提出一个通用的学习框架，能够处理任意多源领域和多目标领域，包括了三、四、五章处理的迁移学习问题。在第七章中，首先扩展第四章的工作到处理多源领域，然后对所有处理多源领域的算法进行实验比较。图 1.4.1 还给出了各个章节采用的关键技术。第三章和第五章主要基于逻辑回归算法，然后用非线性优化方法求解，

第四章是基于非负矩阵分解算法的迁移学习方法，而第六章则是基于 PLSA(Probablity Latent Sementic Analysis)模型，并采用 EM(Expectation Maximization)算法求解。

第二章　迁移学习研究现状

2.1. 引言

迁移学习在 20 世纪 90 年代就被引入到机器学习领域，直到今天已经提出了很多好的算法[Pan, 2010; Jiang, 2008; Bruzzone, 2009; BruzzoneM10]。

Pan 和 Yang[Pan, 2010]针对源领域和目标领域样本是否标注以及任务是否相同或者是否单一对迁移学习进行了划分，如图 2.1.1 和表 2.1.1 所示。

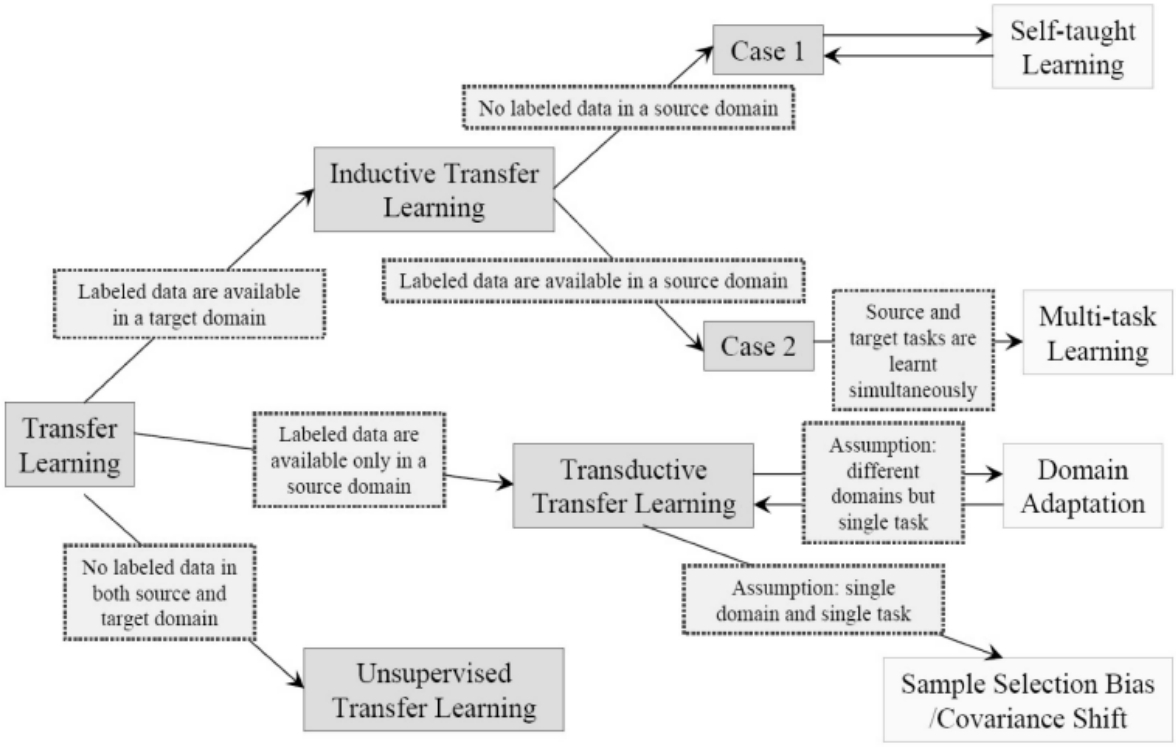


图 2.1.1 迁移学习按照源领域和目标领域样本是否有标注进行划分[Pan, 2010]

表 2.1.1 传统机器学习与各种迁移学习情形之间的关系[Pan, 2010]

Learning Setting		Source and Target Domains	Source and Target tasks
Traditional Machine Learning		the same	the same
Transfer Learning	Inductive Transfer Learning	the same or different but related	different but related
	Unsupervised Transfer Learning	the same or different but related	different but related
	Transductive Transfer Learning	different but related	the same

从图 2.1.1 可以看到，根据源领域和目标领域中是否有标签样本迁移学习划分为三

类，目标领域中有少量标注样本的归纳迁移学习(Inductive Transfer Learning)，只有源领域中有标签样本的直推式迁移学习(Transductive Transfer Learning)，以及源领域和目标领域都没有标签样本的无监督迁移学习。另外还根据源领域中是否有标签样本把归纳迁移学习划分成多任务学习、自学习。表 2.1.1 给出了传统机器学习与各种迁移学习情形之间的关系，以及各种情形下，源领域与目标领域是否相同，源领域与目标领域的任务是否相同。迁移学习是和传统学习相对应的一大类学习方式，传统学习处理源领域和目标领域相同且源领域和目标领域的任务是相同的学习，迁移学习处理除此情形之外的学习，包括：源领域和目标领域的任务相关但不同的归纳迁移学习[Dai, 2007b; Jiang, 2007; Liao, 2005; Lee, 2007; Wang, 2008a; Larence, 2004; Schwaighofer, 2005; Evgeniou, 2004]；源领域和目标领域相关但不相同而源领域和目标领域的任务相同的直推式迁移学习(Transductive Transfer Learning)[Zadrozny, 2004; Huang, 2007; Fan, 2005; Ando, 2005; Blitzer, 2006; Blitzer, 2008; Xing, 2007]。无监督迁移学习与归纳迁移学习类似，不过主要处理源领域和目标领域中都没有标签数据的问题[Dai, 2008; Wang, 2008b]。他们还根据训练样本和测试样本是否来自于同一个领域，把直推式迁移学习划分为样本选择偏差、协方差偏移和领域自适应学习这些相关的子领域。领域自适应学习面对的是任务相同但不同领域的迁移学习问题，即源领域与目标领域的数据边缘分布不同。本文主要针对领域自适应算法进行研究，即源领域和目标领域不同，且只有源领域中有标签数据。

本章主要从迁移学习算法所采用的技术手段对迁移学习研究现状进行介绍。首先介绍与迁移学习极其相关的半监督学习以及多任务学习方法，然后根据迁移学习方法中采用的技术，我们可以把迁移学习方法分为三类：i) 基于特征选择的迁移学习算法研究；ii) 基于特征映射的迁移学习算法研究；iii) 基于权重的迁移学习算法研究。

2.2. 半监督学习

在传统的监督学习中，学习算法通过对大量有标签的训练样本进行学习，从而建立模型用于预测标记新来的没有标签的测试样本。但是随着信息技术、互联网以及存储技术的快速发展，数据量随着指数级增长。人们能够比较容易地收集大量的没有标签的数据，但要获取大量有标签的数据则较为困难，因为这可能需要耗费大量的人力物力。例如，在生物学中进行数据分类，得到一个训练样本的标签往往需要大量的，长时间的，昂贵的实验；在进行 Web 网页推荐时，用户也不愿意花费大量的时间来标记哪些网页是他感兴趣的，因此有标签的网页很少。实际上，在真实世界中通常存在大量的无标签的数据，而有标签的数据则较少。这就需要一种机器学习技术能够利用大量的无标签样本数据以及少量有标签的训练样本进行学习，提高分类任务的准确率。

按照Zhou等人[Zhou, 2006]在文献中的阐述，目前能够利用少量有标签数据和大量没有标签样本数据的技术有三类：半监督学习(Semi-supervised Learning)、直推式学习(Transductive Learning)和主动学习(Active Learning)。这些学习方法都通过大量的无标签样本来辅助少量有标签样本的学习，但它们在思想上又有些不同。半监督学习指的是学习算

法在学习过程中不需要人工干预，基于自身对无标签数据加以利用。而直推式学习，它与半监督学习一样不需要人工干预，但不同的是，直推式学习假设无标签的数据就是最终要用来测试的数据，学习的目的就是在这些数据上取得最佳泛化能力。相对应的，半监督学习在学习时并不知道最终的测试用例是什么。因此，半监督学习考虑的是一个“开放的世界”，即在学习中不知道测试样本是什么，而直推式学习考虑的则是一个“封闭世界”，要测试的样本数据已经参与到学习过程中。如果抛开是否对未知样本进行预测，其实直推式学习可以归结为半监督学习的一种特例。主动学习与半监督学习、直推式学习最大的区别在于它的学习过程需要人工的干预，就是在学习过程通过反馈尽可能地找到那些包含信息量大的样本来辅助少量有标签样本的学习。在传统机器学习中，这三种方法已经得到了广泛应用[Zhu, 2005; Joachims, 1999a; Joachims, 2003; Tong, 2001; Cohn, 1994; Sindhwani, 2005]。多视角学习(Multi-view Learning)也是半监督学习一个很重要的学习任务。Yarowsky[Yarowsky, 1995]和Blum等人[Blum, 1998]认为数据的多视角表示方式可以提高半监督分类学习算法的性能。更一步，文献[Dasgupta, 2001; Abney, 2002; Abney, 2004]用PAC (Probability Approximately Correct)理论分析了联合训练(Co-training)在无标签数据上错误率的上界。

近年来也有很多研究者把这些技术应用到迁移学习领域。文献[施 09]对主动迁移学习模型进行了研究。Shi 等人[Shi, 2008]提出了一种跨领域的主动迁移学习方法，通过似然偏置的大小来选择领域外(out-of domain)有标签的样本。那些能够正确预测领域内(in-domain)数据且高似然偏置的有标签样本被利用，而那些低偏置的样本则通过主动学习进行选择。Liao 等人[Liao, 2005]等提出了一种方法，估计源领域中的每个样本与目标领域中少量标签数据之间的不匹配程度，并把该信息应用到逻辑回归中。本文第三章综合半监督学习的三种正则化技术，流形正则化[Belkin, 2006]、熵正则化[Grandvalet, 2005]以及期望正则化[Mann, 2007]，提出基于混合正则化的迁移学习方法。该方法首先从源领域训练得到一个分类器，然后通过混合正则化在目标领域数据上进行优化。

自学习(Self-taught Learning)[Raina, 2007; Dai, 2008]也是一种利用大量无标签数据来提高给定分类聚类任务性能的方法，自学习被应用于迁移学习中，因为它不要求无标签数据的分布与目标领域中的数据分布相同。Raina 等人 [Raina, 2007]提出了一种自学习的方法，它利用稀疏编码技术对无标签的样本数据构造高层特征，然后少量有标签的数据以及目标领域无标签的样本数据都由这些简洁的高层特征表示。实验表明这种方法可以极大地提高分类任务的准确率。

2.3. 多任务学习

多任务学习是同时对几个相关的问题进行学习的机器学习方法，这些任务共享相同的表示。这种学习方式同样可以得到更好的模型，因为在学习中允许各个任务使用它们之间共性的东西。因此多任务学习[Caruana, 1997; Thrun, 1996]也可以看成是迁移学习早期的研究。Caruana 指出多个任务在使用共同的表示时，可以并行地执行，而且这些任务在学习中相互获利，比单个任务的学习更好。多个任务学习可以应用于许多不同的领域和不同的算法，因此在现实世界中也是非常有用的。

Bakker 等人[Bakker, 2003]运用贝叶斯的方法去估计多个问题所共有的特征参数，从而解决多任务学习的问题。Bai 等人[Bai, 2009]研究学习了多个任务中的非参数共同结构，然后提出了一种算法迭代发现对所有任务都有效的超级特征，最后每个任务的函数估计是这些超级特征的线性组合。文献[Jebara, 2004]利用特征和核函数的选择结合支持向量机来解决多任务学习的问题。Argyriou 等人[Argyriou, 2007]提出了一种针对多任务的空间降维技术，试图寻找一个可以表示所有任务的低维特征空间。类似相关的工作还有[Obozinski, 2006; Gu, 2009]。但多任务学习与迁移学习不同的是，它强调算法在所有任务上都要表现得很好，而迁移学习只强调目标领域上的性能。

2.4. 基于特征选择方法

基于特征选择的迁移学习方法是识别出源领域与目标领域中共有的特征表示，然后利用这些特征进行知识迁移[Jiang, 2007a; Dai, 2007; Zhuang, 2010a]。Jiang 等人[Jiang, 2007a]认为与样本类别高度相关的那些特征应该在训练得到的模型中被赋予更高的权重，因此他们在领域适应问题中提出了一种两阶段的特征选择框架。第一阶段首先选出所有领域(包括源领域和目标领域)共有的特征来训练一个通用的分类器；然后从目标领域无标签样本中选择特有特征来对通用分类器进行精化从而得到适合于目标领域数据的分类器。Dai 等人[Dai, 2007]提出了一种基于联合聚类(Co-clustering)的预测领域外文档的分类方法 CoCC，该方法通过对类别和特征进行同步聚类，实现知识与类别标签的迁移。CoCC 算法的关键思想是识别出领域内(也称为目标领域)与领域外(也称为源领域)数据共有的部分，即共有的词特征。然后类别信息以及知识通过这些共有的词特征从源领域传到目标领域。

2.5. 基于特征映射方法

基于特征映射的迁移学习方法是把各个领域的数据从原始高维特征空间映射到低维特征空间，在该低维空间下，源领域数据与目标领域数据拥有相同的分布[Pan, 2008; Blitzer, 2006; Si, 2010; Si, 2010a]。这样就可以利用低维空间表示的有标签的源领域样本数据训练分类器，对目标测试数据进行预测。该方法与特征选择的区别在于这些映射得到的特征不在原始的特征当中，是全新的特征。

Pan 等人[Pan, 2008]提出了一种新的维度降低迁移学习方法，他通过最小化源领域数据与目标领域数据在隐性语义空间上的最大均值偏差(*Maximun Mean Discrepancy*)，从而求解得到降维后的特征空间。在该隐性空间上，不同的领域具有相同或者非常接近的数据分布，因此就可以直接利用监督学习算法训练模型对目标领域数据进行预测。Gu 等人[Gu, 2009]探讨了多个聚类任务的学习(这些聚类任务是相关的)，提出了一种寻找共享特征子空间的框架。在该子空间中，各个领域的数据共享聚类中心，而且他们还把该框架推广到直推式迁移分类学习。Blitzer 等人[Blitzer, 2006]提出了一种结构对应学习算法

(Structural Corresponding Learning, SCL)，该算法把领域特有的特征映射到所有领域共享的“轴”特征，然后就在这个“轴”特征下进行训练学习。SCL 算法已经被用到词性标注[Blitzer, 2006]以及情感分析[Blitzer, 2007]中。类似的工作还有[Xie, 2009]等。

2.6. 基于权重方法

在迁移学习中，有标签的源领域数据的分布与无标签的目标领域数据的分布是不一样的，因此那些有标签的样本数据并不一定是全部有用的。如何侧重选择那些对目标领域分类有利的训练样本？这就是基于实例的迁移学习所要解决的问题。基于实例的迁移学习通过度量有标签的训练样本与无标签的测试样本之间的相似度来重新分配源领域中样本的采样权重。相似度大的，即对训练目标模型有利的训练样本被加大权重，否则权重被削弱。Jiang 等人[Jiang, 2007]提出了一种实例权重框架来解决自然语言处理任务下的领域适应问题。他们首先从分布的角度分析了产生领域适应问题的原因，主要有两方面：实例的不同分布以及分类函数的不同分布。因此他们提出了一个最小化分布差异性的风险函数，来解决领域适应性问题。Dai 等人[Dai, 2007b] 扩展 Boosting 学习算法到迁移学习中，提出了 TrAdaBoost 算法。在每次迭代中改变样本被采样的权重，即在迭代中源领域中的样本权重被减弱，而有利于模型训练的目标领域中的样本权重被加强。他们还用 PAC 理论分析证明了该算法的有效性。下面简要介绍 TrAdaBoost 算法。

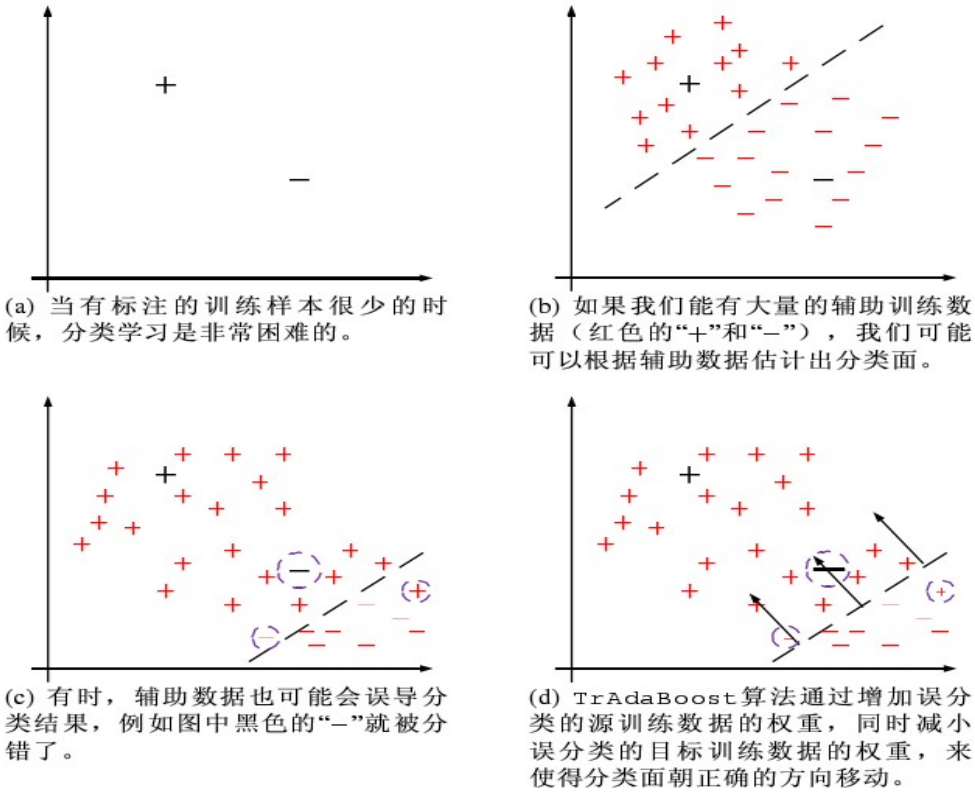


图 2.6.1 关于 TrAdaBoost 算法思想的一个直观示例[Dai, 2007b]

用于迁移学习任务中的源领域数据与目标领域数据虽然分布不同，但是相关的。也就是辅助的源领域中的训练样本存在一部分比较适合用来学习一个有效的分类模型，并且对目标测试样本是适用的。于是 TrAdaBoost 算法的目标就是从辅助的源数据中找出那些适合测试数据的实例，并把这些实例迁移到目标领域中少量有标签样本的学习中去。该算法的关键思想是利用 Boosting 的技术过滤掉源领域数据中那些与目标领域中少量有标签样本最不像的样本数据。其中，Boosting 技术用来建立一种自动调整权重机制，于是重要的源领域样本数据权重增加，不重要的源领域样本数据权重减小。在 TrAdaBoost 中，AdaBoost[Freund, 1997]被用在目标领域中少量有标签样本中，以保证分类模型在目标领域数据上的准确性；而 Hedge(β)[Freund, 1997]被用在源领域数据上，用于自动调节源领域数据的重要度。一个直观 TrAdaBoost 的例子如图 2.6.1 所示。另外对参数加权组合的工作，如[Dredze, 2010]。

除了以上介绍的技术外，最近非负矩阵分解技术也被用到迁移学习中。Li 等人[Li, 2009]提出两阶段的基于约束非负矩阵三因子分解的情感分类(Sentiment Classification)方法，首先通过源领域的学习把文档的类别信息迁移到词特征上，然后把情感分类信息通过词特征迁移到目标领域的文档。他们假设源领域和目标领域共享相同的词特征聚类信息，但由于源领域和目标领域的分布不同使得该假设不尽合理。因此在第四章，本文提出源领域与目标领域之间共享的是词特征聚类与文档类别之间的关联关系，并以此为桥梁进行知识迁移。实验分析表明，该假设更加合理。

根据迁移学习算法是否从多个源领域进行学习，又可以划分为单源领域学习以及多源领域学习。对于单源领域学习[Dai, 2007a; Xing, 2007, Ling, 2008; Wu, 2004]，Ben-David 等人[Ben-David, 2007]分析了领域数据的表示，并提出了一个很好的模型，该模型不仅最小化分类模型在训练数据上的泛化误差，而且最小化源领域与目标领域之间的不同性。Ling 等人[Ling, 2008]提出了一种新的光谱分类算法，该算法通过优化一个目标函数来寻找源领域中的监督信息与目标领域的本质结构之间的最大一致性。Mahmud 等人[Mahmud, 2007; Mann, 2007]从算法信息论的角度来研究迁移学习，该方法度量了不同任务之间的相关性，然后决定多少信息可以做迁移以及怎么迁移这些信息。Xing 等人[Xing, 2007]提出了一种直推式迁移学习方法，该方法首先开发利用所有数据集(包括源领域数据和目标领域数据)上的几何分布结构，然后再利用目标领域上的流形结构。目前很多迁移学习工作主要集中在单个源领域到单个目标领域的迁移分类学习，而对多个源领域的知识进行迁移到单个或多个目标领域还很不多。从一个源领域到目标领域的迁移学习往往不够，容易产生学习偏见，或者知识不够，导致性能不好，因此本文第五章和第六章将对多源领域学习算法进一步研究。Gao 等人[Gao, 2008]提出了一种多模型局部结构映射方案，实际上是对不同源领域训练得到的模型赋予不同的投票权重，而该权重是由预测样本本身的局部分布结构决定的。Gao 等人[Gao, 2009]解决了不同模型的一致性问题。这两个多源领域学习的工作很好地处理了多个模型的集成问题。但这两个方法虽然可以处理多个源领域到单个目标领域的迁移学习，但只是采取了优化的策略对各个源领域训

练得到的子分类器进行优化集成。为了更加深入地挖掘、开发各个源领域数据的内部结构或者数据分布，本文第五章提出一致性正则化框架，在这个框架下，局部的子分类器不仅考虑在源领域上的可利用的局部数据，而且考虑这些由源领域知识得到的子分类器在目标领域上的预测的一致性。实验表明本文提出的算法比 LWE 算法[Gao, 2008]优越。更进一步，第六章提出基于生成模型的挖掘多领域之间共性与特性的迁移学习方法。实验表明该算法比基于判别模型的算法更加有效。

第三章　基于混合正则化的无标签领域归纳迁移学习方法

3.1. 引言

分类学习在智能信息处理中起着关键性的作用，其中包括 Web 网页，图像以及视频的处理等。传统的分类学习研究假设标签数据与无标签数据来源于相同数据分布，但是在实际中无标签数据可能来自于不断变化的但语义相关的不同信息源，因此现有的分类方法不能很好处理这种情况。到目前为止，大多数迁移学习方面的工作都是处理目标领域数据里面有少量标签数据的问题，比如文献[Dai, 2007b; Raina, 2007; Liao, 2005; Yang, 2007; Wu, 2004]。但即使要标注少量目标领域里面的数据也相当困难，因为这也要消耗很大的人力和时间，所以要求迁移学习方法能够处理完全没有标签的目标领域数据。对于这个问题，Xing 等人[Xing, 2007]提出了一种桥接精化(Bridged Refinement)的迁移学习方法，该方法在精化的过程中不断地修正由源领域训练得到的模型在测试数据集上的预测类别，从而获得较高预测准确率。很明显，基于桥接精化的迁移学习方法是一种直推式(Transductive)的学习方法，它不能产生分类器，只能对模型精化中的目标领域数据进行预测，而不能对新来的样本进行直接判别。对于新来的数据，最原始的方法就是重新进行桥接精化的过程，这对于整个学习过程来说，效率很低。

在本章中，首先分析了直推式迁移学习中的桥接精化方法，发现在迭代的过程中预测得到的样本比例在不断发生变化，而且这种类别漂移很大程度上影响了直推式迁移学习桥接精化算法的性能。在一些实际应用中，对应测试样本的实际类别比例(各个类别的样本数除于总的样本数)可以得到，比如通过领域知识[Mann, 2007]或者基于统计估计得到。一个实际的例子，我们可以估计从网上抓取下来的网页，新闻网页大约占 20%，因此可以在桥接精化算法中加入先验知识 — 测试样本的实际类别比例来提高算法的分类性能。

对于目标领域都是无标签数据的情况，我们提出了一种归纳迁移学习算法。该算法与其他方法的不同在于，不仅能够处理完全无标签的目标领域数据，而且能够产生分类模型，对新来的测试样本进行直接预测。它包括两个阶段，首先从源领域数据学习得到一个分类模型 h_s，该模型代表从源领域数据中学到的知识；在第二阶段中，通过把无标签数据加入到模型 h_s 的精化中，从而得到最终模型 h_t 能够很好地对目标领域数据进行预测。我们研究了半监督学习中的几种技术，提出一种混合正则化框架，包括三个正则化准则：流形正则化[Belkin, 2006]、熵正则化[Grandvalet, 2005]以及期望正则化[Mann, 2007]。该框架利用在源领域数据中训练得到的初始模型 h_s 作为第二阶段优化的初始值，

然后通过非线性数值优化技术使得该混合正则化框架收敛到一个局部最优点 h_t。实际文本分类问题的结果表明，本文提出的归纳迁移学习方法是有效的，且优于以往直推式的迁移学习方法。

3.2. 直推式迁移学习

3.2.1 桥接精化方法

这一小节主要描述直推式迁移学习中的桥接精化方法，这个算法主要包括两个阶段迭代的桥接精化过程。

假设概率矩阵为 $\mathbf{T} \in \mathbb{R}_{+}^{n \times |c|}$，其中 $\mathbb{R}_{+}$ 表示非负实数，$|c|$ 表示数据集的类别数，n 为样本的总数，那么 $\mathbf{T}_{ij}$ 表示为第 i 个样本属于第 j 类的概率，K 是近邻个数。$\mathbf{M}$ 为数据集的邻接矩阵，

$$\mathbf{M}_{ij} = \begin{cases} 0 & \text{如果 } \mathbf{x}_j \text{不是} \mathbf{x}_i \text{的} k\text{-近邻}, \\ \dfrac{1}{K} & \text{如果 } \mathbf{x}_j \text{是} \mathbf{x}_i \text{的} k\text{-近邻}。 \end{cases} \tag{3.1}$$

两个样本 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 之间的相似度由余弦距离

$$\cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j}{\| \mathbf{x}_i \| \cdot \| \mathbf{x}_j \|} \tag{3.2}$$

度量。每次迭代中，样本都从近邻的样本中吸收一部分类别信息，同时也保留自身的部分类别信息，迭代公式如下：

$$\mathbf{T}_{i\cdot}^{m+1} = \alpha \sum_{j: x_j \in N_i} \frac{\mathbf{T}_{j\cdot}^{m}}{K} + (1-\alpha)\mathbf{T}_{j\cdot}^{0} \tag{3.3}$$

其中 $\mathbf{T}_{i\cdot}^{m+1}$ 表示概率矩阵 $\mathbf{T}$ 经过 $(m+1)$ 次迭代后得到的第 i 行的值，N_i 是第 i 个样本所有 k-近邻的集合，$0 < \alpha < 1$ 是近邻类别信息与样本自身类别信息之间的平衡参数。以上的迭代公式也可以写成下面矩阵计算形式，

$$\mathbf{T}^{m+1} = \alpha \mathbf{M} \mathbf{T}^{m} + (1-\alpha)\mathbf{T}^{0} \tag{3.4}$$

可以证明得到概率矩阵收敛于下式，

$$\mathbf{T}^{*} = (1-\alpha)(1-\alpha\mathbf{M})^{-1}\mathbf{T}^{0} \tag{3.5}$$

详细的理论分析证明见文献 Wang 等[Wang, 2008c]。

实际上，我们发现该桥接精化过程中每个阶段都需要两个输入参数，邻接矩阵 $\mathbf{M}$ 和

概率矩阵$\mathbf{T}^0$。在第一阶段的输入中，邻接矩阵$\mathbf{M}$是在所有数据集上(包括源领域和目标领域数据)的邻接矩阵，$\mathbf{T}^0$是初始训练模型h_s在所有数据集上(包括源领域和目标领域数据)的预测概率矩阵。在第二阶段，$\mathbf{M}$仅仅是在目标领域数据集上的邻接矩阵，$\mathbf{T}^0$是第一阶段迭代的结果，但只包括目标领域数据集部分。因此，桥接精化方法的主要思想，首先考虑了所有数据$D = D_s \bigcup D_t$上的流形结构，然后是目标领域数据D_t上的流形结构来提高预测结果的准确率。

3.2.2 改进桥接精化方法

我们考查了概率矩阵在迭代精化过程中的性质，发现概率矩阵$\mathbf{T}$的每一行之和保持不变，即$\sum_{j=1}^{|c|}\mathbf{T}_{ij}=1$。但是$\mathbf{T}$的每一列之和$s=\sum_{i=1}^{n}\mathbf{T}_{ij}$却在不断的变化，且$s$在一定程度上反映了预测样本属于类别$j$的样本数。在实际中，待测样本中属于各个类别样本的数目是固定的，也就是期望s趋近于实际的样本数。

对于二类分类问题，

$$r^p=\frac{\sum_{i=1}^{n}\mathbf{T}_{i1}}{\|\mathbf{T}\|}, \tag{3.6}$$

$$r^n=\frac{\sum_{i=1}^{n}\mathbf{T}_{i2}}{\|\mathbf{T}\|} \tag{3.7}$$

分别表示正负类样本在数据集中所占有的类别比例，其中

$$\|\mathbf{T}\|=\sum_{j=1}^{2}\sum_{i=1}^{n}\mathbf{T}_{ij}\ , \tag{3.8}$$

n是数据集的样本数。

如图 3.2.1 所示，问题 10(表 3.5.1 给出了该问题的详细描述)的正样本类别比例r^p($r^n=1-r^p$)随着迭代过程在不断变化。我们可以看到r^p最终收敛到 0.571，与实际的比例值 0.748 有很大的差异，这可能导致算法的性能变差。

在 Xing 等人[Xing, 2007]的方法中，把类别比例都归一化为 1:1，这虽然可以处理类别比例平衡(正负样本类别比例基本相同)的情况，但对于类别比例严重不平衡的情况不适用。考虑在提供实际类别比例的情况下，算法 3.1 给出了加入类别比例归一化的桥接精化方法。在每次迭代中都把迭代得到的类别比例归一化为实际的类别比例。通过这个归一化技术，可以把问题 10 的预测准确率由 80.47%提高到接近 92%，大大提高了预测准确率。

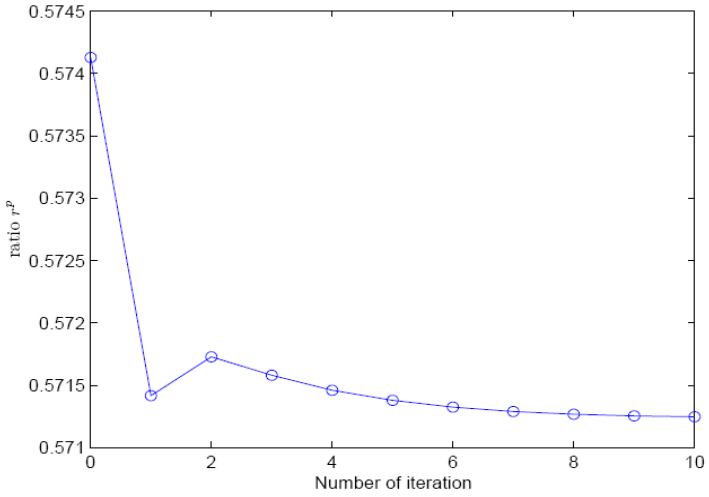


图 3.2.1 问题 10 的正样本类别比例在迭代过程中的变化

算法 3.1: 集成归一化技术的桥接精化方法

输入: 数据集 D，初始未精化的概率矩阵 $\mathbf{T}^0$，平衡参数 α，近邻个数 K，以及所有样本类别的比例先验 $r=(r_1,\cdots,r_{|c|})$，其中 $\sum_{i=1}^{|c|} r_i = 1$。

输出: 精化后的概率矩阵 $\mathbf{T}$。

步骤 1：对每两个样本 $\mathbf{x}_i$ 和 $\mathbf{x}_j$，根据式子(3.2)计算它们之间的相似度；

步骤 2：对每个样本 $\mathbf{x}_i$，求出其 K 近邻 N_i；

步骤 3：求出每个样本的 K 近邻后，构造相邻矩阵 $\mathbf{M}$。

步骤 4：对每个类别 c_j 的预测概率向量做归一化，$\mathbf{T}_{\cdot j} = \mathbf{T}_{\cdot j} / (\| \mathbf{T}_{\cdot j} \| / (r_j \cdot n))$；

步骤 5：对每个样本 $\mathbf{x}_i$，根据式子(3.4)计算第 $t+1$ 次迭代的概率；

对每个类别 c_j 的预测概率向量做归一化，$\mathbf{T}_{\cdot j}^{t+1} = \mathbf{T}_{\cdot j}^{t+1} / (\| \mathbf{T}_{\cdot j}^{t+1} \| / (r_j \cdot n))$；

步骤 6：循环步骤 5，直到预测概率矩阵 $\mathbf{T}$ 收敛。

步骤 7：返回 $\mathbf{T}$。

其中，n 是样本的个数，r_j 是类别 c_j 在样本集 D 中的比例，$\| \mathbf{T}_{\cdot j} \| = \sum_{i=1}^{n} \mathbf{T}_{ij}$。算法的收敛条件是 $|(\| \mathbf{T}^{m+1} \| - \| \mathbf{T}^{m} \|)| < \varepsilon$，$\varepsilon > 0$。

3.3. 基于混合正则化的归纳迁移学习

上一节描述的直推式迁移学习方法不能够产生分类模型，只能对精化过程的样本进行预测，而不能对新来的测试样本直接预测。但是在很多实际应用中，都希望得到一个

最终分类器对新来的样本进行直接预测，因此我们提出了基于混合正则化的两阶段归纳迁移学习方法。该方法可以推广到多类情况，以下仅考虑两类问题。由于我们的算法基于逻辑回归[Davie, 2000]实现，下面将简要描述该分类模型。

3.3.1 逻辑回归

逻辑回归[Davie, 2000]是一种分类学习方法，条件概率 $P(Y \mid X)$ 是在给定样本 X 的情况下，求样本 X 属于类别 Y 的概率，其中 Y 是离散型的值，而 X 是任意的包含离散或者连续型随机变量的向量。逻辑回归通过优化目标函数，在训练数据集上估计参数模型。当 Y 是布尔型时，分类模型如下

$$P(y=\pm 1 \mid \mathbf{x};\mathbf{w}) = \sigma(y\mathbf{w}^{\mathrm{T}}\mathbf{x}) = \frac{1}{1+\exp(-y\mathbf{w}^{\mathrm{T}}\mathbf{x})} \tag{3.9}$$

其中 $\mathbf{w}$ 是参数模型。在最大后验估计原则下，参数模型 $\mathbf{w}$ 通过拉普拉斯先验估计。给定训练数据集 $D=\{\mathbf{x}_i, y_i\}|_{i=1}^{N}$，可以通过优化最大化式子(10)来求参数模型 $\mathbf{w}$，

$$\sum_{i=1}^{N} \log \frac{1}{1+\exp(-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)} - \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{3.10}$$

由于该函数对变量 $\mathbf{w}$ 是一个凹函数，因此可以通过非线性数值优化方法求得全局最优解。当得到参数模型以后，式子(3.9)就可以用来计算待测样本分别属于正负类的概率。

3.3.2 正则化准则

本章提出的混合正则化框架包括三种正则化准则：流形正则化[Belkin, 2006]、熵正则化[Grandvalet, 2005]以及期望正则化[Mann, 2007]。在以往的研究中，这些准则被用于半监督学习中，开发和利用标签和无标签数据的内部本质结构。但在本文中，我们把这些准则用于迁移学习，然后作用于与源领域具有不同数据分布的目标领域数据 $D_t=\{\mathbf{x}_i^{(t)}\}|_{i=1}^{n_t}$ 上(其中 n_t 是目标领域数据集的样本数)。假设参数模型为 $\mathbf{w}$，那么公式(3.9)的函数 σ 同样被用来表示条件概率 $P(y=1 \mid \mathbf{x})$。

流形正则化 Belkin 等人在半监督学习的研究中提出了有效开发和利用标签和无标签数据的流形结构的方法，该准则要求样本的类别标签与其周围样本的类别标签是相似的。我们把该准则运用到完全无标签的目标领域迁移学习中，可以通过最小化以下式子来实现，

$$g_m(\mathbf{w}) = \frac{1}{n_t}\sum_{i=1}^{n_t}[\frac{1}{K}\sum_{k=1}^{K}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_{i_k}) - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)]^2, \tag{3.11}$$

其中 K 是样本 $\mathbf{x}_i$ 的近邻个数，$\mathbf{x}_{i_k}$ 是样本 $\mathbf{x}_i$ 的第 k 个近邻 $(1 \le k \le K)$。我们可以采用任何

一种相似度度量方式来计算样本之间的相邻关系。

熵正则化 Grandvalet 等人提出的熵正则化实现了对样本 $\mathbf{x}_i$ 的预测概率向量 $\mathbf{p}_i=(\mathbf{p}_{i1},\cdots,\mathbf{p}_{i|c|})$ 的熵最小化，其中 $\mathbf{p}_{ij}$ 是样本 $\mathbf{x}_i$ 属于类别 c_j 的概率，$|c|$ 是样本的类别数。熵正则化准则主要是基于每个样本都属于且仅属于一种类别的事实，都希望最终得到真实的概率预测向量。对于二类问题，熵正则化等价于最小化以下式子，

$$g_c(\mathbf{w})=-\frac{1}{n_t}\sum_{i=1}^{n_t}[\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)-\frac{1}{2}]^2 \tag{3.12}$$

期望正则化 Mann 等人提出的期望正则化准则可以使预测得到的结果逼近于一些先验知识，比如样本的实际类别。也就是说，可以使得预测得到的样本类别比例接近于实际的类别比例。形式化表示如下，

$$g_e(\mathbf{w})=\frac{1}{n_t}[\sum_{i=1}^{n_t}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)-r\cdot n_t]^2\ , \tag{3.13}$$

其中 r 是正样本的实际比例。

3.3.3 两个阶段的归纳迁移学习方法

我们提出的归纳迁移学习方法包括两个阶段，首先是训练初始的分类器模型，然后是对初始分类模型的精化。

第一阶段 训练初始分类器模型 h_s。假设源领域的标签数据为 $D_s=\{\mathbf{x}_i^{(s)},y_i^{(s)}\}|_{i=1}^{n_s}$，然后用监督学习算法逻辑回归[Davie, 2000]来学习得到在源数据集上的初始模型 h_s。

第二阶段 通过混合正则化框架精化初始分类器模型 h_s。给出目标领域数据 D_t，我们优化以下目标函数 f：

$$f(\mathbf{w})=\mathbf{w}^{\mathrm{T}}\mathbf{w}+\alpha\cdot g_m+\beta\cdot g_c+\gamma\cdot g_e \tag{3.14}$$

其中 α，β，γ 为这些正则化准则之间的平衡因子，g_m，g_c，g_e 分别为公式(3.11)~(3.13)的定义。

本文提出的混合正则化框架，与半监督学习的不同在于，我们没有考虑源领域标签数据上的 log 似然，如公式(3.10)中的第一项。这是因为在半监督学习中，标签数据和无标签数据都是来自于同种分布，自然地可以共享相同的分类模型。但是在迁移学习中，源领域的数据和目标领域的数据具有不同的数据分布，因此引入源领域数据可能导致性能变差。所以，我们首先在源领域的标签数据上训练初始模型，然后只在目标领域的无标签数据上精化该模型，最后得到优化后的分类模型可以在目标领域上表现得很好且能

直接预测新来的样本。

为了解决该优化问题，我们给出函数 g_m， g_c 以及 g_e 对变量 $\mathbf{w}$ 的偏导数，

$$\begin{aligned}\nabla_{\mathbf{w}} g_m = &\frac{2}{n_t}\cdot\sum_{i=1}^{n_t}\left(\frac{1}{K}\sum_{k=1}^{K}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_{i_k})-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)\right)\cdot \\ &\left(\frac{1}{K}\sum_{k=1}^{K}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_{i_k})(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_{i_k}))x_{i_k}-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)(1-\sigma(\mathbf{w}^{\mathrm{T}}B_i))\mathbf{x}_i\right)\end{aligned}, \tag{3.15}$$

$$\nabla_{\mathbf{w}} g_c = -\frac{2}{n_t}\cdot\sum_{i=1}^{n_t}(\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)-\frac{1}{2})\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i))\mathbf{x}_i, \tag{3.16}$$

$$\nabla_{\mathbf{w}} g_e = \frac{2}{n_t}\left(\sum_{i=1}^{n_t}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)-r\cdot n_t\right)\left(\sum_{i=1}^{n_t}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i))\mathbf{x}_i\right) \tag{3.17}$$

因此目标函数 f 的偏导数为

$$\nabla_{\mathbf{w}} f = 2\cdot\mathbf{w}+\alpha\cdot\nabla_{\mathbf{w}} g_m+\beta\cdot\nabla_{\mathbf{w}} g_c+\gamma\cdot\nabla_{\mathbf{w}} g_e \tag{3.18}$$

由于目标函数 f 既不是凸函数也不是凹函数，因此很难求解得到最优值。但是在给定初始值 h_s 的情况下，可以很容易用非线性数值优化方法得到局部最优解，且这个解是优于初始值的。在本文中，我们采用了共轭梯度方法来求解这个目标函数，详细的共轭梯度优化求解过程见算法 3.2。另外，采用 Matlab 本身自带的函数 ***fminunc***(求无约束最小值点函数)来求解算法 3.2 中步骤 3 的单变量优化问题。

算法 3.2: 共轭梯度法求解混合正则化方法

输入: 初始分类器 $h_s=\mathbf{w}_0$，源领域的数据 $D_s=\{\mathbf{x}_i^{(s)},y_i^{(s)}\}_{i=1}^{n_s}$，近邻个数 K，正样本的类别比例 r，参数 α,β,γ 的值以及误差阈值 ε。

输出: 精化后的分类器 w。

步骤 1：根据式子(3.18)计算偏导数 $\nabla_{\mathbf{w}} f(\mathbf{w}_0)$，如果 $\|\nabla_{\mathbf{w}} f(\mathbf{w}_0)\|<\varepsilon$，则转步骤 7，

否则计算初始的搜索方向 $\mathbf{d}_0$， $\mathbf{d}_0=-\nabla_w f(\mathbf{w}_0)$；

步骤 2： $k=0$；

步骤 3：最小化以下式子，求第 k 次迭代的最佳步长 λ_k，

$$f(\mathbf{w}_k+\lambda_k\mathbf{d}_k)=\min_{\lambda} f(\mathbf{w}_k+\lambda\mathbf{d}_k);$$

步骤 4：得到第 k 次迭代的最佳步长 λ_k 后，计算第 $k+1$ 次的分类器 $\mathbf{w}_{k+1}=\mathbf{w}_k+\lambda_k\mathbf{d}_k$。

步骤 5：计算偏导数$\nabla_{\mathbf{w}} f(\mathbf{w}_{k+1})$，如果$\|\nabla_{\mathbf{w}} f(\mathbf{w}_{k+1})\| < \varepsilon$，则转步骤 7，否则计算第$k+1$次的搜索方向$\mathbf{d}_{k+1}$，$\mathbf{d}_{k+1} = -\nabla_{\mathbf{w}} f(\mathbf{w}_{k+1}) + \mu_k \mathbf{d}_k$；

其中μ_k由 Palak-Ribiere-Polyak 公式计算得到,

$$\mu_k = \frac{\nabla_{\mathbf{w}} f(\mathbf{w}_{k+1})^T [\nabla_{\mathbf{w}} f(\mathbf{w}_{k+1}) - \nabla_{\mathbf{w}} f(\mathbf{w}_k)]}{\nabla_{\mathbf{w}} f(\mathbf{w}_k)^T \nabla_{\mathbf{w}} f(\mathbf{w}_k)}。$$

步骤 6：$k = k+1$，转步骤 3。

步骤 7：返回精化后的分类器$\mathbf{w}$，$h_t = \mathbf{w}$。

3.4. 实验过程和结果

3.4.1 实验数据

我们重新构造数据集 20*Newsgroups*[3]使其符合本文中迁移学习问题的要求，该数据集有两层的层次结构。假设A和B分别表示数据集中顶层的两个类别，A_1，A_2和B_1，B_2分别是属于类别A和B的第二层的子类别。我们构造源领域和目标领域数据集如下，让$A.A_1$和$B.B_1$分别为源领域数据集中的正负类样本；而$A.A_2$和$B.B_2$分别为目标领域数据集中的正负类样本。我们得到 12 个分类问题，如表 3.5.1 所述。同时表 3.5.1 中也列出了正样本在目标领域数据集中的比例r^p。每个文档都用$tf \cdot idf$方法表示成一个向量，文档频率的阈值设置为 5 来选择特征。

3.4.2 性能比较

在本文中，与基于混合正则化框架的归纳迁移学习方法(IHR)比较的算法包括：

- 传统的分类学习算法：SVM[Joachims, 1999]和逻辑回归(LG) [Davie, 2000];
- 直推式迁移学习方法：桥接精化方法(BR)[Xing, 2007]和加入类别比例的强桥接精化方法(PBR)。根据初始预测概率矩阵的不同，桥接精化方法又可以分为BR^{LR}，BR^{SVM}，PBR^{LR}和PBR^{SVM}。BR^{LR}和BR^{SVM}分别表示初始预测概率矩阵由逻辑回归和 SVM 算法预测得到的桥接精化方法，PBR^{LR}和PBR^{SVM}类似。这些直推式方法中，近邻的个数都设置为$K = 70$；SVM 采用线性核，其他的参数都采用默认值[4]。
- 归纳迁移学习方法：基于联合聚类的分类算法 CoCC[Dai, 2007]。

[3] http://people.csail.mit.edu/jrennie/20Newsgroups/

[4] 我们也对分类算法 SVM 和逻辑回归算法的参数进行了调节，发现这并不能进一步提高算法 BR 和 PBR 的性能。文章中采用的默认参数是比较好的，总是可以获得较好的分类性能。这里没有列出详细调参数的实验结果。

— 半监督学习方法[5]：TSVM[Joachims, 1999a]和 SGT[Joachims, 2003]。

算法 IHR，LR，SVM，BR^{LR}，PBR^{LR}，BR^{SVM} 和 PBR^{SVM}：我们比较了以上 7 个算法在 12 个分类问题上的性能，每一种算法在 12 个分类问题上的准确率如图 3.4.1 所示，以及每一种算法在 12 分类问题上的平均性能如表 3.4.1 所示。为了验证算法性能的优越性，我们还做了统计 t 测试(置信度为 95%)，从而发现 1) IHR 算法在统计意义上，相对于算法 LR，SVM，BR^{LR} 以及 BR^{SVM} 有很大的提高；2) IHR 方法相对于算法 PBR^{LR} 和 PBR^{SVM} 的优越性并不是很明显。但是如表 3.4.1 所示，从平均性能上看，IHR 优于算法 PBR^{LR} 和 PBR^{SVM}。

算法 IHR，CoCC[Dai, 2007]，TSVM[Joachims, 1999a]和 SGT[Joachims, 2003]。为了体现出算法之间的可比性和公平性，采用了 Dai 等人[Dai, 2007]论文中的数据(数据的详细描述见其文章中的表 1)，结果如表 3.4.2 所示。从表 3.4.2 可以看到算法 IHR 在所有数据集上的准确率都高于 CoCC，TSVM 和 SGT，再一次验证了 IHR 算法的优越性和有效性。

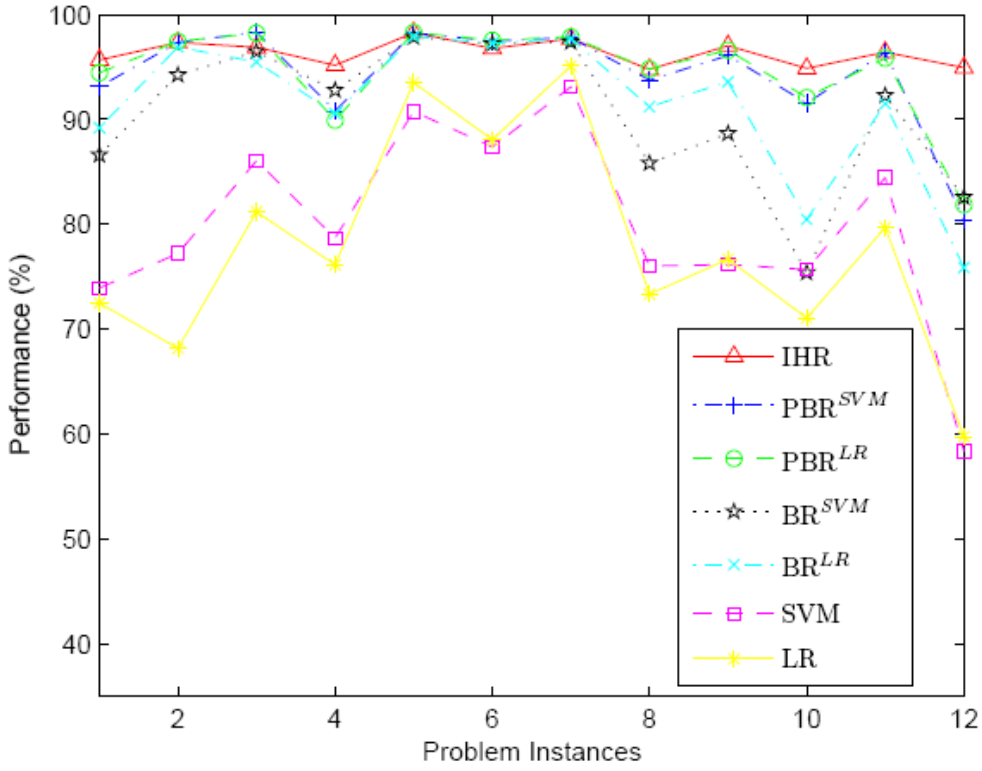


图 3.4.1 算法 IHR，LR，SVM，BR^{LR}，PBR^{LR}，BR^{SVM} 和 PBR^{SVM} 在 12 个分类问题上的性能(%)比较($\alpha = 0.4, \beta = 15, \gamma = 0.12$)

[5] TSVM 和 SGT 的参数设置与 Dai[3]等论文中的设置一致。

表 3.4.1 算法 IHR，LR，SVM，BR^{LR}，PBR^{LR}，BR^{SVM} 和 PBR^{SVM} 在 12 个分类问题上的平均性能(%)比较($\alpha=0.4, \beta=15, \gamma=0.12$)

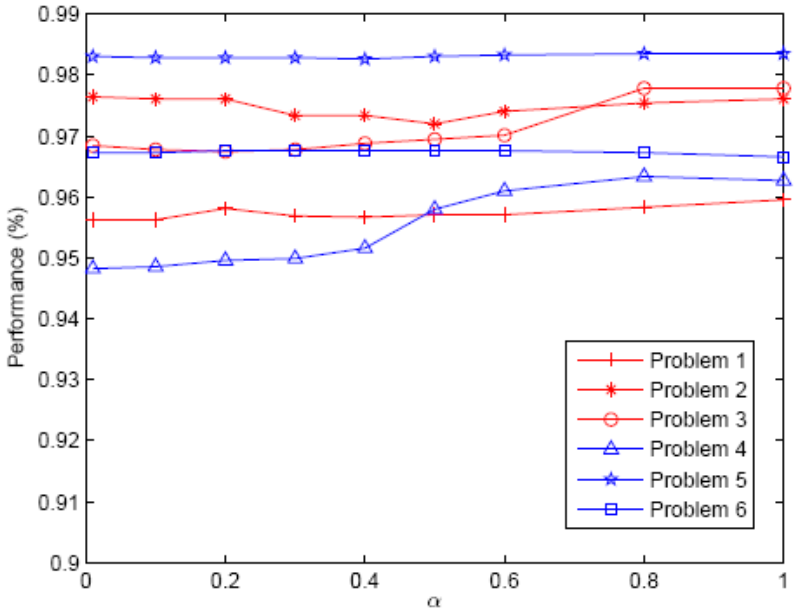
$\overline{\text{LR}}$	$\overline{\text{SVM}}$	$\overline{\text{BR}^{LR}}$	$\overline{\text{BR}^{SVM}}$	$\overline{\text{PBR}^{LR}}$	$\overline{\text{PBR}^{SVM}}$	$\overline{\text{IHR}}$
77.92	79.81	91.46	90.59	94.57	94.27	**96.31**

表 3.4.2 算法 TSVM，SGT，CoCC 和 IHR 之间的性能(%)比较 ($\alpha=0.4, \beta=15, \gamma=0.12$)

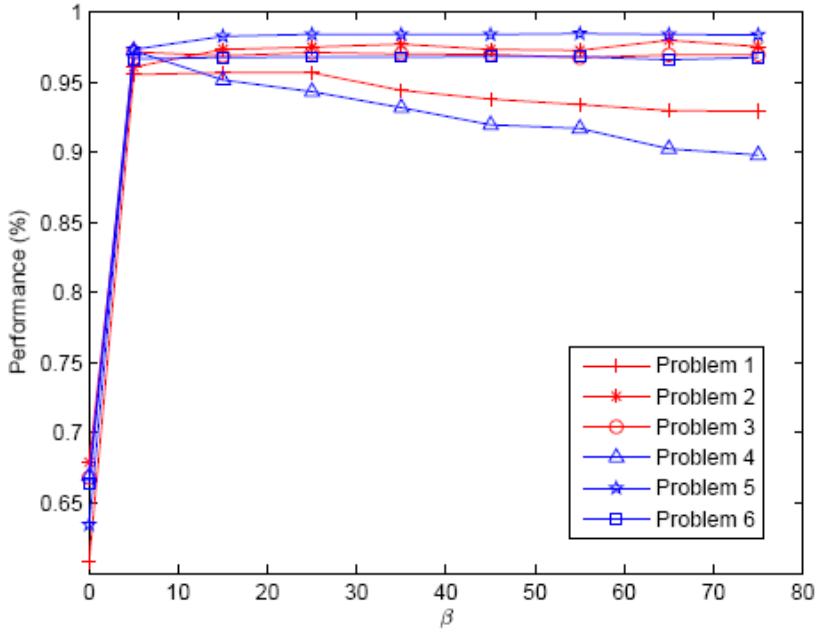
Data set	TSVM	SGT	CoCC	IHR
res vs. talk	96	90.9	96.4	**98.71**
rec vs. sci	93.8	93.8	94.5	**97.15**
comp vs. talk	90.3	97.2	98	**98.28**
comp vs. sci	81.7	72.1	87	**96.65**
comp vs. rec	90.2	95.3	95.8	**97.04**
sci vs. talk	89.2	91.7	94.6	**96.15**

3.4.3 参数影响

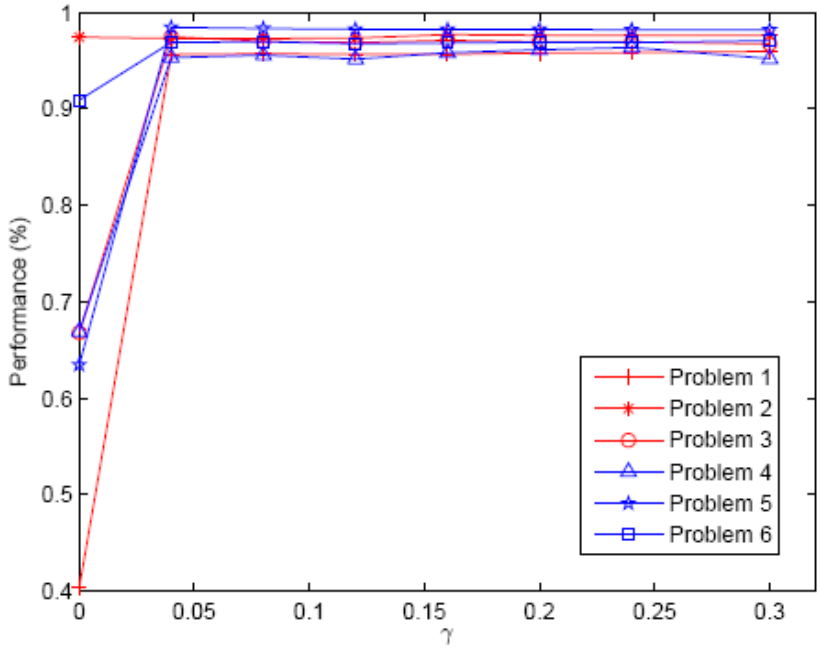
本节还考查了平衡参数 α，β 以及 γ 的不同设置对算法 IHR 的影响，我们选取了 6 个分类问题作为实验数据。在这些参数实验中，我们记录了算法 IHR 的性能随其中任意一个参数变化的影响(另外两个参数取固定值)。结果如图 3.4.2 所示。通过分析，我们发现：



(a) α vs. IHR 的准确率 ($\beta=15, \gamma=0.12$)



(b) β vs. IHR 的准确率（$\alpha = 0.4, \gamma = 0.12$）



(c) γ vs. IHR 的准确率（$\alpha = 0.4, \beta = 15$）

图 3.4.2 参数对算法 IHR 的影响

— 从图 3.4.2(a)，可以看到算法 IHR 的性能对参数 $\alpha \in [0,1]$ 不敏感。从实验结果也可以看到流形正则化对某些分类问题是没有作用的，如问题 5 和 6。我们分析了问题 5 和 6 中出现的现象，发现只要初始模型的预测结果已经有较好的流形结构性，即样本自身的类别已经和近邻样本有很高的一致性，这样流形正则化的作用就很小，甚至不起作用。因此，我们认为影响算法 IHR 性能的不仅是参数的调节，而且跟数据本身的特点也有很大关系。

— 从图 3.4.2(b)和图 3.4.2(c)可以看到参数 β 和 γ 对算法 IHR 有很大的影响，不过

图 3.4.2(c)也表明当γ大于 0.05 的时候，算法的性能是稳定的。

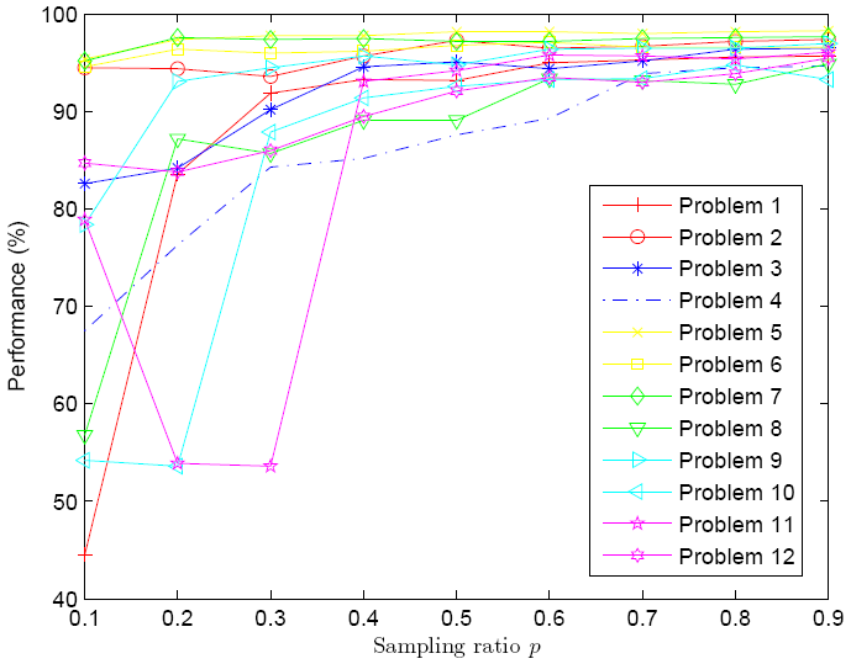
为了证明算法 IHR 对参数设置的健壮性，我们放松参数α，β和γ的取值范围，而不是取以上实验中固定的参数值。经过初步实验，设置参数的取值范围为$\alpha \in (0,1)$，$\beta \in (0,30)$以及$\gamma \in (0,0.5)$，然后评价算法在 12 个分类问题(表 3.5.1 描述)上的性能。我们对参数随机采样m(这里$m=15$)种组合，并在每种参数组合下对所有的 12 分类问题平均算法准确率，结果如表 3.4.3 所示。我们发现该平均性能与参数为($\alpha=0.4$, $\beta=15$, $\gamma=0.12$)时给出的结果几乎一样；而且也看到在所有的参数采样情况下，算法 IHR 的性能都表现得很好，这也再一次验证了算法 IHR 的有效性和健壮性。

表 3.4.3 参数设置对算法 IHR 性能(%)的影响

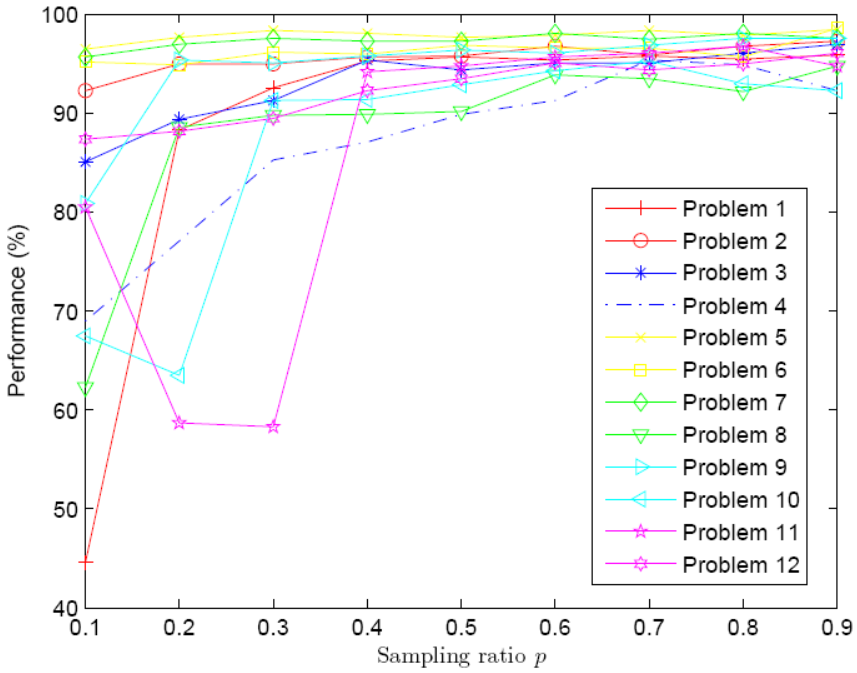
Sample ID	α	β	γ	Problem ID												Average
				1	2	3	4	5	6	7	8	9	10	11	12	
1	0.550	8.897	0.254	95.8	97.0	97.8	97.6	97.8	97.1	97.3	95.2	96.5	93.9	95.5	95.3	96.4
2	0.014	17.411	0.432	95.7	97.4	96.8	94.8	98.2	96.7	97.6	94.1	97.1	95.4	96.4	94.5	96.2
3	0.088	8.604	0.177	95.6	96.8	97.6	97.4	97.7	97.0	97.6	94.7	96.4	94.6	95.4	95.0	96.3
4	0.745	20.812	0.310	96.0	97.5	97.1	95.3	98.3	97.0	97.7	93.9	97.2	95.7	96.6	94.1	96.4
5	0.626	17.952	0.128	95.8	97.3	97.0	94.9	98.4	96.6	97.5	94.9	96.7	95.0	96.4	94.3	96.2
6	0.008	16.624	0.346	95.8	97.5	96.7	94.9	98.2	96.8	97.6	94.2	97.0	95.3	96.4	94.2	96.2
7	0.732	27.048	0.146	95.3	97.6	97.3	93.9	98.4	97.0	97.6	94.1	97.2	95.2	96.6	93.3	96.1
8	0.416	24.654	0.359	95.5	97.5	97.2	94.6	98.4	97.0	97.8	94.2	97.3	94.7	96.5	93.5	96.2
9	0.128	1.819	0.234	93.0	92.2	93.8	94.1	95.1	94.9	95.6	87.7	92.2	86.7	87.9	90.5	92.0
10	0.182	9.711	0.142	95.6	96.9	97.6	96.1	97.9	96.8	97.5	95.4	96.3	92.6	95.8	94.9	96.1
11	0.842	22.251	0.257	95.9	97.5	97.1	95.3	98.4	97.0	97.9	94.6	97.3	95.5	96.8	94.1	96.4
12	0.196	25.486	0.082	95.3	97.4	97.2	93.8	98.5	96.8	97.6	93.8	96.6	95.4	96.3	93.2	96.0
13	0.981	13.580	0.427	95.9	97.5	97.8	97.0	98.1	97.2	97.7	96.3	97.2	95.2	96.2	95.2	96.8
14	0.303	29.857	0.405	94.8	97.8	97.0	93.7	98.4	96.9	97.7	93.8	97.0	92.6	95.8	92.9	95.7
15	0.793	6.347	0.295	95.8	96.6	97.5	98.0	97.5	96.9	97.1	95.7	96.2	93.5	95.0	96.2	96.3
	0.4	15	12	95.7	97.3	96.9	95.2	98.3	96.8	97.7	94.7	97.0	94.9	96.4	94.9	96.3

3.4.4 归纳式学习算法

本文提出的算法 IHR 与直推式算法的不同还在于该算法是归纳式的，可以产生最终分类器对新来的样本进行预测。为了验证算法 IHR 在新来的数据集上的性能，我们依旧采用表 3.5.1 中描述的 12 个分类问题。对于每个分类问题中的目标领域无标签数据，我们随机采样(无放回采样)比例为 p 的数据构成新的数据集D_t^1，剩下的构成数据集D_t^2。数据D_t^1用于对初始模型h_s的精化过程，而数据D_t^2用于测试精化得到的模型h_t的泛化能力。我们记录了算法 IHR 在数据集D_t^1和D_t^2上准确率，同时也记录了不同采样比例 p 下算法 IHR 的分类性能，所有的结果如图 3.4.3 所示。



(a) 采样比例 p vs. D_t^1 上的准确率



(a) 采样比例 p vs. D_t^2 上的准确率

图 3.4.3 算法 IHR 在数据集 D_t^1， D_t^2 上的性能(%) ($\alpha = 0.4$, $\beta = 15$, $\gamma = 0.12$)

从这些结果，我们发现：

- 算法 IHR 在数据集 D_t^1 和 D_t^2 上的分类准确率几乎是一致的；
- 用于精化过程的无标签数据集 D_t^1 越大，算法的泛化能力就越好。当采样比例 $p \geq 0.6$ 时，算法在数据集 D_t^2 上的泛化能力已经达到 90%以上。

该结果表明，算法 IHR 在采样比例大于 60%时，对新来的测试数据就能表现出很好的泛化能力。

3.5. 小结

本章提出了一种归纳迁移学习的方法，解决目标领域中都是无标签数据的情况。首先我们分析了直推式迁移学习方法 — 桥接精化算法，发现样本类别比例在精化过程中的漂移，以及该算法在处理不平衡数据时的局限性。然后在给出类别先验的情况下，我们提出了归一化的方法来解决该问题。第二，提出了一种基于混合正则化的归纳迁移学习方法，包括流形正则化、熵正则化以及期望正则化等三种正则化准则。与直推式迁移学习方法比较，该算法框架可以得到最终分类器对新来的样本进行直接预测，另外，在实际数据集中的分类结果也验证了我们提出的算法的有效性。

但也看到本章提出的基于混合正则化的方法，包含多个正则化项，参数较多，在实际中调节参数并不是很容易，因此下一章我们进一步提出一种基于非负矩阵分解的跨领域分类学习方法，模型更加优越。

表 3.5.1 算法性能评价中的数据描述

数据集	源领域数据 D_s	目标领域数据 D_t	类别比例 r^p
问题 1	comp.graphics, sci.electronics, comp.os.ms-windows.misc, sci.crypt	comp.sys.mac.hardware, sci.med, comp.sys.ibm.pc.hardware, comp.windows.x, sci.space	0.60
问题 2	rec.autos, talk.politics.guns, rec.motorcycles, talk.politics.misc	rec.sport.baseball, rec.sport.hockey, talk.politics.mideast	0.68
问题 3	rec.autos, sci.med, rec.sport.baseball, sci.space	rec.motorcycles, rec.sport.hockey, sci.crypt	0.67
问题 4	rec.autos, sci.med, rec.sport.baseball, sci.space	rec.motorcycles, rec.sport.hockey, sci.electronics	0.67
问题 5	talk.religion.misc, talk.politics.mideast, comp.sys.mac.hardware, comp.graphics	comp.windows.x, comp.sys.ibm.pc.hardware, talk.politics.guns, talk.politics.misc, comp.os.ms-windows.misc	0.63
问题 6	comp.graphics, rec.sport.hockey, comp.sys.ibm.pc.hardware, rec.motorcycles	comp.windows.x, rec.autos, comp.os.ms-windows.misc	0.66
问题 7	comp.graphics, rec.sport.hockey, comp.sys.ibm.pc.hardware, rec.motorcycles	comp.windows.x, rec.sport.baseball, comp.os.ms-windows.misc	0.66
问题 8	sci.electronics, talk.politics.misc, sci.med, talk.religion.misc	sci.crypt, talk.politics.guns, sci.space	0.68
问题 9	sci.electronics, talk.politics.misc, sci.med, talk.religion.misc	sci.crypt, talk.politics.mideast, sci.space	0.68
问题 10	comp.os.ms-windows.misc, sci.crypt, comp.graphics, sci.electronics	comp.sys.ibm.pc.hardware, sci.med, comp.sys.mac.hardware, comp.windows.x	0.75
问题 11	comp.graphics, sci.med, comp.windows.x, sci.space	comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, sci.crypt	0.75
问题 12	sci.crypt, talk.politics.guns, sci.electronics	sci.med, talk.politics.mideast, talk.politics.misc, talk.religion.misc	0.30

第四章　一种有效挖掘词特征聚类与文档类别关联关系的迁移学习算法

4.1. 引言

在迁移学习文本分类中，源领域数据与目标领域数据在原始词特征上分布不一致，也就是说它们可能会采用不同的词特征来表示同一个语义概念。但我们发现不同的领域数据，其词特征聚类(又称词特征概念)与文档类别(又称文档聚类)之间的关联关系可能是一样的。比如，表示词特征概念“Computer Science”的词有‘hardware’，‘software’，‘program’，‘programmer’，‘disks’以及‘rom’等，但是这些词在不同的领域中可能频率相差很大。在关于硬件公司的新闻网页中，‘hardware’，‘disks’以及‘rom’可能是高频词，而相反，在关于软件公司的新闻网页中，‘software’，‘program’，以及‘programmer’更可能是高频词。因此不同的领域表示同一个概念的词特征差异很大，这就会导致用原始特征训练得到的分类器可能是不可靠的。如果我们能够找出各个领域的词特征概念，并用它们来预测样本的类别，那么就会比直接用原始特征要可靠和有效。从上面的例子可以看到，一个网页不管是来自于哪一个领域，只要其包含特征概念“Computer Science”，那么该网页就是属于计算机相关的文档类。基于此，词特征概念与文档类别的之间的关联关系可能是领域无关的，这就使得我们可以利用它做桥梁把知识从源领域迁移到目标领域，实现对目标领域数据的准确预测。

直观的例子。为了进一步说明本文的动机，我们扩展了 Li 等人[Li, 2008]文章的例子，如图 4.1.1 所示。其中图 4.1.1(a)部分给出了 4 个构造的句子，D1 和 D2 属于类别信息检索(IR)，而 D3，D4 属于类别计算机视觉(CV)。假设 D1 和 D3 是来自于有标签的源领域数据，D2 和 D4 是来自无标签目标领域。去掉停用词，所有的文档都可以表示成原始词特征上的向量，如图 4.1.1(b)部分。可以看到在原始特征上，用 *k*-紧邻算法做分类预测，不能对无标签数据 D2 和 D4 进行预测。但是我们发现如果这些词特征可以聚成一些聚类，如图 4.1.1(c)部分所示，比如词特征概念“Learning”包括词特征“Clustering”和“Classification”。不难发现，如果以词特征概念作为特征，我们可以正确地预测 D2 和 D4。图 4.1.1(d)部分给出了源领域与目标领域中词特征聚类和文档类别之间的共现矩阵，可以看到两个共现矩阵是一样的。这个直观的例子表明，基于词特征概念的分类比基于原始词特征更加稳定、可靠，即不同的领域可能共享相同的词特征聚类与文档类别之间的关联关系，因此我们可以利用它为桥梁把源领域中的知识迁移到目标领域。

本章基于非负矩阵三因子分解算法，提出了一种有效挖掘词特征聚类与文档类别关联关系的跨领域学习算法。下一节中，我们简单介绍矩阵分解技术，以及形式化本章所提出的学习框架。

D1: An Algorithm for Hyperlink Clustering
D2: Algorithms for Webpage Classification
D3: Texture Clustering Algorithms
D4: An Algorithm for Illumination Classificaiton

(a) A Synthetic Dataset

		Learning		Graphics		Web	
		w1	w2	w3	w4	w5	w6
		Clustering	Classification	Illumination	Texture	Webpage	Hyperlink
IR	D1	1	0	0	0	0	1
	D2	0	1	0	0	1	0
Vision	D3	1	0	0	1	0	0
	D4	0	1	1	0	0	0

(b) Vector Representation for Dataset Based on Raw Words

	Learning	Graphics	Web
D1	1	0	1
D2	1	0	1
D3	1	1	0
D4	1	1	0

(c) Vector Representation for Dataset Based on Concepts

		Learning	Graphics	Web	
Source Domain	IR	1	0	1	S1
	Vision	1	1	0	
Target Domain	IR	1	0	1	S2
	Vision	1	1	0	

(d) The CO-Occurrence of Word Cluster and Document Class on Source and Target Domains

图 4.1.1 一个直观的例子。基于词特征概念的分类比基于原始词特征更加稳定、可靠，即不同的领域可能共享相同的词特征聚类与文档类别之间的关联关系

4.2. 矩阵分解技术和问题形式化

4.2.1 矩阵分解技术

非负矩阵分解算法(NMF)已经广泛应用于维度降低，模式识别，分类以及聚类[Ding, 2006, Lee, 2001, Guillamet, 2002, Guillamet, 2003, Sha, 2003, Wang, 2008, Li, 2009a]等领域。Lee 和 Seung[Lee, 2001]提出非负矩阵分解算法分解多变量数据，并给出两种乘法算法(Multiplucative Algorithm)进行求解。他们还利用一个辅助函数证明了这两种求解算法的单调收敛性，后续有很多工作扩展该模型到不同的应用中。Guillamet 等人[Guillamet, 2003]扩展非负矩阵算法到加权的情况(WNMF)来提高数据表示能力，实验表明 WNMF 比 NMF 和主成分分析(PCA)，在图像分类方面取得了显著提高。Ding 等人[Ding, 2006]首先分析了 2-因子分解和 3-因子分解之间的关系，然后提出了基于正交非负矩阵分解的聚类算法，并从实验上表明双向正交非负矩阵分解可以有效地对输入数据矩阵同时进行行和列进行聚类。Wang 等人[Wang, 2008]提出了一种基于矩阵分解的半监督聚类方法。

事实上，概率隐性语义分析(Probabilistic Latent Semantic Analysis, PLSA)[Hofmann, 2001]和隐性狄利克雷分配模型(Latent Dirichlet Location, LDA)[Blei, 2003]也可以看成是一种非负矩阵分解方法[Gaussier, 2005]。以上工作与我们的工作不同的是，他们都假设词特征聚类与文档聚类具有相同的语义空间，即设置相同的隐含主题[Hofmann, 2001]。本文提出的方法 MTrick，词特征聚类和文档聚类可以有不同的语义空间，这样词特征聚类和文档类别之间的关系就可以很好地表示出来。

近年来，把矩阵分解应用到迁移学习中也得到了初步的研究。Li 等人[Li, 2009, Li, 2010]提出了两阶段的基于约束非负矩阵三因子分解的情感分类(Sentiment Classification)方法，首先通过源领域的学习把文档的类别信息迁移到词特征上，然后把情感分类信息通过词特征迁移到目标领域的文档。而且文献[Li, 2010]中的方法需要目标领域中有少部分有标签的样本数据，而本章提出的方法不需要任何目标领域中的样本信息。这两个方法的缺点在于，他们假设源领域和目标领域共享相同的词特征上的信息，但是就像上面提到的，不同领域可能用不同的词特征来表示同一个词特征概念。Li 等人[Li, 2009a]提出了一种处理跨领域协同过滤(Collaborative Filtering)的新方法，他们单独分解源领域和目标领域的数据矩阵。也就是说先从第一步学习得到共享信息，然后作为第二步的输入，从而实现知识的迁移。本章中的方法是一个联合优化框架，同时优化分解源领域和目标领域的数据矩阵，实验表明这种联合优化框架，性能更加优越。

矩阵的三因子分解最基本的公式如下：

$$\mathbf{X}_{m\times n} = \mathbf{F}_{m\times k_1}\mathbf{S}_{k_1\times k_2}\mathbf{G}^{\mathrm{T}}_{n\times k_2} \tag{4.1}$$

其中 m 和 n 分别为特征与样本的个数，k_1 和 k_2 分别为词特征聚类和文档类别的个数，$\mathbf{F}_{m\times k_1}$ 表示词特征的后验类别概率，$\mathbf{G}_{n\times k_2}$ 表示文档的后验类别概率，$\mathbf{S}_{k_1\times k_2}$ 则可以看成数据矩阵 $\mathbf{X}$ 的压缩。实际上 $\mathbf{S}_{k_1\times k_2}$ 可以看成是特征聚类与文档类别层面上的联合概率分布矩阵。前面提到在 Li 等人[Li, 2009]的工作中，他们把词的后验类别概率信息 $\mathbf{F}_{m\times k_1}$ 作为桥信息，从而实现知识从源领域到目标领域的迁移。我们发现，不同领域的词特征后验类别概率信息 $\mathbf{F}_{m\times k_1}$ 可能是不同的。比如电子产品这个类别，源领域数据中可能出现了电脑这个词特征，但目标领域却出现了打印机这个词特征，显然这两个领域的词特征后验类别概率是不同的。进一步可以发现，基于词特征聚类以及文档类别层面上的联合概率矩阵 $\mathbf{S}$ 是可以不变的，因此源领域数据和目标领域完全可以通过共享词特征概念和类别层面的信息来实现知识的迁移。基于以上观测，我们的研究任务是开发利用词聚类与文档类别的联合关系 $\mathbf{S}$，来实现跨领域的文档分类。

4.2.1 问题形式化

对于源领域中的联合概率分布矩阵 $\mathbf{X}_s \in \mathbb{R}_+^{m\times n_s}$ ($\mathbf{X}_s = \dfrac{\mathbf{Y}_s}{\sum_{i,j}\mathbf{Y}_{s_{(ij)}}}$，$\mathbf{Y}_s$ 是源领域中的词-文档频率共现矩阵，以下 $\mathbf{X}_t$ 类似定义)，其中 m 是词特征个数，n_s 是源领域中样本的个数，可以形式化以下优化问题：

$$\begin{aligned}&\min_{\mathbf{F}_s,\mathbf{S}_s,\mathbf{G}_s} \|\mathbf{X}_s-\mathbf{F}_s\mathbf{S}_s\mathbf{G}_s^{\mathrm{T}}\|^2+\frac{\alpha}{n_s}\cdot\|\mathbf{G}_s-\mathbf{G}_0\|^2,\\ &s.t.\quad \sum_{j=1}^{k_1}\mathbf{F}_{s_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{s_{(ij)}}=1,\end{aligned} \tag{4.2}$$

其中 α 是平衡参数，$\mathbf{G}_0$ 包含源领域中的样本标签信息，当第 i 个样本属于第 j 时，$\mathbf{G}_{0_{(ij)}}=1$；否则 $\mathbf{G}_{0_{(ik)}}=0(k\neq j)$。在这个优化问题中，$\mathbf{G}_0$ 作为监督信息来优化得到最终后验概率分布矩阵 $\mathbf{F}_s,\mathbf{G}_s,\mathbf{S}_s$ (它们的含义见上一小节引言部分)。

同样对于目标领域中的联合概率分布矩阵 $\mathbf{X}_t \in \mathbb{R}_+^{m\times n_t}$，$n_t$ 是目标领域中样本的个数，可以形式化优化问题如下：

$$\begin{aligned}&\min_{\mathbf{F}_t,\mathbf{G}_t} \|\mathbf{X}_t-\mathbf{F}_t\mathbf{S}_0\mathbf{G}_t^{\mathrm{T}}\|^2\\ &s.t.\quad \sum_{j=1}^{k_1}\mathbf{F}_{t_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{t_{(ij)}}=1,\end{aligned} \tag{4.3}$$

其中词特征聚类与文档类别之间的关系矩阵 $\mathbf{S}_0$ 由求解式子(4.2)得到。当得到 $\mathbf{F}_t,\mathbf{G}_t$ 后，就可以通过以下式子得到样本的预测类别，

$$index_i=\arg\max_j \mathbf{G}_{t_{(ij)}} \tag{4.4}$$

如果联合优化式子(4.2)和(4.3)，可以得到以下优化问题：

$$\begin{aligned}&\min_{\mathbf{F}_s,\mathbf{G}_s,\mathbf{S},\mathbf{F}_t,\mathbf{G}_t} \|\mathbf{X}_s-\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\|^2+\frac{\alpha}{n_s}\cdot\|\mathbf{G}_s-\mathbf{G}_0\|^2+\beta\cdot\|\mathbf{X}_t-\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\|^2,\\ &s.t.\quad \sum_{j=1}^{k_1}\mathbf{F}_{s_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{s_{(ij)}}=1,\sum_{j=1}^{k_1}\mathbf{F}_{t_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{t_{(ij)}}=1,\end{aligned} \tag{4.5}$$

其中参数 $\alpha\geq 0$，$\beta\geq 0$ 是平衡因子，关系矩阵 $\mathbf{S}$ 是两个联合概率分解的共享因子，这样 $\mathbf{S}$ 其实是把知识从源领域到目标领域迁移的桥梁。下面主要关注如何求解优化问题(4.5)，两步优化问题(4.2)和(4.3)可以类似得到求解。

4.3. 优化问题求解算法

对于优化问题(4.5)，我们提出一个迭代算法进行求解。根据矩阵的运算规则，式子(4.5)可以重新写成式子(4.6)，

$$\begin{aligned}
&L(\mathbf{F}_s,\mathbf{G}_s,\mathbf{S},\mathbf{F}_t,\mathbf{G}_t)\\
&=Tr(\mathbf{X}_s^{\mathrm{T}}\mathbf{X}_s-2\mathbf{X}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}+\mathbf{G}_s\mathbf{S}^{\mathrm{T}}\mathbf{F}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}})\\
&+\frac{\alpha}{n_s}\cdot Tr(\mathbf{G}_s\mathbf{G}_s^{\mathrm{T}}-2\mathbf{G}_s\mathbf{G}_0^{\mathrm{T}}+\mathbf{G}_0\mathbf{G}_0^{\mathrm{T}})+\beta\cdot Tr(\mathbf{X}_t^{\mathrm{T}}\mathbf{X}_t-2\mathbf{X}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}+\mathbf{G}_t\mathbf{S}^{\mathrm{T}}\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}),\\
&s.t.\quad \sum_{j=1}^{k_1}\mathbf{F}_{s_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{s_{(ij)}}=1,\sum_{j=1}^{k_1}F_{t_{(ij)}}=1,\sum_{j=1}^{k_2}G_{t_{(ij)}}=1.
\end{aligned}\tag{4.6}$$

对目标函数L求偏导，

$$\frac{\partial L}{\partial \mathbf{F}_s}=-2\mathbf{X}_s\mathbf{G}_s\mathbf{S}^{\mathrm{T}}+2\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s\mathbf{S}^{\mathrm{T}},\tag{4.7}$$

$$\frac{\partial L}{\partial \mathbf{G}_s}=-2\mathbf{X}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}+2\mathbf{G}_s\mathbf{S}^{\mathrm{T}}\mathbf{F}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}+\frac{2\alpha}{n_s}\cdot(\mathbf{G}_s-\mathbf{G}_0),\tag{4.8}$$

$$\frac{\partial L}{\partial \mathbf{S}}=-2\mathbf{F}_s^{\mathrm{T}}\mathbf{X}_s\mathbf{G}_s+2\mathbf{F}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s-2\beta\cdot\mathbf{F}_t^{\mathrm{T}}\mathbf{X}_t\mathbf{G}_t+2\beta\cdot\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\mathbf{G}_t,\tag{4.9}$$

$$\frac{\partial L}{\partial \mathbf{F}_t}=-2\beta\cdot\mathbf{X}_t\mathbf{G}_t\mathbf{S}^{\mathrm{T}}+2\beta\cdot\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\mathbf{G}_t\mathbf{S}^{\mathrm{T}},\tag{4.10}$$

$$\frac{\partial L}{\partial \mathbf{G}_t}=-2\beta\cdot\mathbf{X}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S}+2\beta\cdot\mathbf{G}_t\mathbf{S}^{\mathrm{T}}\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S}.\tag{4.11}$$

目标函数L是一个非凸函数，因此很难用非线性优化方法得到全局最优解。我们提出一个迭代算法可以得到局部最优解，各个变量的迭代式子如下，

$$\mathbf{F}_{s_{(ij)}}\leftarrow\mathbf{F}_{s_{(ij)}}\cdot\sqrt{\frac{(\mathbf{X}_s\mathbf{G}_s\mathbf{S}^{\mathrm{T}})_{(ij)}}{(\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s\mathbf{S}^{\mathrm{T}})_{(ij)}}},\tag{4.12}$$

$$\mathbf{G}_{s_{(ij)}}\leftarrow\mathbf{G}_{s_{(ij)}}\cdot\sqrt{\frac{(\mathbf{X}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}+\frac{\alpha}{n_s}\cdot\mathbf{G}_0)_{(ij)}}{(\mathbf{G}_s\mathbf{S}^{\mathrm{T}}\mathbf{F}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}+\frac{\alpha}{n_s}\cdot\mathbf{G}_s)_{(ij)}}},\tag{4.13}$$

$$\mathbf{F}_{t_{(ij)}}\leftarrow\mathbf{F}_{t_{(ij)}}\cdot\sqrt{\frac{(\mathbf{X}_t\mathbf{G}_t\mathbf{S}^{\mathrm{T}})_{(ij)}}{(\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\mathbf{G}_t\mathbf{S}^{\mathrm{T}})_{(ij)}}},\tag{4.14}$$

$$\mathbf{G}_{t_{(ij)}}\leftarrow\mathbf{G}_{t_{(ij)}}\cdot\sqrt{\frac{(\mathbf{X}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S})_{(ij)}}{(\mathbf{G}_t\mathbf{S}^{\mathrm{T}}\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S})_{(ij)}}},\tag{4.15}$$

然后通过以下式子归一化$\mathbf{F}_s,\mathbf{G}_s,\mathbf{F}_t,\mathbf{G}_t$，使得其满足约束条件

$$\mathbf{F}_{s_{(i\cdot)}} \leftarrow \frac{\mathbf{F}_{s_{(i\cdot)}}}{\sum_{j=1}^{k_1}\mathbf{F}_{s_{(ij)}}}, \tag{4.16}$$

$$\mathbf{G}_{s_{(i\cdot)}} \leftarrow \frac{\mathbf{G}_{s_{(i\cdot)}}}{\sum_{j=1}^{k_2}\mathbf{G}_{s_{(ij)}}}, \tag{4.17}$$

$$\mathbf{F}_{t_{(i\cdot)}} \leftarrow \frac{\mathbf{F}_{t_{(i\cdot)}}}{\sum_{j=1}^{k_1}\mathbf{F}_{t_{(ij)}}}, \tag{4.18}$$

$$\mathbf{G}_{t_{(i\cdot)}} \leftarrow \frac{\mathbf{G}_{t_{(i\cdot)}}}{\sum_{j=1}^{k_2}\mathbf{G}_{t_{(ij)}}}. \tag{4.19}$$

关系矩阵S的迭代更新公式如下：

$$\mathbf{S}_{(ij)} \leftarrow \mathbf{S}_{(ij)} \cdot \sqrt{\frac{(\mathbf{F}_s^{\mathrm{T}}\mathbf{X}_s^{\mathrm{T}}\mathbf{G}_s + \beta\cdot\mathbf{F}_t^{\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}\mathbf{G}_t)_{(ij)}}{(\mathbf{F}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s + \beta\cdot\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\mathbf{G}_t)_{(ij)}}}, \tag{4.20}$$

详细的迭代算法如算法 4.1 所示。

算法 4.1: 基于非负矩阵三因子分解跨领域分类方法(MTrick)

输入: 源领域的联合概率矩阵$\mathbf{X}_s \in \mathbb{R}_+^{m\times n_s}$以及其标签信息$\mathbf{G}_0$；目标领域的联合概率矩阵$\mathbf{X}_t \in \mathbb{R}_+^{m\times n_t}$，平衡参数$\alpha,\beta$的值以及误差阈值$\varepsilon>0$；迭代算法的最大迭代次数$max$。

输出: $\mathbf{F}_s,\mathbf{G}_s,\mathbf{F}_t,\mathbf{G}_t$和$\mathbf{S}$。

步骤 1：初始化矩阵$\mathbf{F}_s^{(0)},\mathbf{G}_s^{(0)},\mathbf{F}_t^{(0)},\mathbf{G}_t^{(0)}$和$\mathbf{S}^{(0)}$；

步骤 2：根据式子(4.6)计算初始目标函数值$L^{(0)}$；

步骤 3：$k:=1$；

步骤 4：根据式子(4.12)和(4.16)分别更新和归一化$\mathbf{F}_s^{(k)}$；

步骤 5：根据式子(4.13)和(4.17)分别更新和归一化$\mathbf{G}_s^{(k)}$；

步骤 6：根据式子(4.14)和(4.18)分别更新和归一化$\mathbf{F}_t^{(k)}$；

步骤 7：根据式子(4.15)和(4.19)分别更新和归一化$\mathbf{G}_t^{(k)}$；

步骤 8：根据式子(4.20)更新$\mathbf{S}^{(k)}$；

步骤 9：根据式子(4.6)计算目标函数值$L^{(k)}$，如果$|L^{(k)}-L^{(k-1)}|<\varepsilon$，则转到步骤 11；

步骤 10：$k:=k+1$，若$k<max$则转步骤 4。

步骤 11：输出$\mathbf{F}_s^{(k)},\mathbf{G}_s^{(k)},\mathbf{F}_t^{(k)},\mathbf{G}_t^{(k)}$和$\mathbf{S}^{(k)}$。

4.4. 算法收敛性分析

这一节分析提出的迭代算法的收敛性，首先分析$\mathbf{F}_s$而其它变量$\mathbf{G}_s,\mathbf{S},\mathbf{F}_t,\mathbf{G}_t$固定的情况。我们构造拉格朗日函数如下，

$$\Psi(\mathbf{F}_s) = \| \mathbf{X}_s - \mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}} \|^2 + Tr[\lambda(\mathbf{F}_s\mathbf{u}^{\mathrm{T}} - \mathbf{v}^{\mathrm{T}})(\mathbf{F}_s\mathbf{u}^{\mathrm{T}} - \mathbf{v}^{\mathrm{T}})^{\mathrm{T}}], \tag{4.21}$$

其中 $\lambda \in \mathbb{R}^{m\times m}$，$\mathbf{u} \in \mathbb{R}^{1\times k_1}$，$\mathbf{v} \in \mathbb{R}^{1\times m}$（$\mathbf{u}$ 和 $\mathbf{v}$ 是所有元素都为 1 的向量），式子的第二项是拉格朗日约束条件。

根据矩阵运算规则，$\| \mathbf{X}_s - \mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}} \|^2 = Tr(\mathbf{X}_s^{\mathrm{T}}\mathbf{X}_s - 2\mathbf{X}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}} + \mathbf{G}_s\mathbf{S}^{\mathrm{T}}\mathbf{F}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}})$，所以有

$$\frac{\partial \Psi}{\partial \mathbf{F}_s} = -2\mathbf{X}_s\mathbf{G}_s\mathbf{S}^{\mathrm{T}} + 2\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s\mathbf{S}^{\mathrm{T}} + 2\lambda\mathbf{F}_s\mathbf{u}^{\mathrm{T}}\mathbf{u} - 2\lambda\mathbf{v}^{\mathrm{T}}\mathbf{u}. \tag{4.22}$$

引理 4.1：使用迭代式子(4.23)，目标函数(4.6)将会单调下降。

$$\mathbf{F}_{s_{(ij)}} \leftarrow \mathbf{F}_{s_{(ij)}} \cdot \sqrt{\frac{(\mathbf{X}_s\mathbf{G}_s\mathbf{S}^{\mathrm{T}} + \lambda\mathbf{v}^{\mathrm{T}}\mathbf{u})_{(ij)}}{(\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s\mathbf{S}^{\mathrm{T}} + \lambda\mathbf{F}_s\mathbf{u}^{\mathrm{T}}\mathbf{u})_{(ij)}}}. \tag{4.23}$$

证明：为了证明引理 4.1，我们首先引入辅助函数的定义。

定义 4.1(辅助函数)：如果函数 $H(\mathbf{Y}, \widetilde{\mathbf{Y}})$ 是 $\Gamma(\mathbf{Y})$ 的辅助函数[Lee, 2001]，那么对于任意 $\mathbf{Y}$，$\widetilde{\mathbf{Y}}$ 它满足，

$$H(\mathbf{Y}, \widetilde{\mathbf{Y}}) \geq \Gamma(\mathbf{Y}), \quad H(\mathbf{Y}, \mathbf{Y}) = \Gamma(\mathbf{Y}), \tag{4.24}$$

然后定义，

$$\mathbf{Y}^{(t+1)} = \mathrm{argmin}_{\mathbf{Y}} H(\mathbf{Y}, \mathbf{Y}^{(t)}). \tag{4.25}$$

则有

$$\Gamma(\mathbf{Y}^{(t)}) = H(\mathbf{Y}^{(t)}, \mathbf{Y}^{(t)}) \geq H(\mathbf{Y}^{(t+1)}, \mathbf{Y}^{(t)}) \geq \Gamma(\mathbf{Y}^{(t+1)}). \tag{4.26}$$

即最小化 $H(\mathbf{Y}, \mathbf{Y}^{(t)})$（$\mathbf{Y}^{(t)}$ 固定)也就是最小化函数 Γ。这样我们可以构造函数 Ψ 的辅助函数如下，

$$\begin{aligned} H(\mathbf{F}_s, \mathbf{F}'_s) = &-2\sum_{ij}(\mathbf{X}_s\mathbf{G}_s\mathbf{S}^{\mathrm{T}})_{(ij)}\mathbf{F}'_{s_{(ij)}}(1 + \log\frac{\mathbf{F}_{s_{(ij)}}}{\mathbf{F}'_{s_{(ij)}}}) + \sum_{ij}(\mathbf{F}'_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s\mathbf{S}^{\mathrm{T}})_{(ij)}\frac{\mathbf{F}_{s_{(ij)}}^2}{\mathbf{F}'_{s_{(ij)}}} \\ &+ \sum_{ij}(\lambda\mathbf{F}'_s\mathbf{u}^{\mathrm{T}}\mathbf{u})_{(ij)}\frac{\mathbf{F}_{s_{(ij)}}^2}{\mathbf{F}'_{s_{(ij)}}} - 2\sum_{ij}(\lambda\mathbf{v}^{\mathrm{T}}\mathbf{u})_{(ij)}\mathbf{F}'_{s_{(ij)}}(1 + \log\frac{\mathbf{F}_{s_{(ij)}}}{\mathbf{F}'_{s_{(ij)}}}). \end{aligned} \tag{4.27}$$

显然，当 $\mathbf{F}'_s = \mathbf{F}_s$ 时，$H(\mathbf{F}_s, \mathbf{F}'_s) = \Psi(\mathbf{F}_s)$ 成立。而且可以根据文献[Ding, 2006]的证明方法得到 $H(\mathbf{F}_s, \mathbf{F}'_s) \geq \Psi(\mathbf{F}_s)$。因此在固定 $\mathbf{F}'_s$ 的情况下，最小化 $H(\mathbf{F}_s, \mathbf{F}'_s)$，其偏导数如下，

$$\frac{\partial H(\mathbf{F}_s,\mathbf{F'}_s)}{\partial \mathbf{F}_{s_{(ij)}}} = -2(\mathbf{X}_s\mathbf{G}_s\mathbf{S}^{\mathrm{T}})_{(ij)}\frac{\mathbf{F'}_{s_{(ij)}}}{\mathbf{F}_{s_{(ij)}}} + 2(\mathbf{F'}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s\mathbf{S}^{\mathrm{T}} + \lambda\mathbf{F'}_s\mathbf{u}^{\mathrm{T}}\mathbf{u})_{(ij)}\frac{\mathbf{F}_{s_{(ij)}}}{\mathbf{F'}_{s_{(ij)}}} - 2(\lambda\mathbf{v}^{\mathrm{T}}\mathbf{u})_{(ij)}\frac{\mathbf{F'}_{s_{(ij)}}}{\mathbf{F}_{s_{(ij)}}}. \tag{4.28}$$

由偏导为 0 得到，

$$\Rightarrow \mathbf{F}_{s_{(ij)}} = \mathbf{F'}_{s_{(ij)}} \cdot \sqrt{\frac{(\mathbf{X}_s\mathbf{G}_s\mathbf{S}^{\mathrm{T}} + \lambda\mathbf{v}^{\mathrm{T}}\mathbf{u})_{(ij)}}{(\mathbf{F'}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\mathbf{G}_s\mathbf{S}^{\mathrm{T}} + \lambda\mathbf{F'}_s\mathbf{u}^{\mathrm{T}}\mathbf{u})_{(ij)}}}. \tag{4.29}$$

引理 1 得证。

还有一个问题就是如何确定拉格朗日乘子λ。其实这里λ的作用是使得所求得的解满足约束条件，因此采用一个简单的归一化技术来满足约束，从而可以忽略λ。我们用式子(4.12)和(4.16)大致等价式子(4.29)，而省略掉λ的求解。对于求解变量$\mathbf{G}_s,\mathbf{F}_t,\mathbf{G}_t,\mathbf{S}$的收敛性，可以类似分析。最后我们得到定理 4.1。

定理 4.1(收敛性)：通过算法 4.1 中的迭代，目标函数(4.6)单调下降且收敛。

根据引理 4.1 以及[Lee, 2001]中的乘法算法迭代准则，算法每次迭代都会降低目标函数，而且目标函数具有下界 0，所以保证算法最终收敛。

4.5. 实验过程和结果

本节将展示大量的实验结果来证明本章所提出算法 MTrick 的有效性，主要集中在两类问题和三类问题。实验中，文档聚类的个数设置为实际样本个数。

4.5.1 实验数据

表 4.5.1 数据集 20*Newsgroups* 中的四大类，以及对应的四个小类

大类	每个大类所对应的四个小类
comp	*comp. graphics*, *comp.sys.mac.hardware*, *comp.sys.ibm.pc.hardware*, *comp.os.ms-windows.misc*
rec	*rec.autos*, *rec.motorcycles*, *rec.baseball*, *rec.hockey*
sci	*sci.crypt*, *sci.med*, *sci.electronics*, *sci.space*
talk	*talk.politics.guns*, *talk.politics.mideast*, *talk.politics.misc*, *talk.religion.misc*

数据集 20*Newsgroups*[6]是评价文本分类的标准数据集，包括大约 20,000 个文档，根据不同的主题被划分为 20 个小类，每个小类包含的文档数差不多。由于某些主题比较相

[6] http://people.csail.mit.edu/jrennie/20Newsgroups/

近又可以组成一些大类，比如 *sci.srypt*，*sci.electronics*，*sci.med* 和 *sci.space* 这四个小类构成一个大类 *sci*。本章选用该数据集的四个大类作为我们的实验数据，如表 4.5.1 所示。这样该数据集包含两层结构，每个大类下面都有对应的四个小类。

两类分类问题：我们选择三个大类 *sci*，*talk*，*rec* 来构造两类分类问题，因此任意选择两大类数据(分别为正类和负类)，可以构造三个数据集的两类分类实验，包括 *sci vs. talk*，*rec vs. sci* 和 *rec vs. talk*。我们构造两类问题如下，以数据集 *sci vs. talk* 为例，分别从大类 *sci* 以及 *talk* 各选一个小类构成源领域数据，而同样从两个大类中各选一个小类构成目标领域数据，注意源领域数据和目标领域数据的文档集不相交。这样构造的分类问题符合迁移学习问题，(1) 源领域数据与目标领域数据分布不相同，因为它们来自于不同的小类，即不同的主题；(2) 源领域数据与目标领域数据是相关的，因为它们来自于相同的大类。通过这样的构造方法，每个数据集可以构造 144($P_4^2 \cdot P_4^2$)个两类分类问题，三个数据集总共144×3 个分类问题。

三类分类问题：构造三类分类问题与两类问题类似，随机选择三个大类，可以构造 4 个数据集，包括 *comp vs. rec vs. sci*，*comp vs. rec vs. talk*，*comp vs. sci vs. talk* 和 *rec vs. sci vs. talk*。对于每个数据集可以构造 1728($P_4^2 \cdot P_4^2 \cdot P_4^2$)个三类分类问题，在本节的实验中，我们随机抽取 100 个分类问题，4 个数据集总共有100×4 。

为了进一步验证我们的算法，还在数据集 *Reuters-21578*[7] 上做了比较实验，并且采用了 Gao 等人[Gao, 2008]文章中构造的分类任务。

4.5.2 比较算法和实现细节

比较算法：与MTrick比较的算法包括：(1) 监督学习方法逻辑回归(LG)[Davie, 2000]，LibSVM[Chang, 2001]以及支持向量机算法(SVM)[Boser, 1992]；(2) 半监督学习算法(TSVM)[Joachims, 1999a]；(3) 跨领域分类方法，Dai 等人[Dai, 2007]提出的基于联合聚类方法 CoCC 以及 Gao 等人[Gao, 2008]提出的局部加权集成方法 LWE；另外还比较了本章中提出的两步优化方法，即分两步分解数据矩阵的式子(4.2)和(4.3)，称为 MTrick0。

实现细节：我们提出的求解优化框架算法是一个迭代算法，变量$\mathbf{F}_s, \mathbf{F}_t, \mathbf{G}_s, \mathbf{G}_t, \mathbf{S}$ 的初始化如下，(1) $\mathbf{F}_s$ 和$\mathbf{F}_t$ 初始化为 PLSA[Hofmann, 1999]的聚类结果，即$\mathbf{F}_{s_{(ij)}}$ 和$\mathbf{F}_{t_{(ij)}}$ 都被初始化为 PLSA[8](在所有的数据上做 PLSA，包括源领域数据和目标领域数据)的输出 $P(z_j \mid w_i)$；(2) $\mathbf{G}_s$ 被初始化为其真实的样本类别信息；(3) $\mathbf{G}_t$ 则被初始化为从源领域数据学习得到的监督分类模型在目标领域数据上的预测结果，实验中采用监督学习方法逻

[7] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[8] 其 Matlab 实现见，http://www.kyb.tuebingen.mpg.de/bs/people/pgehler/code/index.html

辑回归(LG)[Davie, 2000]；(4) 关联关系矩阵**S**中的每个元素赋值为同一个数值，但满足 $\sum_{i,j}\mathbf{S}_{(ij)}=1$。

由于 PLSA 算法有一个随机初始化的过程，因此本节实验中列出的结果是进行三次实验的平均值。我们采用 $tf-idf$ 特征来构造词—文档矩阵**Y**，然后转换为概率分布矩阵**X**作为算法的输入，另外采用文档频率值 15 来选择词特征。实验中采用的参数 $\alpha=1$，$\beta=1.5$，$\varepsilon=10^{-11}$，$max=100$ 以及词特征聚类个数 $k_1=50$。对于比较算法 CoCC[Dai, 2007]和 LWE[Gao, 2008]都采用他们文章中推荐的参数。

4.5.3 结果比较

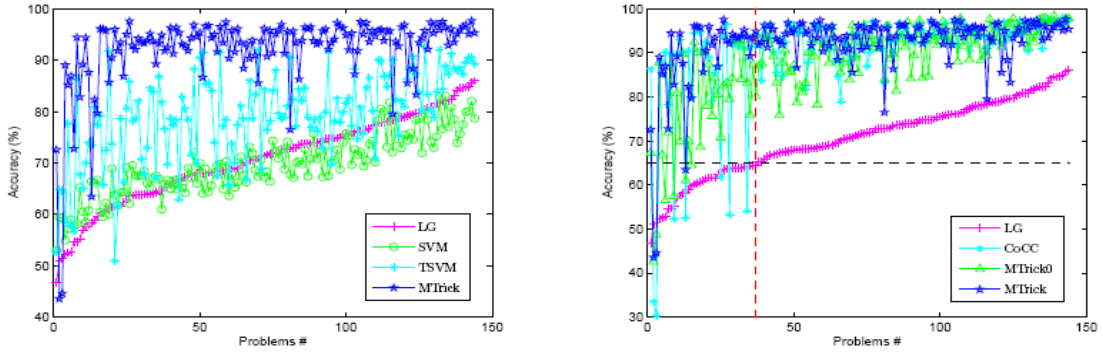
两类分类实验：对于两类分类实验，我们比较了 *20Newsgroup* 上144×3个分类问题以及 *Reuters-21578* 上 3 个分类问题。*20Newsgroup* 上的实验结果如图 4.5.1，图 4.5.2 和图 4.5.3 所示。每个数据集的 144 个问题按照 LG 算法准确率的升序排列，这也一定程度上反映这些分类问题的迁移学习难度，LG 准确率越小，迁移学习难度越大，反之越容易。

从实验结果可以看到，

— MTrick 大大优越于监督学习算法 LG 和 SVM，这表明监督学习方法并不能很好地处理迁移学习问题；

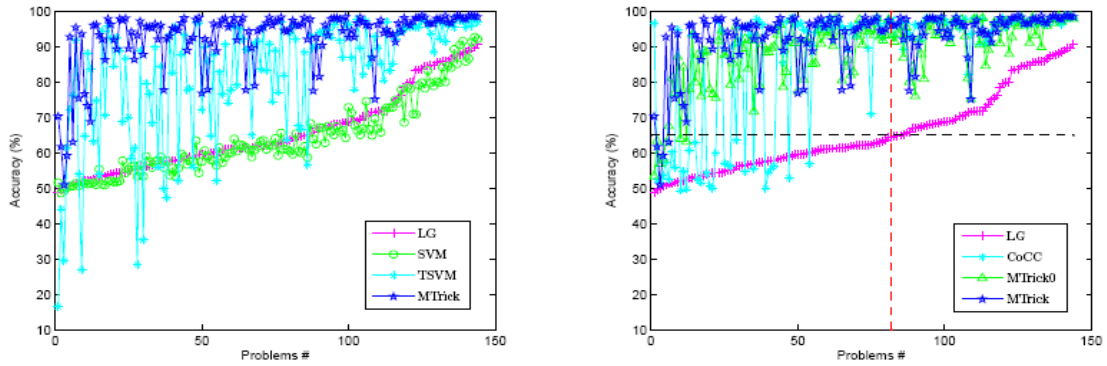
— MTrick 同样也比半监督学习方法 TSVM 表现得好，因为半监督学习方法也假设目标领域无标签数据与源领域标签数据服从独立同分布条件；

— 从图 4.5.1(b)，图 4.5.2(b)和图 4.5.3(b)可以看到，当 LG 算法表现较好时(LG 准确率高于 65%)，算法 MTrick 和 CoCC 表现类似；而当 LG 算法表现较差时(LG 准确率低于 65%)，MTrick 却比跨领域分类方法 CoCC 好很多，这表明 MTrick 更能处理学习问题较难的情况，具有更强的迁移学习能力；



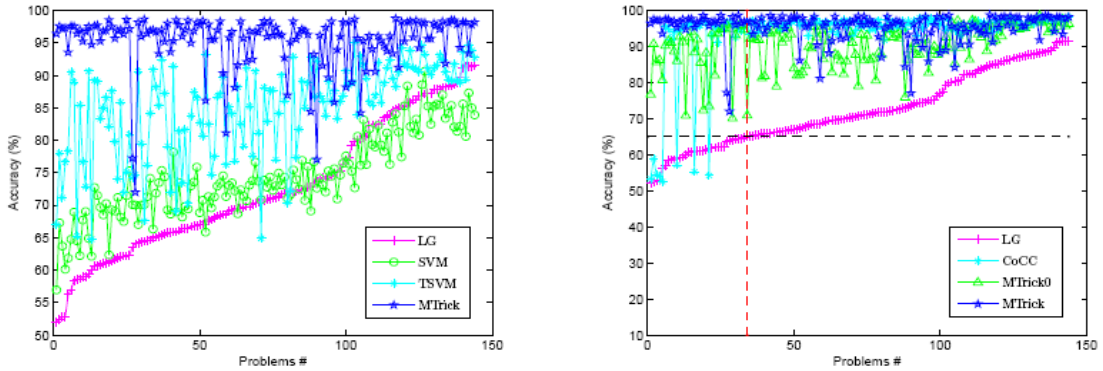
(a) MTrick vs. LG, SVM, TSVM on data set *sci vs.* (b) MTrick vs. MTrick0, CoCC on data set *sci vs. talk*
talk

图 4.5.1 算法 MTrick，LG，SVM，TSVM，CoCC 以及 MTrick0 在数据集 *sci vs talk* 上的比较



(a) MTrick vs. LG, SVM, TSVM on data set *rec vs. sci* (b) MTrick vs. MTrick0, CoCC on data set *rec vs. sci*

图 4.5.2 算法 MTrick，LG，SVM，TSVM，CoCC 以及 MTrick0 在数据集 *rec vs sci* 上的比较



(a) MTrick vs. LG, SVM, TSVM on data set *rec vs.* (b) MTrick vs. MTrick0, CoCC on data set *rec vs. talk*

图 4.5.3 算法 MTrick，LG，SVM，TSVM，CoCC 以及 MTrick0 在数据集 *rec vs talk* 上的性能(%)比较

— MTrick 也比两步优化算法 MTrick0 好，这说明联合优化算法比分两步优化要优越，更能找到最优解。

表 4.5.2 算法 MTrick，LG，SVM，TSVM，CoCC 以及 MTrick0 在三个数据集(两类分类问题)上的平均准确率(%)比较

Data sets		LG	SVM	TSVM	CoCC	MTrick0	MTrick
sci vs. talk	*L*	59.09	62.88	72.13	81.09	76.90	**86.52**
	R	74.21	71.70	81.58	93.41	91.28	**93.71**
	Total	70.64	69.62	79.35	90.50	87.88	**92.01**
rec vs. sci	*L*	57.42	56.78	75.73	79.69	85.39	**90.44**
	R	75.76	73.48	91.66	**96.18**	93.50	95.53
	Total	65.57	64.20	82.81	87.02	88.99	**92.70**
rec vs. talk	*L*	60.28	67.64	79.82	85.62	87.62	**95.57**
	R	76.29	76.52	86.52	**96.14**	91.19	95.09
	Total	72.49	74.42	84.94	93.66	90.35	**95.21**

为了更加直观的体现所提出的算法 MTrick 的优越性，表 4.5.2 给出了所有算法在三个数据集上的平均准确率比较，表中的 *L*，*R* 分别表示 LG 准确率低于和高于 65%的分类问题的平均值，而 *Total* 表示所有 144 个问题的平均值。可以清楚地看到 MTrick 比所有的比较算法算法好很多，特别是当分类问题较难的情况(LG 准确率低于 65%)。*t* 测试在 95%的置信度上也表明，MTrick 大大优越于所有比较的算法。

我们还在数据集 *Reuters-21578* 上对算法 MTrick，LG，SVM，TSVM，CoCC 以及 LWE 进行了比较，其任务描述和实验结果见表 4.5.3 和表 4.5.4。表 4.5.4 中的实验结果再一次验证了 MTrick 的有效性。

表 4.5.3 在数据集 *Reuters-21578* 上的分类任务描述

Data sets	Source-Domain D_s	Target-Domain D_t
orgs vs. people	document from a set of sub-categories	document from a different set of sub-categories
orgs vs. place		
people vs. place		

表 4.5.4 算法 MTrick，LG，SVM，TSVM，CoCC 以及 LWE 在数据集 *Reuters-21578* 上的性能比较(%)

Data Sets	LG	SVM	TSVM	CoCC	LWE	MTrick
orgs vs. people	74.92	74.25	73.80	79.79	79.67	**80.80**
orgs vs. place	71.91	69.99	69.89	74.18	73.04	**76.77**
people vs. place	58.03	59.05	58.43	66.94	68.52	**69.02**

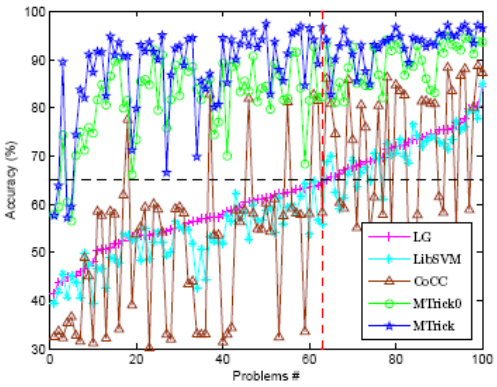
三类分类实验：为了说明本章提出的算法 MTrick 可以直接处理多类分类问题，我们构造了 4 个数据集(见实验数据小节中的描述)，总共 100×4 个分类问题。所有的实验结果都列在表 4.5.5 和图 4.5.4。

表 4.5.5 算法 MTrick，LG，LibSVM，CoCC 以及 MTrick0 在四个数据集(三类分类问题)上的平均准确率(%)比较

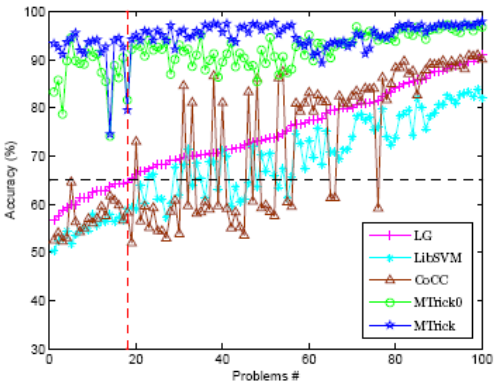
	comp vs. rec vs. sci			*comp vs. rec vs. talk*			*comp vs. sci vs. sci*			*rec vs. sci vs. talk*		
	L	*R*	*Total*	*L*	*R*	*Total*	*L*	*R*	*Total*	*L*	*R*	*Total*
LG	55.53	72.49	61.81	61.54	77.52	74.65	59.44	69.10	62.63	64.88	70.34	56.45
LibSVM	52.35	71.11	59.29	55.17	71.19	68.31	55.51	64.76	58.56	64.03	64.91	52.31
CoCC	50.53	72.46	58.65	56.99	74.47	71.33	35.12	52.32	40.80	46.30	57.58	48.33
MTrick0	81.20	89.53	84.28	87.14	92.91	91.87	76.95	88.64	80.81	81.00	90.78	81.83
MTrick	**87.31**	**93.22**	**89.50**	**91.41**	**95.51**	**94.77**	**80.12**	**92.12**	**84.08**	**89.48**	**93.58**	**88.60**

比较的算法中 LibSVM，MTrick0 可以直接处理多类数据，而 LG 和 CoCC 采用一对多法(one-versus-rest)。所有的实验结果表明 MTrick 可以直接处理多类数据，而且分类性能比算法 LG，LibSVM，CoCC，MTrick0 好，MTrick 算法甚至可以达到 80%的准确率

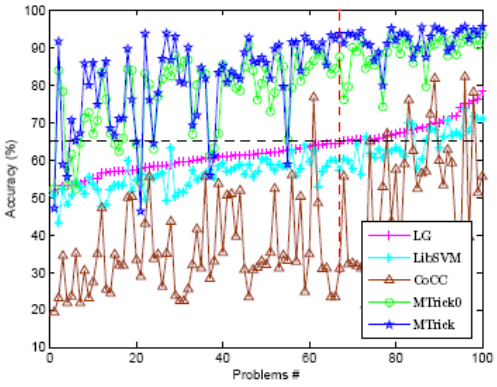
即使监督学习算法 LibSVM 和 LG 的分类效果仅仅跟随机猜测相当。我们还看到当迁移学习问题非常困难的时候，CoCC 算法出现了负迁移。所有的这些结果证明 MTrick 算法是非常有效的。



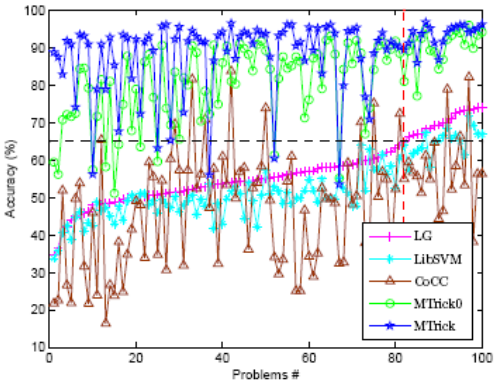
(a) MTrick vs. LG, LibSVM, MTrick0 on data set *comp vs. rec vs. sci*



(b) MTrick vs. LG, LibSVM, MTrick0 on data set *comp vs. rec vs. talk*



(c) MTrick vs. LG, LibSVM, MTrick0 on data set *comp vs. sci vs. talk*



(d) MTrick vs. LG, LibSVM, MTrick0 on data set *rec vs. sci vs. talk*

图 4.5.4 算法 MTrick，LG，LibSVM，CoCC 以及 MTrick0 在三类分类问题上的性能(%)比较

4.5.4 分析 MTrick 输出的词特征聚类

MTrick 不仅能够对目标领域数据进行分类，还能够输出源领域和目标领域数据中词特征的聚类信息。为了说明本章的研究动机，不同领域词特征聚类信息不一样但是相关，因此这一小节分析输出的$\mathbf{F}_s$和$\mathbf{F}_t$。假设源领域和目标领域中的词特征都聚成k_1类，然后对于每一个聚类都选出N (这里$N=20$)个最有代表性的词，比如A_i和B_i分别表示源领域和目标领域第$i\,(1\le i\le k_1)$个聚类的代表性词特征集合，而C_i表示初始的 PLSA 聚类结果的第i个聚类，那么我们定义以下两个准则，

$$r_1 = \frac{1}{k_1}\sum_{i=1}^{k_1}\frac{|I_i|}{|C_i|}, \tag{4.30}$$

$$r_2 = \frac{1}{k_1}\sum_{i=1}^{k_1}\frac{|U_i \cap C_i|}{|C_i|}, \tag{4.31}$$

其中 $I_i = A_i \cap B_i$， $U_i = A_i \cup B_i$，对于数据集 *sci vs. talk* 中的每个分类问题，都记录这两个值 r_1 和 r_2，实验结果如所示。曲线 r_1 表明虽然源领域和目标领域词特征聚类包含的词特征不一样，但还是相关的，因为它们共享一些相关的代表性词特征；r_2 曲线则表明源领域和目标领域中代表性词特征的并集与初始 PLSA 在所有数据上做聚类的结果非常相似。也就是说，通过我们的算法 MTrick，可以发现不同的领域词特征聚类是不一样的，因为它们用不同的关键词来表示同一个词特征概念；另一方面却相关，因为共享一些代表性的词特征。

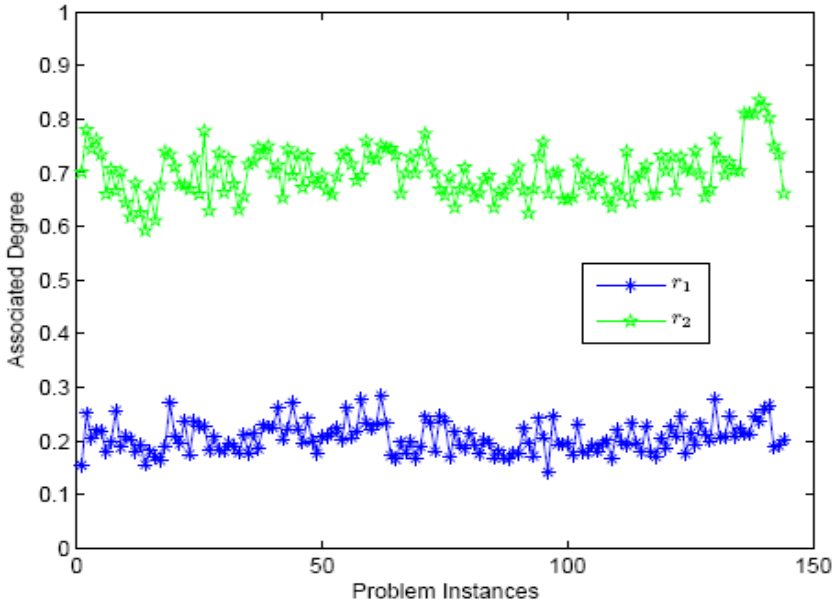


图 4.5.5 数据集 *sci vs. talk* 上的 r_1 和 r_2 值

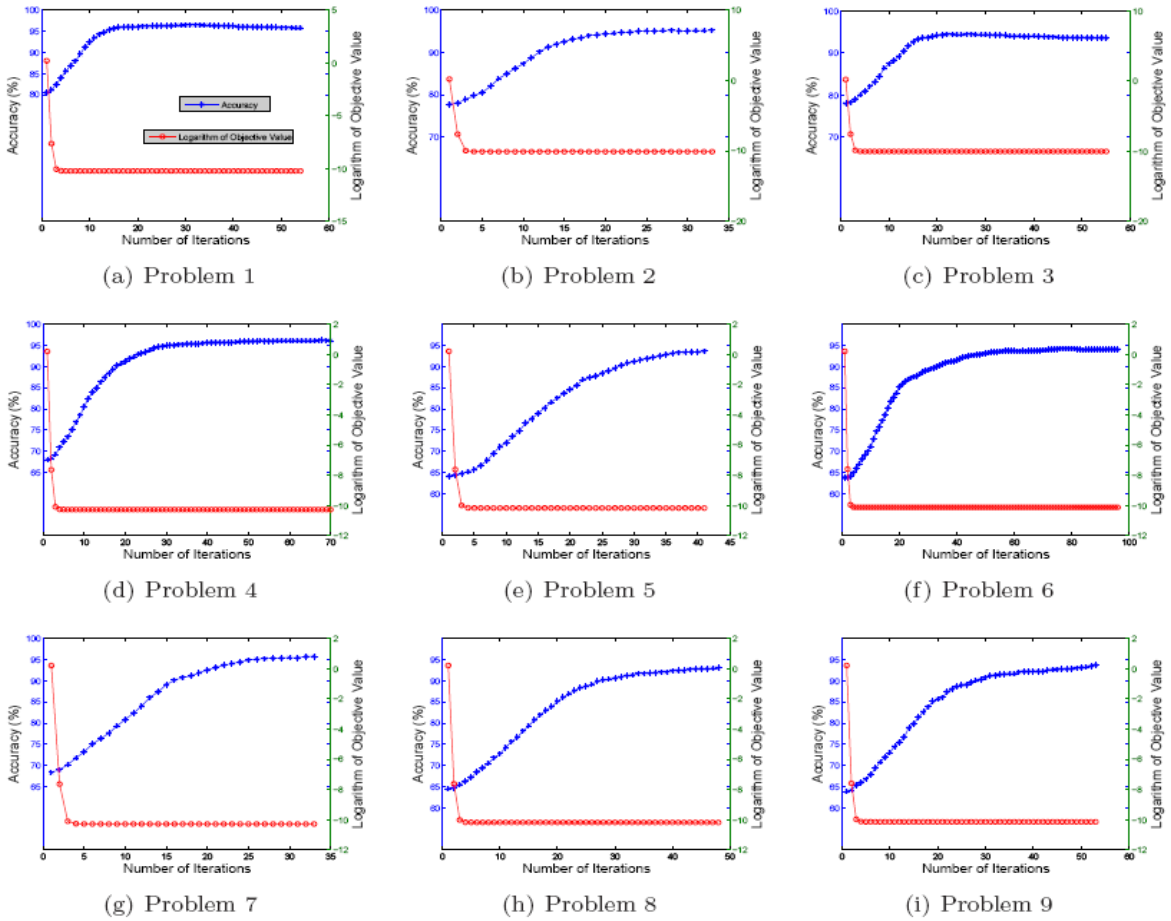
4.5.5 参数影响

我们考察了不同参数设置对 MTrick 算法的影响，MTrick 中有三个参数 α, β 以及词特征聚类个数 k_1。经过初步实验确定各个参数的取值范围 $\alpha \in [1,10]$， $\beta \in [0.5,3]$，$k_1 \in [10,100]$，然后随机采样 10 种参数组合，实验数据是数据集 *sci vs. talk* 上随机选择的 10 个分类问题，所有的实验结果如表 4.5.6 所示。第 12 和 13 行分别表示 10 个问题

的平均准确率和方差，最后一行是实验中采取的参数设置。从表 4.5.6 中的结果可以看到，所有参数情况下，平均值几乎一样，而且方差也很小，因此我们的算法 MTrick 非常稳定，在一定的取值范围内对参数不敏感。

表 4.5.6 参数选择对算法 MTrick 性能(%)的影响

Sampling ID	α	β	*kl*	Problem ID									
				1	2	3	4	5	6	7	8	9	10
1	2.44	2.39	58	92.34	94.28	95.37	88.47	94.99	92.43	95.24	92.04	91.69	95.32
2	7.45	1.69	83	93.05	94.35	97.00	88.47	95.28	92.69	94.91	91.76	92.33	95.30
3	6.92	0.96	38	95.92	94.70	97.33	90.90	95.01	89.32	94.47	90.45	89.99	95.63
4	2.67	1.65	15	94.39	95.53	96.02	90.53	95.42	92.59	94.55	90.92	90.02	95.28
5	5.61	2.45	72	91.58	95.07	94.79	87.83	95.34	93.17	94.99	91.24	91.75	95.28
6	3.63	2.32	32	93.59	94.12	94.98	89.98	95.57	92.90	94.49	91.83	91.24	95.09
7	2.30	1.57	21	92.72	94.46	96.47	89.77	94.84	92.43	94.49	91.34	91.46	96.23
8	7.53	0.72	52	95.80	94.12	97.52	91.13	95.40	89.55	94.35	89.71	90.12	95.47
9	1.88	1.50	26	95.57	94.14	96.90	90.70	95.71	92.69	94.93	91.53	90.08	95.08
10	7.95	1.18	92	94.54	95.02	97.51	89.75	95.28	92.18	94.55	91.38	92.28	95.70
Variance				2.351	0.236	1.089	1.370	0.073	1.897	0.085	0.496	0.913	0.119
Mean				93.95	94.58	96.39	89.75	95.28	92.00	94.70	91.22	91.10	95.44
This paper	1	1.5	50	93.77	94.42	94.99	90.33	95.05	93.12	95.96	93.84	90.90	95.66



(a) Problem 1 (b) Problem 2 (c) Problem 3

(d) Problem 4 (e) Problem 5 (f) Problem 6

(g) Problem 7 (h) Problem 8 (i) Problem 9

图 4.5.6 MTrick 算法收敛性，准确率、目标值与迭代次数的关系

4.5.6 算法收敛性

前面已经从理论上分析了 MTrick 算法的收敛性，这里再次用实验验证算法的收敛性。从数据集 *sci vs. talk* 中任意选择 9 个问题，其实验结果如图 4.5.6，*x*-坐标表示迭代次数，左 *y*-坐标表示准确率，右 *y*-坐标表示式子(4.6)目标值的对数。从结果可以看到，准确率基本上随着迭代次数的增加，先增加然后稳定，基本保持不变；而目标值则是先下降很快，最后也趋于收敛状态，与理论分析一致。

4.6. 小结

本章首先首先分析了不同领域之间虽然在原始词特征层面上数据分布不同，但是词特征聚类和文档类别之间的关联关系可能是领域独立的，即可以通过该关联关系在源领域和目标领域之间进行知识迁移。我们提出基于非负矩阵分解的跨领域学习方法 MTrick，该算法可以同时分解源领域和目标领域的概率分布矩阵，而通过共享词特征聚类和文档类别之间的关联关系来实现对目标领域数据的分类预测。我们开发了一个迭代算法对 MTrick 进行求解，并从理论上分析了该迭代算法的收敛性。实验结果表明本章提出的算法可以很好地解决迁移学习问题，并且优越于所有比较的算法。

基于非负矩阵分解的跨领域学习方法，虽然能够取得很好的分类性能，但是缺乏概率解释，第六章将介绍基于生成模型的跨领域学习方法，具有更加完美的概率解释。下一章介绍从多个源领域的跨领域学习方法，能够避免从单个源领域学习的偏差，而且从多个源领域学习，分类性能可以进一步提高。

第五章　基于一致性正则化的多源跨领域学习框架

5.1. 引言

前面很多关于迁移学习的工作都主要集中在单个源领域到单个目标领域的学习[Dai, 2007b; Dai, 2007; Xing, 2007]。在实际应用中，有标签的训练样本可能来自多个源领域，而且这些源领域数据虽然分布不同，但是却语义相关，下面给出两个例子作为本章的研究动机。

在 **Web 网页分类**中，假设我们下载了一个高校的所有 Web 网页，并想要从所有的网页中找出关于课程的网页。如果利用传统的分类方法来把有关课程的网页分出来，那么就需要花费很大的人工劳动对足够多的网页进行标注作为训练集，从而学习得到一个好的分类模型。这对于巨大的网页量来说，不太可能。在这种情况下，如果有其他高校的网页，已经大部分被标注，那么就可以利用这些已经标注的数据集进行训练，然后对目标高校的网页进行分类。但是这些已标注的网页不能直接用于传统的分类学习方法，因为不同高校的数据集分布可能不同。比如，一个网页中表示阅读材料的词项可能包括“Required Reading List”，“TextBooks”，“Reference”等等。那么不同的高校可能用不同词来描述同一个概念，就会导致数据分布不一致，这样传统的机器学习方法就很难处理这种情况。因此，我们考虑一个高校作为一个源领域的知识，那我们的目标就是找出目标高校的新的关于课程的网页，即本章所讨论的多源领域迁移学习问题。

图 5.1.1 四个主要媒体的视频镜头(CCTV, CBS, CNN and NBC)

迁移学习在**视频概念检测**中也有很大应用前景，就是把建立在多个源视频数据集上的语义概念检测模型泛化到目标领域的数据集上。在电视节目频道中，由于视觉特征以

及语义内容所产生的“语义缺口”，不同节目频道的视频语义信息是不同分布的。如图 5.1.1 所示，很明显看到图中的 4 个视频镜头 CCTV，CBS，CNN 和 NBC 属于不同电视节目，但是却有相似的视觉特征。如果把每个节目频道当成一个领域，那么数据集 TRECVID 数据集[Smeaton, 2003]也是一个多源领域学习问题的视频数据集。由于数据分布不匹配在多媒体领域更加严重[Yang, 2007]，那么从单个源领域的学习就可能表现不佳。利用本章提出的方法可以很容易地从多个源领域的视频数据集迁移学习知识，然后用于目标领域的视频语义检测。

以上举的两个例子，训练集来自于多个源领域数据，并且语义相关，但是属于不同的数据分布。如果把所有的源领域的数据融合起来，形成一个大的源领域，这样就可以利用现有一个源领域到另一个源领域的迁移学习算法。但是在数据合并的过程中不同数据分布之间所隐含的信息就有可能丢失，这个重要信息对于理解所有源领域的共同特性起着非常关键的作用。还有，这些源领域的数据在地理位置上可能是分布式的，数据的融合需要消耗网络带宽，大大加强了对网络负载的需求；另外，有时候为了保护数据的隐私性，是不允许各个源领域的原始数据在网络上传输和最终的合并。因此，本章还把一致性正则化算法推广到分布式实现，主从节点上通信主要是一些中间结果和统计信息，这样减轻了网络通信负载以及一定程度上保护了数据的隐私。

假设有 m 个源领域的数据集表示为 $D_s^1,\cdots,D_s^m$，一个目标领域的数据集表示为 D_t。有标签的各个源领域的数据集之间以及与没有标签的目标领域的数据集之间在地理位置上分布式存储。我们假设所有的数据集上的样本类别个数一样，即源领域与目标领域的数据集有相同的标签集 C；另外进一步假设，这些数据集之间虽然具有不同的数据分布，但却是语义相关，即相似的数据特征描述相似的类别信息。在这些假设下，我们的目标是利用从 m 个源领域学到的知识对目标领域中的数据进行正确分类。图 5.1.2 描述了从多个源领域到目标领域的迁移学习示意图。

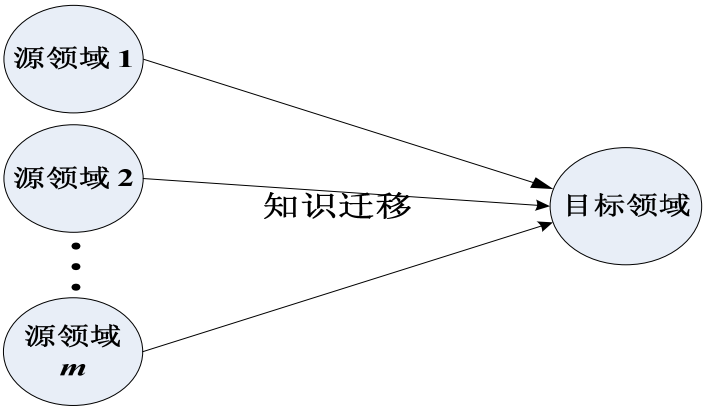


图 5.1.2 多源领域迁移学习示意图

如前面所描述，每个源领域的数据分布与目标领域的数据分布可能都不同，若局部子分类器 $h^l(l=1,\cdots,m)$ 是由数据集上 D_s^l 训练得到的分类器，$D_s^1,\cdots,D_s^m$ 上的数据都有标

签，而 D_t 上的数据都没有标签。如果仅仅简单地利用从单个源领域的数据学习得到的分类器对目标数据集进行分类，由于数据分布的不匹配就有可能得到极差的预测结果，而且所有局部分类器分类得到的结果就有可能大相径庭，产生高方差。实际上，目标数据集上的每个样本的标签都是本身固定的，局部分类器的预测结果多样性不利于集成分类器的分类性能。这种预测不同性就预示了算法的优化方向，使得局部的分类器对目标数据集的预测结果尽可能一致。

如何解决对多个源领域的知识进行迁移学习，并把它运用于新的目标领域？主要有以下两个难点，也是本章要解决的问题：

(1) 怎么利用多个不同源领域的不同分布来促进在目标领域的预测效果，以及如何优化使得局部分类器对目标数据集具有尽可能一致的预测效果；

(2) 为了保护原始数据的隐私性，如何扩展该算法到分布式环境中实现。

针对以上问题，本章提出了旨在解决从多个源领域到单个目标领域的迁移学习问题的一致性正则化框架，基础分类算法是逻辑回归(Logistic Regression)模型(也可以利用其他指数模型作为基础分类模型)。对于第一个问题，我们提出了最大一致性正则化方法，各个局部分类器学习过程中结合利用目标领域中的样本信息。最后每个源领域都会得到一个局部分类器，该分类器既考虑了在本身数据集上的分类性能，也考虑了与其他局部分类器在目标数据集上的预测一致性。由于源领域与目标领域的知识相互影响，学习后各个局部分类器不仅保持了自己的独立性，也表现出了在目标领域上的共同性质。对于第二个问题的解决，我们设计的算法可以分布式运行，该算法实现具有主次节点的分布式体系，目标领域作为主节点，而其他的源领域作为从节点。为了得到最终的局部分类器，算法的每次迭代只要在主次节点之间传递一些中间结果和统计信息，而不用传递各个源领域的原始数据，有利于保护数据的隐私。

5.2. 一致性度量和问题形式化

这一小节首先定义一致性度量准则，然后介绍问题的形式化。

5.2.1 一致性度量

为了对一致性进行度量的准则进行定义，首先定义基于概率分布向量的 *shannon* 熵(*Shannon Entropy*)，然后利用该 *shannon* 熵来度量所有子分类器对某个样本预测结果的一致性程度。

定义 5.1 (*Probability Distribution Vector*) $\mathbf{p} \in \mathbb{R}_{+}^{|C|}$ 是一个预测概率分布向量，满足 $\sum_{i=1}^{|C|} \mathbf{p}_i = 1$，$|C|$ 表示样本的类别数，$\mathbf{p}_i$ 表示样本属于第 i 类的概率。

定义 5.2 (*Shannon Entropy*) 假设 $\mathbf{p} \in \mathbb{R}_{+}^{|C|}$ 是一个概率分布向量，对于概率分布向量 $\mathbf{p}$ 的

shannon 熵定义如下：

$$\mathrm{E}(\mathbf{p}) = \sum_{i=1}^{|C|} \mathbf{p}_i \log \frac{1}{\mathbf{p}_i} \tag{5.1}$$

假设有 m 个分类器 $\mathbf{H} = \left\{h^l\right\}_{l=1}^{m}$，每个分类器对一个样本的预测都输出一个预测概率分布向量 $\mathbf{p}^l$，在这些分类器上的平均预测分布向量通过式(5.2)计算：

$$\overline{\mathbf{p}} = \frac{\sum_{l=1}^{m} \mathbf{p}^l}{m} \tag{5.2}$$

通过上面定义的 *shannon* 熵，可以算出这些分类器在样本上的预测结果的一致性程度。让我们来看一个例子，有 3 个分类器和 3 个类别的分类问题，在表 5.2.1 中列出了由三个分类器 h^1, h^2, h^3 对样本 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ 预测的概率分布向量，以及它们对应的平均概率分布向量。从表 5.2.1 可以看到，这些分类器对于样本 $\mathbf{x}_1$ 的预测一致性最高，达到了一致，预测为类别 1 的概率都为 1。因此当它们的一致性程度达到最大的时候，这些预测结果的平均概率分布向量的熵 $\mathrm{E}(1,0,0)$ 达到最小。另一方面，对于第三个样本 $\mathbf{x}_3$，三个分类器预测的结果分别是以 1 的概率属于第 1，2，3 类，这些结果相互之间都不相同，一致性程度达到最小，即平均概率分布向量的熵达到最大 $\mathrm{E}(\frac{1}{3},\frac{1}{3},\frac{1}{3})$。基于此，可以利用熵取负值作为分类器对样本预测结果一致性程度的度量值。

表 5.2.1 熵和概率分布向量的一致性度量

Instance	$\mathbf{p}^1$ by h^1	$\mathbf{p}^2$ by h^2	$\mathbf{p}^3$ by h^3	$\overline{\mathbf{p}}$	Entropy	Consensus
$\mathbf{x}_1$	(1,0,0)	(1,0,0)	(1,0,0)	(1,0,0)	0	0
$\mathbf{x}_2$	(0.7,0.25,0.05)	(0.8,0.1,0.1)	(0.6,0.25,0.5)	(0.7,0.2,0.1)	1.16	-1.16
$\mathbf{x}_3$	(1,0,0)	(0,1,0)	(0,0,1)	(033,0.33,0.33)	1.59	-1.59

定义 5.3 (*Entropy Based Consensus Measure*) 给出 m 个预测概率分布向量 $\mathbf{p}^1, \mathbf{p}^2, \cdots, \mathbf{p}^m$，这些向量的一致性程度的度量定义如下：

$$C_e(\mathbf{p}^1, \mathbf{p}^2, \cdots, \mathbf{p}^m) = -\mathrm{E}(\overline{\mathbf{p}}) \tag{5.3}$$

其中 E 是式子(5.1)定义的 *shannon* 熵，$\overline{\mathbf{p}}$ 是由式子(5.2)计算得到的平均预测概率分布向量。由于只考虑两个一致性度量值的大小，因此允许它们的值是负数。在上面的定义下，很容易得到分类器对样本 $\mathbf{x}_2$ 的预测结果的一致性程度度量值为 $-\mathrm{E}(0.7, 0.2, 0.1)$。

由于熵计算的复杂性，对于两类情况的概率分布向量，一致性度量可以简化为如下式子：

$$C_s(\mathbf{p}^1,\mathbf{p}^2,\cdots,\mathbf{p}^m)=(\bar{\mathbf{p}}_1-\bar{\mathbf{p}}_2)^2=(\bar{\mathbf{p}}_1-(1-\bar{\mathbf{p}}_1))^2=(2\bar{\mathbf{p}}_1-1)^2 \tag{5.4}$$

我们认为当仅仅比较两个度量值的大小，且在两类情况下，C_e 与 C_s 对一致性程度的度量是等价的。在后面的一致性正则化具体实现以及实验中将采用式子(5.4)来度量所有分类器对样本的预测一致性。实际上该一致性度量准则具有两个优势，1) 最大化模型在目标领域上的预测一致性；2) 最小化熵，即使得尽可能以最大概率预测为某一类。实验中将会验证这两者的作用。

5.2.2 问题形式化

给出 m $(m>1)$ 个已标注的源领域数据集 $D_s^1,\cdots,D_s^m$ 作为训练集，第 l 个源领域的数据集表示为 $D_s^l=\{(\mathbf{x}_i^l,y_i^l)\}|_{i=1}^{n^l}$，其中 y_i^l 是样本 $\mathbf{x}_i^l$ 的真实标签，n^l 是第 l 个源领域数据集的样本数；那么未标注的目标领域数据集表示为 $D_t=\{(\mathbf{x}_i)\}|_{i=1}^{n}$，$n$ 是目标领域数据集的大小。根据前面的假设，数据集 $D_s^1,\cdots,D_s^m$ 以及 D_t 之间的分布不同，但是相关。我们的目标是从这些源领域迁移学习到一个分类模型，使得它能够对目标领域的未标注数据进行正确分类。

如果我们从 m 个不同源领域局部地训练得到分类器 $h^1,\cdots,h^m$，那么它们都是仅仅基于局部训练数据得到的分类模型，最理想的情况就是这些局部分类器对目标领域的数据预测结果完全的一致，且预测结果都是以 100%的置信度把目标数据样本分到某一类。但由于 $D_s^1,\cdots,D_s^m$ 以及 D_t 之间的分布不同，因此这些模型在某些程度上对目标数据的预测结果通常会有很大的不同，这就为使得局部分类器模型在目标数据集上的预测结果一致性提供了优化的空间，这也是本章提出一致性正则化框架的动机。本章把一致性度量结合到监督学习的标准框架中，通过优化式子(5.5)后，就可以得到 m 个优化后的模型 $h'^1,\cdots,h'^m$：

$$\sum_{l=1}^{m}P(h^l\mid D_s^l)+\theta\cdot Consensus(h^1,\cdots,h^m\mid D_t) \tag{5.5}$$

其中 $P(h^l\mid D_s^l)$ 是在观测数据集 D_s^l 下，假设 h^l 给出的预测概率值，而 $Consensus(h^1,\cdots,h^m\mid D_t)$ 是 m 个分类模型 $h^1,\cdots,h^m$ 在目标数据集 D_t 上预测结果一致性程度的度量值，θ 是平衡参数，控制一致性正则化的作用。在式子(5.5)的第一项中，每个模型 h^l 应用于本身对应的源领域训练数据集上，而第二项正则化项正是各个局部模型

连接的桥梁，实现互相协作最优化，最后达到预测结果尽可能一致。因此，该正则化框架不仅考虑分类器$h^1,\cdots,h^m$最大化在各自源领域数据上的最大后验，而且最大化在目标数据集上预测的一致性程度。通过这种方式得到的模型，不仅保持了在各个对应源领域数据集上的独立性，也表现了在目标领域上的共同特性。

给出一个源领域的数据集$D_s^l=\{(\mathbf{x}_i^l,y_i^l)\}|_{i=1}^{n^l}$，是独立同分布的。最大化式子(5.5)中的第一项$P(h^l\mid D_s^l)$可以进一步扩展如下：

$$\begin{aligned}\max P(h^l\mid D_s^l)&=\max\frac{P(h^l,D_s^l)}{P(D_s^l)}\\&=\max P(h^l,D_s^l)=\max P(D_s^l\mid h^l)P(h^l)\\&=\max P(h^l)\cdot\prod_{i=1}^{n^l}P(y_i^l\mid\mathbf{x}_i;h^l)=\max(\log(P(h^l)+\sum_{i=1}^{n^l}\log P(y_i^l\mid\mathbf{x}_i;h^l))\end{aligned}\tag{5.6}$$

对于一致性程度度量$Consensus(h^1,\cdots,h^m\mid D_t)$，定义为$m$个模型在目标领域所有数据上预测结果一致性度量值的总和，

$$Consensus(h^1,\cdots,h^m\mid D_t)=\sum_{i=1}^{n}C_e(\mathbf{p}_i^1,\cdots,\mathbf{p}_i^m)\tag{5.7}$$

其中，$\mathbf{p}_i^l$表示在源领域数据集D_s^l训练得到的模型h^l对目标领域数据集D_t中第i个样本的预测概率分布向量。

5.3. 基于逻辑回归的一致性正则化实现

本章采用第三章中介绍的逻辑回归函数形式来实现本章提出的一致性正则化算法，且主要集中两类分类问题。但根据前面熵度量一致性程度的定义，一致性正则化框架可以直接扩展到多类情况。

根据上一节的问题形式化，我们可以通过最大化以下式子来得到一致性正则化框架下的m个逻辑模型$\mathbf{w}^1,\cdots,\mathbf{w}^m$，

$$\begin{aligned}g_e(\mathbf{w}^1,\cdots,\mathbf{w}^m)=&\sum_{l=1}^{m}(\sum_{i=1}^{n^l}\log P(y_i^l\mid\mathbf{x}_i^l;\mathbf{w}^l)-\frac{\lambda^l}{2}\mathbf{w}^{l\mathrm{T}}\mathbf{w})\\&-\theta\cdot\sum_{i=1}^{n}\mathrm{E}(\frac{\sum_{l=1}^{m}P(y=-1\mid\mathrm{x}_i;\mathbf{w}^l)}{m},\frac{\sum_{l=1}^{m}P(y=1\mid\mathrm{x}_i;\mathbf{w}^l)}{m})\end{aligned}\tag{5.8}$$

其中，条件概率P是逻辑函数，其定义见式子(3.9)，E是*shannon*熵，$\mathbf{w}$是逻辑回归线性模型。为了简化以上函数对熵求导的复杂性，对于2类问题用式子(5.4)中的C_s取代C_e，因此式子(5.8)可以改写如下：

$$g_s(\mathbf{w}^1,\cdots,\mathbf{w}^m)=\sum_{l=1}^{m}(\sum_{i=1}^{n^l}\log P(y_i^l\mid \mathbf{x}_i^l;\mathbf{w}^l)-\frac{\lambda^l}{2}\mathbf{w}^{l^{\mathrm{T}}}\mathbf{w}) \\ +\theta\cdot\sum_{i=1}^{n}(2\cdot\frac{\sum_{l=1}^{m}P(y=1\mid \mathrm{x}_i;\mathbf{w}^l)}{m}-1)^2 \tag{5.9}$$

为了优化上述函数，求它们最大值得到m个模型，对上式函数g_s中每个逻辑模型$\mathbf{w}^l$求偏导得到下式，

$$\nabla_{\mathbf{w}^l}(g_s)=\frac{\partial g_s}{\partial \mathbf{w}^l}=\nabla_{sn}^l(\mathbf{w}^l,D_s^l)+\nabla_{mn}^l(\mathbf{w}^1,\cdots,\mathbf{w}^m,D_t) \tag{5.10}$$

根据函数σ的定义$\sigma(y\mathbf{w}^{\mathrm{T}}\mathbf{x})=\frac{1}{1+\exp(-y\mathbf{w}^{\mathrm{T}}\mathbf{x})}$，上式中的$\nabla_{sn}^l,\nabla_{mn}^l$可以分别表示成如下式子，

$$\nabla_{sn}^l(\mathbf{w}^l,D_s^l)=\sum_{i=1}^{n^l}(1-\sigma(y_i^l\mathbf{w}^{l^{\mathrm{T}}}\mathbf{x}))y_i^l\mathbf{x}_i^l-\lambda^l\mathbf{w}^l\ , \tag{5.11}$$

$$\nabla_{mn}^l(\mathbf{w}^1,\cdots,\mathbf{w}^m,D_t)=\frac{4\theta}{m^2}\sum_{i=1}^{n}(2\cdot\sum_{k=1}^{m}\sigma(\mathbf{w}^{k^{\mathrm{T}}}\mathbf{x}_i)-m)(1-\sigma(\mathbf{w}^{l^{\mathrm{T}}}\mathbf{x}_i))\sigma(\mathbf{w}^{l^{\mathrm{T}}}\mathbf{x}_i)\mathbf{x}_i \tag{5.12}$$

显然目标函数(5.9)是一个非凸非凹函数，因此很难用非线性优化方法得到最优解。不过可以通过求目标函数的偏导数，然后利用非线性最优化函数求得局部最优解，本章采用共轭梯度[Ruszczynski, 2006]方法，且$\mathbf{w}^1,\cdots,\mathbf{w}^m$的初始值由各个源领域用监督学习算法逻辑回归[Davie, 2000]学习得到。详细的算法伪代码如算法 5.1 所示，步骤 4 中单变量的优化可以用传统的优化方法求解，本算法采用 Matlab 自带的函数*fminunc*。

算法 5.1: 共轭梯度法求解一致性正则化的集中式版本

输入: 有标签源领域数据集$D_s^1,\cdots,D_s^m$，无标签目标领域数据集D_t，单元矩阵$\mathbf{Q}\in\mathbb{R}^{N\times N}$（$N$是样本的维度)；误差阈值$\varepsilon>0$；迭代算法的最大迭代次数$max$。

输出: m个分类模型$\mathbf{w}^1,\cdots,\mathbf{w}^m$。

步骤 1：每个源领域利用逻辑回归算法在本地数据上初始化$\mathbf{w}_0^l$；

步骤 2：$k:=1$；

步骤 3：对于$l=1,\cdots,m$，用式子(5.10)计算$\nabla_{\mathbf{w}^l}(g_s)$，然后设置搜索方向

$$\begin{cases} \mathbf{d}_0^l = \nabla_{\mathbf{w}_0^l}(g_s) \\ \mathbf{d}_{k+1}^l = \nabla_{\mathbf{w}_{k+1}^l}(g_s) + \alpha_k \mathbf{d}_k^l \\ \alpha_k = -\nabla_{\mathbf{w}_{k+1}^l}^{\mathrm{T}}(g_s)\mathbf{Q}\mathbf{d}_k^l / \mathbf{d}_k^{l\,\mathrm{T}} Q \mathbf{d}_k^l \end{cases} \tag{5.13}$$

如果$\sum_{l=1}^{m} \| \nabla_{\mathbf{w}_k^l}(g_s) \| < \varepsilon$，则转到步骤 6;

步骤 4：优化以下式子，计算最佳搜索步长$\gamma \geq 0$，

$$u_s(\gamma) = g_s(\mathbf{w}_k^1 + \gamma \mathbf{d}_k^1, \cdots, \mathbf{w}_k^m + \gamma \mathbf{d}_k^m) \tag{5.14}$$

对于$l = 1, \cdots, m$，用式子(2.18)计算$\mathbf{w}_{k+1}^l$，

$$\mathbf{w}_{k+1}^l = \mathbf{w}_k^l + \gamma \mathbf{d}_k^l \tag{5.15}$$

步骤 5：$k := k+1$，若$k < max$则转步骤 3。

步骤 6：输出$\mathbf{w}_k^1, \cdots, \mathbf{w}_k^m$。

5.4. 一致性正则化算法的分布式实现

本章提出的一致性正则化算法的另一个优势在于可以在分布式环境下实现，只传递一些中间结果，而不是原始数据，以保护各个源领域数据的安全性，隐私性等。在这一节中，我们探讨如果扩展集中式算法到分布式学习算法，使得该算法能够处理源领域数据$D_s^1, D_s^2, \cdots, D_s^m$与目标领域数据$D_t$在地理上分布的情况。在这个分布式环境中，把各个包含源领域原始数据的数据节点当作从节点，表示为$sn^1, sn^2, \cdots, sn^m$，而包含目标领域数据的节点做主节点，表示为mn。

回顾下上一节中式子(5.10)求得的偏导数，可以发现它由两部分组成，第一部分$\nabla_{sn}^l(\mathbf{w}^l, D_s^l)$只与局部模型$\mathbf{w}^l$和数据集$D_s^l$有关，因此可以在包含该源领域数据的从节点上计算完成；而第二项$\nabla_{mn}^l(\mathbf{w}^1, \cdots, \mathbf{w}^m, D_t)$只与各个源领域数据训练得到的模型$\mathbf{w}^1, \cdots, \mathbf{w}^m$和目标领域的数据集$D_t$有关，因此各个从节点$sn^l (l = 1, \cdots, m)$只要把分别训练得到的模型$\mathbf{w}^l (l = 1, \cdots, m)$和计算得到的$\nabla_{sn}^l$发送给主节点$mn$，主节点收到这些中间结果后就可以直接计算$\nabla_{\mathbf{w}^l}(g_s) = \nabla_{sn}^l + \nabla_{mn}^l$，在计算的过程中不用传输源领域中的原始数据。

通过上述分析，在优化过程中的每一步，只要在主从节点间进行一些简单的通信，传输中间结果，就可以分布式实现整个一致性正则化优化算法，算法 5.2 详细描述了分布式算法。实际上算法 5.2 只是算法 5.1 的分布式实现，因此它们可以优化得到相同的分类模型。

算法 5.2: 共轭梯度法求解一致性正则化的分布式版本

输入: 有标签源领域数据集 $D_s^1,\cdots,D_s^m$ 在不同的从节点 $sn^1,\cdots,sn_m$ 上，无标签目标领域数据集 D_t 在主节点 mn 上，误差阈值 $\varepsilon>0$；迭代算法的最大迭代次数 max；迭代步长常数 γ。

输出: m 个分类模型 $\mathbf{w}^1,\cdots,\mathbf{w}^m$。

步骤 1：每个从节点 sn^l $(l=1,\cdots,m)$利用逻辑回归算法在本地数据 D_s^l 上初始化 $\mathbf{w}_0^l$，然后发送初始模型 $\mathbf{w}_0^l$ 以及 $\nabla_{sn}^l(\mathbf{w}_0^l,D_s^l)$ $(l=1,\cdots,m)$到主节点 mn；

步骤 2：$k:=1$；

步骤 3：主节点用式子(5.12)计算梯度 $\nabla_{\mathrm{w}^l}(g_s)$ $(l=1,\cdots,m)$，然后根据式子(5.13)设置搜索方向；如果 $\sum_{l=1}^m\|\nabla_{\mathbf{w}_k^l}(g_s)\|<\varepsilon$，则转到步骤 6，否则利用输入常数 γ 和式子(5.15)计算 $\mathbf{w}_{k+1}^l(l=1,\cdots,m)$;

步骤 4：主节点发送 $\mathbf{w}_{k+1}^l(l=1,\cdots,m)$给从节点，然后从节点各自计算 $\nabla_{sn}^l(\mathbf{w}_0^l,D_s^l)$ 后，重新发送给主节点;

步骤 5：$k:=k+1$，若 $k<max$ 则转步骤 3。

步骤 6：输出 $\mathbf{w}_k^1,\cdots,\mathbf{w}_k^m$。

我们还分析了该循环优化过程中的通信代价，在算法 5.2 中，每个从节点 sn^l 实际上只需要传送一个向量 ∇_{sn}^l (第一次迭代还需要传输最原始的模型 $\mathbf{w}^l$)给主节点，然后主节点 mn 返回更新后的模型给对应的从节点，那么当迭代 k 次后，算法收敛，总的通信代价为 $(2k+1)\sum_{l=1}^m|\mathbf{w}^l|$。

从上面的分析中，可以看到在算法的实现中，只要传输一些中间结果和统计信息，比如 $\nabla_{sn}^l(l=1,\cdots,m)$，而不是传输原始数据，这从一定程度上有效地保护了数据的隐私性。

5.5. 为什么一致性正则化有用？

本节从理论上分析一致性正则化的有效性，即最大化任何两个分类器的一致性，可以提高分类器的性能。

为了简化讨论，我们仍然考虑两类分类的情况(该分析可以泛化到多类分类的情况)，分类类标为 1 和-1。从 m 源领域学习得到 m 个模型 $h^1,\cdots,h^m$，则任意两个模型预测不一致的概率表示为 $\Pr(h^i \neq h^j)\,(i,j\in\{1,\cdots,m\},i\neq j)$，这里说明下 $\Pr(\cdot)$ 只针对目标领域数据定义。假设变量 Y 表示样本的目标类别，则我们有如下三个定义：

定义 5.4 (*Nontrivial Classifier*) 如果分类器 h 满足以下条件，

$$\Pr(h=u\,|\,Y=u)>\Pr(h=\bar{u}\,|\,Y=u)\,, \tag{5.16}$$

其中 $u\in\{1,-1\}$，且 $\bar{u}$ 是 u 的补，则称分类器 h 是一个不平凡分类器(Nontrivial Classifier)。

对于两类分类问题，不平凡条件可以表示如下，

$$\Pr(h=u\,|\,Y=u)>\frac{1}{2}\quad \text{或者}\quad \Pr(h\neq u\,|\,Y=u)\leq\frac{1}{2} \tag{5.17}$$

定义 5.5 (*Nonperfect Classifier*) 如果分类器 h 对目标领域数据集的预测准确率小于 100%，则称之为不完美分类器(Nonperfect Classifier)。

定义 5.6 (*Conditional Independent Classifiers*) 如果分类器 $h^1,\cdots,h^m$ 满足条件独立，则有，

$$\Pr(h^i=u\,|\,h^j=v,Y=y)=\Pr(h^i=u\,|\,y=y) \tag{5.18}$$

其中 $u,v,y\in\{1,-1\}$，$i,j\in\{1,\cdots,m\},i\neq j$。

如果分类器满足以上三个条件，不平凡、不完美以及条件独立，则有如下定理成立：

定理 5.1：如果分类器满足条件，不平凡、不完美以及条件独立，则分类器之间的不一致程度是分类器预测错误率的严格上界。

证明：分类模型 h^i 的分类错误率为

$$\begin{aligned}\Pr(h^i\neq Y)&=\Pr(h^i=1,Y=-1)+\Pr(h^i=-1,Y=1)\\&=\Pr(h^i=1,h^j=-1,Y=-1)+\Pr(h^i=1,h^j=1,Y=-1)\\&\quad+\Pr(h^i=-1,h^j=-1,Y=1)+\Pr(h^i=-1,h^j=1,Y=1)\end{aligned}\,, \tag{5.19}$$

两个分类器预测不一致的概率则为

$$\begin{aligned}\Pr(h^i\neq h^j)&=\Pr(h^i=1,h^j=-1)+\Pr(h^i=-1,h^j=1)\\&=\Pr(h^i=1,h^j=-1,Y=-1)+\Pr(h^i=1,h^j=-1,Y=1)\\&\quad+\Pr(h^i=-1,h^j=1,Y=-1)+\Pr(h^i=-1,h^j=1,Y=1)\end{aligned}\,, \tag{5.20}$$

其中，$i, j \in \{1,\cdots,m\}, i \neq j$。

为了验证$\Pr(h^i \neq Y) < \Pr(h^i \neq h^j)$，我们只需要证明以下不等式成立，

$$\begin{aligned}&\Pr(h^i=1,h^j=1,Y=-1)+\Pr(h^i=-1,h^j=-1,Y=1)\\&<\Pr(h^i=1,h^j=-1,Y=1)+\Pr(h^i=-1,h^j=1,Y=-1)\end{aligned} \tag{5.21}$$

根据式子(5.18)和贝叶斯准则，式子(5.21)可以写成

$$\begin{aligned}&\Pr(h^i=1\mid Y=-1)\Pr(h^j=1,Y=-1)\\&+\Pr(h^i=-1\mid Y=1)\Pr(h^j=-1,Y=1)\\&<\Pr(h^i=1\mid Y=1)\Pr(h^j=-1,Y=1)\\&+\Pr(h^i=-1\mid Y=-1)\Pr(h^j=1,Y=-1)\end{aligned} \tag{5.22}$$

通过定义 5.4 和 5.5，有以下不等式成立，

$$\Pr(h^i=1\mid Y=-1) < \Pr(h^j=-1\mid Y=-1) \tag{5.23}$$

$$\Pr(h^i=-1\mid Y=1) < \Pr(h^j=1\mid Y=-1) \tag{5.24}$$

$$\Pr(h^i=-1,Y=1) > 0 \tag{5.25}$$

$$\Pr(h^i=1,Y=-1) > 0 \tag{5.26}$$

因此，式子(5.21)成立。最后可以得到

$$\Pr(h^i \neq Y) < \Pr(h^i \neq h^j) \tag{5.27}$$

证明完毕。

通过最大化目标函数(5.5)，可以得到优化后的分类模型$f^1,\cdots,f^m$，则有以下定理成立。

定理 5.2：如果分类器满足条件，不平凡、不完美以及条件独立，则优化后的分类器的准确率是分类器一致性的上界。

证明：根据定理 5.1 有

$$\Pr(f^i \neq Y) < \Pr(f^i \neq f^j) \tag{5.28}$$

那么，

$$\begin{aligned}1-\Pr(f^i=Y)&<1-\Pr(f^i=f^j)\\\Pr(f^i=Y)&>\Pr(f^i=f^j)\end{aligned} \tag{5.29}$$

证明完毕。

定理 5.1 和 5.2 都表明，本章提出的一致性正则化框架可以有效地降低分类器的分类错误率，提高学习性能。

5.6. 实验过程和结果

本节介绍了如何设计实验来评价本文提出的算法有效性，在实验中，我们主要关注从 3 个源领域到一个目标领域的 2 类迁移学习分类问题。实验数据包括文本数据和图像数据。

5.6.1 实验数据

类似于文献[Dai, 2007]以及第四章中数据准备方法，我们构造了从多个源领域进行迁移学习的数据集。要求选择的数据集至少应该有两层的层次结构，即有几个大类，然后大类下面又有若干子类(该实验中每个大类包含 4 个小类)。假设数据集中有两个大类，表示为 A 和 B，那么 $A_1,\cdots,A_4$ 和 $B_1,\cdots,B_4$ 分别表示它们的下一层的 4 个小类，这 4 个小类中，其中 3 个小类作为源领域数据，另外 1 个作为目标领域数据。我们构造训练集和测试集如下，令 A_{a_i} 和 B_{b_i} ($a_i,b_i \in \{1,\cdots,4\}$)分别为第 i 个领域中正负类样本实例，$D_i = A_{a_i} \bigcup B_{b_i}$，在构造的 3 个源领域和 1 个目标领域中，$A_{a_i}$ 和 B_{b_i} 当且仅当只出现一次。这样构造以后，不同领域中的正类样本相关但不相同，因为它们都是属于相同的大类 $A(B)$ 却属于不同的小类，同样不同领域中的负类样本也满足这种情况。以上构造出来的 4 个领域，它们具有不同的数据分布，但是相关。在实验中，可以选择任意 1 个领域数据作为目标领域，而其他的 3 个作为源领域。因此，给出两个大类 A 和 B，各自包含 4 个小类 $A_1,\cdots,A_4$ 和 $B_1,\cdots,B_4$，可以构造 $96(4\cdot P_4^4)$ 种 3 个源领域 1 个目标领域的迁移学习两类分类问题。

文本数据。本实验中仍然采用数据集 20*Newsgroups*，如第四章的表 4.5.1 所描述，该数据集具有两层结构，包含四个大类 *comp*，*rec*，*sci* 以及 *talk*，且每个大类下面都有对应的四个小类。我们构造了两个数据集 *sci vs. talk* 和 *comp vs. talk*，每个数据集可以构造 96 种 3 个源领域 1 个目标领域的两类分类问题，因此总共有 192 个分类问题。另外采用文档频率值 5 来选择词特征。

图像数据。对于图像数据集，我们从 COREL 图像库[9]中构造了一组数据集，该数据集包含两个大类 *flower* 和 *traffic*，其中包括 *flower* 包括 *flower.sunflower*，*flower.rose*，*flower.lotus* 以及 *flower.tulip* 四个小类，而 *traffic* 包括四个小类 *traffic.aviation*，*traffic.bus*，*traffic.boat* 和 *traffic.dogsled*。图 5.6.1 给出了每一小类中图像的示意图，表 5.6.1 给出了该图像数据集的详细描述。对于每个图像，提取 87 维的特征，包括 36 维颜色直方图特征[Zhang, 2001]以及 51 维 SILBP 纹理直方图特征[Shi, 2008a]。

[9] http://wang.ist.psu.edu/docs/related.shtml

图 5.6.1 图像数据集中每一小类图像的示意图

表 5.6.1 图像数据集的详细描述

flower	*flower.sunflower*	*flower.rose*	*flower.lotus*	*flower.tulip*
No. of samples	85	100	66	100
traffic	*traffic.aviation*	*traffic.bus*	*traffic.boat*	*traffic.dogsled*
No. of samples	100	100	100	100
No. of features	87	87	87	87

5.6.2 比较算法和实现细节

比较算法：本实验中比较的分类算法可以分成两类，把数据收集到中心节点的集中式算法和处理数据地理上分布的分布式算法。实验中采用逻辑回归算法[10][Davie, 2000]为监督学习算法。

分布式算法：可以分布式实现的算法有两种，第一种是最简单的分布式集成算法，即每个源领域训练得到一个分类器，然后把分类模型传到中心节点，进行等权重加权集成，表示为DE；第二种是本章提出的算法 5.2，所有优化后的子分类器也是通过等权重加权得到最后的结果，称之为DCR。

集中式算法：集中式算法指先把数据集中到中心节点，再处理的分类算法。第一种是把所有源领域数据合并在一起，训练得到得到一个模型，表示为CT；其实一致性正则化算法也可以处理单个源领域的情况($m=1$，不用计算平均概率预测向量)，我们把所有的领域数据合并起来，然后用算法 5.1 进行处理，称之为CCR_1；第三种就是算法 5.1 中的一致性正则化框架考虑多个源领域的情况，称之为CCR (实验中只考虑三个源领域的情况，也表示为CCR_3)，后面实验中将会分析CCR_1和CCR_3的区别；我们还比较了跨领域学习算法CoCC [Dai, 2007]，以及半监督学习算法TSVM [Joachims, 1999a]和

[10] http://research.microsoft.com/$\sim$minka/papers/logreg/

SGT [Joachims, 2003]。

从上面可以看到，本章提出的算法包括DCR，CCR_1以及CCR_3，比较算法包括DE，CT，CoCC，TSVM和SGT。这里需要注意的是当目标函数(5.9)中的参数θ的值为0时，一致性正则化不起作用，只相当于考虑各自源领域数据的上最大后验，DCR退化为DE，CCR_1退化为CT。另外算法5.2中如果取比较小的迭代步长常数γ，除了通信代价外，DCR与CCR_3预测分类结果基本相同，因此以下的实验结果中就只以CCR_3表示这两种算法。

实现细节：我们用几种非线性优化方法进行求解算法5.1，比如拟牛顿法，共轭梯度法等，结果表明最后得到的模型非常相似，但是共轭梯度法收敛速度最快，因此后续实验中都采用共轭梯度法进行优化。初步实验结果表明，一致性正则化框架(5.9)对参数$\lambda^l\,(l=1,\cdots,m)$不敏感，因此对于文本数据，设置$\lambda^l=145\,(l=1,\cdots,m)$；对于图像数据，设置$\lambda^l=0.05\,(l=1,\cdots,m)$。最大迭代次数设置$max$为200，误差率$\varepsilon=0.1$，控制一致性正则化作用的平衡参数$\theta$的取值范围为[0, 0.25]，本实验中间隔0.05采样。比较算法CoCC，TSVM和SGT的参数采用文献[Dai, 2007]的参数设置。

采用准确率衡量所有算法的预测性能。假设c为样本真实类别的映射函数，f是训练得到的分类器对样本进行预测的函数。那么评价准则准确率定义如下：

$$acc=\frac{|\{\mathbf{x}\,|\,\mathbf{x}\in D_t\wedge c(\mathbf{x})=f(\mathbf{x})\}|}{|D_t|} \tag{5.30}$$

当采用式子(5.4)定义一致性度量时，我们还定义了多个分类器对目标领域所有数据的预测一致性度量，若h是源领域中训练得到的初始分类器，它对应于把每个样本预测为相应类别的概率，那么m个分类器$h_1,\cdots,h_m$在目标领域所有数据上预测一致性的度量定义如下，

$$consensus=\frac{\sum_{\mathbf{x}\in D_t}\sqrt{C_s(h_1(\mathbf{x}),\cdots,h_m(\mathbf{x}))}}{|D_t|} \tag{5.31}$$

其中C_s由式子(5.4)定义，很明显当所有的分类器都分类一致且都是以概率1预测样本属于某一类时，则$consensus$达到最大值1。

为了突出本节中的丰富实验，首先总结以下实验中所要展示的结果，及其每部分的结果所要解决的问题：

一 为了验证本章提出的一致性正则化框架的有效性，1) 比较了算法CCR_3，DE和

CT；2) 比较了算法CCR_3， CoCC， TSVM 和SGT；

- 从实验上分析了为什么一致性正则化方法可以提高分类性能的来源，主要有，1) 在多源领域情况下，开发源领域之间的分布不同性对算法性能的提升具有很重要的作用；2) 一致性正则化框架不仅有最大化预测一致性的作用，还有最小化熵的作用(最小化熵的作用是使样本以最大概率被预测到某一类别)，这是另外两个来源。
- 一致性正则化算法是归纳式学习算法，可以泛化到新来的未标记样本，只要有足够的无标签目标领域数据用于训练；
- 最后，考察了共轭梯度下降法的收敛情况，一般要求迭代算法要有良好的收敛性质。

5.6.3 结果比较

算法CCR_3， DE 和CT 的比较：从前面的数据描述中可以看到实验中有三组数据，两组文本数据和一组图像数据，每组数据可以构造 96 个分类问题，而对于每个分类问题在每种θ参数设置下，我们都记录下一致性正则化方法的预测准确率和优化前的初始分类器的预测一致性度量值。表 5.6.2 中给出了数据集*sci vs. talk*中的一个分类问题的结果。这里没有列出所有96×3 个表格，但它们的结果基本类似。后面会分析所有所有问题的实验结果。

表 5.6.2 算法CCR_3和CCR_1的准确率(%)和一致性度量值

θ	The classifier on D_s^1		The classifier on D_s^2		The classifier on D_s^3		Consensus	CCR_3	CCR_1
	Acc. on D_s^1	Acc. on D_t	Acc. on D_s^2	Acc. on D_t	Acc. on D_s^3	Acc. on D_t		Acc. on D_t	Acc. on D_t
0	100	71.10	99.95	55.75	100	72.10	0.2775	73.94	71.99
0.05	100	92.14	99.95	90.61	99.94	93.20	0.6459	93.46	74.47
0.1	100	93.14	99.95	92.67	99.94	92.93	0.7385	93.30	76.11
0.15	100	93.83	99.95	93.46	99.89	93.88	0.7861	93.72	77.90
0.2	100	93.25	99.95	92.99	99.89	93.20	0.8146	93.25	80.43
0.25	100	93.35	99.95	92.99	99.89	93.14	0.8349	93.20	81.12

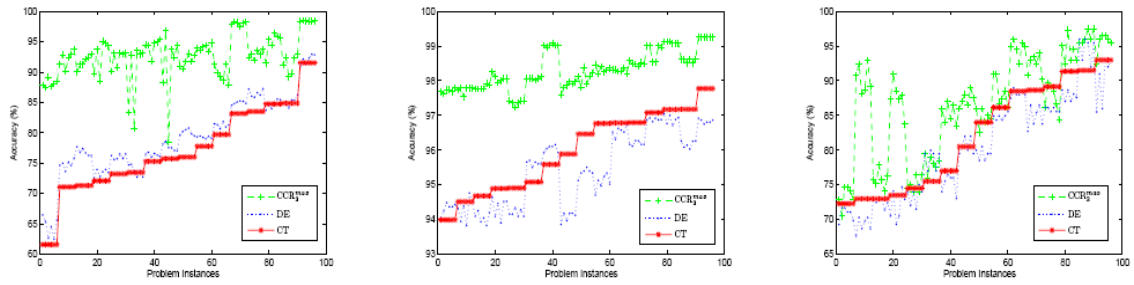
对于每一个θ值，算法一致性正则化得出 3 个分类器对应于相应的源领域，优化后得到的分类器分别对相应源领域的数据以及目标领域数据集进行测试，结果由表 5.6.2 中第 2 列到第 7 列给出；第 8 列给出了优化前三个初始分类器预测一致性的度量值，最后第 9，10 列分别给出了CCR_3(CCR_3是所有分类器等权重投票集成的结果)和CCR_1两种算法的准确率。如前面所述，当$\theta=0$(不考虑一致性正则化的作用)时，第 1 行第 9 列的准确率值 73.94%也是算法DE 的准确率；第 1 行第 10 列的准确率值 71.99%也是算法CT 的准确率。

从表 5.6.2 中可以看到，1) 当$\theta \neq 0$时，算法CCR_3总是比算法DE，CT都好，这充分说明CCR_3可以很好地处理迁移学习问题；1) 当$\theta \neq 0$时，优化后得到的分类器，在自己本身数据集上进行测试，总能得到很好的性能(接近于100%)，而且对目标领域的数据进行测试，都有非常大的提高。例如，分类器2在优化之前对目标领域的数据预测准确率只有55.75%，而优化后的准确率则达到90%以上，提高相当明显；3) 当θ值增大时，分类器预测一致性的度量值也随着增加，当达到一定值后，这些分类器对所有的样本都给出相同的分类结果，因此每个分类器与所有分类器集成的结果CCR_3给出的结果基本一致。

为了更有力地验证本文提出的算法的有效性，对所有96×3分类问题得到的96×3个表，每个表抽取4个统计值做比较：①算法DE的准确率；②算法CT的准确率；③CCR_3在$\theta \in [0.05, 0.25]$范围内采样得到的平均准确率，记为$\overline{CCR_3}$；④CCR_3在$\theta \in [0.05, 0.25]$范围内采样得到的最好准确率，记为CCR_3^{max}。对于以上每个表求得的4个值在对应数据集的96个问题上做平均，结果如图 5.6.2，图 5.6.3 和表 5.6.3 所示。图 5.6.2，图 5.6.3 中96个分类问题按照CT算的准确率值按从小到大排序。

表 5.6.3 三个数据集上对应96个问题的平均准确率(%)比较

Data set	CCR_3^{max}	$\overline{CCR_3}$	DE	CT
sci vs. talk	92.66	90.42	79.21	77.19
comp vs. talk	98.29	98.08	95.43	95.97
flower vs. traffic	87.54	85.89	80.50	81.97

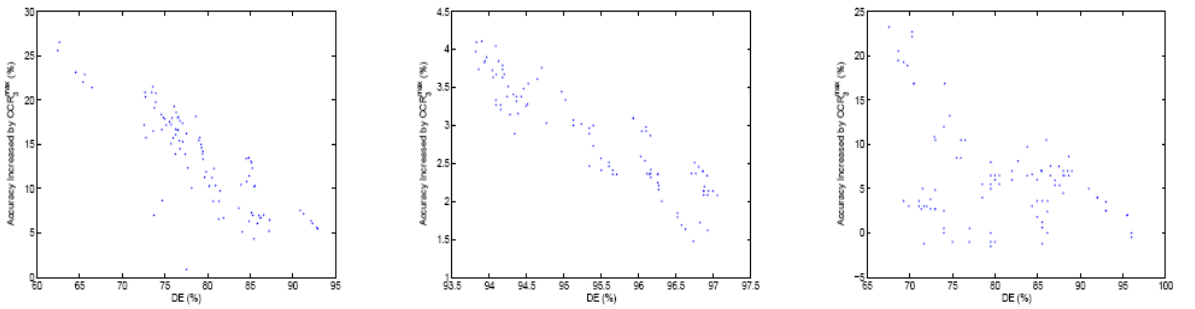


(a) CCR_3^{max} vs. DE and CT on *sci vs. talk* (b) CCR_3^{max} vs. DE and CT on *comp vs. talk* (c) CCR_3^{max} vs. DE and CT on *flower vs. traffic*

图 5.6.2 算法CCR_3^{max}，DE和CT在三个数据集上的性能(%)比较

从图 5.6.2 中可以看到CCR_3^{max}总是比DE和CT好，这说明算法CCR_3^{max}的有效性。在图 5.6.3 中，横坐标x轴表示算法DE的准确率，而纵坐标y轴表示CCR_3^{max}相对于DE

算法性能的提高程度。从图 5.6.3(a)和 图 5.6.3(b)中可以看到随着DE的准确率的上升，CCR_3^{max} 提高性能的程度逐渐下降，这与预测的目标相符合，因为当DE的准确率越高，最初的模型对目标领域数据预测的一致性程度就越高，这样优化的空间就越小，一致性正则化算法提高一致性程度就越小。虽然图像数据集的数据量比较少，但也表现出同样的趋势，如图 5.6.3(c)所示。表 5.6.3 中给出了三个数据集上 4 个统计值在 96 个分类问题上的平均值，可以看到算法CCR_3^{max} 总是表现出最好的分类性能。而且相对于DE，准确率在数据集 *sci vs. talk* 上，从 79.21%提高到 92.66%。



(a) DE vs. CCR_3^{max} on *sci vs. talk* (b) DE vs. CCR_3^{max} on *comp vs. talk* (c) DE vs. CCR_3^{max} on *flower vs. traffic*

图 5.6.3 三个数据集上CCR_3与DE之间的关系

算法CCR_3，CoCC，TSVM和SGT的比较：为了实现这些算法之间的比较，采用了文献[Dai, 2007]中的数据，并改造该数据集使得其满足多个源领域的情形，即把单个源领域的数据集划分成多个源领域数据集，数据集划分的详细描述如表 5.6.4 所示。

表 5.6.4 算法CCR_3，CoCC，TSVM和SGT比较中的数据描述

Data Set	D_s		D_t
	D_s^1	D_s^2	
comp vs. sci	*comp.graphics sci.crypt*	*comp.os.ms-windows.misc, sci.electronics*	*comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, sci.med, sci.space*
rec vs. talk	*rec.autos talk.politics.guns*	*rec.motorcycles talk.politics.misc*	*rec.sport.baseball, talk.religion.misc, rec.sport.hockey, talk.politics.mideast*
rec vs. sci	*rec.autos sci.space*	*rec.sport.baseball sci.med*	*rec.motorcycles, rec.sport.hockey, sci.crypt, sci.electronics*
sci vs. talk	*sci.electronics talk.religion.misc*	*sci.med talk.politics.misc*	*sci.crypt, talk.politics.guns, sci.space, talk.politics.mideast*

实验结果如表 5.6.5 所示。从表 5.6.5 中可以看到在这些分类任务中，除第 4 种情况

算法CoCC比CCR_3稍微好点，其它情况下$\overline{CCR_3}$，CCR_3^{max}都比算法CoCC准确率高。所有的情况下，CCR_3都优于算法TSVM和SGT。

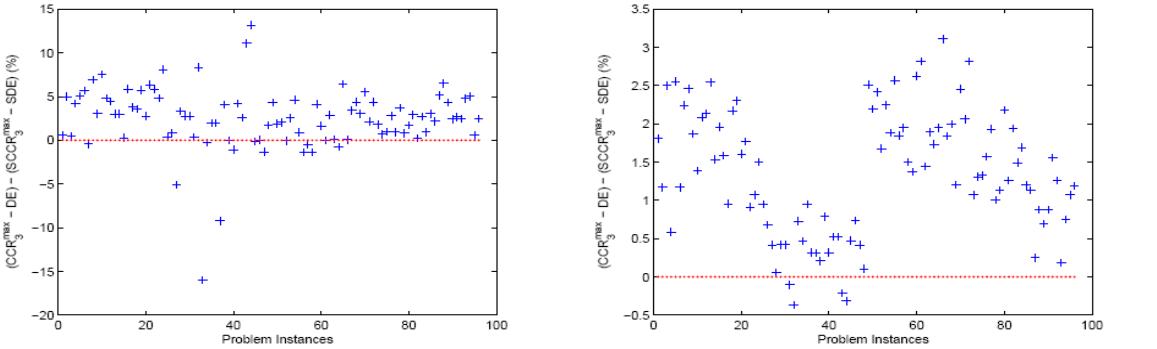
表 5.6.5 算法CCR_3，CoCC，TSVM和SGT之间准确率(%)的比较

Data Set	TSVM	SGT	CoCC	$\overline{CCR_3}$	CCR_3^{max}
comp vs. sci	81.7	72.1	87.0	91.4	93.1
rec vs. talk	96.0	90.9	96.5	97.8	98.0
rec vs. sci	93.8	93.8	94.5	96.3	96.7
sci vs. talk	89.2	91.7	94.6	92.8	93.6

5.6.4 性能提高的来源

本小节从实验上分析为什么一致性正则化方法可以提高分类性能的来源。

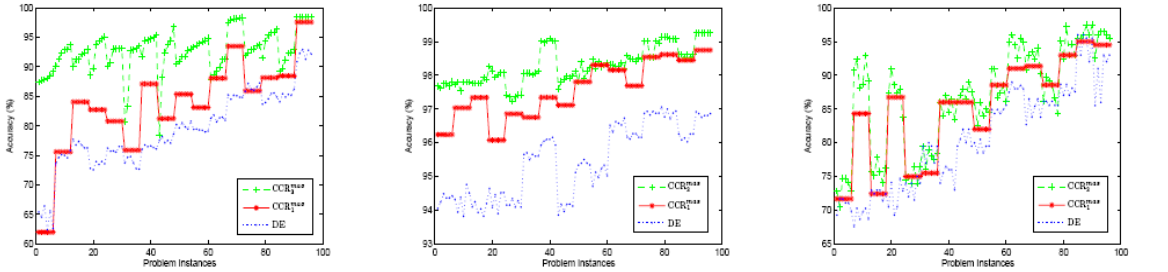
有效开发源领域之间的分布不同性。在一致性正则化框架中，不同源领域之间的不同性非常重要，但是通常很难精确度量领域之间的分布不同性。为了简化讨论，我们主要考虑两个层次的分布不同程度，1) 不同源领域之间具有很大的分布不同性，如 5.6.1 节中源领域的构造；2) 不同源领域之间具有很小的分布不同性。如果把 5.6.1 节中所有的源领域数据合并在一起称为D_s，然后重新随机划分该数据集成m (这里$m=3$)个源领域，表示为$SD_s^1,\cdots,SD_s^m$。可以看到合并后又重新划分的源领域，它们之间的分布不同性大大减少了。用本章提出的一致性正则化框架对数据$SD_s^1,\cdots,SD_s^m$进行处理，同样对每个表抽取 4 个值，分别表示为SDE，SCT，$SCCR_3^{max}$以及$\overline{SCCR_3}$，其中SCT=CT。图 5.6.4 给出了合并前后一致性正则化算法性能提高的差异，可以看到合并前算法性能提高程度比合并后重新划分的情况要好很多，*t*-测试以 95%的置信度表明这种优越性是统计突出的。结果表明有效开发领域之间的分布不同性可以提高学习性能。



(a) Problem Instances vs. $(CCR_3^{max}$ - DE) - $(SCCR_3^{max}$ - SDE) on *sci vs. talk*　(b) Problem Instances vs. $(CCR_3^{max}$ - DE) - $(SCCR_3^{max}$ - SDE) on *comp vs. talk*

图 5.6.4 有效开发源领域之间分布不同性。源领域数据合并前后一致性正则化算法性能提高的差异

最大化预测一致性以及最小化熵。如前面定义一致性度量准则时描述，一致性正则化具有两个作用，最大化预测一致性以及最小化熵。这里比较了算法CCR_3^{max}，CCR_1^{max}以及DE(CT的性能与DE类似)。在算法CCR_1^{max}中，只有一个源领域，因此只有最小熵作用，而没有最大化预测一致性的作用。三个数据集上的结果如图 5.6.5 所示。

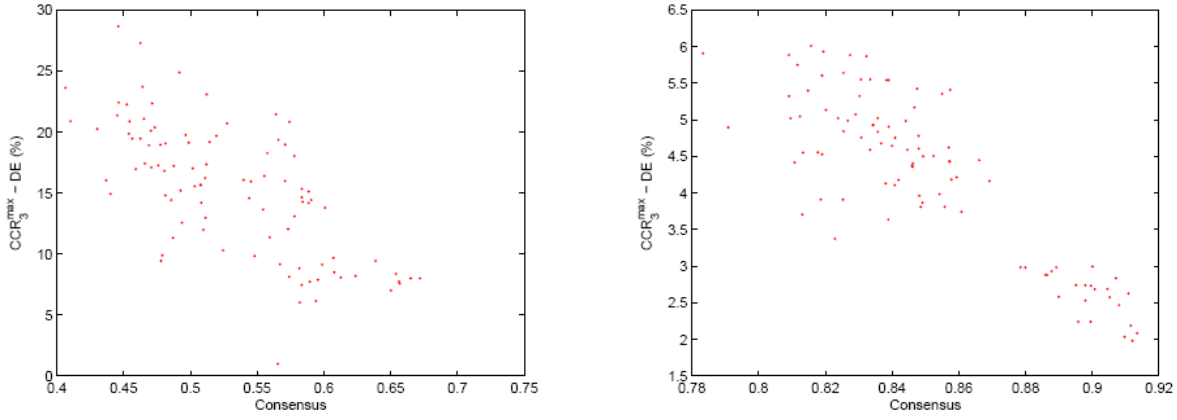


(a) CCR_3^{max} vs. CCR_1^{max} and DE on *sci vs. talk* (b) CCR_3^{max} vs. CCR_1^{max} and DE on *comp vs. talk* (c) CCR_3^{max} vs. CCR_1^{max} and DE on *flower vs. traffic*

图 5.6.5 算法CCR_3^{max}，CCR_1^{max}和DE在三个数据集上的比较

可以看到，1) 算法CCR_1^{max}优于DE，表明最小化熵是有效的；2) 算法CCR_3^{max}大大优越于CCR_1^{max}，这充分说明了最大化预测一致性的有效性。实验再一次验证了一致性正则化框架应用于多源领域是非常有效的。

一致性与算法性能提高之间的关系。为了突出一致性正则化算法的主题，我们考察了初始分类器一致性与性能提高之间的关系，结果如图 5.6.6 所示。横坐标x轴表示初始分类模型在目标领域数据上的预测一致性，而纵坐标y轴表示CCR_3^{max}相对于DE算法性能的提高。从图 5.6.6 中可以看到性能提高量随着一致性度量值的增加而减小，即一致性程度越高，优化空间越小。因此从某种程度上要求初始分类器具有多样性，这与提升(Boosting)算法的要求一致，比如 Bagging[Leskes, 2008]，Adaboost[Dietterich, 2000]等。

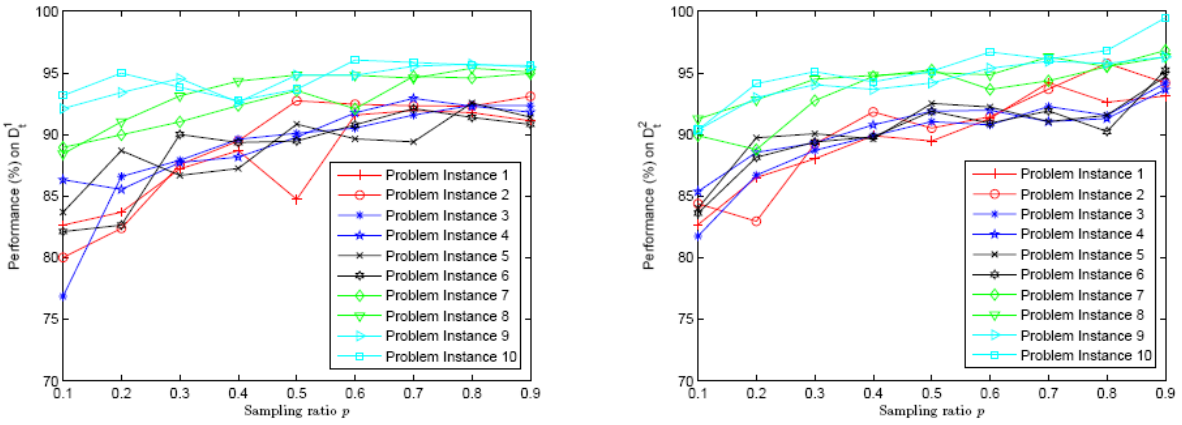


(a) Consensus vs. CCR_3^{max} - DE on *sci vs. talk* (b) Consensus vs. CCR_3^{max} - DE on *comp vs. talk*

图 5.6.6 一致性与算法性能提高之间的关系

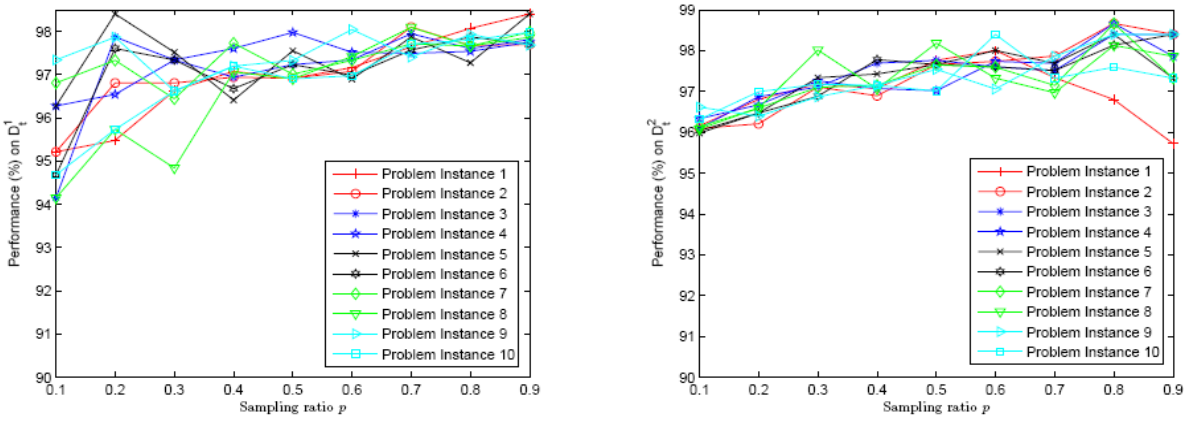
5.6.5 归纳式学习算法

一致性正则化算法是归纳式的，因此可以产生最终分类器对新来的样本进行预测。为了验证算法CCR_3^{max}在新来的数据集上的性能，采用文本数据作为实验数据，分别从数据集 *sci vs. talk* 和 *comp vs. talk* 中各随机选择 10 个分类问题。对于每个分类问题中的目标领域无标签数据，我们随机采样(无放回采样)比例为 p 的数据构成新的数据集 D_t^1，剩下的构成数据集 D_t^2。数据 D_t^1 用于训练过程，而数据 D_t^2 用于测试优化后得到的算法。实验中记录了算法CCR_3^{max}在数据集 D_t^1 和 D_t^2 上准确率，同时也记录了不同采样比例 p (其中 $p\in[0.1,0.9]$，且以间隔 0.1 采样)下算法CCR_3^{max}的分类性能，所有的结果如图 5.6.7 和图 5.6.8 所示。



(a) Performance of CCR_3^{max} on *sci vs. talk*　　(b) Performance of CCR_3^{max} on *sci vs. talk*

图 5.6.7 在不同的采样比例 p 下，算法CCR_3^{max}在数据集 *sci vs. talk* 上的泛化能力



(a) Performance of CCR_3^{max} on *comp vs. talk*　　(b) Performance of CCR_3^{max} on *comp vs. talk*

图 5.6.8 在不同的采样比例 p 下，算法CCR_3^{max}在数据集 *comp vs. talk* 上的泛化能力

从图 5.6.7 和图 5.6.8，我们看到，1) 用于训练的无标签数据集 D_t^1 越大，则优化后输出的分类器泛化能力越强；2) 当 $p \geq 0.6$ 时，算法 CCR_3^{max} 在数据集 D_t^1 和 D_t^2 上的分类准确率基本上达到一致。该结果表明，算法 CCR_3^{max} 在大于 60%的样本用于训练时，对新来的测试数据能表现出很好的泛化能力。

5.6.6 算法收敛性

为了考察一致性正则化算法 CCR_3 的收敛性，我们选择了数据集 *sci vs. talk* 中的 6 个分类问题作为实验数据。图 5.6.9 展示了其实验结果，其中横坐标 x 轴表示迭代次数，而纵坐标 y 轴表示算法 CCR_3 在参数 $\theta = 0.25$ 时的准确率。从图中可以看到算法的性能随着不断迭代而提高，收敛速度也非常快，基本上可以在 30 次内收敛。这充分说明了该算法具有很好的收敛性质。

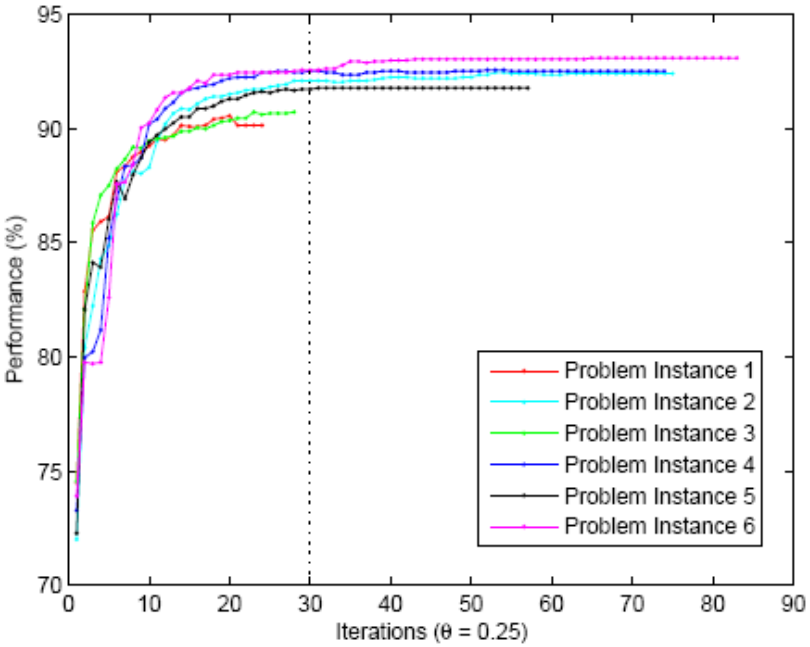


图 5.6.9 算法 CCR_3 的收敛性

5.7. 小结

本章对多源领域到单个目标领域的迁移学习问题进行了研究，其中这些源领域数据与目标领域虽然数据分布不同，但是语义上相关。基于此，提出了一致性正则化框架，有效地挖掘多源领域之间那种分布不同性来提升目标领域数据上的分类预测性能。在该框架下，我们设计了一个分布式学习算法，其中局部的子分类器不仅考虑了在源领域上的可利用的局部数据，而且考虑了这些由源领域知识得到的子分类器在目标领域上的预

测的一致性。更进一步，为了处理各个源领域数据在地理上分布的情况，提出了一致性正则化的分布式实现，可避免收集各个领域数据到中心节点，而只是传递一些统计变量，一定程度上减轻了数据信息的隐私性担忧。还从理论上分析了一致性正则化的有效性。最后，在三个数据集的实验充分验证了本章提出算法的有效性。系统的实验还分析了算法提升性能的来源，算法的收敛性等。

但是一致性正则化算法还不能处理多个目标领域的情况，因此下一章将介绍能够同时处理多个源领域和多个目标领域的跨领域学习算法。

第六章　基于生成模型的挖掘多领域之间共性与特性的跨领域分类方法

6.1. 引言

如第四章所述，在迁移学习文本分类中，源领域数据与目标领域数据在原始词特征上分布不一致，也就是说它们可能会采用不同的词特征来表示同一个语义概念。但不同的领域数据，其词特征聚类(又称词特征概念)与文档类别(又称文档聚类)之间的关联关系可能一样。第四章提出基于非负矩阵分解的跨领域学习算法，有效地挖掘词特征聚类与文档聚类之间的关联关系，从而实现准确地对目标领域数据进行预测。但基于矩阵分解的方法缺乏概率解释，且不容易同时处理多源领域和多目标领域的情况。本章进一步开发高层语义词特征概念，提出基于生成模型的挖掘多领域之间共性(Commonality)与特性(Distinction)的跨领域分类学习方法，具有比较完美的概率解释，而且能够同时处理多个源领域和多个目标领域数据。

对于词特征概念，可以表示为关于词特征的多项式分布 $p(w|y)$，而且该分布通常依赖于不同的领域数据。比如给定词特征概念“product”，如果该概念来自于领域 HP 公司，则其概率 $p("printer"|"product")$ 以及 $p("LaserJet"|"product")$ 就会比较大，因为 HP 公司生产打印机；相反，如果该概念来自于生产笔记本的 IBM 公司，那么概率值 $p("printer"|"product")$、$p("LaserJet"|"product")$ 很小，而概率 $p("thankpad"|"product")$ 和 $p("thankcenter"|"product")$ 会很大。在后面的实验中，我们列出了不同领域中表示词特征概念的关键词特征，可以发现不同领域用不同的词特征来描述同一词特征概念。我们还发现不管词特征概念出现在哪里，它都会指向与该词特征概念有密切联系的文档类。也就是说，只要一个新闻文档包含词特征概念“product”，那么不管该新闻文档属于哪个领域，HP 公司或者 IBM 公司，该文档都极有可能属于类别“product announcement”，而不是“financial scandal”。换句话说，词特征概念 y 与文档类别 z 之间的联合概率 $p(y,z)$ 是领域独立的，不依赖于领域数据。

从上面的例子中可以看到，$p(w|y)$ 和 $p(y,z)$ 分别对于词特征概念 y 的两个层面，即外延(Extension)和内涵(Intension)。通常，概念的外延是指独立的对象各自拥有的特征，而概念内涵是指所有对象共享的一些特征[11]。根据概念外延和内涵的一般定义，以下给

[11] http://www.philosophypages.com/lg/e05.htm. In general, the extension of a concept is just the collection of individual objects to which it is correctly applied, while the intension of a concept is the set of features which are shared by everything to which it applies.

出词特征概念外延和内涵的定义，

定义 6.1 (*Extension of Word Concept*) 词特征概念 y 的外延是指一个词 w 能够表示该词特征概念的程度，表示为 $p(w|y)$。

也就是说概率 $p(w|y)$ 的值越大，词 w 越能表示该词特征概念 y。

定义 6.2 (*Intension of Word Concept*) 词特征概念 y 的内涵解释为它与文档类别 z 之间的关系，用 $p(y,z)$ 表示词特征概念与文档类别之间的联合概率分布。

也就是说概率 $p(y,z)$ 越大，那么一个文档如果包含词特征概念 y，它属于文档类别 z 的概率就越大。这样高层语义特征词概念特征 y 可以看成类别 z 的固有特征，与领域无关。类似词特征概念的定义，同样可以定义文档概念(document concept)的外延和内涵，分别为 $p(d|z)$ (关于文档的多项式分布)和 $p(y,z)$。在本章中，只考虑一个文档类别就等价于一个文档概念(通常一个文档类别可以有多个文档概念)，因此在本章中文档概念与文档类别交替使用。为使得本章定义的几个术语“Distinction”，“Commonality”，“Extension”以及“Intension”，更加清晰和易于理解，总结它们之间的关系于表 6.1.1。从表中可以看到，领域特性包括词特征概念的外延和文档概念的外延，而领域共性包含词特征概念的内涵和文档概念的内涵。如前面的例子，对于领域 HP 公司和 IBM 公司，它们的特性为 1) 描述词特征概念“product”的关键词特征不一样；2) 在文档类别“product announcement”中的新闻文档不一样。另一方面，领域共性就是共享的词特征概念与文档类别之间的联合概率，如 $p(\text{"product","product announcement"})$。

表 6.1.1 术语“Distinction”，“Commonality”，“Extension”和“Intension”之间的关系

Domain Distinction		Domain Commonality	
Extension of word concept $p(w\|y)$	Extension of document concept $p(d\|z)$	Intension of word concept $p(y,z)$	Intension of document concept $p(y,z)$

通过以上定义，我们认为在一般情况下，词特征概念和文档概念的外延依赖于不同领域，而其内涵则独立于领域本身。因此本章的目标是挖掘领域之间的这种共性和特性，从而实现对目标领域数据的正确分类。

首先，我们提出一个统计生成模型协同对偶 PLSA(Collaborative Dual-PLSA，简称 CD-PLSA)，来同时挖掘领域之间的共性和特性。CD-PLSA 模型的主要思想如图 6.1.1 所示。有 s 个源领域和 t 个目标领域(s 和 t 为任意正整数)，表示为图 6.1.1 中大的矩形框，

左边为源领域，右边为目标领域，c_i 表示第 i 个领域。每个大矩形框中又包含两个小的矩形框，分别为各个领域词特征概念的外延和文档概念的外延。领域的特性包括所有的外延，而领域的共性则是它们共享的词特征概念与文档概念之间的联合概率分布，即图中的六边形所示。实际上源领域中的数据是有标记的，即源领域中文档概念的外延已知，可以作为整个模型的监督信息，如图中的实心圆圈所示。这些监督信息通过领域之间的共性实现知识的迁移，领域的共性起到桥的作用，最后实现对目标领域数据的分类预测。为了求解 CD-PLSA 模型，我们提出了一个 EM 算法。更进一步，为了处理各个领域数据是地理上分布的情况，也提出了 EM 算法的分布式实现，不用传递原始数据，只需要传递中间一些统计变量，一定程度上保护数据的隐私性。

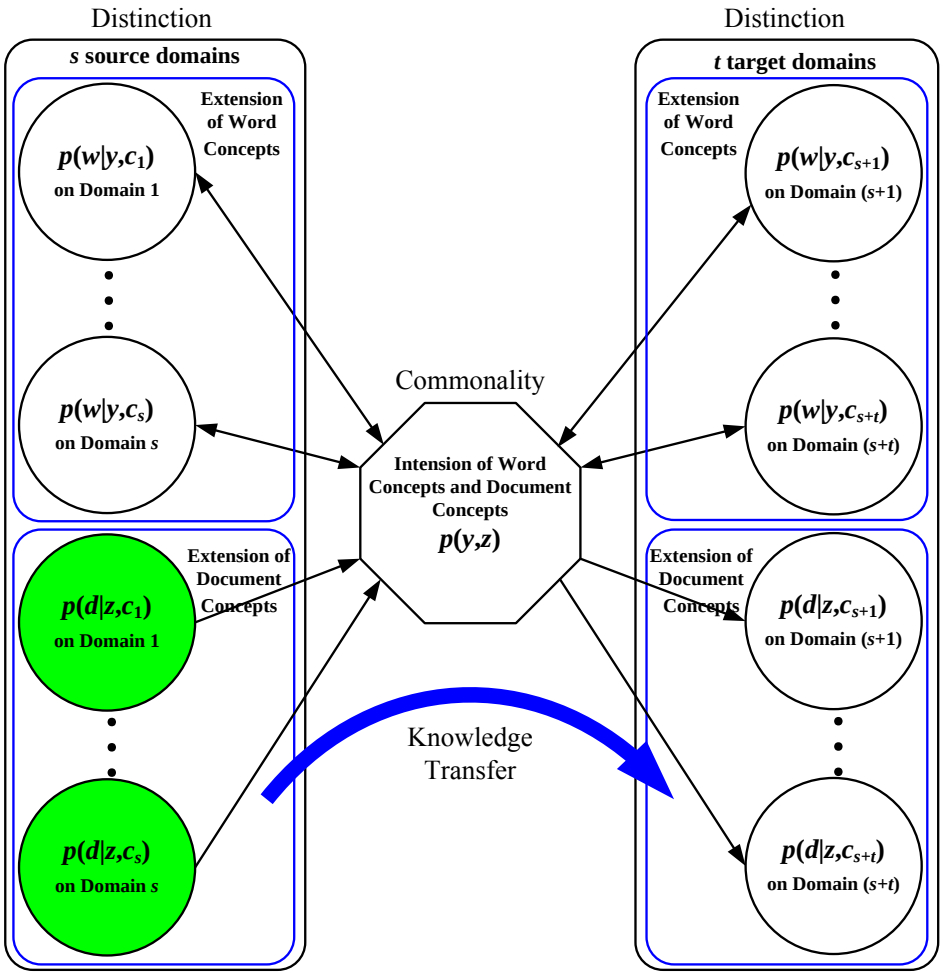


图 6.1.1 不同领域概念的外延与内涵

在第二阶段，我们进一步开发目标领域自身的内部结构。通过第一阶段对 CD-PLSA 模型的求解，可以得到所有领域共享的词特征概念以及文档概念的内涵 $p(y,z)$。实际上由于领域之间的分布不一致性，求解 CD-PLSA 模型得到的各个领域共享的内涵 $p(y,z)$，并不一定是各个目标领域的内涵。因此在第二阶段，我们提出进一步精化第一步的结果，

不过只用各个目标领域自身的数据。总之，我们提出两阶段的跨领域文本分类算法，第一阶段在所有的领域数据上协同训练一个生成模型(CD-PLSA)，产生结果领域共性 $p(y,z)$，以及领域特性 $p(w|y)$， $p(d|z)$。第二阶段用各个目标领域自身的数据，对第一阶段相对应的输出进行精化，称这个两阶段的方法为 RCD-PLSA。

6.2. 预备知识和问题形式化

本节首先介绍了概率隐性语义分析模型(PLSA)，以及其扩展模型，对偶概率隐性语义分析模型(Dual-PLSA，简称 DPLSA)。然后对本章提出的跨领域文本分类问题进行形式化。

6.2.1 预备知识

概率隐性语义分析模型(PLSA)[Hofmann, 1999]是对共现数据矩阵进行分析的统计模型。给定词—文档同现矩阵 $\mathbf{O}$，其元素 $\mathbf{O}_{w,d}$ 表示词特征 w 在文档 d 中出现的频率，那么 PLSA 使用混合模型对 $\mathbf{O}$ 进行建模，其中隐性主题(Latent topic)表示为变量 y，

$$p(w,d)=\sum_{y}p(w|y)p(d|y)p(y) \tag{6.1}$$

图 6.2.1(a)给出了 PLSA 模型的图模型，在所有变量 w， d， y 的参数 $p(w|y)$，$p(d|y)$ 以及 $p(y)$ 可以通过求解最大似然问题得到，一般采用 EM 算法求解。

在 PLSA 模型中，文档和词特征共享相同的隐性变量 y，其实在实际中，文档和词特征通常表现不同的组织和结构，因此它们可能需要不同的隐性变量来进行刻画，例如用隐性变量 y 来刻画词特征，隐性变量 z 来刻画文档。图模型如图 6.2.1(b)所示，由于包含两个隐性变量，称为对偶 PLSA(本章中简称 D-PLSA)。

给定词—文档同现矩阵 $\mathbf{O}$，类似式子(6.1)，得到以下混合模型，

$$p(w,d)=\sum_{y,z}p(w,d,y,z)=\sum_{y,z}p(w|y)p(d|z)p(y,z) \tag{6.2}$$

其中关于所有变量 w， d， y， z 的参数 $p(w|y)$， $p(d|z)$ 以及 $p(y,z)$ 同样可以由 EM 算法求解最大似然问题得到。其实这些参数中 $p(w|y)$ 和 $p(d|z)$ 分别为词特征概念 y 以及文档概念 z 的外延，而且 $p(y,z)$ 是它们的内涵。

D-PLSA 由文献[Jiho, 2009]提出，并且用于聚类。其实我们发现该模型中，词特征概念与文档概念有各自的隐性变量，因此可以很容易通过文档概念 z 的外延 $p(d|z)$ 注入

标签信息，在本章中，z 其实是文档的类别。这样，D-PLSA 可以是一个半监督的分类模型。在实验中还会详细描述该模型。

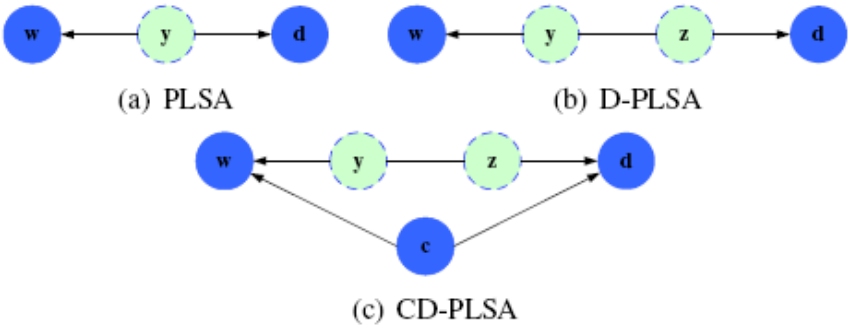


图 6.2.1 图模型 PLSA，D-PLSA 和 CD-PLSA

6.2.2 问题形式化

基于 D-PLSA，我们提出一个统计生成模型处理多领域数据的跨领域文本分类算法。给出 $s+t$ 个领域数据，表示为 $D=(D_1,\cdots,D_s,D_{s+1},\cdots,D_{s+t})$，不失一般性，假设前面 s 个领域为带标签的源领域数据，而后 t 个领域为无标签的目标领域数据。简单地，对于每个领域都可以得到领域自身词特征概念和文档概念的外延和内涵，这样就可以得到 $s+t$ 个概念的内涵。为了使得所有的领域共享词特征概念和文档概念的内涵，即只得到一个内涵，那么就需要隐性变量 y 和 z 独立于变量 c (表示数据领域)，因此我们提出了图 6.2.1(c)的图模型，满足以下要求 1) 隐性变量 y 和 z 条件独立于变量 c；2) 词特征 w 依赖于变量 y 和 c；3) 文档 d 依赖于变量 z 和 c。根据图 6.2.1(c)中的图模型，所有变量的联合概率分布如下，

$$p(w,d,y,z,c)=p(w|y,c)p(d|z,c)p(y,z)p(c) \tag{6.3}$$

共现矩阵 $\mathbf{O}_c$ 表示第 c 个领域的词-文档共现矩阵，其元素 $\mathbf{O}_{w,d,c}$ 表示三元组 (w,d,c)，第 c 个领域中词 w 在文档 d 中出现的频率。若 $\mathbf{Z}$ 表示潜在变量 y,z，$\mathbf{X}$ 表示所有领域数据，则形式化问题为最大对数似然如下，

$$\log p(\mathbf{X}|\theta)=\log\sum_{\mathbf{Z}} p(\mathbf{Z},\mathbf{X}|\theta) \tag{6.4}$$

其中 θ 包括所有的参数 $p(y,z)$，$p(w|y,c)$，$p(d|z,c)$ 和 $p(c)(1\le c\le s+t$，总共有 $s+t$ 个领域数据)。

这里需要注意，尽管各个领域词特征概念 y 的外延 $p(w|y,c)$ 不同，但如第四章所述，

它们在一定程度上语义相关，这样才能协同训练得到一个共享的概念内涵 $p(y,z)$。实验中将会直观上展示不用领域词特征概念的外延既不同又相关。因为本章提出的模型是协同训练 D-PLSA，所以称我们的模型为协同对偶 PLSA，简称 CD-PLSA。下一节介绍 EM 算法求解式子(6.4)中的最大对数似然问题。

6.3. EM 算法求解

6.3.1 EM 算法

EM(Expectation-Maximization)算法[Dempster, 1977; Broman, 2004]通过最大化式子(6.4)的下界 L_0 (根据 Jensen 不等式)，

$$L_0 = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\{\frac{p(\mathbf{Z},\mathbf{X}\mid\theta)}{q(\mathbf{Z})}\} \tag{6.5}$$

其中 $q(\mathbf{Z})$ 可以是任意函数，这里取 $q(\mathbf{Z}) = p(\mathbf{Z}\mid\mathbf{X};\theta^{\text{old}})$，并把它代入式子(6.5)得到，

$$\begin{aligned} L_0 &= \sum_{\mathbf{Z}} p(\mathbf{Z}\mid\mathbf{X};\theta^{\text{old}}) \log p(\mathbf{Z},\mathbf{X}\mid\theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}\mid\mathbf{X};\theta^{\text{old}}) \log p(\mathbf{Z}\mid\mathbf{X};\theta^{\text{old}}) \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}\mid\mathbf{X};\theta^{\text{old}}) \log p(\mathbf{Z},\mathbf{X}\mid\theta) + const \end{aligned} \tag{6.6}$$

从式子(6.6)可以看到，我们只需要考虑对数联合概率 $q(\mathbf{Z}) = p(\mathbf{Z},\mathbf{X};\theta)$ 和隐性变量的后验概率 $q(\mathbf{Z}) = p(\mathbf{Z}\mid\mathbf{X};\theta^{\text{old}})$。根据图 6.2.1(c)以及 *d*-分割(*d*-separation)准则，

$$\log p(\mathbf{Z},\mathbf{X}\mid\theta) = \log \prod_n p(\mathbf{Z}_n,\mathbf{X}_n\mid\theta) = \sum_n \log p(\mathbf{Z}_n,\mathbf{X}_n\mid\theta) \tag{6.7}$$

其中 $\mathbf{X}_n$ 和 $\mathbf{Z}_n$ 分别表示 $\mathbf{X}$ 和 $\mathbf{Z}$ 的第 n 个元素。类似地，

$$p(\mathbf{Z}\mid\mathbf{X};\theta) = \prod_m p(\mathbf{Z}_m\mid\mathbf{X};\theta) = \prod_m p(\mathbf{Z}_m\mid\mathbf{X}_m;\theta) \tag{6.8}$$

令 L 为式子(6.6)中的非常数项，则

$$\begin{aligned} L &= \sum_{\mathbf{Z}} p(\mathbf{Z}\mid\mathbf{X};\theta^{\text{old}}) \log p(\mathbf{Z},\mathbf{X}\mid\theta) = \sum_{\mathbf{Z}} \prod_m p(\mathbf{Z}_m\mid\mathbf{X}_m;\theta^{\text{old}}) \sum_n \log p(\mathbf{Z}_n,\mathbf{X}_n\mid\theta) \\ &= \sum_n \sum_{\mathbf{Z}} \prod_m p(\mathbf{Z}_m\mid\mathbf{X}_m;\theta^{\text{old}}) \log p(\mathbf{Z}_n,\mathbf{X}_n\mid\theta) \\ &= \sum_n \sum_{\mathbf{Z}_n} \sum_{\mathbf{Z}_{-n}} \prod_{m\neq n} p(\mathbf{Z}_m\mid\mathbf{X}_m;\theta^{\text{old}}) \cdot p(\mathbf{Z}_n\mid\mathbf{X}_n;\theta^{\text{old}}) \log p(\mathbf{Z}_n,\mathbf{X}_n\mid\theta) \\ &= \sum_n \sum_{\mathbf{Z}_n} p(\mathbf{Z}_n\mid\mathbf{X}_n;\theta^{\text{old}}) \log p(\mathbf{Z}_n,\mathbf{X}_n\mid\theta) \cdot \sum_{\mathbf{Z}_{-n}} \prod_{m\neq n} p(\mathbf{Z}_m\mid\mathbf{X}_m;\theta^{\text{old}}) \\ &= \sum_n \sum_{\mathbf{Z}_n} p(\mathbf{Z}_n\mid\mathbf{X}_n;\theta^{\text{old}}) \log p(\mathbf{Z}_n,\mathbf{X}_n\mid\theta) \end{aligned} \tag{6.9}$$

现在把观测数据 $\mathbf{X}$ 的每一项 $\mathbf{X}_n$ 即 (w,d,c)，以及隐性变量 $\mathbf{Z}$ 的每一项 $\mathbf{Z}_n$ 即 (y,z) 写成详细的形式，

$$\begin{aligned} L &= \sum_{w,d,c} \mathbf{O}_{w,d,c} \sum_{y,z} p(y,z \mid w,d,c;\theta^{\text{old}}) \cdot \log p(y,z,w,d,c \mid \theta) \\ &= \sum_{y,z,w,d,c} \mathbf{O}_{w,d,c}\ p(y,z \mid w,d,c;\theta^{\text{old}}) \cdot \log[p(y,z)p(w \mid y,c)p(d \mid z,c)p(c)] \end{aligned} \tag{6.10}$$

其中 $\mathbf{O}_{w,d,c}$ 表示词，文档以及领域这三者的同现频率。那么可以构造 EM 算法的 E 步如下，

$$p(y,z \mid w,d,c;\theta^{\text{old}}) = \frac{p(y,z)p(w \mid y,c)p(d \mid z,c)p(c)}{\sum_{y,z} p(y,z)p(w \mid y,c)p(d \mid z,c)p(c)} \tag{6.11}$$

以下开始构造 EM 算法的 M 步，最大化函数 L，首先抽取关于参数 $p(w \mid y,c)$ 的项，

$$L_{[p(w|y,c)]} = \sum_{y,z,w,d,c} \mathbf{O}_{w,d,c}\ p(y,z \mid w,d,c;\theta^{\text{old}}) \cdot \log p(w \mid y,c) \tag{6.12}$$

通过拉格朗日函数，加入约束 $\sum_{w} p(w \mid y,c) = 1$ 得到，

$$\frac{\partial[L_{[p(w|y,c)]} + \lambda(1 - \sum_{w} p(w \mid y,c))]}{\partial p(w \mid y,c)} = 0 \tag{6.13}$$

则

$$p(w \mid y,c) = \frac{\sum_{z,d} \mathbf{O}_{w,d,c}\ p(y,z \mid w,d,c;\theta^{\text{old}})}{\lambda} \tag{6.14}$$

考虑前面的约束 $\sum_{w} p(w \mid y,c) = 1$，所以

$$\begin{aligned} 1 = \sum_{w} p(w \mid y,c) &= \frac{\sum_{w}\sum_{z,d} \mathbf{O}_{w,d,c}\ p(y,z \mid w,d,c;\theta^{\text{old}})}{\lambda} \\ \Rightarrow \lambda &= \sum_{w}\sum_{z,d} \mathbf{O}_{w,d,c}\ p(y,z \mid w,d,c;\theta^{\text{old}}) \end{aligned} \tag{6.15}$$

这样可得到参数 $p(w \mid y,c)$ 的迭代公式如下，

$$\hat{p}(w \mid y,c) = \frac{\sum_{z,d} \mathbf{O}_{w,d,c} p(y,z \mid w,d,c;\theta^{\text{old}})}{\sum_{z,w,d} \mathbf{O}_{w,d,c} p(y,z \mid w,d,c;\theta^{\text{old}})}. \tag{6.16}$$

类似地，

$$\hat{p}(d \mid z,c) = \frac{\sum_{y,w} \mathbf{O}_{w,d,c} p(y,z \mid w,d,c;\theta^{\text{old}})}{\sum_{y,w,d} \mathbf{O}_{w,d,c} p(y,z \mid w,d,c;\theta^{\text{old}})} \tag{6.17}$$

$$\hat{p}(y,z)=\frac{\sum_{w,d,c}\mathbf{O}_{w,d,c}p(y,z\mid w,d,c;\theta^{\text{old}})}{\sum_{y,z,w,d,c}\mathbf{O}_{w,d,c}p(y,z\mid w,d,c;\theta^{\text{old}})} \tag{6.18}$$

$$\hat{p}(c)=\frac{\sum_{y,z,w,d}\mathbf{O}_{w,d,c}p(y,z\mid w,d,c;\theta^{\text{old}})}{\sum_{y,z,w,d,c}\mathbf{O}_{w,d,c}p(y,z\mid w,d,c;\theta^{\text{old}})} \tag{6.19}$$

6.3.2 注入监督信息

这一小节介绍怎么注入源领域中的标签信息。我们需要解决两个问题，1) 怎样加入源领域中的标签信息；2) 算法收敛得到所有的参数后，如何对目标领域数据进行分类。

不失一般性，假设前 s 个领域为源领域，则可以通过对概率 $p(d\mid z,c)(1\le c\le s)$的赋值而达到加入标签信息的目的。若 $\mathbf{L}^c\in[0,1]^{n_c\times m}$ 为第 c 个领域的真实标签信息，n_c 为第 c 个领域的样本数，m 为类别数。当文档 d 属于类别 z_0 时 $\mathbf{L}^c_{d,z_0}=1$，否则 $\mathbf{L}^c_{d,z}=0\,(z\ne z_0)$。我们使用式子(6.20)归一化 $\mathbf{L}^c$ 使其满足概率条件，并把它赋值给 $p(d\mid z,c)(1\le c\le s)$，

$$\mathbf{N}^c_{d,z}=\frac{\mathbf{L}^c_{d,z}}{\sum_d\mathbf{L}^c_{d,z}} \tag{6.20}$$

在迭代过程中，$p(d\mid z,c)(1\le c\le s)$保持不变，因为其包含源领域的标签信息，而 $p(d\mid z,c)(s+1\le c\le s+t)$随着 EM 迭代不断变化直到收敛。当 EM 算法收敛后，得到所有的参数 $p(y,z)$，$p(w\mid y,c)$，$p(d\mid z,c)$ 和 $p(c)\,(1\le c\le s+t)$。最后通过计算得到 $p(d\mid z,c)(s+1\le c\le s+t)$对目标领域数据进行赋值，

$$\begin{aligned}p(z\mid d,c)&=\frac{p(z,d,c)}{p(d,c)}\propto p(z,d,c)=p(d\mid z,c)p(z,c)\\&=p(d\mid z,c)p(z)p(c)=p(d\mid z,c)p(c)\sum_y p(y,z)\\&\propto p(d\mid z,c)\sum_y p(y,z)\end{aligned} \tag{6.21}$$

最后各个目标源领域中样本的标签预测如下，

$$\arg\max_z p(z\mid d,c) \tag{6.22}$$

详细的跨领域学习算法 CD-PLSA 如算法 6.1 中描述。值得注意的是我们的算法可以同时处理多个源领域与目标目标领域的分类问题。

算法 6.1: 基于生成模型的跨领域学习分类方法(CD-PLSA)

输入： $s+t$ 个领域数据，$D_1,\cdots,D_s,D_{s+1},\cdots,D_{s+t}$，前面 s 个领域为带标签的源领域数据，而后 t 个领域为无标签的目标领域数据；迭代次数，T；词特征的聚类个数，Y。

输出： 目标领域中文档的预测信息。

步骤 1：初始化。$p^{(0)}(w|y,c)$ 初始化为 PLSA 的输出(在所有的数据上做 PLSA)，$p^{(0)}(d|z,c)$ 的初始化在实验部分详细描述，随机初始化 $p^{(0)}(y,z)$。

步骤 2：$k:=1$。

步骤 3：对于 $c:=1\to s+t$，

根据式子(6.11)更新 E 步中的 $p^{(k)}(y,z|w,d,c;\theta^{\text{old}})$。

步骤 4：对于 $c:=1\to s+t$，

根据式子(6.16)更新 M 步中的 $p^{(k)}(w|y,c)$。

步骤 5：对于 $c:=s+1\to s+t$，

根据式子(6.17)更新 M 步中的 $p^{(k)}(d|z,c)$。

步骤 6：根据式子(6.18)更新 M 步中的 $p^{(k)}(y,z)$。

步骤 7：根据式子(6.19)更新 M 步中的 $p^{(k)}(c)$。

步骤 8：$k:=k+1$，若 $k<T$ 则转步骤 3。

步骤 11：各个目标领域中的文档类别通过式子(6.21)和(6.22)预测得到。

6.3.3 精化 CD-PLSA

CD-PLSA 算法输出各个领域的词特征概念的外延 $p(w|y,c)$，文档概念的外延 $p(d|z,c)$，以及所有领域共享的词特征概念和文档概念的内涵 $p(y,z)$。在 CD-PLSA 模型中，假设所有的领域共享相同的词特征概念和文档概念的内涵 $p(y,z)$，但由于各个领域的分布不同，CD-PLSA 所求得概念内涵 $p(y,z)$ 有可能并不完全是各个领域自身的概念内涵。因此，在这个步骤中我们提出进一步进化 CD-PLSA 模型的结果，即对每一个目标领域 c ($s+1\le c\le s+t$)，只利用自身的局部数据进一步迭代更新参数 $p(y,z|w,d,c)$，$p(w|y,c)$，$p(d|z,c)$ 以及 $p(y,z)$，根据式子(6.11)，(6.16)，(6.17)和(6.18)。这些参数的

输入为第一步骤 CD-PLSA 的输出。实验表明，经过局部精化，可以进一步提高 CD-PLSA 的分类性能。称这两阶段的方法为 RCD-PLSA。

6.4. CD-PLSA 算法的分布式实现

我们还扩展 CD-PLSA 模型的 EM 求解算法到分布式版本，该版本能够处理源领域数据 $D_1,\cdots,D_s$ 和目标领域数据 $D_{s+1},\cdots,D_{s+t}$ 地理上分布的情况，即不能把所有的领域数据集中到中间节点，由于数据安全或者隐私考虑。

在这个分布式环境下，需要一个中间节点 mn 和 $s+t$ 个从节点，每个从节点储存一个领域数据，所有从节点表示为 $sn^{(1)},\cdots,sn^{(s+t)}$。从算法 6.1 可以发现，1) 对于参数 $p(y,z|w,d,c)$，$p(w|y,c)$，$p(d|z,c)$ 的迭代计算可以在各个从节点 $sn^{(c)}$ 局部计算；2) $p(y,z)$ 可以在主节点上计算。假设，

$$\Delta_{y,z}^{(c)}=\sum_{w,d}\mathbf{O}_{w,d,c}\,p(y,z|w,d,c;\theta^{\text{old}}) \tag{6.23}$$

$$V^{(c)}=\sum_{y,z,w,d}\mathbf{O}_{w,d,c}\,p(y,z|w,d,c;\theta^{\text{old}}) \tag{6.24}$$

则

$$p(y,z)=\frac{\sum_{c}\Delta_{y,z}^{(c)}}{\sum_{y,z,c}\Delta_{y,z}^{(c)}},\qquad p(c)=\frac{V^{(c)}}{\sum_{c}V^{(c)}} \tag{6.25}$$

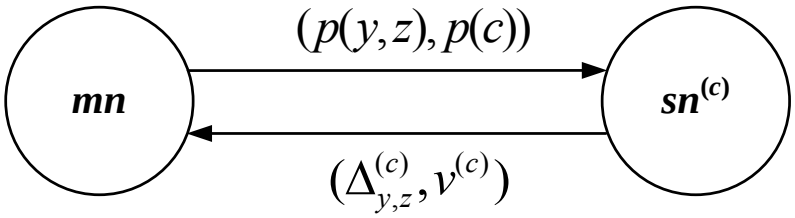


图 6.4.1 主从节点之间中间结果的传递情况

在每一次迭代中，主节点 mn 首先传送 $p(y,z)$，$p(c)$ 给从节点，然后每个从节点 $sn^{(c)}(1\le c\le s+t)$ 局部计算 $p(y,z|w,d,c)$，$p(w|y,c)$，$p(d|z,c)$，$\Delta_{y,z}^{(c)}$ 以及 $V^{(c)}$；完成这些计算后，从节点 $sn^{(c)}$ 传递局部中间结果 $\Delta_{y,z}^{(c)}$ 和 $V^{(c)}$ 给主节点；最后主节点根据式子(6.25)更新 $p(y,z)$，$p(c)$，并把它们传送给从节点，开始下一次迭代。可以看到整个迭代过程中，只传递了一些中间的结果，如 $p(y,z)$，$p(c)$，$\Delta_{y,z}^{(c)}$ 和 $V^{(c)}$，而不是传输领域数据，这从一定程度上保护了数据的安全隐私。主节点与从节点之间的通信情况如图

6.4.1 所示。假设迭代T次 EM 算法收敛，文档类别的个数为C，词特征概念的个数为Y，则整个通信负载为$2T\cdot(s+t)\cdot(Y\cdot C+1)$(参数$p(y,z)$和$\Delta_{y,z}^{(c)}$的大小均为$Y\cdot C$)，可见通信量非常小，因此该分布式算法非常高效。

6.5. 实验过程和结果

这一节设计系统的实验来证明 CD-PLSA 模型的有效性。对于两类分类问题，每个问题有 4 个领域，分成两种情况，1) 1 个源领域加上 3 个目标领域；2) 3 个源领域加上 1 个目标领域。对于三类分类问题，每个问题也有 4 个领域，不过这里只列出了 3 个源领域 1 个目标领域的情况。

6.5.1 实验数据

本章仍然采用 *20Newsgroup* 数据集作为评价 CD-PLSA 的数据集，其详细描述见表 4.5.1，构造分类问题类似于第五章。数据集 20*Newsgroups* 包含四个大类 *comp*，*rec*，*sci* 以及 *talk*，若任意取两个大类构造两类问题，则可以构造 6(C_4^2)个数据集，它们是 *rec vs sci*，*comp vs. sci*，*sci vs. talk*，*comp vs. rec*，*comp vs. talk* 以及 rec vs. talk。同样每个数据集可以构造 96 种 4 个领域的两类分类问题，我们考虑两种情况，包括 1 个源领域 3 个目标领域，以及 3 个源领域 1 个目标领域的两类分类问题。

对于三类分类问题，我们构造了 4(C_4^3)个数据集，即从 4 个大类中任意选择三个大类。对于每个数据集，可以构造 2304($4\cdot P_4^4\cdot P_4^4$)中 4 个领域的三类分类问题。这里只考虑了多个源领域的情况，即 3 个源领域 1 个目标领域。实验中对每个数据集只随机采用了 100 个分类问题，因此总共有100×4个三类分类问题。

6.5.2 比较算法和实现细节

比较算法。与本章提出的算法 CD-PLSA，RCD-PLSA 比较的算法包括，1) 监督学习算法逻辑回归(LG)[Davie, 2000]以及 LibSVM[Chang, 2001]; 2) 经典的跨领域分类算法 CoCC[Dai, 2007]，局部加权集成(Local Weighted Ensemble，LWE)[Gao, 2008]，以及桥接精化方法(BR)[Xing, 2007]；3) 6.2 节描述的 D-PLSA 算法。在该算法中，所有的数据合并起来组成一个的大的领域，即没有区分样本的来源。实验表明如果不考虑样本来自于哪个领域，会导致性能的降低。

这里需要强调下，对于不能处理多个源领域的方法，采用如下的方式，每个源领域和目标领域训练得到一个模型，然后最后做等权重加权平均集成。对算法 CoCC，LG，LibSVM 以及 BR 都采用相似的处理方法。

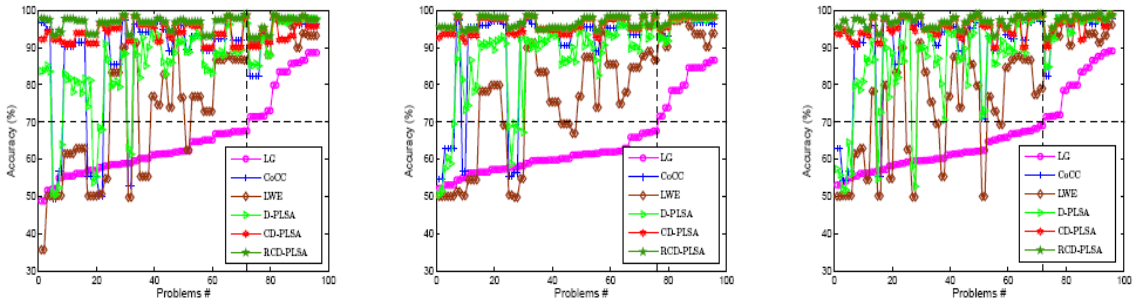
实现细节。由于 D-PLSA，CD-PLSA 算法都有一个随机初始化的过程，因此实验中

列出的结果是进行三次实验的平均值。初步的实验表明，当词特征聚类的个数Y在一定范围内时，如$[2^5, 2^8]$，CD-PLSA 对其不敏感，因此设置$Y = 64$。算法 D-PLSA，CD-PLSA 以及 RCD-PLSA 的迭代次数设为 50。比较算法 CoCC，LWE 和 BR 的参数设置为它们论文中推荐的参数。

6.5.3 实验结果

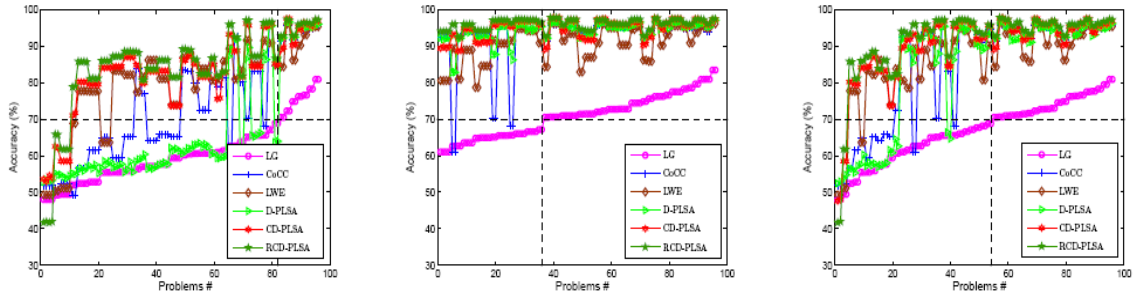
为了验证我们的算法 CD-PLSA，RCD-PLSA 的有效性，1) 在两类分类实验中，比较了算法 LG，CoCC，LWE 以及 D-PLSA；2) 在三类分类实验中，比较了 LG，LibSVM，BR 以及 D-PLSA。

两类分类—多目标领域。如前面的实验数据中所描述，我们构造了 6 个数据集对两类分类问题进行实验，每个数据集又可以构造 96 种 1 个源领域和 3 个目标领域的分类问题，总共有 576(96×6)个分类问题。所有的实验结果如图 6.5.1 到图 6.5.6 所示。每个图中有三个小图，分别表示三个目标领域，每个小图上 96 个问题按照 LG 算法的准确率按升序排列，这也一定程度上反应迁移学习的难度，LG 算法准确率越低，迁移学习难度越大，否则就会比较容易。



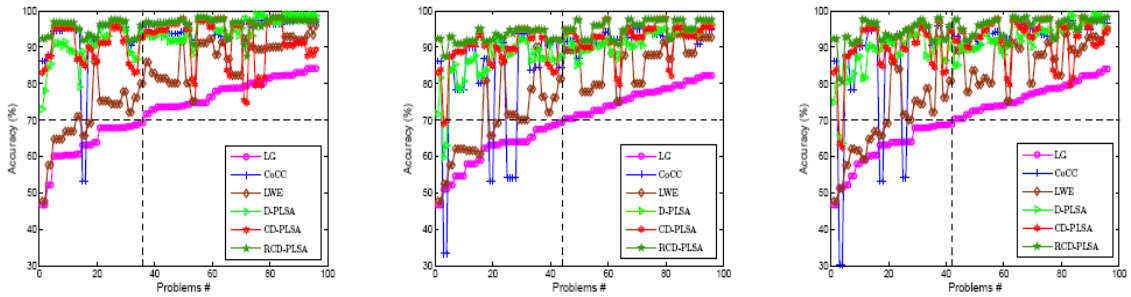
(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 1 (b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 2 (c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 3

图 6.5.1 CD-PLSA，RCD-PLSA 与其他算法在数据*rec vs. sci* 上的比较(多目标领域)



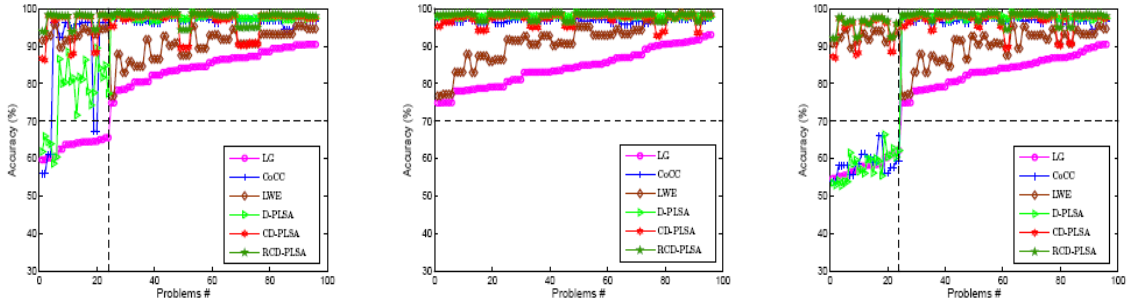
(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 1 (b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 2 (c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 3

图 6.5.2 CD-PLSA，RCD-PLSA 与其他算法在数据*comp vs. sci* 上的比较(多目标领域)



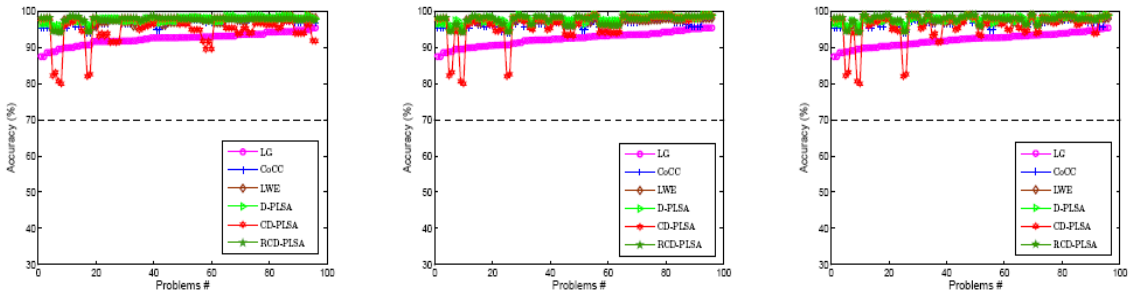
(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 1 (b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 2 (c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 3

图 6.5.3 CD-PLSA，RCD-PLSA 与其他算法在数据 *sci vs. talk* 上的比较(多目标领域)



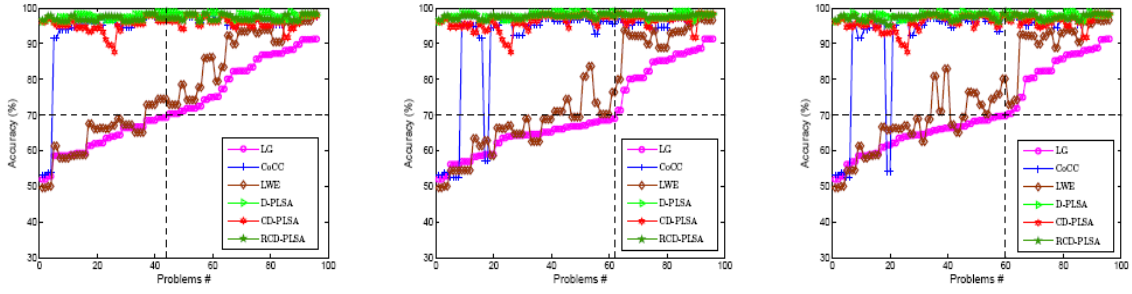
(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 1 (b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 2 (c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 3

图 6.5.4 CD-PLSA，RCD-PLSA 与其他算法在数据 *comp vs. rec* 上的比较(多目标领域)



(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 1 (b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 2 (c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 3

图 6.5.5 CD-PLSA，RCD-PLSA 与其他算法在数据 *comp vs. talk* 上的比较(多目标领域)



(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 1 (b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 2 (c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on Target Domain 3

图 6.5.6 CD-PLSA，RCD-PLSA 与其他算法在数据 *rec vs. talk* 上的比较(多目标领域)

表 6.5.1 CD-PLSA，RCD-PLSA 与其他算法的平均准确率(%)比较(多目标领域)

Data Sets			LG	CoCC	LWE	D-PLSA	CD-PLSA	RCD-PLSA
rec vs. sci	Target-1	*Left*	60.33	86.32	69.78	83.16	93.98	**96.59**
		Right	80.82	93.70	93.47	94.31	94.09	**96.47**
		Total	65.46	88.17	75.70	85.95	94.00	**96.56**
	Target-2	*Left*	59.72	89.34	73.24	86.23	95.82	**96.72**
		Right	80.88	96.47	94.29	96.72	97.64	**97.93**
		Total	64.13	90.82	77.62	88.42	96.20	**96.97**
	Target-3	*Left*	61.00	89.62	72.49	84.70	95.39	**96.97**
		Right	81.11	95.55	94.02	95.94	96.32	**97.50**
		Total	66.03	91.10	77.87	87.51	95.62	**97.10**
comp vs. sci	Target-1	*Left*	57.93	69.10	77.64	62.54	80.64	**82.61**
		Right	75.70	94.14	91.56	94.74	94.54	**95.64**
		Total	60.52	72.75	79.67	67.23	82.66	**84.51**
	Target-2	*Left*	64.66	89.52	88.44	92.71	94.08	**95.92**
		Right	74.70	94.95	92.36	95.29	95.65	**96.36**
		Total	70.93	92.91	90.89	94.33	95.06	**96.19**
	Target-3	*Left*	61.02	77.30	82.05	76.86	86.53	**88.40**
		Right	74.36	94.64	92.65	94.30	95.02	**95.99**
		Total	66.86	84.89	86.68	84.49	90.25	**91.72**
sci vs. talk	Target-1	*Left*	63.15	91.38	70.36	89.99	90.40	**95.04**
		Right	78.22	95.95	87.84	95.60	92.39	**96.58**
		Total	72.57	94.24	81.29	93.50	91.64	**96.00**
	Target-2	*Left*	61.35	80.52	69.36	85.49	88.85	**92.83**
		Right	76.42	92.97	87.14	92.29	93.18	**95.76**
		Total	69.52	87.26	78.99	89.17	91.20	**94.41**
	Target-3	*Left*	62.04	85.42	69.89	87.43	89.45	**93.76**
		Right	76.96	94.57	87.57	93.41	93.09	**96.27**
		Total	70.44	90.57	79.84	90.79	91.50	**95.17**
comp vs. rec	Target-1	*Left*	63.26	87.10	93.10	76.71	95.29	**97.26**
		Right	85.15	96.86	90.46	97.71	96.59	**97.85**
		Total	79.68	94.42	91.12	92.46	96.26	**97.70**
	Target-2	*Left*	–	–	–	–	–	–
		Right	84.38	96.81	91.60	98.05	97.21	**98.18**
		Total	84.38	96.81	91.60	98.05	97.21	**98.18**
	Target-3	*Left*	58.02	58.70	92.38	58.11	94.28	**95.75**
		Right	83.67	96.81	90.02	97.90	96.94	**98.05**
		Total	77.26	87.28	90.61	87.95	96.27	**97.47**
comp vs. talk	Target-1	*Left*	–	–	–	–	–	–
		Right	92.51	96.80	97.43	**97.82**	95.00	97.80
		Total	92.51	96.80	97.43	**97.82**	95.00	97.80
	Target-2	*Left*	–	–	–	–	–	–
		Right	92.44	96.40	97.24	**97.95**	96.24	97.81
		Total	92.44	96.40	97.24	**97.95**	96.24	97.81
	Target-3	*Left*	–	–	–	–	–	–
		Right	92.19	96.77	97.50	97.83	96.04	**97.96**
		Total	92.19	96.77	97.50	97.83	96.04	**97.96**
rec vs. talk	Target-1	*Left*	62.63	91.78	64.32	**97.59**	95.07	97.09
		Right	81.34	96.89	88.08	**98.01**	96.11	97.33
		Total	72.77	94.54	77.19	**97.82**	95.64	97.22
	Target-2	*Left*	62.91	88.77	65.66	**97.94**	95.85	97.62
		Right	84.39	96.26	92.53	**97.97**	96.83	97.85
		Total	70.52	91.42	75.17	**97.95**	96.20	97.70
	Target-3	*Left*	63.48	90.02	66.86	**97.79**	95.64	97.37
		Right	83.62	96.61	91.16	**97.90**	96.39	97.62
		Total	71.03	92.49	75.98	**97.83**	95.92	97.46

从这些结果中，我们观测到，

— *t* 统计测试以 95%的置信度表明，CD-PLSA 在所有 596 个分类问题上优于所有比较的算法。进一步我们发现 CD-PLSA 的优越性更加明显，特别是当 LG 算法的准确率低于 70%的时候。表 6.5.1 列出了所有算法在 6 个数据集上的平均性能。表中的 *Left*，*Right* 分别表示 LG 准确率低于和高于 70%的分类问题的平均准确率，而 *Total* 表示所有 96 个问题的平均准确率。可以清楚地看到 CD-PLSA 在 *left* 行比其他比较算法的优越性要比 *right* 行好很多。这说明我们的算法无论什么情况都可以表现得很好，特别是当分类问题较难的情况(LG 准确率低于 70%)，这说明 CD-PLSA 算法有比较好的迁移学习能力。

— CD-PLSA 算法要比 D-PLSA 算法好很多。原因是在 D-PLSA 算法，所有的领域数据被合并成一个大的领域，这就无法挖掘领域之间的共性和特征。只有超过两个领域才能有效地挖掘领域之间的共性和特征，也就说合并所有的数据会牺牲算法的性能。从另一方面也说明了 CD-PLSA 模型中引入的领域变量 c 是非常有用的，它把数据按照领域分成几个聚类，这有利于提高 CD-PLSA 模型的能。

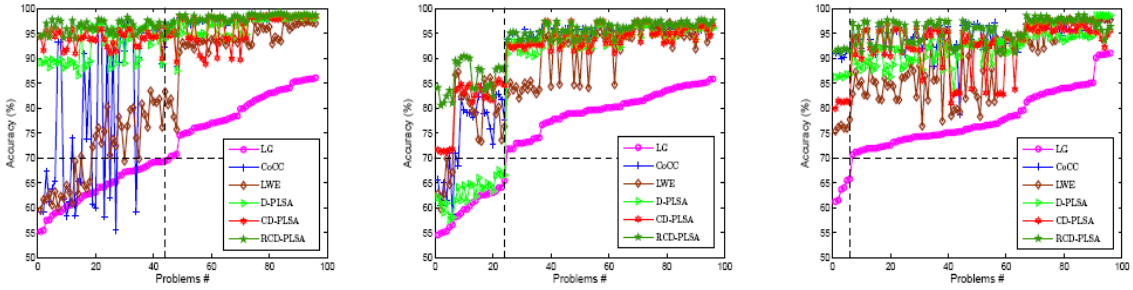
— RCD-PLSA 通常可以进一步提高 CD-PLSA 算法的准确率。也就是说，CD-PLSA 输出的所有领域共享的内涵通常不是目标领域自身的内涵，因此进一步的局部精化是有效的。

所有的这些结果都证明了 CD-PLSA 可以很好的处理多目标领域的跨领域分类学习问题，且非常有效。

两类分类—多源领域。对于多源领域的情况，仍在前面的 6 组数据上进行实验，每个分类问题有 3 个源领域和 1 个目标领域。所有的实验结果如图 6.5.7 所示。类似于多目标领域的情况，表 6.5.2 列出了所有算法的平均准确率比较。

表 6.5.2 CD-PLSA，RCD-PLSA 与其他算法的平均准确率(%)比较(多源领域)

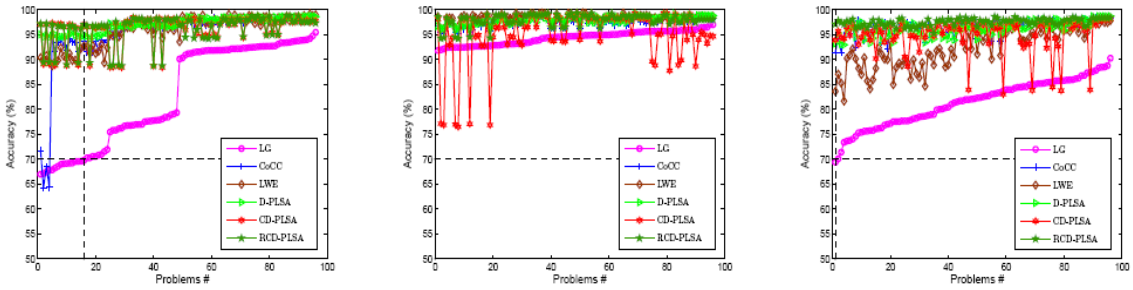
Data Sets		LG	CoCC	LWE	D-PLSA	CD-PLSA	RCD-PLSA
rec vs. sci	*Left*	64.01	80.07	71.41	92.03	94.06	**96.27**
	Right	79.84	97.70	93.62	96.77	96.46	**98.18**
	Total	72.42	89.44	83.21	94.55	95.33	**97.28**
comp vs. sci	*Left*	60.15	74.88	76.77	63.21	80.54	**85.93**
	Right	79.91	95.58	92.14	94.38	94.51	**96.02**
	Total	74.97	90.41	88.30	86.59	91.02	**93.50**
sci vs. talk	*Left*	63.62	91.05	76.30	86.45	81.05	**91.62**
	Right	78.20	94.76	89.78	92.40	92.84	**95.68**
	Total	77.29	94.53	88.94	92.03	92.42	**95.43**
comp vs. rec	*Left*	68.61	86.59	90.46	94.75	94.39	**95.82**
	Right	85.76	97.11	96.90	97.50	96.53	**97.85**
	Total	82.90	95.36	**98.84**	97.04	96.18	98.79
comp vs. talk	*Left*	–	–	–	–	–	–
	Right	94.37	97.41	**98.39**	97.89	95.06	97.96
	Total	94.37	97.41	**98.39**	97.89	95.06	97.96
rec vs. talk	*Left*	69.39	91.39	83.58	93.18	93.93	**97.09**
	Right	81.49	96.60	92.63	96.20	94.98	**97.50**
	Total	81.36	96.56	92.54	96.16	94.96	**97.50**



(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on data set *rec vs. sci*

(b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on data set *comp vs. sci*

(c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on data set *sci vs. talk*



(d) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on data set *comp vs. rec*

(e) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on data set *comp vs. talk*

(f) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, LWE, CoCC, LG on data set *rec vs. talk*

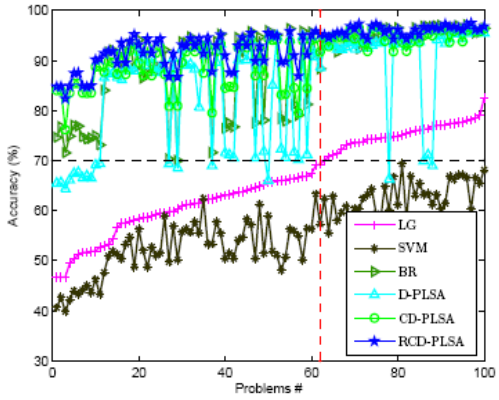
图 6.5.7 CD-PLSA，RCD-PLSA 与其他算法在 6 个数据集上的比较(多源领域)

类似地，可以得到与多目标领域相似的结果。所有的实验结果再次表明 CD-PLSA 模型优越于所有比较的算法，而且从表 6.5.2 看到，RCD-PLSA 总是比 CD-PLSA 好，这充分说明了局部精化有效性。不过也必须看到在图 6.5.7(e)和图 6.5.7(f)中，CD-PLSA 在某些分类问题上出现了过拟合，当 LG 准确率高于 80%的情况。我们猜测此时目标领域与源领域的内涵可能差别很大，所以导致过拟合，从而牺牲了准确率。不过从实验结果可以看到，CD-PLSA 的微小不足可以由局部精化 RCD-PLSA 来弥补。

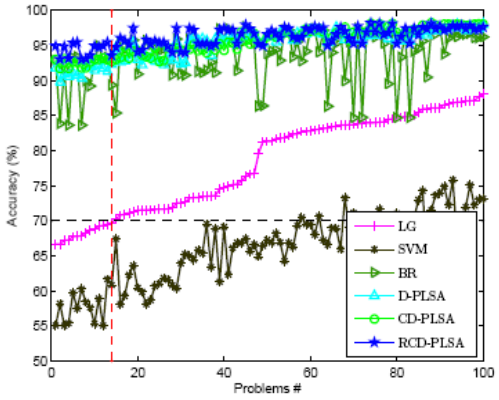
三类分类—多源领域。为了说明 CD-PLSA 算法可以直接处理多类分类问题，我们对算法 CD-PLSA，RCD-PLSA，LG，LibSVM，BR 以及 D-PLSA 在 4 个数据上(详细是数据描述见实验数据部分)进行了比较。算法 LibSVM 和 BR 可以直接处理多类数据，而 LG 采用 1 对多的方式。所有的实验结果如图 6.5.8 和表 6.5.3 所示。我们发现这些三类分类问题的结果与两类分类问题基本一致。CD-PLSA 准确率比所有的比较算法都高，除了在数据集 *comp vs. rec vs. talk* 上与 BR 算法相当。同样，CD-PLSA 可以处理迁移学习问题较难的情况，而且 RCD-PLSA 也比 CD-PLSA 有了显著提高。

表 6.5.3 CD-PLSA，RCD-PLSA 与其他算法的平均准确率(%)比较

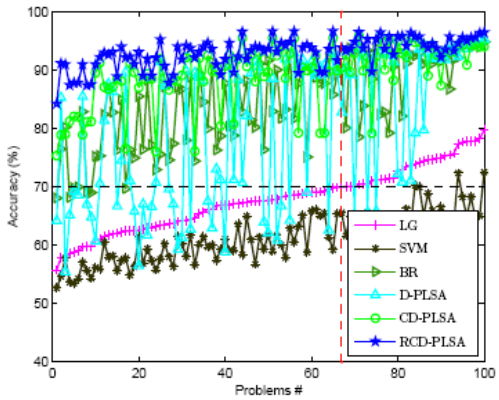
	comp vs. rec vs. sci			*comp vs. rec vs. talk*			*comp vs. sci vs. talk*			*rec vs. sci vs. talk*		
	Left	*Right*	*Total*	*Left*	*Right*	*Total*	*Left*	*Right*	*Total*	*Left*	*Right*	*Total*
LG	60.00	75.39	65.85	68.19	80.00	78.31	64.87	73.84	67.83	61.33	72.04	63.26
LibSVM	52.06	62.88	56.17	57.78	67.18	65.86	59.31	64.85	61.14	50.90	57.63	52.12
BR	85.83	95.21	89.40	92.59	96.10	95.61	81.68	90.02	84.43	89.84	95.39	90.84
D-PLSA	81.30	91.30	85.10	91.30	95.87	95.23	76.11	87.65	79.91	88.14	93.80	89.16
CD-PLSA	88.46	94.89	90.90	90.33	93.99	93.48	87.62	92.49	89.23	91.75	94.02	92.16
RCD-PLSA	**91.17**	**96.04**	**93.02**	**94.31**	**96.52**	**96.21**	**92.09**	**94.69**	**92.95**	**94.57**	**96.13**	**94.85**



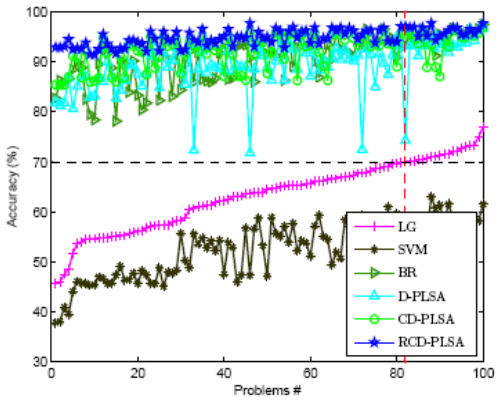
(a) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, BR, LibSVM and LG on data set *comp vs. rec vs. sci*



(b) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, BR, LibSVM and LG on data set *comp vs. rec vs. talk*



(c) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, BR, LibSVM and LG on data set *comp vs. sci vs. talk*



(d) Comparison among RCD-PLSA, CD-PLSA, D-PLSA, BR, LibSVM and LG on data set *rec vs. sci vs. talk*

图 6.5.8 CD-PLSA，RCD-PLSA 与其他算法在 4 个数据集上的比较

6.5.4 词特征概念的理解

这一节中，我们从实验结果分析各个领域词特征概念的外延之间的不同与相关性。对于每个领域 c 的概念 y，列出其代表性的 N (这里 $N=20$)个关键词根据概率 $p(w|y,c)$，表 6.5.4 中列出 3 个词特征概念的外延。可以看到它们是相关的，因为这些关键词表示相同的语义概念。但这些外延又是不同的，比如第三个词特征概念“Space Science”，第一个领域用关键词“rocket”，“ESA”(Europea Space Agency)以及“satellite”等描述，而第二个领域则用关键词“acceleration”，“NASA”和“earth”等描述。这些结果从直观上表明了算法 CD-PLSA 可以有效地挖掘领域之间的共性与特性。

6.5.5 算法执行时间

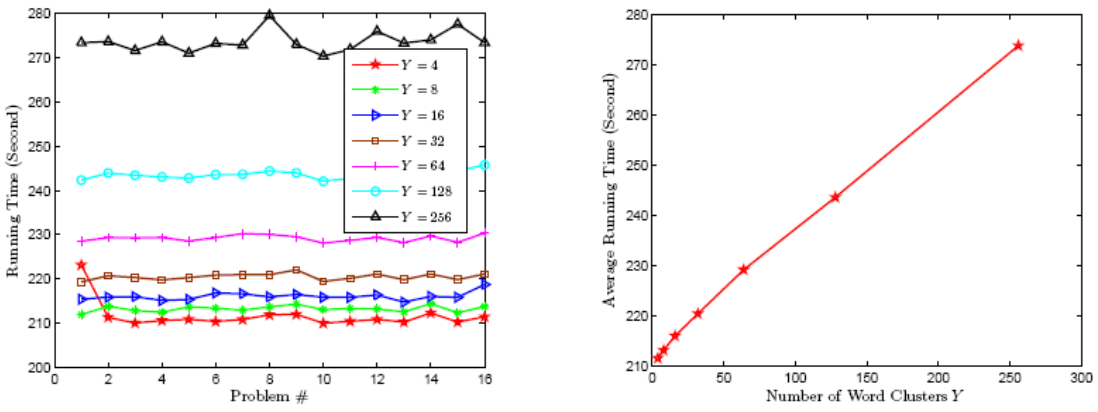
我们还考察了 CD-PLSA 算法的有效性。从数据集 *rec vs. sci* 中随机 16 个分类问题，然后记录这 16 个问题在不同数目的词特征聚类下，CD-PLSA 的运行时间，其结果如图 6.5.9 所示。从图 6.5.9(a)可以看到，CD-PLSA 算法运行非常快，非常高效，基本上可以

在 240 秒内执行完，当词特征聚类个数为 64 且词特征数为 7500 的条件下。图 6.5.9(b)考察了 CD-PLSA 在 16 个分类问题上的平均执行时间与词特征聚类个数 Y 的关系，可以看到执行时间基本上与 Y 成线性比的关系，因此具有很好的扩展性。

实验小结。本节中做了大量的实验，主要贡献，1) 验证了 CD-PLSA 算法的有效性，能够处理迁移学习较难的分类问题，迁移学习能力强；2) CD-PLSA 优于 D-PLSA，这说明我们的算法可以有效挖掘领域之间的共性和特性以提高预测性能；3) RCD-PLSA 优于 CD-PLSA，说明局部精化是非常有效的；4) 从直观上分析了每个领域词特征概念的关键词表示，从而加深对词特征概念的理解；5) 考察了算法 CD-PLSA 的执行时间，表明其高效性。

表 6.5.4 每个领域词特征概念的关键词

Associated with Concept: *Space Science*	Domain 1	rocket, esa, assist, frank, af, thu, helsinki, ron, atlantic, jet, observer, satellite, venus, sei, min, ir, russia, stars, star, ray
	Domain 2	relay, km, rat, pixel, command, elements, arc, acceleration, nasa, earth, fuse, ground, bulletin, pub, anonymous, faq, unix, cit, ir, amplifier
	Domain 3	from, earth, science, word, pictures, years, center, data, national, dale, nasa, gif, reports, mil, planet, field, jpl, ron, smith, unix
	Domain 4	service, archive, unit, magnetic, thousands, technology, information, arc, keys, faq, probes, ir, available, gov, embedded, tens, data, system, unix, mil
Associated with Concept: *Computer Science*	Domain 1	support, astronomer, near, thousands, million, you, vnet, copy, ad, bright, lab, idea, data, hardware, engines, ibm, project, soviet, software, program
	Domain 2	legally, schemes, protected, bytes, mq, disks, patch, registers, machine, pirates, install, card, rom, screen, protection, disk, ram, tape, mb, copy
	Domain 3	discomfort, friend, normal, self, tests, programmer, steve, state, program, lab, you, your, jon, my, headache, trial, she, pain, page, trials
	Domain 4	wcs, cipher, scheme, brute, user, file, encryption, message, serial, decryption, crypto, keys, cryptosystems, skipjack, plaintext, secure, key, encrypted, nsa, des
Associated with Concept: *Car*	Domain 1	saves, power, was, at, disappointment, al, europeans, will, ny, north, their, they, deal, best, year, sports, cs, new, series, gm
	Domain 2	crash, price, vehicle, insurance, handling, gas, xs, dealer, cruiser, leather, buy, latech, fj, paint, ride, buying, bmw, engine, car, honda
	Domain 3	or, value, they, wade, good, car, better, best, three, performance, more, runner, than, average, dl, extra, base, cs, al, year
	Domain 4	dealer, camry, saab, engine, eliot, requests, mazda, liter, mustang, diesel, wagon, nissan, mileage, byte, saturn, toyota, si, cars, car, db



(a) The Running Time of CD-PLSA on Data Set *rec vs. sci* (b) The Relationship between Running Time and Word Clusters Y

图 6.5.9 CD-PLSA 的执行时间

6.6. 小结

本章研究和开发了词特征概念以及文档概念的外延和内涵。各个领域中词特征概念的外延以及文档概念的外延不同，是依赖于领域本身的；而各个领域中词特征概念和文档概念的内涵可能相同，独立于领域。我们提出两阶段的跨领域文本分类算法，首先提出 CD-PSLA 模型以协同训练的方式有效挖掘不同领域之间的共性和特性；然后进一步开发目标领域本身的内部结构，精化第一阶段的结果，分类性能得到进一步加强。本章还扩展 CD-PLSA 算法到分布式实现形式，处理各领域数据地理上分布的情况，一定程度上保护领域数据的安全、隐私性。实验结果表明本章提出的算法优于所有的比较算法，且具有较强的迁移学习能力，可以处理迁移学习比较难的分类问题。

下一章将系统地比较从多个源领域学习的跨领域学习算法。

第七章　多源领域跨领域迁移学习算法比较

7.1. 引言

在以往的工作中，迁移学习算法主要集中在单个源领域到单个目标领域的学习[Dai, 2007b; Dai, 2007; Xing, 2007; Zhuang, 2009]。在实际应用中，有标签的训练样本可能来自多个源领域，而且这些源领域数据虽然分布不同，但是却语义相关。另一方面，从一个源领域到目标领域的迁移学习往往不够，容易产生学习偏见，或者知识不够，导致性能不好。例如：要想学习好泛函分析，仅有代数学的基础是不够的，可能还得有几何学、微积分学以及函数论等相关学科的基础。

本文对从多源领域的跨领域学习进行了研究，比如第五章和第六章，本章将对一些从多源领域的迁移学习算法进行系统的比较。首先，我们扩展第四章基于非负矩阵的跨领域方法 MTrick，使之能同时处理多源领域。然后，对第六章提出的方法 CD-PLSA 进行改进。最后，对这些多源领域算法进行系统的比较。

7.2. 扩展 MTrick 到多源领域学习

7.2.1　处理单源领域的 MTrick

对于源领域中的联合概率分布矩阵 $\mathbf{X}_s \in \mathbb{R}_+^{m\times n_s}$ ($\mathbf{X}_s = \dfrac{\mathbf{Y}_s}{\sum_{i,j}\mathbf{Y}_{s_{(ij)}}}$， $\mathbf{Y}_s$ 是源领域中的词-文档频率共现矩阵，以下 $\mathbf{X}_t$ 类似定义)，目标领域中的联合概率分布矩阵 $\mathbf{X}_t \in \mathbb{R}_+^{m\times n_t}$，其中 m 是词特征个数，n_s 是源领域中样本的个数，n_t 是目标领域中样本的个数，可以形式化以下联合优化问题：

$$\begin{aligned}&\min_{\mathbf{F}_s,\mathbf{G}_s,\mathbf{S},\mathbf{F}_t,\mathbf{G}_t} \|\mathbf{X}_s-\mathbf{F}_s\mathbf{S}\mathbf{G}_s^{\mathrm{T}}\|^2+\frac{\alpha}{n_s}\cdot\|\mathbf{G}_s-\mathbf{G}_0\|^2+\beta\cdot\|\mathbf{X}_t-\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\|^2,\\ &s.t.\quad \sum_{j=1}^{k_1}\mathbf{F}_{s_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{s_{(ij)}}=1,\sum_{j=1}^{k_1}\mathbf{F}_{t_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{t_{(ij)}}=1,\end{aligned} \tag{7.1}$$

其中参数 $\alpha\ge 0$，$\beta\ge 0$ 是平衡因子，k_1 和 k_2 分别为词特征聚类个数和文档类别个数，关系矩阵 $\mathbf{S}$ 是两个联合概率分解的共享因子，这样 $\mathbf{S}$ 其实是把知识从源领域到目标领域迁移的桥梁。$\mathbf{F}$ 表示词特征的后验类别概率，$\mathbf{G}$ 表示文档的后验类别概率。$\mathbf{G}_0$ 包含源领域中的样本标签信息，当第 i 个样本属于第 j 时，$\mathbf{G}_{0_{(ij)}}=1$；否则 $\mathbf{G}_{0_{(ik)}}=0(k\ne j)$。在

这个优化问题中，$\mathbf{G}_0$作为监督信息来优化得到最终后验概率分布矩阵$\mathbf{F}_s,\mathbf{G}_s,\mathbf{F}_t,\mathbf{G}_t$以及联合关系矩阵$\mathbf{S}$。

7.2.2 处理多源领域的 MTrick

给出N $(N>1)$个已标注的源领域数据集$D_s^1,\cdots,D_s^N$作为训练集，可得到N个联合概率分布矩阵$\mathbf{X}_s^l \in \mathbb{R}_+^{m\times n_s^l}$ $(l=1,\cdots,N)$，目标领域中的联合概率分布矩阵$\mathbf{X}_t \in \mathbb{R}_+^{m\times n_t}$，其中$m$是词特征个数，$n_s$是源领域中样本的个数，$n_t$是目标领域中样本的个数，可以形式化以下处理多个源领域的联合优化问题：

$$\begin{aligned}&\min_{\mathbf{F}_s^1,\cdots,\mathbf{F}_s^N,\mathbf{G}_s^1,\cdots,\mathbf{G}_s^N,\mathbf{S},\mathbf{F}_t,\mathbf{G}_t}\sum_{l=1}^{N}(\|\mathbf{X}_s^l-\mathbf{F}_s^l\mathbf{S}\mathbf{G}_s^{l^{\mathrm{T}}}\|^2+\frac{\alpha}{n_s^l}\cdot\|\mathbf{G}_s^l-\mathbf{G}_0^l\|^2)+\beta\cdot\|\mathbf{X}_t-\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\|^2,\\ &s.t.\quad\sum_{j=1}^{k_1}\mathbf{F}_{s_{(ij)}}^l=1,\sum_{j=1}^{k_2}\mathbf{G}_{s_{(ij)}}^l=1\ (l=1,\cdots,N),\qquad\sum_{j=1}^{k_1}\mathbf{F}_{t_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{t_{(ij)}}=1,\end{aligned}\tag{7.2}$$

其中参数$\alpha\geq 0$，$\beta\geq 0$是平衡因子，关系矩阵$\mathbf{S}$是两个联合概率分解的共享因子，$\mathbf{F}_s^l$和$\mathbf{G}_s^l$分别为第l个源领域中词特征和文档的后验概率，$\mathbf{G}_0^l$为第l个源领域中的标签信息。初步实验表明式子(7.2)中的中间项对算法性能影响不大，特别是迭代算法初始化$\mathbf{G}_s^l$为$\mathbf{G}_0^l$时，算法最后求得的局部最优解即$\mathbf{G}_s^l=\mathbf{G}_0^l$，即对参数$\alpha$不敏感。因此简化优化问题(7.2)如下，

$$\begin{aligned}&\min_{\mathbf{F}_s^1,\cdots,\mathbf{F}_s^N,\mathbf{S},\mathbf{F}_t,\mathbf{G}_t}\sum_{l=1}^{N}\|\mathbf{X}_s^l-\mathbf{F}_s^l\mathbf{S}\mathbf{G}_0^{l^{\mathrm{T}}}\|^2+\beta\cdot\|\mathbf{X}_t-\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\|^2,\\ &s.t.\quad\sum_{j=1}^{k_1}\mathbf{F}_{s_{(ij)}}^l=1\ (l=1,\cdots,N),\sum_{j=1}^{k_1}\mathbf{F}_{t_{(ij)}}=1,\sum_{j=1}^{k_2}\mathbf{G}_{t_{(ij)}}=1\end{aligned}\tag{7.3}$$

类似于第四章中推导 MTrick 处理单个源领域的方法，可以得到以下每个变量的迭代公式，

$$\mathbf{F}_{s_{(ij)}}^l\leftarrow\mathbf{F}_{s_{(ij)}}^l\cdot\sqrt{\frac{(\mathbf{X}_s^l\mathbf{G}_0^l\mathbf{S}^{\mathrm{T}})_{(ij)}}{(\mathbf{F}_s^l\mathbf{S}\mathbf{G}_0^{l^{\mathrm{T}}}\mathbf{G}_0^l\mathbf{S}^{\mathrm{T}})_{(ij)}}},\tag{7.4}$$

$$\mathbf{F}_{t_{(ij)}}\leftarrow\mathbf{F}_{t_{(ij)}}\cdot\sqrt{\frac{(\mathbf{X}_t\mathbf{G}_t\mathbf{S}^{\mathrm{T}})_{(ij)}}{(\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\mathbf{G}_t\mathbf{S}^{\mathrm{T}})_{(ij)}}},\tag{7.5}$$

$$\mathbf{G}_{t_{(ij)}}\leftarrow\mathbf{G}_{t_{(ij)}}\cdot\sqrt{\frac{(\mathbf{X}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S})_{(ij)}}{(\mathbf{G}_t\mathbf{S}^{\mathrm{T}}\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S})_{(ij)}}},\tag{7.6}$$

然后通过以下式子归一化$\mathbf{F}_s^l(l=1,\cdots,N),\mathbf{F}_t,\mathbf{G}_t$，使得其满足约束条件

$$\mathbf{F}_{s_{(i\cdot)}}^l \leftarrow \frac{\mathbf{F}_{s_{(i\cdot)}}^l}{\sum_{j=1}^{k_1}\mathbf{F}_{s_{(i\cdot)}}^l}, \tag{7.7}$$

$$\mathbf{F}_{t_{(i\cdot)}} \leftarrow \frac{\mathbf{F}_{t_{(i\cdot)}}}{\sum_{j=1}^{k_1}\mathbf{F}_{t_{(ij)}}}, \tag{7.8}$$

$$\mathbf{G}_{t_{(i\cdot)}} \leftarrow \frac{\mathbf{G}_{t_{(i\cdot)}}}{\sum_{j=1}^{k_2}\mathbf{G}_{t_{(ij)}}}. \tag{7.9}$$

关系矩阵S的迭代更新公式如下：

$$\mathbf{S}_{(ij)} \leftarrow \mathbf{S}_{(ij)} \cdot \sqrt{\frac{\mathbf{A}_{(ij)}}{\mathbf{B}_{(ij)}}} \tag{7.10}$$

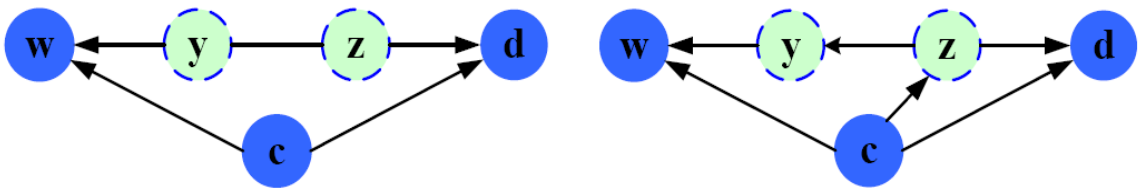
其中，

$$\mathbf{A} = \sum_{l=1}^{N}\mathbf{F}_s^{l^{\mathrm{T}}}\mathbf{X}_s^{l^{\mathrm{T}}}\mathbf{G}_0^l + \beta\cdot\mathbf{F}_t^{\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}\mathbf{G}_t \tag{7.11}$$

$$\mathbf{B} = \sum_{l=1}^{N}\mathbf{F}_s^{l^{\mathrm{T}}}\mathbf{F}_s^l\mathbf{S}\mathbf{G}_0^{l^{\mathrm{T}}}\mathbf{G}_0^l + \beta\cdot\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t\mathbf{S}\mathbf{G}_t^{\mathrm{T}}\mathbf{G}_t \tag{7.12}$$

算法的详细迭代过程与算法 4.1 相似，这里不再累述。

7.3. 改进 CD-PLSA



(a) CD-PLSA　　(b) Improved CD-PLSA

图 7.3.1 图模型 CD-PLSA 和改进 CD-PLSA

基于 D-PLSA，第六章提出一个统计生成模型处理多领域数据的跨领域文本分类算法。给出$s+t$个领域数据，表示为$D=(D_1,\cdots,D_s,D_{s+1},\cdots,D_{s+t})$，不失一般性，假设前面$s$个领域为带标签的源领域数据，而后$t$个领域为无标签的目标领域数据。根据图 7.3.1(a)中的图模型，所有变量的联合概率分布如下，

$$p(w,d,y,z,c) = p(w\,|\,y,c)p(d\,|\,z,c)p(y,z)p(c) \tag{7.13}$$

共现矩阵$\mathbf{O}_c$表示第c个领域的词-文档共现矩阵，其元素$\mathbf{O}_{w,d,c}$表示三元组(w,d,c)，

第 c 个领域中词 w 在文档 d 中出现的频率。

在第六章提出的 CD-PLSA 模型中，假设各个领域词特征概念和文档概念的内涵 $p(y,z)$ 一样。下面分析下这个联合概率 $p(y,z)=p(y|z)p(z)$，包括两部分，条件概率 $p(y|z)$ 和领域中各文档类别的概率 $p(z)$。很明显各个领域中文档类别的概率 $p(z)$ 不一样，因此为使得内涵 $p(y,z)$ 一样，条件概率 $p(y|z)$ 也必须不一样。在前面的例子中，如果一个新闻文档包含词特征概念“product”，那么该文档极有可能属于类别“product announcement”，而不是“financial scandal”。这是一个条件概率的形式，因此在改进版本中认为各个领域共享条件概率 $p(y|z)$ 更加贴切和合理，图 7.3.1(b)给出了改进 CD-PLSA 图模型。在该图模型中文档类别的概率 $p(z)$ 不再独立于领域，而是依赖于不同的领域，表示为 $p(z|c)$，而各领域共享的共性是条件概率 $p(y|z)$，根据图 7.3.1(b)可以写出所有变量的联合概率分布如下，

$$p(w,d,y,z,c)=p(w|y,c)p(d|z,c)p(y|z)p(z|c)p(c) \tag{7.14}$$

所有参数的求解仍然用 EM 算法，详细推导与第六章类似。这里给出所有参数的 EM 算法迭代公式。

E 步：

$$p(y,z|w,d,c;\theta^{\text{old}})=\frac{p(w|y,c)p(d|z,c)p(y|z)p(z)p(c)}{\sum_{y,z}p(w|y,c)p(d|z,c)p(y|z)p(z)p(c)} \tag{7.15}$$

M 步：

$$\hat{p}(w|y,c)=\frac{\sum_{z,d}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})}{\sum_{z,w,d}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})}. \tag{7.16}$$

$$\hat{p}(d|z,c)=\frac{\sum_{y,w}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})}{\sum_{y,w,d}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})} \tag{7.17}$$

$$\hat{p}(y,z)=\frac{\sum_{w,d,c}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})}{\sum_{y,w,d,c}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})} \tag{7.18}$$

$$\hat{p}(z|c)=\frac{\sum_{y,w,d}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})}{\sum_{y,z,w,d}\mathbf{O}_{w,d,c}p(y,z|w,d,c;\theta^{\text{old}})} \tag{7.19}$$

$$\hat{p}(c)=\frac{\sum_{y,z,w,d}\mathbf{O}_{w,d,c}\,p(y,z\mid w,d,c;\theta^{\text{old}})}{\sum_{y,z,w,d,c}\mathbf{O}_{w,d,c}\,p(y,z\mid w,d,c;\theta^{\text{old}})} \tag{7.20}$$

由于本章提出的是改进的 CD-PLSA 算法，因此表示为CD-PLSA*。同样也可以对CD-PLSA*输出的结果进行局部精化，表示为RCD-PLSA*。从后面的实验结果中可以看到，由于CD-PLSA*进一步考虑了文档类别的概率 $p(z)$ 依赖于各个不同的领域，因此相对CD-PLSA 有了性能上的提高。而值得注意的是RCD-PLSA*相对于CD-PLSA*的提高并不明显，这也充分说明了改进后的CD-PLSA 假设所有的领域数据共享条件概率 $p(y\mid z)$ 更加合理。

7.4. 实验过程和结果

7.4.1 实验过程

实验数据。为了评价所有的多源领域算法，我们从 6.5.1 节构造的 6 个数据集中选择两个数据集，包括 *comp vs. sci* 和 *rec vs. talk*。每个数据集有 96 种 3 个源领域 1 个目标领域的两类分类问题，因此总共有 192 个分类问题。可以看到这两个数据集比较有代表性，在数据 *comp vs. sci* 上构造的分类问题比较难，可以从监督学习算法逻辑回归(LG)[Davie, 2000]的准确率中看出来，因为其准确率比较低；而数据集 *rec vs. talk* 上的分类问题则比较容易，因为 LG 准确率在 96 个分类问题都超过 69%。详细的数据描述可见前面章节。

比较算法。本章比较的算法，除了本文提出的多源领域算法，1) 基于一致性正则化的多源领域学习方法，CCR_3(实验中记录CCR_3^{max}和CCR_3^{mean}的结果)；2) 基于非负矩阵分解的多源领域学习方法，MTrick；3) 有效挖掘领域共性与特性的方法，CD-PLSA 和 RCD-PLSA；4) 改进的 CD-PLSA 算法，CD-PLSA*和RCD-PLSA*；还有监督学习方法逻辑回归 LG，以及跨领域学习算法CoCC [Dai, 2007]和LWE [Gao, 2008]。

由于算法 LG 以及 CoCC 不能直接处理多源领域的情况，因此对于每个源领域训练一个模型，最后进行等权重加权平均。

实现细节。一致性正则化方法CCR_3采用第五章推荐的参数，即控制一致性正则化作用的参数 $\lambda=145$；基于非负矩阵分解的多源领域学习方法 MTrick，$\beta=1.5$；CD-PLSA*和RCD-PLSA*的参数设置与改进前一样。对于CoCC 和LWE 采用其原文章

推荐的参数。

7.4.2 实验结果

所有算法在两个数据集上的比较结果如表 7.4.1 所示。表中列出了每组数据集在 96 个问题上的平均准确率，其中 *Left*，*Right* 分别表示 LG 准确率低于和高于 70%的分类问题的平均准确率，而 *Total* 表示所有 96 个问题的平均准确率。所有比较的算法，除了 LG 是传统监督学习算法，其他算法都为跨领域学习方法。

表 7.4.1 多源领域跨领域分类算法的性能(%)比较

	comp vs. sci			*rec vs. talk*		
	Left	*Right*	*Total*	*Left*	*Right*	*Total*
LG	60.15	79.91	74.97	69.39	81.49	81.36
CoCC	74.88	95.58	90.41	91.39	96.60	96.56
LWE	76.77	92.14	88.30	83.58	92.63	92.54
CCR_3^{mean}	53.76	92.75	83.01	94.65	96.86	96.84
CCR_3^{max}	56.57	93.79	84.48	95.11	97.56	97.54
MTrick	83.40	95.71	92.63	97.27	97.45	97.45
CD-PLSA	80.54	94.51	91.02	93.93	94.98	94.96
RCD-PLSA	85.93	96.02	93.50	97.09	97.50	97.49
CD-PLSA*	82.92	95.49	92.32	96.71	97.34	97.33
RCD-PLSA*	84.81	96.01	93.21	96.78	97.65	97.64

从表中的实验结果可以看到，

- 在逻辑回归算法 LG 准确率不太低(大于 70%)的情况下，所有的迁移学习算法都是有效的。这也说明了迁移学习问题不太难时，可以比较容易用迁移学习算法提高其分类性能。但是，如果 LG 准确率比较低(低于 70%)，迁移学习较难的情况下，有些迁移学习算法就会出现负迁移，比如一致性正则化方法，负迁移就很严重，这也表明在实际中各个源领域训练得到的子分类器的条件独立假设很难满足。
- 本文提出的方法 MTrick，CD-PLSA，RCD-PLSA，CD-PLSA* 以及 RCD-PLSA* 比算法 CoCC，LWE 优越，特别是在迁移学习问题较难的情况下(LG 准确率低于 70%)，这说明了本文提出的这些方法具有较强的迁移学习能力。
- 算法 MTrick，CD-PLSA 性能差不太多，因为它们都考虑了不同领域之间的共性和特性。不过这两个算法有两点区别，1) 基于生成模型的 CD-PLSA 具有比较完美的概率解释；2) 两个模型的约束条件不同，如 MTrick 中的约束条件为 $\sum_y p(y \mid w) = 1$，而 CD-PLSA 中的约束条件是 $\sum_w p(w \mid y) = 1$。

- 改进后的算法CD-PLSA* 比CD-PLSA 准确率高，且与RCD-PLSA 准确率相当。这表明改进后假设各个领域之间共享条件概率 $p(y|z)$ 比共享联合概率 $p(y,z)$ 更加合理，因为CD-PLSA* 进一步考虑了不同领域中样本类别的分布 $p(z)$ 是不一样的。局部精化RCD-PLSA* 对CD-PLSA* 算法提高不明显，也充分说明共享条件概率 $p(y|z)$ 假设比较有效和合理。

7.5. 小结

本章首先扩展第四章的基于非负矩阵方法到处理多源领域，然后改进第六章提出的模型，最后对多源领域的学习算法进行了比较，包括对本文提出的所有多源领域学习算法，基于联合聚类的跨领域学习方法(CoCC)以及局部加权集成方法(LWE)。实验表明虽然基于一致性正则化方法有完整的理论分析，但条件独立假设太强，容易产生负迁移。MTrick， CD-PLSA 表现差不多，且比以往的迁移学习CoCC， LWE 优越。改进后的模型CD-PLSA* 比 CD-PLSA 更加有效和合理。

第八章　结束语

近十几年来，迁移学习算法已经引起了广泛的关注和研究，因为传统的机器学习需要的两个基本假设：(1) 用于学习的训练样本与新的测试样本满足独立同分布的条件；(2) 必须有足够可利用的训练样本才能学习得到一个好的分类模型；在实际应用中往往无法满足。本文主要针对迁移学习中的文本分类算法进行研究，研究的问题从简单到复杂，提出的算法也从浅到深。依次研究了从单个源领域到单个目标领域的学习问题，从多个源领域到单个目标领域的学习问题，从单个源领域到多个目标领域的学习问题以及从多个源领域到多个目标领域的学习问题。提出了四个跨领域文本分类算法，最后系统地多源领域跨领域学习算法进行了比较。取得的成果如下：

(1) 提出基于混合正则化的无标签领域归纳迁移学习方法。该方法解决目标领域无标签数据以及源领域数据是不同分布的分类问题，且建立一个归纳分类模型对新来的目标数据进行预测。研究学习了几种半监督学习技术，并把它们应用到迁移学习中，提出一种基于混合正则化框架的归纳迁移学习算法。其中包括目标领域分布结构的流形正则化，预测概率的熵正则化，以及类别比例的期望正则化。这个框架被用于从源领域到目标领域学习的归纳模型中。实验表明，加入类别先验可以避免类别比例漂移问题，其提高算法的准确率，且我们提出的算法比所比较的算法优越。

(2) 提出一种有效挖掘词特征聚类与文档类别关联关系的迁移学习算法。跨领域分类学习的目标是在源领域数据与目标领域数据具有不同数据分布的情况下，把从有标签源领域学习到的知识适应到无标签目标领域中。我们发现，虽然在原始词特征上，源领域与目标领域的数据分布不同，但是不同领域词特征聚类(词概念)与文档类别之间的关联关系可能是一样的。因此，开发可以这种与领域独立的关联关系，并且作为源领域与目标领域之间知识迁移的桥梁。即我们提出了同时分解源领域与目标领域数据矩阵的联合优化框架，其中共享词特征聚类(词概念)与文档类别之间的关联关系。为了求解该优化框架，我们提出一个迭代算法，并从理论上分析了其收敛性。实验结果表明本文提出的算法，a) 可以很好地解决迁移学习问题，并且优越于所有比较的算法。b) 更能处理学习问题较难的情况，具有更强的迁移学习能力。

(3) 提出基于一致性正则化的多源跨领域学习框架。在该框架下，局部的子分类器不仅考虑了在源领域上的可利用的局部数据，而且考虑了这些由源领域知识得到的子分类器在目标领域上的预测的一致性。更进一步，我们理论上分析了一致性正则化的有效性。最后，为了处理各个源领域数据在地理上分布的情况，提出了一致性正则化的分布式实现，可避免收集各个领域数据到中心节点，而只是传递一些统计变量，一定程度上减轻了数据信息的隐私性担忧。在实验中，我们 a) 验证了一致性正则化方法的有效性；b) 分析了一致性正则化方法可以提高分类性能的来源；c) 考察了算法的收敛性等。

(4) 给出基于生成模型的挖掘多领域共性与特性的跨领域分类方法。这一章从生成模型的角度研究多领域学习，有效挖掘多领域间的共性与特性。区别于概率隐性语义分析模型(PLSA)，只有一个隐性变量，我们提出的 CD-PLSA 模型有两个隐性变量 y 和 z，分别表示词特征概念和文档类别。不同领域间的共性把它们的特性联系起来，并且作为知识迁移的桥梁。提出一个 EM 算法来求解 CD-PLSA 模型，并实现了处理领域数据分布在不同节点的分布式算法。实验结果表明 CD-PLSA 算法优于所有比较的算法，且具有较强的迁移学习能力，可以处理迁移学习比较难的分类问题。还有可以同时处理多源领域与目标领域的分类问题。

(5) 系统地对本文提出的几种多源领域跨领域学习算法进行比较。首先扩展基于非负矩阵的跨领域方法 MTrick，使之能同时处理多源领域。然后，对 CD-PLSA 方法进行改进。实验表明多种多源跨领域学习算法各有优缺点，但都比传统监督学习算法性能优越。本文提出的多源领域学习算法也比以往的跨领域学习算法 CoCC，LWE 表现得好，且能处理迁移学习问题比较难的情况，具有较强的迁移学习能力。

本文对迁移学习中文本分类算法进行了深入研究。对所提出的算法都给出了详细的算法思想，并给出了详细的算法流程，最后都用系统、丰富的实验验证了所提出算法的有效性，有些算法还给出了一定假设条件下的形式化证明。但迁移学习作为一个新兴的研究领域，还有很多问题值得我们进一步的研究。

以下指出几点进一步的研究方向，首先，迁移学习的应用领域广泛，包括文本、图像分类，情感分类，强化学习，排序学习，度量学习，人工智能规划，文本、图像聚类，协同过滤以及基于传感器定位估计等，本文只针对文本分类问题进行研究。第二，关于迁移学习的理论研究还很缺乏，研究可迁移学习条件，目前这方面的研究还很少，主要集中在算法的研究。第三，研究迁移学习算法，如何避免负迁移？什么情况下的迁移学习是安全的。最后，目前的研究主要还是集中在研究领域，数据量小而且测试数据非常标准，应使所研究的算法在海量的数据中得到实际应用。

参考文献

[Abney, 2002] Abney A. Bootstrapping. // *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg PA: Association for Computational Linguistics, 2002: 360－367.

[Abney, 2004] Abney A. Understanding the Yarowsky Algorithm. // *Journal of Computational Linguistics*, Cambridge, MA, USA: MIT Press, 2004: 365—395.

[Ando, 2005] Ando R K, Zhang T. A high-performance semi-supervised learning method for text chunking [C]. // *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA: ACL, 2005: 1—9.

[Argyriou, 2007] Argyriou A, Evgeniou T, Pontil M. Multi-task Feature Learning [C]. Thrun S, Saul L K, Scholkopf B. // *Proceedings of Advances in Neural Information Processing Systems 19*, Cambridge: MIT Press, 2007: 243－272.

[Bai, 2009] Bai J, Zhou K, Xue G R, et al. Multi-task Learning for Learning to Rand in Web Search [C]. // *Proceedings of 18th ACM Conference on Information and Knowledge Management*, New York: ACM Press, 2009: 1549－1552.

[Bakker, 2003]Bakker B, Heskes T. Task clustering and gating for bayesian multitask learning [J]. *The Journal of Machine Learning Research*, 2003(4):83－99.

[Bel, 2003] Bel N, Koster C H A, Villegas M. Cross-Lingual Text Categorization [C]. // *Proceedings of European Conference on Digital Libraries,* Berlin: Springer-Verlag, 2003: 126－139.

[Ben-David, 2007] Ben-David S, Blitzer J, Crammer K, et al. Analysis of Representations for Domain Adaptation [C]. *// Proceedings. of Advances in Neural Information Processing Systems 19,* Cambridge: MIT Press , 2007: 137－144.

[Belkin, 2006] Belkin M, Niyogi P, Sindhwani V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples [J]. *Journal of Machine Learning Research*, 2006(7): 2399－2434.

[Blei, 2003] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. *Journal of Machine Learning Research*, 2003, 3: 993－1022.

[Blitzer, 2006] Blitzer J, McDonald R, Pereira F. Domain Adaptation with Structural Correspondence Learning [C]. // *Proceedings. of the International Conference on Empirical Methods in Natural Language Processing*, Stroudsburg PA: Association for Computational Linguistics, 2006: 120－128.

[Blitzer, 2007] Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, Boom-boxes and

Blenders: Domain Adaptation for Sentiment Classification [C]. // *Proc. of 45th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg PA: Association for Computational Linguistics, 2007: 440－447.

[Blitzer, 2008] Blitzer J, Crammer K, Kulesza A, et al. Learning Bounds for Domain Adaptation [C]. // *Proceedings. of Advances in Neural Information Processing Systems 20,* Cambridge: MIT Press , 2008: 129－136.

[Blum, 1998] Blum, Mitchell T. Combining Labeled and Unlabeled Data with Co-training [C]. // *Proceedings of the eleventh annual conference on Computational learning theory*, New York, NY, USA: ACM, 1998: 92—100.

[Boser, 1992] Boser B E, Guyou I, Vapnik V. A training Algorithm for Optimal Margin Classifiers [C]. *Proceedings of the 5th Annual Aorkshop on Computational learning theory*, New York, NY, USA: ACM, 1992: 144－152.

[Breiman, 1984] Breiman L, Friedman J, Stone C J, et al. Classification and Regression Trees [M]. *Wadsworth Int'l Group*, 1984.

[Broman, 2004] Borman S. The Expectation Maximization Algorithm: A Short tutorial [R]. *Available at http://www.seanborman.com/publicaitons*, 2004.

[Bruzzone, 2009] Bruzzone L, Marconcini M. Toward the Automatic Updating of Land-Cover Maps by a Domain-Adaptation SVM Classifier and a Circular Validation Strategy [J]. *IEEE Transactions on Geoscience and Remote Sensing In Geoscience and Remote Sensing*, 2009, 47(4): 1108－1122.

[Bruzzone, 2010] Bruzzone L, Marconcini M. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(5): 770－787.

[Caruana, 1997] Caruana R. Multitask Learning [J]. *Machine Learning of Special issue on Inductive Transfer*, Berlin: Springer-Verlag, 1997, 28(1): 41－75.

[Chang, 2001] Chang C C, Lin C, J. LibSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/cjli/libsvm, 2001.

[Cohn, 1994] Cohn D, Atlas L, Ladner R. Improving Generalization with Active Learning [J]. *Machine Learning,* Berlin: Springer-Verlag, 1994, 15(2): 201－221.

[Cover, 1967] Cover T, Hart P E. Nearest neighbor pattern classification [J]. *IEEE Transaction on Information Theory*, Los Vaqueros: IEEE Computer Society, 1967, 13(1): 21－27.

[Dai, 2007] Dai W Y, Xue G R, Yang Q, et al. Co-clustering based Classification for Out-of-domain Documents [C]. // *Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining,* New York: ACM Press,

2007: 210－219.

[Dai, 2007a] Dai W Y, Xue G R, Yang Q, et al. Transferring Naive Bayes Classifiers for Text Classification [C]. // *Proceedings of 22nd Conference on Artificial Intelligence,* California 94025: AAAI Press, 2007: 540－545.

[Dai, 2007b] Dai W Y, Yang Q, Xue G R, et al. Boosting for Transfer Learning [C]. // *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann, 2007: 193－200.

[Dai, 2008] Dai W Y, Yang Q, Xue G R, et al. Self-taught Clustering [C]. // *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann, 2008: 200－207.

[Dai, 2008a] Dai W Y, Chen Y Q, Xue G R, et al. Translated Learning: Transfer Learning across Different Feature Spaces [C]. *// Proceedings. of Advances in Neural Information Processing Systems 20,* Cambridge: MIT Press, 2008: 353－360.

[Dai, 2009] Dai W Y, Jin O, Xue G R, et al. Eigen Transfer: a Unified Framework for Transfer Learning [C]. // *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann, 2009: 193－200.

[Dasgupta, 2001] Dasgupta S, Littman M L, Mcalleste D. PAC Generalization Bounds for Co-Training [C]. *// Proceedings. of Advances in Neural Information Processing Systems 13,* Cambridge: MIT Press , 2001: 375－382.

[Davie, 2000] Davie H, Stanley L. Applied Logistic Regression [M], New York, 2000.

[Dempster, 1977] Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm [J]. Journal of the Royal Statistical Society, Series B, 1977, 39(1): 1-38.

[Dietterich, 2000] Dietterich T G. Ensemble Methods in Machine Learning [C]. // *Proceedings of the First International Workshop on Multiple Classifier Systems*, London, UK: Springer-Verlag, 2000: 1－15.

[Ding, 2006] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix tri-factorizations for clustering [C]. // *Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining,* New York: ACM Press, 2006: 126－135.

[Dredze, 2010] Dredze M, Kulesza A, Crammer K. Multi-domain learning by confidence-weighted parameter combination [J]. *Journal of Machine Learning*, 2010, 79(1-2): 123－149.

[Duan, 2009] Duan L X, Tsang Ivor W, Xu D, et al. Domain Adaptation from Multiple Sources via Auxiliary Classifiers [C]. // *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA: ACM,

2009: 289—296.

[Evgeniou, 2004] Evgeniou T, Pontil M. Regularized Multi--task Learning [C]. // *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, 2004: 109—117.

[Fan, 2005] Fan W, Davidson I, Zadrozny B, et al. An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias [C]. // *Proceedings of the 5th International Conference on Data Mining*, Los Vaqueros: IEEE Computer Society, 2005: 605—608.

[Freund, 1997]Freund Y, Schapire R E. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting [J]. *Journal of Computer and System Sciences*, 1997, 55(1): 119－139.

[Friedman, 1996] Friedman M, Goldszmidt T, Building Classifers Using Bayesian Networks [C]. // *Proceedings of 22nd Conference on Artificial Intelligence*, California 94025: AAAI Press, 1996: 1277－1284.

[Gao, 2008] Gao J, Fan W, Jiang J, et al. Knowledge Transfer via Multiple Model Local Structure Mapping [C]. // *Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press, 2008: 283－291.

[Gao, 2009] Gao J, Fan W, Sun Y Z, et al. Heterogeneous Source Consensus Learning via Decision Propagation and Negotiation [C]. // *Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press, 2009: 339－348.

[Gaussier, 2005] Gaussier E, Goutte C. Relation between PLSA and NMF and Implications [C]. // *Proceedings of the 28th SIGIR Conference*, New York: ACM Press, 2005: 601－602.

[Grandvalet, 2005] Grandvalet Y, Bengio Y. Semi-supervised Learning by Entropy Minimization [C]. *// Proceedings. of Advances in Neural Information Processing Systems 17,* Cambridge: MIT Press, 2005: 529－536.

[Gu, 2009] Gu Q Q, Zhou J. Learning the Shared Subspace for Multi-Task Clustering and Transductive Transfer Classification [C]. // *Proceedings of the 9th International Conference on Data Mining*, Los Vaqueros: IEEE Computer Society, 2009: 159－168.

[Guillamet, 2002] Guillamet D, Vitri`a J. Non-negative Matrix Factorization for Face Recognition [C]. // *Proceedings of the 5th Catalonian Conference on AI: Topics in Artificial Intelligence*, London UK: Springer-Verlag, 2002: 336－344.

[Guillamet, 2003] Guillamet D, Vitri`a J, Schiele B. Introducting a Weighted Non-negative

Matrix Factrorization for Image Classification [J]. *Pattern Recognition Letter*, 2003, 24(14): 2447－2454.

[Hastie, 2001] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [M]. Second Edition. Berlin: Springer-Verlag, 2001.

[He, 2002] He Q, Shi Z Z, Ren L A. The classification method based on hyper surface [C]. // *Proceedings of the International Joint Conference on Neural Networks*, Los Vaqueros: IEEE Computer Society, 2002: 1499－1503.

[He, 2003] He Q, Shi Z Z, Ren L A, et al. A Novel Classification Based on Hypersurface [J]. *International Journal of Mathematical and Computer Modeling*, 2003, 38(3-4): 395－407.

[He, 2008] He Q, Zhao X R, Shi Z Z. Minimal Consistent Subset for Hyper Surface Classification Method [J]. *International Journal of Pattern Recognition and Artificial Intelligence*, 2008, 22(1): 95－108.

[Hofmann, 1999] Hofmann T. Probabilistic Latent Semantic Analysis [C]. // *In Proc. of 15th Conference on Uncertainty in Artificial Intelligence*, San Francisco: Morgan Kaufmann, 1999: 289－296.

[Hofmann, 2001] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis [J]. *Journal of Machine Learning*, 2001, 42(1-2): 177－196.

[Huang, 2007] Huang J Y, Smola A J, Gretton A, et al. Correcting Sample Selection Bias by Unlabeled Data [C]. // *Proceedings. of Advances in Neural Information Processing Systems 19,* Cambridge: MIT Press, 2007: 601－608.

[Jebara, 2004] Jebara T. Multi-task Feature and Kernel Selection for SVMs [C]. *// Proceedings of 21th International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, 2004: 55－62.

[Jiang, 2007] Jiang J, Zhai C X. Instance Weighting for Domain Adaptation in NLP [C]. // *In Proc. of 45th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg PA: *Association for Computational Linguistics*, 2007: 264－271.

[Jiang, 2007a] Jiang J, Zhai C X. A two-stage approach to domain adaptation for statistical classfiers [C]. *In Proc. of 16th ACM Conference on Information and Knowledge Management*, New York: ACM Press, 2007: 401－410.

[Jiang, 2008] Jiang J. Domain Adaptation in Natural Language Processing [D]. Illinois, Computer Science in the Graduate College of the University of Illinois at Urbana-Champaign. 2008.

[Jiho, 2009] Jiho Y, Choi S J. Probabilistic Matirx Tri-factorization [C]. *// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*,

Los Vaqueros: IEEE Computer Society, 2009: 1553—1556.

[Joachims, 1999] Joachims T. Making Large-scale SVM Learning Practical [C]. // *Proceedings of Advances in Kernel Methods*, Cambridge: MIT Press, 1999: 169—184.

[Joachims, 1999a] Joachims T. Transductive Inference for Text Classification Using Support Vector Machines [C]. // *Proceedings of 16th International Conference on Multimedia,* Augsburg Germany, New York: ACM Press, 1999: 200－209.

[Joachims, 2003] Joachims T. Transductive Learning via Spectral Graph Partitioning [C]. // *Proceedings of 16th International Conference on Multimedia,* Augsburg Germany, New York: ACM Press, 2003: 290－297.

[Larence, 2004] Lawrence N D, Platt J C. Learning to Learn with the Informative Vector Machine [C]. *Proceedings of the 21st International Conference on Machine Learning*, New York, NY, USA: ACM, 2004: 65—72.

[Lee, 2001] Lee D D, Seung H S, Algorithms for non-negative matrix factorization [C]. // *Proceedings. of Advances in Neural Information Processing Systems 13,* Cambridge: MIT Press, 2001: 556－562.

[Lee, 2007] Lee S, Chatalbashev V, Vickrey D, et al. Learning A Meta-level Prior for Feature Relevance from Multiple Related Tasks. // *Proceedings of the 24th International Conference on Machine Learning*, New York, NY, USA: ACM, 2007: 489—496.

[Leskes, 2008] Leskes B, Torenvliet L. The Value of Agreement, A New Boosting Algorithm [J]. *Journal of Computer System Sciences*, 2008, 74(4): 557－586.

[Liao, 2005] Liao X J, Xue Y, Carin L, Logistic Regression with an Auxiliary Data Source [C]. *// Proceedings of 22nd International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, 2005: 505－512.

[Li, 2008] Li T, Ding C, Zhang Y, et al. Knowledge Transformation from Word Space to Document Space [C]. // *Proceedings of the 31st SIGIR Conference*, New York: ACM Press, 2008: 187－194.

[Li, 2009] Li T, Sindhwani V, Ding C, et al. Knowledge Transformation from for Cross-Domain Sentiment Classification [C]. // *Proceedings of the 31st SIGIR Conference*, New York: ACM Press, 2009: 716－717.

[Li, 2009a] Li B, Yang Q, Xue X Y. Can Movies and Books Collaborate? Cross-Domain Collaborative Filtering for Sparsity Reduction [C], // *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, California 94025: AAAI Press, 2009: 2052－2057.

[Li, 2010] Li T, Sindhwani V, Ding C, et al. Bridge Domains with Words: Opinion

Analysis with Matrix Tri-factorizations [C]. // *Proceedings of the 8th SIAM Conference on Data Ming*, Philadelphia: SIAM Press, 2010: 293－302.

[Ling, 2008] Ling X, Dai W Y, Xue G R, et al. Spectral Domain-Transfer Learning [C]. // *Proceedings of 14th ACM International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press, 2008: 488－496.

[Luo, 2008] Luo P, Zhuang F Z, Xiong H, et al. Transfer Learning from Multiple Source Domains via Consensus Regularization. // *Proceedings of 17th ACM Conference on Information and Knowledge Management*, New York: ACM Press, 2008: 103－112.

[Mahmud, 2007] Mahmud M M H, Ray S. Transfer Learning Using Kolmogorov Complexity: Basic Theory and Empirical Evaluations [R]. Department of Computer Science, University of Illinois at Urbana-Champaign, 2007.

[Mahmud, 2007a] Mahmud M M H. On Universal Transfer Learning [C]. // *Proceedings of 18th International Conference on Algorithmic Learning Theory*, Sendai, Japan, 2007: 135—149.

[Mann, 2007] Mann G S, McCallum A. Simple, Robust, Scalable Semi-supervised Learning via Expectation Regularization [C]. // *Proceedings of 24th International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, 2007: 593－600.

[Obozinski, 2006] Obozinski G, Taskar B, Jordan M I. Multi-task Feature Selection [R]. Department of Statistics, University of California, Berkeley, 2006.

[Pan, 2008] Pan S J, Kwok J T, Yang Q, Transfer Learning via Dimensionality Reduction [C]. // *Proceedings of the 23rd Conference on Artificial Intelligence*, California 94025: AAAI Press, 2008: 677－682.

[Pan, 2010] Pan S J, Yang Q. A Survey on Transfer Learning [J]. *IEEE Transaction on Data Engineering* , 2010, 22(10): 1345－1359.

[Quinlan, 1993] Quinlan J R. C4.5: Programs for Machine Learning [M]. San Francisco CA: Morgan Kaufmann Publishers Inc., 1993.

[Raina, 2007] Raina R, Battle A, Lee H, et al. Self-taught Learning: Transfer Learning from Unlabeled Data [C], // *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann, 2007: 759－766.

[Rosenstein, 2005] Rosenstein M T, Marx Z, Kaelbling L P. To Transfer or Not to Transfer [C]. // *Proceedings of Neural Information Processing Systems 2005 workshop on Inductive Transfer: 10 years Later*, Cambridge: MIT Press, 2005.

[Ruszczynski, 2006] Ruszczynski A. Nonlinear Optimization [M]. Princeton: Princeton University Press, 2006: 1-464.

[Samarth, 2006] Samarth S, Sylvian R. Cross Domain Knowledge Transfer Using Structured Representations [C]. // *Proceedings of 21nd Conference on Artificial Intelligence*, California 94025: AAAI Press, 2006: 506－511.

[Schwaighofer, 2005] Schwaighofer A, Tresp V, Yu K. Learning Gaussian process kernels via hierarchical Bayes [C]. // *Proceedings of Advances in Neural Information Processing Systems 17*, Cambridge: MIT Press, 2005: 1209—1216.

[Sha, 2003] Sha F, Saul L K, Lee D D. Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines [C]. // *Proceedings of Advances in Neural Information Processing Systems 15*, Cambridge: MIT Press, 2003: 1041－1048.

[Shi, 2008] Shi X X, Fan W, Ren J T. Actively Transfer Domain Knowledge [C]. // *Proceedings of The European Conference on Machine learning and Knowledge Discovery in DataBases,* Berlin: Springer-Verlag, 2008: 342－357.

[Shi, 2008a] Shi Z P, Ye F, He Q, Shi Z Z. Symmetric Invariant LBP Texture Description and Application for Image Retrieval [C]. // *Proceedings of the 2008 Congress on Image and Signal Processing*, Los Vaqueros: IEEE Computer Society, 2008: 825－829.

[Si, 2010] Si S, Tao D C, Chan K P. Evolutionary Cross-domain Discriminative Hessian Eigenmaps [J]. *IEEE Transactions on Image Processing*, 2010, 19(4): 1075－1086.

[Si, 2010a] Si S, Tao D C, Geng B. Bregman Divergence-based Regularization for Transfer Subspace Learning [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(7): 919－942.

[Sindhwani, 2005] Sindhwani V, Niyogi P. A Co-regularized Approach to Semi-supervised Learning with Multiple Views [C]. // *Proceedings of the ICML Workshop on Learning with Multiple Views*, San Francisco: Morgan Kaufmann, 2005: 74－79.

[Smeaton, 2003] Smeaton A, Over P. TRECVID: Benchmarking the Effectiveness of Information Retrieval Tasks on Digital Video [C]. // *Proceedings of the 2nd International Conference on Image and Video Retrieval*, Berlin, Heidelberg: Spring-Verlag, 2003: 19－27

[Thrun, 1996] S. Thrun. Is learning the *n*-th Thing Any Easier Than Learning The First? [C]. // *Proceedings of Advances in Neural Information Processing Systems 8*, Cambridge: MIT Press, 1996: 640－646.

[Tong, 2001] Tong S, Chang E. Support Vector Machine Active Learning for Image Retrieval [C]. // *Proceedingss of 9th ACM International Conference on Multimedia*, New

York: ACM Press, 2001: 107－118.

[Vapnik, 1998] Vapnik V N. Statistic Learning Theory [M]. New York: Wiely-Interscience, 1998.

[Wang, 2008] Wang F, Li T, Zhang C S. Semi-supervised Clustering via Matrix Factorization [C]. // *Proceedings of the 8th SIAM Conference on Data Ming*, Philadelphia: SIAM Press, 2008: 1－12.

[Wang, 2008a] Wang C, Mahadevan S. Manifold Alignment Using Procrustes Analysis [C]. // *Proceedings of 25th International Conference on Machine Learning*, New York, NY, USA: ACM, 2008: 1120—1127.

[Wang, 2008b] Wang Z, Song Y Q, Zhang C S. Transferred Dimensionality Reduction [C]. // *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg: Springer-Verlag, 2008: 550—565.

[Wang, 2008c] Wang F, Zhang C S. Label Propagation through Linear Neighborhoods [J]. *IEEE Transations on Knowledge and Data Engineering*, 2008, 20(1): 55－67.

[Wu, 2004] Wu P C, Dietterich T G. Improving SVM Accuracy by Training on Auxiliary Data Sources [C]. // *Proceedings of 21th International Conference on Machine Learning*, San Francisco: Morgan Kaufmann, 2004: 871－878.

[Xie, 2009] Xie S H, Fan W, Peng J, et al. Latent Space Domain Transfer between High Dimensional Overlapping Distributions [C]. // *Proceedings of ACM Conference on World Wide Web*, New York: ACM Press, 2009: 91–100.

[Xing, 2007] Xing D K, Dai W Y, Xue G R, et al. Bridged Refinement for Transfer Learning [C], // *Proceedings of 11th European Conference on Practice of Knowledge Discovery in Databases*, Berlin: Springer-Verlag, 2007: 324－335.

[Xue, 2008] Xue G R, Dai W Y, Yang, Q, et al. Topic-bridged PLSA for Cross-domain Text Classification [C]. // *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, 2008: 627－634.

[Yang, 2007] Yang J, Yan R, Hauptmann A G, Cross-domain Video Concept Detection Using Adaptive SVMs [C], // *Proc. of 24th International Conference on Multimedia*, New York: ACM Press, 2007: 188－197.

[Yarowsky, 1995] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods [C]. // *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 1995: 189—196.

[Zadrozny, 2004] Zadrozny B. Learning and Evaluating Classifiers under Sample Selection Bias [C]. // *Proceedings of the Twenty-first International Conference on*

Machine learning, New York, NY, USA: ACM, 2004: 114—121.

[Zhai, 2004] Zhai C X, Velivelli A, Yu B. A Cross-collection Mixture Model for Comparative Text Mining [C]. // *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, 2004: 743—748.

[Zhang, 2001] Zhang L. The Research on Human-Computer Cooperation in Content-based Image Retrieval [D]. Beijing: Computer Science, Tsinghua University. 2001.

[Zhou, 2006] Z.H. Zhou. Learning with Unlabeled Data and Its Application to Image Retrival [C]. // *In Proc. of 9th Pacific Rim International Conference On Artificial Intelligence*, Berlin: Springer-Verlag, 2006: 5－10.

[Zhu, 2005] Zhu X J. Semi-supervised Learning Literature Survey [R]. Department of Computer Sciences, University of Wisconsin, Madison, 2005.

[Zhuang, 2009] Zhuang F Z, Luo P, He Q, et al. Inductive Transfer Learning for Unlabeled Target-domain via Hybrid Regularization [J]. *Chinese Science Bulletin*, 2009, 54(14): 2470－2478.

[Zhuang, 2010] Zhuang F Z, Luo P, Xiong H, et al. Cross-domain Learning from Multiple Sources: A Consensus Regularization Perspective [J]. *IEEE Transactions On Knowledge And Data Engineering*, 2010, 22(12): 1664－1678.

[Zhuang, 2010a] Zhuang F Z, Luo P, Xiong H, et al. Exploiting Associations between Word Clusters and Document Classes for Cross-domain Text Categorization [C]. // *Proceedings of the 10th SIAM Conference on Data Ming*, Philadelphia: SIAM Press, 2010: 13－24.

[戴, 08] 戴文渊. 基于实例和特征的迁移学习算法研究[D]. 上海: 上海交通大学计算机科学与工程系. 2008.

[施, 09] 施潇潇. 主动迁移学习模型的研究与应用[D]. 中山: 中山大学计算机应用技术, 2009.

致　谢

岁月匆匆，五年的硕博生涯即将画上句号。值此博士论文完成之际，谨向所有关心和帮助我的老师、同学、朋友和家人表示真挚的感谢。

首先衷心地感谢的是我的导师何清研究员，是他在五年前仔细考察录取了我，使我有机会在计算所继续深造学习。在这五年的时间里，何老师在研究上大力倡导创新，带领我们进行科学攻关，在研究工作中给予了孜孜不倦的指导。在生活上给予了无微不至的关怀，特别是他总是从学生的角度考虑帮助我们。他严谨细致、一丝不苟的作风和平易近人的品质一直是我学习、生活中的榜样；他循循善诱的教导和不拘一格的创新研究思路给予我无尽的启迪，才有这个博士论文的顺利完成。感谢何老师提供良好的学习与科研环境以及留学机会，在此我向他表示崇高的敬意和衷心的感谢。

我还要感谢史忠植研究员，史老师严谨的治学态度、渊博的专业知识与敏锐的学术洞察力让我真正领会到了一个科学家的风范，他的敬业精神更是我一生学习的典范。

我要特别感谢罗平师兄，四年前何老师启发我们就迁移学习方向展开合作研究，可以说他对我学术上的具体指导和帮助伴随着我的整个博士研究。这四年来，通过共同讨论，他在学习工作上给了我很大的帮助，不仅是在疑难问题上给予详细解答，而且对研究方向的选择也给出了很好建议。没有他亲力亲为的帮助和指点，我不能这么顺利完成我的博士研究。他在生活上也给了我很大的帮助，在此我要对他表示诚挚的感谢。感谢美国新泽西 Rugter 大学的 Hui Xiong 教授，他在英文写作上也给了很大的帮助。感谢 HP 实验室的沈志勇博士，我们的合作也同样的愉快。这里也一并感谢 HP 实验室平时帮助过我的人，包括 Yuhong Xiong，Min Wang，Yong Zhao 等。

感谢中科院计算所智能科学课题组的胡宏副研究员和施智平老师，与他们的讨论也

让我受益匪浅。特别感谢胡兰平女士和田卫平女士，感谢她们在学习和工作上给予的支持和帮助，她们的工作使我感到实验室的温馨。感谢中科院计算所智能信息处理重点实验室的老师，他们是曹存根研究员、眭跃飞研究员、金芝研究员等。感谢计算所研究生部李琳老师、宋守礼老师、周世佳老师、冯刚老师等在生活中给我的关心和帮助。特别感谢宋守礼老师和我本科的辅导员重庆大学计算机学院的封传银老师、本科毕业设计的导师重庆大学计算机学院的何中市教授，是你们五年前的极力推荐和帮助才让我有机会进入计算所。

感谢与我同在机器学习与数据挖掘课题组的全体同学，我很高兴和珍惜与你们一起相处、工作的时光，不管是学习上还是项目上的合作都是那么的愉快。他们是我的师兄弟姐妹：赵秀荣、刘秋阁、赵卫中、谭庆、李金成、马旭东、罗文娟、李宁、李婷婷、杜长营、王群、尚田丰、敖翔、马云龙、董智等。还要感谢智能科学课题组一起学习和工作过的其他同学，大家的和谐相处使我在学习生活中保持愉快的心情，也感谢他们在工作和生活中的帮助与支持。他们是曾立、马慧芳、林欢欢、李志欣、刘曦、陈明、谭力、彭晖、郭立君、张志勇、张素兰、邱莉榕、石川、王茂光、黄瑞、罗杰文、石志伟、常亮、林芬、杨来、万常林、史俊、史春奇、余清、张子云、李志清、蒙祖强老师、苏变萍老师、崔志华老师、张冬蕾、张大鹏、牛温佳、曹鹏、王竹晓、陈立民、王晓峰、杨鲲、韩旭、张颖、杨来、叶飞、赵晨轶、陈坤荣、杨兴华、董琪、刘伟民、王喜顺、许新征等。感谢我的室友朱晏同学和肖斌同学，几年的相处使我们之间不仅仅是普通同学之间的情谊，我们一起营造宿舍良好的气氛，有一个舒心的环境生活学习，感谢你们。

感谢香港科技大学的杨强教授和明尼苏达大学双城分校的 George Karypis 教授，由于你们的邀请，我才有机会到香港科技大学和明尼苏达大学进行访问学习，是你们提供了让我拓宽视野，增长见识的机会，非常感谢你们。

感谢在我成长过程中陪伴和帮助过我的所有人。我要感谢的人很多，这里不能一一列举。不过我要特别感谢我初中的班主任杨文权老师，谢谢他一直以来的鼓励和帮助。感谢高中的所有老师和同学，是你们在我最艰难的时候帮助了我，让我顺利度过难关。

最后要感谢我的父母、兄弟姐妹和家人，是你们默默的支持和殷切的希望，不管是精神还是物质，都时刻激励着我，是我一直努力向前的动力。你们一直为我而自豪，今天我要说为拥有你们而感到骄傲。这里要特别感谢我亲爱的妈妈，她是世界上最勤劳最伟大的妈妈。这几十年来，她辛勤劳动，无私地奉献着这个家，从来都没有怨言，她不仅供我读书，还教会了我许多做人的道理。妈妈，没有您就没有我的今天，就没有这份完满的博士论文答卷，儿子感谢您。感谢上苍，九年前爸爸的意外事故，没有造成任何遗憾，让我始终拥有一个幸福健康美满的家庭。博士生涯即将结束了，准备参加工作，希望这是我人生的新起点，希望我的明天更加美好。

作者简历

姓名：庄福振　　性别：男　　出生日期：1983.07.18　　籍贯：福建漳州

2006.9 – 2011.7　　中国科学院计算技术研究所计算机软件与理论专业硕博研究生
2002.9 – 2006.7　　重庆大学计算机学院计算机软件与理论专业本科生

【攻读博士学位期间发表或录用的论文】

期刊论文：

[1] **Zhuang Fuzhen**, Luo Ping, Shen Zhiyong, Xiong Yuhong, He Qing, Shi Zhongzhi, Xiong Hui. Mining Distinction and Commonality across Multiple Domains using Generative Model for Text Classification [J]. ***IEEE Transactions on Knowledge and Data Engineering*** (TKDE), 2011, Accepted. (SCI, EI Source)

[2] **Zhuang Fuzhen**, Luo Ping, Xiong Hui, Xiong Yuhong, He Qing, Shi Zhongzhi. Cross-domain Learning from Multiple Sources: A Consensus Regularization Perspective [J]. ***IEEE Transactions on Knowledge and Data Engineering*** (TKDE), 2010, 22(12): 1664-1678. (impact factor (2009): 2.285) (SCI, EI)

[3] **Fuzhen Zhuang**, Ping Luo, Qing He, Zhongzhi Shi. Inductive Transfer Learning for Unlabeled Target-domain via Hybrid Regularization [J]. ***Chinese Science Bulletin***, 2009, 54(14): 2470-2478. (impact factor (2009): 0.917) (SCI)

[4] **Zhuang Fuzhen**, Luo Ping, Xiong Hui, He Qing, Xiong Yuhong, Shi Zhongzhi. Exploiting Associations between Word Clusters and Document Classes for Cross-domain Text Categorization [J]. ***Statistical Analysis and Data Mining***, Wiley, 2011, 4(1): 100-114. (this is an invited version of the best of SDM2010) (EI)

[5] **Fuzhen Zhuang**, Ping Luo, Qing He, Zhongzhi Shi. Inductive Transfer Learning for Unlabeled Target-domain via Hybrid Regularization [J]. ***Chinese Science Bulletin***, 2009, 54(11): 1618-1625. (In Chinese)

[6] He Qing, Du Changying, Wang Qun, **Zhuang Fuzhen**, Shi Zhongzhi. A Parallel Incremental Extreme SVM Classifier [J]. ***Neuro Computing***, 2010. (accepted) (SCI Source).

[7] Lin Huanhuan, **Zhuang Fuzhen**, Wang Wenjie, Shi Zhongzhi. A New Aggregator for Vertical Search Engine [J]. *Computer Simulation*, 2009, 26(5): 129-133. (In Chinese) (EI)

会议论文：

[1] **Zhuang Fuzhen**, Luo Ping, Shen Zhiyong, He Qing, Xiong Yuhong, Shi Zhongzhi, Xiong

Hui. Collaborative Dual-PLSA: Mining Distinction and Commonality across Multiple Domains for Text Classification [C]. In: ***Proceedings of the ACM 19th Conference on Information and Knowledge Management*** (CIKM'10). 2010: 359-368. (**among the 8 best paper candidates, student travel award**) (EI)

[2] **Zhuang Fuzhen**, Luo Ping, Shen Zhiyong, He Qing, Xiong Yuhong, Shi Zhongzhi. D-LDA: A Topic Modeling Approach without Constraint Generation for Semi-Defined Classification [C]. In: ***Proceedings of the 10th IEEE International Conference on Data Mining*** (ICDM'10). 2010: 709-718. (EI)

[3] **Zhuang Fuzhen**, Luo Ping, Xiong Hui, He Qing, Xiong Yuhong, Shi Zhongzhi. Exploiting Associations between Word Clusters and Document Classes for Cross-domain Text Categorization [C]. In: ***Proceedings of the SIAM International Conference on Data Mining*** (SDM'10). 2010: 13-24. (**among the 12 best paper candidates**) (EI Source)

[4] **Zhuang Fuzhen**, He Qing, Shi Zhongzhi. Feature Transformation for Efficiently Improving Performance of HSC [C]. In: *Proceedings of the 7th International Conference on Machine Learning and Cybernetics* (ICMLC'08). 2008: 423-428. (EI)

[5] **Zhuang Fuzhen**, He Qing, Shi Zhongzhi. Multi-Agent based Automatic Evaluation System for Classification Algorithms [C]. In: *Proceedings of International Conference on Information Automation* (ICIA'08). 2008: 264-269. (EI)

[6] He Qing, **Zhuang Fuzhen**, Li Jincheng, Shi Zhongzhi. Parallel Implementation of Classification Algorithms based on MapReduce [C]. In: *Proceedings of the 5th International Conference on Rough Set and Knowledge Technology* (RSKT'10). 2010: 655-662. (EI)

[7] He Qing, **Zhuang Fuzhen**, Shi Zhongzhi. The Data Selection Criteria for HSC and SVM Algorithms [C]. In: *Proceedings of the 4th International Conference on Natural Computation* (ICNC'08). 2008: 384-388. (EI)

[8] He Qing, **Zhuang Fuzhen**, Zhao Xiurong, Shi Zhongzhi. Enhanced Algorithm Performance for Classification Based on Hyper Surface Using Bagging and Adaboost [C]. In: *Proceedings of the 6th International Conference on Machine Learning and Cybernetics* (ICMLC'07). 2007: 3624-3629. (EI)

[9] Ma Xudong, Luo Ping, **Zhuang Fuzhen**, He Qing, Shi Zhongzhi, Shen Zhiyong. Combining Supervised and Unsupervised Models via Unconstrained Probabilistic Embedding [C]. In: ***Proceedings of the 22nd International Joint Conferences on Artificial Intelligence*** (IJCAI'11), 2011. (EI source) (Accepted)

[10] He Qing, Wang Qun, **Zhuang Fuzhen**, Tan Qing, Shi Zhongzhi. Parallel CLARANS Clustering Based on MapReduce [C]. In: *Proceedings of the 3rd International Conference on Machine Learning and Computing* (ICMLC'11). 2011: V1-236-V1-240. (EI source)

[11] He Qing, Li Tingting, **Zhuang Fuzhen**, Shi Zhongzhi. Frequent Term based Peer-to-Peer

Text Clustering [C]. In: *Proceedings of the 3rd International Symposium on Knowledge Acquisition and Modeling* (KAM'10). 2010: 352-355. (EI source)

[12] Luo Wenjuan, **Zhuang Fuzhen**, He Qing, Shi Zhongzhi. Effectively Leveraging Entropy and Relevance for Summarization [C]. In: *Proceedings of the 6th Asia Information Retrieval Society Conference* (AIRS'10). 2010: 241-250. (EI source)

[13] Li Tingting, **Zhuang Fuzhen**, He Qing. A noise handling method for hyper surface classification [C]. In: *Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD'10). 2010: 1484-1488. (EI)

[14] He Qing, Luo Wenjuan, **Zhuang Fuzhen**, Shi Zhongzhi. Local Bayesian Based Rejection Method for HSC Ensemble [C]. In: *Proceedings of the 7th International Symposium on Neural Networks* (ISNN'10). 2010: 404-412. (EI)

[15] He Qing, Ma Xudong, **Zhuang Fuzhen**, Shi Zhongzhi. The Effect of Scale Transformation for Hyper Surface Classification Method [C]. In: *Proceedings of the 8th International Conference on Machine Learning and Cybernetics* (ICMLC'09). 2009: 1856-1860. (EI)

[16] Luo Ping, **Zhuang Fuzhen**, Xiong Hui, Xiong Yuhong, He Qing. Transfer Learning From Multiple Source Domains via Consensus Regularization [C]. In: ***Proceedings of the ACM 17th Conference on Information and Knowledge Management*** (CIKM'08), 2008: 103-112. (EI)

【评审中论文】

[1] **Zhuang Fuzhen**, George Karypis, Xia Ning, He Qing, Shi Zhongzhi. Multi-view Learning via Probabilistic Latent Semantic Analysis [J]. Submitted to *Information Sciences* (INS). Under review

【攻读博士学位期间参加的科研项目】

[1] 国家自然科学基金面上项目“分布式计算环境下的并行数据挖掘算法与理论研究”(项目编号：No. 60975039)

[2] 国家自然科学基金重点项目“WEB 搜索与挖掘的新理论与方法”(项目编号：No. 60933004)

[3] 国家 973 项目子课题“非结构化(图像)信息的内容理解与语义表征”(项目编号：No.2007CB311004)

[4] 国家自然科学基金面上项目“基于超曲面的覆盖分类算法与理论研究”(项目编号：No. 60675010)

[5] 北京市自然科学基金“海量高维、多类数据分类法研究及其应用”(项目编号：No. 4052025)