软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):2811−2813 [doi: 10.13328/j.cnki.jos.005353]

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

复杂环境下的机器学习研究专刊前言*

胡清华[1], 张道强[2], 张长水[3]

[1](天津大学 计算机科学与技术学院,天津 300350)

[2](南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

[3](清华大学 自动化系,北京 100084)

通讯作者: 张长水, E-mail: zcs@mail.tsinghua.edu.cn

中文引用格式: 胡清华,张道强,张长水.复杂环境下的机器学习研究专刊前言.软件学报,2017,28(11):2811−2813. http://www.jos.org.cn/1000-9825/5353.htm

机器学习作为人工智能的核心技术,已经在许多领域得到应用,并发挥了重要作用.国务院近日印发了《新一代人工智能发展规划》,提出了面向 2030 年我国新一代人工智能发展的指导思想、战略目标、重点任务和保障措施,部署构筑我国人工智能发展的先发优势,加快建设创新型国家和世界科技强国.机器学习技术已得到了学术界和工业界空前的关注.随着其应用范畴的不断拓展,开放环境下、复杂场景中的探索式学习、多任务协同学习等更具挑战性的任务显得尤为重要.为了应对这些挑战,有必要根据学习任务特性提出更灵活、更鲁棒、更自主、自进化的学习机制.本专刊选题为复杂环境下的机器学习研究,将突出目前对多源异构等复杂数据进行不确定性建模的机器学习方法及应用研究.

专刊公开征文,共征得投稿 43 篇(其中包括第 16 届中国机器学习学术会议(CCML 2017)推荐的 10 篇高质量论文).这 43 篇论文通过了特约编辑的形式审查,有 39 篇论文进入到评审阶段.特约编辑先后邀请了 72 位机器学习相关领域的专家参与审稿工作,每篇投稿至少邀请 2 位专家进行评审.稿件评审历经 7 个月,经初审、复审、CCML 2017 会议宣读和终审这 4 个阶段,最终有 22 篇论文入选本专刊.这 22 篇论文可以大致分为如下 3 个部分.

(一) 不确定性数据处理与建模

不确定性数据处理与建模是复杂环境下机器学习研究的支撑技术.

《不一致数据上精确决策树生成算法》针对不一致数据问题,对决策树生成算法进行改进,使其能够直接对不一致数据进行分类.对约束条件中特征对分类结果的影响进行了多方面的衡量,并依此调整特征的影响因子,使得决策树的节点分割更加精确,分类效果更优.

《基于随机抽样的模糊粗糙约简》将随机抽样引入传统的模糊粗糙集,使得属性约简的效率大幅提升.该约简方法能够保持统计下近似定义统计辨识度不变的属性子集,用抽样方法计算统计辨识度的样本估计值,基于此估计值可以对统计属性重要性进行排序,从而设计出一个快速的适用于大规模数据的序约简算法.

《一种基于密度的分布式聚类方法》提出了一种高效的基于密度的分布式聚类方法 MRCSDP.通过将数据拆分成若干数据块,采用分布式计算得到各数据块的局部密度,将局部密度整合为全局密度,进而计算聚类中心.所提出的方法能够更快速地处理大规模数据聚类问题.

《基于标记与特征依赖最大化的弱标记集成分类》提出了一种基于标记与特征依赖最大化的弱标记集成分类方法 EnWL.EnWL 方法在高维数据的特征空间多次利用近邻传播聚类方法,获得代表性的特征子集,从而降低噪声和冗余特征的干扰.该方法在每个特征子集上训练一个基于标记与特征依赖最大化的半监督多标记分类器,最后通过投票集成这些分类器实现多标记分类.

《一种利用关联规则挖掘的多标记分类算法》在经典关联规则算法的基础上进行改进,提出了基于矩阵分

* 收稿时间: 2017-09-25

治的频繁项集挖掘算法 MDC-FIM.结合标记之间的关联规则,从全局和局部两个角度出发对原有数据进行更新,在此基础上可以直接应用现有的多标记学习算法得到新的多标记分类模型.

《卷积神经网络特征重要性分析及增强特征选择模型》研究如何显式表达神经网络中的特征重要性,提出了基于感受野的特征贡献度分析方法.将神经网络特征选择与传统特征评价方法进行对比分析,将传统特征评价方法对特征重要性的理解结合入神经网络的学习过程中,辅助深度神经网络进行特征选择.

《基于显露模式的数据流贝叶斯分类算法》利用模式在对立类数据集合中的支持度信息,提出了基于显露模式的适应流数据应用的贝叶斯分类模型,使用一个简单的混合森林结构来维护内存中事务的项集,并采用一种快速的模式抽取机制来提高算法速度.

《记忆神经网络的研究与发展》介绍了强监督模型和弱监督模型的记忆神经网络和各自应用场景以及处理方式,总结了两类主要模型的优缺点,对两类模型的发展及其模型创新进行了简要综述,总结了各类新模型在处理自然语言过程中所起的关键作用.

(二) 多源异构等复杂数据的学习

多源异构等复杂数据是复杂环境下机器学习的重点研究对象.

《多源数据融合高时空分辨率晴雨分类》面向多源数据高时空分辨率晴雨估计问题,提出雷达、卫星和地面观测因子多视角构建方法,构建了 VisCAPPI、VisPPI、VisSat 和 VisGround 晴雨估计视角以及它们的组合视角,并设计了能够融合以上多源数据的多视角方法 MVWRF.

《一种基于局部分类精度的多源在线迁移学习算法》提出了一种基于局部分类精度的多源在线迁移学习方法,计算新样本与目标领域已有样本之间的距离,在各源领域分类器中挑选局部精度最高的分类器,并将其与目标领域分类器进行加权组合,从而实现多个源领域知识到目标领域的迁移学习.

《一种基于直推判别字典学习的零样本分类方法》针对零样本分类问题,提出了一种直推式的字典学习方法.首先通过构建判别字典学习模型,对带标签的可见类别样本的视觉特征和类别语义特征建立映射关系模型;然后针对可见类别和未见类别不同引起的域偏移问题,提出了一个基于直推学习的修正模型.

《面向认知的多源数据学习理论和算法研究进展》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.

《混杂数据的多核几何平均度量学习》提出了一种基于几何平均的混杂数据度量学习方法.采用不同的核函数将数值型数据和符号型数据分别映射到可再生核希尔伯特空间,从而避免了特征的高维度带来的负面影响.另外,提出了一个基于几何平均的多核度量学习模型,将混杂数据的度量学习问题转化为求黎曼流形上的两个点的中心点的问题.

《面向复杂时间序列的 k 近邻分类器》为了有效对齐并分类复杂时间序列,提升基于时序对齐的 k 近邻分类器的分类效果,提出了一种具有辨别性的局部加权动态时间扭曲方法,用于发现同类复杂时间序列的共同点,以及异类序列间的不同点.通过迭代学习时间序列对齐点的正例集与负例集获取复杂时间序列中每个特征的辨别性权重.

(三) 机器学习在复杂任务中的分析与应用

机器学习在复杂任务中的分析与应用在这次投稿中得到了广泛关注.

《用于糖尿病视网膜病变检测的深度学习模型》针对糖尿病眼底病变筛查工作,采用两级深度卷积神经网络,对原始照片进行特征提取和结果分类.实验结果表明,模型的输出结果与医生诊断结果之间具有高度的一致性.

《一种基于 RNN 的社交消息爆发预测模型》利用基于深度循环神经网络对社交消息的传播过程进行建模,对社交网络中消息的爆发进行预测.该方法能够自动学习消息传播过程的速率函数,不需要手动定义消息传播速率的特征函数,具有较强的数据场景适应性.

《基于前缀投影技术的大规模轨迹预测模型》针对海量移动对象轨迹数据,结合频繁序列模式发现的思

想,提出了基于前缀投影技术的轨迹预测模型 PPTP.首先挖掘频繁轨迹模式,构造投影数据库并递归挖掘频繁前序轨迹模式;然后进行轨迹匹配,以不同频繁序列模式作为前缀增量式扩展生成频繁后序轨迹,将大于最小支持度阈值的最长连续轨迹作为结果输出.

《移动应用安全生态链构建方法》从移动应用软件安全威胁发生的源头、途径和终端出发,实现了基于编程风格的源代码作者溯源追踪、移动应用安全加固及渠道监测、基于深度学习的移动应用安全检测,构建了移动应用安全生态链,保障了用户个人信息安全.

《基于代价敏感间隔分布优化的软件缺陷定位》提出了代价敏感的间隔分布优化(CSMDO)损失函数,并将代价敏感的间隔分布优化层应用到深度卷积神经网络中,能够良好地处理软件缺陷数据的不平衡性,提高缺陷定位的准确度.

《烟花算法优化的软子空间 MR 图像聚类算法》针对 MR 图像在聚类分割中易受随机噪声影响的问题,提出了一种基于烟花算法的软子空间 MR 图像聚类算法.该算法采用结合界约束与噪声聚类的目标函数,并提出了新的隶属度计算方法.另外,为了提升结果的鲁棒性,在聚类过程中引入自适应演化算法来平衡局部与全局搜索.

《miRNA 与疾病关联关系预测算法》针对预测 miRNAs 与疾病的关联关系问题,首先构建 miRNA-疾病双层网络模型,依据 miRNA 的功能相似度对其进行基于密度的聚类,进而将二分网络投影应用于聚类后的 miRNAs 及疾病集合构成的 miRNA-疾病双层子网中,结果表明其能够有效提升预测精度.

《一种加权稠密子图社区发现算法》利用复杂网络中的连接紧密程度或者可信度意义的权重等先验信息,提出了基于加权稠密子图的重叠聚类算法(OCDW).综合考虑网络拓扑结构及真实网络中边权重的影响,给出了一种网络中边的权重定义方法;进而给出种子节点选取方式和权重更新策略;最终得到聚类结果.

本专刊主要面向机器学习、数据挖掘等相关领域的研究人员,反映了我国学者在复杂环境下机器学习领域最新的研究进展.在此,我们要特别感谢《软件学报》编委会、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《软件学报》编辑部和中国人工智能学会机器学习专委会的各位老师从征稿启事发布、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,感谢专刊的评审专家、编辑和读者,希望本专刊能够对机器学习相关领域的研究工作有所促进.

胡清华(1976－),男,湖南娄底人,博士,教授,博士生导师,CCF 专业会员,在国际期刊和会议发表论文 150 余篇,研究成果应用于空间天气预报、智能无人驾驶和设备故障诊断等领域.主要研究领域为多模态学习,不确定性建模.

张道强(1978－),男,博士,教授,博士生导师,CCF 专业会员,担任学术期刊《PLOS ONE》和《自动化学报》的编委,主要研究领域为机器学习,模式识别,脑影像分析.

张长水(1965－),男,博士,教授,博士生导师,CCF 高级会员,担任学术期刊《Pattern Recognition》和《计算机学报》的编委,曾在国际期刊发表论文 100 余篇,在顶级会议上发表论文 50 多篇,主要研究领域为机器学习,模式识别,计算视觉.