软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(11):2749−2751 [doi: 10.13328/j.cnki.jos.004909]

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

大数据时代的机器学习研究专刊前言*

何晓飞[1], 郭茂祖[2], 张敏灵[3]

[1](计算机辅助设计与图形学国家重点实验室(浙江大学),浙江 杭州 310058)

[2](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[3](东南大学 计算机科学与工程学院,江苏 南京 210096)

通讯作者: 何晓飞, E-mail: xiaofeihe@cad.zju.edu.cn

中文引用格式: 何晓飞,郭茂祖,张敏灵.大数据时代的机器学习研究专刊前言.软件学报,2015,26(11):2749−2751. http://www.jos.org.cn/1000-9825/4909.htm

进入 21 世纪以来,科学研究与社会生活各个领域中的数据正在以前所未有的速度产生并被广泛收集与存储.如何实现数据的智能化处理从而充分利用数据中蕴含的知识与价值,已成为当前学术界与产业界的共识.机器学习作为一种主流的智能数据处理技术,是实现上述目标的核心途径.传统机器学习研究通常假设数据具有相对简单的特性,如数据来源单一、概念语义明确、数据规模适中、结构静态稳定等.当数据具有以上简单特性时,基于现有的机器学习理论与方法可以有效地实现数据的智能化处理.然而,在大数据时代背景下,数据往往体现出多源异构、语义复杂、规模巨大、动态多变等特殊性质,为传统机器学习技术带来了新的挑战.本专刊选题为"大数据时代的机器学习研究",反映我国学者在大数据机器学习等领域的部分近期研究成果.

专刊公开征文历经两轮,共征得投稿 50 篇.此外,专刊组稿与第 15 届中国机器学习会议(CCML 2015)合作,从 399 篇会议投稿中遴选出了 8 篇高质量论文.上述稿件涉及大数据机器学习在理论、算法以及应用等诸多方面的研究内容.特约编辑先后邀请了 60 余位机器学习及相关领域的专家参与审稿工作,每篇投稿邀请 2 位专家进行评审.稿件评审历经 6 个月,经初审、复审、CCML 2015 会议宣读和终审各个阶段,最终有 21 篇论文入选本专刊.

首先,如何面向大数据的特殊性质,设计适于大数据分析处理的学习算法是大数据机器学习研究的重点.

《一种减小方差求解非光滑问题的随机优化算法》证明了一类重要的非光滑损失随机优化算法 COMID 具有方差形式的收敛速率,通过引入方差减小策略得到一种适于求解大规模机器学习问题的随机优化算法 α-MDVR.

《一种异构直推式迁移学习算法》针对特征空间异构的迁移学习问题,采用无监督匹配源领域和目标领域特征空间的方法学习映射函数,将源领域已标注数据迁移至目标领域,以提升目标领域未标注数据的直推泛化性能.

《覆盖学习的道路优化算法》考察基于李群连通性的多联通覆盖学习算法,基于 Fisher 投影的思想改进该算法中最优覆盖道路可能存在交叉或邻域重合的问题,在图像分类等问题上取得了更好的分类性能.

《一种大数据环境中分布式辅助关联分类算法》针对大数据环境下的辅助分类问题,提出了一种分布式辅助关联分类算法,充分考察分布式数据集在空间上存在的类别分布差异以及时间上存在的类别动态迁移现象.

《用于多标记学习的分类器圈方法》考察分类器链方法在标记相关性建模时存在的标记次序选择问题,提出了一种基于分类器圈结构的多标记学习方法,通过圈结构依次迭代训练与各标记对应的分类器.

《大数据的密度统计合并算法》针对聚类算法在大数据环境下的可扩展性问题,提出一种基于抽样的大数据密度统计合并算法 DSML.该算法将特征视为独立随机变量,并根据独立有限差分不等式获得统计合并判定准则.

* 收稿时间: 2015-10-01

《求解大规模谱聚类的近似加权核 *k*-means 算法》证明了基于归一化图划分的谱聚类与加权核 *k*-means 聚类在目标函数上具有等价性,并基于此提出了一种适于大数据谱聚类问题的近似加权核 *k*-means 算法.

《一种基于近邻表示的聚类方法》提出了一种基于近邻表示的数据压缩表示方法,通过寻找样本在训练集中的 *k* 最近邻获得样本的压缩 0-1 向量表示.实验结果表明,该方法在保持聚类性能的同时可有效降低存储空间开销.

《普适性核度量标准比较研究》考察核方法的核函数选择及参数优化问题,总结了 5 种流行的普适性核度量标准,并基于模拟数据与真实数据对这 5 种普适性核度量标准的性质和有效性进行了对比分析.

其次,作为智能数据处理的核心技术,机器学习对互联网、计算机视觉等大数据应用起着广泛的支撑作用.

《大数据环境下移动对象自适应轨迹预测模型》考察大数据环境下移动对象海量轨迹预测问题,提出了一种基于密度聚类技术的隐马尔科夫自适应轨迹预测模型 SATP,并设计实现了相应的移动对象轨迹预测系统.

《基于多尺度时间递归神经网络的人群异常检测》针对密集场景中的人群异常检测所面临的人群密度大、变化快、存在大量遮挡等挑战,提出了一种基于多尺度时间递归神经网络的人群异常事件检测和定位方法.

《视频人脸识别中判别性联合多流形分析》考察不受控环境下的视频人脸识别问题,基于类间流形与类内流形分别表示视频图像集的平均脸信息与原始图像信息,提出了一种判别性联合多流形视频人脸识别方法.

《基于差值局部方向模式的人脸特征表示》针对人脸识别中的人脸特征表示问题,提出了一种基于差值局部方向模式的人脸特征表示方法 DLDP,比 LBP 等经典局部表示方法包含更为丰富的结构及细节信息.

《基于 Pivots 选择的有效图像块描述子》针对核描述子方法生成图像块特征时计算复杂度高的缺陷,提出了一种基于不完整 Cholesky 分解自动筛选少量标志性图像块的高效图像块特征表示算法.

《基于多学习器协同训练模型的人体行为识别方法》针对传统行为识别方法需依赖大量有标记训练样本的问题,提出了一种基于多学习器协同训练模型的半监督人体行为识别方法,可有效利用未标记样本提升识别性能.

《一种大数据环境下的在线社交媒体位置推断方法》针对位置数据在社交媒体大数据中呈现的稀疏性,通过分析本地词语、社交关系与地理位置三者之间的关系,提出了一种基于用户生成内容的位置推断方法.

此外,机器学习对数据挖掘、计算智能等大数据相关领域的研究起着积极的促进作用.

《面向模式图变化的增量图模式匹配》考察数据图不变而模式图发生变化的增量图模式匹配问题,通过结合改进的 GPM 算法以及面向模式图的增边及减边操作,实现了适于大数据图的增量图模式匹配算法 PGC_IncGPM.

《基于评分矩阵局部低秩假设的成列协同排名算法》针对现有协同过滤方法过于关注评分预测准确性的问题,基于评分矩阵的局部低秩假设,提出了一种结合协同过滤与排名学习技术的成列协同排名算法 LLCR.

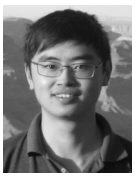
《带间隔约束的 Top-*k* 对比序列模式挖掘》考察间隔约束条件下的对比序列模式挖掘问题,通过引入 top-*k* 约束来解决传统对比序列模式挖掘需预设支持度阈值的问题,并基于剪枝启发策略提升算法执行效率.

《大数据分析中的计算智能研究现状与展望》对大数据分析中的计算智能方法进行综述,从人工神经网络、模糊系统、演化计算和群体智能这 3 个方面讨论了大数据计算智能的已有工作以及面临的主要问题.

《教育数据挖掘研究进展综述》对大数据背景下教育数据挖掘的近期研究工作从数据来源、研究方法、研究结果等方面做了介绍及对比分析,并对现有研究工作的不足及未来发展趋势进行了讨论.

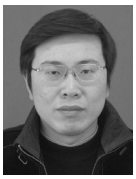
本专刊主要面向机器学习、人工智能、数据挖掘等相关领域的研究人员,反映了我国学者在大数据机器学习等领域的最新研究进展.在此,我们要特别感谢《软件学报》编委会对专刊工作的指导和帮助,感谢编辑部各位老师从征稿启示发布、审稿专家邀请至评审意见汇总、论文定稿、修改及出版所付出的辛勤工作和汗水,感谢专刊评审专家及时、耐心、细致的评审工作.此外,我们还要感谢向本专刊踊跃投稿的作者对《软件学报》的信任.

最后,感谢专刊的读者们,希望本专刊能够对相关领域的研究工作有所促进.



何晓飞(1978－),男,四川成都人,博士,教授,博士生导师.现任 CAAI 机器学习专业委员会常务委员,CCF 人工智能与模式识别专业委员会委员.曾获国家杰出青年科学基金(2011 年),担任 TKDE,TCYB,CVIU 等多个学术期刊编委.主要研究领域为机器学习,人工智能,计算机视觉.

郭茂祖(1966－),男,博士,教授,博士生导师.现任 CAAI 机器学习专业委员会常务委员,CCF 人工智能与模式识别专业委员会委员.曾获黑龙江省杰出青年科学基金(2006 年),担任中国机器学习会议(CCML 2015)大会主席.主要研究领域为机器学习,计算生物学,图像理解.



张敏灵(1979－),男,博士,教授.现任 CAAI 机器学习专业委员会秘书长,CCF 人工智能与模式识别专业委员会委员.曾获 CCF 优秀博士学位论文(2008 年),国家自然科学基金优秀青年科学基金(2012 年)等.主要研究领域为人工智能,机器学习,数据挖掘.