软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(4):1153−1155 [doi: 10.13328/j.cnki.jos.006484]

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

面向开放场景的鲁棒机器学习专刊前言 *

陈恩红[1], 李宇峰[2], 邹 权[3]

[1](中国科学技术大学 计算机科学与技术学院, 安徽 合肥 230026)

[2](南京大学 人工智能学院, 江苏 南京 210023)

[3](电子科技大学 基础与前沿研究院, 四川 成都 611731)

通信作者: 陈恩红, E-mail: cheneh@ustc.edu.cn; 李宇峰, E-mail: liyf@nju.edu.cn; 邹权, E-mail: zouquan@nclab.net

中文引用格式: 陈恩红, 李宇峰, 邹权. 面向开放场景的鲁棒机器学习专刊前言. 软件学报, 2022,33(4):1153−1155. http://www.jos.org.cn/1000-9825/6484.htm

近年来, 随着学术界与工业界在机器学习和人工智能领域越来越多的投入和关注, 相关技术获得飞速发展, 机器学习已经被应用到社会生活的方方面面, 并产生巨大社会价值.

机器学习模型主要依赖大量高质量数据的封闭训练, 随着机器学习模型付诸于开放场景, 例如, 数据分布的变化、数据特征的变化、数据标记的偏差、任务目标的变化、恶意样本的攻击、设备能力的受限等, 其往往面临模型失效、性能不佳等风险隐患. 基于此, 研究人员亟需探索开放场景下的鲁棒机器学习模型. 具体而言, 包括分布变化的机器学习、弱监督学习、模型复用、表示学习、强化学习、对抗学习、迁移学习以及更多实际领域问题中的应用等. 为此, 我们组织了面向开放场景的的鲁棒机器学习专刊.

通过两轮征稿共收到投稿 53 篇. 特约编辑先后邀请了 60 余位国内机器学习领域的知名专家参与审稿工作, 每篇投稿至少由 2 位专家进行评审. 最终有 23 篇论文被本专刊录用. 录用论文涉及开放域机器学习、任务关系利用与优化、新型特征表示与聚类、面向特定领域的机器学习模型等, 一定程度上反映了我国在该专题下的研究水平. 根据主题, 本专刊论文大致可分为 4 组.

(1) 开放域机器学习

《考虑多粒度类相关性的对比式开放集识别方法》发现已知类的数据表示是开放集识别的关键因素, 基于此设计多粒度损失函数, 通过约束类间关系, 增强开放集数据表示能力和识别精度.

《基于自监督知识的无监督新集域适应学习》针对开放场景下任务缺乏类别信息的问题, 提出了基于源域样本对比知识的无监督域适应方法, 利用图自监督分类损失, 缓解域间无共享类别的负迁移现象.

《伪标签不确定性估计的源域无关鲁棒域自适应》针对开放场景下源域数据难以获取的问题, 提出了面向不确定伪标签的估计方法, 依据预训练的源域模型和无标签目标域数据, 提升域适应鲁棒性.

《开放环境多分布特性的局部敏感哈希检索方法》面向开放环境的多种分布特性, 设计了基于 Laplacian 算子的局部敏感哈希搜索方法, 分析了低投影密度区间分割的有效性能适应大规模高维检索的鲁棒性需求.

《双标签监督的几何约束对抗训练》设计了一种针对几何结构的新型约束方法, 保持自然样本与对抗样本的特征空间分布一致性进行对抗训练, 有效提升了模型的防御能力.

《基于 AdaGrad 的自适应 NAG 方法及其最优个体收敛性》提出了一种自适应的加速梯度方法, 面向约束非光滑凸优化问题, 具有最优的个体收敛速率.

(2) 任务关系利用与优化

《标签推荐方法研究综述》从任务内容的多样性、标签之间的相关性以及用户偏好的差异性, 对标签推荐主流方法进行了梳理和剖析, 并展望了当前标签推荐领域面临的主要挑战.

《基于标记因果顺序挖掘的多标记分类方法》针对多标记学习的标记关系挖掘, 从标记关系因果角度出

* 收稿时间: 2021-09-28

发, 提出了基于标记因果顺序的两种新型标记关系挖掘和利用方法, 有效提升了学习性能.

《利用标注者相关性的深度生成式众包学习》结合深度神经网络的优势与标注者相关性, 提出了一种深度生成式众包学习方法, 实现了完全贝叶斯推断, 能够自适应匹配数据及模型复杂度.

《概念漂移数据流半监督分类综述》从问题设置到模型划分对概念漂移数据流半监督分类的学习方法进行了梳理, 剖析了数据分布变化、分类模型更新策略等因素对性能的影响.

(3) 新型特征表示与聚类

《特征演化的置信-加权学习方法》针对特征更替和消失问题, 提出了基于二阶信息的特征演化置信-加权学习算法, 融合利用新旧特征进行集成, 促进模型性能的鲁棒提升.

《基于随机近邻嵌入的判别性特征学习》利用随机近邻嵌入, 提出了一种同时考虑数据判别信息与拓扑结构的特征学习方法, 有效促进特征的表示能力.

《基于可辨识矩阵的完全自适应 2D 特征选择算法》提出了基于可辨识矩阵的自适应特征选择算法, 迭代式选择重要特征与剔除冗余特征, 减少人工参与, 取得有竞争力性能.

《基于多阶近邻融合的不完整多视图聚类算法》针对不完整多视图聚类性能欠佳的问题, 提出了基于多阶近邻扩散融合的聚类算法, 将不同阶的深层结构信息进行非线性融合, 提升算法鲁棒性和灵活性.

《CMvSC: 知识迁移下的深度一致性多视图谱聚类网络》同时考虑单视图局部不变与多视图全局一致的特点, 提出了深度一致性多视图谱聚类网络, 提升谱聚类的泛化能力与可扩展能力.

《基于 K 近邻和优化分配策略的密度峰值聚类算法》面向簇间密度分布差异, 提出了一种利用 K 近邻和优化分配策略的密度峰值聚类算法, 在大量数据集和多组度量指标上表现出鲁棒的性能.

(4) 面向特定领域的机器学习模型

《类脑超大规模深度神经网络系统》借鉴人脑功能分区以模块化构建神经网络模型的思路, 提出了一种受人脑功能机制启发的大规模神经网络系统, 具备多通道协同处理、知识存储迁移、持续学习等能力.

《ReChorus: 综合高效易扩展的轻量级推荐算法框架》从常规推荐、序列推荐到引入知识图谱、时间动态性的推荐等方面, 实现了综合、高效、易扩展的轻量级推荐算法框架.

《面向知识产权的科技资源画像构建方法》从科技数据获取、命名实体识别等方面, 利用双向长短时循环网络和知识图谱, 实现了一种面向知识产权的科技资源画像构建方法.

《基于数据场聚类的共享单车需求预测模型》面向共享单车调度中天气、时间等动态因素, 利用多特征 LSTM 提出了基于站点聚类的共享单车需求预测算法, 显著促进性能提升.

《基于预测编码的样本自适应行动策略规划》面向强化学习状态表征困难和数据效率低下的问题, 提出了基于预测编码的样本自适应行动策略规划, 在全国兵棋推演比赛中取得优异效果.

《面向手术器械语义分割的半监督时空 Transformer 网络》面向手术视频数据标注成本高昂的问题, 提出了带有时空 Transformer 的半监督分割框架, 在手术器械分割挑战赛数据集上取得明显提升.

《基于 StarGAN 和类别编码器的图像风格转换》针对生成对抗网络对精细图像特征描述能力的不足, 提出了能够结合小样本和图像风格转化的生成对抗网络模型, 提升图像风格转化能力.

本专刊主要面向机器学习、数据挖掘等多个领域的研究人员和工程技术人员, 反映了我国学者在机器学习领域最新的研究进展. 感谢《软件学报》编委会及编辑部、中国人工智能学会机器学习专委会对专刊工作的指导和帮助, 感谢专刊全体评审专家及时、耐心、细致的评审工作, 感谢踊跃投稿的所有作者. 希望本专刊能够对我国开放场景下的鲁棒机器学习研究有所促进.

陈恩红(1968－), 男, 博士, 教授, 博士生导师, 中国科学技术大学大数据学院执行院长, 计算机科学与技术学院副院长, 中科大智慧城市研究院(芜湖)院长, 大数据分析及应用安徽省重点实验室主任, 安徽省计算机学会理事长. CCF 会士, CAAI 会士, 国家杰出青年基金获得者, 科技部重点领域创新团队负责人, 中组部“万人计划”科技创新领军人才.《IEEE Transactions on Knowledge and Data Engineering》《IEEE Transactions on System Man and Cybernetics: System》《ACM Transactions on Intelligent Systems and Technology》《Knowledge and Information System》等期刊编委. 曾获 KDD 2008 最佳应用论文奖、KDD 2018 最佳学生论文奖、ICDM 2011 最佳研究论文奖, 教育部自然科学一等奖.

李宇峰(1983－), 男, 博士, 南京大学人工智能学院副教授, 博士生导师. 入选 CCF 优博、CCF 首届青年人才发展计划.从事机器学习、数据挖掘的研究, 相关成果在领域重要期刊会议发表论文 60 余篇. 现为中国人工智能学会机器学习专委会秘书长. 担任 IEEE BigComp20、MLA20、CCML21 共同程序主席, ACML21 Journal Track 共同主席, ACML19 Tutorial 和 ACML18 Workshop 共同主席; 担任《Machine Learning》《Neural Network》等国际期刊编委, 《Frontiers of Computer Science》青年编委; 长期担任 AAAI、IJCAI、ICML 等重要国际会议领域主席/资深程序委员.

邹权(1982－), 男, 博士, 电子科技大学基础与前沿研究院教授, 博士生导师. 主要研究领域为生物信息学、机器学习和字符串算法. 目前担任《Current Bioinformatics》主编和多个 SCI 期刊的编委; 入选科睿唯安 2018、2019 年全球高被引学者; 2019 年获得国家自然科学基金优秀青年基金资助; 其中代表作发表在《Bioinformatics》《PLOS Computational Biology》《RNA》等知名学术期刊上. 相关论文被《Nature》子刊引用; 率先采用 MapReduce 并行框架和字符串算法突破了多序列比对难题的计算瓶颈, 相关软件被美国、欧洲、印度科学院院士高度评价, 并被中国科学院官网、新浪科技等媒体报道; 提出的集成分类算法不但是学术期刊 Neurocomputing 官网下载次数最多的热点论文之一, 而且得到产业化应用, 用于百度贴吧的反作弊系统, 受到百度主题研究项目资助和百度公司官方报导.