软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2473−2475 [doi: 10.3724/SP.J.1001.2013.04487]

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

统计学习研究与应用专刊前言*

高 阳[1], 陈松灿[2]

[1](计算机软件新技术国家重点实验室(南京大学),江苏 南京 210046)

[2](南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

通讯作者: 高阳, E-mail: gaoy@nju.edu.cn

中文引用格式: 高阳,陈松灿.统计学习研究与应用专刊前言.软件学报,2013,24(11):2473−2475. http://www.jos.org.cn/1000-9825/4487.htm

机器学习是计算机科学、认知科学、数学、统计学、控制、人工智能等诸多学科的交叉领域.机器学习研究自二十世纪五六十年代的感知机开始起步,经过符号学习、神经网络等技术后,逐步发展成为包含统计机器学习、集成机器学习、强化机器学习的研究领域,并在近年来又涌现出迁移学习、流形学习、概率图模型等新的研究方向.机器学习除了自身研究范围不断延展、研究深度不断深化、基础理论逐步完善之外,其应用领域也不断扩大,甚至成为某些学科的基础性研究工具和支撑技术.与此同时,机器学习技术在软件工程、计算机网络、信息安全、生物信息学、信息检索和多媒体等领域中的成功应用,也极大地推动了该研究领域的发展.特别是在机器学习研究者 Lesie Valiant,Judea Pearl 分别因计算学习理论和概率图模型的创造性工作而获得 2011 年度和 2012 年度的图灵奖以后,统计机器学习研究又迎来了一个新的热潮.

目前随着研究的深入和应用的扩展,除了原有的学习泛化问题之外,机器学习也不断面临理论和技术层面上新的挑战.例如,大数据问题、弱一致性假设问题、弱标记问题、可理解性问题、代价敏感问题等等.本专刊选题为“统计学习研究与应用”,将突出目前机器学习研究中的几个热点技术,如概率图模型、稀疏性学习以及应用研究.

专刊组稿与第 14 届中国机器学习会议(CCML 2014)合作.通过会议,我们从总共 420 篇的会议投稿中遴选出了 7 篇高质量的论文.同时,通过专刊公开征文获得 67 篇投稿.这些稿件的内容覆盖了统计机器学习的几个主要研究方向和应用.专刊的审稿严格按照专刊审稿要求进行,特约编辑先后邀请了 40 余位机器学习及相关领域的专家参与评审,每篇邀请 3 位专家进行评审,并经历了初审、复审、会议宣读和终审各阶段.整个流程历经半年,最终有 23 篇文章入选本专刊.

《概率图模型研究进展综述》和《稀疏学习优化问题的求解综述》是两篇约稿的综述文章.前者系统地介绍了概率图模型的表示、推理和学习,着重讨论了贝叶斯网络、马尔可夫网络和因子图及其两类代表性模型:条件随机场和主题模型,并进一步分析了近似推理的加速.后者针对一类基于 L1 正则化的稀疏性学习问题,从建模到优化,从批优化到在线优化等进行了层次性的总结和分析,并侧重介绍了坐标优化、在线优化和随机优化等方法的特点、适用范围及其优劣.

《关系分类的学习界限研究》是一篇很有新意的理论性文章.文中提出了一个称为关系维的概念,用其度量关系模型中关联数据的能力,由此导出了关系分类的学习界限.

对支持向量机的研究仍处于不断的深入之中.《L_p 范数约束的多核半监督支持向量机学习方法》讨论了一种多核框架,采用双层优化过程来优化决策函数和核组合权系数.《动态粒度支持向量回归机》提出了一种改进的粒度支持向量机,实现了多层次的动态粒化方法.《加权光滑 CHKS 孪生支持向量机》采用了 CHKS 的光滑函数结合数据点的权重以处理孪生支持向量机的异常点.《基于统计学习分析多核间性能干扰》是一篇有趣

* 收稿时间: 2013-09-06

的一般性的模型学习文章,对上述文章中的模型选择会有启发性.该文通过核间的性能干扰,分析了核间的相互影响及改进模型性能的可能途径,借助主成分线性回归多核性能干扰模型的讨论,阐明了相关问题.

针对不平衡数据问题,《一种基于聚类的 PU 主动文本分类方法》通过聚类技术获得可信反例以平衡类数据,从而提高了文本的分类准确度.针对代价敏感问题,《具有 Fisher 一致性的代价敏感 Boosting 算法》通过对 Logit 损失函数的代价敏感改造,发展出了代价敏感的 Boosting 算法,并保证能够收敛于代价敏感贝叶斯决策.

结合数据的内在结构是实现有效学习的重要手段.《一种引入成对代价的子类判别分析》将数据的聚类结构信息引入现有子类判别分析中,获得了更有效的判别性能.《一种基于数据流的软子空间聚类算法》通过结合子空间和软聚类方法,设计了适合流数据分析的聚类方法.《属性加权的类属型数据非模聚类》针对类属性数据的特点,定义了相关的相异度度量,进而发展出一种类属性数据的软子空间聚类算法.《辅助信息自动生成的时间序列距离度量学习》首先借助动态时间弯曲自动生成样本间的成对约束信息,指导序列度量的学习,进而基于所学度量达到有效改善时间序列聚类性能的目的.《等谱流形学习算法》利用等谱流形所具有的内在结构的一致性,通过类别信息构建出相关邻接图,由此实现了等谱流形特性下的数据结构信息利用和对分类性能的提升.

强化学习是一类基于马尔可夫决策过程的顺序决策技术.《一种基于自生成样本学习的奖赏塑形方法》通过在学习过程中收集的样本,构造奖赏塑形函数,以替代真实的奖赏函数来加速学习过程.《一种高斯过程的带参近似策略迭代算法》则利用高斯过程对带参数的值函数进行建模,并通过其直接进行动作选择以加速学习过程.

向自然学习是孕育新的智能学习方法的有效途径之一,相应算法能够帮助统计学习中复杂模型的优化搜索.《逐维改进的布谷鸟搜索算法》通过对一类布谷鸟生存行为的仿生,设计出了一种相应的优化搜索算法.

社会网络分析和协同过滤是近年来统计学习所关注的面向应用的重要研究内容.《一种基于随机块模型的快速广义社区发现算法》通过对广义随机块收敛节点和边参数的裁剪,实现了一种能够快速发现问题结构发现的算法.《社群演化的稳健迁移估计及演化离群点检测》针对社群演化离群点问题,提出了社群迁移矩阵概念,进而讨论了相关性质.《基于时序行为的协同过滤推荐算法》则通过分析时间序列上相关用户(产品)的结构关系,实现了一种基于概率矩阵分解的协同过滤推荐算法.

《专家证据文档识别无向图模型》应用无向图模型,通过融合专家的多种信息来源,包括独立页面特征和文档间关联特征,实现了有效的专家证据文档识别.

人脸检测和图像分类同样是统计学习所面对的典型应用问题.《基于局部区域稀疏编码的人脸检测》通过对人脸每个局部区域的字典学习,结合区域检测和投票以实现有效的人脸检测.《融合显著信息的 LDA 极光图像分类》则针对特定的极光图像,利用图像的谱残差显著图等语义信息,提出了一种融合极光图像显著信息的潜在主题方法.

作为一期与全国性学术会议合作的专刊,对于特约编辑而言,面临着会议投稿与专刊投稿之间的平衡.有部分优秀稿件,由于容量等原因,非常遗憾无法列入本专刊发表.我们要特别感谢《软件学报》编委会和编辑部,从专刊的立项到征稿启事的发布,从审稿专家的邀请到评审意见的汇总,乃至最后的定稿、出版与发行,他们付出了大量的劳动和辛勤的汗水.

我们还要感谢各位审稿专家.在专刊文章的多轮评审过程中,评审专家都在非常紧迫的时间内返回了高质量的评审意见.这些意见,均对论文的出发点、创新点,乃至论文的理论推导、实验验证进行了非常细致和中肯的评阅.这些评审意见极大地提高了本专刊论文的质量.

最后,我们要感谢读者们,希望本专刊能够为你们提供有益的参考.

高阳(1972－),男,江苏淮阴人,博士,教授,博士生导师.现任南京大学计算机科学与技术系副主任,中国计算机学会高级会员,中国计算机学会大数据专家委员会委员,中国计算机学会人工智能与模式识别专业委员会委员,中国人工智能学会粗糙集与软计算专业委员会副主任,中国人工智能学会机器学习专业委员会秘书长,江苏省计算机学会人工智能专业委员会副主任.先后入选教育部新世纪优秀人才计划、江苏省"333 高层次人才培养工程"和江苏省"六大人才高峰"等计划.主要研究领域为机器学习、人工智能、大数据处理.

E-mail: gaoy@nju.edu.cn

陈松灿(1962－),男,博士,教授,博士生导师.现任中国人工智能机器学习专委会副主任委员,中国计算机学会高级会员.曾发表 SCIE 学术论文 100 多篇,其中有 3 篇论文先后获得 Pattern Recognition 的双年度最佳论文提名奖(Honorable Mentions).已培养毕业博士生 30 名,其中有两位分别获得 2006 年和 2011 年全国百篇优博论文提名奖.主要研究领域为模式识别、机器学习、智能(大)数据分析.

E-mail: s.chen@nuaa.edu.cn