Exploring graph-based clustering and outlier detection algorithms

Li, J.

摘要

在大数据时代，从复杂且相互关联的数据集中提取有意义的信息，仍然是各应用领域面临的关键挑战。本论文研究聚类与离群点检测方法以支持数据分析。传统方法（如 k-均值、DBSCAN 和统计异常评分）严重依赖欧氏距离、密度假设或全局阈值，往往难以捕捉高维、含噪或动态演化数据中错综复杂的关系和上下文异常。基于图的方法（如最小生成树方法和谱模型）通过利用结构特性来揭示隐藏模式和偏差，提供了一种强有力的替代方案。我们的研究通过系统性的方法，证明了基于图的技术相较于其他一些方法的优越性，并强调了其在不依赖深度学习或神经网络情况下的适应性、高效性及现实世界适用性。

基于最小生成树的聚类擅长检测传统方法所遗漏的不规则形状簇和层次结构。通过构建以节点表示数据点、边编码相似性（例如通过 kNN 或高斯核）的图，诸如谱聚类之类的算法通过分析拉普拉斯图（一种捕捉全局连接模式的矩阵）来划分节点。与假设球形簇的 k-均值不同，谱聚类能适应复杂的几何结构，使其非常适合边界非线性的任务，如图像分割。在数据稀疏或分布不均的情况下，这些方法的表现优于基于密度的方法（如 DBSCAN），而传统方法在这些情况下往往会使簇分裂或合并无关组。

离群点检测同样受益于图结构。传统统计方法（如 Z-score、IQR）基于与均值的全局偏差来标记异常，常常忽略在其局部环境中正常但在全局上下文中异常的上下文离群点。谱分析离群点检测方法引入了一种新颖的基于图的方法来识别高维空间中的离群点。通过构建 k 近邻图并分析其拉普拉斯矩阵的特征值分布，SAOD 能根据离群点与众不同的谱模式有效区分离群点和内围值。该方法采用多元高斯核密度估计来量化离群程度。在真实数据上的实验表明，SAOD 对于高维数据普遍存在的应用（如医疗诊断和网络安全）尤其有价值。

许多离群点检测方法无法处理具有变化密度或不规则形状的数据集，并且需要事先了解离群点比例，这限制了它们在医疗应用中的实用性。自适应迷你 MST 离群点检测通过开发一种新颖的归一化距离度量并构建局部迷你 MST，解决了

密度依赖性方法的局限性，从而能够在不知晓离群点分布先验知识的情况下进行有效的离群点识别。该算法新颖的距离度量是根据局部数据密度动态进行归一化，从而能够在具有不同簇形状和密度的复杂数据集中精确识别离群点。在电子健康记录数据上的验证证明了 MMOD 的优越性。

基于聚类的离群点检测常常受到噪声敏感性和任意中心点选择的影响，导致在复杂数据集中出现误报。基于归一化 MST 与中心点选择的离群点检测将最小生成树的鲁棒性与创新的噪声感知中心点选择过程相结合。该方法通过在归一化的 MST 中进行最长边切割来迭代地划分数据，同时其先进的中心点选择算法考虑了聚类核心中的噪声污染。

本论文所展示的研究确立了基于图的聚类和离群点检测作为现代数据分析不可或缺的工具，在处理复杂、互联数据集方面提供了无与伦比的灵活性。通过推进理论基础、优化计算效率并展示实际影响，本工作架起了学术创新与实际部署之间的桥梁。未来的方向包括将这些方法扩展到流数据进行实时分析，以及开发可解释框架以增强对基于图的决策的信任，确保其在日益数据驱动的世界中的持续相关性。