软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):2971−2991 [doi: 10.13328/j.cnki.jos.005348]

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

面向认知的多源数据学习理论和算法研究进展*

杨 柳[1], 于 剑[2], 刘 烨[3,4], 詹德川[5]

[1](天津大学 计算机科学与技术学院,天津 300350)

[2](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)

[3](脑与认知科学国家重点实验室(中国科学院 心理研究所),北京 100101)

[4](中国科学院大学 心理学系,北京 100049)

[5](计算机软件新技术国家重点实验室(南京大学),江苏 南京 210023)

通讯作者: 于剑, E-mail: jianyu@bjtu.edu.cn

摘 要: 多源数据学习在大数据时代具有极其重要的意义.目前,多源数据学习算法研究远远超前于多源数据学习理论研究,经典的机器学习理论难以应用于多源数据学习,更难以提供多源数据学习算法在实际应用中的理论保障.从学习的最终目的是知识这一认知切入点出发,对人类学习的认知机理、机器学习的三大经典理论(计算学习理论、统计学习理论和概率图理论)以及多源数据学习算法设计这 3 个方面的研究进展进行总结,最后给出未来研究方向的思考.

关键词: 统计学习理论;模式分类;特征空间;认知心理

中图法分类号: TP181

中文引用格式: 杨柳,于剑,刘烨,詹德川.面向认知的多源数据学习理论和算法研究进展.软件学报,2017,28(11):2971−2991. http://www.jos.org.cn/1000-9825/5348.htm

英文引用格式: Yang L, Yu J, Liu Y, Zhan DC. Research progress on cognitive-oriented multi-source data learning theory and algorithm. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):2971−2991 (in Chinese). http://www.jos.org.cn/1000-9825/5348.htm

Research Progress on Cognitive-Oriented Multi-Source Data Learning Theory and Algorithm

YANG Liu[1], YU Jian[2], LIU Ye[3,4], ZHAN De-Chuan[5]

[1](School of Computer Science and Technology, Tianjin University, Tianjin 300350, China)

[2](Beijing Key Laboratory of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing 100044, China)

[3](State Key Laboratory of Brain and Cognitive Science (Institute of Psychology, The Chinese Academy of Sciences), Beijing 100101, China)

[4](Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China)

[5](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: In the age of big data, learning from multi-source data plays an important role in many real applications. To date, plenty of multi-source data learning algorithms have been proposed, however, they pay little attention to the fundamental theoretic laws. Meanwhile, it is hard for the classical machine learning theories to govern all learning systems, and to further provide a theoretical support for multi-source learning algorithms. From the perspective of knowledge acquisition through learning, a survey is given on the research progress of three key problems: the human cognitive mechanism, three classical machine learning theories (such as computational

* 基金项目: 国家自然科学基金(61632004, 61773198, 61702358)
Foundation item: National Natural Science Foundation of China (61632004, 61773198, 61702358)
本文由复杂环境下的机器学习研究专刊特约编辑胡清华教授、张道强教授、张长水教授推荐.
收稿时间: 2017-05-14; 修改时间: 2017-06-16; 采用时间: 2017-08-23

learning theory, statistical learning theory, and probabilistic graphical model), and the design of multi-source learning algorithms. Future theoretical research issues of multi-source data learning also presented and investigated.

Key words: statistical learning theory; pattern classification; featurespace; coginitive psychology

随着人们收集、存储、传输、管理数据的能力日益提高,各行各业已经从多种渠道/信道收集并积累了大量的数据资源.如《Nature》于 2008 年 9 月出版了一期 Big Data 专刊[1],《Science》在 2011 年 2 月推出了 Dealing with Data 专刊[2],列举了在生物信息、交通运输、金融、互联网等多领域,多源数据在科学研究中扮演着越来越重要的角色.大数据的特点之一是混杂性(variety),数据的混杂性和数据的采集源十分相关,正是由于实际应用数据来源于多种渠道,使得对复杂对象复杂应用的描述具有多源性.2016 年 1 月,《Nature》以封面论文的形式,介绍了 Google DeepMind 开发的人工智能程序 AlphaGo 击败欧洲围棋冠军樊麾,引起了公众的热议,该程序使用了深度学习,其数据也是多源的[3].DeepMind 的创始人 DemisHassabis 在 AAAI 2016 大会报告上明确指出,多源数据学习技术是未来 DeepMind 和谷歌的发力方向[4].

在多源数据学习算法设计和应用方面,机器学习研究者已经进行了先驱性的工作:将监督多源学习技术应用于人脸识别领域对人脸表情进行识别[5];应用于自然语言处理领域对多语种文本进行分析,如国际学术会议(ACL)每年都接受多篇相关工作;考虑到和单源情境一样,在多源情境下,标记样本的数量也可能比较少,可以引入半监督学习技术并且取得了良好的效果[6].另外,在图像聚类[7]、视频分割[8]、事件识别[9]、视频检索[10]等方面也有与多源信息处理相关的研究.

相对于机器学习算法(包括单源和多源)的研究,机器学习理论研究的进展要艰难得多.机器学习传统上分为监督学习与无监督学习.对于监督学习,最经典的学习理论包括计算学习理论和统计学习理论.2010 年图灵奖得主 Leslie Valiant 于 1984 年开创了计算学习理论,即概率近似正确(probably approximately correct,简称 PAC)学习理论[11].Vapnik 在 20 世纪 70 年代初步提出了以 VC 维为重要概念的统计学习理论,并于 1995 年设计出著名的支持向量机方法[12].Vapnik 提出的 VC 维理论可以将统计机器学习理论与 PAC 学习理论进行有效整合,对于监督学习算法给出了许多深刻的理论结果.对于无监督学习中的典型学习范式之一的聚类分析,2003 年,美国三院院士 Kleinberg 给出了一个极为悲观的理论结论[13],即著名的聚类不可能性定理.因此,正如 Valiant 在其 2013 年发表的文献[14]中所说,PAC 理论是关于监督学习的理论模型,而对无监督学习建立类似的理论尚未成功;图灵奖得主 Pearl 系统发展的概率图理论[15]可以处理能够用概率解释的机器学习算法,对于不是基于概率的机器学习算法(如模糊逻辑等)则无能为力.

图灵奖得主 Valiant 在 2016 年 1 月份接受《Quanta Magazine》期刊记者Pavlas采访时明确指出,机器学习与人类学习机理是一致的[16].为了让机器能够尽可能近似地模拟人类的学习能力,从多个数据源同时获取信息进行学习是必需的.无论是人类学习还是机器学习,学习的终极目的都是从有限的数据中学习到知识.而知识的基本单位是概念.毫无疑问,人类通过多源感知来认识世界.非常年幼的儿童通过视、听等多感觉通道,已经可以非常自如地进行类别学习获得概念,在学龄前已经具有非常好的类别学习能力,并掌握了大量概念[17].甚至昆虫、鸟类、哺乳动物也具有跨通道的类别学习能力[18],而且某些鸟类和哺乳类动物可以掌握抽象概念[19].另外,多模态的知觉学习可以促进单模态的知觉学习[20].目前,认知心理学提出的知觉组织原则、概念表征理论(如基于相似性的原型理论、样例理论)已经为机器学习算法提供了有益的认知机理.一个自然的问题是,人类强大而灵活的多源数据学习能力及其机理是什么?其认知机理是否能够为机器学习的理论和算法提供启迪,促进机器的学习能力?Jordan 等人于 2015 年在《Science》上发表的文献[21]中提到,机器学习的研究问题之一是发现统管机器、人、和生物的学习规律.

本文首先研究人类的认知机理,然后分析人类和机器学习的认知机理的相似性,研究机器学习理论,最后在机器学习理论的指导下,设计多源学习算法,如图 1 所示.众所周知,知识由各种概念组成,概念是构成人类知识世界的最小单元.人们必须借助概念才能理解世界和认知世界.认知科学总是假设概念在人的心智中是存在的,概念在人心智中的表示称为认知表示.当人们心中有了概念时,必然使用这些概念对世界上的对象进行归类.在

本文中,类与概念具有相同的语义,实际上,模式与类也有同样的语义.归类是人类的一项最重要而且最基本的认知能力,人类是依赖于相似性将对象归类的.从直观上说,人们之所以将某个对象归为某个类,是因为该对象最像该类;反之,如果某个对象最像某个类,则该对象应该归为该类.

Fig.1 Framework of the main contents

图 1 主要内容的框架图

机器学习的基本任务是获取知识,因此最终输出结果为知识(可以是显性知识,也可以是隐性知识).因此,机器学习研究如何从数据中学习概念,通过学习也拥有类似人类的归类能力.机器学习和人类的学习机理是一致的[16],主要是通过相似性对对象进行归类的,归类遵循的原则应该是:归哪类,像哪类;像哪类,归哪类.目前,关于人类的概念表示研究结果适合于机器学习,通过学习人类的认知机理,可以指导机器具有更好的学习能力.

现在机器学习问题众多,虽然各种学习算法层出不穷,但它们都有一定的理论支撑.例如,对于分类算法(包括单源和多源)来说,人们通常期望学到的类标预测函数在测试数据上的性能能够满足需求,即学习算法的泛化能力要好.泛化错误率反映了学习方法的泛化能力,学习的类表示具有更小的泛化错误率说明该类表示更有效.泛化错误率不要太大,最好控制在实际应用中可以容忍的泛化错误率以内.如果泛化错误率大于容忍的错误率,这样的情形在概率意义下发生的可能性也是取决于实际应用的需要.这就是 PAC 理论,它为分类算法提供了理论保障.本文将针对人类的认知机理、机器学习理论和多源学习算法展开系统的综述,并探索未来的研究方向.

1 人类多源学习的认知机理

概念是构成人类知识世界的最小单元,人们必须借助概念才能理解世界和认知世界.现代认知科学提出的概念内涵表示理论有原型理论、样例理论和知识理论.原型理论认为,一个概念可由一个原型来表示,一个原型可以是一个实际的或者虚拟的对象样例,通常假设为概念的最理想代表.样例理论认为,概念不可能由一个对象样例来代表,应该由多个样例来表示.更进一步地,认知科学家发现,各种人类文明中,单一概念不可能独立于特定的文明之外而存在,由此形成了概念的知识理论.在知识理论中,认为概念是特定知识框架(文明)的一个组成部分.但是无论如何,认知科学总是假设概念在人的心智中是存在的.人们可以使用这些概念对世界上的对象进行归类.归类是人类的一项最重要而且最基本的认知能力,归类正确与否,明确显示了人是否掌握了与该类对应的概念.

对人类的认知系统而言,通过视觉、听觉、触觉、嗅觉、本体感觉这 5 种感觉通路来学习知识,积累经验.在知觉学习的基础上形成类别,并以概念的表征形式储存在记忆中.人类天生具有多源数据学习的能力,具有多个感知系统的优势显而易见,每种感觉在不同的场景下发挥最优的效用,多种感觉联合起来就增大了检测和识别有意义的事件或物体的可能性[22].目前,有关人类的多模态信息融合研究主要集中在感觉信息的融合以及注意在信息融合中的作用,而在高水平的认知阶段如何进行多模态的信息融合的研究较少.研究如何根据人类多

源认知的机理指导机器学习多源数据,是非常有必要的.因此,本节将介绍人类多源学习的认知机理.知觉、记忆、判断是人类认知过程的 3 个基本阶段[23],如图 2 所示.本节将按照这 3 个阶段,回顾国内外关于类别知觉、概念表征和类别学习的多模态信息融合的研究现状.

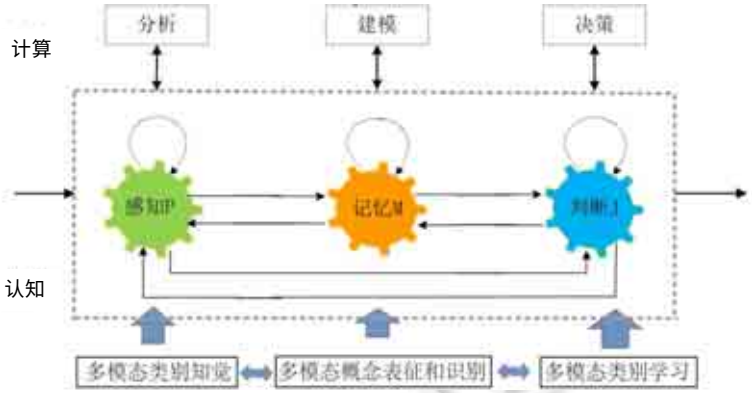

Fig.2 Framework of cognitive mechanism based on perception, memory, and judgment

图 2 基于感知、记忆和判别的认知机理框架图

1.1 多模态的类别知觉

类别学习是人类重要的认知功能,类别知识的获得可以简化人们的认知过程,在人们完成许多认知任务的过程中起着重要作用.类别知觉表征是类别学习的第 1 步,通过对类别特征的感知和记忆形成关于类别结构的表征[24],并影响后续类别加工[25].类别决策是类别学习的第 2 步,在类别表征的基础上对样例做出相应的归类决策.类别决策中存在着明显的类别知觉效应,即与类内样例的辨别相比,类间样例的辨别正确率更高,反应时间更短,这是类别知觉最重要的特点[26].面孔属于自然类别,可以天然地从多个维度分类,例如表情.人们对高兴、悲伤、愤怒、恐惧、厌恶、惊讶这 6 种基本表情的知觉过程中都存在类别知觉效应[27].

人们对类别的知觉表征包括基于样例的特征维度分析和基于样例整体相似性的整体加工[28].相似性在分类的原型理论、样例理论、规则理论中都起着重要的作用.基于样例整体相似性的加工是类别学习的一种重要表征方式.儿童在类别的知觉表征中更容易基于整体相似性进行加工,较少关注到特定的特征维度.当知道某个对象有某个特征,要求判断另一个对象是否也具有该特征时,儿童在很大程度上把注意力放在两个对象的知觉的相似性上[29].类别知觉表征方式与类别结构有密切关系[30].在类别学习中,当各个特征维度之间独立分离,特征之间难以整合在一起时,类别结构的表征和识别通常是建立规则(rule-based)[30];当特征维度之间能够整合,类别结构的表征通常是相对抽象的样例整体相似性.表情的类别知觉研究通常通过人为操纵刺激材料的物理差异来获得类内和类间刺激,以评估类别界限.然而在表情的类别知觉中,表情的类别界限是异于物理刺激的,因此,表情的分类可能基于相对抽象的整体相似性.

在日常生活中,面对复杂的外部环境,我们接触的并非只是单一感觉通道的信息,大脑能够通过获取不同感觉通道(例如视觉、听觉、嗅觉、触觉等)的信息对复杂的外部环境进行感知.不同的感觉信息可被大脑有效地合并为统一、连贯、稳定的知觉信息[31],在人脑中按照类别表征,而不是按照连续的物理信号表征的.这种整合的类别特征不同于初级的物理特征,通常是一种抽象特征,即类别内的个体具有某种抽象的整体相似性.通过多种感知模式的信息传入,对各个感觉通道信息综合分析比较并分配权重,可以有效弥补信息传入不足或者环境干扰等因素的影响,最终达到对事物最优化的感知[31].例如,视觉在刺激的空间定位上具有优势,而听觉则在时间定位方面更强,多重感觉则可以提供互补信息以最优化信息加工[32].信息在时间和空间上的接近或匹配,是构成视听整合加工优势的必要条件,在相近时间或空间位置出现的不同通道的信息容易被整合在一起[33],进而形成对客体或事件的知觉.

近年来有关多感觉通道信息整合的神经机制研究发现,很多以往被认为是单一感觉通道信息加工的皮层区实际上可能是多感觉通道的,如听皮层对触觉和视觉刺激均存在反应;单纯听觉刺激可以激活初级视皮层,单纯视觉刺激也可以激活初级听皮层.越来越多的对感觉通道间交互的研究挑战了传统所认为的初级感觉皮层是功能独立的和感觉特异的观点[34].多感觉整合加工的神经机制是一个涉及多个脑区的网络,大脑中不仅具有特异性感觉加工脑区,而且具有专门负责多感觉信息整合加工的功能特异性脑区[35].研究结果也表明,早期生活的学习和可塑性是高度多重感觉的,多感觉通道的学习比单一感觉通道的学习更有效率[36].这也说明多通道的类别学习也可能具有某种优势.这种优势的前提也是不同感觉刺激之间的匹配或表征相似性.Hahn 等人[37]提出了表征变形理论,认为类别个体相似性可以通过客体的表征间相互转换和变形来表示客体的相似性.但如果类别的多通道信息是相对独立分离的特征维度,那么多通道信息整合也可能干扰人们对客体的类别知觉.关于维度特征相似性的研究通常假设人们在分类中依据的类别标准有清晰、明确的界限,但特征维度融合的整体相似性可能是具有模糊性的.因此,多通道信息的融合如何形成整体相似性表征、在多通道类别知觉起什么作用,仍然有待深入研究.

1.2 多模态的概念表征

近年来,多模态的感觉运动信息在人类概念表征中的作用得到越来越多证据的支持,有关物体识别的研究表明,与物体相关的感觉运动信息(sensory-motor information)可能参与了物体的表征和识别过程.首先,神经心理学的研究发现,可操作性(manipulability)很可能是造成物体识别的范畴特异性现象的重要原因[38];其次,针对正常被试的行为研究表明,操作动作表征可以在物体加工过程中自动激活,并且会影响随后的物体识别和动作反应[39].另外,神经成像的研究也证实:在识别可操作物体时,与操作动作相关的运动脑区会自动激活[40].同时,研究表明,存在两种不同的操作动作:一种是结构性操作(structural manipulation),另一种是功能性操作(functional manipulation)[41].结构性操作仅指“拿起并移动物体”,如拿起并移动锤子;功能性操作则针对使用物体功能的操作,如使用钢笔.研究结果表明,两种操作动作表征在功能和神经基础上都存在差异.

基于上述研究,研究者开始以不同形式提出,概念表征本质上基于感觉和运动系统,即具身认知(embodied cognition)或及地认知假设(grounded cognition)[42].概念表征的具身认知理论[42]强调感觉和运动系统直接参与了概念的加工和表征,因而感觉和运动信息也是概念构成的重要部分.因此,考察与物体相关的操作动作信息在物体识别中的作用具有重要的理论意义和应用价值.首先,相关研究将为物体知识是如何表征在人脑中这一关键理论问题提供实证研究的证据.长久以来,认知心理学的研究都认为知识是以抽象的符号表征在头脑中,而近年来的物体识别研究结果表明,与感觉、运动通道相关的信息会直接参与物体的表征和识别,尤其是操作动作信息.其次,传统的物体识别聚焦于物体本身的感觉和知觉属性,但是在识别与人交互的物体时,人操作物体的动作信息也会为物体分类识别提供一定的帮助.这种多模态的信息将为机器学习的理论和算法提供新的视角.

1.3 多模态的类别学习

类别学习是人类智能的一种基本认知过程,它是指为将事物分为不同类别而形成特定记忆痕迹的过程[43].类别学习反映了认知加工中的抽象过程,它使得人们可以举一反三、触类旁通.也就是人们在加工具体的视觉、听觉等多通道感觉信息时,会将外界刺激信息抽象为不同的类别,从而使人们可以对未见过的物体或情境做出快速反应.近年来,根据类别规则是否可以用言语来描述,有研究者将类别学习分为外显类别学习和内隐类别学习[44].该理论认为:外显类别学习依赖于工作记忆和注意,通过假设检验过程获得有关相关类别的概念知识,前扣带回和前额叶皮层是外显类别学习的关键脑区;内隐类别学习不依赖于工作记忆和注意,通过对物体表征的熟悉度,或者相关的程序记忆,或者记忆中的样例等来进行分类,纹状体在内隐类别学习中起着关键作用.不过,目前有关多通道的类别知识在人脑中是如何表征的仍不清楚.

综上所述,研究者已经发现了许多人类学习的认知机理.例如,对于知觉学习,人们发现非目标相似性损害知觉学习,但目标相似性促进知觉学习的迁移;对于概念表征,人们发现信息不仅以抽象符号的形式存储,也可能存在模态特异性存储;对于类别学习,人们发现相似性和特征显著性在类别学习中起着重要作用.人类是依赖

于相似性将对象归类的[45].有认知实验结果表明,不仅儿童是基于相似性表示类的,甚至基于相似性的类表示在发育过程中也是默认设置的[46].归类遵循的原则应该是:归哪类,像哪类;像哪类,归哪类.也就是说,人们心里想的归类结果要与客观的归类结果一致.机器学习的基本任务是获取知识,研究如何从数据中学习概念,希望通过学习也拥有类似于人类的归类能力.机器学习和人类的学习机理都是主要通过相似性对对象进行归类的.目前,关于人类的概念表示研究结果适合于机器学习,通过学习人类的认知机理,可以指导机器具有更好的学习能力.尽管多模态学习是人类学习的基本特性,但当前研究大多探讨单模态的学习机理,并且这些学习机理是否适合机器学习还有待研究.

2 机器学习理论

学习理论的研究对机器学习的发展起着重要的支撑和指导作用[47].主流的学习理论大致可以分为 3 个方向:图灵奖得主 Valiant 提出的计算学习理论、Vapnik 提出的统计学习理论以及图灵奖得主 Pearl 系统发展的概率图学习理论.前两种理论主要应用于监督学习,下面我们将分别加以论述.

2.1 计算学习理论

机器学习中的一个基本问题是:一个学习问题是否是可学习的,即什么样的任务对学习是可行的,需要多少训练样本才能学习获得最优分类器,其样本复杂度是多少?计算学习理论[11,14]对这一系列问题给出了完整的回答,又称为 PAC 学习理论或概率近似正确理论.传统的 PAC 学习理论关注一致分类器、不一致分类器中一般概念的可学习性,回答了需要多少训练样本才能确保得到较低的错误率这一问题.从这个角度来看,PAC 理论主要是为有监督学习提供理论支撑.

监督学习的目标是,能够从合理数量的训练数据中通过合理的计算学习获取知识.然而机器学习面临的现实情况是:除非对每个可能的数据进行训练,否则总会存在多个假设,使得真实错误率不为 0,即分类器无法保证和目标函数完全一致.另外,由于学习算法事先并不知道概念类的真实存在,因此假设概念集合和目标概念集合通常是不同的.PAC 理论通过放松一致性条件来弱化对分类器的要求.具体来讲,对每个假设和目标函数,不要求有监督学习方法的输出错误率为 0,只要求错误率被限制在某常数ε范围内,ε可为任意小;同时,不要求分类器对所有任意抽取的数据都能成功地预测,只要求其失败概率被限定在某个常数δ的范围内,δ可取任意小.

在 PAC 理论中,样本复杂度的上下界一直都没有很好地匹配,总存在 $\log(1/\varepsilon)$的间隙.已有不少学者根据 PAC 理论研究不同的监督学习算法所需要的训练样本数量.Ehrenfeucht 等人[48]给出的每种 PAC 算法的样本复杂度下界为$\Omega((1/\varepsilon)(d+\log(1/\delta)))$,而 Blumer 等人[49]给出的一致性算法的样本复杂度上界为$O((1/\varepsilon)(d\log(1/\varepsilon)+\log(1/\delta)))$.Auer 等人[50]证明了这个上界对于一些一致性算法是非常紧的,他们证明的最优因子与 $\log(1/\varepsilon)$不同,因此没有办法显示任意的一致性算法是最优的.也有不少学者关注是否存在一个“最佳 PAC 算法”的问题.Warmuth[51]针对是否存在一个样本复杂度下界为$\Omega((1/\varepsilon)(d+\log(1/\delta)))$的最优 PAC 算法进行了探讨,Haussler 等人[52]采用 1-inclusion 图算法对预测进行了分析.Hanneke[53]展示了上界中额外的因子 $\log(1/\varepsilon)$,可以由与概念类关联的不一致系数的对数来代替.随后,Darnstadt[54]证明了对于交叉封闭的类,闭包算法是一个最优的 PAC 算法.Simon[55]指出:每个一致性算法,甚至是子最优的算法,都会包括一个 PAC 算法家族$(l_K)_{K\geq 1}$,所需的训练样例数目超过普通因子 $l_K(1/\varepsilon)$的下界(l_K表示第 K 次迭代的对数),该算法家族和最优 PAC 算法是非常接近的.Simon[55]提出了一种基于多数投票的分类方法.该算法复杂度的上界从对数因子降为一个缓慢增长的边界.Hanneke[56]提出了一种基于多数投票的递归算法,边界中彻底消除了对数因子.

最初的 PAC 学习理论中,只证明了对于有限假设空间情况下 PAC 的有效性,然而这种有限假设空间很难满足实际应用的需求.并且 PAC 只能为有监督学习提供理论支撑,提出 PAC 理论的 Valiant 在 2013 年指出,PAC 学习还不能应用到无监督学习中[14].

2.2 统计学习理论

现实学习任务所面临的假设空间通常是无限的,Vapnik 等人[57]定义了 VC 熵和 VC 维,成为研究小样本情况

下统计规律及统计决策问题的核心概念.并以此为基础,相继提出了著名的大数定律和机器学习结构风险最小化原则[58].Vapnik 将机器学习问题描述成通过样本数据寻找依赖关系的问题:在假设观测的样本服从一定的未知分布规律并且不估计分布规律的前提下,机器学习研究基于训练样本的目标函数和基于分布的函数之间的关系.由此可见,该理论也是为监督学习服务的.同时,大数定律确定的非渐进界过分保守,更不幸的是,由于数学原理复杂,长期没有找到将这些理论付诸实践的好方法,以致这些成果在相当长的一段时间没有得到重视.直到根据支持向量机(support vector machine,简称 SVM)发展出间隔(margin)理论之后[12],统计学习理论的框架才基本完成.

1998 年,哥德尔奖得主 Schapire 教授等人[59]提出了著名的间隔理论,用于理解 Boosting 理论研究中最重要的问题,即 AdaBoost 算法为何不易陷入过拟合.间隔理论对这个问题给出了合理的解释,其核心是,Boosting 方法的泛化界能够通过间隔来刻画.Berkeley 权威统计学家 Breiman 教授[60]改进了 Schapire 的理论结果,得出了一个更紧的上界,然而实验发现,这两种理论与实验结果自相矛盾.Wang 等人[61]对 Boosting 间隔理论进行了较为深入的研究,得到了一系列新的理论结果并设计了相关算法.Gao 等人[62]对间隔理论做了进一步的深入研究,通过证明新的经验 Bernstein 界,给出 Boosting 间隔分布理论,从而有效支撑了 Boosting 间隔理论.在此基础上,Zhang 等人[63]提出了优化间隔分布的支持向量机方法.

与经典 PAC 理论假设空间有限不同,VC 维的泛化误差边界是分布无关、数据独立的[64],即,对任何数据分布都成立.这使得基于 VC 维的可学习性分析结果具有一定的普适性.但是由于没有考虑数据自身,基于 VC 维得到的泛化误差边界通常比较松.由于假设空间的大小是无限的,因此无法使用假设空间的大小来表示其复杂度.VC 维在假设空间上引入额外的度量,并且 VC 维与样本数据集的分布是无关,因此在分析学习模型的泛化能力时显得过于保守[65].

研究者在对统计学习理论的深入研究中,注意到数据依赖复杂性度量的重要性[66],将 Rademacher 复杂度引入到了学习模型的泛化能力分析研究中[67,68].Rademacher 复杂度通过测量一个假设集能够拟合随机噪点的程度来表达假设空间的复杂度.与 VC 维不同的是,它不需要引入额外度量,并且在一定程度上考虑了数据分布,并基于 Rademacher 复杂度进一步研究关于函数空间的泛化误差边界[69].Bartlett 等人[70]指出,对学习模型泛化能力起关键作用的是具有较小方差的函数所构成假设空间的子空间,而不是整个假设空间中的函数,给出了局部 Rademacher 复杂度[70,71]的概念.Zhivotovskiy[72]给出了一种替代局部化的思路,提出了基于固定点局部经验熵的复杂度度量方法.研究者还通过 Rademacher 复杂度分析了多标签学习算法[73]、多模态度量学习算法[74]的泛化性能以及深度神经网络中 Dropout 的泛化界[75]等.

VC 维、Margin 理论和 Rademacher 复杂度已经与计算学习理论进行了有效的融合,解决了计算学习理论最初只对有限假设空间问题有效的弊端.然而统计学习理论和计算学习理论一样,主要分析有监督学习算法,对无监督学习无法提供有效的理论支撑.

2.3 概率图理论

计算学习理论和统计学习理论主要用来支撑有监督学习.由 Pearl 开发出来的用图来表示变量概率依赖关系的概率图理论,为多种机器学习方法(有监督、无监督等)提供了较为宽泛的理论支撑,Pearl 也因此而获得了图灵奖.概率图理论主要通过结合概率论与图论的知识,利用图来表示与模型有关的变量的联合概率分布.概率图理论分为概率图模型表示理论、概率图模型学习理论和概率图模型推理理论[76,77],近年来已成为不确定性推理的研究热点,在人工智能、机器学习和计算机视觉等领域具有广阔的应用前景[78].

概率图模型的表示由参数和结构两部分组成,基本的概率图模型可以大致分为两个类别:贝叶斯网络(Bayesian network)[79]和马尔可夫随机场(Markov random field)[80].其主要区别在于,前者采用有向无环图来表达因果关系,后者采用无向图(undirected graph)来表达变量间的相互作用.这种结构上的区别导致了它们在建模和推断方面的一系列微妙的差异.基于概率图模型的学习[81]分为概率网络的参数学习和结构学习算法.参数学习假设在已知网络结构的情况下,从数据中学习每个变量的概率分布.概率分布的形式一般需要预先指定,只需利用一定的策略估计概率分布的参数.

概率图模型的推理是利用联合概率分布,在已知网络结构和证据的情况下回答查询问题.概率图模型的推理方法主要有精确推理和近似推理两大类.其中,精确推理按照概率公式回答查询问题,从而得到精确的查询结果.但是,精确推理难以处理大型的复杂概率图模型.近似推理是在近似模型上推理,得到近似的结果,可以用来处理大型复杂的网络.近似推理的方法主要包括基于平均场逼近的变分推理、信念传播和蒙特卡罗采样.在指数族分布所组成的贝叶斯网络推断中,研究者引入基于平均场逼近的变分推理,通过一个容易计算的上界逼近原模型的 log partition 函数,从而有效地降低了优化的复杂度.被广泛采纳的期望最大化(expectation-maximization,简称 EM)算法[82]就属于这类方法.早期在进行树状结构统计推断时,Pearl 提出了信念传播方法[83];随后,Murphyd 等人[84]把这种算法扩展到有环的模型.Mooij 等人[85]在此基础上提出了迭代信念传播,给出理论解释并刻画出它在各种设定下的收敛条件.Wainwright 等人[86]采用混合树来逼近任意的图模型,对信念传播进行了有效的推广.Minka[87]提出了期望传播,将信念传播成功地推广至更一般的概率图.Johnson 等人[88]提出的 Walk Sum 分析成功地刻画了信念传播在高斯场上的收敛条件,这也是后来提出的多种改进型信念传播方法的理论依据.与上述基于优化的方法不同,蒙特卡罗方法通过对概率模型的随机模拟运行来收集样本,然后,通过收集到的样本来估计变量的统计特性[89].

概率图理论有很多好的性质[81],它提供了一种简单的可视化概率模型的方法,有利于设计和开发新模型;通过对图的深入研究,了解概率模型的性质;用于表示复杂的推理和学习运算,简化了数学表达.然而,概率图理论需要假设事先知道数据的分布情况,而且概率模型的估计过程是一个近似 NP 难的问题,这对实际问题来说极具挑战性,现在发展的算法大多是基于近似估计.

总而言之,现在的机器学习理论各有不足,见表 1,有必要研究一个覆盖更多机器学习算法的机器学习理论.已有不少工作向这方面努力,如 PAC-Bayes 理论[90−92],其对部分无监督学习算法(如 co-clustering 和密度估计[93,94])也进行了研究,但 PAC-Bayes 理论依然是在概率统计的意义下进行研究,其复杂度比经典的机器学习理论更高一些.客观地说,现今的 3 大经典机器学习理论远远超过了一个正常的七八岁儿童可以理解的程度,难以期望一个正常儿童使用这些优美但却复杂的学习理论来进行物体识别.但是,一个不可否认的事实是:一个正常儿童能够识别日常生活中的很多类别,具有很高的学习能力.因此,有必要深入研究人类认知的学习机理和现有机器学习算法的共性,提供一套既可以统管机器学习、人类学习甚至组织学习,又符合人类机理的学习理论.

Table 1 Theories in machine learning

表 1 机器学习理论

主要理论		主要特点	不足
计算学习理论	可能近似正确(PAC)理论[11,14,48−56]	确定了若干假设类别,判断它们能否从多项式数量的训练样例中学习得到.定义了一个对假设空间复杂度的自然度量,由它可以界定归纳学习所需的训练样本数目.	只能有限假设空间下证明 PAC 的有效性,无法应用到无限假设空间的情形.PAC 只能为有监督学习提供理论支撑.
统计学习理论	VC 维[57,58]	将机器学习问题描述成通过样本数据寻找依赖关系的问题.VC 维的泛化误差边界是分布无关、数据独立的.VC 理论和算法关系不大,它刻画的是集合的复杂程度.	该理论是为监督学习服务的
	Margin 理论[12]	Margin 理论主要刻画算法	
	Rademacher 复杂度[67−75]	通过测量一个假设集能够拟合随机噪点的程度来表达假设空间的复杂度,不需要引入额外度量,并且在一定程度上考虑了数据分布	
概率图理论	概率图模型表示理论[79,80]	用图来表示变量概率依赖关系的概率图理论,为有监督、无监督等提供了较为宽泛的理论支撑	需要假设事先知道数据的分布情况,而且概率模型的估计过程是一个近似 NP 难的问题
	概率图模型学习理论[81]	用来学习概率网络的参数和结构	
	概率图模型推理理论[82−89]	在已知网络结构和证据的情况下,回答查询问题.主要方法有精确推理和近似推理.	

3 多源数据学习

人类通过多源感知认知世界,类似地,计算机也应该通过多源数据进行知识发现.多源数据学习起源于Yarowsky 等人[95]利用多个数据源完成学习任务的工作.研究者可以在以上机器学习理论的支撑下设计各种多源学习算法,例如,概率图理论为基于概率主题的多源学习模型提供了理论保障.多源数据学习算法一般满足一致性原则和互补性原则[96],围绕传统数据挖掘任务,如特征降维、聚类、分类、半监督等,目前已涌现出众多的多源数据学习方法.这些方法往往采用 3 种方式进行多源数据学习:前期数据准备阶段融合、中期学习阶段融合、后期学习结果融合.

单纯的数据准备阶段融合是将多个数据源合并为一个大数据源.这种简单的融合模式忽略了数据源之间的冗余性,从而导致学习效果不够理想.后期学习结果融合主要有集成学习[97],通过构建并结合多个学习器来完成学习任务.然而,每个学习器的结果往往依赖于数据特征的好坏,因此,本文将重点介绍中期学习阶段,相关技术主要包括多视图学习和迁移学习,异同点见表 2.下面将分别论述.

Tabel 2 Similarities and differences of multi-view learning and transfer learning

表 2 多视图学习和迁移学习的异同点

	相同之处	区别
多视图学习[96,98]	研究的对象都是描述数据样本的多个特征集,基于一致性原则或互补性原则使学习系统具有更强的泛化能力	多视图学习面对多个数据源,不区分源和目标领域,数据在各个视图上特征的维度一般不相同,但是通常需要各个视图对应同样的样本.多视图学习更加侧重于多数据源之间的协同学习.
迁移学习[99−101]		迁移学习一般把数据分为源领域和目标领域.主要任务是从源领域向目标领域迁移知识,辅助目标领域的学习.

3.1 多视图学习

多视图是指从不同信息源搜集到的数据,可以看作是对同一事物从不同的角度或者不同途径的描述,每种描述有与之对应的属性集(即视图).多视图学习的目的是:利用多个视图的数据探索多视图之间的联系,并改善学习性能[98].根据具体的技术方法,多视图学习方法可分为[96]以协同训练(co-training)为代表的方法、多核学习方法和基于子空间学习的方法.

协同训练[102]通过不断最大化数据在不同数据源之间的一致性,充分发挥不同数据源的优势.在此基础上,研究者又提出了许多新的算法和变种:基于生成模型的 EM 参数估计方法[103]、贝叶斯无向图模型以及协同训练核的高斯过程分类器[104];基于图的多视图谱聚类方法[105];对单视图构造多个分类器来提高泛化性能的方法 Tri-training[106]和 co-forest 算法[107];通过 RKHS 理论的联合正则化(co-regularization)学习方法[108];两个视图训练直推式的 SVM 分类器方法[109];co-labeling 的多视图弱标记学习方法[110].协同训练的方法是多个视图后期结合的方式,在每一个视图上分别训练分类器,不同视图被看作是相互独立的,强制不同视图在输出上保持一致.

多核学习是选用一组合适的核以及核结合方式进行多视图学习[111].最常用的核结合方法包括线性与非线性方式.线性结合一般采用直接相加核和加权相加核将多个核结合到一起,非线性结合的方式有指数幂方式或者能量方式.研究者通过半正定问题[112]、二阶锥规划问题[113]、半无限线性规划问题[114]、自适应的 l_2 范数正则项[115]、group-Lasso 分组结构[116]来进行多核学习.多核学习的方法采用多视图中期结合的方式,在每个视图上分别计算独立的核,然后将多个核结合起来.

基于子空间学习的方法将多个视图直接结合到一起,然后利用潜在子空间进行学习.该方式是寻求多个视图共享的一个潜在子空间.典型相关分析(canonical correlation analysis,简称 CCA)[117]和非线性的 KCCA[118]已经成为对多视图数据进行子空间学习的基本工具,可以进行多视图聚类[119]和回归[120].有研究者进行了 KCCA 有限样本的统计分析[121]、分析了 KCCA 的收敛率[122].基于子空间学习的方法还有:基于概率生成模型的有 Corr-LDA 模型[123]、基于 Fisher 判别分析[124]、基于子空间的偏多视图聚类方法[125]、部分共享隐因子学习模型[126].Farquhar 等人提出的 SVM-2K 算法[127]将 KCCA 与支持向量机组合到一个优化问题中,并且证明了该算法的泛化误差界.Yin 等人[128]提出了利用不完整视图的数据学习子空间的方法.Xu 等人提出了大间隔多视图信

息瓶颈算法[129]和多视图完好无损空间方法[130],探索多视图的内蕴子空间.

近年来,一些研究者把深度学习引入到多视图学习中,主要采用的是共享子空间方法[131−139].Ngiam 等人[131]提出了基于自编码器的多模态深度模型来学习深层共享表示,Srivastava 等人[132,133]提出了基于 RBM 的多模态深度模型来学习多源数据的深层共享表示.Su 等人[134]和 Elhoseiny 等人[135]提出了 CNN 多视图学习模型.Andrew 等人[136]和 Benton 等人[137]提出了深度 CCA 模型,但是基于 CCA 扩展的模型,只能应用到两个视图的数据上,在具体应用中受到一定的限制.虽然深度学习在多视图学习中取得了一定的成功,但是在理论方面还需要进一步研究.主流的深度学习方法可解释性比较差,对于相同的输入,不同人由于参数设置的不同,会得到差别较大的结果,因此对于使用者来说,调参是一个大问题.

研究者对多视图学习的方法进行了理论的分析.Sun 等人[92]利用 PAC-Bays 对多视图学习进行了分析.对于协同训练方法,Du 等人[140]提出了验证充足性假设和独立性假设是否满足的方法.然而,这两个条件在实际问题中很难满足,为了使协同训练更适用于实际问题,研究了 PAC 学习的协同训练框架[141]、弱依懒性假设[142]、膨胀性假设[143]、视图间大差异性理论[144]等.Sindhwani 等人[145]给出了协同正则算法的 Rademacher 复杂度.为了保证多核学习的学习性能,一些研究者对多核学习的边界进行了深入探讨,展示了给定 k 个核的错误边界[112]、多核学习算法的 Rademacher 界[146],采用候选核集合的度量熵积分和伪维度估计了经验 Rademacher chaos 复杂度[147]、研究了 l_p 范数多核学习的局部 Rademacher 复杂度上界[148].对于子空间方法的理论研究,Xu 等人通过研究多视图特征的一致性,降低了 Rademacher 复杂度[129];证明了 Rademacher 泛化误差界可以通过不同视角之间的互补性得到改善[130].这些理论研究为多视图方法的应用与发展提供了较好的理论保障.

多视图学习的方法总结见表 3.多视图学习面对多个数据源,侧重于多数据源之间的协同学习.通常情况下,多视图学习的假设是,不同视图上的数据是完整且对应的,然而这个假设通常不能成立.迁移学习则没有数据一一对应的约束,一般把数据分为源领域和目标领域,主要是从源领域向目标领域迁移知识.

Tabel 3 Methods of multi-view learning

表 3 多视图学习方法

主要方法	描述	代表性工作
协同训练方法	不断最大化数据在不同数据源之间的一致性,充分发挥不同数据源的优势	协同训练方法[102]、EM参数估计方法[103]、协同训练核的高斯过程分类器[104]、基于图的多视图谱聚类方法[105]、Tri-training[106]、co-forest算法[107]、RKHS的联合正则化学习方法[108]、直推式的SVM分类器方法[109]、co-labeling的多视图弱标记学习方法[110]
多核学习方法	选用一组合适的核以及核结合方式进行多视图学习	多核学习方法[111]、通过半正定问题[112]、二阶锥规划问题[113]、半无限线性规划问题[114]、自适应的l_2范数正则项[115]、group-Lasso分组结构[116]进行多核学习的方法
子空间学习方法	将多个视图直接结合到一起,然后利用潜在子空间进行学习	典型相关分析CCA[117]、非线性的KCCA[118]、基于概率生成模型的Corr-LDA模型[123]、基于Fisher判别分析[124]、基于子空间的偏多视图聚类方法[125]、部分共享隐因子学习模型[126]、基于深度学习的方法[131−139]

3.2 迁移学习

Woodworth 等人在 1901 年从心理学和教育学的角度提出了学习迁移的理论[149].从人类智能的角度来看,人类可以在不同领域之间和不同问题之间进行相关知识的迁移或转化.我们希望机器可以在搜集到的多种数据源之间,像人类一样进行知识迁移或转化,这就是机器学习领域中的迁移学习,或称归纳迁移、领域适配[150].对迁移学习的研究始于 1995 年 NIPS-95 关于“学会学习(learning to learn)”的专题研讨会[151].迁移学习强调的是,在不同但是相关的领域、任务和分布之间进行知识的迁移.

按照源领域和目标领域的特征空间是否相同,迁移学习可分为同构空间下的迁移学习和异构空间下的迁移学习.同构空间下的迁移学习是指源领域和目标领域的特征空间相同,代表性工作有基于主题的隐含语义分析算法[152]、谱分析算法[153]、基于流形结构的算法[154]、潜变量的核空间模型[155]、自学习聚类算法[156]、Tradaboosting 算法[157]等.异构空间下的迁移学习[158]也称为翻译学习,在两个完全不同的特征空间下,解决目标领域和源领域的学习任务,其主要任务是构建异构特征空间之间的关联关系.主要模型有风险最小化框架[158]、

概率隐含语义分析方法[159,160]、协同矩阵分解技术[161]、积极迁移学习框架[162].

根据源领域和目标领域是否有标注样本[99],可以把迁移学习分为归纳迁移学习、直推式迁移学习和无监督迁移学习.归纳迁移学习[157,163−166]的目标领域中有少量标注样本,根据源领域中是否有标注样本,还可以把归纳迁移学习划分成多任务学习(源领域中有标注数据)、自学习(源领域中没有标注数据).直推式迁移学习[167−170]是只有源领域中有标签样本,源领域和目标领域的数据相关但不相同,两个领域的任务相同.无监督迁移学习[156,171]处理的是源领域和目标领域中都没有标签数据的问题.

按照迁移的内容对迁移学习进行划分,主要有特征表示迁移、实例迁移、参数迁移和关联关系迁移.特征表示迁移期望联合表示的特征优于只基于目标领域中数据的特征表示[172],代表性工作有自学习聚类算法[156]、基于概率的隐含语义分析算法[152]、基于流形结构的算法[154]等.实例迁移是指从源领域训练数据中抽取适合测试数据的实例,迁移到目标领域增加训练数据数目.经典模型包括 Tradaboosting 算法[157]、基于隐含稀疏领域迁移方法[173]等.参数迁移假设源领域和目标领域的一些参数是共享的,代表性工作有基于高斯过程 Gaussian Process(GP)的模型[174,175]以及结合层次贝叶斯 hierarchical Bayesian(HB)框架的模型[176,177].关联关系迁移的方法通常数据表示为关联关系,例如社会网络数据.通过源领域和目标领域的数据,迁移它们之间的关联关系,代表性工作有利用 Markov Logic 网络进行关系的迁移[178−180].

以上介绍的迁移学习算法形式各异,例如基于矩阵分解、概率主题模型等.表 4 对迁移学习方法进行了总结.

Table 4 Description of different kinds of transfer learning strategies

表 4 不同迁移学习方法的描述

迁移学习		描述	代表性工作
按照源领域和目标领域的特征空间是否相同划分	同构迁移学习	源领域和目标领域的特征空间相同,可以把与目标领域相关的源领域中的数据直接应用到目标领域中	基于主题的隐含语义分析算法[152]、谱分析算法[153]、自学习聚类算法[156]、Tradaboosting算法[157]、基于深度学习的方法[181−183]等
	异构迁移学习	源领域和目标领域的特征空间不同,通常需要学习异构的源领域和目标领域之间的关系,可以直接进行特征映射,也可以映射到共同的子空间中	风险最小化框架[158]、概率隐含语义分析方法[159,160]、协同矩阵分解技术[161]、积极迁移学习框架[162]、基于深度学习的方法[100,184,185]等
按照源领域和目标领域中是否有标注数据划分	归纳迁移学习	目标领域中有少量标注数据,根据源领域中是否有标注数据,可以把归纳迁移学习划分成多任务学习(源领域中有标注数据)、自学习(源领域中没有标注数据)	Tradaboosting算法[157]、基于Procrustes的流形对齐方法[164]、基于深度学习的方法[183]等
	直推式迁移学习	只有源领域中有标注样本,源领域和目标领域的数据相关但不相同,两个领域的任务相同	Structural correspondence学习模型(SCL)[167]、Bridged refinement模型[168]、基于深度学习的方法[182,186]等
	无监督迁移学习	源领域和目标领域中都没有标签数据的问题	自学习聚类算法STC[156]、迁移判别分析TDA[172]等
按照迁移的内容划分	特征表示迁移学习	期望源领域和目标领域学到的联合表示特征优于只基于目标领域中数据的特征表示	基于主题的隐含语义分析算法[152]、基于流形结构的算法[154]、自学习聚类算法[156]等、基于深度学习的方法[184,187,188]
	实例迁移学习	从源领域训练数据中抽取一些实例,迁移到目标领域增加训练数据数目	Tradaboosting算法[157]、基于重建的隐含稀疏领域迁移方法[173]等
	参数迁移学习	挖掘源领域和目标领域的共享参数,用于目标领域中	基于高斯过程Gaussian Process(GP)的模型[174,175]以及结合层次贝叶斯hierarchical Bayesian(HB)框架的模型[176,177]等
	关联关系迁移学习	通过源领域和目标领域的数据,迁移它们之间的关联关系	利用Markov Logic网络迁移的模型[178−180]等

这些模型大都是浅层结构,随着具有多层结构的深度学习在一些领域获得了成功,研究者把深度学习引入到迁移学习模型中,通过构建具有很多隐层的结构,学习更有用的特征,最终提升迁移学习在目标领域中任务的性能[100,181−193].Bengio 等人[181]研究了无监督预训练特征的有效性,将其应用到迁移学习场景下.Glorot 等人[182]将不同领域的数据输入到叠加降噪自动编码器中,学习更加健壮的特征,对源领域和目标利用中的样本进行重

新表示.Oquab 等人[183]首先利用源领域中已标注的样本训练 CNN,然后增加适应层缩小两个领域之间的差异,并且利用目标领域中已标注样本微调训练的 CNN.Yosinski 等人[189]量化了深层神经网络中每一层特征的可迁移性.Zhuang 等人[190]结合深度自动编码器进行迁移,通过最小化源领域和目标领域的隐藏层的 KL 距离获得领域不变的特征.Long 等人[186]提出了联合自适应网络的结构,应对目标领域中已标注样本数量较少的问题.Sun 等人[187]和 Rozantsev 等人[187]提出了深度领域自适应的方法.文献[100,184,185]设计了基于深度学习的异构迁移学习算法,应用到人脸识别、图像-文本迁移中.深度学习在迁移学习上获得了较好的性能,但是在其可解释性和参数调整方面,仍然需要进一步研究.

研究者从理论层面对迁移学习进行了研究,Ben-David 等人[194]基于 VC 维对领域适应性问题给出了推广性的界.对于有限 VC 维的情况,可用文献[195]中提出的方法,从有限个样本估计适应推广能力.Blitzer 等人[169]从源数据加权组合获得模型,并给出在特定的经验风险最小化的情形下的误差率.Ben-David 等人[196]研究了分类器能够在目标领域很好地完成分类任务的条件.Mansour 等人[197,198]针对任意目标分布给出了基于源领域和目标领域之间 Rényi 散度的领域推广误差,提出通过加权经验分布可较为准确地反映目标领域分布.Zhang 等人[199]提出一种新的框架来分析典型的领域适应学习过程的理论性质,分别开发了 Hoeffding 型、Bennett 型和 McMiarmid 型偏差不等式,提出了基于 Rademacher 复杂度的泛化边界,并分析了渐近收敛性和学习过程的收敛速度.Kumagai[200]分析了参数迁移学习的 margin 边界.然而,正如庄福振等人[101]所指出的,尽管研究者对迁移学习已经进行了一些理论尝试,但还远远不够,尤其是需要深入开展迁移学习有效性的理论研究.

通过以上分析可以发现,不同多视图学习和迁移学习的算法设计理论各不相同.然而从认知的角度考虑,多源学习方法的总体目标是为了学习隐藏于多源数据下的共同知识.已有的多源学习方法只是根据具体问题设计的具体算法,这些方法没有共性的约束,不能指导新的学习算法设计.

4 未来研究方向的思考

虽然在基础理论研究和应用领域,多源数据学习已经成为研究热门且存在一些较为成熟的技术,但是现有的多源数据学习算法对应的学习任务差别巨大,其学习算法的表示也严重碎片化,彼此形式差别极大.早在 2004 年,周志华就指出[201],机器学习“以 Tom Mitchell 的经典教科书为例,很难看到基础学科(例如数学、物理学)教科书中那种贯穿始终的体系,也许会让人感到这不过是不同方法和技术的堆砌”.

机器学习算法(包括单源和多源数据学习算法)的表示碎片化和形式差别化,对研究如何统一机器学习算法的理论带来很大的困难.如果像 Vapnik 那样将机器学习问题看作是一个基于经验值的函数估计问题[202],则会失去对学习问题的内在约束,对很多机器学习算法的设计启发性不足.更重要的是,如果将机器学习问题看作一个基于经验值的函数估计问题,则几乎完全隐蔽了对学习问题所具有的共同内蕴认知性质,即学习是为了完成一个认知任务(从数据中形成知识).多源数据是由单源数据组合而成的,多源数据学习也是一个典型的认知任务,希望像人一样进行多源学习.因此,应该将多源数据学习所具有的共同内蕴的认知性质挖掘出来.由于现今的单源数据学习理论对于这一点研究不足,而现有的多源数据学习方法往往是由单源数据学习方法扩展而来,所以面临着与单源数据学习同样的理论薄弱问题,很难解决多源数据学习算法的理论性能问题.根据上述分析,我们认为,多源数据的理论和算法的未来研究方向包括如下几个方面.

(1) 目前,人类学习的认知机理存在着两个关键问题:一是学习的认知机理不一致,不能确定是相似性还是简单性起主导作用;二是尽管多模态学习是人类学习的基本特性,但当前研究大多探讨单模态的学习机理.在人类知觉学习、类别学习和概念表征以及机器学习中,相似性都起到了重要作用.因此,未来的研究热点之一是以相似性为中心,基于单模态相似性来研究多模态相似性的整合机制,从而得到适合于机器学习使用的人类多源数据学习的认知机理,即概念表征的认知机理.

(2) 为了从认知上统一各种学习算法,研究者已经做出了很多努力,提出了许多理论——PAC 学习理论、统计学习理论、概率图理论等.现在的学习理论都只覆盖了部分学习算法.因此,两位机器学习领域的指标性人物 Jordan 和 Mitchell 在《Science》上提出了机器学习的一个重要挑战:能否建立一个能够统管所有机器、人和生

物的学习理论?机器学习的认知目的是从数据中得到知识,而知识的基本单位是概念.因此,未来的研究热点之一是如何利用概念表征的认知机理进行概念的统一表示理论,进而将学习算法进行统一表示,特别是将多源数据学习进行统一表示,在此基础之上,发现学习算法的基本认知假设,即机器学习公理化研究.

(3) 现在的多源数据学习算法大致的设计思路有 3 种:前期融合算法,强调特征融合;后期融合算法,强调结果融合;中期融合算法,强调子空间共性.这些思路都需要设计概念表示理论,即学习算法的认知基本假设.未来的研究热点之一是在机器学习公理化的研究基础上,研究多源的学习算法设计原则和学习算法评估原则,研究一系列多源数据学习算法设计.

References:

[1] Nature. Big data. 2008. http://www.nature.com/news/specials/bigdata/index.html

[2] Science. Special online collection: Dealing with data. 2011. http://www.sciencemag.org/site/special/data/

[3] Silver D, Huang A, Maddison CJ. Mastering the game of Go with deep neural networks and tree search. Nature, 2016,529(7587): 484−489. [doi: 10.1038/nature16961]

[4] http://www.aaai.org/Conferences/AAAI/2016/aaai16speakers.php#Hassabis

[5] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor J. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 2001,18(1):32−80. [doi: 10.1109/79.911197]

[6] Cohen I, Cozman F, Sebe N, Cirelo M, Huang T. Semisupervised learning of classifiers: Theory, algorithms and their applications to human- computer interaction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004,26(12):1553−1567. [doi: 10.1109/TPAMI. 2004.127]

[7] Bekkerman R, Jeon J. Multi-Modal clustering for multimedia collections. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2007. 1−8. [doi: 10.1109/CVPR.2007.383223]

[8] Wang J, Duan L, Liu Q, Lu H, Jin J. A multimodal scheme for program segmentation and representation in broadcast video streams. IEEE Trans. on Multimedia, 2008,10(3):393−408. [doi: 10.1109/TMM.2008.917362]

[9] Cristani M, Bicego M, Murino V. Audio-Visual event recognition in surveillance video sequences. IEEE Trans. on Multimedia, 2007,9(2):257−267. [doi: 10.1109/TMM.2006.886263]

[10] Liu J, Lai W, Hua XH, Huang Y, Li S. Video search re-ranking via multi-graph propagation. In: Proc. of the 15th ACM Int'l Conf. on Multimedia. 2007. 208−217. [doi: 10.1145/1291233.1291279]

[11] Valiant L. A theory of the learnable. Communications of the ACM, 1984,27(11):1134−1142. [doi: 10.1145/1968.1972]

[12] Cortes C, Vapnik VN. Support vector networks. Machine Learning, 1995,20(3):273−297. [doi: 10.1007/BF00994018]

[13] Kleinberg J. An impossibility theorem for clustering. In: Proc. of the Advances in Neural Information Processing Systems. 2003. 463−470.

[14] Valiant L. Probably approximately correct: Nature's Algorithms for Leaning and Prospering in a Complex World. New York: Basic Books, 2013.

[15] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo: Morgan Kaufmann Publishers, 1988.

[16] Pavlas J. Searching for the algorithms underlying life. Quanta Magazine, 2016. https://www.quantamagazine.org/the-hidden-algorithms-underlying-life-20160128/

[17] Snedeker J. Word learning. In: Binder MD, Hirokawa N, Windhorst U, eds. Encyclopedia of Neuroscience. Springer-Verlag, 2009. 503−508.

[18] Urcuioli PJ, Wasserman EA, Zentall TR. Associative concept learning in animals: Issues and opportunities. Journal of the Experimental Analysis of Behavior, 2014,101(1):165−170. [doi: 10.1002/jeab.62]

[19] Daniel TA, Cook RG, Katz JS. Temporal dynamics of task switching and abstract-concept learning in pigeons. Front Psychol, 2015,6:1−8. [doi: 10.3389/fpsyg.2015.01334]

[20] Gingras G, Rowland BA, Stein BE. The differing impact of multisensory and unisensory integration on behavior. Journal of Neuroscience, 2009,29(15):4897−4902. [doi: 10.1523/JNEUROSCI.4120-08.2009]

[21] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science, 2015,349(6245):255−260. [doi: 10.1126/science.aaa8415]

[22] Stein BE, Stanford TR. Multisensory integration: Current issues from the perspective of the single neuron. Nature Reviews Neuroscience, 2008,9(4):255−266. [doi:10.1038/nrn2331]

[23] Fu X, Cai L, Liu Y, Jia J, Chen W, Yi Z, Zhao G, Liu Y, Wu C. A computational cognition model of perception, memory, and judgment. Science China (Information Sciences), 2014,57(3):1−15. [doi: 10.1007/s11432-013-4911-9]

[24] Seger CA. How do the basal ganglia contribute to categorization? Their role in generalization, response selection, and learning via feedback. Neuroscience and Biobehavioral Reviews, 2008,32(2):265−278. [doi: 10.1016/j.neubiorev.2007.07.010]

[25] Seger CA, Miller EK. Category learning in the brain. Annual Review of Neuroscience, 2010,33:203−219. [doi: 10.1146/annurev.neuro.051508.135546]

[26] Fugate JM. Categorical perception for emotional faces. Emotion Review, 2013,5(1):84−89. [doi: 10.1177/1754073912451350]

[27] Young AW, Rowland D, Calder AJ, Etcoff NL, Seth A, Perrett DI. Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. Cognition, 1997,63:271−313. [doi: 10.1016/S0010-0277(97)00003-6]

[28] Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM. A neuropsychological theory of multiple systems in category learning. Psychological Review, 1998,105:442−481. [doi: 10.1037/0033-295X.105.3.442]

[29] Flavell JH. Cognitive Development. Englewood Cliffs: Prentice-Hall, 1985.

[30] Ashby FG, OBrien JB. Category learning and multiple memory systems. Trends in Cognitive Sciences, 2005,9:83−89. [doi: 10.1016/j.tics.2004.12.003]

[31] Ernst MO, Bülthoff HH. Merging the senses into a robust percept. Trends in Cognitive Sciences, 2004,8(4):162−169. [doi: 10.1016/j.tics.2004.02.002]

[32] Alais D, Newell FN, Mamassian P. Multisensory processing in review: From physiology to behaviour. Seeing and Perceiving, 2010,23(1):3−38. [doi: 10.1163/187847510X488603]

[33] Wassenhove van V, Grantand KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia, 2007,45(3):598−607. [doi: 10.1016/j.neuropsychologia.2006.01.001]

[34] Driver J, Noesselt T. Multisensory interplay reveals crossmodal influences on ‘sensory-specific’ brain regions, neural responses, and judgments. Neuron, 2008,57(1):11−23. [doi: 10.1016/j.neuron.2007.12.013]

[35] Ursino M, Cuppini C, Magosso E. A neural network for learning the meaning of objects and words from a featural representation. Neural Networks, 2015,63:234−253. [doi: 10.1016/j.neunet.2014.11.009]

[36] Shams L, Seitz AR. Benefits of multisensory learning. Trends in Cognitive Sciences, 2008,12(11):411−417. [doi: 10.1016/j.tics.2008.07.006]

[37] Hahn U, Chater N, Riehardson L. Similarity as transformation. Cognition, 2003,87(1):1−32. [doi: 10.1016/S0010-0277(02)00184-1]

[38] Lin N, Guo QH, Han ZZ, Bi Y. Motor knowledge is one dimension for concept organization: Further evidence from a Chinese semantic dementia case. Brain and Language, 2011,119(2):110−118. [doi: 10.1016/j.bandl.2010.07.001]

[39] Bub DN, Masson ME, Cree GS. Evocation of functional and volumetric gestural knowledge by objects and words. Cognition, 2008,106(1):27−58. [doi: 10.1016/j.cognition.2006.12.010]

[40] Vingerhoets G. Knowing about tools: Neural correlates of tool familiarity and experience. Neuroimage, 2008,40(3):1380−1391. [doi: 10.1016/j.neuroimage.2007.12.058]

[41] Buxbaum LJ, Kalénine S. Action knowledge, visuomotor activation, and embodiment in the two action systems. Annals of the New York Academy of Sciences, 2010,1191(1):201−218. [doi: 10.1111/j.1749-6632.2010.05447.x]

[42] Barsalou LW. Grounded cognition. Annual Review of Psychology, 2008,59:617−645. [doi: 10.1146/annurev.psych.59.103006.093639]

[43] Ell SW, Ing AD, Maddox WT. Critrial noise effects on rule-based category learning: The impact of delayed feedback. Attention, Perception, and Psychophysics, 2009,71(6):1263−1275. [doi: 10.3758/APP.71.6.1263]

[44] Ashby FG, Maddox WT. Human category learning 2.0. The Year in Cognitive Neuroscience, 2010,1224(1):147−161. [doi: 10.1111/j.1749-6632.2010.05874.x]

[45] Hahn U. Similarity. Wiley Interdisciplinary Reviews: Cognitive Science, 2014,5(3):271−280. [doi: 10.1002/wcs.1282]

[46] Kloos H, Sloutsky VM. What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. Journal of Experimental Psychology: General, 2008,137(1):52−72. [doi: 10.1037/0096-3445.137.1.52]

[47] Lelkes AD. Algorithms and complexity results for learning and big data [Ph.D. Thesis]. University of Illinois at Chicago, 2017.

[48] Ehrenfeucht A, Haussler D, Kearns M, Valiant L. A general lower bound on the number of examples needed for learning. Information and Computation, 1989,82(3):247−261. [doi: 10.1016/0890-5401(89)90002-3]

[49] Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Learnability and the Vapnik-Chervonenkis dimension. Journal of the Association on Computing Machinery, 1989,36(4):929−965. [doi: 10.1145/76359.76371]

[50] Auer P, Ortner R. A new PAC bound for intersection-closed concept classes. Machine Learning, 2007,66(2-3):151−163. [doi: 10.1007/s10994-006-8638-3]

[51] Warmuth M. The optimal PAC algorithm. In: Proc. of the 17th Annual Conf. on Learning Theory. 2004. 641−642. [doi: 10.1007/978-3-540-27819-1_45]

[52] Haussler D, Littlestone N, Warmuth MK. Predicting {0,1} functions on randomly drawn points. Information and Computation, 1994,115(2):284−293.

[53] Hanneke S. Theoretical foundations of active learning [Ph.D. Thesis]. Carnegie Mellon University, 2009.

[54] Darnstadt M. The optimal PAC bound for intersection-closed concept classes. Information Processing Letters, 2014,115(4):458−461. [doi: 10.1016/j.ipl.2014.12.001]

[55] Simon HU. An almost optimal PAC algorithm. In: Proc. of the Int'l Conf. on Machine Learning Workshop. 2015. 1552−1563.

[56] Hanneke S. The optimal sample complexity of PAC learning. Journal of Machine Learning Research, 2016,17(38):1−15.

[57] Vapnik VN, Chervonenkis AY. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications, 1971,16(2):264−280. [doi: 10.1137/1116025]

[58] Vapnik VN. Statistical Learning Theory. Wiley-Interscience, 1989.

[59] Schapire R, Freund Y, Bartlett P, Lee WS. Boosting the margin a new explanation for the effectiveness of voting methods. Annals of Statistics, 1998,26:1651−1686. [doi: 10.1214/aos/1024691352]

[60] Breiman L. Prediction games and arcing algorithms. Neural Computation, 1999,11(7):1493−1517. [doi: 10.1162/089976699300016106]

[61] Wang L, Sugiyama M, Jing Z, Yang C, Zhou ZH, Feng J. A refined margin analysis for boosting algorithms via equilibrium margin. Journal of Machine Learning Research, 2011,12:1835−1863.

[62] Gao W, Zhou ZH. On the doubt about margin explanation of boosting. Artificial Intelligence, 2013,203:1−18. [doi: 10.1016/j.artint.2013.07.002]

[63] Zhang T, Zhou ZH. Large margin distribution machine. In: Proc. of the 20th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. 2014. 313−322. [doi: 10.1145/2623330.2623710]

[64] Hanneke S. Refined error bounds for several learning algorithms. Journal of Machine Learning Research, 2016,17(135):1−55.

[65] Wu XX, Zhang JP. Researches on Rademacher complexities in statistical learning theory: A survey. Acta Automatica Sinica, 2017,43(1):20−39 (in Chinese with English abstract).

[66] Shawe-Taylor J, Bartlett PL, Williamson RC, Anthony M. Structural risk minimization over data-dependent hierarchies. IEEE Trans. on Information Theory, 1998,44(5):1926−1940. [doi: 10.1109/18.705570]

[67] Bartlett PL, Mendelson SR. Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 2003,3:463−482.

[68] Gnecco G, Sanguineti M. Approximation error bounds via Rademacher's complexity. Applied Mathematical Sciences, 2008,4(2):153−176.

[69] Koltchinskii V, Panchenko D. Rademacher processes and bounding the risk of function learning. In: Proc. of the High Dimensional Probability II. Birkhäuser Boston: Cambridge, 2000. 443−457. [doi: 10.1007/978-1-4612-1358-1_29]

[70] Bartlett PL, Bousquet O, Mendelson S. Local Rademachercomplexities. The Annals of Statistics, 2005,33(4):1497−1537. [doi: 10.1214/009053605000000282]

[71] Koltchinskii V. Local Rademacher complexities and oracle inequalities in risk minimization. Annals of Statistics, 2006,34(6): 2593−2656. [doi: 10.1214/009053606000001019]

[72] Zhivotovskiy N, Hanneke S. Localization of VC classes: Beyond local Rademacher complexities. In: Proc. of the Int'l Conf. on Algorithmic Learning Theory. 2016. 18−33. [doi: 10.1007/978-3-319-46379-7_2]

[73] Xu C, Liu T, Tao D, Xu C. Local Rademacher complexity for multi-label learning. IEEE Trans. on Image Processing, 2016,25(3): 1495−1507. [doi: 10.1109/TIP.2016.2524207]

[74] Lei YW, Ying YM. Generalization analysis of multi-modal metric learning. Analysis and Applications, 2016,14(4):503−521. [doi: 10.1142/S0219530515500104]

[75] Gao W, Zhou ZH. Dropout Rademacher complexity of deep neural networks. Science China (Information Sciences), 2016,59(7): 1−12. [doi: 10.1007/s11432-015-5470-z]

[76] Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 2008,1(1-2):1−305. [doi: 10.1561/2200000001]

[77] Koller D, Friedman N. Probabilistic Graphical Models. MIT Press, 2009.

[78] Jordan MI. Graphical models. Statistical Science, 2004,19(1):140−155. [doi: 10.1214/088342304000000026]

[79] Pearl J. Asymptotic properties of minimax trees and game-searching procedures. Artificial Intelligence, 1980,14(2):113−138.

[80] Jensen FV. An Introduction to Bayesian Networks. Springer-Verlag, 1996.

[81] Jordan MI, ed. Learning in Graphical Models. Kluwer: The Netherlands, 1998.

[82] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1977,39(1):1−38.

[83] Pearl J. Reverend Bayes on inference engines: A distributed hierarchical approach. In: Proc. of the 2nd National Conf. on Artificial Intelligence. 1982. 133−136.

[84] Murphy KP, Weiss Y, Jordan MI. Loopy belief propagation for approximate inference: An empirical study. In: Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence. 1999. 467−475.

[85] Mooij J, Kappen H. Sufficient conditions for convergence of the sum-product algorithm. IEEE Trans. on Information Theory, 2007, 53(12):4422−4437. [doi: 10.1109/TIT.2007.909166]

[86] Wainwright MJ, Jaakkola TS, Willsky AS. Tree-Reweighted belief propagation and approximate ML estimation by pseudo-moment matching. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2003. 1−8.

[87] Minka TP. Expectation propagation for approximate Bayesian inference. In: Proc. of the Conf. on Uncertainty in Artificial Intelligence. 2001. 362−369.

[88] Johnson J, Malioutov D, Willsky A. Walk-Sum interpretation and analysis of Gaussian belief propagation. In: Proc. of the Advances in Neural Information Processing Systems. 2006. 1−8.

[89] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 1970,57(1):97−109. [doi: 10.1093/biomet/57.1.97]

[90] McAllester DA. Some PAC Bayesian theorems. In: Proc. of the Annual Conf. on Computational Learning Theory. 1998,51: 230−234.

[91] Seldin Y, Tishby N. PAC-Bayesian analysis of co-clustering and beyond. Journal of Machine Learning Research, 2010,11: 3595−3646.

[92] Sun S, Shawe-Taylor J, Mao L. PAC-Bayes analysis of multi-view learning. Information Fusion, 2017,35:117−131. [doi: 10.1016/j.inffus.2016.09.008]

[93] Seldin Y, Tishby N. A PAC-Bayesian approach to unsupervised learning with application to co-clustering analysis. Journal of Machine Learning Research, 2010,3:1−46. [doi: 10.1561/2200000016]

[94] Higgs M, Shawe-Taylor J. A PAC-Bayes bound for tailored density estimation. In: Proc. of the Algorithmic Learning Theory. 2010. 148−162. [doi: 10.1007/978-3-642-16108-7_15]

[95] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proc. of the Association for Computational Linguistics. 1995. 189−196. [doi: 10.3115/981658.981684]

[96] Xu C, Tao DC, Xu C. A survey on multi-view learning. arxiv: 1304.5634, 2013.

[97] Zhou ZH. Ensemble Methods: Foundations and Algorithms. Boca Raton: Chapman and Hall/CRC, 2012.

[98] Zhao J, Xie X, Xu X, Sun S. Multi-View learning overview: Recent progress and new challenges. Information Fusion, 2017,38: 43−54. [doi: 10.1016/j.inffus.2017.02.007]

[99] 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]

[100] Saxena S, Verbeek J. Heterogeneous face recognition with cnns. In: Proc. of the European Conf. on Computer Vision Workshop on Transferring and Adapting Source Knowledge in Computer Vision. 2016. 483−491. [doi: 10.1007/978-3-319-49409-8_40]

[101] 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]

[102] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proc. of the Workshop on Computational Learning Theory. 1998. 92−100. [doi: 10.1145/279943.279962]

[103] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training. In: Proc. of the 9th Int'l Conf. on Information and Knowledge Management. 2000. 86−93. [doi: 10.1145/354756.354805]

[104] Yu S, Krishnapuram B, Rosales R, Rao RB. Bayesian co-training. Journal of Machine Learning Research, 2011,12:2649−2680.

[105] Kumar A, Daume H. A co-training approach for multi-view spectral clustering. In: Proc. of the Int'l Conf. on Machine Learning. 2011. 393−400.

[106] Zhou ZH, Li M. Tri-Training: Exploiting unlabeled data using three classifiers. IEEE Trans. on Knowledge and Data Engineering, 2005,17(11):1529−1541. [doi: 10.1109/TKDE.2005.186]

[107] Li M, Zhou ZH. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans, 2007,37(6):1088−1098. [doi: 10.1109/TSMCA.2007.904745]

[108] Sindhwani V, Rosenberg DS. An RKHS for multi-view learning and manifold co-regularization. In: Proc. of the 25th Int'l Conf. on Machine Learning. 2008. 976−983. [doi: 10.1145/1390156.1390279]

[109] Li G, Chang K, Hoi S. Multiview semi-supervised learning with consensus. IEEE Trans. on Knowledge and Data Engineering, 2012,24(11):2040−2051. [doi: 10.1109/TKDE.2011.160]

[110] Xu X, Li W, Xu D, Tsang IW. Co-Labeling for multi-view weakly labeled learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016,38(6):1113−1125. [doi: 10.1109/TPAMI.2015.2476813]

[111] Gönen M, Alpaydın E. Multiple kernel learning algorithms. Journal of Machine Learning Research, 2011,12:2211−2268.

[112] Lanckriet GR, Cristianini N, Bartlett P. Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research, 2004,5:27−72.

[113] Bach FR, Lanckriet GR, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proc. of the 21st Int'l Conf. on Machine Learning. 2004. 1−8. [doi: 10.1145/1015330.1015424]

[114] Sonnenburg S, Rätsch G, Schäfer C. Large scale multiple kernel learning. Journal of Machine Learning Research, 2006,7: 1531−1565.

[115] Rakotomamonjy A, Bach F, Canu S, Grandvalet Y. SimpleMKL. Journal of Machine Learning Research, 2008,9:2491−2521.

[116] Subrahmanya N, Shin YC. Sparse multiple kernel learning for signal processing applications. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(5):788−798. [doi: 10.1109/TPAMI.2009.98]

[117] Hotelling H. Relations between two sets of varieties. Biometrika, 1936,28(3-4):321−377. [doi: 10.1093/biomet/28.3-4.321]

[118] Akaho S. A kernel method for canonical correlation analysis. arXiv preprint cs/0609071, 2006.

[119] Chaudhuri K, Kakade SM, Livescu K. Multi-View clustering via canonical correlation analysis. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. ACM Press, 2009. 129−136. [doi: 10.1145/1553374.1553391]

[120] Kakade SM, Foster DP. Multi-View regression via canonical correlation analysis. In: Proc. of the Learning Theory. 2007. 82−96. [doi: 10.1007/978-3-540-72927-3_8]

[121] Hardoon DR, Shawe-Taylor J. Convergence analysis of kernel canonical correlation analysis: Theory and practice. Machine Learning, 2009,74(1):23−38. [doi: 10.1007/s10994-008-5085-3]

[122] Cai J, Sun HW. Convergence rate of kernel canonical correlation analysis. In: Proc. of the Science China Mathematics. 2011. 1−10. [doi: 10.1007/s11425-011-4245-2]

[123] Blei DM, Jordan MI. Modeling annotated data. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2003. 127−134. [doi: 10.1145/860435.860460]

[124] Diethe T, Hardoon DR, Shawe-Taylor J. Multiview fisher discriminant analysis. In: Proc. of the Advances in Neural Information Processing Systems Workshop on Learning from Multiple Sources. 2008. 1−8.

[125] Li SY, Jiang Y, Zhou ZH. Partial multi-view clustering. In: Proc. of the Association for the Advancement of Artificial Intelligence. 2014. 1968−1974.

[126] Liu J, Jiang Y, Li Z, Zhou ZH, Lu H. Partially shared latent factor learning with multiview data. IEEE Trans. on Neural Networks and Learning Systems, 2014,26(6):1233−1246. [doi: 10.1109/TNNLS.2014.2335234]

[127] Farquhar J. Two view learning: SVM-2K, theory and practice. In: Proc. of the Advances in Neural Information Processing Systems. 2005. 355−362.

[128] Yin Q, Wu S, Wang L. Unified subspace learning for incomplete and unlabeled multi-view data. Pattern Recognition, 2017,67: 313−327. [doi: 10.1016/j.patcog.2017.01.035]

[129] Xu C, Tao D, Xu C. Large-Margin multi-view information bottleneck. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014,36(8):1558−1572. [doi: 10.1109/TPAMI.2013.2296528]

[130] Xu C, Tao D, Xu C. Multi-View intact space learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,37(12): 2531−2544. [doi: 10.1109/TPAMI.2015.2417578]

[131] Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proc. of the 28th Int'l Conf. on Machine Learning. 2011. 689−696.

[132] Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines. In: Proc. of the Advances in Neural Information Processing Systems. 2012. 2222−2230.

[133] Srivastava N, Salakhutdinov R. Learning representations for multimodal data with deep belief nets. In: Proc. of the Int'l Conf. on Machine Learning Workshop on Representation Learning. 2012. 1−8.

[134] Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-View convolutional neural networks for 3D shape recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 945−953. [doi: 10.1109/ICCV.2015.114]

[135] Elhoseiny M, El-Gaaly T, Bakry A, Elgammal A. A comparative analysis and study of multiview CNN models for joint object categorization and pose estimation. In: Proc. of the 33rd Int'l Conf. on Machine Learning. 2016. 888−897.

[136] Andrew G, Arora R, Bilmes J, Livescu K. Deep canonical correlation analysis. In: Proc. of the Int'l Conf. on Machine Learning. 2013. 1247−1255.

[137] Benton A, Khayrallah H, Gujral B, Reisinger D, Zhang S, Arora R. Deep generalized canonical correlation analysis. arXiv preprint arXiv:1702.02519, 2017.

[138] Wang W, Arora R, Livescu K, Bilmes J. On deep multi-view representation learning: Objectives and optimization. arXiv preprint arXiv: 1602.01024, 2016.

[139] Yang X, Ramesh P, Chitta R, Madhvanath S, Bernal E A, Luo J. Deep multimodal representation learning from temporal data. arXiv preprint arXiv:1704.03152, 2017.

[140] Du J, Ling CX, Zhou ZH. When does co-training work in real data? IEEE Trans. on Knowledge and Data Engineering, 2011,23(5): 788−799. [doi: 10.1109/TKDE.2010.158]

[141] Dasgupta S, Littman M, McAllester D. PAC generalization bounds for co-training. In: Proc. of the Neural Information Processing Systems. 2001. 375−382.

[142] Abney S. Bootstrapping. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. 2002. 360−367.

[143] Balcan M, Blum A, Yang K. Co-Training and expansion: Towards bridging theory and practice. In: Proc. of the Advances in Neural Information Processing Systems. 2004. 89−96.

[144] Wang W, Zhou ZH. A new analysis of co-training. In: Proc. of the Int'l Conf. on Machine Learning. 2010. 1135−1142.

[145] Sindhwani V, Niyogi P, Belkin M. A co-regularization approach to semi-supervised learning with multiple views. In: Proc. of the Int'l Conf. on Machine Learning Workshop on Learning with Multiple Views. 2005. 74−79.

[146] Hussain Z, Shawe-Taylor J. Improved loss bounds for multiple kernel learning. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2011. 370−377.

[147] Ying Y, Campbell C. Generalization bounds for learning the kernel. In: Proc. of the 22nd Annual Conf. on Learning Theory. 2009. 1–9.

[148] Kloft M, Blanchard G. The local rademacher complexity of l_p-norm multiple kernel learning. Arxiv preprint arXiv: 1103.0790, 2011.

[149] Woodworth RS, Thorndike EL. The influence of improvement in one mental function upon the efficiency of other functions. Psychological Review, 1901,8(3):247–261.

[150] Long MS. Transferlearing: Problems and methods [Ph.D. Thesis]. Beijing: Tsinghua University, 2014 (in Chinese).

[151] http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html

[152] Xue GR, Dai WY, Yang Q, Yu Y. Topic-Bridged PLSA for cross-domain text classification. In: Proc. of the 31st Int'l ACM SIGIR Conf. on Research and Development on Information Retrieval. 2008. 627–634. [doi: 10.1145/1390334.1390441]

[153] Ling X, Dai WY, Xue GR, Yang Q, Yu Y. Spectral domain-transfer learning. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2008. 488–496. [doi: 10.1145/1401890.1401951]

[154] Pan SJ, Ni X, Sun J, Yang Q, Chen Z. Cross-Domain sentiment classification via spectral feature alignment. In: Proc. of the 19th Int'l Conf. on World Wide Web. 2010. 751–760. [doi: 10.1145/1772690.1772767]

[155] Gao XB, Wang XM, Li XL, Tao D. Transfer latent variable model based on divergence analysis. Pattern Recognition, 2011, 44(10-11):2358–2366. [doi: 10.1016/j.patcog.2010.06.013]

[156] Dai WY, Yang Q, Xue GR, Yu Y. Self-Taught clustering. In: Proc. of the Int'l Conf. on Machine Learning. 2008. 200–207. [doi: 10.1145/1390156.1390182]

[157] Dai WY, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Proc. of the Int'l Conf. on Machine Learning. 2007. 193–200. [doi: 10.1145/1273496.1273521]

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

[159] Yang Q, Chen Y, Xue GR, Dai W, Yu Y. Heterogeneous transfer learning for image clustering via the social Web. In: Proc. of the Joint Conf. of the Annual Meeting of the ACL and the Int'l Joint Conf. on Natural Language Processing. 2009. 1–9.

[160] Zhuang ZF, Luo P, Shen ZY, He Q, Xiong Y, Shi Z, Xiong H. Collaborative dual-PLSA: Mining distinction and commonality across multiple domains for Classification. In: Proc. of the Int'l Conf. on Information and Knowledge Management. 2010. 359–368.

[161] Yang L, Jing LP, Ng KM. Robust and non-negative collective matrix factorization for text-to-image transfer learning. IEEE Trans. on Image Processing, 2015,24(12):4701–4714. [doi: 10.1109/TIP.2015.2465157]

[162] Yang L, Hanneke S, Carbonell J. A theory of transfer learning with applications to active learning. Machine Learning, 2013,90(2): 161–189. [doi: 10.1007/s10994-012-5310-y]

[163] Jiang J, Zhai CX. Instance weighting for domain adaptation in NLP. In: Proc. of the 45th Annual Meeting of the Association for Computational Linguistics. 2007. 264–271.

[164] Wang C, Mahadevan S. Manifold alignment using procrustes analysis. In: Proc. of the Int'l Conf. on Machine Learning. 2008. 1120–1127. [doi: 10.1145/1390156.1390297]

[165] Chang WC, Wu Y, Liu H, Yang Y. Cross-Domain kernel induction for transfer learning. In: Proc. of the Association for the Advancement of Artificial Intelligence. 2017. 1–7.

[166] Segev N, Harel M, Mannor S, Crammer K, El-Yaniv R. Learn on source, refine on target: A model transfer learning framework with random forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016,39(9):1811–1824. [doi: 10.1109/TPAMI. 2016.2618118]

[167] 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. 2006. 120–128.

[168] 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. 2007. 324–335. [doi: 10.1007/978-3-540-74976-9_31]

[169] Blitzer J, Crammer K, Kulesza A, Pereira F, Wortman J. Learning bounds for domain adaptation. In: Proc. of the Advances in Neural Information Processing Systems. 2008. 129–136.

[170] Zen G, Porzi L, Sangineto E, Ricci E, Sebe N. Learning personalized models for facial expression analysis and gesture recognition. IEEE Trans. on Multimedia, 2016,18(4):775−788. [doi: 10.1109/TMM.2016.2523421]

[171] Wang Z, Song YQ, Zhang CS. Transferred dimensionality reduction. In: Proc. of the European Conf. on Machine Learning and Knowledge Discovery in Databases. 2008. 550−565. [doi: 10.1007/978-3-540-87481-2_36]

[172] Jiang W, Chung F. Transfer spectral clustering. In: Proc. of the 2012 European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2012. 789−803. [doi: 10.1007/978-3-642-33486-3_50]

[173] Zhang L, Zuo ZW, Zhang D. LSDT: Latent sparse domain transfer learning for visual adaptation. IEEE Trans. on Image Processing, 2016,25(3):1177−1191. [doi: 10.1109/TIP.2016.2516952]

[174] Lawrence ND, Platt JC. Learning to learn with the informative vector machine. In: Proc. of the 21st Int'l Conf. on Machine Learning. 2004. 1−8. [doi: 10.1145/1015330.1015382]

[175] Bonilla E, Chai KM, Williams C. Multi-Task Gaussian process prediction. In: Proc. of the 20th Advances in Neural Information Processing Systems. 2008. 153−160.

[176] Schwaighofer A, Tresp V, Yu K. Learning Gaussian process kernels via hierarchical Bayes. In: Proc. of the 17th Advances in Neural Information Processing Systems. 2005. 1209−1216.

[177] Evgeniou T, Pontil M. Regularized multi-task learning. In: Proc. of the 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. 2004. 109−117. [doi: 10.1145/1014052.1014067]

[178] Mihalkova L, Huynh T, Mooney RJ. Mapping and revising Markov logic networks for transfer learning. In: Proc. of the 22nd Association for the Advancement of Artificial Intelligence. 2007. 608−614.

[179] Mihalkova L, Mooney RJ. Transfer learning by mapping with minimal target data. In: Proc. of the Association for the Advancement of Artificial Intelligence Workshop Transfer Learning for Complex Tasks. 2008. 31−36.

[180] Davis J, Domingos P. Deep transfer via second-order Markov logic. In: Proc. of the 26th Int'l Conf. on Machine Learning. 2009. 217−224. [doi: 10.1145/1553374.1553402]

[181] Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: Proc. of the Int'l Conf. on Machine Learning Workshop on Unsupervised and Transfer Learning. 2012. 17−36.

[182] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proc. of the 28th Int'l Conf. on Machine Learning. 2011. 513−520.

[183] Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 1717−1724. [doi: 10.1109/CVPR.2014.222]

[184] Chen WY, Hsu TM, Tsai YH. Transfer neural trees for heterogeneous domain adaptation. In: Proc. of the European Conf. on Computer Vision. 2016. 399−414. [doi: 10.1007/978-3-319-46454-1_25]

[185] Wang L, Li Y, Lazebnik S. Learning deep structure-preserving image-text embeddings. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5005−5013. [doi: 10.1109/CVPR.2016.541]

[186] Long MS, Wang JM, Jordan MI. Deep transfer learning with joint adaptation networks. CoRR, vol. arXiv:1605.06636, 2016.

[187] Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation. In: Proc. of the European Conf. on Computer Vision Workshop on Transferring and Adapting Source Knowledge in Computer Vision. 2016. 443−450. [doi: 10.1007/978-3-319-49409-8_35]

[188] Rozantsev A, Salzmann M, Fua P. Beyond sharing weights for deep domain adaptation. CoRR, vol. arXiv:1603.06432, 2016.

[189] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Proc. of the Advances in Neural Information Processing Systems. 2014. 3320−3328.

[190] Zhuang F, Cheng X, Luo P, Pan SJ, He Q. Supervised representation learning: Transfer learning with deep autoencoders. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence. 2015. 4119−4125.

[191] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big Data, 2016,3(1):1−40. [doi: 10.1186/s40537-016-0043-6]

[192] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, MarchandM, Lempitsky V. Domain adversarial training of neural networks. Journal of Machine Learning Research, 2016,17(59):1−35.

[193] Liu MY, Tuzel O. Coupled generative adversarial networks. In: Proc. of the Advances in Neural Information Processing Systems. 2016. 469−477.

[194] 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. 2007. 137−144.

[195] Kifer D, Ben-David S, Gehrke J. Detecting change in data streams. In: Proc. of the 30th Int'l Conf. on Very Large Data Bases. 2004. 180−191.

[196] Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Machine Learning, 2010,79(1-2):151−175. [doi: 10.1007/s10994-009-5152-4]

[197] Mansour Y, Mohri M, Rostamizadeh A. Multiple source adaptation and the Rényi divergence. In: Proc. of the 25th Conf. on Uncertainty in Artificial Intelligence. 2009. 367−374.

[198] Mansour Y, Mohri M, Rostamizadeh A. Domain adaptation: Learning bounds and algorithms. In: Proc. of the 22nd Annual Conf. on Learning Theory. 2009.

[199] Zhang C, Zhang L, Fan W, Ye J. Generalization bounds for representative domain adaptation. arXiv preprint arXiv: 1401.0376, 2014.

[200] Kumagai W. Learning bound for parameter transfer learning. In: Proc. of Advances in Neural Information Processing Systems. 2016. 2721−2729.

[201] Zhou ZH. Pervasive machine learning. In: Proc. of Department of Information Science of National Natural Science Foundation of China on Major Issues in Basic Theory of Intelligent Science and Technology. 2004 (in Chinese with English abstract).

[202] Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 2000. [doi: 10.1007/978-1-4757-3264-1]

附中文参考文献:

[65] 吴新星,张军平.Rademacher 复杂度在统计学习理论中的研究:综述.自动化学报,2017,43(1):20−39.

[101] 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]

[150] 龙明盛.迁移学习问题与方法研究[博士学位论文].北京:清华大学,2014.

[201] 周志华.普适机器学习.见:国家自然科学基金委员会信息科学部"智能科学技术基础理论重大问题"高层研讨会.2004.

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

刘烨(1979－),女,博士,副研究员,CCF 专业会员,主要研究领域为认知心理学,情感计算.

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.

詹德川(1982－),男,博士,副教授,CCF 专业会员,主要研究领域为机器学习,数据挖掘.