交互式可视化标注与主动学习：实验比较

Mohammad CHEGINI[1,2]，Jürgen BERNARD[3]，Jian CUI[2]，Fatemeh CHEGINI[4]，
Alexei SOURIN[2]，Keith ANDREWS[5]，Tobias SCHRECK[1]
[1] 格拉茨技术大学计算机图形与知识可视化研究所，奥地利格拉茨，8010
[2] 南洋理工大学计算机科学与工程学院，新加坡，639798
[3] 英属哥伦比亚大学信息可视化研究组，加拿大温哥华，V6T1Z4
[4] 马克斯-普朗克气象研究所，德国汉堡，20146
[5] 格拉茨技术大学互动系统与数据科学研究所，奥地利格拉茨，8010

摘要：监督式机器学习方法可自动分类新数据，且对数据分析非常有帮助。监督式机器学习的质量不仅依赖于使用的算法类型，也依赖于用于训练分类器的标注数据集的质量。训练数据集中的标注实例通常依赖于专业分析人员的手工选择与注释，且通常是一个单调与耗时的过程。标签可以在学习过程中为主动学习算法提供有用的输入，以自动确定数据实例的子集。交互式可视化标注技术是有前景的选择，它提供有效的视觉概览，分析人员可从中同时查看数据记录与选择项目标签。将分析人员置于循环中，生成的分类器可得到更高准确率。虽然交互式可视化标注技术的初步结果在某种意义上有前景的，考虑到用户标注可改善监督式学习，但是该技术的许多方面仍有待探索。本文使用 mVis 工具标注一个多元数据集以比较 3 种交互式可视化技术（相似图、散点矩阵与平行坐标图）以及主动学习。结果表明 3 种交互式可视化标注技术的分类准确率均高于主动学习算法，相对于散点矩阵与平行坐标图，用户主观上更偏爱使用相似图标注。用户也可以根据使用的可视化技术采用不同标注策略。

关键词：交互式可视化标注；主动学习；可视分析
https://doi.org/10.1631/FITEE.1900549