中国科学: 信息科学 2017 年 第 47 卷 第 11 期: 1443–1444
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
机器学习专题

编者按

随着信息科学技术发展进入大数据时代, 人们在科学研究、生产消费、社会生活等各个领域中积累的数据正以前所未有的速度增长. 如何实现数据的智能化处理, 从而充分发掘利用数据中蕴含的知识与价值, 已成为学术界、工业界乃至各国政府的普遍共识. 作为一种主流的智能数据处理技术, 机器学习是实现上述目标的核心途径. 美国政府 2016 年 10 月发布了以机器学习为核心的国家人工智能研究与发展战略规划, 国务院 2017 年 7 月印发的《新一代人工智能发展规划》中将机器学习研究作为发展新一代人工智能基础理论和共性关键技术的重点任务. 机器学习近年来在理论、方法及应用等诸多方面取得了令人瞩目的研究进展, 正如 *Science* 近期发表的综述文章所称, 机器学习是当前发展最迅速的信息科学技术领域之一.

在大数据时代背景下, 数据分析任务往往面临模态多源异构、语义复杂多变、规模巨大迅增、环境动态开放等特性, 为机器学习技术的发展带来了新的挑战. 为及时反映国内同行在机器学习研究方面的最新成果, 进一步推动我国机器学习及相关领域的创新发展, 《中国科学: 信息科学》特组织本期“机器学习专题”. 此外, 专题组稿与 2017 中国计算机学会人工智能会议 (CCFAI 2017)、第十六届中国机器学习会议 (CCML 2017) 合作, 从 799 篇会议投稿中遴选高质量论文. 特约编辑先后邀请多位机器学习及相关领域的专家参与审稿工作, 稿件评审历经 7 个月, 最终有 8 篇论文入选本专题.

一方面, 如何面向特定任务和数据特点, 构建具有强泛化性能的学习模型是机器学习研究的核心内容.

王双成等人的论文 “小时间序列动态完全 Bayesian 集成分类器研究” 针对提高连续属性小时间序列分类的可靠性问题, 提出了一种动态完全贝叶斯集成分类模型 DFBEC. 该模型采用动态完全贝叶斯分类器增加属性变量为类变量提供的信息量, 并结合核函数概率联合密度估计、平滑参数优化、时序递进分类标准、分类器选择与平均等技术实现训练与泛化之间的均衡. 在宏观经济小时间序列数据上的实验表明, 基于 DFBEC 模型的小时间序列转折点预测具有良好的分类性能.

杜航原等人的论文 “一种基于优化模型的演化数据流聚类方法” 针对现有数据流演化聚类方法仅适于硬聚类且忽略聚类模型与概念漂移间相关性的问题, 提出了一种基于优化策略的模糊最大熵数据流演化聚类算法. 该算法利用模糊隶属度和信息熵定义聚类问题的优化目标, 并在优化模型获得最优解的情况下进行概念漂移检测. 在人造和真实数据集上的实验结果表明, 该文所提算法在有效控制计算量和存储空间的情况下, 其聚类精度和概念漂移检测精度均显著优于对比聚类算法.

吴西竹等人的论文 “领域知识指导的模型重用” 针对训练数据匮乏情况下如何重用已有模型辅助当前学习任务求解的问题, 提出了一种领域知识指导的模型重用框架 MRDK. 该框架将已有的预训练机器学习模型视为黑盒, 在不需要已有模型结构信息的情况下, 使用领域知识对已有模型进行筛选和结合以适应环境的变化. 在蛋白质功能预测与图像分类两个具体任务上的实验表明, MRDK 框架通过重用具有相同输入特征的已有模型, 可以显著提升当前任务上的模型性能.

吴英杰等人的论文 “指数衰减模式下基于矩阵机制的差分隐私流数据发布算法” 针对指数衰减模式下的流数据统计发布问题, 提出了一种基于矩阵机制的差分隐私流数据发布算法 DMFDA. 该算法

引用格式: 张敏灵, 周志华. 机器学习专题编者按. 中国科学: 信息科学, 2017, 47: 1443–1444, doi: 10.1360/N112017-00226

利用树状数组设计有效的策略矩阵与还原矩阵构造方法, 利用对角矩阵对策略矩阵进行调整以提高发布精度, 并利用对角矩阵结构特性对其对角元素快速求解以适应流数据发布的实时性要求. 真实数据上的实验结果表明, DMFDA 相比同类流数据指数衰减算法具有更高的数据发布精度且能适应流数据的高效发布需求.

另一方面, 应用驱动是机器学习研究的重要特征, 机器学习现已成为求解许多应用问题的基础支撑技术.

郭茂祖等人的论文 “一种基于多组学生物网络的癌症关键模块挖掘方法” 针对癌症关键基因模块挖掘问题, 从数据集成角度提出了一种集成多组学生物数据的癌症关键模块挖掘方法. 该方法通过引入 IncRNA 组学数据扩大多组学数据集成方法的广度, 在此基础上构造关键基因的异常调控网络, 挖掘与肺鳞癌相关的关键基因模块及其所影响的异常调控基因集合. 实验结果表明, 对于最终所得包含 15 个基因的两个关键基因模块, 能够很好地区分高低风险组并具有良好的预后性能.

乔少杰等人的论文 “一种基于空间编码技术的轨迹特征提取方法” 针对大规模移动对象轨迹数据的特征提取问题, 提出了一种基于空间编码技术的轨迹特征提取方法. 该方法基于 GeoHash 区域编码技术形成轨迹点空间索引结构 GeoHashTree, 通过计算角度变化点以及基于密度的提取点聚类得到特征点集合, 对角度变化点进行深层次特征提取. 大规模 GPS 数据集上的实验结果表明, 该文所提算法在保证聚类结果准确性的前提下可显著降低算法时间开销, 并能准确找到角度变化点实现特征点的有效挖掘.

胡海峰等人的论文 “基于多示例多标记迁移学习的蛋白质功能预测” 针对新完成测序物种的蛋白质功能预测任务, 将其抽象为多示例多标记迁移学习问题并提出了相应的学习框架 TR-MIML. 该框架将源域与目标域数据集中的多示例样本转化为单示例样本, 通过最小化投影空间源域样本与目标域样本中心点距离赋予源域样本不同权值, 并随机挑选少量目标域标记样本与源域样本结合生成多示例多标记训练集. 在两个新完成测序物种上的实验结果表明, TR-MIML 框架有助于相应物种的蛋白质功能预测.

郝占刚等人的论文 “基于监督联合去噪模型的社交网络链接预测” 针对社交网络链接预测问题, 提出了一种基于有监督矩阵去噪模型的链接预测算法. 该算法通过综合利用现有用户特征信息和链接信息, 通过求解权重矩阵范数最小化问题实现已有社交网络至理想社交网络的映射函数训练, 完成链接预测任务. 5 个真实数据集上的实验结果表明, 在小规模和大规模社交网络上本文所提算法均可显著提高链接预测的精度, 在保持较高计算效率的条件下可提供更准确的朋友推荐.

本专题主要面向机器学习领域的研究人员, 反映了我国学者在机器学习等领域研究的前沿进展. 在此, 我们要特别感谢《中国科学: 信息科学》编委会对专题工作的指导和帮助, 感谢编辑部各位老师从征稿通知发布、论文评审与意见汇总、论文定稿、修改及出版所付出的辛勤工作和汗水, 感谢专题评审专家及时、耐心、细致的评审工作. 最后, 感谢专题的读者们, 希望本专题能够对相关领域的研究工作有所促进.

特约编辑: 张敏灵　东南大学

周志华　南京大学