软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(11):2762−2780 [doi: 10.13328/j.cnki.jos.004892]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种异构直推式迁移学习算法*

杨 柳[1,2,3], 景丽萍[1], 于 剑[1]

[1](交通数据分析与挖掘北京市重点实验室(北京交通大学),北京 100044)
[2](河北大学 数学与信息科学学院,河北 保定 071000)
[3](河北省机器学习与计算智能重点实验室(河北大学),河北 保定 071000)
通讯作者: 景丽萍, E-mail: lpjing@bjtu.edu.cn

摘 要: 目标领域已有类别标注的数据较少时会影响学习性能,而与之相关的其他源领域中存在一些已标注数据.迁移学习针对这一情况,提出将与目标领域不同但相关的源领域上学习到的知识应用到目标领域.在实际应用中,例如文本-图像、跨语言迁移学习等,源领域和目标领域的特征空间是不相同的,这就是异构迁移学习.关注的重点是利用源领域中已标注的数据来提高目标领域中未标注数据的学习性能,这种情况是异构直推式迁移学习.因为源领域和目标领域的特征空间不同,异构迁移学习的一个关键问题是学习从源领域到目标领域的映射函数.提出采用无监督匹配源领域和目标领域的特征空间的方法来学习映射函数.学到的映射函数可以把源领域中的数据在目标领域中重新表示.这样,重表示之后的已标注源领域数据可以被迁移到目标领域中.因此,可以采用标准的机器学习方法(例如支持向量机方法)来训练分类器,以对目标领域中未标注的数据进行类别预测.给出一个概率解释以说明其对数据中的一些噪声是具有鲁棒性的.同时还推导了一个样本复杂度的边界,也就是寻找映射函数时需要的样本数.在 4 个实际的数据库上的实验结果,展示了该方法的有效性.

关键词: 异构迁移学习;直推式迁移学习;异构特征空间;映射函数
中图法分类号: TP181

中文引用格式: 杨柳,景丽萍,于剑.一种异构直推式迁移学习算法.软件学报,2015,26(11):2762−2780. http://www.jos.org.cn/1000-9825/4892.htm

英文引用格式: Yang L, Jing LP, Yu J. Heterogeneous transductive transfer learning algorithm. Ruan Jian Xue Bao/Journal of Software, 2015,26(11):2762−2780 (in Chinese). http://www.jos.org.cn/1000-9825/4892.htm

Heterogeneous Transductive Transfer Learning Algorithm

YANG Liu[1,2,3], JING Li-Ping[1], YU Jian[1]

[1](Beijing Key Laboratory of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing 100044, China)
[2](College of Mathematics and Information Science, Hebei University, Baoding 071000, China)
[3](Key Laboratory of Machine Learning and Computational Intelligence (Hebei University), Baoding 071000, China)

Abstract: The lack of labeled data affects the performance in target domain. Fortunately, there are ample labeled data in some other related source domains. Transfer learning allows knowledge to be transferred from source domains to target domain. In real applications, such as text-image and cross-language transfer learning, the feature spaces of source and target domains are different, that is heterogeneous transfer learning. This paper focuses on heterogeneous transductive transfer learning (HTTL), an approach to improve the performance of unlabeled data in target domain by using some labeled data in heterogeneous source domains. Since the feature spaces of source domains and target domain are different, the key problem is to learn the mapping functions between the heterogeneous source domains and target domain. This paper proposes to learn the mapping functions by unsupervised matching in the different feature spaces.

* 基金项目: 国家自然科学基金(61375062, 61370129); 高等学校博士学科点专项科研基金(20120009110006); 中央高校基本科研业务费专项基金(2014JBM029); 河北省科技厅科技计划(13210347); 河北省教育厅资助项目(QN20131006); CCF-腾讯科研基金

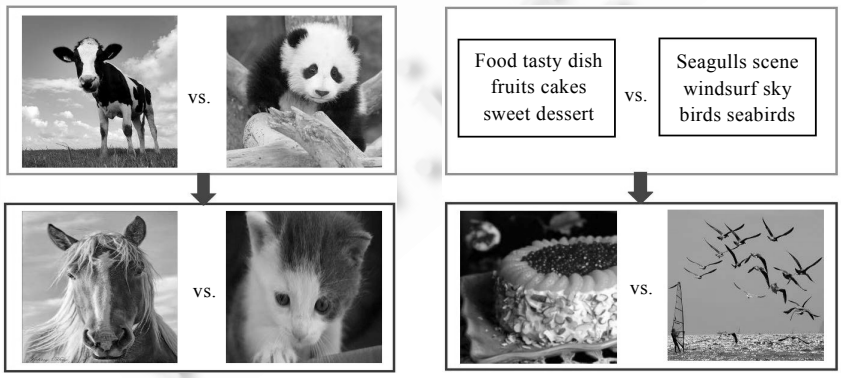
收稿时间: 2015-02-28; 修改时间: 2015-05-11, 2015-07-24; 定稿时间: 2015-08-26

The data in source domains can be re-represented with the mapping functions and transferred to the target domain. Thus, in target domain, there are some labeled data which come from the source domains. Standard machine learning methods such as support vector machine can be used to train classifiers for predicting the labels of unlabeled data in target domain. Moreover, a probabilistic interpretation is derived to verify the robustness of the presented method over certain noises in the utility matrices. A sample complexity bound is given to indicate how many instances are needed to adequately find the mapping functions. The effectiveness of the proposed approach is verified by experiments on four real-world data sets.

Key words: heterogeneous transfer learning; transductive transfer learning; heterogeneous feature space; mapping function

随着社会信息化的飞速发展,出现在人类日常生活和工作中的各种信息,如文本、图像、语音和视频等多媒体数据急剧增长.这些数据具有数量巨大、内容和形式多样等特点,用户从这些庞大繁杂的数据中获得有用的信息是非常困难的.分类技术可以对信息进行有效的组织管理,有利于快速而准确地定位信息,从而帮助用户获得有意义的数据.但是在很多领域中,有类别标注的训练样本非常短缺,如果对数据进行标注,需要投入大量的时间和资金成本.当没有大量已标注好的训练样本时,通过传统机器学习方法建立的分类器就不能达到令人满意的预测性能.针对这一情况,人们根据日常生活的经验,很自然地会想到,即使某一领域没有或者只有少数的训练样本,能否利用从相关领域学到的知识来帮助这个领域进行学习.迁移学习(transfer learning),或称归纳迁移、领域适配[1−14]就很好地解决了这个问题,它能够对先前领域中学到的知识进行识别,并将其应用到一个全新的领域中去.先前领域称为源领域,全新领域称为目标领域.在某种程度上,迁移学习源自于心理学上的“学习能力迁移”.如图 1(a)所示,学会区分“牛”和“熊猫”可以帮助学习区分“马”和“猫”.“牛”和“马”虽然相关,但是它们毕竟属于不同类的动物,那么它们的特征分布是不相同的.因此,迁移学习同时也放宽了训练和测试数据服从同一分布这一假设,重点关注如何最大程度地结合源领域知识来改善目标领域的学习性能.当它在目标领域中没有足够多的训练样本,却有很多与训练样本相关但特征分布不同的其他领域信息时,是非常有效的.

如果源领域和目标领域的数据属于同一特征空间,只是特征分布不同,这一类迁移学习的方法叫做同构迁移学习[1,7],例如图 1(a),源和目标领域的数据都是图像.而在很多实际应用中,尤其针对多源异构的大数据,相关源领域数据与目标领域数据往往属于不同特征空间,如图 1(b)的右下图所示:给定关于飞鸟的图片,计算机很难识别出里面包含天空、大海和飞鸟等,而可以通过其他手段(例如图片的标题、伴随文字等)得到一些与该图片相关的文本信息,见图 1(b)上面的文字,文字包含较多的语义信息,因此可以结合文本信息对图像进行语义理解.这一类方法处理的数据处于不同的特征空间,称为异构迁移学习[15,16].需要考虑如何计算源领域与目标领域的相关性,并且研究如何把有用的知识从源领域应用到目标领域中.



(a) 同构迁移学习 (b) 异构迁移学习

Fig.1 Homogeneous and heterogeneous transfer learning

图 1 同构和异构迁移学习

异构迁移学习是一门交叉学科,涉及多个不同的领域.目前,某些领域的技术方法比较成熟并且数据很多;

而某些领域技术难以达到较高的性能,并且数据较难得到.因此,异构迁移学习有广泛的应用前景.例如,图像与文本的信息融合(图像标注、分类和检索等)是异构迁移学习常用的应用之一.传统的图像标注是利用人工标注或者半自动标注,如果利用异构迁移学习自动地把有用的文本知识迁移到图像中,就可以对图像进行自动标注,从而节省大量的人力资源.图像分类比文本分类要难,如果结合相对成熟的文本技术来进行图像分类,就可以有效地提高图像分类的效率和准确率,图像检索经常会结合图像的标题、标注和伴随文本等信息来提高检索的性能.另一个常见的应用是跨语言学习:有些语言(例如英语)是很普遍的国际化语言,针对它的研究就相对比较多;然而有些小语种语言由于其使用群体人数较少且应用范围较小,因此针对它们的研究也比较少.各种语言之间有词典可以把它们的特征空间对应起来,这样,异构迁移学习就可以把研究相对较多的语言中的一些技术迁移到研究相对较弱的小语种中去,提高这些较弱语言的研究水平.与传统的学习相比,异构迁移学习难度更大,应用范围更广.

本文研究异构直推式迁移学习,也就是利用源领域中已标注的数据来辅助提高目标领域中数据的预测性能.第 1 节介绍和异构迁移学习相关内容.第 2 节介绍直推式的异构迁移学习.第 3 节是算法的鲁棒性和样本复杂度分析.第 4 节是实验结果和对比分析.最后是对全文的总结.

1 异构迁移学习

异构迁移学习中,关键的问题是源领域和目标领域中的数据处在不同的特征表示空间中,这也正是异构迁移学习最为挑战以及与其他学习模式不同的地方.在异构特征空间进行迁移学习,通常必须依赖领域特定的先验知识,包括特征空间之间的关联关系(如双语词典,如图 2(a)所示)、多模数据每个视图之间的对应关系(如网页中的文本和图像,如图 2(b)所示)、类别对应关系(如图 2(c)所示),等等.如果没有这类先验知识,则难以进行异构迁移学习.

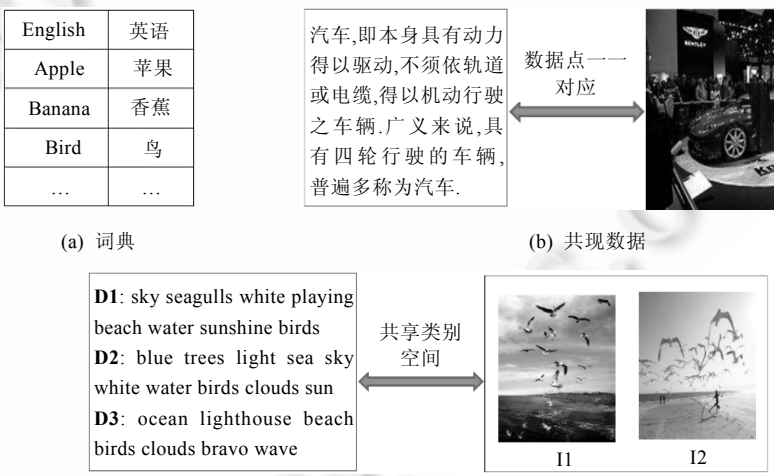


Fig.2 Bridge between source and target domains

图 2 源领域与目标领域之间的桥梁

如何建立特征空间之间的关联关系,一个比较直观的想法是,将一个特征空间中的样本直接翻译到另一个特征空间中去.这个方案在某些领域是成功的,如跨语言学习[17−20]、词典,如图 2(a)所示.文献[10,21−24]采用自动翻译将源领域语言翻译到目标领域语言中去,然后再处理所有语言在同一空间的问题,从而将异构迁移学习转化为同构迁移学习.Mahmud[10]提出翻译学习的方法,可以使用某一个特征空间上的训练样本学习分类模型,并对另一个特征空间上的测试样本进行预测.该方法依赖于不同特征空间之间的翻译模型,可以通过跨语言字

典或多模数据推导而来.Ling 等人[21]提出了一种训练数据为英文网页、测试数据为中文网页的方法.他们通过信息瓶颈理论来寻找公共信息,通过一些现有的翻译软件,例如 Google 翻译,把中文文档翻译成英文文档,然后使用信息论的技术,将训练文本与目标的分类文本一起做信息编码.Wu 等人[22]提出使用翻译的方法解决跨语言学习的问题:首先,不同语言的文本被翻译成同一种语言;然后,在学到的隐语义空间上,使用近邻算法对文本进行分类.Platt 等人[23]提出同时从不同语言的训练样本(源于互译平行语料库、机器翻译或者其他方法)中学习语言相关特征的投影映射,从而将不同特征空间映射到同一个与语言无关的高层抽象空间,然后,利用典型相关分析的方法得到不同语言间的关联关系.在学习到语言无关的抽象空间后,来自不同语言的所有文档都可以映射到共享子空间中.Shi 等人[24]提出给定标注时,翻译条件的概率通过由跨语言字典、源领域数据和目标领域数据进行期望最大化来协同学习得到.这一类异构迁移学习的学习算法大部分都需要使用一个翻译器将数据转换到同一个特征表示之中,因此只适用于完成跨语言的任务.

然而,对于更加一般的异构迁移学习,例如图像与文本间的异构迁移学习,机器翻译就不是那么容易了.因此,必须设计更一般的算法,能够不借助机器翻译的翻译学习算法.可以在有多个特征空间表示的数据基础上,来构建不同特征空间沟通的桥梁.这些由多个领域表示的数据是共现数据,虽然有时它们不能直接用来做分类任务的训练样本,但是可以用来构建异构特征空间之间的特征映射.在实际应用中有很多共现数据,例如百度图像搜索,用户输入的查询与搜索引擎返回的图像结果就是共现数据(如图 2(b)所示).这时,可以认为这些图像与用户查询在文本信息语义上非常相关.另外一个常用共现数据的例子是社交网站上的数据,例如 Flickr 这样的网站,用户上传的大量图片及其文字描述同样可以作为共现数据[11,15,16,25−29].如图 2(b)所示,一个文档对应一个图像,虽然文本词和视觉词的表示方式不同,但是它们都是说明一个事物.

多视角学习(multi-view learning)[28,29]是一种处理多个领域中共现数据的方法.它可以利用数据之间的一一对应关系来连接源领域和目标领域,但是多视角学习需要每个样本都有不同空间的表示.另外还有一些异构迁移学习的方法[11,15,16,25−27],可以利用共现数据学习到不同领域之间特征空间的对应关系.Dai 等人[11]提出了一个称作翻译学习的新的学习框架,通过使用风险最小化框架来解决数据属于不同特征领域之间的异构迁移学习问题.Yang 等人[15]提出一个无监督的概率模型.它是基于图像和文本的概率隐含语义分析模型.它把共现数据和目标领域内的数据统一到一个概率模型中,目的是为了提高目标领域内的聚类性能.Zhu 等人[16]提出针对图像-文本数据,采用共同矩阵分解技术,同时对“图像特征-单词标注”共同出现矩阵与“文档-单词标注”矩阵进行分解,产生一个更好的异构特征语义空间,将目标领域中的数据用低维特征重新表示,它是为了改善目标领域内数据分类的性能.实验结果表明,使用这种更深入挖掘高阶语义关联的模型在现实的图像分类任务上非常有效,对图像分类效果有很大的提升.Ng 等人[25,26]提出一个有监督的共同训练模型,分别计算领域内和领域之间的相似度矩阵,然后把多个领域(包含所有的源领域和目标领域)的信息整合到一个模型中,采用随机游走的方法,获得所有领域内未标注数据的分类结果.Yang 等人[27]提出一个无监督的共同聚类的模型,利用共现数据计算领域之间的相似度,然后利用谱聚类的思想对所有领域内的数据一起进行聚类.这些方法都是基于共现数据的,也就是需要一些在所有不同空间进行特征表示的数据,并且它们在建立多个领域之间的关系时没有利用到类信息.

文献[30−34]中关注于异构迁移学习这个问题,希望解决当训练数据在一个特征空间(如文本)而测试数据在另一个特征空间(如图像)的问题,同时,这两个特征空间的样例之间可能没有对应关系,但是它们共享同样的类别空间.例如图 2(c):3 个文档和两个图像没有一一对应关系,但它们都是与“飞鸟”相关的.这类异构迁移学习与其他种类学习模式的区别在于:原来学习问题上的很多限制被放宽了,使得不同特征空间中的数据已经无需再一一对应起来.

根据源领域和目标领域中是否有标签样本,迁移学习被划分为 3 类[1],如图 3 所示:目标领域中有少量标注样本的归纳迁移学习(inductive transfer learning)[35]、只有源领域中有标注样本的直推式迁移学习(transductive transfer learning)[36]以及源领域和目标领域都没有标注样本的无监督迁移学习(unsupervised transfer learning)[37].这个划分同样适用于异构迁移学习,前提是目标领域和源领域的特征空间不同.归纳异构迁移学习的基本假设是源领域中有大量已标注的样本,而目标领域中有少量已标注的样本,典型的模型有文献[30−34,

38−40].Harel 和 Mannor[30]采用谱方法并且结合类信息将源数据和目标数据建立对应关系,对源领域的数据进行重新表示,然后训练分类器,以便对目标领域中的数据进行分类.Wang 等人[31]结合类别信息和特征信息建立了两个相似度矩阵和一个不相似度矩阵,然后利用谱分解的方法学到源领域和目标领域到一个新的低维空间的映射关系.新的空间里面保持了数据点原有的类别和特征信息.所有数据点通过新的表示之后,即,都在同一个特征空间里面,因此可以进行传统机器学习的方法完成分类或者聚类任务.Duan 等人[32]借用支持向量机(support vector machine,简称 SVM)的思想,结合类信息学习,把源领域和目标领域中的数据都映射到一个新的低维空间,然后,结合各自空间的特征组成新的特征表示.Li 等人[33]又对此方法进行半监督扩展,学习映射的时候结合了未标注的目标领域内的数据.Zhou 等人[34]提出:首先对源领域和目标领域内的数据进行各自的 SVM 学习,然后利用学到的分类边界信息结合稀疏的思想,直接学习两个领域的特征映射关系.Shi 等人[39,40]提出的异构映射(heterogeneous mapping,简称 HeMap)模型将源领域和目标领域的数据映射到同一空间,既能很好地保持原有数据的本质结构,又能最大化源领域和目标领域的相似性.但是该方法需要用样本抽样的方法将源领域和目标领域中的样本个数保持一致,当训练样本较少时,从源领域迁移到目标领域中的信息量较少,性能会受影响.以上这些方法需要源领域和目标领域中都有类别信息,它们不能处理目标领域中没有已标注类的数据.本文研究直推式异构迁移学习,也就是源领域中包含已标注的样本,而目标领域中所有数据都没有标注信息.在已存在的模型中,都不能解决这一类问题.

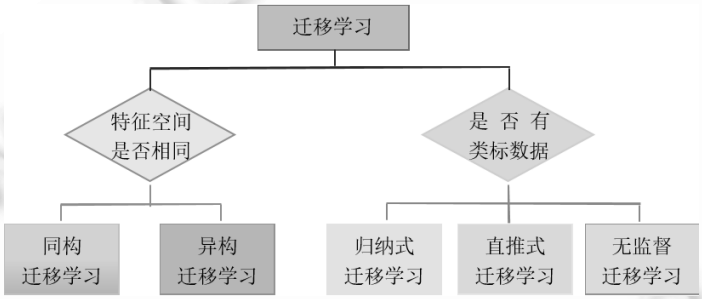


Fig.3 Taxonomy system of transfer learning

图 3 迁移学习分类体系

2 异构直推式迁移学习

本节将介绍异构直推式迁移学习(heterogeneous transductive transfer learning,简称 HTTL),即,源领域存在大量标注数据,而目标领域仅存在无类标数据的迁移学习问题.本节首先形式化一个源领域和一个目标领域的多类分类学习模型,然后推广到多个源领域和一个目标领域的学习任务上.

2.1 问题描述

假设有一个有类标的源领域数据库合和一个无类标的目标领域数据库合,目标领域与源领域的特征空间是不相同的,但是它们预测的类别空间是相同的.目的是在两个领域数据的帮助下建立一个分类器,并且尽可能准确地预测目标领域中所有样本的分类标记.源领域样本的数目是 $n^{(s)}$,特征空间的维数表示为 $m^{(s)}$,目标领域样本的数目是 $n^{(t)}$,特征空间的维数表示为 $m^{(t)}$,源领域和目标领域中类的个数都是 c.源领域定义的数据库是 $\{(x_k^{(s)}, y_k)\}_{k=1}^{n^{(s)}}$,其中,$x_k^{(s)} \in \mathbb{R}^{m^{(s)}}$ 是源领域中的第 k 个样本,y_k 是样本 x_k 对应的分类类别.目标领域定义的无标记的数据库是 $\{x_k^{(t)}\}_{k=1}^{n^{(t)}}$,其中,$x_k^{(t)} \in \mathbb{R}^{m^{(t)}}$ 是目标领域中的第 k 个样本.学习目的是对目标领域中的数据进行类别预测.把所有的源领域内数据和类别信息以及目标领域内的数据分别连接在一起,形成矩阵 $X^{(s)} \in \mathbb{R}^{m^{(s)} \times n^{(s)}}$,$Y^{(s)} \in \mathbb{R}^{c \times n^{(s)}}$ 和 $X^{(t)} \in \mathbb{R}^{m^{(t)} \times n^{(t)}}$.

2.2 无监督特征映射学习

异构迁移学习的重点是在异构的特征空间之间建立联系,因为两个空间的特征属性完全不同,因此需要学习映射矩阵来关联两个完全异构的空间.本节将只利用源领域和目标领域的数据矩阵 $X^{(s)}$和 $X^{(t)}$来学习映射矩阵,没有用到类别信息,因此是无监督的.

首先需要进行预处理,使得每个领域内的数据处于同一个取值范围内.可以采用的方法是对每个领域内的数据进行归一化处理,这样使得所有数据都处于相同的范围内.需要对所有领域内的数据进行操作,即,包括源领域和目标领域中已标注类别的和未标注类别的所有数据.进行完数据归一化处理之后,对源领域和目标领域内的所有数据进行去均值操作,使得每个领域内的数据的均值变成 0.假设$\mu^{(s)}$是源领域数据的均值,$\mu^{(t)}$是目标领域数据的均值.通过以下两个公式去均值,得到数据新表示 $\tilde{X}^{(s)}$ 和 $\tilde{X}^{(t)}$.

$$\tilde{X}^{(s)} = X^{(s)} - \mu^{(s)} \tag{1}$$

$$\tilde{X}^{(t)} = X^{(t)} - \mu^{(t)} \tag{2}$$

经过预处理之后,就可以对两个领域内的数据进行匹配.这步操作的目的是对预处理之后的数据建立映射关系,因为两个空间的维数不同,且每个空间的坐标系统不同.本文选择的方法是对每个领域内数据进行坐标变换,使得变换后的坐标都能够表达该领域内的主要特征信息.可以借鉴特征重构的方法来寻找关键特征,也就是对每个领域内的数据进行重新表示,方法有很多种,其中,主成分分析的方法可以通过旋转变换坐标系,使得变换之后各个主成分之间的方向是正交的[41].它有很好的物理意义,因为通过主成分分析之后,找到前 $h(h\leqslant \min(m^{(s)},m^{(t)}))$个方向是方差最大的方向,这样就能在新空间上更好地区分开不同类别的样本.对不同领域进行这种坐标变换,表达的意义都是能够使样本更加具有可分性.因此,在这样的新空间之间建立映射关系是比较合理的.

通过以下步骤得到新转换之后的数据表示:对源领域内的数据进行奇异值分解(singular value decomposition,简称 SVD),选择前 h 个最大的特征值对应的特征向量作为旋转之后的前 h 个主方向.由前 h 个主方向形成的矩阵标记为 $D^{(s)}\in\mathbb{R}^{m^{(s)}\times h}$, 这样,源领域内的数据就可以在这 h 个主方向上进行新的表示.同理,对目标领域内数据进行主成分分析,可以得到前 h 个主方向形成的矩阵标记为 $D^{(t)}\in\mathbb{R}^{m^{(t)}\times h}$. 得到源领域和目标领域旋转之后的主方向之后,就可以来建立两个领域之间的映射关系.采用的方法是把源领域的前 h 个主成分映射到目标领域中,然后计算与目标领域主方向的距离,具体公式为最小化以下目标函数:

$$\left.\begin{aligned}&\arg\min_{R} \| RD^{(s)} - D^{(t)} \|_F^2\\ &\text{s.t. } R^TR = I\end{aligned}\right\} \tag{3}$$

其中,$\|\cdot\|_F$ 是 Frobenius 范数,用来表示距离; $R\in\mathbb{R}^{m^{(t)}\times m^{(s)}}$ 是两个领域的映射矩阵.对公式(3)中的目标函数进行展开,得到:

$$\| RD^{(s)} - D^{(t)} \|_F^2 = tr(D^{(s)T}R^TRD^{(s)} - 2D^{(t)T}RD^{(s)} + D^{(t)T}D^{(t)}) .$$

$tr(\cdot)$是矩阵的迹,$D^{(t)T}D^{(t)}$为常数项,并且有约束条件 $R^TR=I$,则 $D^{(s)T}R^TRD^{(s)}$也是常数项.那么,问题(3)转化成最大化下面的形式:

$$\left.\begin{aligned}&\arg\max_{R} tr(D^{(t)T}RD^{(s)})\\ &\text{s.t. } R^TR = I\end{aligned}\right\} \tag{4}$$

因为 $D^{(s)}$和 $D^{(t)}$是由 h 个主成分方向组成,也就是 $D^{(s)}=[v_1^{(s)},...,v_h^{(s)}]$ 和 $D^{(t)}=[v_1^{(t)},...,v_h^{(t)}]$ ($v_l^{(s)}$ 和 $v_l^{(t)}$ (l=1,…,h) 分别是源领域和目标领域的第 l 个主成分方向),为了得到更直观的解释,可以把公式(4)转化成:

$$\begin{aligned}&\arg\max_{R} \sum_{l=1}^{h} v_l^{(t)T}Rv_l^{(s)}\\ &\text{s.t. } R^TR = I.\end{aligned}$$

问题(3)是最小化两个领域主方向的夹角,通过公式(4)转换成了最大化两个领域主方向的内积之和.由于 $tr(D^{(t)T}RD^{(s)})=tr(RD^{(s)}D^{(t)T})$,可将公式(4)变为

$$\left.\begin{array}{l}\arg\max\limits_{R} tr(RD^{(s)}D^{(t)T}) \\ \text{s.t. } R^TR=I\end{array}\right\} \tag{5}$$

为了求解公式(5),引入 Procrustes 分析[42],对 $D^{(s)}D^{(t)T}$ 进行奇异值分解(SVD),得到 USV^T.令 $Z=V^TRU$,则

$$tr(RD^{(s)}D^{(t)T})=tr(RUSV^T)=tr(ZS)=\sum_{\tau=1}^{m}z_{\tau\tau}\sigma_\tau \tag{6}$$

其中,σ_τ是对 $D^{(s)}D^{(t)T}$ 进行 SVD 分解的第τ个奇异值.由于 $Z^TZ=U^TR^TVV^TRU=I$,因此 $\sum_{\tau=1}^{m}z_{\tau\tau}\sigma_\tau\leqslant\sum_{\tau=1}^{m}\sigma_\tau$, 则

$$\max tr(RD^{(s)}D^{(t)T})=\sum_{\tau=1}^{m}\sigma_\tau=tr(S).$$

也就是 $Z=I$,可以得到 $R=VU^T$.

2.3 异构直推迁移模型

通过映射矩阵 R,使得两个领域进行匹配,然后,把处理之后的源领域内数据加上目标领域内的均值,这样就把源领域数据和目标领域数据建立了映射关系,统一到一个空间中.这样,通过 R 得到源领域数据 $\tilde{X}^{(s)}$ 在目标领域内的表示为 $\hat{X}^{(s)}=R\tilde{X}^{(s)}+\mu^{(t)}$. 具体步骤可以总结如算法 1.

算法 1. 异构直推迁移学习.

输入:目标领域数据矩阵 $X^{(t)}$,源领域数据矩阵和类别矩阵 $X^{(s)},Y^{(s)}$.

输出:目标领域数据预测结果 $Y^{(t)}$.

(1) $\tilde{X}^{(t)}=X^{(t)}-\mu^{(t)}$.

(2) $\tilde{X}^{(s)}=X^{(s)}-\mu^{(s)}$.

(3) 构建矩阵 $D^{(t)},D^{(s)}$.

(4) SVD 分解 $D^{(s)}D^{(t)T}=USV^T$.

(5) $R=VU^T$.

(6) $\hat{X}^{(s)}=R\tilde{X}^{(s)}+\mu^{(t)}$.

(7) 在源领域数据 $\{\hat{X}^{(s)},Y^{(s)}\}$ 上训练 SVM 分类器.

(8) 利用在源领域数据建立的分类器,对目标领域内所有数据 $X^{(t)}$进行预测,得到结果分类 $Y^{(t)}$.

在异构迁移学习中,源领域和目标领域的特征空间不同,因此特征维数也不相同,也就是 $m^{(s)}\neq m^{(t)}$,因此,R 不是一个方阵. $D^{(s)}\in\mathbb{R}^{m^{(s)}\times h}$,$D^{(t)}\in\mathbb{R}^{m^{(t)}\times h}$, 如果 $m^{(t)}<m^{(s)}$,则把矩阵 $D^{(t)}$的后面补充 $m^{(s)}-m^{(t)}$行;反之,则把矩阵 $D^{(s)}$的后面补充 $m^{(t)}-m^{(s)}$行.令 $m=\max(m^{(s)},m^{(t)})$,这样,R 就是一个 $m\times m$ 的方阵.

2.4 时间复杂度分析

异构直推迁移学习中主要的步骤是构建矩阵 $D^{(t)},D^{(s)}$,即,对 $\tilde{X}^{(t)}$ 和 $\tilde{X}^{(s)}$ 分别进行 SVD 分解,时间复杂度分别是 $O(hT^{(t)}N_{nz}^{(t)})$ 和 $O(hT^{(s)}N_{nz}^{(s)})$ [43].其中,h 是降维之后的维度,$T^{(t)}$和 $T^{(s)}$分别是对 $\tilde{X}^{(t)}$ 和 $\tilde{X}^{(s)}$ 进行 SVD 时的 Lanczos 迭代次数, $N_{nz}^{(t)}$ 和 $N_{nz}^{(s)}$ 是 $\tilde{X}^{(t)}$ 和 $\tilde{X}^{(s)}$ 中的非零个数.另外一个主要步骤是对 $D^{(s)}D^{(t)T}$ 进行 SVD 分解, Lanczos 迭代次数为 T,非零个数为 N_{nz},则时间复杂度是 $O(hTN_{nz})$.其中,$h,T^{(t)},T^{(s)},T$ 都不大,如果矩阵是稀疏的, $N_{nz}^{(t)},N_{nz}^{(s)}$ 和 N_{nz} 也不太大,那么,时间复杂度不是太高.

2.5 扩展到k个领域

算法 1 是针对一个源领域和一个目标领域,如果有多个(假设有 Q 个)源领域数据 $\{X^{(j)}\}_{j=1}^{Q}$,也可以同时转换到一个目标领域 $X^{(t)}$中去.这种情况下,多个有类标标注数据的源领域可以更好地帮助目标领域的数据进行分类预测.与一个源领域类似,第一步需要把所有的数据都进行去均值,求主成分方向,这样得到 Q 个源领域矩阵 $\{D^{(j)}\}_{j=1}^{Q}$ 和一个目标领域矩阵 $D^{(t)}$.然后通过以下公式,把所有源领域的特征空间转换到目标领域中:

$$\left.\begin{aligned}&\arg\min_{\{R^{(j)}\}_{j=1}^{Q}}\sum_{j=1}^{Q}\|R^{(j)}D^{(j)}-D^{(t)}\|_F^2\\&\text{s.t. } R^{(j)T}R^{(j)}=I\end{aligned}\right\}\tag{7}$$

求 $R^{(j)}$的方法与算法 1 中求 $R^{(s)}$的方法类似,是对 $D^{(j)}D^{(t)T}$ 进行 SVD 分解,$D^{(j)}D^{(t)T}=U^{(j)}S^{(j)}V^{(i)T}$,得到 $R^{(j)}=V^{(j)}U^{(j)T}$.那么 $\hat{X}^{(j)}=(X^{(j)}-\mu^{(j)})R^{(j)T}+\mu^{(t)}$,这样就得到了源领域数据 $X^{(j)}$在目标领域中的新表示 $\hat{X}^{(j)}$.然后,可以利用传统分类方法进行学习.

3 算法分析

本节中将分析噪声对 HTTL 模型的影响,证明该方法的鲁棒性;同时,将分析样本复杂度情况.

3.1 鲁棒性能分析

本节将研究在源领域的旋转矩阵 $D^{(s)}$中加入噪声Δ时对模型的影响.噪声Δ可以是任意的,不需要指定它服从的分布,只是利用它的边界值分析 HTTL 模型的可信度.证明的思路是:如果可以把源领域中的噪声转换为在原始问题的基础上加一个依赖于噪声的常数,该噪声Δ只对边界有一个可加效果,这样就可以证明源领域中的噪声不影响系统的性能.

假定Δ是在矩阵 $D^{(s)}$上的随机不确定的噪声,它服从一个未知的分布$\Delta\sim\Gamma$.这个不确定性可以由问题(3)的机会约束来描述[44,45],即,把这个问题转化为以一定的概率满足约束条件;或者说违反约束的概率不能超过一个给定的值,这个值为违反概率.如果违反概率设定得合理,就可以在不违法原则的情况下得到较为理想的解:

$$\left.\begin{aligned}&\min_{R^TR=I,\pi}\pi\\&P_{\Delta\sim\Gamma}\{\|R(D^{(s)}+\Delta)-D^{(t)}\|_F\leqslant\pi\}\geqslant1-\eta\end{aligned}\right\}\tag{8}$$

其中,$\eta\in[0,1]$是期望的置信度.尽管机会约束问题有较好的直观概率形式,但是它的求解不太容易[46].因此.可以采用公式(8)的近似问题,假定 $\rho^*=\inf_{\pi}\{P_{\Delta\sim\Gamma}(\|\Delta\|_F\leqslant\pi)\geqslant1-\eta\}$,得到:

$$\|R(D^{(s)}+\Delta)-D^{(t)}\|_F\leqslant\max_{\|\Delta\|_F\leqslant\rho^*}\|R(D^{(s)}+\Delta)-D^{(t)}\|_F\tag{9}$$

则问题(8)可以转化成下面的最小最大化问题:

$$\min_{R^TR=I}\max_{\|\Delta\|_F\leqslant\rho^*}\|R(D^{(s)}+\Delta)-D^{(t)}\|_F\tag{10}$$

这样,需要证明的是:在不确定的噪声集合 $U=\{\Delta|\|\Delta\|_F\leqslant\rho^*\}$的基础上,公式(10)使原始的映射问题变得更加鲁棒.

定理 1. 公式(10)等价于 $\min_{R^TR=I}\|RD^{(s)}-D^{(t)}\|_F+\rho^*$.

证明:公式(10)中的 Frobenius 范数可以转换成迹的形式:

$$\max_{\|\Delta\|_F\leqslant\rho^*}\|R(D^{(s)}+\Delta)-D^{(t)}\|_F=\max_{\|\Delta\|_F\leqslant\rho^*,\|A\|_F\leqslant1}tr(A^T(R(D^{(s)}+\Delta)-D^{(t)})\tag{11}$$

其中,$A=R(D^{(s)}+\Delta)-D^{(t)}$.那么,公式(11)可转换成:

$$\max_{\|\Delta\|_F\leqslant\rho^*,\|A\|_F\leqslant1}tr(A^T(R(D^{(s)}+\Delta)-D^{(t)}))=\max_{\|V\|_F\leqslant1}\left(tr(A^T(RD^{(s)}-D^{(t)}))+\max_{\|\Delta\|_F\leqslant\rho^*}tr(A^TR\Delta)\right)\tag{12}$$

对于公式(12)中第 2 项 $\max_{\|\Delta\|_F\leqslant\rho^*}tr(A^TR\Delta)$,利用 Cauchy-Schwartz 不等式,结合约束 $R^TR=I$,可以得到:

$$\max_{\|\Delta\|_F\leqslant\rho^*}tr(A^TR\Delta)\leqslant\max_{\|\Delta\|_F\leqslant\rho^*}\|A\|_F\|R\Delta\|_F=\rho^*\|A\|_F\tag{13}$$

假定$\Delta^*=R^TA/\|A\|_F$,可以得到:

$$\max_{\|\Delta\|_F\leqslant\rho^*}tr(A^TR\Delta)\geqslant tr(A^TR\Delta^*)\tag{14}$$

$\|\Delta^*\|_F=\rho^*$,则 $tr(A^TRR^TA)/\|A\|_F^2=\rho^*$, 也就是 $tr(A^TRR^TA)/\|A\|_F=\rho^*\|A\|_F$.那么 $tr(A^TR\Delta^*)=\rho^*\|A\|_F$,因此,

$$\max_{\|\Delta\|_F \leqslant \rho^*} tr(A^T R\Delta) \geqslant \rho^* \| A \|_F \tag{15}$$

通过公式(13)和公式(15),可以得到 $\max_{\|\Delta\|_F \leqslant \rho^*} tr(A^T R\Delta) = \rho^* \| A \|_F$,带入公式(12):

$$\max_{\|\Delta\|_F \leqslant \rho^*} \| R(D^{(s)} + \Delta) - D^{(t)} \|_F = \max_{\|\Delta\|_F \leqslant \rho^*, \|A\|_F \leqslant 1} (tr(A^T (RD^{(s)} - D^{(t)})) + \rho^* \| A \|_F) = \| RD^{(s)} - D^{(t)} \|_F + \rho^* \tag{16}$$

由此,定理 1 得证. □

由定理 1 可知,可以把源领域中任意形式的噪声转换为在原始问题的基础上加一个依赖于噪声的常数,这样,该噪声Δ只对边界有一个可加效果,因此可以证明源领域中的噪声不影响系统的性能,该模型具有鲁棒性.

3.2 样本复杂度分析

本节将讨论样本复杂度的边界.

假定 1. 每个领域的数据由 c 个高斯分布混合而成(c 是类的个数), $x \sim \sum_i^c w_i f_i(x)$, 其中,$f_i(x)$服从均值为$\mu_i$,方差为$\Sigma_i$的高斯分布,也就是$f_i(x) \sim N(\mu_i, \Sigma_i)$,每个类的混合系数 w_i 满足 $\sum_i^c w_i = 1$. 假定每个主成分$\|E((xx^T)\| \leqslant 1$.

定理 2. 如果假定 1 成立,每个领域对于$\delta \in [0,1]$,假设样本的个数满足:

$$n \geqslant \sum_{i=1}^c C \frac{mh^2}{\varepsilon_i^2} \log^2\left(\frac{32mh^2}{\varepsilon_i^2}\right) \log^2\left(\frac{4hm}{\delta}\right) \tag{17}$$

那么, $P(\| \tilde{R} - R \| \leqslant \varepsilon) \geqslant 1 - \delta$,其中, $\tilde{R}$ 是通过算法 1 步骤(5)估计的旋转矩阵,m 是特征维数,C 是一个常量.

在证明定理 2 之前,先证明以下的引理:

引理 1. 如果假定 1 成立,每个领域对于$\delta \in [0,1]$,假设第 i 类样本的个数满足:

$$n_i \geqslant C \frac{2m}{\varepsilon} \log\left(\frac{m}{\delta}\right) \tag{18}$$

那么, $P(\| \tilde{\mu} - \mu \| \leqslant \varepsilon) \geqslant 1 - \delta$,其中, $\tilde{\mu}$ 和μ分别是每个主成分估计的和实际的均值.

证明:使用 $\sigma_{\max}^2 = \max_l(\sigma_l^2)$ 是各个主成分方向中的最大方差.σ_l 是样本在第 l 个主成分方向的标准差.

对 $|\tilde{\mu} - \mu|$ 的每个主成分应用 Chernoff 边界[47],得到如果样本个数满足 $n_i \geqslant C \frac{2\sigma_{\max}^2 m}{\varepsilon_i^2} \log\left(\frac{m}{\delta}\right)$,则

$$P(\| \tilde{\mu} - \mu \| \leqslant \varepsilon) \geqslant 1 - \delta.$$

由假定 1 得到 $\sigma_{\max}^2 \leqslant 1$,则该引理成立. □

引理 2. 假设数据 X 由 n 个样本组成,每个样本 x_k 服从一维均值为μ方差为σ^2 的高斯分布,则

$$P\left(| x_k - \mu | \leqslant \sigma \sqrt{2\log\left(\frac{n}{\delta}\right)}\right) \geqslant 1 - \delta, \forall x_k \in X \tag{19}$$

引理 3. 假设 $x_1,\ldots,x_n$ 是服从多项分布的独立随机向量集合,那么,

$$P\left(\| x_k \| \leqslant \| \mu \| + \sigma \sqrt{2m\log\left(\frac{nm}{\delta}\right)}\right) \geqslant 1 - \delta \tag{20}$$

证明:根据三角不等式:

$$\|x_k\| - \|\mu\| \leqslant |\|x_k\| - \|\mu\|| \leqslant \|x_k - \mu\| \tag{21}$$

对向量 x_k 的第 l 个主成分,应用引理 2,得到:

$$P\left(| x_k - \mu_l | \geqslant \frac{\varepsilon}{\sqrt{m}}\right) \leqslant n \exp\left(-\frac{1}{2}\frac{\varepsilon^2}{\sigma^2 m}\right) \leqslant \frac{\delta}{m} \tag{22}$$

合并 m 个主成分,得到:

$$P\left(\|x_k\|-\|\mu\|\leqslant\|x_k-\mu\|\leqslant\sigma\sqrt{2m\log\left(\frac{nm}{\delta}\right)}\right)\geqslant 1-\delta \tag{23}$$

得证. □

引理 4. 假设 X 是一个随机样本集合,服从高斯分布,协方差矩阵是Σ,均值为μ=0.定义 $\tilde{\Sigma},\tilde{\mu}$ 是估计的协方差矩阵和均值.那么$\delta\in[0,1]$,如果样本大小为

$$n\geqslant C\frac{m}{\varepsilon_2^2}\log^2\left(\frac{2m}{\varepsilon_2^2}\right)\log^2\left(\frac{2m}{\delta}\right) \tag{24}$$

那么,

$$P(\|\tilde{\Sigma}-\Sigma\|\leqslant\varepsilon_1+\varepsilon_2)\geqslant 1-\delta \tag{25}$$

证明:

$$\|\tilde{\Sigma}-\Sigma\|\leqslant\|\mu\mu^T-\tilde{\mu}\tilde{\mu}^T\|+\left\|\frac{1}{n}\sum_{k=1}^{n}x_kx_k^T-E(xx^T)\right\| \tag{26}$$

因为μ=0,因此第 1 个成分的边界是$\|\tilde{\mu}\|^2$,应用引理 1,当 $n_1\geqslant\frac{2m}{\varepsilon_1}\log\left(\frac{2m}{\delta}\right)$时,得到:

$$P(\|\tilde{\mu}\|^2\leqslant\varepsilon_1)\geqslant 1-\frac{\delta}{2} \tag{27}$$

第 2 个主成分的边界由协方差矩阵的集中不等式来限定[48].应用引理 3 和假定 1,得到:

$$\|x_k\|\leqslant\sqrt{2m\log\left(\frac{n_2m}{\delta}\right)}.$$

由文献[48],定义 $b=\tilde{C}\sqrt{2m\log\left(\frac{n_2m}{\delta}\right)}\sqrt{\frac{\log(n_2)}{n_2}}$,其中,$\tilde{C}$ 为常数,可得:

$$P\left(\left\|\frac{1}{n}\sum_{k=1}^{n}x_kx_k^T-E(xx^T)\right\|\leqslant e\right)\geqslant 1-2\exp(-ce^2/b^2).$$

用 $t^2=a^2\log\left(\frac{2}{\delta}\right)\Big/c$ 带入 e^2 和 $a^2=\varepsilon_2^2c^2\Big/\log\left(\frac{2}{\delta}\right)$带入 e^2,则

$$a=\frac{\varepsilon_2c}{\sqrt{\log\left(\frac{2}{\delta}\right)}}\geqslant\tilde{C}\frac{\sqrt{2m\log\left(\frac{n_2m}{\delta}\right)\log(n_2)}}{\sqrt{n_2}} \tag{28}$$

其中,n_2 满足:

$$n_2\geqslant C\frac{m}{\varepsilon_2^2}\log^2\left(\frac{2m}{\varepsilon_2^2}\right)\log^2\left(\frac{2m}{\delta}\right) \tag{29}$$

最终的边界选择均值估计(27)和方差估计(29)中较大的一个 n_2,因此可得公式(24). □

下面证明定理 2.

经过去均值和归一化处理,每个主成分的均值均为 0.应用引理 1,第 j 个领域中第 i 类的样本复杂度是 $n_i^{(j)}\geqslant\frac{2m^{(j)}}{\varepsilon^2}\log\left(\frac{m^{(j)}}{\delta}\right)$,其中,$n^{(j)}$为 $n^{(s)}$或者 $n^{(t)}$,分别表示源领域和目标领域中样本的个数.应用三角不等式,得到:

$$\|\tilde{R}-R\|_F=\|\tilde{V}\tilde{U}^T-VU^T\|_F\leqslant\|V\|\|\Delta U\|+\|\Delta V\|\|\Delta U\|+\|\Delta V\|\|U\| \tag{30}$$

其中,$\Delta V=\tilde{V}-V,\Delta U=\tilde{U}-U.V,\tilde{V},U,\tilde{U}$ 是对 $D^{(s)}D^{(t)T}$ 和 $\tilde{D}^{(s)}\tilde{D}^{(t)T}$ 进行 SVD 分解,$D^{(s)}D^{(t)T}=USV^T$, $\tilde{D}^{(s)}\tilde{D}^{(t)T}=\tilde{U}\tilde{S}\tilde{V}^T$:

$$\|\tilde{R}-R\|_F\leqslant C\|D^{(s)}D^{(t)T}-\tilde{D}^{(s)}\tilde{D}^{(t)T}\|_F \tag{31}$$

由三角不等式得到:

$$\| D^{(s)}D^{(t)T} - \tilde{D}^{(s)}\tilde{D}^{(t)T} \|_F \leqslant \| D^{(s)}D^{(t)T} - D^{(s)}\tilde{D}^{(t)T} \|_F + \| D^{(s)}\tilde{D}^{(t)T} - \tilde{D}^{s}\tilde{D}^{(t)T} \|_F \tag{32}$$

通过 F 范数的次可乘性得到:

$$\begin{aligned}\| D^{(s)}D^{(t)T} - D^{(s)}\tilde{D}^{(t)T} \|_F + \| D^{(s)}\tilde{D}^{(t)T} - \tilde{D}^{s}\tilde{D}^{(t)T} \|_F &\leqslant \sqrt{h}\,\| D^{(t)T} - \tilde{D}^{(t)T} \|_F + \sqrt{h}\,\| D^{(s)} - \tilde{D}^{(s)} \|_F \\ &= \sqrt{h}(\| \Delta D^{(s)} \|_F + \| \Delta D^{(t)} \|_F).\end{aligned}$$

当列的个数 $h<d$ 时,需要补充一些零列.

定义 $v_l^{(j)}$ 和 $\tilde{v}_l^{(j)}(l=1,...,h)$ 是 $D^{(j)}$和 $\tilde{D}^{(j)}$ 的第 h 个特征向量,那么 $\| \Delta D^{(j)} \|_F^2 = \sum_{l=1}^{h} \| \tilde{v}_l^{(j)} - v_l^{(j)} \|_F^2$.

定义 $E^{(j)} = \Sigma^{(j)} - \tilde{\Sigma}^{(j)}$, 使用扰动理论[49],得到 $\| \tilde{v}_l^{(j)} - v_l^{(j)} \| \leqslant C \| E^{(j)} \|$.

应用引理 4 计算边界 $E^{(j)}$,用 $\frac{\varepsilon_1^{(j)} + \varepsilon_2^{(j)}}{4h}$ 代替$\varepsilon_1+\varepsilon_2$,用 $\frac{\delta}{2h}$ 代替δ,如果每类样本数是:

$$n^{(j)} \geqslant C\frac{mh^2}{(\varepsilon_2^{(j)})^2}\log^2\left(\frac{32mh^2}{(\varepsilon_2^{(j)})^2}\right)\log^2\left(\frac{4hm}{\delta}\right) \tag{33}$$

那么,所有类的样本数为公式(17),则

$$P\left(\| \Sigma^{(j)} - \tilde{\Sigma}^{(j)} \| \leqslant \frac{\varepsilon_1^{(j)} + \varepsilon_2^{(j)}}{4h}\right) \geqslant 1 - \frac{\delta}{2h} \tag{34}$$

由于 $h \geqslant 1$,可得:

$$P\left(\| \Delta D^{(j)} \| \leqslant \frac{\varepsilon_1^{(j)} + \varepsilon_2^{(j)}}{4\sqrt{h}}\right) \geqslant 1 - \frac{\delta}{2} \tag{35}$$

得到最后的边界,令 $\varepsilon = \frac{1}{4}\sum_{j=s,t} \varepsilon_1^{(j)} + \varepsilon_2^{(j)}$:

$$P(\| D^{(s)}D^{(s)T} - \tilde{D}^{(t)}\tilde{D}^{(t)T} \|_F \leqslant \varepsilon) \geqslant 1 - \delta \tag{36}$$

因此,定理 2 可证. □

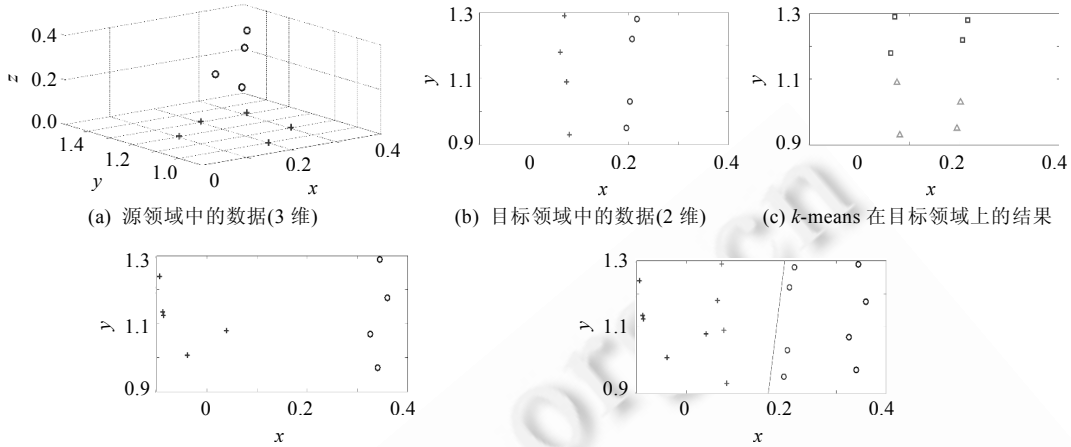
4 实验结果

本节在一个模拟数据和 4 个实际应用数据库上进行系统性的实验,证明 HTTL 模型的有效性.在源领域中建立的分类器,本文采用的是 SVM 方法.下面分别对模拟实验、真实数据库、基准方法以及实验结果进行介绍.

4.1 模拟实验

为了验证本文方法在目标领域中数据量较少或所含信息匮乏时的结果,首先构造具有已有类标标注的源领域数据库.如图 4(a)所示,用十字和圆形来表示在 3 维源领域空间中的两类样本.然后构造数据较少且没有监督信息的目标领域数据库.图 4(b)展示了 2 维特征空间的目标领域数据,并标出了真正的类别.模型的假设是源领域和目标领域共享类别空间,因此也用十字和圆形来表示不同的类别.

图 4(c)展示了 *k*-means 的聚类结果,可以看出,由于目标领域中样本数量较少,得到的结果与实际的标注结果是有一定差异的.图 4(d)展示了源领域(3 维空间)中的数据,通过 HTTL 得到的在目标领域中(2 维空间)的新表示.在此基础之上,学到的分类器如图 4(e)中的边界线所示.基于这个分类边界,可以得到目标领域中的分类结果,与实际的标注结果是一致的.由此可见,本文所提出的 HTTL 算法充分利用了源领域中的监督信息,从而对目标领域中完全无监督的数据进行归类有很好的指导作用.



(a) 源领域中的数据(3 维) (b) 目标领域中的数据(2 维) (c) *k*-means 在目标领域上的结果

(d) 源领域数据在目标领域中的新表示 (e) 由新表示建立的分类器以及对目标领域中数据的预测

Fig.4 Experiment on synthetic data

图 4 模拟数据上的实验结果

4.2 真实数据库

- NUS-WIDE

该数据库中的图像来自 Flickr,每个图像对应一些文字说明,共包含 81 个类.按照文献[25,26]介绍的方法,本文从中选择 6 个类(birds,food,boats,flower,tower,sun),两两组建成 15 个两类任务.每个任务包括图像 600 幅,每幅图像的视觉特征用词袋(bag of words,简称 BOW)表示,特征维数为 500 维.文档有 1 200 个,每个文档由 1 000 维 0-1 特征组成,其中,每个维度上对应的词在该文档中出现即为 1,否则为 0.

- BDGP(berkeley drosophila genome project)

它是一个 5 类的生物数据库[50],其中每类代表一个基因生长的阶段.共包含 2 500 幅图像,每个类别各有 500 幅.每幅图像有 1 750 维的视觉特征,同时,有生物学家标注的 79 维文本词,每维文本特征和 NUS-WIDE 一样,也是 0-1 值.

- LabelMe

LabelMe[51]是风景图像库,包含 8 类(coast,forest,highway,insidecity,mountain,opencountry,street,tallbuilding)共 2 366 幅图像,其中,属于 highway 的有 266 幅,其余每类各包含 300 幅.每幅图像由 BOW 模型生成 240 维视觉特征组成,同时对应 809 维由词频构成的文本特征,平均每个文档有 7 个词,其中,最高词频为 36.

- 跨语言分类数据

该数据库包括 3 种语言的文档(英语、法语和西班牙语)[25,26],所有文档都来自于 Google 和 Wikipedia.每一种语言的文档分为两个类别(birds 和 animals).该数据库包括英语文档 3 415 篇,法语 2 511 篇和西班牙语 3 113 篇,其中,3 种语言属于“birds”类的分别是 1 525 篇、1 500 篇和 1 520 篇,其余文档属于“animals”类别.每个语言的特征都由词频表示,特征维数是 5 000.其中,英语、法语和西班牙语平均每个文档有 42,45 和 29 个词,最高词频都为 20.

4.3 基准方法

在现有的异构迁移学习模型中,文献[30–34,39,40]需要目标领域中也有类别标注的数据,文献[15,16,25–27]需要有共现数据.其中,原始的 HeMap[39,40]模型在分类时需要源领域和目标领域都有标注数据,但是它可以在无监督的情况下学到两个映射矩阵,将源领域和目标领域分别映射到同一空间.因此,可以利用学习到的映射矩阵对该模型进行改变.在新空间下,将已标注的源领域数据作为训练集,用以学习分类器,然后对未标注的目标领域数据进行分类.本文将修改后的 HeMap 作为基准的异构迁移学习的方法.另外,如果对目标领域中没有类别

标注数据进行归类,只能采用聚类的方法,因此也将最常用在单个领域中的 k-means 和谱聚类方法(spectral clustering,简称 SC)作为对比的方法.

采用目标领域数据上的准确率(accuracy)作为评价指标:

$$Accuracy=\frac{|x_k:k=1,\ldots,n^{(t)},f(x_k^{(t)})=y(x_k^{(t)})|}{n^{(t)}} \tag{37}$$

其中, $y(x_k^{(t)})$ 是测试样例 $x_k^{(t)}$ 的真实标签, $f(x_k^{(t)})$ 是待测学习算法为样例 $x_k^{(t)}$ 预测的标签.准确率是目标领域中预测正确的样本个数与所有样本的比值,准确率越大,性能越好.在实验中,对于每个数据库的每个设置情况下进行 10 次,用平均准确率对结果进行评价.

4.4 实验结果分析

本节给出 HTTL 和基准方法在各分类任务上的准确率,并对实验结果进行了分析.

4.4.1 NUS-WIDE

这个数据库展示了 HTTL 在两类图像-文本数据上的性能,其中,图像数据为目标领域,文本数据为源领域.目的是借助已标注的文本作为训练集,对目标领域中的 600 幅图像进行分类.

(1) 参数影响

HTTL 方法中只有一个参数 h,用来控制各个领域降维之后的维度.从 15 个任务中选择了一个数据库“boats vs tower”来展示参数 h 对分类准确率的影响.从源领域的每个类中选择了 25 个作为训练样本,对目标领域中图像进行分类,准确率如图 5 所示.从图中可以看出,分类准确率开始时随着 h 的增大而增加,当 h>10 以后开始下降.h 在区间 4~10 时可以取得较好的结果,这个维数远远小于文本和图像原始的维度 1 000 和 500,说明在较低的维度上就能得到本质的特征.

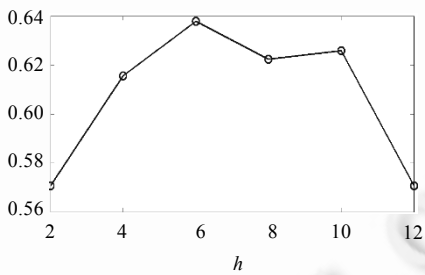


Fig.5 Effect of different parameters (h) on HTTL with a NUS-WIDE binary classification task (boats vs. tower)

图 5 不同参数 h 对 HTTL 在 NUS-WIDE(boats vs. tower)上分类准确率的影响

(2) 源领域数据个数的影响

本节将测试源领域中样本个数的变化对目标领域的分类结果的影响,源领域中每类样本的个数由 5 增加到 25.由于 HeMap 在源领域中样本个数较少时性能较差,只列出了该方法在源领域训练样本为 25 的结果,同时,列出了 k-means 和 SC 方法作为基准进行对比.HTTL,HeMap,k-means 和 SC 在 15 个分类任务上 10 次的平均准确率和标准差见表 1,观察到,HTTL 一般比标准学习方法能获得更好的分类准确率.例如,HTTL 在源领域训练样本为 25 时的平均准确率是 0.62,而 HeMap 在训练样本为 25 时以及 k-means 和 SC 的平均准确率分别为 0.57,0.54 和 0.55,HTTL 分别比 HeMap,k-means 和 SC 提高了 7.5%,13.9%和 12.3%.

从实验结果可以看出,HTTL 和 HeMap 结合了文本数据,图像分类性能得到了较好的改善.这说明虽然文本与图像数据来自不同的特征空间,但是与图像相关的文本能够对图像进行归类有很好的帮助.图像的视觉特征包含较少的语义信息,对计算机来说较难理解.相对于图像来说,文本含有较多的语义信息,对图像语义理解很有帮助.因此,结合了文本信息的图像更容易被计算机进行正确的归类.HeMap 在一些数据(例如 flowers-food)上

的性能比 *k*-means 和 SC 还要差,这是因为 HeMap 模型需要把源领域和目标领域中的样本个数进行补齐,而在训练样本比较少时不能体现出该方法的优势,也就是从源领域中获取的信息较少.而 HTTL 没有这个限制,即使是在训练文本数据比较少时(例如有 5 个训练文本数据),也比传统的聚类方法有很好的提升效果(*k*-means 为 9.0%,SC 为 7.4%).

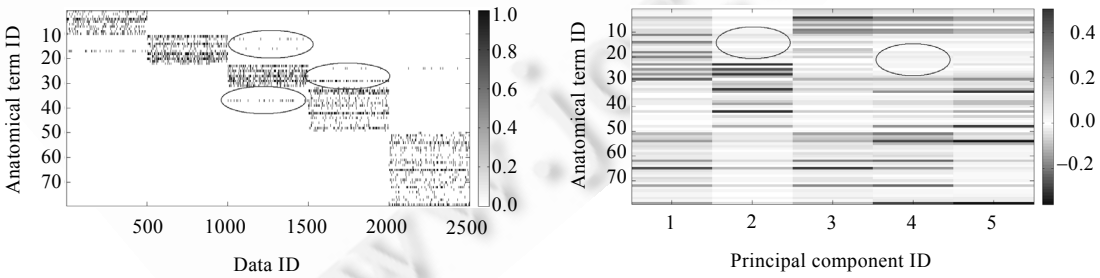
Table 1 Mean and standard deviations of classification accuracies on NUS-WIDE data set

表 1 NUS-WIDE 数据库的平均分类准确率及标准差

Task	*k*-means	SC	源领域中每类训练样本的个数					
			HeMap	HTTL				
			25	5	10	15	20	25
birds-sun	0.61±0.00	0.62±0.00	0.65±0.08	0.67±0.10	0.67±0.09	0.68±0.09	0.68±0.08	0.69±0.08
birds-food	0.53±0.00	0.55±0.00	0.53±0.02	0.56±0.01	0.57±0.01	0.57±0.01	0.58±0.01	0.59±0.01
birds-tower	0.51±0.00	0.50±0.00	0.51±0.02	0.53±0.08	0.54±0.08	0.54±0.07	0.55±0.02	0.56±0.03
birds-boats	0.53±0.01	0.54±0.00	0.57±0.05	0.57±0.06	0.58±0.05	0.58±0.06	0.58±0.06	0.59±0.05
birds-flowers	0.51±0.00	0.52±0.00	0.54±0.01	0.53±0.02	0.55±0.02	0.56±0.01	0.56±0.01	0.56±0.01
boats-food	0.56±0.00	0.56±0.00	0.60±0.07	0.61±0.06	0.62±0.09	0.62±0.08	0.62±0.07	0.63±0.05
boats-sun	0.55±0.00	0.56±0.00	0.61±0.05	0.62±0.08	0.63±0.05	0.63±0.06	0.64±0.02	0.64±0.06
boats-tower	0.50±0.01	0.51±0.00	0.58±0.03	0.60±0.05	0.61±0.05	0.61±0.06	0.63±0.01	0.64±0.02
boats-flowers	0.52±0.00	0.53±0.00	0.57±0.06	0.58±0.07	0.59±0.07	0.60±0.06	0.60±0.05	0.61±0.04
flowers-food	0.52±0.00	0.52±0.00	0.51±0.01	0.53±0.02	0.54±0.02	0.55±0.02	0.56±0.02	0.56±0.01
flowers-sun	0.60±0.00	0.61±0.00	0.63±0.07	0.65±0.09	0.66±0.09	0.67±0.09	0.67±0.09	0.68±0.09
flowers-tower	0.50±0.00	0.51±0.00	0.58±0.05	0.60±0.07	0.61±0.04	0.61±0.08	0.62±0.07	0.62±0.04
food-sun	0.61±0.00	0.62±0.00	0.60±0.07	0.65±0.10	0.66±0.09	0.67±0.09	0.67±0.10	0.67±0.09
food-tower	0.55±0.01	0.56±0.00	0.55±0.05	0.58±0.11	0.59±0.10	0.60±0.11	0.61±0.09	0.61±0.10
sun-tower	0.51±0.00	0.52±0.00	0.57±0.05	0.58±0.04	0.58±0.03	0.59±0.04	0.60±0.02	0.61±0.03

4.4.2 BDGP

第 2 个数据库 BDGP 也是图像-文本数据,但是它包含 5 类数据,因此它展示了 HTTL 在多类数据上的性能.与 NUS-WIDE 相同,图像和文本数据分别为目标领域和源领域.BDGP 数据库中,5 个类别分别对应基因生长的 5 个阶段.图 6(a)展示了所有 2 500 个数据在文本领域 79 个特征上的分布情况(图中每 500 个数据对应一个类别),可以看出,数据中含有噪声(例如,用椭圆标示出的数据会影响分类性能).HTTL 首先对源领域和目标领域的数据进行主成分分析,然后在主成分的基础之上建立关联.图 6(b)展示了源领域的文本特征与前 5 个主成分(从前 6 个最小特征值去掉最小的主成分)的分布情况,可以看出,其中也包含一些噪声(例如用椭圆标示出的部分).因此,该实验测试源领域中的训练样本带有噪声时,对目标领域中数据分类性能影响的情况.



(a) 原始数据($X^{(s)}$) (b) 原始文本特征与 5 个主成分的关系矩阵($D^{(s)}$)

Fig.6 Illustration of BDGP data in source (text) domain

图 6 BDGP 源领域(文本)数据展示

从 2 500 幅图像中随机选择了 2 000 幅图像作为测试样本,每次从剩余的 500 幅图像对应的文本中,每类随机选择 5~25 个文档作为训练数据.对于每个固定的训练样本大小,HTTL 和 HeMap 利用已标注的文本对目标领域中的图像进行分类.所有方法都进行 10 次,记录平均准确率.表 2 展示了 *k*-means,SC 的结果以及 HeMap 和

HTTL 结合 5~25 个文本数据的分类结果.

从表 2 可以看出,随着文本数据个数的增加,*k*-means 和 SC 只利用目标领域中的数据,结果保持不变;而 HeMap 和 HTTL 分类准确率会随着增加.这说明训练文本数据越多,对图像进行分类的指导作用就越大.而 HeMap 在训练样本为 5~20 时性能比 SC 差一些,这是因为在源领域样本个数较少时,HeMap 模型不能从源领域中获取较多有用信息.根据对文本领域数据的分析(如图 6 所示),虽然在引入源领域数据中含有噪声,但是 HTTL 的分类性能相比单个领域的 *k*-means,SC 方法仍然有所提高,这说明 HTTL 对源领域数据中的噪声具有鲁棒性.

Table 2 Mean and standard deviations of classification accuracies on BDGP data set

表 2 BDGP 数据库的平均分类准确率及标准差

方法		源领域中每类训练样本的个数				
		5	10	15	20	25
异构迁移学习方法	HTTL	0.508±0.031	0.517±0.037	0.526±0.031	0.532±0.036	0.539±0.031
	HeMap	0.437±0.028	0.453±0.022	0.465±0.021	0.473±0.027	0.483±0.021
目标领域内的聚类方法	*k*-means			0.442±0.011		
	SC			0.479±0.006		

由表 2 可知,随着源领域中训练样本个数的增加,HTTL 的分类性能会逐渐提高.也就是说,较多的训练样本能够达到较好的分类性能.由定理 2 可知,当样本达到一定数量时(公式(17)),可以达到 $P(\|\tilde{R}-R\|\leqslant\varepsilon)\geqslant 1-\delta$.例如,对于 BDGP 数据库中的源领域(文本)数据,m=79,实验中 h=5.由于不能得到真正的 R 值,因此也不能得到实际的 $\|\tilde{R}-R\|$ 值,只能估计 $\|\tilde{R}-R\|$ 的取值范围为 $0\sim 2\tilde{R}$.根据实验中记录的 R 值,如果想得到 0.54 的准确率,那么计算出每类需要的样本数大约为 24.通过表 2 可以看出,当源领域中每类选择 25 个训练样本时,HTTL 的平均分类准确率为 0.539,基本与定理 2 中的理论结果保持一致.然而定理 2 给出的只是一个理论上的边界,也就是当样本数满足一定条件时,能够获得一定的分类性能.但是给出的边界往往比较松,也就是当样本数少于给定的边界时,也可能会获得相同的性能.

4.4.3 LabelMe

LabelMe 是风景图像数据库,从每类中选择 200 幅图像(8 类共 1 600 幅图像)作为目标领域的数据,每类从其余的图像中随机选择 20 幅图像对应的文本组成源领域数据库合,任务是对目标领域内的 1 600 幅图像进行正确归类.每个图像对应一些文本描述,随机选择了两个类别(Coast 和 Forest)的样本图像和文本如图 7 所示.由于 LabelMe 原始的文本描述是人工标注,包含的噪声较少,为了测试噪声对 HTTL 方法的影响,在源领域的文本中随机选择了 5~20%的文档加入噪声,图 7 中也展示了对原始数据加入了一些噪声词的文本数据,例如第 1 幅图像中加入了噪声词“car”和“sun”.由于源领域中加入了噪声,那么也会给经过主成分分析之后得到的 $D^{(s)}$带来一些噪声.

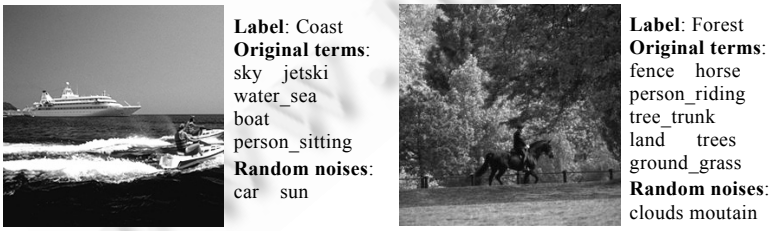


Fig.7 Samples of LabelMe data set

图 7 LabelMe 数据库中的样本数据

表 3 展示了 *k*-means,SC 以及 HeMap 和 HTTL 在没有噪声(0%)和加入不同比例噪声(5~20%)的分类结果.从表 3 可以看出,由于引入了源领域数据,HeMap 和 HTTL 比 *k*-means,SC 的分类性能有所提高,这说明源领域中已有类别标注的文本数据对图像进行分类有指导作用.随着源领域引入了噪声比例的增加,HeMap 和 HTTL 的

性能有所下降,其中,HeMap 下降得较多,而 HTTL 的分类性能下降得较少.这说明,如果源领域数据中含有噪声,HTTL 受影响的程度比较小.

Table 3 Mean and standard deviations of classification accuracies with different noise proportions on LabelMe data set

表 3 LabelMe 数据库加入不同噪声比例的平均分类准确率及标准差

方法		噪声比例(%)				
		0	5	10	15	20
异构迁移学习方法	HTTL	0.505±0.032	0.498±0.038	0.493±0.030	0.489±0.034	0.481±0.031
	HeMap	0.457±0.029	0.449±0.026	0.437±0.023	0.426±0.025	0.411±0.023
目标领域内的聚类方法	*k*-means	0.403±0.021				
	SC	0.438±0.017				

4.4.4 跨语言分类

跨语言数据库是 3 个领域(English,French,Spanish)的两类数据,它展示了 HTTL 在多个源领域数据上的性能.对于每个领域,*k*-means 和 SC 作为基准的方法,它们只利用目标领域中的数据.HeMap 和 HTTL 结合了源领域中的已标注的数据,对目标领域的数据进行分类.以 English 作为目标领域的数据为例,目标是对该领域内的 2 511 篇文档进行归类,HTTL 可以分别利用 French 和 Spanish 作为源领域,也可以利用 French 和 Spanish 一起作为源领域.HeMap 模型只能处理一个源领域和一个目标领域的数据.每次从每个源领域的每个类别中选取 5~25 个样本组成训练集.对于每个固定的训练样本大小,进行 10 次得到的平均准确率见表 4.由于 HeMap 在源领域数据较少时性能较差,因此只列出了源领域中训练样本个数为 25 的结果.

Table 4 Mean and standard deviations of classification accuracies on three-cross-language data set

表 4 3 个领域的跨语言数据库的平均分类准确率及标准差

Target	*k*-means	SC	Source	源领域中每类训练样本的个数					
				HeMap	HTTL				
				25	5	10	15	20	25
English	0.57±0.02	0.58±0.00	French	0.60±0.04	0.59±0.04	0.62±0.03	0.63±0.03	0.64±0.01	0.65±0.01
			Spanish	0.59±0.03	0.59±0.03	0.61±0.03	0.63±0.03	0.64±0.02	0.65±0.02
			All	–	0.61±0.03	0.63±0.04	0.64±0.03	0.66±0.02	0.67±0.02
French	0.58±0.03	0.59±0.00	English	0.60±0.02	0.60±0.04	0.60±0.03	0.62±0.03	0.63±0.02	0.64±0.02
			Spanish	0.60±0.03	0.60±0.04	0.62±0.04	0.63±0.03	0.64±0.02	0.65±0.01
			All	–	0.61±0.04	0.62±0.03	0.64±0.03	0.65±0.03	0.67±0.02
Spanish	0.57±0.03	0.58±0.00	English	0.60±0.03	0.59±0.03	0.61±0.03	0.63±0.03	0.64±0.02	0.64±0.02
			French	0.61±0.02	0.60±0.03	0.61±0.03	0.64±0.03	0.65±0.02	0.65±0.02
			All	–	0.60±0.04	0.62±0.03	0.65±0.03	0.66±0.02	0.67±0.02

从表 4 可以看出,HTTL 随着源领域中训练数据个数的增加,分类准确率会相应地增加.对于每种语言,如果加入了其他语言作为训练集,分类性能有所提高.例如,对于 Spanish,*k*-means 和 SC 的结果分别是 0.57 和 0.58.而 HTTL 只利用 5 个 English 和 French 文档作为指导时,准确率分别变成了 0.59 和 0.60;当利用 25 个训练样本时,准确率分别变成了 0.64 和 0.65;当 25 个 English 和 French 一起作为源领域数据时,准确率变成了 0.67.这说明随着训练样本的增多,性能会得到提高;同时,随着源领域个数的增加,性能也会有所改善.虽然不同语言的特征空间不同,但是由于这些文档的类别空间一致,也就是它们的高层语义空间是一致的,因此各个语言之间进行知识的迁移,可以提高某一个领域中的分类性能.由于 HeMap 模型中需要源领域和目标领域数据个数一致的限制,当对应源领域数据个数较少时,效果提升不是特别明显.而 HTTL 模型则没有这个约束限制,同时对源领域中数据具有鲁棒性,因此获得了较好的性能.

5 结束语

本文针对源领域中有已标注数据、目标领域中没有标注数据,且源领域和目标领域的特征空间不相同,提

出了异构直推式迁移学习 HTTL 模型,采用匹配源领域和目标领域的特征空间的方法来学习映射函数.首先,进行对每个领域内的数据进行归一化处理;然后是对两个领域内的数据通过去均值,建立映射关系,统一到目标领域特征空间中.这样,把源领域中的数据重表示之后迁移到目标领域中去,在目标领域中就包含了来自源领域中已标注的数据.本文采用支持向量机方法来训练分类器,以对目标领域中未标注的数据进行类别预测.该模型可直接推广到多个源领域和一个目标领域的学习任务上.本文展示了该模型对数据中的一些噪声是具有鲁棒性的,并且还推导了一个样本复杂度的边界.在一个模拟数据和 4 个实际的数据库上的结果表明,:该方法通过把已标定的异构数据迁移到目标领域中,可以改善目标领域中为未标定数据的性能.

HTTL 完成的任务是在目标领域中没有任何标注数据的情况下,结合源领域中的已标注数据,对该领域中的数据进行归类.而本文提出的从源领域和目标领域到共同空间的映射模型是无监督的,也就是在学习时没有用到源领域中的类别标注信息.今后的工作将考虑如何改进异构迁移学习模型,使它们能够充分利用源领域数据和类别信息来建立源领域和目标领域之间的关联关系,从而从源领域中获得更准确的知识,使得目标领域中的分类性能得到更好的提高.

References:

[1] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans on Knowledge and Data Engineering, 2010,22(10):1345−1359. [doi: 10.1109/TKDE.2009.191]

[2] Zhuang FZ, He Q, Shi ZZ. Survey on transfer learning research. Ruan Jian Xue Bao/Journal of Software, 2015,26(1):26−39 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4631.htm [doi: 10.13328/j.cnki.jos.004631]

[3] Liao XJ, Xue Y, Carin L. Logistic regression with an auxiliary data source. In: Proc. of the 22nd Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2005. 505−512.

[4] Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning. In: Proc. of the Int'l Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2006. 120−128.

[5] Samarth S, Sylvian R. Cross domain knowledge transfer using structured representations. In: Proc. of the 21st Conf. on Artificial Intelligence. AAAI Press, 2006. 506−511.

[6] Ben-David S, Blitzer J, Crammer K, Pereira F. Analysis of representations for domain adaptation. In: Proc. of the Advances in Neural Information Processing Systems 19. Cambridge: MIT Press, 2007. 137−144.

[7] Dai WY, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Proc. of the 24th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2007. 193−200.

[8] Dai WY, Xue GR, Yang Q, Yu Y. Transferring naive Bayes classifiers for text classification. In: Proc. of the 22nd Conf. on Artificial Intelligence. AAAI Press, 2007. 540−545.

[9] Xing DK, Dai WY, Xue GR, Yu Y. Bridged refinement for transfer learning. In: Proc. of the 11th European Conf. on Practice of Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2007. 324−335. [doi: 10.1007/978-3-540-74976-9_31]

[10] Mahmud MMH. On universal transfer learning. In: Proc. of the 18th Int'l Conf. on Algorithmic Learning Theory. 2007. 135−149. [doi: 10.1007/978-3-540-75225-7_14]

[11] Dai WY, Chen YQ, Xue GR, Yang Q, Yu Y. Translated learning: Transfer learning across different feature spaces. In: Proc. of the Advances in Neural Information Processing System. Cambridge: MIT Press, 2008. 353−360.

[12] Luo P, Zhuang FZ, Xiong H, Xiong YH, He Q. Transfer learning from multiple source domains via consensus regularization. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2008. 103−112. [doi: 10.1145/1458082.1458099]

[13] Duan LX, Tsang IW, Xu D, Chua TS. Domain adaptation from multiple sources via auxiliary classifiers. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. New York: ACM Press, 2009. 289−296. [doi: 10.1145/1553374.1553411]

[14] Zhuang FZ, Luo P, Xiong H, Xiong YH, He Q, Shi ZZ. Cross-Domain learning from multiple sources: A consensus regularization perspective. IEEE Trans. on Knowledge And Data Engineering, 2010,22(12):1664−1678. [doi: 10.1109/TKDE.2009.205]

[15] Yang Q, Chen Y, Xue GR, Yu Y. Heterogeneous transfer learning for image clustering via the social Web. In: Proc. of the ACL/IJCNLP. 2009. 1−9.

[16] Zhu Y, Chen Y, Lu Z, Pan SJ, Xue GR, Yu Y, Yang Q. Heterogeneous transfer learning for image classification. In: Proc. of the 26th Conf. on Artificial Intelligence. AAAI Press, 2011. 1304−1309.

[17] Bel N, Koster CHA, Villegas M. Cross-Lingual text categorization. In: Proc. of the European Conf. on Digital Libraries. Berlin: Springer-Verlag, 2003. 126−139. [doi: 10.1007/978-3-540-45175-4_13]

[18] Olsson JS, Oard DW, Hajič J. Cross-Language text classification. In: Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2005. 645−646.

[19] Wei B, Pal C. Cross lingual adaptation: An experiment on sentiment classifications. In: Proc. of the 48th Annual Meeting of the Association of Computational Linguistics. 2010. 258−262.

[20] Prettenhofer P, Stein B. Cross-Language text classification using structural correspondence learning. In: Proc. of the 48th Annual Meeting of the Association of Computational Linguistics. 2010. 1118−127.

[21] Ling X, Xue GR, Dai W, Jiang Y, Yang Q, Yu Y. Can Chinese Web pages be classified with English data source? In: Proc. of the 17th Int'l Conf. on World Wide Web. 2008. 969−978. [doi: 10.1145/1367497.1367628]

[22] Wu Y, Oard DW. Bilingual topic aspect classification with a few training examples. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2008. 203−210. [doi: 10.1145/1390334.1390371]

[23] Platt J, Toutanova K, Yih WT. Translingual document representations from discriminative projections. In: Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing. 2010. 251−261.

[24] Shi L, Mihalcea R, Tian M. Cross language text classification by model translation and semi-supervised learning. In: Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing. 2010. 1057−1067.

[25] Ng M, Wu QY, Ye YM. Co-Transfer learning via joint transition probability graph based method. In: Proc. of the 1st Int'l Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining. 2012. 1−8. [doi: 10.1145/2351333.2351334]

[26] Ng M, Wu QY, Ye YM. Co-Transfer learning using coupled Markov chains with restart. IEEE Intelligent Systems, 2014,29(4): 26−33. [doi: 10.1109/MIS.2013.32]

[27] Yang L, Jing LP, Yu J. Heterogeneous co-transfer spectral clustering. In: Proc. of the Rough Sets and Knowledge Technology. 2014. 352−363. [doi: 10.1007/978-3-319-11740-9_33]

[28] Liu JL, Wang C, Gao J, Han JW. Multi-View clustering via joint nonnegative matrix factorization. In: Proc. of the SIAM Int'l Conf. on Data Mining. 2013. 252−260.

[29] Kummar A, Rai P, Daume H. Co-Regularized multi-view spectral clustering. In: Proc. of the Advances in Neural Information Processing Systems. 2011. 1413−1421.

[30] Harel M, Mannor S. Learning from multiple outlooks. In: Proc. of the Int'l Conf. on Machine Learning. 2011. 401−408.

[31] Wang C, Mahadevan S. Heterogeneous domain adaptation using manifold alignment. In: Proc. of Int'l Joint Conf. on Artificial Intelligence. 2011. 1541−1546. [doi: 10.5591/978-1-57735-516-8/IJCAI11-259]

[32] Duan LX, Xu D, Tsang IW. Learning with augmented features for heterogeneous domain adaptation. In: Proc. of the Int'l Conf. on Machine Learning. 2012. 711−718.

[33] Li W, Duan LX, Xu D, Tsang IW. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014,36(6):1134−1148. [doi: 10.1109/TPAMI.2013.167]

[34] Zhou JT, Tsang IW, Pan SJ, Tan MK. Heterogeneous domain adaptation for multiple classes. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2014. 1095−1103.

[35] Wei FM, Zhang JP, Yan C, Yang J. CFSFP: Transfer learning from long texts to the short. Applied Mathematics & Information Sciences, 2014,8(4):2033−2044.

[36] Dai WY, Xue GR, Yang Q, Yu Y. Co-Clustering based classification for out-of-domain documents. In: Proc. of the 13th ACM Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007. 210−219. [doi: 10.1145/1281192.1281218]

[37] Dai WY, Yang Q, Xue GR, Yu Y. Self-Taught clustering. In: Proc. of the 24th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2008. 200−207. [doi: 10.1145/1390156.1390182]

[38] Samanta S, Tirumarai SA, Das S. Cross-Domain clustering performed by transfer of knowledge across domains. In: Proc. of the 2013 IEEE 4th National Conf. on Computer Vision, Pattern Recognition, Image Processing and Graphics. 2013. 1−4. [doi: 10.1109/NCVPRIPG.2013.6776213]

[39] Shi X, Liu Q, Fan W, Yu PS, Zhu RX. Transfer learning on heterogeneous feature spaces via spectral transformation. In: Proc. of the Int'l Conf. on Data Mining. 2010. 1049−1054. [doi: 10.1109/ICDM.2010.65]

[40] Shi X, Liu Q, Fan W, Yu PS. Transfer across completely different feature spaces via spectral embedding, IEEE Trans. on Knowledge and Data Engineering, 2013,24(4):906−918. [doi: 10.1109/TKDE.2011.252]

[41] Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika, 1936,1(3):211−218. [doi: 10.1007/BF02288367]

[42] Gower JC, Dijksterhuis GB. Procrustes Problems. Oxford University Press, 2004.

[43] Rokhlin V, Szlam A, Tygert M. A randomized algorithm for principal component analysis. SIAM Journal on Matrix Analysis and Applications, 2009,31(3):1100−1124. [doi: 10.1137/080736417]

[44] Charnes A, Cooper WW. Chance-Constrained programming. Management Science, 1959,6(1):73−79. [doi: 10.1287/mnsc.6.1.73]

[45] Doppa JR, Yu J, Tadepalli P, Getoor L. Chance-Constrained programs for link prediction. In: Proc. of the Advances in Neural Information Processing Systems Workshop on Analyzing Networks and Learning with Graphs. 2009. 1−8.

[46] Shapiro A, Dentcheva D, Ruszczynski A. Lectures on stochastic programming: Modeling and theory. Society for Industrial and Applied Mathematics and Mathematical Programming Society, 2009.

[47] Mitzenmacher M, Upfal E. Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, 2005.

[48] Rudelson M, Vershynin R. Sampling from large matrices: An approach through geometric functional analysis. Journal of the ACM, 2007,54(4):21−40. [doi: 10.1145/1255443.1255449]

[49] Stewart GW, Sun JG. Matrix Perturbation Theory. Academic Press, 1990.

[50] Cai X, Wang H, Huang H, Ding C. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. Bioinformatics, 2012,28(12):i16−i24. [doi: 10.1093/bioinformatics/bts220]

[51] Russell B, Torralba A, Murphy K, Freeman W. Labelme: A database and Web-based tool for image annotation. International Journal of Computer Vision, 2008,77(1):157−173. [doi: 10.1007/s11263-007-0090-8]

附中文参考文献:

[2] 庄福振,何清,史忠植.迁移学习研究进展.软件学报,2015,26(1):26−39. http://www.jos.org.cn/1000-9825/4631.htm [doi: 10.13328/j.cnki.jos.004631]

杨柳(1980－),女,河北保定人,博士生,主要研究领域为机器学习,数据挖掘.

于剑(1969－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为机器学习,数据挖掘.

景丽萍(1978－),女,博士,教授,博士生导师,CCF 会员,主要研究领域为机器学习,数据挖掘.