中山大学

硕士学位论文

主动迁移学习模型的研究与应用

姓名：施潇潇

申请学位级别：硕士

专业：计算机应用技术

指导教师：任江涛

20090520

论文题目：主动迁移学习模型的研究与应用
专　　业：计算机应用技术
硕 士 生：施潇潇
指导教师：任江涛 讲师

摘 要

分类是数据挖掘领域的一个重要技术。在数据独立同分布的假设下，分类技术根据已有的带有类别标签的训练样本建立分类模型，并利用该模型尽量准确地对新的数据样本进行预测与分类。但是在实际应用中，满足独立同分布条件的训练样本往往相当缺乏，造成分类模型的准确率下降。近年来，为了解决训练样本不足的问题，学者们提出了主动学习和迁移学习两类方法。主动学习的目的是选取少量的具有代表性的数据样本，并由领域专家给这些样本标注类别标签，以使其成为训练样本。然后，主动学习可以用这少量的具有代表性的训练样本建立准确率高的分类模型，从而减少了对训练样本的数量的需求。另一类方法叫做迁移学习。迁移学习的目的是借助来自其他领域的，不满足数据独立同分布的训练样本，帮助目标领域建立分类模型，从而减少了对目标领域的训练样本的要求。

但是，在解决训练样本缺乏的问题上，主动学习和迁移学习各有不足。一方面，主动学习对训练样本的需求依然很大，造成某些领域获得训练样本的代价依然很高；迁移学习虽然可以以零代价获得训练样本，但是从其他领域迁移过来的训练样本有可能与目标领域分布差异很大，造成“负迁移”，即使得分类模型的准确率下降。为了更好地解决训练样本不足的问题，在这两类方法的基础上，本文提出了主动迁移学习的模型。本文结合主动学习的思想，解决负迁移的问题；并利用了迁移学习来降低主动学习中获得训练样本的代价。理论和实验证明本文提出的模型能有效地避免负迁移，提高分类的准确率，并有效地降低获得训练样本的代价。

另外，为了说明主动迁移学习模型的有效性，本文还以文本分类为例，分析了迁移学习的实际例子，并结合向量准换、特征选择、降维等技术，解决了文本

分类所面临的训练样本严重缺乏所导致的分类效果不理想的问题。基于文本挖掘的实验也证明了该算法能有效地避免负迁移，并且有比较好的实用性与比较理想的分类效果。

关键词：分类；主动学习；迁移学习；文本挖掘

Title: The Research and Application of Active Transfer Learning Model

Major: Computer Application and Technology

Name: Xiaoxiao Shi

Supervisor: Prof. Jiangtao Ren

Abstract

Classification is one of the important techniques in data mining. With the assumption that all the data are independently identical distributed (i.e., i.i.d.), classification works well to build an accurate classifier from the labeled training data, to predict the unlabeled test data. However, we usually encounter the situations when we do not have sufficient labeled training data satisfying the i.i.d. assumption in practice, and finally decrease the accuracy of classification. To resolve the problem, active learning and transfer learning are two separate solutions proposed in recent years. Active learning aims at selecting a small amount of representative data and querying their class labels from the domain experts with some costs. The goal is to build an accurate classifier with this small amount of informative data in order to alleviate the need for labeled samples. Another solution is called transfer learning, which can borrow labeled examples from related source domain, to help build the classifier in target domain. Though the borrowed examples may not be i.i.d., transfer learning decreases the need for labeled data in the target domain.

However, there are still disadvantages of these two solutions to solve the problem of label deficiency. On the one hand, active learning still requires some labeled examples; while in some cases, it is still difficult and expensive to obtain labeled examples. On the other hand, although transfer learning can borrow labeled examples from other domain without any cost, it has the risk of "negative transfer" when the data from other domain are too different from that of the target domain. This

will finally cause a significant drop of the learning accuracy. To better solve the label deficiency problem, we propose an active transfer learning model. The proposed model solves the problem of "negative transfer" by applying the idea of active learning; while at the same time, decrease the cost of obtaining examples by the idea of transfer learning. Theoretical and empirical results show that, the proposed approach not only increases learning accuracy, but also decreases the cost of obtaining labeled examples.

In addition, we extend the proposed active transfer learning model to the field of text classification. By applying the techniques of vertex generation, feature selection, etc, we alleviate the problem of label deficiency in text mining. And the application shows that the proposed approach can effectively avoid negative transfer and increase the learning accuracy.

Key Words: Classification, Active Learning, Transfer Learning, Text Mining

论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：2009年 5 月 20 日

学位论文使用授权声明

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅，有权将学位论文的内容编入有关数据库进行检索，可以采用复印、缩印或其他方法保存学位论文。

学位论文作者签名：

导师签名：

日期：2009 年 5 月 20 日

第1章 绪论

本章将介绍本文的研究背景及其意义，并对相关的研究工作进行概述，最后将介绍本文的内容安排。

1.1 研究的背景与意义

近十几年来，人们利用信息技术生产和搜集数据的能力大幅度提高。千千万万个数据库被用于商业管理、政府办公、科学研究和工程开发等等，要想使数据真正成为一个公司的资源，只有充分利用它为公司自身的业务决策和战略发展服务才行，否则大量的数据可能成为包袱，甚至成为垃圾。因此，面对人们被数据淹没却饥饿于知识的挑战，数据挖掘和知识发现技术[1]应运而生，并得以蓬勃发展，越来越显示出其强大的生命力。

数据挖掘[1][2]（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。人们把标准数据看作是形成知识的源泉，就像从矿石中采矿一样。标准数据可以是结构化的，如关系数据库中的数据。也可以是半结构化的，如文本图形、图像数据、甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的、也可以是非数学的、可以是演绎的、也可以是归纳的、挖掘出的知识可以被用于信息管理、查询优化、决策支持、过程控制等。还可以用于数据自身的维护。因此，数据挖掘是一门很广义的交叉学科，涉及人工智能技术，统计技术与数据库技术等多种技术。它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。

分类是数据挖掘中一种重要的挖掘任务和挖掘方法。它根据数据库中带有类别标签的数据样本(也称为“训练样本”)建立分类模型(也称为“分类器”)，并利用分类模型预测新的数据样本的分类标号。例如，可以建立一个分类模型，对银行贷款的安全或风险进行分类；还可以建立分类模型根据潜在顾客的收入和职

业，预测他们在营销产品上的花费。许多分类和预测方法已经广泛地被机器学习，专家系统，统计学和神经生物学所应用。

迄今为止人们提出了很多分类算法，例如基于决策树模型的分类方法（如ID3[3]，C4.5[4]等等），基于分割的方法（如SVM[5]），数据拟合方法（如Logicstic拟合法[6]），还有结合图模型的产生式分类方法(Graphical Model[7])。不同的算法有着不同的应用背景，有的适用于大规模数据集，有的适用于有时序特性的数据，有的算法思想简单，有的算法有严密的数学背景。总的来说，算法都试图从不同的途径实现对数据集进行高效、可靠的分类。

但是，这些分类模型的可靠性依靠于一个严格的假设：用于建立模型的训练样本与新的数据样本满足独立同分布的条件[8]。在数据挖掘的实际应用中，人们发现满足这一条件的训练样本往往很缺乏，而且很难获得。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；在文本分类领域，人们发现已有的训练样本远远不足够建立一个可靠的分类模型，而标注大量文档往往需要高薪聘请大量的专家，造成获得训练样本的代价很高。总而言之，人们一方面需要大量的训练样本以建立准确率高的分类模型；而另一方面，获得大量的训练样本在很多实际应用中几乎不可能[9]。

面对训练样本缺乏的问题，学者们提出了两种不同的解决方案。第一种方案叫做主动学习(active learning)[10]。该方法的主要思想是主动地选择少量的，核心的数据作为训练样本。例如其中的一类重要的方法是选择接近分类边界（decision boundary）的数据作为训练样本[10]。其目的是使得这个新的训练样本能提供比较多的信息给分类器，从而产生更好的分类边界。总而言之，主动学习方法的核心是如何选择少量的，信息量大的数据点作为训练样本，以用最少的训练样本建立准确率高的分类模型。

另外一种方法叫做迁移学习（transfer learning）[11]。该方法是近年来数学挖掘领域各个重要学术会议如KDD, ICDM等的热门话题。迁移学习与传统机器学习方法的主要区别是前者没有数据独立同分布的假设。于是，迁移学习旨在借助来自其它领域的训练样本，更好地对本领域的数据进行分类。其思想来源于人类的学习经验：例如，骑摩托车的人可以很快地学会开汽车，因为骑摩托车的交通规则，对运动物体的距离感等知识可以迁移到学开汽车；熟悉下五子棋的人可

以很快地学会下围棋，因为棋盘的结构，规则等都相似。于是，尽管目标领域的训练样本很缺乏，迁移学习可以从大量的来自其它领域的数据集中取得大量的训练样本，以更好地训练分类器。例如，文献[12]利用数据特征的相似性实现了文本挖掘中的知识迁移，文献[13]利用分类器训练参数的相似性实现了知识的迁移等等。

由于对迁移学习的研究目前还在起步阶段，它产生了不少热门课题。其中一个重要的研究问题是如何避免“负迁移”[14]，即当来自其它领域的数据与当前数据具有很大的差异时，如何保证迁移学习不会降低分类的准确率。由于负迁移问题还没有可行的解决方案，笔者希望把主动学习与迁移学习的思想集合起来，提出主动迁移学习的框架。主动迁移学习一方面降低负迁移的风险，另一方面降低获得训练样本的代价。笔者希望新的模型能被真正用到实际的研究与生产中，并有比较好的实用性。

1.2 相关研究

本文涉及的课题主要与主动学习和迁移学习相关。对主动学习的研究主要分为基于池与基于流的两类主动学习背景。其中基于池的算法假设主动学习模型可以从一个样本池中选取最好的训练样本，而基于流的算法假设主动学习模型只能判断当前的样本是否应该被设定为训练样本，而不能对其他样本进行存取。因此，可以把基于池的算法看成是全局算法，而基于流的算法为局部算法。本文所涉及的主动学习策略均为基于池的方法。

虽然对迁移学习的研究刚刚起步，其是近年来各个顶级会议的研究热点，而与迁移学习问题相关的研究方向主要有两个：

1、样本选择偏见(Sample Selection Bias)。当训练样本的产生存在特征，类别或者综合偏见时，就会产生对应的样本选择偏见[15]。例如在医疗诊断系统的采样中，也许只记录了区域内病人的医疗信息。当该系统要在区域外运行时，很可能由于训练样本不具代表性而使得系统准确率下降。样本选择偏见与迁移学习的共同点是训练样本和测试样本均不满足独立同分布的假设。但是迁移学习理论上可以主动地接纳这些分布不一致的数据；而对样本选择偏见的研究则是如何避免

分布的不一致。这两个研究方向虽然有类似的假设，但是动机不同。

2、多任务学习(Multiple Task Learning)[13]。对多任务学习问题的研究也是比较热门的研究方向。其基本假设是同时学习多个任务比只学习一个任务的学习效果（如分类效果）要更好。例如要对各个学校的学生进行分类，则可以把来自不同学校的训练样本（学生资料）综合起来训练一个基本的分类模型，然后在该分类模型的基础上再对各个学校生成有特色的分类器。由于训练样本来自多个任务，与迁移学习相似，多任务学习需要解决的一个重要问题是如何处理数据非独立同分布的情况。

近年来在各个国际会议上也提出了不少优秀的迁移学习的算法。总的来说，这些算法主要解决的问题是如何在数据非独立同分布的情况建立分类模型。大概可以分成三大类：

1、基于采样的知识迁移方法。文献[16]对传统的 Adaboost 算法框架进行了修改，以使得采样后的训练样本能提高分类的准确率。文献[16]则对各个数据样本赋予不同的权值，使得与测试样本分布相似的训练样本获得较大的权值，最后进行一次采样并根据新的训练集构造新的分类器。

2、基于公共特征的知识迁移方法。这类方法又可以分为两个方向。第一个方向是先寻找公共的特征子空间（如利用特征选择或共聚类的方法），然后通过特征子空间构造分类模型[12]。另外一类方法是寻找经过特征的投影空间，其基本的思路是（1）在该投影空间中，训练样本与测试样本有相似的分布；（2）在该空间中，训练样本具有很明显的分割[17]。

3、参数迁移方法。该类方法主要用于 Logistic 拟合方法或者产生式分类模型中。其基本假设是分布相似的数据训练出来的分类器也应该比较相似[18]。这类方法被广泛应用于多任务学习问题中。

尽管对迁移学习的研究涌现了不少算法，但是这些算法主要着重于解决“在数据分布相似”的情况下，如何进行知识的迁移。而文献[14]提出了负迁移的概念：当训练样本的分布与测试样本的分布差别很大时，知识的迁移也许会对分类的结果造成不好的影响，如降低分类的准确率。这个时候就说发生了负迁移。而负迁移的存在也可以在日常生活中发现。例如，骑自行车的人学三轮车的时候往往会不习惯三轮车相反的重心而造成翻车。由于负迁移是个新的概念，当前还没

有很好的解决框架。笔者在本文希望结合主动学习的思想，提出主动迁移学习的框架，以降低迁移学习中负迁移的风险，并同时降低获得训练样本的代价。

1.3 本文的内容安排

本文的研究内容主要有以下几个方面：

1、提出并实现了主动迁移学习框架 ACTRAK，从理论上证明了（1）与传统的主动学习相比，该框架能降低获得训练样本的代价。（2）与传统的迁移学习相比，该框架能有效地降低负迁移的风险。接着，笔者通过对比的方法进行实验分析，说明新的算法框架比以往的主动学习算法和迁移学习算法在有效性和稳定性上的提高。

2、在文本分类的领域使用主动迁移学习框架 ACTRAK，并通过实验分析说明算法的有效性与实用性。

本文主体共分为以下几大部分：

第一部分 绪论，介绍了课题的背景和意义，概括了课题具体所涉及的理论及相应的特点；简述了所涉及技术的发展和目前的研究成果等，最后列出了本论文的主要研究内容和结构。

第二部分 数据挖掘的相关概念，介绍了数据挖掘的起源，讨论了数据挖掘领域的相关概念和方法，和数据结构和数据类型。并重点介绍分类了数据挖掘中的主要分类算法，最后对分类结果的评价方法进行了讨论。

第三部分 主动学习理论和迁移学习理论，讨论了主动学习和迁移学习的思想以及其对训练样本不足情况下的分类分析的影响，主要介绍几种主流的主动学习和迁移学习策略。

第四部分 主动迁移学习，先介绍了主动学习和迁移学习在解决训练样本不足情况下的分类问题的缺点。然后提出主动迁移学习的基本思想与算法步骤，分析它的理论基础。最后会用实验对比分析主动迁移学习框架的效果。

第五部分 主动迁移学习在文本挖掘中的应用，讨论了文本分类在数据挖掘中的特殊性，通过对文本数据的变化处理后，把第五章提出的主动迁移算法应用于文本分类。并用实验分析说明应用的效果。

第六部分 总结与展望，总结了本文的研究过程及研究效果，并讨论了对该研究进一步深化的展望。

第2章 数据挖掘的相关概念

随着信息技术的快速发展和信息搜集能力的日益提高，产生了海量的数据。这些海量的数据或者是以静态的形式存储在企业的物理存储器上，或者是不被存储而瞬时出现的动态数据。面对如此丰富的海量数据，传统的数据处理方法和能力已经远远不能满足实际的需求。面对日趋激烈的市场竞争，人们需要从这些蕴涵着丰富决策信息的数据中抽取能帮助人们进行决策的知识。在需求的强烈驱动下，数据挖掘技术应运而生。

本章依次介绍了数据挖掘的定义及功能，数据挖掘中的数据结构，并重点介绍了数据挖掘中的分类算法以及对分类方法的评价标准。

2.1 数据挖掘的概念及功能

数据挖掘（Data Mining）是指从大量的数据(结构化和非结构化)中提取有用的信息和知识的过程。在这个定义中，要求数据源应该是大量的、真实的、含有噪音的；所发现的信息和知识是潜在的并隐藏在大量数据背后的，是用户感兴趣的、可理解、可运用的知识。所以，数据挖掘有时也被人们称为知识挖掘、知识提取、知识发现等 [3]。

被挖掘的原始数据其形式是复杂多样化的，有结构化数据，异构化的数据，半结构化数据，甚至是非结构化的数据(包括多媒体数据)。使用的数据挖掘方法也是丰富多彩的，如有数学方法，也有非数学的方法；有演绎的方法，也有归纳的方法；有聚类的方法，也有分类的方法；有关联的方法，也有孤立点分析方法；有对文本数据的挖掘方法，也有对复杂数据的挖掘方法；等等。被挖掘出来的数据和知识的用途是广泛的，可以用于知识管理、信息服务的知识推送，企业竞争、客户关系管理以及用于决策支持和过程控制等领域。

数据挖掘的本质就是知识发现，但不要认为这里所指的知识发现是发现放之四海而皆被的真理，也不是去发现新的物质或新的自然科学定理，更不是利用计算机证明某个定理是否正确。实际上，它所有发现的知识都是隐藏在大量数据之

中的关联信息，所有的知识都是有特定前提和约束条件的，是面向特定领域的，而且，这些知识还要能够易于被用户理解，能用自然语言表达所发现的结果。

一般而言，数据挖掘的功能与挖掘的目标数据类型是相关的。某些功能只能应用在某种特定的数据类型上，而某些功能则可以应用在多个不同类型的数据库上。对于数据挖掘任务的确定，必须综合考虑数据挖掘功能、要挖掘的数据类型和用户的兴趣。

数据挖掘的功能主要包括以下几个方面：概念描述、关联分析、分类、聚类、偏差检测和时序演变分析。数据挖掘功能一般可以分为描述和预测两类。描述性挖掘分析主要用来刻画数据集合的一般特性；预测性挖掘则是根据当前数据进行分析推算，从而达到预测的目的。

2.1.1 概念描述

概念描述（concept description）就是通过对与某类对象关联数据的汇总、分析和比较，对此类对象的内涵进行描述，并概括这类对象的有关特征。这种描述是汇总的、简洁的和精确的，当然也是非常有用的知识。例如：关系数据库中的一个关系(即一个表)代表了一个对象集，其中的每个元组可以看作是一个对象。每个对象有一个惟一标示和数个属性值。在一个或一组属性上取值相同的对象构成一个对象类。

概念描述分为特征性描述和区别性描述。前者描述某类对象的共同特征，后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性；生成区别性描述则涉及目标类和对比类中对象的共性。

特征性描述是目标类数据的一般特征或特性的汇总。基本方法有两种：基于数据立方体的 OLAP 方法和面向属性的归纳方法（AOI，Attribute Oriented Induction）。OLAP 方法中涉及到对数据立方体的上卷操作，其实质就是一种交互式的、由用户控制的、按照指定维的层次向上汇总的过程。由此，人们可以发现汇总后的、处于更高概念层次的目标类知识。面向属性的归纳方法的主要思想是，首先建立对象集属性的概念层次，然后在较高的概念层上对原始数据进行抽象，并发现和表示知识，就可以得到关于对象类的较高级的知识。与 OLAP 方法不同，AOI 方法不必每一步都与用户交互. 并且可以自动建立静态或动态的

概念层次结构。

区别性描述是将目标类数据的一般特性与一个或多个对比类的数据的一般特性进行比较。而这种比较必须是在具备可比性的两个或多个类之间进行的。区别性描述所采用的方法与特征性描述相似。例如，对研究生和本科生的特征进行比较。可能会发现研究生的年龄较大，成绩优秀；而本科生的年龄较小，成绩优秀的所占比例不大。

2.1.2 关联分析

关联分析（association analysis）就是从大量的数据中发现项集之间有趣的关联、相关关系或因果结构以及项集的频繁模式。数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。例如，某两个或多个变量的某个固定取值组合频繁的出现(所谓频繁项集)，就可以认为这个固定取值的组合表示了一种关联规则。可以认为这个固定取值的组合表示了一种关联规则。一般而言，这种规则可以这样表述："80%包含项 A、B 和 C 的记录同时也包含项 D 和 E"。而百分比 80%称为这条规则的置信度（Confidence），它可以衡量规则的确定性；还有一个度量用来衡量规则的有用性，称为支持度（Support）。这两个度量的定义公式如下：

Confidence ＝ 同时包含项 A、B、C、D 和 E 的记录数÷同时包含 A、B 和 C 的记录数

Support ＝ 同时包含项 A、B、C、D 和 E 的记录数÷总记录数

通常的数据挖掘系统使用最小置信度和最小支持度作为阈值来筛选有价值或有兴趣的关联规则，用户可以自行设定阈值，以调整挖掘结果。

关联可分为简单关联(例如，购买尿布的顾客中有 90%的人同时购买啤酒)和因果关联。根据关联规则所涉及变量的多少，可以分为多维关联规则和单维关联规则。例如，"所有顾客中，有 2%的年龄为 20—29 岁，年收人为 20 000～29000 元的顾客购买 CD 机，这个年龄和收入组的顾客购买 CD 机的可能性为 60%"。这一规则涉及到年龄、收入和购买三个变量(即三维)，可称为多维关联规则。而在规则"所有购买记录中，有 1%同时购买了计算机和打印机，购买了计算机的同时购买打印机的可能性为 60%"中，由于只涉及到购买事物这一个变量，所

以称为单维关联规则。

以上涉及的是结构化数据的关联分析，对于非结构化或半结构化的文本数据，也可以进行一种基于关键字或词的关联分析。文本关联分析是这样一个过程：首先，对文本数据库中的文档数据进行处理，得到每个文档的关键字或词集合。然后把每一文档的惟一标示及其关键字或词集合当作一个数据对象，对这个数据对象集合应用结构化数据关联挖掘第法，可以发现经常连续出现或紧密相关的关键字或词。通过文本关联分析，可以找出词或关键字间的关联，这往往是一种有趣的规则。

2.1.3 分类

信息的有监督学习，或称分类（classification），是信息处理的重要组成部分。事实上，它是人们对信息最自然而然的处理。信息分类将信息或数据有序地聚合在一起，有助于人们对事物的全面和深人了解。根据处理对象的不同，信息分类可以分为结构化数据分类和文本数据分类两种。

结构化数据分类的过程为：对于给定的一个对象集合，用一组标记（即一组具有不同特征的类别）来为每一个对象进行归类，然后找出描述并区分这些类的模型(或函数)，利用这个模型可以预测类标记未知的对象类。这种模型可能是显式的，如一组规则定义；或者是隐式的，如一个数学模型或公式。目前有很多种分类方法，典型的有线性回归方法、决策树方法、（IF-THEN）规则方法、神经网络方法和支持向量机方法等。分类不仅可以预测数据对象的类标记，还可以用来预测某些空缺或未知的数据值。

文本数据分类是一种重要的文本挖掘工作，鉴于越来越多的电子文档已经无法出人工处理，自动文本分类的研究是十分有意义的。文本分类的处理过程与结构化数据分类处理过程相似。首先，把一组预先归类的文档作为训练集，对训练集进行分析得出分类模型，然后就可以利用这些模型对未分类的文档进行归类。由于本文的研究工作是基于有监督学习，即分类分析的，因此笔者将在第 2.4 节对分类方法进行详细的介绍。

2.1.4 聚类

无监督学习即聚类（clustering），它是一种特殊的分类，与分类分析法不同，聚类分析是在预先不知道欲划定类的情况下(如没有预定的分类表、没有预定的类目)，根据信息相似度原则进行信息集聚的一种法。聚类的目的是根据最大化类内的相似性、最小化类间的相似性这一原则合理的划分数据集合，并用显式或隐式的方法描述不同的类别。因此，聚类的意义也在于将观察到的内容组织成类分层结构，把类似的事物组织在一起。通过聚类，人们能够识别密集的和稀疏的区域，因而发现全局的分布模式，以及数据属性之间的有趣的关系。

聚类也分为结构化数据聚类和文本数据聚类两种。结构化数据聚类指的是这样一个过程：它将物理或抽象对象的集合分组成为由类似的对象组成的多个类。聚类分析的主要方法是基于距离的统计方法。另外，机器学习领域的自组织特征映射（SOM）方法也被用于聚类分析。聚类分析的应用很广泛，例如，可以根据公司客户的基本信息发现不同的客户群，并且用购买模式来刻画不同的客户群的特征。

2.1.5 偏差检测

偏差检测（deviation detection）就是对数据库中的偏差数据进行检测和分析。数据库中的数据常有一些异常记录，它们与其他数据的一般行为或模型不一致。这些数据记录就是偏差（deviation），也叫孤立点。偏差的产生可能是某种数据错误造成的，也可能是数据变异所固有的结果。从数据库中检测这些偏差很有意义，例如在欺诈探测中，偏差可能顶示着欺诈行为。因此，偏差检测和分析就成为一个有趣的数据挖掘任务。

偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。

偏差检测的主要问题在于：偏差点与数据库记录之间不一致的标准如何确定；以及如何找到一个有效的方法来发现这样的偏差点。

偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别。基于计算机的偏差检测算法大致有三类：统计学方法，基于距离的方法和基于偏移的方法。

例如，偏差检测可以发现信用卡欺骗。通过检测一个给定账号的支付记录，如果发现存在着某个付款数额比一般的付款数额高出很多的付费记录，则可能是信用卡欺诈。

2.1.6 时序演变分析

数据的时序演变分析（temporal evolution analysis）是针对事件或对象行为随时间变化的规律或趋势，并以此来建立模型。它主要包括时间序列数据分析、序列或周期模式匹配和基于类似性的数据分析。例如，对股票市场交易数据进行时序演变分析，则可能得到这样的规则：若 AT&T 股票连续上涨两天且 DEC 股票不下跌，那么第三天 IBM 股票上涨的可能性为 75%。

文本数据中所涉及到的事件、对象、时间及地点等一般的关系，已在人们的记忆里形成了一些固定的范畴和关系结构，发掘出这些结构就可以发现文本数据所反映的事物发展变化的时间顺序，以此作为理解文本的一条重要线索。这就是文本数据的时序分析。

2.2 数据挖掘中的数据结构

由于本文主要研究分类方法，因此本节具体介绍分类分析的实际数据库应用中，分类的数据结构及其特点。

2.2.1 主要数据结构

假设要分类的数据集合包含了 n 个数据对象，这些数据对象可能表示人、房子、文档、DNA 序列等。许多基于内存的分类算法选择如下两种有代表性的数据结构：

1、数据矩阵（data matrix）：它用 p 个变量（属性）来表现 n 个对象，例如用年龄、身高、体重、性别、种族等属性来表示一个“人”。这种数据结构是关系表的形式，或者看成$n \times p$(n个对象$\times$ p个对象) 的矩阵。通常，最后一个特征为对应对象的类别标签。

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

2、相异度矩阵（dissimilarity matrix）：存储 n 个对象两两之间的近似性，表现形式是一个 $n \times n$ 维的矩阵。而对象的类别标签一般需要另外用一个一维向量存储。

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \cdots & \cdots & \cdots & \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix}$$

在这里 $d(i,j)$ 表示对象 i 和 j 之间的相异性的量化表示，通常它是一个非负的数值有 $d(i,j)=d(j,i)$，$d(i,i)=0$，并且当对象 i 和 j 越相似或接近，其值越接近 0；两个对象越不同，其值越大。

数据矩阵经常被称为二模矩阵（two-mode），而相异度矩阵被称为单模（one-mode）矩阵，这是因为前者的行和列代表不同的实体，后者的行和列代表相同的实体，许多分类算法以相异度矩阵为基础。如果数据是用数据矩阵的形式表示的，在使用该类算法之前要将其化为相异度矩阵。

2.2.2 数据类型

分类分析起源于统计学，传统的分析方法大多是在数值类型数据的基础上研究的。然而数据挖掘的对象复杂多样，要求分类分析的方法不仅能够对属性为数值类型的数据进行，而且要适应数据类型的变化。一般而言，在数据挖掘中，对象属性经常出现的数据类型有：区间标度变量，二元变量，标称型、序数型和比例标度型变量以及混合类型的变量。

1、区间标度变量。区间标度变量是一个粗略线性标度的连续度量。典型的例子则包括重量和高度，经度和纬度坐标，以及大气温度等。为了将数据或对象集合划分成不同类别，必须定义差异性或相似性的测度来度量同一类别之间数据的相似性和不属于同一类别数据的差异性。同时要考虑到数据的多个属性使用的是不同的度量单位，这些将直接影响到分类分析的结果，所以在计算数据的相似

性之前先要进行数据的标准化。

对于一个给定的有 n 个对象的 m 维（属性）数据集，主要有两种标准化方法：

平均绝对误差 S_p：

$$S_p = \frac{1}{n}\sum_{i=1}^{n}\left|x_{ip} - m_p\right|$$

这里 x_{ip} 表示的是第 i 个数据对象在属性 p 上的取值，m_p 是属性上的平均值，即

$$m_p = \frac{1}{n}\sum_{i=1}^{n}x_{ip}$$

标准化度量值 Z_p：

$$Z_p = \frac{x_{ip} - m_p}{S_p}$$

这个平均的绝对误差 S_p 比标准差 σ_p 对于孤立点具有更好的鲁棒性。在计算平均绝对偏差时，属性值与平均值的偏差 $\left|x_{ip} - m_p\right|$ 没有平方，因此孤立点的影响在一定程度上被减小了。

数据标准化处理以后就可以进行属性值的相似性测量，通常是计算对象间的距离。对于 n 维向量 x_i 和 x_j，有以下几种距离函数：

欧氏距离：

$$D(x_i, x_j) = \left\|x_i - x_j\right\| = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \tag{3.1}$$

曼哈顿距离：

$$D(x_i, x_j) = \sum_{k=1}^{n}\left|x_{i,k} - x_{j,k}\right| \tag{3.2}$$

一般化的明氏距离：

$$D_m(x_i, x_j) = \left[\sum_{k=1}^{m}(x_{i,k} - x_{j,k})^m\right]^{\frac{1}{m}} \tag{3.3}$$

当 m=2 时，明氏距离 D_2 即为欧氏距离；当 m=1 时，明氏距离 D_1 即为曼哈顿

距离。

对于欧氏距离和曼哈顿距离满足以下条件：

（1）$D(x_i,x_j)\geq 0$ ：距离是一个非负数值。

（2）$D(x_i,x_j)=0$ ：对象与自身的距离是零。

（3）$D(x_i,x_j)=D(x_j,x_i)$ ：距离函数具有对称性。

（4）$D(x_i,x_j)\leq D(x_i,x_k)+D(x_k,x_j)$ ：x_i到x_j的距离不会大于x_i到x_k和x_k到x_j的距离之和（三角不等式）。

2、二元变量。二元变量只有两个状态：0和1。其中二元变量又分为对称的二元变量和不对称的二元变量。前者是指变量的两个状态不具有优先权，后者对于不同的状态其重要性是不同的。对于二元变量，度量两个变量的差异度由简单匹配系数（对于对称的情况）和Jaccard系数（对于非对称情况）决定。设两个对象x_i和x_j，q是属性值在两个对象中都为1的属性个数，r是属性值在x_i中为1而在x_j中为0的属性个数，s是属性值在x_i中为0而在x_j中为1的属性个数，t是属性值在两个对象中都为0的属性个数。

简单匹配系数：

$$d(x_i,x_j)=\frac{r+s}{q+r+s+t}$$

Jaccard系数：

$$d(x_i,x_j)=\frac{r+s}{q+r+s}$$

3、标称型、序数型和比例标度型变量。标称变量是二元变量的推广，它可以有多个状态值，状态之间是无序的。具有这种数据类型的属性也称分类（categorical）属性。它的差异度可用简单匹配法来计算：

$$d(x_i,x_j)=\frac{p-m}{p}$$

其中m是对象x_i和x_j中匹配的属性个数，而p是全部属性个数。

4、混合类型的变量。以上讨论了各种数据类型和它们差异度的计算方法，

在实际数据库中，对象是由混合类型的变量描述的。在实际分类分析中，将不同的类型属性组合在同一个差异度矩阵中进行计算。设数据包含 m 个不同类型的属性，对象x_i和x_j之间的差异度定义为：

$$d(x_i,x_j)=\frac{\sum_{p=1}^{m}\delta_{ij}^{(p)}d_{ij}^{(p)}}{\sum_{p=1}^{m}\delta_{ij}^{(p)}}$$

其中如果x_{ip}或x_{jp}缺失，或$x_{ip}=x_{jp}=0$，且变量是不对称二元变量，则指示项$\delta_{ij}^{(p)}=0$；否则$\delta_{ij}^{(p)}=1$。

如果属性 p 是二元变量或标称变量：如果$x_{ip}=x_{jp}$，$d_{ij}^{(p)}=0$ 否则，$d_{ij}^{(p)}=1$。

如果属性 p 是序数型或比例标度型变量：将其转化为区间标度变量值对待。

2.3 数据挖掘中的分类算法

本小节，笔者将分别介绍数据分类的几类重要模型，如贝叶斯分类、判定树归纳、支持向量机等。

2.3.1 贝叶斯分类

贝叶斯分类算法是基于贝叶斯定理的一种统计学分类算法[23]。它们可以预测类成员关系的可能性，如给定样本属于一个特定类的概率。如果出现类别重叠现象，贝叶斯方法算法采用两种方法处理这种情况：一是选择后验概率最大的类别，二是选择效用函数最大（或损失最小）的类别。朴素贝叶斯分类算法概述如下：

设每个训练样本用一个n维特征向量X={x1,x2,……,xn}表示，分别描述训练样本的n个属性A1,A2,……,An的属性值。假定有m个类C1,……,Cm。给定一个未知类别的样本X，朴素贝叶斯分类算法将预测X属于具有最高后验概率（条件X下）的类。也就是说，朴素贝叶斯分类将未知的样本分配给类Ci，当且仅当$P(C_i|X)>P(C_j|X)$，1≤j≤m，j≠i。$P(C_i|X)$最大的类Ci称为最大后验假定。

根据贝叶斯定理,

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{3.4}$$

因为P(X)对所有的类都为常数，所以只需要$P(X \mid C_i)P(C_i)$最大即可将x分配给类Ci。

类的先验概率可以用$P(C_i) = \frac{s_i}{s}$计算，其中s_i是类C_i中的训练样本数，而s是训练样本总数。如果训练样本的属性很多，计算$P(X \mid C_i)$的开销可能非常大，为了降低计算$P(X \mid C_i)$的开销，可以做类条件独立的朴素假定。给定样本的类标号，假定属性值相互独立，即在属性间不存在依赖关系。这样,

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \tag{3.5}$$

概率$P(x_k \mid C_i)$, k=1,……,n, 可以由训练样本估值。

简而言之，为了对未知样本X分类，朴素贝叶斯分类算法对每个类C_i计算出$P(X \mid C_i)P(C_i)$。样本X被分配到类C_i，当且仅当$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j)$，$1 \leqslant j \leqslant m, j \neq i$。

朴素贝叶斯分类算法可以与决策树和神经网络分类算法相媲美。用于大型数据库，贝叶斯分类算法也表现出高准确性与高速度。理论上讲，与其他所有的分类算法相比，朴素贝叶斯分类算法具有最小的错误率。然而实践中并非总是如此，这是由于对其应用的假定（如类条件独立性）的不准确性，以及缺乏可用的概率数据造成的。另外，朴素贝叶斯分类算法没有规则输出。目前，出现了很多降低独立性假设的贝叶斯分类算法，如贝叶斯信念网络等。贝叶斯分类算法还可以用来为不直接使用贝叶斯定理的其他分类算法提供理论判据。

2.3.2 决策树分类算法

决策树学习是一种逼近离散值函数的算法，对噪声数据有很好的健壮性，且能够学习析取表达式，是最流行的归纳推理算法之一，已经成功应用到医疗诊断、

评估贷款申请的信用风险、雷达目标识别、字符识别、医学诊断和语音识别等广阔领域。

决策树分类算法使用训练样本集合构造出一棵决策树，从而实现了对样本空间的划分。当使用决策树对未知样本进行分类时，由根结点开始对该样本的属性逐渐测试其值，并且顺着分枝向下走，直到到达某个叶结点，此叶结点代表的类即为该样本的类。例如，图 2.1 即为一棵决策树，它将整个样本空间分为三类。如果一个样本属性 A 的取值为 a2,属性 B 的取值为 b2，属性 C 的取值为 c1 那么属于类 1。

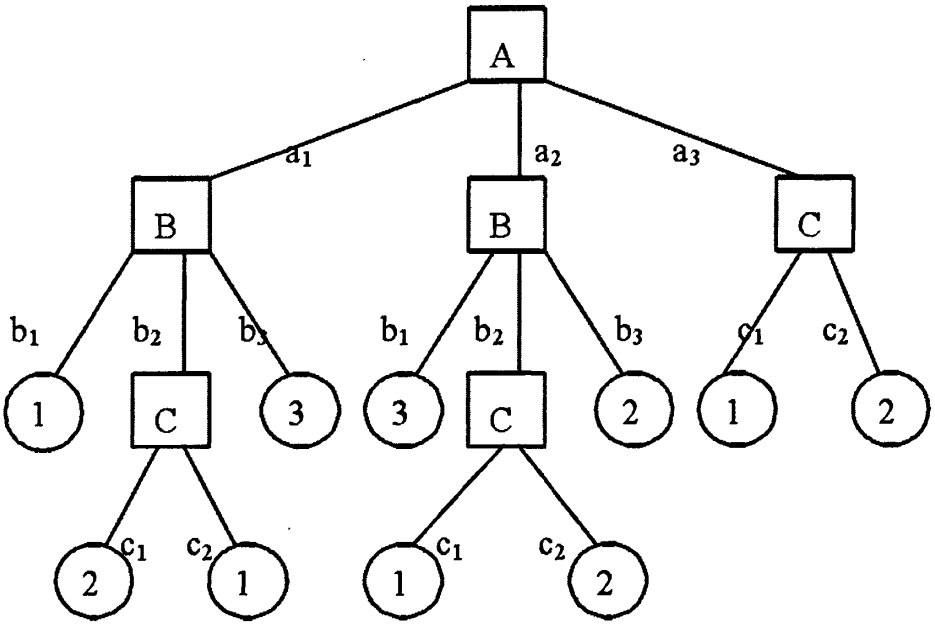

图 3-1 一个决策树的例子

决策树的构造是一个贪婪的，从上到下，各个击破的递归过程。决策树构造算法的关键在于属性评价标准的确定和划分结点所采用的形式。参考文献[22]中讨论了不同的属性评价标准，指出属性评价标准大致可以分为三类：①基于信息论，②基于距离，③基于独立性。没有哪种属性评价标准明显优于其他的标准。按照划分结点所采用的形式，决策树可以分为两类，基于单变量划分的决策树和基于多变量划分的决策树。下面，笔者与其中一种决策树算法 ID3 为例说明决策树算法的思路。大部分优秀算法，如 C4.5, J48 都是以 ID3 算法为基础的。

在决策树学习算法的各种算法中最有影响的是 Quinlan 提出的 ID3 算法[3]。

ID3 算法概述如下：

算法 3-1：Generate_decision_tree 由给定的训练数据产生一棵决策树

输入：训练样本 samples，由离散值属性表示；候选属性的集合 attribute_list。

输出：一棵决策树。

算法：

（1）创建结点 S；

（2）if samples 都在同一个类 C

then 返回 S 作为叶结点，以类 C 标记；

（3）if attribute_list 为空

then 返回 S 作为叶结点，标记为 samples 中最普通的类；

（4） 选择 attribute_list 中具有最高信息增益的属性 test_attribute；

（5） 标记结点 S 为 test_attribute；

（6）for each test_attribute 中的已知值 ai；

（7） 由结点 S 长出一个条件为 test_attribute=ai 的分支；

（8） 设 si 是 samples 中 test_attribute=ai 的样本的集合；

（9） if si 为空

then 加上一个树叶，标记为 samples 中最普通的类；

（10）else 加上一个由 Generate_decision_tree(si，attribute_list－test_attribute) 返回的结点；

ID3 算法使用信息增益作为属性评价标准。设结点 S 中有 s 个训练样本，有 m 个不同的类别 Ci(i=1,……,m),设 si 是 Ci 中的样本数。对于一个给定的样本分类所需的期望信息的计算公式为：

$$I(s_1,s_2,s_3,\cdots s_m)=-\sum_{i=1}^{m}p_i\log_2(p_i) \tag{3.6}$$

其中 p_i 是任意样本属于 C_i 的概率，用 $\frac{s_i}{s}$ 估计。

设属性 A 有 v 个不同的值 $\{a_1,a_2,\cdots,a_v\}$。用属性 A 将 S 划分为 v 个孩子结点 $\{S_1,S_2,\cdots S_v\}$，S_j 由属性 A 的值为 a_j 的样本构成。设 s_{ij} 为 S_j 中类别为 C_i 的样本数。根据由 A 划分成子集的期望信息的计算公式为·

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I(s_{1j},s_{2j},\cdots,s_{mj}) \tag{3.7}$$

那么在属性 A 上进行分枝所获得的信息增益为：

$$G(A)=I(s_1,s_2,\cdots,s_m)-E(A) \tag{3.8}$$

ID3 算法的优点是：算法理论清晰，方法简单，学习能力强。其缺点是：只对较小的数据集有效，且对噪声比较敏感，当训练数据集加大时，决策树可能随时改变。

2.3.3 支持向量机

支持向量机（SVM）[19]建立在计算学习理论的结构风险最小化原则之上。其主要思想是针对两分类问题，首先用内积函数定义的非线性变换将样本空间变换成一个高维空间，然后在高维空间上寻找一个（广义）最优分类面作为两类的分割，以保证最小的分类错误率。支持向量机的分类函数形式上类似于一个神经网络，输出是中间节点的线性组合，每个中间节点对应一个支持向量。

对于数据挖掘而言，需要处理的是海量数据，SVM 的传统算法显然不适合，需要采用其他更优化的算法。目前，研究人员已经提出了几种更有效的算法。Osuna 等人提出了一种分解算法，该算法首先选择一个小的工作集，在这个工作集上进行优化，然后从工作集中移走一个样本，加入一个不满足 Kuhn-Tucker 条件的样本，再进行优化，重复进行。因为工作集相对较小，所以每一步都很容易得到 QP 问题的最优解。而且，在同一时间，算法的运行只需要有限的内存消耗。在 Osuna 提出的分解方法的基础上，通过适当的选择工作集进行优化，Joachims 提出了 SVMlight 算法，该方法能够在同一时间处理一定数量的样本，从而也能有效的解决内存消耗。Platt 提出的 SMO 算法，其算法的特点是能够在同一时间处理两个样本例子；此外，还有 OlviL.Mangasarian 等提出的 LSVM、ASVM、SOR、SSVM 等[5]。

2.4 分类结果的评价

分类算法的评价标准包括：

1、预测的准确率：这涉及到模型正确地预测新的或先前没见过的数据的类标号能力。保持和 k-折交叉确认是两种常用的评估分类准确率的方法，在保持方法中，给定数据随机的划分成两个独立的集合：训练集和测试集。通常，三分之二的数据分配到训练集，其余三分之一分配到测试集。使用训练集导出分类法，其准确率用测试集评估。在 k-折交叉确认中，初始数据被划分为 k 个互不相交的子集或“折” S1，S2，……，Sk，每个折的大小大致相等，训练和测试进行 k 次，在第 i 次迭代，Si 用作测试集，其余的子集都用于训练分类法。准确率估计是 k 次迭代正确分类数除以初始数据中的样本总数。评价分类准确率的其它的方法包括引导和留一。

2、速度：涉及到产生和使用模型的计算花费。

3、强壮性：这涉及给定噪声数据或具有空缺值的数据，模型正确预测的能力。

4、可伸缩性：这涉及给定大量的数据，有效的构造模型的能力。

5、可解释性：这涉及学习模型提供的理解和洞察的层次。

2.5 本章小结

本章先介绍了数据挖掘的相关概念与其功能，重点介绍了关联分析，数据分类，数据聚类，偏差检测，时序分析等等。另外，本章也介绍了在实际应用中数据挖掘的相关数据结构及常用的分析方法。

接着，本章重点介绍了数据挖掘中的分类方法。分类也叫做有监督学习，它是数据挖掘研究的一个重点和热点。本章分析和比较了主要的分类算法，如贝叶斯方法，决策树方法及支持向量机方法等。最后还介绍了对分类算法的衡量标准。

第3章 数据挖掘中的主动学习及迁移学习

在数据挖掘中，传统的分类算法（也叫做“有监督的学习”）建立在一个严格的假设下：训练样本与测试样本必须独立同分布。但是，在实际的应用中，满足这一条件的训练样本往往相当缺乏。例如在生物信息学领域，获得一个训练样本往往需要大量的，昂贵的实验。面对这种困境，人们想了两类方法来解决训练样本的缺乏。第一种方法称为主动学习（Active Learning）。在主动学习中，分类器能够主动地选择包含信息量大的未标注样例并将其交由专家进行标注，然后置入训练集进行训练，从而在训练集较小的情况下获得较高的分类正确率，这样可以有效地降低构建高性能分类器的代价。第二种方法称为迁移学习（Transfer Learning）。其基本思路是借助来自其他领域的训练样本帮助本领域分类器的训练，以解决本领域训练样本不足的问题。但是，在迁移学习中，一个核心的问题是来自其他领域的训练样本与本领域的样本不满足独立同分布的条件。于是，迁移学习的核心是解决数据非独立同分布的问题。下面，笔者将分别对主动学习和迁移学习的主要算法进行简单的概述，并对这两种解决训练样本不足的方法进行分析与评价。

3.1 主动学习的概述

主动学习方法主要的思路是从大量的未标签的数据样本中，选择出少量的信息量大的数据（样例），由领域专家以一定得耗费进行标注。于是，根据主动学习中选择数据样例的策略的不同，可以分为以下两种方法：基于池的样例选择算法和基于流的样例选择算法。以下分别介绍。

3.1.1 基于池的样例选择算法

基于池的样例选择算法是当前研究得最为充分的。本文也采用了基于池的样例选择算法的相关技术。按照选择的标准不同，该类算法可以分为以下几类：

1、 基于不确定度缩减的方法

这类方法选择那些当前基准分类器最不能确定其分类的样例进行标注。这种方法以信息熵作为衡量样例所含信息量大小的度量，而信息熵最大的样例正是当前分类器最不能确定其分类的样例。从几何角度看,这种方法优先选择靠近分类边界的样例，所以又可以称为最近边界方法。这种方法可以应用于任何形式的基准学习器，如 logistic regression [23]、隐马尔可夫模型[24]、支撑向量机[25]以及归纳逻辑编程[26]等。它在大多数问题上能取得比随机选择更好的性能，但有可能采集到孤立点。

2、 基于版本空间缩减的方法

这类方法选择那些训练后能够最大程度缩减版本空间的样例进行标注。在二值分类问题中，这类方法选择的样例总是差不多平分版本空间，其思想来源于二分搜索。这包括 QBC[27], QBoost[28]和 Active Decorate[29]等。

QBC 算法[27]从版本空间中随机选择若干假设构成一个委员会，然后选择委员会中的假设预测分歧最大的样例进行标注。评价分歧度有如下标准:投票熵、Jensen-Shannon 分歧度[30]、Kullback-Leibler 分歧度[31]等。Fruend 等人[27]给出了 QBC 方法的严格理论证明。为了优化委员会的构成,增强其多样性，QBoost[28]和 Active Decorate[29]算法分别采用 Bagging,AdaBoost 和 Decorate 等成熟的分类器集成算法从版本空间中产生委员会。

3、 基于泛化误差缩减的方法

这类方法试图选择那些能够使未来泛化误差最大程度减小的样例。其一般过程为：首先选择一种损失函数用于估计未来错误率，然后将未标注样例集中的每一个样例都作为下一个可能的选择，分别估计其能给基准分类器带来的误差缩减，选择估计误差缩减最大的那个样例进行标注。当前针对不同的基准分类器提出相应的算法，如朴素贝叶斯[32]、贝叶斯网络[33]、最近邻算法[34]等。这种方法直接针对分类器性能的最终评价指标，理论上具有很好的效果，但计算量较大，同时损失函数的精度对性能的影响也至关重要。

4、其他方法

包括各种难以归入以上分类的主动学习算法，包括 COMB[35]、多视图(view)主动学习[36]、预聚类主动学习[37]等。COMB 算法[35]组合 3 种不同的主动学习器，迅速切换到当前性能最好的学习器从而使选择样例尽可能高效。多视图主动学习

[36]用于学习问题为多视图学习的情况，选择那些使不同视图的预测分类不一致的样例进行学习。其中，视图是指样例中足够做出分类判断的特征集合。这种方法对处理高维的主动学习问题非常有效,但不适用于低维问题。预聚类主动学习[37]认为基于不确定度缩减的方法会忽略样例的先验分布，而样例的分布恰恰有可能蕴涵丰富的信息。因此，首先运行聚类算法，然后选择样例时优先选取最靠近分类边界的样例和最能代表聚类的样例（即聚类中心）。

3.1.2 基于流的样例选择算法

基于池的算法大多可以通过调整以适应基于流的情况。但由于基于流的算法不能对未标注样例逐一比较，需要对样例的相应评价指标设定阈值，当提交给选择引擎的样例评价指标超过阈值，则进行标注。但这种方法需要针对不同的任务调整阈值，所以难以作为一种成熟的方法投入使用。

QBC 算法[27]也曾用于解决基于流的主动学习问题。样例以流的形式连续提交给选择引擎,选择引擎选择那些委员会(此处委员会只由两个成员分类器组成)中的成员分类器预测不一致的样例进行标注。

不同于 Freund 等人[27]的工作，Cesa-Bianchi 等人[38]针对的问题为:流中的样例独立同分布(不要求均匀分布)地取自 Rd 上的单位球面，其标注由一个二值线性概率函数生成且线性系数未知。他提出的算法维持一个对当前训练集的最小平方估计，该估计随着算法迭代的进行而逐步更新，当新样例自流中提交给选择引擎时，计算对新样例的最小平方估计的边缘值，当该边缘值小于一个阈值时，标注该样例，而该阈值随着算法迭代的进行逐步调整。Cesa-Bianchi 等人[38]证明，随着提交给选择引擎的样例增多，该算法需要标注的样例呈对数级地减少。

3.2 迁移学习的概述

传统的数据挖掘或机器学习的有监督学习方法建立在一个严格的假设下：训练样本与测试样本满足独立同分布。但是，在实际问题中，人们往往无法得到足够的满足要求的训练样本。为了解决这个问题，上一小节介绍了主动学习的思想。主动学习旨在选择一小部分具有代表性，或信息量大的数据给专家去标注以获得

类别标签。如果数据含有足够的信息量，则少量的花费可以换来分类精度的提高。

另一个方法称为迁移学习。在迁移学习中，人们可以借助来自其他领域，其他分布的训练样本来构建当前的分类器。于是，在迁移学习中，数据不一定满足独立同分布的要求。因此大多数传统的分类器在这种情况下的运行效果均不理想。近两年来，关于迁移学习的课题在数据挖掘的各个国际顶级会议（如 KDD, ICML, SDM, ECML 等）上得到了广泛的关注。下面笔者先介绍迁移学习里面的基本概念。

3.2.1 迁移学习的基本概念与定义

在数据挖掘和机器学习领域，迁移学习是近年兴起的研究问题。笔者在这一小节先对迁移学习的基本概念和定义进行介绍。

首先，图 4-1 是迁移学习的总体框架[11]。

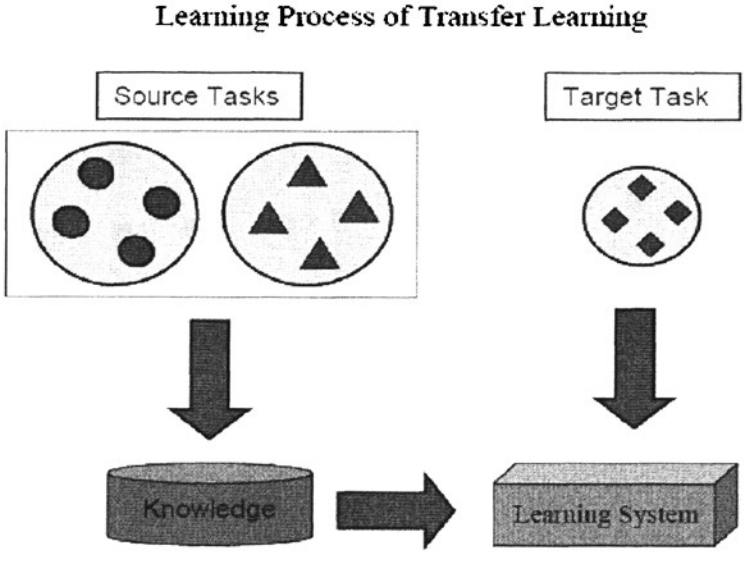

图 4-1 迁移学习的框架

在这里，有几个概念需要介绍：

（1）目标领域，或叫目标任务（Target Task or Target Domain）。是指当前要进行分类操作的数据集。目标领域包含待分类的数据，而迁移学习的目标就是最小化目标领域里数据的分类错误率。

（2）源领域，或叫源任务（Source Tasks or Source Domains）。是指其他的与目标领域不同的数据集。例如，目标领域是关于体育类和政治类的新闻文档；源

领域可能是关于娱乐和教育类的文档，也可能是关于军事和农业的文档。而迁移学习的作用就是要利用这些看似不同的来自源领域的数据集来帮助目标领域数据的学习（分类）。

（3）知识迁移。图4-1中的知识可以有不同的理解。例如，知识可以是部分的训练样本，这就延伸出基于采样的知识迁移方法；知识还可以是一个好的特征空间，这也延伸出基于特征的知识迁移方法；知识还可以是训练的模型，于是有了基于训练参数的知识迁移方法。

（4）迁移学习的目标：与传统的机器学习目标相同，经验风险最小化：

$$\theta^* = \arg\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\in P}[l(x,y,\theta)]$$

其中θ^*是训练模型的参数，L是分类的损失函数[20]。但迁移学习在优化经验风险时会利用来自源领域的知识。

（5）负迁移。由于源领域的数据与目标领域的数据的分布可能差异很大，因此某些迁移学习也许会使得分类的结果变坏，即分类的准确率下降。具体而言，如下面的公式所示：

$$\varepsilon[\mathcal{N}(\mathcal{U})] < \varepsilon[\mathcal{T}(\mathcal{U})]$$

该公式的具体意义是不采用迁移学习的分类错误率比采用迁移学习后的分类错误率还要小。在这种情况下，迁移学习带来了不好的效果（Negative result）。文献[14]把这种现象称为负迁移。

总的来说，关于迁移学习的算法可以分为两大类：一类为假设当前领域已经包含少量的独立同分布的训练样本，文献[11]把这类算法称为归纳式迁移学习（Inductive Transfer Learning）；另外一类为假设目标领域不包含任何训练样本，文献[11]把这类算法称为推导式迁移学习（Transductive Transfer Learning）。下面，分别对这两类迁移学习方法进行简要的介绍。

3.2.2 归纳式迁移学习

在归纳式迁移学习中，目标领域（或本领域）已经包含有部分的独立同分布的训练样本。于是，在该类方法中，人们可以利用这一小部分的独立同分布的训练样本，从其他领域中选取有用的知识以进行迁移。根据迁移的知识分类，又可以分为以下几种方法：

1、基于采样的归纳式迁移。

这类方法的主要思想是从其他领域的数据样本中进行采样，以使得采样后的数据集与目标领域的数据的分布相似。其中一个典型的方法称为 Tradaboost[16]。该方法是在 Adaboost[21]的基础上，加入知识迁移的因素。具体的思路是把数据集分成两部分来看待：一部分数据来自本领域，这部分数据的采样权值的变化与传统的 Adaboost 一样，即被错分的样本，其权值变大；另一部分的数据来自其他领域，这部分数据的采样权值的变化则是，被错分的样本，其权值变小。Tradaboost 的具体算法如下所示：

算法 4-1： Tradaboost

输入：目标领域的训练数据集 Td，目标领域的测试数据 S，源领域数据集 Ts，初始的权值 W，最大循环次数 N，一个分类器 Learner。

输出：各个测试数据的分类结果及置信度。

算法

For t=1:N

（1） $\mathbf{p}^t = \mathbf{w}^t / (\sum_{i=1}^{n+m} w_i^t)$;

（2） 对样本数据集按照分布 p 进行采样，并利用采样后的数据集构建分类器；

（3） 计算错误率

$$\epsilon_t = \sum_{i=n+1}^{n+m} \frac{w_i^t \cdot |h_t(x_i) - c(x_i)|}{\sum_{i=n+1}^{n+m} w_i^t}$$

（4）令 $\beta_t = \epsilon_t/(1-\epsilon_t)$; $\beta = 1/(1+\sqrt{2\ln n/N})$

（5）跟新采样权值：

$$w_i^{t+1} = \begin{cases} w_i^t \beta^{|h_t(x_i)-c(x_i)|}, & 1 \le i \le n \\ w_i^t \beta_t^{-|h_t(x_i)-c(x_i)|}, & n+1 \le i \le n+m \end{cases}$$

End for;

（6）输出分类的可信度：

$$h_f(x) = \begin{cases} 1, & \prod_{t=\lceil N/2 \rceil}^{N} \beta_t^{-h_t(x)} \ge \prod_{t=\lceil N/2 \rceil}^{N} \beta_t^{-\frac{1}{2}} \\ 0, & \text{otherwise} \end{cases}$$

同类的方法还有文献[39]。这类方法的好处是可以借助已有的样本，对其他领域的样本进行选择，以达到提高分类准确率的目的。

2、基于特征的归纳式迁移。

基于特征的迁移方法大概有两种思路。一为对全部数据集进行特征选择，以使得目标领域和来自其他领域的数据在特征子空间中分布足够相似，类别分离性足够强。另外一类方法是把全部的数据映射到一个新的空间，例如核空间[40]，spectrum 空间[17]等等。同样，这类方法的目的是找到一个新的映射空间使得不同的数据集在这个空间里面拥有很好的类别区分性。例如，其中一个典型的方法是优化以下的目标[11]：

$$\begin{aligned} \arg\min_{A,U} \quad & \varepsilon(A,U)=\sum_{t\in\{T,S\}}\sum_{i=1}^{n_t} L(y_{t_i},\langle a_t, U^T x_{t_i}\rangle)+\gamma\|A\|_{2,1}^2 \\ s.t. \quad & U\in \mathbf{O}^d \end{aligned}$$

其中 S 和 T 表达源领域和目标领域。A 是一个特征变换的参数矩阵，U 是一个坐标基向量，L 是损失函数。该表达式同时优化了三个变量：目标领域和源领域的映射$U^T X_T, U^T X_S$，参数矩阵 A。这个优化函数可以被化为凸函数来进一步求解。

3、基于参数的归纳式迁移。

大部分归纳迁移算法都假设来自不同领域的分类器会有相类似的模型参数。例如，这些参数满足某种分布或者共享某些隐含参数（Latent parameters）。例如，文献[41]对支持向量机进行改进，使得不同领域的 SVM 具有相似的参数。其优化目标如下：

$$\begin{aligned} \min_{w_0,v_t,\xi_{t_i}} \quad & \left\{J(w_0,v_t,\xi_{t_i})=\sum_{t\in\{S,T\}}\sum_{i=1}^{n_t}\xi_{t_i}+\frac{\lambda_1}{2}\sum_{t\in\{S,T\}}\|v_t\|^2+\lambda_2\|w_0\|^2\right\} \\ s.t. \quad & y_{t_i}(w_0+v_t)\cdot x_{t_i}\geq 1-\xi_{t_i},\ \xi_{t_i}\geq 0,\ i\in\{1,2,...,n_t\}\ and\ t\in\{S,T\}. \end{aligned}$$

其中，v 是参数 w 的一个偏移量。该方法是直接对 SVM 的优化目标进行改变，以使得来自不同领域的 SVM 的参数向量很相似。

3.2.3 推导式迁移学习

在推导式迁移学习（Transductive Transfer Learning）框架中，目标领域没有

任何训练样本，因此，知识的迁移会更加困难。与归纳式迁移一样，推导式迁移也分为基于采样，基于特征两类方法。由于目标领域不含有任何训练样本，因此目标领域无法建立本领域的训练模型，于是推导式迁移学习没有基于训练参数的迁移方法。下面，笔者对基于采样和基于特征的知识迁移方法分别进行介绍。

1、基于采样的知识迁移。

在传统的数据挖掘和机器学习模型中，大部分分类方法旨在最小化损失，即优化以下函数：

$$\theta^* = \arg\min_{\theta\in\Theta} \mathbb{E}_{(x,y)\in\mathcal{P}}[l(x,y,\theta)]$$

但是，在推导式迁移学习中，由于在目标领域没有训练样本，无法直接优上面的目标函数。于是，文献[42]对优化函数进行了变换，利用源领域的训练样本来估计目标领域的损失，目标函数变为：

$$\begin{aligned}\theta^* &= \arg\min_{\theta\in\Theta} \sum_{(x,y)\in\mathcal{D}_S} \frac{\mathcal{P}(\mathcal{D}_T)}{\mathcal{P}(\mathcal{D}_S)}\mathcal{P}(\mathcal{D}_S)l(x,y.\theta) \\ &\approx \arg\min_{\theta\in\Theta} \sum_{i=1}^{n_S} \frac{\mathcal{P}_T(x_{T_i},y_{T_i})}{\mathcal{P}_S(x_{S_i},y_{S_i})} l(x_{S_i},y_{S_i},\theta).\end{aligned}$$

其中，T 和 S 分别代表目标和源领域。于是，只需要对每个源领域的训练样本计算权值$\frac{\mathcal{P}_T(x_{T_i},y_{T_i})}{\mathcal{P}_S(x_{S_i},y_{S_i})}$，即可以用源领域的训练样本的损失来估计目标领域的分类损失。而为了计算训练样本的权值，往往有不同的方法，如文献[43]结合了特征映射的方法，先找到一个特征核空间，再在核空间中计算样本的权值；文献[42]则假设目标空间和源空间的后验概率 P(y|x)是相同的，并结合采样的方法生成新的数据集和构建训练模型。

2、基于特征的知识迁移。

基于特征的知识迁移方法假设不同领域的特征存在相似性，因此，这类方法通常只适用于特定的问题。例如，文献[12]提出了文本挖掘中的知识迁移方法。它利用了词在各类文档中的相似性，用共聚类（Co-clustering）的方法进行了知识的迁移。该方法主要的框架如图 4-2 所示。

该方法主要利用词的特性作为桥梁，建立了不同领域（Di 和 Do）的相似性，并通过词的聚类进行类别标签的迁移。文献[12]也用实验证明该方法对文本挖掘

具有良好的适用性，可大幅度提高分类的准确率。

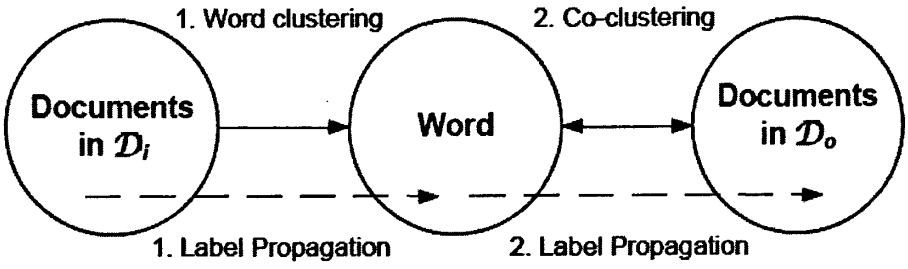

图 4-2 本文挖掘中的迁移学习

3.3 本章小结

本章介绍了主动学习和迁移学习的基本概念与典型方法。主动学习和迁移学习是解决训练样本不足的两类方法。其中主动学习通过少量的花费以获得更多的训练样本；迁移学习则通过从其他领域（源领域）中获得更多的训练样本。对主动学习的当前研究主要分为基于池和基于流的样本选择方法；而对迁移学习的研究主要分为归纳式和推导式方法。下一章节，笔者将分析主动学习和迁移学习在解决样本不足问题上的缺陷，并提出主动迁移学习的框架。

第4章 主动迁移学习模型

传统的有监督学习需要大量的满足独立同分布的训练样本。然而，在实际应用中，获得训练样本往往相当的昂贵。上一章节介绍了两个不同的方法来解决训练样本不足的问题：主动学习和迁移学习。主动学习的研究重点是如何选择少量的具有足够信息量的数据作为训练样本，以使得获得训练样本的总代价降低；迁移学习的研究重点是如何借助来自其他领域（源领域）的训练样本，来帮助目标领域的模型的建立，因此降低了对本领域训练样本的需求。但是，主动学习与迁移学习均有各自的缺陷，以使得其仍然无法很好地运用到实际应用中。下面，笔者将对它们的缺陷进行分析，并提出主动迁移的框架以对这两种分开的方法进行改善。

4.1 主动迁移学习的研究背景

主动学习与迁移学习是解决训练样本不足的两种不同的策略。但是，它们各自存在缺陷。

4.1.1 主动学习的缺陷

主动学习的目的是从数据源选取少量的数据作为训练样本，以从总体上降低获得类别标签的代价。总的来说，主动学习的框架如图 5-1 所示。在图 5-1 中，主动学习的目的是要对本领域（In-domain），或者称为“目标领域”的数据集进行学习（分类）。由于缺乏训练样本，主动学习的学习器（Learner）会从本领域的数据集中选择出少量的，具有代表性的数据，并把这些数据交给领域的专家（Domain Experts）。领域专家便对这些数据进行类别的标注，以使得这部分数据成为训练样本。但是，聘请领域专家对数据进行标注还是得有代价。例如，在文本挖掘领域，著名的网络公司如 yahoo!，Google 等会花大量的资金聘请专家对网页的类别进行标注；在生物信息学领域，获得一个数据的类别标签更需要领域专家进行大量的长期的实验。因此，主动学习的一个明显的缺陷是它还是需

要有获得训练数据的代价。

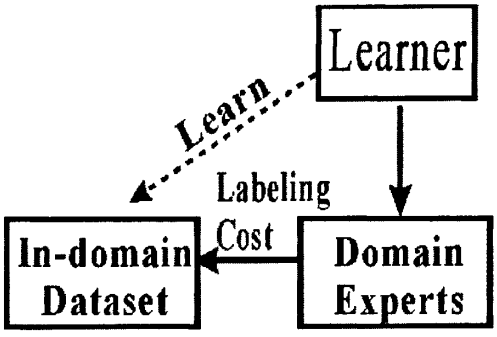

图 5-1 主动学习的框架

图 5-2 是一个主动学习在文本挖掘中的例子。如图所示，250 个训练样本所能达到的分类准确率只有 64%。因此，如果要吧分类的准确率提高到 80%，从图 5-2 中估计，需要接近 500 个训练样本。因此，主动学习仍然对训练样本的需求很大。在某些领域，如生物学，训练样本缺乏的问题仍然得不到很好的解决。

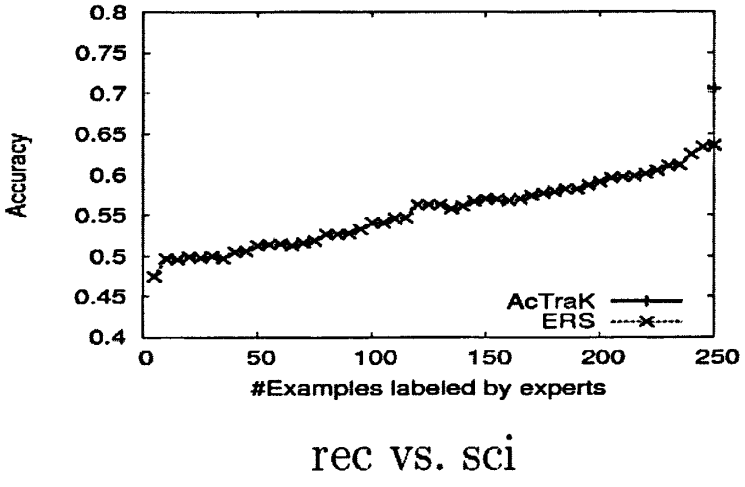

图 5-2 文本挖掘中的主动学习

4.1.2 迁移学习的缺陷

迁移学习是解决训练样本缺乏的另外一种策略。其算法框架如图5-3所示。迁移学习与传统的有监督学习的目标相同，要尽量提高对本领域（In-domain）的学习的精度。由于缺乏训练样本，迁移学习可以从源领域或者其他领域（Out-of-domain）里面借助训练样本，以帮助本领域的学习。

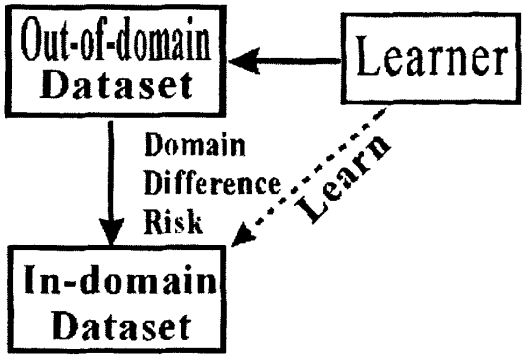

图 5-3 迁移学习的框架

但是，迁移学习有一个明显的缺点：当本领域（In-domain）与其他领域（Out-of-domain）的数据集差异很大时，知识的迁移可能大幅度降低学习的精度，发生“负迁移”现象[14]。在这里，笔者把这种缺陷称为“领域相异性风险”（Domain Difference Risk）。

如上所述，虽然主动学习与迁移学习是解决训练样本不足的两种相对有效的方法，但是主动学习对训练样本的需求依然很大，造成获得训练样本的代价仍然比较高，而迁移学习存在领域相异性风险，无法保证学习的效果。因此，无论主动学习还是迁移学习，均无法在实际应用中有效地解决分类方法中的训练样本缺乏的问题。本文将结合主动学习与迁移学习的思想，提出主动迁移学习模型，以更好地解决训练样本不足的问题。

4.2 主动迁移学习的具体框架

本小节，笔者将介绍主动迁移学习的模型。首先，笔者将介绍主动迁移学习的基本思想以及大致的框架；然后，笔者将进一步介绍迁移学习的理论基础；最后，笔者将用实验数据来说明迁移学习的可靠性。

4.2.1 基本思想

主动学习和迁移学习是解决训练样本缺乏的不同的策略。但是，如上以一小节所分析，主动学习和迁移学习仍然存在缺陷，以至于这两种策略无法很好地运用到实际应用中。

为了弥补这两种策略的缺陷，笔者提出主动迁移学习的策略。在这里，笔者先用日常生活中的主动迁移的例子来说明它的基本思想。在日常的学习中，如果我们有了数学的基础，我们可以更好地理解物理的知识。但是，并不是所有的物理知识都可以从数学中迁移过来。例如一般的数学课本不包含牛顿的三大定律。于是，为了学习这一部分的知识，我们还要去向老师请教。在这个过程中，我们既利用到了迁移学习，也利用到了主动学习的策略。重要的是，在这个过程中，由于我们主动地选择哪些知识可以从数学中迁移过来，哪些知识需要向老师请教，于是一方面，与迁移学习相比，我们有老师的指导，所以避免了“负迁移”；另一方面，与主动学习相比，部分知识可以从数学中迁移过来而不必请教老师，于是我们降低了学习的代价。这个就是日常生活中人的主动迁移学习策略。

我们也可以用图5-4所示的框架来描述主动迁移学习策略。如图所示，为了获得某个数据的类别标签，我们应该先利用迁移学习对该数据进行评估。如果迁移学习能很好地判断该数据的类别，则我们可以直接地给数据标号。但是如果迁移学习不能很好地分类，我们就把该数据交给领域的专家，由专家对数据进行标注。该数据获得标签后，变成为了训练数据。在这个过程中，与传统的主动学习相比，部分训练数据的获得不需要经过领域专家，于是进一步降低了类别标签获得的代价；与传统的迁移学习相比，我们可以把不确定的数据交给领域专家，以避免“负迁移”。

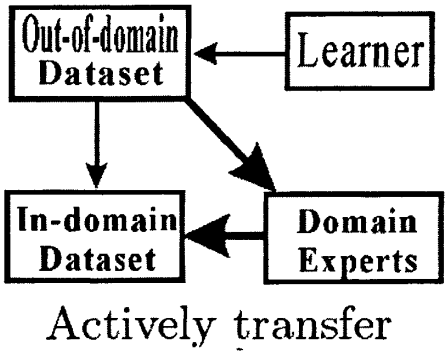

图 5-4 主动迁移学习的框架

4.2.2 主要框架

上一小节介绍了主动迁移学习的思想，这一小节将介绍主动迁移学习的算法框架。

算法5-1：主动迁移学习框架AcTraK: ACtively TRAnsfer Knowledge

输入：目标领域数据集U；源领域数据集L；给专家标注的最大样本数N

输出：分类器*l*

算法流程：

（1） n←0；

（2） Repeat

（3） X ← 用传统的主动学习器从U中选择一个样本点；

（4） 构建迁移学习器T；（将在下一小节中具体介绍）；

（5） 用迁移学习器T对X进行分类；

（6） 计算决策函数 F(x) （将在下面具体介绍）；

（7） If F(x)=0

（8） 用迁移学习器T给X加类别标签；

（9） Else

（10） 由领域专家给X加类别标签；

（11） n←n+1；

（12） End

（13） $L \leftarrow L \cup \{(X, y)\}$；

（14） Until n>N

（15） 以L为新的训练集，训练并输出分类器*l*；

如该算法所示，先由传统的主动学习器从无类别标签的数据集U中选择一个具有代表性的样本点X。然后给该样本点加上类别标签有两个途径：第一个途径为由迁移学习的分类器提供预测的类别标签，即先建立一个迁移学习分类器，并用该分类器对X的预测类别作为X的标签。另外一个途径就是把X交给领域专家，像传统的主动学习一样，由专家对样本点进行标注。这两种途径中，由主动学习进行标注的话，可以保证样本点的类别标签准确无误，但是这种标注会产生耗费；而由迁移学习进行标注的话，可以避免类别标注的耗费，但是却有类别标签错误的风险。因此，平衡这两种途径的核心是算法中的决策函数 F(x)。决策函数的

主要作用为，当迁移学习的分类器的分类结果可信度强时，由迁移学习分类器提供类别标签，否则，由领域专家提供类别标签。算法5-1的时间复杂度为$O(Nn^2)$，其中算法的第3步为主动学习中的样本选择子算法，通常其复杂度为$O(n^2)$。下面，笔者将分别对主动迁移学习框架中的分类器和算法中的决策函数进行介绍。

4.2.3 主动迁移学习中的迁移分类器

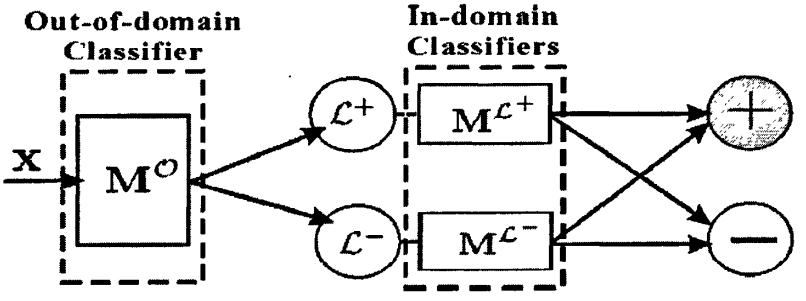

图 5-5 主动迁移框架中的分类器 T

在主动迁移学习的算法框架中，被选出来的数据点 X 先被迁移学习分类器 T 进行分类。而迁移学习分类器可借助来自其他领域的训练样本对 X 进行预测。在本小节，笔者将介绍一个简单的迁移分类器，并介绍该迁移分类器的良好的理论特性。

图 5-5 描述了迁移学习分类器的结构模型。其中 O 是来自源领域的数据集，L 是来自目标领域的数据集，而$\mathcal{L}^+ = \{\mathbf{x}|\mathbf{x} \in \mathcal{L} \wedge \mathbf{M}^O(\mathbf{x}) = \text{“+”}\}$ 且 $\mathcal{L}^- = \{\mathbf{x}|\mathbf{x} \in \mathcal{L} \wedge \mathbf{M}^O(\mathbf{x}) = \text{“-”}\}$，即$\mathcal{L}^+$是被分类器$M^O$分为正类别“+”的数据集；$\mathcal{L}^-$是被分类器$M^O$判断为负类别“-”的数据集。那么，主动迁移学习的分类器的算法流程如下所示：

算法 5-2：迁移学习分类器

输入：待分类数据 x，源领域数据 O，目标领域数据 L，普通的分类器 Learner

输出：x 的类别标签及其分类置信度。

算法流程：

（1）用来自源领域的数据O建立分类器M^O，并用M^O对数据X进行分类，得到两个概率$P(\mathcal{L}^+|\mathbf{x}, \mathbf{M}^{\mathcal{O}})$和$P(\mathcal{L}^-|\mathbf{x}, \mathbf{M}^{\mathcal{O}})$；

（2）用$\mathcal{L}^+$与$\mathcal{L}^-$分别建立分类器$\mathbf{M}^{\mathcal{L}+}$与$\mathbf{M}^{\mathcal{L}-}$，并分别用这两个分类器对X进行分类，得到概率$P(+|\mathbf{x}, \mathbf{M}^{\mathcal{L}^+})$和$P(+|\mathbf{x}, \mathbf{M}^{\mathcal{L}^-})$；

（3）数据X被分类器T判断为正例“+”的概率为：

$$P_T(+|\mathbf{x}) = P(\mathcal{L}^+|\mathbf{x}, \mathbf{M}^{\mathcal{O}}) \times P(+|\mathbf{x}, \mathbf{M}^{\mathcal{L}^+}) + P(\mathcal{L}^-|\mathbf{x}, \mathbf{M}^{\mathcal{O}}) \times P(+|\mathbf{x}, \mathbf{M}^{\mathcal{L}^-})$$

当忽略掉分类概率对分类器的明显的依赖性，上面的式子还可以化成一个简单的形式：

$$P_T(+|\mathbf{x}) = P(+|\mathcal{L}^+, \mathbf{x}) \times P(\mathcal{L}^+|\mathbf{x}) + P(+|\mathcal{L}^-, \mathbf{x}) \times P(\mathcal{L}^-|\mathbf{x}) \quad \text{（公式 5-1）}$$

于是，如果$P_T(+|\mathbf{x}) > 0.5$，X为正例，否则，X为负例。这个迁移分类器简单，而且有比较好的性质。

推论1 假设$\varepsilon 1$和$\varepsilon 2$分别是分类器$\mathbf{M}^{\mathcal{L}+}$和$\mathbf{M}^{\mathcal{L}-}$的错误率（Error rate），那么分类器T的错误率ε_t满足以下性质：

$$\min(\varepsilon_1, \varepsilon_2) \leq \varepsilon_t \leq \frac{1}{2}(\varepsilon_1 + \varepsilon_2)$$

证明：$\forall \mathbf{x} \in \mathcal{U}$，假设X的正确标签为“+”且$P(\mathcal{L}^+|\mathbf{x}) \geq P(\mathcal{L}^-|\mathbf{x})$，那么X被错分为“-”的概率为：

$$\begin{aligned}\varepsilon_t(\mathbf{x}) &= P(-|\mathbf{x}) \\ &= P(\mathcal{L}^-|\mathbf{x}) \times P(-|\mathcal{L}^-, \mathbf{x}) + (1 - P(\mathcal{L}^-|\mathbf{x})) \times P(-|\mathcal{L}^+, \mathbf{x}) \\ &= P(-|\mathcal{L}^+, \mathbf{x}) + P(\mathcal{L}^-|\mathbf{x}) \times (P(-|\mathcal{L}^-, \mathbf{x}) - P(-|\mathcal{L}^+, \mathbf{x}))\end{aligned}$$

且

$$P(-|\mathcal{L}^-, \mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{L}^-, -)P(\mathcal{L}^-, -)}{P(\mathcal{L}^-|\mathbf{x})P(\mathbf{x})} > \frac{P(\mathbf{x}|\mathcal{L}^+, -)P(\mathcal{L}^+, -)}{P(\mathcal{L}^+|\mathbf{x})P(\mathbf{x})} = P(-|\mathcal{L}^+, \mathbf{x})$$

那么，

$$P(-|\mathbf{x}) \geq P(-|\mathcal{L}^+, \mathbf{x}) = \min(P(-|\mathbf{x}, \mathbf{M}^{\mathcal{L}^+}), P(-|\mathbf{x}, \mathbf{M}^{\mathcal{L}^-}))$$

另外，因为$P(\mathcal{L}^+|\mathbf{x}) \geq P(\mathcal{L}^-|\mathbf{x})$，我们可以得到$0 \leq P(\mathcal{L}^-|\mathbf{x}) \leq \frac{1}{2}$。那么，

$P(-|\mathbf{x}) \leq$

$P(-|\mathbf{x}, \mathbf{M}^{\mathcal{L}^+}) + \frac{1}{2}(P(-|\mathbf{x}, \mathbf{M}^{\mathcal{L}^-}) - P(-|\mathbf{x}, \mathbf{M}^{\mathcal{L}^+}))$

$= \frac{1}{2}(P(-|\mathbf{x}. \mathbf{M}^{\mathcal{L}^+}) + P(-|\mathbf{x}, \mathbf{M}^{\mathcal{L}^-}))$，

于是，

$$\min(\varepsilon_1, \varepsilon_2) \leq \varepsilon_t \leq \frac{1}{2}(\varepsilon_1 + \varepsilon_2)$$

上面的推论证明了算法 5-2 中的分类器能较好地保证分类的准确率（或描述为降低分类的错误率）。有了这个迁移学习分类器 T，我们可以得到样本 X 的预测标签，我们下面再来介绍算法 5-1 中的决策函数 F（x）。

4.2.4 决策函数

上一小节介绍了算法 5-1 中的迁移学习分类器 T。这一小节，笔者将介绍算法 5-1 中的决策函数 F(x)来决定迁移学习分类器的分类结果的可信度是否足够高。首先，笔者设置的决策函数应有如下性质：如果决策函数 F(x)>0，那么迁移分类器的分类结果将被看做 X 的类别标签；否则，将由领域专家给出 X 的类别标签。于是决策函数 F(x)是主动学习框架中极为重要的一环。

首先，我们可以考虑有如下几种情况，我们不应该采用分类器 T 给出的分类结果：

（1）如果分类器 T 给出的分类结果与目标领域的分类器的分类结果不一致；

（2）如果分类器 T 的分类置信度很低；

（3）如果训练数据集相当的小。

在这些情况下，分类器 T 给出的分类结果都是不可信任的，而在这些情况下我们更应该由领域专家给数据样本 X 进行标注，以避免“负迁移”。假设目标领域的分类器为$\mathbf{M}^{\mathcal{L}}$，那么我们可以定义分类器 T 的分类风险评价函数$\theta(\mathbf{x})$为：

$$\theta(\mathbf{x}) = \Big(1 + \alpha(\mathbf{x})\Big)^{-1}$$

$$\alpha(\mathbf{x}) = \Big(1 - [\![\mathbf{M}^{\mathcal{L}}(\mathbf{x}) \neq \mathbf{T}(\mathbf{x})]\!]\Big) \cdot \mathbf{P}_T(\mathbf{T}(\mathbf{x}) = y|\mathbf{x}) \cdot \exp(-\frac{1}{|\mathcal{L}|})$$

其中当π为真时，$[\![\pi]\!] = 1$。分类风险评价$\theta(\mathbf{x})$的设置源于上述的三点考虑，因此分类风险的值越小，迁移学习器的分类结果越可靠。因此，决策函数 F(x)可以被定义为：

$$\mathbb{F}(\mathbf{x}) = \begin{cases} 0 & if\ R > \theta(\mathbf{x}) \\ 1 & otherwise \end{cases}$$

其中 R 是一个随机数。这个式子的意思是，以概率$\theta(\mathbf{x})$，F(x)的值为 0。

4.3 主动迁移学习的相关推论

上面的三个小节介绍了主动迁移学习框架 AcTraK（Actively Transfer Knowledge），并对 AcTraK 里面的分类器及决策函数做了较为详细的介绍。这个小节主要分析主动学习框架 AcTraK 的几个重要的性质。

推论 2：在算法 AcTraK 中，设ε_t为迁移学习分类器 T 的错误率，N 为给领域专家标注的样本的最大个数，那么，AcTrak 的错误率ε满足：

$$\varepsilon \leq \frac{\varepsilon_t^2}{1 + (1 - \varepsilon_t) \times \exp(-|N|^{-1})}$$

证明：

由算法流程容易得到$\varepsilon \leq (\varepsilon_t)^2(1-\theta(\mathbf{x}))$，那么，

$$\theta(\mathbf{x}) = \frac{1}{1+(1-\varepsilon_t)e^{-\frac{1}{|\mathcal{L}|}}} \geq \frac{1}{1+(1-\varepsilon_t)e^{-\frac{1}{N}}},$$

则

$$\varepsilon \leq$$

$$\varepsilon_t^2(1-\theta(\mathbf{x})) \leq \frac{\varepsilon_t^2 \times (1-\varepsilon_t) \times \exp(-|N|^{-1})}{1+(1-\varepsilon_t) \times \exp(-|N|^{-1})} \leq \frac{\varepsilon_t^2}{1+(1-\varepsilon_t) \times \exp(-|N|^{-1})}。$$

推论 3：在算法 AcTrak 中，设εt为迁移学习分类器 T 的错误率，ε_i为领域内分类器的错误率，令$\alpha = \varepsilon_t + \varepsilon_i$，那么，AcTrak 通过领域专家给数据样本标注的概率为:

$$P[Query] \leq \alpha + \frac{1-\alpha}{1+(1-\varepsilon_t)\times \exp(-\frac{1}{|N|})}$$

证明:

由算法流程，容易得到

$P[Query] = \varepsilon_i(1-\varepsilon_t)$

$_+\varepsilon_t(1-\varepsilon_i)+[\varepsilon_t\varepsilon_i+(1-\varepsilon_t)(1-\varepsilon_i)]\theta(\mathrm{x}) = \theta(\mathrm{x})+(\varepsilon_t+\varepsilon_i-2\varepsilon_t\varepsilon_i)(1-\theta(\mathrm{x}))\leq$

$\alpha+(1-\alpha)\theta(\mathbf{x}) \leq \alpha + \frac{1-\alpha}{1+(1-\varepsilon_t)\times\exp(-\frac{1}{|N|})}$。

推论 2 证明了 AcTrak 的错误率的上限，证明了 AcTrak 对降低负迁移风险的能力；而推论 3 证明了 AcTrak 与传统的主动学习相比，降低了获得训练样本的成本。下面，笔者将从实验上证明 AcTrak 的这两个优点。

4.4 实验研究

本节将用实验方法来证明主动迁移算法 AcTraK 的有效性。

4.4.1 实验数据的说明

为了明显地评估算法的有效性，本文在这一小节自动生成了几个数据集。并主要从下面两个方面评估 AcTraK:（1）与传统的迁移学习相比较，AcTraK 能否很好地降低负迁移的风险；（2）与传统的主动学习相比较，AcTraK 能否降低获得样本的成本。因此，笔者生成了如图 5-6 所示的五个数据集。

如图 5-6 所示，五个数据集均包含两个类别，在图上分别用正方形“□”与三角型“△”表示。而 U 为目标领域的数据集，它包含两个初始的训练样本，用菱形“◇”来显示。O1 到 O4 是四个源领域数据集。其中 O1 和目标领域数据

集 U 有比较相似的分布；而 O2 和 U 已经有比较明显的差异；O3 又叫做“异或”数据集，它跟目标领域数据集 U 的差异最大；O4 是个有趣的数据集，它和 U 的分布很相似，但是类别标签完全相反。那么，由观察猜想，传统的迁移学习在数据集 O2~O4 中均有可能产生负迁移而降低分类准确率。

另外，为了更好地评价算法 AcTrak，笔者在实验中使用了两个典型的对比算法，一个是主动学习算法 ERS[44]，一个是迁移学习算法 Tradaboost[16]。

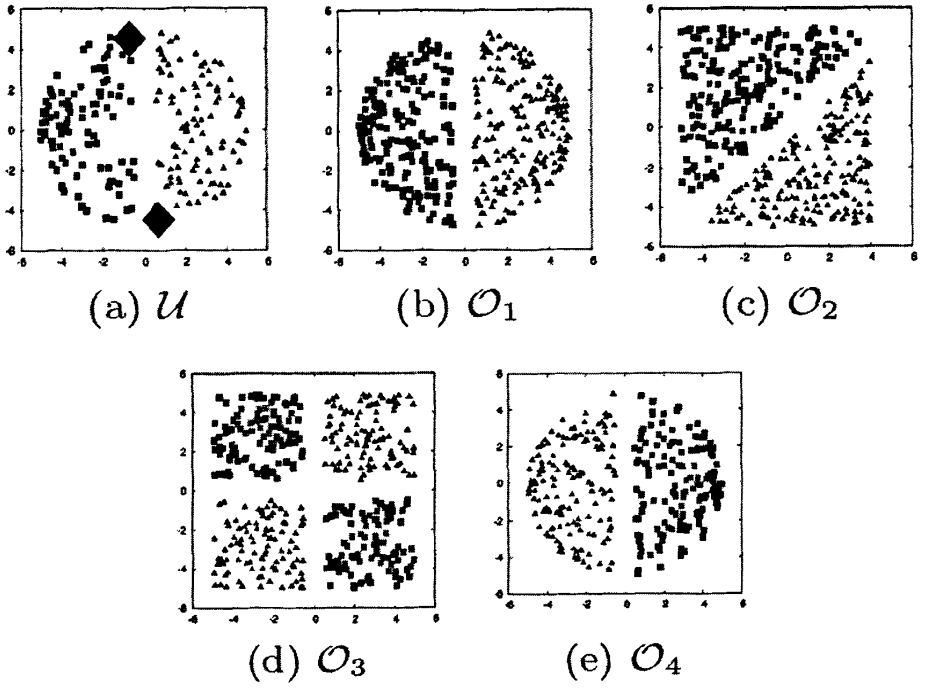

图 5-6 测试数据

4.4.2 实验机器的说明

本实验运行在安装有 Elipse 的 PC 机上，该 PC 机的基本配置如下：AMD2000+（1.6G）的 CPU，256M 的内存。所有算法用 JAVA 编写，所有实验均在 Elipse 上运行。

4.4.3 实验结果的检验方法

由于数据集都带有数据的分类标签，本文采用了分类准确性的计算方法评价

算法的结果。

在分类领域一般有三类命中率：正命中率、负命中率和总体命中率，详见公式（5-3-5-1）。

$$\begin{cases} r_p = num_p / Num_p \\ r_n = num_n / Num_n \\ r_a = (num_p + num_n)/(Num_p + Num_n) \end{cases}$$

其中r_p为正命中率；r_n为负命中率；r_a为总体命中率；num_p表示被分类器正确分类的正类样本数目；num_n表示被分类器正确分类的负类样本数目；Num_p表示数据集中正类样本总数；Num_n表示数据集中负类样本总数。在本实验中，采用正命中率r_p来评价算法的结果。

4.4.4 实验结果及分析

实验数据如图 5-6 所示，笔者用了四个不同的源领域数据 O1~O4，分别对目标领域数据 U 进行分类。实验有四组结果，笔者把四组结果都用图表示出来。其中每幅图的横坐标是被领域专家所标注的样本的个数，纵坐标是分类的正命中率（Accuracy）。那么，随着被标注的样本个数的增加，正命中率一般也会随着增加。但是，对比算法 Tradaboost 是一传统的迁移学习算法，它会完全地利用源领域的数据对目标领域的数据集进行分类。因此，图 5-7~图 5-10 中描述 Tradaboost 的结果都是直线。

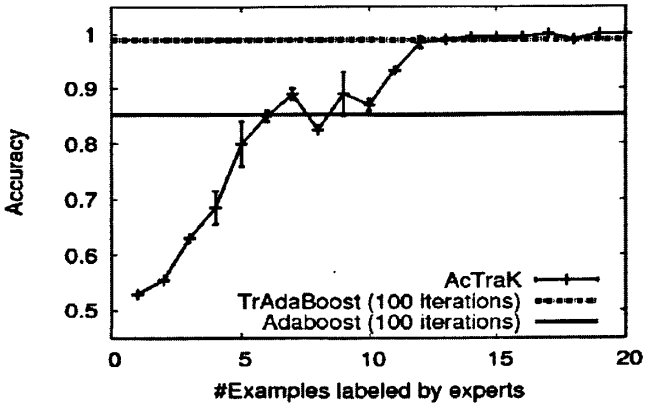

图 5-7 实验结果 1（源领域数据为 O_1）

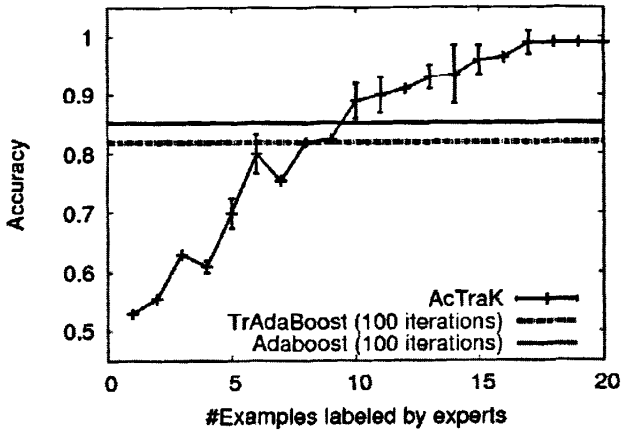

图 5-8 实验结果 2（源领域数据为 0_2）

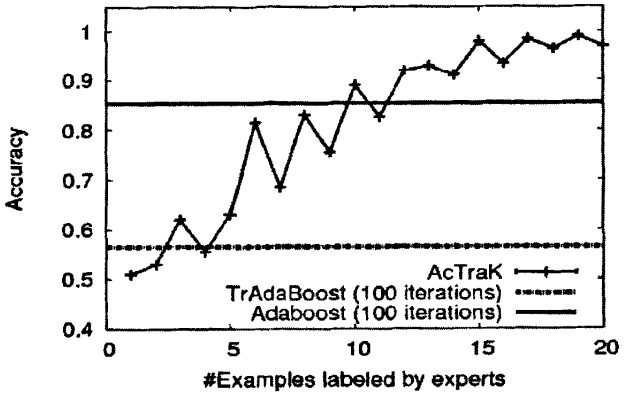

图 5-9 实验结果 3（源领域数据为 0_3）

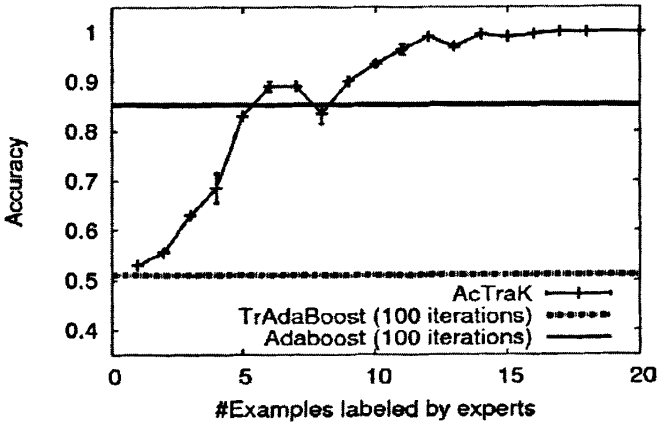

图 5-10 实验结果 4（源领域数据为 0_4）

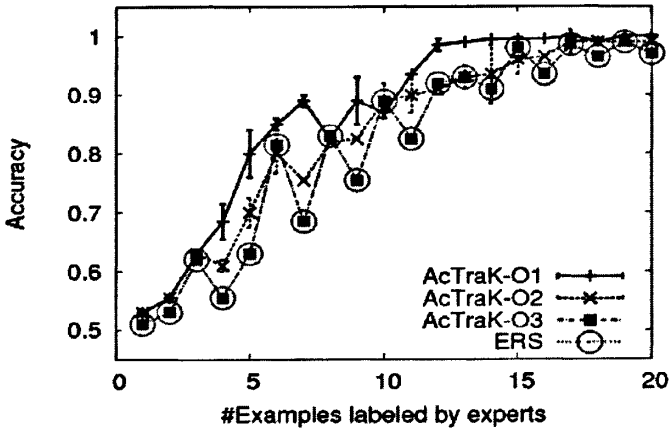

图 5-11 与主动学习器的实验比较

首先，图 5-7 描述了利用 O1 来对 U 就行分类的结果曲线。如图 5-6(a)所示，U 与 O1 有很相似的分布，因此，迁移学习在这个数据集中的效果比较好，Tradaboost 的准确率达到 99%。但是，当源领域数据集与目标领域数据集有差异时，迁移学习的效果开始变差。例如，当源领域数据集为 O2 时，Tradaboost 的准确率只有 84%；而当源领域数据集为 O3 时，它的准确率只有 57%；甚至当源领域数据集为 O4 时，虽然该数据集与目标数据有几乎相同的分布，但是由于它与目标领域数据集的类别标签相反，造成 Tradaboost 的准确率只有 50%左右。由这几个实验结果可以清楚地观察到负迁移对传统的迁移学习影响很大。

但是，同样是这样的数据集，我们提出的算法 AcTraK 却很好地保证了分类的精确度。事实上，每个实验结果的分类正命中率最后都超过了 90%。这充分说明了主动迁移学习可以有效地降低“负迁移”的风险。

另外，我们还把 AcTraK 与传统的主动学习器 ERS 进行了比较。结果如图 5-11 所示。从结果上看，当源数据集与目标数据集分布比较接近时，AcTraK 能用更少的资源达到更好的效果。例如，当源数据集是 O1 时，AcTraK 只用了 7 个专家标注的样本就达到了 90%的正命中率；而 ERS 需要用到 15 个这样的样本。因此，主动迁移学习与传统的主动学习相比较，前者能更好地降低获得成本的花费，达到更好地分类效果。

4.5 本章小结

本章首先介绍了主动学习与迁移学习的优缺点。迁移学习不需要获得样本的成本，但是它有“负迁移”的风险；主动学习没有“负迁移”的风险，但是它可能需要很大的耗费来获得训练样本。接着，本章提出并介绍了主动迁移学习框架。主要介绍了主动迁移学习的基本思想及算法。最后，本章从理论证明和实验证明的角度对主动迁移学习算法 AcTraK 进行了评价，并显示 AcTraK 能很好地降低“负迁移”的风险，同时降低获得成本的代价。

第5章　主动迁移学习在文本分类中的应用

文本分类是一个有着广泛应用领域的研究问题。目前流行的文本分类方法主要为传统的有监督学习，即通过大批量已有类别标签的训练数据建立分类器，并通过分类器对剩余的数据进行分类[3]。但是，由于文本挖掘的训练数据的获得相当困难，文献[16]在文本挖掘中引入迁移学习的技术，利用相近领域的训练样本帮助目标领域的文本的分类，并取得了一定得成效。但是，文献[16]依然假设源领域与目标领域有极为相近的数据分布。在本小节，笔者将展示在文本挖掘领域，迁移学习存在“负迁移”的风险，造成传统的迁移学习方法在文本挖掘的实际应用中并没有帮助分类准确率的提高。同时，笔者将用实验的方法证明主动迁移学习在文本挖掘中的优势。

5.1　文本分类的基本原理与方法

文本分类可以描述如下：对一个给定的文本集合$D=\{d_1, d_2, \cdots, d_n\}$及文本集合的类别$C=\{c_1, c_2, \cdots, c_k\}$，其中$c_i < D (i=1,2,\cdots,k)$，使得$\forall d_i$ $(d_i \in D)$，$\exists$ c_j $(c_j \in C)$ 有$d_i \in c_j$。并假设我们已知一个子集$D_l \subset D$，且对$\forall d_l \in D_l$，其类别标签已知。那么利用D_l对$D-D_l$进行分类，并同时使得错误率达到最小。

文本集合D的特征空间就是出现在该文本集合中的所有的词的集合，表示为$T_D=\{t \mid \forall t \in d, \quad \forall d \in D\}$。通常文本集合D都具有相当的规模，因此$T_D$是非常大的，这其中会包括这样的一些词，用 frequency (t) 表示词t出现在整个文本集合中的频度，则有：

$$\text{frequency}(t) \leqslant \varepsilon_1 \text{ 或 } \text{frequency}(t) \geqslant \varepsilon_2$$

ε_1 是接近0的一个值，ε_2 是接近1的一个值。也即是说，词 t 要么在D中极

少的文本中出现,要么在D中绝大多数的文本中出现, 这些词对文本的区分不会产生太大的作用, 反而会影响到分类的效果, 所以进行文本的特征选取, 找到能最好的区分文本的词的集合。本文使用 *tfidf* 作为特征选取度量。该度量方法是给每个文本中的每个词一个权重, 权重的计算[28]如下:

$$tfidf(d,t) = tf(d,t) \times \log\frac{n}{df(t)}$$

其中, $tf(d,t)$ 是词 t 在文本 d 中的词频, $df(t)$ 是D中包含词 t 的所有文本的数目, n是文本集合D的大小。在文本分类中, 通常使用向量空间模型来表示每个文本[28] , 在这种模型中, 每个文本 d 被认为是向量空间内的一个向量。在这里,我们用了 $tf-idf$ 词权重模型,则每个文本就被表示为:

$$< tf_1 \log(n/df_1),\ tf_2 \log(n/df_2),\ \cdots,\ tf_m \log(n/df_m) >$$

其中tf_i是文本中的第i个词的频度, df_i是包含第i个词的文本的数目。此外, 为了解决文本的长度不同的问题, 每个文本向量的模都被标准化为1。设:

$$d_{tfidf} = < tf_1 \log(n/df_1),\ tf_2 \log(n/df_2),\ \cdots,\ tf_m \log(n/df_m) >$$

则标准化之后,表示为:

$$d_{tfidf} = < \frac{tf_1 \log(n/df_1)}{\sqrt{(tf_1 \log(n/df_1))^2 + \ldots + \sqrt{(tf_n \log(n/df_n))^2}}}, \cdots,$$

$$\frac{tf_n \log(n/df_n)}{\sqrt{(tf_1 \log(n/df_1))^2 + \ldots + \sqrt{(tf_n \log(n/df_n))^2}}} >$$

于是有 $\| d_{tfidf} \| = 1$。

5.2 主动迁移学习在文本挖掘中的应用

本小节将介绍经过向量转化后，文本的数据如何进行迁移学习的实验。并把上一章节介绍的主动迁移框架运用到文本挖掘中。

5.2.1 特征选择与降维

定义词t在文本d_j中的频率记为f_j,则对词t来说,用下式度量t的质量[45]:

$$q(t)=\sum_{j\in A} f_j^2 - \frac{1}{n_1}\left[\sum_{j\in A} f_j\right]^2$$

其中$|A| = n_1$,词t至少出现过一次的文本组成了集合A中的元素,且$f_j > 1$。在得到了每个词t的$q(t)$之后,将词按照$q(t)$的大小排序,选择排在前$\alpha\times 100\%$的词作为特征词,从而实现降维。

5.2.2 算法流程

根据上面得到的文本数据，主动迁移学习的算法框架如下：

算法6-1：文本挖掘中的主动学习算法

输入：目标领域文本数据集U；源领域文本数据集L；给专家标注的最大文本数N

输出：分类器l

算法流程：

（1） 根据上两小节的讨论，把U与L的文本数据转化成数据向量；

（2） n←0；

（3） Repeat

（4） X ← 用传统的主动学习器从U中选择一个具有代表性的文本；

（5） 构建迁移学习器T；

（6） 用迁移学习器T对文本X进行分类；

（7） 计算决策函数 F(x) ；

（8） If F(x)=0

（9） 用迁移学习器T给X加文本标签；

（10） Else

（11） 由领域专家给X加文本标签；

（12） n←n+1；

（13） End

（14）　$L \leftarrow L \cup \{(X, y)\}$；

（15）　Until n>N

（16）　以L为新的训练集，训练并输出文本分类器l；

算法6-1的框架来源于上一章介绍的主动迁移学习框架AcTraK，不过算法6-1更针对于文本挖掘。下面，笔者将通过实验的分析解析如何建立文本实验数据，如何进行迁移学习，并进一步说明迁移学习在文本挖掘中的“负迁移”。

5.3 实验研究

为了评估算法的有效性，笔者以20NewsGroup[46]数据集为例，说明文本挖掘中迁移学习实验的设置。

5.3.1 数据集说明

20Newsgroup是一个包含20，000篇newsgroup文档的数据集，一共包含20类不同的文档标签。这20个类别如表6-1所示：

表 6-1　20Newsgroup 中的 20 个类

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

如表6-1所示，这20个类别中有的类别的关联性特别强。如Comp类别下分成了graphics, os, sys, windows四个子类别，而sys又继续分成ibm和mac两个子类别。因此，comp.sys.ibm.pc.hardware与comp.sys.mac.hardware的文档相当的相似，这也为迁移学习打下了基础。

文献[12]利用了这一特性，在文本挖掘中实现了迁移学习。其数据集构建如下：

表6-2　基于20Newsgroup的迁移学习实验的设计

Data Set	$\mathcal{D}_i$	$\mathcal{D}_o$
comp vs sci	comp.graphics comp.os.ms-windows.misc sci.crypt sci.electronics	comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x sci.med sci.space
rec vs talk	rec.autos rec.motorcycles talk.politics.guns talk.politics.misc	rec.sport.baseball rec.sport.hockey talk.politics.mideast talk.religion.misc
rec vs sci	rec.autos rec.sport.baseball sci.med sci.space	rec.motorcycles rec.sport.hockey sci.crypt sci.electronics
sci vs talk	sci.electronics sci.med talk.politics.misc talk.religion.misc	sci.crypt sci.space talk.politics.guns talk.politics.mideast
comp vs rec	comp.graphics comp.sys.ibm.pc.hardware comp.sys.mac.hardware rec.motorcycles rec.sport.hockey	comp.os.ms-windows.misc comp.windows.x rec.autos rec.sport.baseball
comp vs talk	comp.graphics comp.sys.mac.hardware comp.windows.x talk.politics.mideast talk.religion.misc	comp.os.ms-windows.misc comp.sys.ibm.pc.hardware talk.politics.guns talk.politics.misc

如图，文献[12]生成了6组数据集，每组数据集又包含一个源领域数据集Di和一个目标领域数据集Do。这两个数据集的文档共同拥有相同的母类别。例如源领域数据集包含comp.graphics，comp.os.ms-windows.misc等文档集合，目标领域数据集包含comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x等文档集合。于是这些文档都属于类别标签为comp的文档。而且由于文档所属的子类别不一样，这些文档的分布存在差异。因此，这样的数据集满足了迁移学习的要求，即数据不满足独立同分布，但是关联性较强。

在文本中，笔者也将采用上述的数据集结构。与文献[12]不同的是，目标领域的数据集一开始只包含2个类别不同的训练样本，而其他的均为测试样本。其数据集特性如下所示：

表 6-3　数据规格描述

Data Set	K-L	Documents			SVM	
		$\|\mathcal{D}_i\|$	$\|\mathcal{D}_o\|$	$\|\mathcal{W}\|$	$\mathcal{D}_i$-$\mathcal{D}_o$	$\mathcal{D}_o$-CV
rec vs talk	1.102	3,669	3,561	19,412	0.233	0.003
rec vs sci	1.021	3,961	3,965	18,152	0.212	0.007
comp vs talk	0.967	4,482	3,652	17,918	0.103	0.005
comp vs sci	0.874	3,930	4,900	18,379	0.317	0.012
comp vs rec	0.866	4,904	3,949	18,903	0.165	0.008
sci vs talk	0.854	3,374	3,828	20,057	0.226	0.009

在表6-3中，第一列为数据集组的名称。第二列为数据集组里面源领域数据集和目标领域数据集的Kullback-Leibler Divergence的值。该值常被用来衡量两个数据分布的差异，而独立同分布的数据集的KL值为0。第三列为数据集组的统计信息。如“rec vs. talk”的数据集组一共包含3669篇源领域文档，3561篇目标领域的文档，一共有19412个词。第四列为数据集组的分类特性。“Di-Do”表示用源领域数据集Di直接建立一个SVM分类器，再用该分类器对目标领域数据集Do进行分类的分类错误率。该值大则说明迁移学习的负迁移的风险大。而“Do-CV”的值是用10-折叠法对目标数据集Do进行错误率的预测。该值小则说明数据集越容易分割，即分类的复杂度比较小。

5.3.2 评价方法

由于本文提出的主动迁移学习框架从理论上证明一方面比迁移学习更准确；另一方面比主动学习的耗费更少。因此，笔者将从准确率和耗费两个方面对算法框架 AcTrak 进行评价。

首先，准确率的定义笔者采用正命中率的定义。而对耗费的衡量，在主动学习中，一般采用类似于图 6-1 的图形对算法进行评价。图 6-1 的横坐标轴为被领域专家所标注的训练样本的个数。一般来说，被标注的训练样本越多，产生的标注耗费也越多。而纵坐标为学习准确率，以 100%为上限。于是一个理想的主动学习器应该标注较少的训练样本，达到较高的分类准确率。如图 6-1 所示，主动学习算法 A1 明显比算法 A2 要优秀。而为了更加方便地量化算法 A1 比 A2 的优秀程度，笔者定义了一个叫做“IEA”（Integral Evaluation on Accuracy，卷积准确率）的函数。

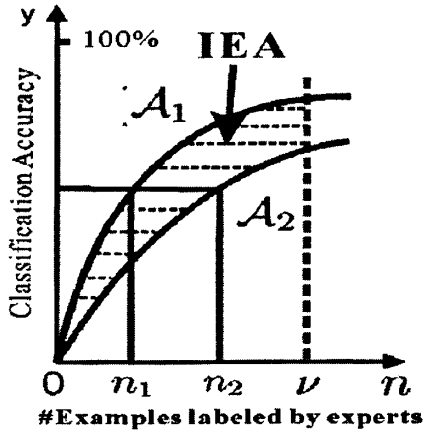

图 6-1 主动学习的量化标准 IEA

定义 1 已知一个分类器 M，两个主动学习算法 A1 及 A2。并定义 A(n)为利用主动学习算法 A 选择 n 个训练样本后，分类器 M 的准确率。那么，IEA 定义为：

$$\mathbf{IEA}(\mathcal{A}_1, \mathcal{A}_2, \nu) = \int_0^{\nu} (\mathcal{A}_1(n) - \mathcal{A}_2(n)) d\mathbf{n} = \sum_{n=0}^{\nu} (\mathcal{A}_1(n) - \mathcal{A}_2(n)) \Delta n$$

如定义 1 所示，IEA 实际上表达的是主动分类器 A1，A2 的学习曲线与直线 n=v 所围成的图形的面积。该部分面积越大，则说明 A1 比 A2 越优秀。容易得到，如果 IEA 为负值，则表明 A2 比 A1 优秀。笔者将利用 IEA 对主动学习框架 AcTrak 与主动学习器 ERS[44]进行对比。

5.3.3 实验结果分析

表 6-4 文本分类的实验结果

Dataset	SVM	TrAdaBoost(100 iterations)	AcTraK	IEA(AcTraK, ERS, 250)
rec vs. talk	60.2%	72.3%	**75.4%**	+0.91
rec vs. sci	59.1%	67.4%	**70.6%**	+1.83
comp vs. talk	53.6%	74.4%	**80.9%**	+0.21
comp vs. sci	52.7%	57.3%	**78.0%**	+0.88
comp vs. rec	49.1%	77.2%	**82.1%**	+0.35
sci vs. talk	57.6%	71.3%	**75.1%**	+0.84

实验结果如表 6-4 所示。一方面，主动迁移学习算法 AcTrak 在大多数情况下比传统的迁移学习算法 TrAdaBoost[16]的准确率要高。例如，在数据集“Comp VS.

Sci”中，TrAdaBoost的准确率只有57.3%，而AcTraK的准确率达到78%。这个结果得益于AcTraK能主动地选择可以迁移的知识与不可迁移的知识，最后达到避免负迁移的目的。另一方面，主动迁移算法 AcTraK 比传统的主动学习算法ERS[44]更加节省耗费。我们可以由表 6-4 看到，所有的 IEA 值皆为正值，说明AcTraK在这个文本挖掘的数据集中，比ERS算法更加优秀。

5.4 本章小结

本文主要针对文本分类所面临的维数灾难，稀疏向量等问题，提出了面向文本分类的主动迁移学习算法，通过运用特征选择及降维、稀疏向量筛除等方法进行改进。另外，本章还介绍了文本分类中的迁移学习的例子，并利用实验证明主动迁移学习算法能较为有效地提高文本分类的准确率、并有效地节省获得训练样本的耗费。

第6章 总结与展望

本章将对本文的内容进行总结，并介绍主动迁移学习框架的缺点及其研究的展望。

6.1 总结

数据挖掘是指从大量的数据(结构化和非结构化)中提取有用的信息和知识的过程。它源于大型数据系统的广泛使用和把数据转换成有用知识的迫切需要。数据挖掘是人工智能、机器学习、数学统计等技术的有机结合，其功能主要包括以下几个方面：概念描述、关联分析、有监督学习、无监督学习、偏差检测和时序演变分析。

有监督学习（分类分析）是数据挖掘中的一个重要的分析方法。分类分析的主要目的是利用已有类别标签的训练样本建立分类器，对没有类别标签的数据样本进行分类，并使得分类尽可能地准确。但是，传统的有监督学习要求训练样本与待分类的数据样本独立同分布，这个条件使得在现实应用中往往很缺乏满足条件的训练样本。

为了解决训练样本缺乏的问题，学者们提出了主动学习和迁移学习两种不同的框架。主动学习的目的是选择尽量少的数据样本作为训练样本，并使得分类的准确率尽量地高。因此，主动学习的研究重心是如何选择有代表性，信息量大的少量的数据样本。迁移学习的目的则是利用来自不同领域的训练样本来帮助目标领域的分类器的建立。于是在迁移学习中，一个重要的问题是来自源领域的训练样本与目标领域的数据样本并不满足独立同分布，因此传统的有监督学习方法并不适用于迁移学习。近年来，学者们也提出了不少关于迁移学习的算法。

但是，在解决训练样本缺乏的问题上，主动学习与迁移学习各有利弊。例如，主动学习对训练样本的要求依然很大，因而造成获得样本的代价在某些领域依然相当高；而迁移学习获得训练样本虽然是零代价，但是它有负迁移的风险。于是，笔者提出了主动迁移学习的框架，其基本思想为主动地选择可以迁移的知识，这

部分的知识可通过源领域以零代价获得，而其他的不可迁移学习的知识，则由主动学习获得。本文从理论和实验上证明了主动迁移学习算法不仅仅有效地提高了学习的准确率，还较好地降低了获得样本的代价。另外，本文还以文本分类为例，说明了主动迁移学习框架在实际应用中的意义。实验证明，结合文本分类中的特征降维，向量变换等技术，主动迁移学习框架能很好地用到实际应用中。

6.2 研究展望

虽然本文提出的主动迁移学习算法 AcTraK 可以比较大幅度地提高了分类的准确率，降低了获得训练样本的代价，但是笔者认为该算法的优化还可以继续下去。主要有三个大方向：

（1）效率问题。

由于主动迁移学习算法 AcTraK 需要对全局的待选的样本进行测试，以选择出足够具有代表性的数据样本，所以 AcTraK 的效率并不高。其中一个解决的方法是只对一个比较小的样本集里面的样本进行测试，或者结合聚类的方法同时选择多个具有代表性的训练样本，以提高算法的效率。

（2）算法思想的扩展。

AcTraK 只是主动迁移思想的一个具体算法。笔者认为算法的思想可以作进一步的扩展，即如何更好地使主动学习与迁移学习结合起来解决训练样本缺乏的问题。

（3）算法应用的扩展。

因为 AcTraK 有比较高的分类准确率，所以笔者认为可以把该算法应用在图像分类、组合分类器技术或者多任务学习的技术上。

参考文献

[1] Usama M. Fayyad, Ramasamy Uthurusamy. Data Mining and Knowledge Discovery in Databases (Introduction to the Special Section). Communications of the ACM (CACM). Volume 39, 1996 (11): 24~26

[2] G. Piatetsky-Shapiro. The Data-Mining Industry Coming of Age. IEEE Intelligent Systems, 2001: 32~34

[3] J.Han, M. Kamber. 数据挖掘概念与技术(影印版). 北京:高等教育出版社. 2001.5

[4] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

[5] John Shawe-Taylor, Nello Cristianini. Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000

[6] Hilbe, Joseph M. Logistic Regression Models. Chapman & Hall/CRC Press, 2009

[7] M. I. Jordan, Ed., Learning in Graphical Models, Adaptive Computation and Machine Learning, Cambridge, MA: MIT Press, 1999

[8] David J. C. MacKay, Information Theory, Inference & Learning Algorithms, Cambridge University Press, 2002

[9] **Xiaoxiao Shi**, Wei Fan, and Jiangtao Ren, Actively Transfer Domain Knowledge, 2008 European Confernce on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML/PKDD08), 2008

[10] Burr Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009

[11] Sinno Jialin Pan and Qiang Yang, A Survey on Transfer Learning, Computer Sciences Technical Report HKUST-CS08-08, 2008

[12] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, Yong Yu, Co-clustering based classification for out-of-domain documents, KDD'07, 2007

[13] Rich Caruana. Multitask learning. Machine Learning, 28(1):41–75, 1997

[14] Michael T. Rosenstein, Zvika Marx, and Leslie Pack Kaelbling. To transfer or not to transfer. In a NIPS-05 Workshop on Inductive Transfer: 10 Years Later, 2005

[15] Jiangtao Ren, **Xiaoxiao Shi**, Wei Fan, and Philip S. Yu "Type Independent Correction of Sample Selection Bias via Structural Discovery and Re-balancing", 2008 SIAM International Conference on Data Mining (SDM'08), 2008

[16] Wenyuan Dai, Qiang Yang, Guirong Xue, and Yong Yu. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning, 2007

[17] Xiao Ling, Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Spectral domain-transfer learning. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008

[18] Xuejun Liao, Ya Xue, and Lawrence Carin. Logistic regression with an auxiliary data source. In Proceedings of the 21st International Conference on Machine Learning, 2005

[19] Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning, Springer. 2001

[20] Mitchell, T., Machine Learning, McGraw Hill. 1997

[21] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer. 2006

[22] Sreerama, K.Murthy.Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery, 1998, 2:345-389.

[23] 龙军，殷建平，祝恩，赵文涛. 主动学习研究综述. 计算机研究与发展. 45(Suppl.): 300~304, 2008

[24] T Shcheffer, S Wrobel. Active learning of partially hidden Markov models. In Proceedings of the ECML/PKDD-2001, 2001

[25] G Schohn, D Cohn·Less is more: Active learning with support vector machines. In Proceeding of the 17th Int'l Conf on Machine Learning, 2000

[26] C Tohompson, M E Califf, R Mooney. Active learning for natural language

parsing and information extraction. In Proceedings of the 16th Int'l Conf on Machine Learning, 1999

[27] HS Seung, M Opper, H Sompolinsky. Query by committee· Annual Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, United States, 1992

[28] N Abe, H Mamitsuka. Query learning strategies using boosting and bagging. In Proceedings of the 15th Int'l Conf on Machine Learning, 1998·1-10

[29] P Melville, R J Mooney. Diverse ensembles for active learning· In Proceedings of the 21th Int'l Conf on Machine Learning, 2004

[30] P Melville, S M Yang, M Saar-Tsechansky,et al. Active learning for probability estimation using Jensen-Shannon divergence. The 16th European Conf on Machine Learning, 2000

[31] F Pereira, N Tishby, L Lee. Distributional clustering of English words. In: Proceedings of the 31st ACL. Morristown, NJ, USA: Association for Computational Linguistics, 1993

[32] N Roy, A McCallum. Toward optimal active learning through sampling estimation of error reduction. In Proceedings of the 18th Int'l Conf on Machine Learning, 2001

[33] S Tong, D Koller. Active learning for parameter estimation in Bayesian networks. In Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000

[34] M Lindenbaum, S Markovitch, D Rusakov. Selective sampling for nearest neighbor classifiers. Machine Learning, 2004, 54(2): 125-152

[35] Y Baram, R El-Yaniv, K Luz. Online choice of active learning algorithm. In Proceedings of the 20th Int'l Conf on Machine Learning, 2003

[36] I Muslea, S Minton, C A Knoblock. Active learning with multiple view. Journal of Artificial Intelligence Research, 2006, 27: 203-233

[37] H T Nguyen, A Smeulders. Active learning using pre-clustering. The 21st Int'l Conf on Machine Learning, 2004

[38] N Cesa-Bianchi, A Conconi, C Gentile. Learning probabilistic linear-threshold classifiers via selective sampling. In Proceedings of the 16th COLT. Berlin:

Spriniger, 2003

[39] Steffen Bickel, Michael Br¨uckner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In Proceedings of the 24th international conference on Machine learning, 2007

[40] Tony Jebara. Multi-task feature and kernel selection for svms. In Proceedings of the 21st International Conference on Machine Learning, 2004

[41] Pengcheng Wu and Thomas G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In Proceedings of the 21st International Conference on Machine Learning, 2004

[42] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In Proceedings of the 21st International Conference on Machine Learning, 2004

[43] Jiayuan Huang, Alex Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Sch¨olkopf. Correcting sample selection bias by unlabeled data. In Proceedings of the 19th Annual Conference on Neural Information Processing Systems. 2007

[44] Roy, N., McCallum, A. Toward optimal active learning through sampling estimation of error reduction. In: Proc. of ICML 2001, 2001

[45] Dhilloni, Kogan J, N Icholas C. Feature Selection and Document Clustering. 2002 CAD IP Research Symposium Proceedings. 2002

[46] Ken Lang, 20 Newsgroup. http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.

硕士期间论文及科研情况

[2008] **Xiaoxiao Shi**, Wei Fan, and Jiangtao Ren, "Actively Transfer Domain Knowledge", 2008 European Confernce on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML/PKDD08), 2008

[2008] Jiangtao Ren, **Xiaoxiao Shi**, Wei Fan, and Philip S. Yu "Type Independent Correction of Sample Selection Bias via Structural Discovery and Re-balancing", 2008 SIAM International Conference on Data Mining (SDM'08), Atlanta, GA, 2008.

后记

本论文是在任江涛老师的指导下完成的，在此对任老师表示衷心的感谢！同时，也衷心感谢来自美国 IBM Waston 实验室的范伟博士和来自伊利诺伊大学芝加哥分校的 Philip S. Yu 教授对该项研究的指导和支持。