基于轻量化聚类模型的高校教学数据挖掘与应用实践

曹秀平
青岛恒星科技学院，山东 青岛 266100
DOI: 10.61369/TACS.2026090037

摘　　要： 随着高等教育改革的不断推进，教育数字化也逐步成为高校教育教学改革的重要方向。在此背景下，如何有效挖掘和应用高校教务管理、在线教学平台等系统中的海量数据，也成了广大高校在数字化建设与改革进程中亟待思考的问题。本文立足这一背景，就基于轻量化聚类模型的高校教学数据挖掘与应用路径进行了探讨，旨在为广大高校教学管理的精准化与智能化提供可行技术路径。

关 键 词： 高校；轻量化聚类；教学数据；挖掘与应用

University Teaching Data Mining and Application Practice Based on Lightweight Clustering Model

Cao Xiuping
Hengxing University, Qingdao, Shandong 266100

Abstract: With the continuous advancement of higher education reform, educational digitalization has gradually become an important direction for teaching reform in universities. Under this background, how to effectively mine and apply massive data from university educational administration management, online teaching platforms and other systems has become an urgent issue for universities in the process of digital construction and reform. Based on this context, this paper discusses the teaching data mining and application paths of universities based on the lightweight clustering model, aiming to provide feasible technical approaches for the precision and intelligence of university teaching management.

Keywords: universities; lightweight clustering; teaching data; mining and application

当前，各类教务系统、数字教学平台在高校得到了广泛应用，它们每天都产生大量的教育数据（包括学生成绩、网络行为、课堂出勤等）[1]。这些数据如果可以得到有效的挖掘与应用，将会为高校个性化教学、教学评估等工作的高质量开展提供科学的决策依据。聚类分析属于教育数据挖掘的重要技术手段之一，可以将具备类似属性的学生自动分成若干群组，揭示数据内在结构。轻量化聚类模型通过算法优化以及数据结构精简，能够在保证聚类质量的同时，有效降低计算开销，助力教育教学数据的有效挖掘与利用[2]。对此，深入探究基于轻量化聚类模型的高校教学数据挖掘与应用路径正当其时，对于提升高等教育教学质量有着重要现实意义。

一、轻量化聚类模型的内涵与核心特征

（一）轻量化聚类模型

轻量化聚类模型（Lightweight Clustering Model）指的是不显著牺牲聚类质量的同时，借助算法优化以及数据结构压缩等方式来有效减少时间和空间复杂度的一种聚类分析方法。它的“轻量化”主要体现在三个方面，即处理速度快、部署门槛低以及资源占用少[3]。在高校教学数据挖掘中，该模型的应用核心任务是自动分群学生的多维特征数据，让同一群组的学生能够有着比较高的相似性，同时让不同群组的学生差异性明显，以此来解释其内在的结构。相较于传统的聚类算法来说，轻量化模型关注资源的有效利用与高效聚类，让数据挖掘能够真正下沉到学校服务器甚至是个人计算机之上，以此来有效降低高校教学数据挖掘、分析与利用的门槛和成本。

（二）代表性轻量化聚类算法及其教育适配性

目前，在教育数据挖掘领域主要有两类代表性轻量化聚类方法。首先，是BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）算法。它属于增量式的层次聚类法，核心在于引入聚类特征树（CF-Tree）结构，以样本数、线性求和、平方求和作为三元组来描述数据子集，把原始数据压缩成少数个聚类特征节点[4]。新数据出现时，算法只需要在CF-Tree中沿着路径更新相应的节点，而不需要重新处理所有的历史数据。增量处理机制使得BIRCH可以在一次数据扫描中完成聚类，计算复杂度是线性级别的，比需要保存全部数据的算法内存占用更低。相关数据显示，在同样的数据集上，BIRCH计算时间比K-means和GMM要少很多，特别适合于大样本教育数据分析。其次，是改进型K-means算法。它是从另一个角度进行轻量化的。传统K-means的主要计算瓶颈就是迭代过程中要反复

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

计算所有的样本和聚类中心距离[5]。改进策略有三种，第一，用Canopy算法事先确定初始聚类中心和簇数目，减少无效迭代，第二，用特征选择或者降维方法减少参与计算的数据维度，第三，用群体智能优化算法（改进的布谷鸟搜索、遗传算法等）加快中心点寻优。采用Gap值和ISSA K-means相结合的方式对学生成绩数据按照学期进行切分后，可以实现学生群体精准分类以及学业表现预测。两种轻量化聚类模型的共同优点在于，它们在普通服务器上即可完成百万级数据记录的聚类分析工作，不需要依赖于大规模的计算机群，这也为高校教学数据的常态化挖掘和应用奠定了坚实的技术基础。

二、基于轻量化聚类模型的高校教学数据挖掘与应用路径

（一）数据治理：多源教学数据的采集、整合与预处理

高校的教学数据类型较为多样，格式也各有不同，因此，在挖掘与应用前应当充分做好基础治理工作，为系统化的数据采集奠基[6]。首先，高校的教学数据主要来源于教学管理系统（存储学生的基本信息、课程成绩以及选课记录等信息）、在线教学平台（存储学生的作业完成情况、学习时长、讨论次数等信息）、图书馆系统（存储学生的图书借阅情况）等，这些数据平台分属于不同的系统，因此要先依托数据结构与仓库技术来对其进行跨系统数据整合与汇聚，如可以以学生编号为主键进行关联整合。其次，在数据处理过程中应当注重以下几点：一是缺失值的处理。对于那些关键字段缺失比例较低的数据，应当采用众数填充的方式来弥补，而如果缺失的比例较高，则可以考虑进行剔除处理；二是异常值的检测，如果某一项数据（如成绩、行为记录）等在逻辑上明显不正常，则将其列为异常数值；三是数据规范化处理，依托Z-score标准化或Min-Max归一化等方式来对差异较大的数据进行尺度统一，以此来减少量纲差异的影响；四是类别特征的编码，对于分类变量（如学生的年级、专业与性别等），可以借助独热编码或者标签编码来对其进行数值转换。在此基础上，可以借助K-prototype等支持混合类型数据的聚类方法来对数值型和类别型混合的教学数据进行补充处理。再者，教学数据治理涉及多个部门，因此，高校也应建立常态化的数据治理机制，明确教务、学工以及信息中心等各部门的职责以及相关共享规范，从而保证数据采集的及时性、准确性与一致性，为后续的聚类分析提供保障。

（二）特征工程：面向聚类目标的特征提取与降维

原始的教学数据在经过采集预处理之后，依然会存在冗余多、维度高等一些问题，如果这时直接运用聚类算法的话，不但可能增加计算开销，而且也可能淹没那些关键、有用的判别信息。而特征工程便是应对这一问题的有效途径，其主要旨在保留数据核心判别信息的同时，最大化地降低特征维度，从而促进聚类质量的有效提升。首先，是特征提取。基于高等教育特点，在原始数据中构造相应的高层特征，如可以将学生的在线平台学习时长、讨论情况、视频观看情况等融合为“积极性”特征；将学生各个学期成绩的标准差作为“稳定性”特征；将学生作业完成情况、课堂出勤情况等融合为“规范性”特征等，以此来为后续的结果解释提供便利[7]。其次，是特征降维。这里主要是借助统一流形逼近与投影（UMAP）或者主成分分析（PCA）等方式来开展。前者在保留数据全局和局部结构方面有着重要作用，因此比较适合于聚类前的数据处理，后者主要依托线性变换，在低位空间中映射高位特征，保留方差最大的主成分，进而在降低维度的基础上，保证数据整体结构。而在使用二者进行特征降维的同时，可以依托BIRCH算法来减少计算所需的时间并提升聚类质量。再者，在特征降维后，应当基于聚类效果的轮廓系数或者累计方差贡献率来确定特征维度，如果其在2-10维之间便可以满足大部分的教学数据聚类需求。最后需要注意的是，特征工程应当始终围绕聚类目标来灵活展开，如以学生分类管理为目标，则应当兼顾学生成绩、日常行为等维度特征；如果以学业预警为目标，则可以提取学生在学习过程中的成绩、作业完成情况等特征。特征选择完成之后，可以依托热力图、散点图等方式来检验区分度，以此来保证数据的有效性，精准反映学生群体的内在差异。

（三）算法实施：轻量化聚类模型的选型、训练与质量评估

对于轻量化聚类而言，算法实施是其核心所在。首先，要充分做好算法选型工作。例如，当教学数据集规模较大时（百万级以上），可以运用BIRCH算法来对数据子集展开概要表示。期间，对原始数据进行压缩处理，使其成为少量聚类特征节点，可在单次扫描中进行聚类操作，计算复杂度为线性级别；当教学数据集规模中等时，则可以运用K-means算法——通过Canopy算法预先确定初始聚类中心与簇数量以减少无效迭代，也可以运用群体智能优化算法来加速中心点寻优[8]。其次，是模型训练，如在BIRCH算法应用过程中，阈值参数控制子簇的最大半径，直接影响着聚类的粒度。如果阈值较大，那么子簇将会变得较为粗放，虽然计算效率得到了保证，但是精度可能反向下降；分支因子则影响着各个节点的最大子节点数、树结构的复杂度和实际内存占用。高校教学数据轻量化聚类的阈值一般设在0.1-0.5之间，分支因子设置在50-100之间。在此基础上，可以依托轮廓系数以及肘部法则来对簇的数量进行明确。再者，是质量评估。该环节是保证轻量化聚类质量的重心所在，而当前比较实用的评估指标包括Calinski-Harabasz指数（基于簇内与簇间离散度比值，数值越大则说明效果越好）、Davies-Bouldin指数（对簇间相似度进行评估，数值越小则说明簇间区分越清晰）、轮廓系数（衡量簇间分离度和簇内紧密性的综合指标，取值如果越接近于1则说明其聚类质量越好）。评估结果的应用一方面可以判断是否达到预期效果，另一方面也可以作为算法参数调整的重要依据，如果评估指标不理想的话，则可以通过回溯检查、参数调整等方式来进行优化。

（四）应用转化：聚类结果的教学场景落地与闭环优化

应用转化是基于轻量化聚类模型的高校教学数据挖掘的价值体现，对此，首先是结果解释与可视化呈现，基于聚类结果可以依托雷达图等方式来展现具体结果，揭示学生群体的内部共同特征并赋予其具体的教育语义标签[9]。其次，是干预策略设计。即

 | Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

结合不同标签的群体来制定针对性的干预策略，如针对那些学习困难的学生，可以设计学业指导、教学辅导等策略；针对那些存在偏科的群体，可以设计专科补习方案；针对那些课堂参与度不高的群体，可以设计相应的激励措施等等。通过针对性的干预策略设计，来让高等教育教学可以有的放矢地开展，推动因材施教理念的落实。再者，是闭环优化机制的建立，即高校应当加快建设轻量化聚类模型的闭环优化机制，每学期基于历史教学数据来对学生进行轻量化聚类与划分，学期中可以基于群体特征开展差异化、个性化的教学干预；期末则可以基于行为改善、成绩提升等方面的变化来评估策略效果，在此基础上，依托这些新数据来开展下一轮的聚类分析，进而形成“聚类分析—策略实施—效果评估—模型迭代”的完整闭环[10]。此外，聚类分析还可以在高等教育多个领域和维度进行有效运用，如可以在资源配置等方面应用，结合学生群体的个体差异，为其安排个性化的学习资源，设计差异化的课程安排等等；在教育质量评估方面则可以基于聚类分析来识别教师的教学效果，为教师优化教学模式奠基，推动高等教育质量更上一层楼。

综上所述，随着数字化时代来临，高等教育也掀起了数字化的改革浪潮。在此背景下，基于轻量化聚类模型来有效挖掘与应用教学数据，也成了高校教育改革的重要举措。对此，广大高校也应积极推进新算法、新模型的应用，深挖轻量化聚类模型的赋能价值，从而借助现代技术提升高校教学管理的精准化与智能化水平。

参考文献

[1] 周宏谦 . 基于模糊聚类最大树算法的高校课堂教学质量在线评估方法 [J]. 无线互联科技 ,2025,22(15):39-42.

[2] 巨金香 , 张福泉 , 黄锐 . 基于烟花算法优化 k 均值聚类的教学质量评估模型 [J]. 济南大学学报 (自然科学版),2022,36(6):755-760.

[3] 虞烨青 . 基于教学数据挖掘的课堂教学过程优化研究 [D]. 湖北师范大学 ,2025.

[4] 毛德军 , 刘骏洲 , 宋树成 , 等 . 基于数据挖掘的地面领航教学质量评价模型研究 [J]. 中国军转民 ,2024,(7):40-41.

[5] 单红宇 . 数据挖掘技术在高校教学质量监控与评估系统中的应用研究 [J]. 办公自动化 ,2022,27(8):41-43.

[6] 熊志杰 . 数据挖掘技术在学业预警系统中的应用 [J]. 集成电路应用 ,2025,42(10):260-261.

[7] 赵俊 . 基于数据挖掘技术的本科生学业预警模型研究 [J]. 大数据时代 ,2025,(3):65-69.

[8] 钟婷 . 数据挖掘在高校内部教学质量监控中的应用研究 [D]. 四川师范大学 ,2020.

[9] 刘亮亮 . 数据挖掘技术在高校教学质量评估中的应用 [J]. 教育信息化论坛 ,2019,3(12):9-10.

[10] 周红志 , 郝文清 , 董阜平 . 基于数据挖掘的高校教学质量评估体系的构建 [J]. 文教资料 ,2019,(28):171-172+132.

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License