

基于 K-means 聚类 的学生画像分析与构建

郝慧君, 叶展, 王周, 张静也, 陈伟

温州商学院, 浙江 温州 325000

DOI:10.61369/EST.2024050020

摘 要 : 随着教育数字化转型的深入推进, 本研究基于 K-means 聚类算法构建多维学生画像, 旨在为精准化教育供给提供数据支持。通过整合学生在学业表现、学习投入、职业生涯规划等多维度数据研究构建的可视化画像体系为个性化教学策略制定提供了新路径, 尤其在课程资源适配、学习监督机制优化等方面具有实践价值, 响应了国家教育信息化战略对数据驱动教学改革的政策要求。

关 键 词 : 聚类; 学生画像; 数据驱动; 教育改革

Analysis and Construction of Student Portraits Based on K-means Clustering

Hao Huijun, Ye Zhan, Wang Zhou, Zhang Jingye, Chen Wei

Wenzhou Business College, Wenzhou, Zhejiang 325000

Abstract : As the digital transformation of education advances, this study constructs multi-dimensional student profiles using the K-means clustering algorithm to provide data support for precise educational provision. By integrating multidimensional data on academic performance, learning engagement, and career planning, the visualized profile system offers a new approach to personalized teaching strategies. It particularly has practical value in course resource adaptation and the optimization of learning supervision mechanisms, responding to the policy requirements of the national education informatization strategy for data-driven teaching reform.

Keywords : clustering; student portrait; data-driven; education reform

引言

随着教育改革的不断深化, 借助大数据技术, 将数字化手段融入教学, 提升教育质量成为新趋势。在 2015 年的国际教育信息化大会上, 习近平总书记提出了“推动教育变革和创新, 构建网络化、数字化、个性化、终身化的教育体系”。近年来, 教育部出台多项相关政策和行动纲要也明确指出, 要利用大数据等技术“为教师、教学管理人员等进行教学决策、改善教学计划、提高教学质量、保证教学效果提供参考依据”, “完善教育管理信息化顶层设计, 全面提高利用大数据支撑保障教育管理、决策和公共服务的能力”^[1-2]。

学生画像来源于“用户画像”, 是由交互设计之父艾伦·库珀 (Alan Cooper) 提出, 建立在一系列真实数据之上的目标用户模型, 是真实用户的虚拟代表。目前, 已有的画像研究聚焦学生的学业水平, 对学生未来生涯规划及缺乏探讨。基于此, 本研究采用收集学生学业、消费和职业生涯规划等方面的数据, 采用聚类分析等机器学习算法对学生画像进行可视化展示, 结合在校学生特点, 构建了学生画像的流程框架, 通过可视化技术呈现学生个体画像和群体画像, 并在此基础上对构建的学生展开具体的个性化教学策略^[3]。

一、国内外研究现状

(一) 学生画像

通过分析学生在校产生的数据, 描绘出学生个体和群体画像, 分析学生特点, 精准服务学生, 提升高校教育的精细化, 提升教学质量。其中, 张娜通过记录并分析学生各种在校活动的数据, 将学生信息抽象成标签, 帮助教师快速识别学生, 督促其养成良好的学习习惯, 制定个性化的帮扶措施, 改进教学设计^[4]。赵

国亮以高校学生成绩数据和消费数据为例, 实现用户画像模型并围绕高校学生自画像的研究前景, 挖掘高校数据价值, 为后期勤工助学、爱好培养、社团选拔、学生挂科预警、学生未在校预警等提供参考^[5]。丁凯借助大数据、云计算、智慧校园等技术, 实现对大学生日常生活轨迹和思想政治状况的信息化收集和存储, 通过对个体偏好、行为习惯、价值倾向等内容的分析和预测, 能够全方位、多角度地探究需求的层次性和多样性, 精准定位个性化的精神与文化需求^[6]。

课题信息: 本文系浙江省教育厅科研项目资助项目, 项目名称: 工程教育认证背景下基于学生画像的教育精准供给研究 (编号: Y202352174)。

二、基于精准教育供给的学生画像流程框架

学习者在学习过程中产生的行为数据是具有多维度的，显示着学习者多方面的特性，并且各数据之间也存在内容的相关性，学生画像的设计是基于学习者的个体差异及群体共性^[7-8]。本研究收集学生在校产生的多维数据，采用定量学生画像构建的方法，具体指导学生画像的分析、刻画和呈现（如图1）。



图1 学生画像构建流程图

（一）学生画像标签

学习者是学习活动的主体，学习者所具有的认知的、情感的、社会的特征都将对学习的信息加工过程产生影响^[9]。学习者的特征主要分为两个方面：智力因素和非智力因素。其中，智力因素有关的特征主要包括知识基础、认知能力和认知结构变量，非智力因素有关的特征包括兴趣、动机、情感、意志和性格。所以在设计学习者画像前，可以从学习者的智力因素和非智力因素出发，了解学习者的认知能力与结构、兴趣偏好、学习表现、学习结果等各方面特征。因此，选用学习媒介偏好、学习风格为标签确定学习者的个体特征；以学习投入为标签确定学习者的学习表现；以知识点掌握水平为标签确定学习者的学习结果。

学习指标：{ 学习媒介偏好；学习风格；学习投入；学习结果 }

同时，在构建学生画像时，应综合多维度的考虑学生在校，了解学生的“真实”，所以，在学生画像的标签设定时，不能简单的从学习方面刻画，还应该考虑学生在校对自己的职业生涯规划并需要结合在学校社交、消费等方面。

非学习指标：{ 职业生涯 }

（二）画像构建

K-Means 算法是一种无监督的机器学习算法，广泛应用于数据聚类。在学生画像中，K-Means 可以帮助我们理解和分组学生的不同特征和行为模式。K-Means 算法是将整个数据集划分为 K 个簇群，通过计算数据点之间的距离来确定它们之间的相似度，使得同一簇群内的数据点尽可能相似，而不同簇群之间的数据点则具有较大的差异。这种基于距离的相似度度量方式，使得 K-Means 算法能够有效地识别出数据集中的内在结构和群组。

三、学生画像的构建与呈现

（一）学生画像的关键属性

1. 学习媒介。随着科技的进步，知识的获取方式日趋多样化，从传统的纸质材料、传统课堂到电子讲义、视频教程以及在线互动课程等，每一种知识的获取方式都为学生提供了独特的学习体验。因此，学习媒介偏好的标签可以描述为：媒介偏好：{ 线下课堂；在线课堂；混合式课堂 } × { 纸质材料（书籍、笔记）；电子材料（包括互联网查询）}。

2. 学习风格。英国教育专家皮特·霍尼（Peter Honey）和阿伦·芒福德（Alan Mumford）认识到量身定制学习方法的价值，并结合学习者在学习过程中的行为与风格，开发了 Honey & Mumford 学习风格模型，确定了四种不同的学习风格：实践型、理论型、反思型和实用型。教师可以根据学生的不同学习风格设定个性化学习需求，提供更为贴切的教学方法和材料，从而提高教学效果。因此，学习风格的标签可以描述为：学习风格：{ 实践型；实用型；反思型；理论型 }

3. 学习投入。学习者的学习投入一般分为行为、情感、认知三个维度。行为投入指学习者在学习过程中参与学习活动的程度，包括参与学校内外的学业或非学业活动；情感投入又称心理投入，指学习者对学业任务或他人的积极情感反应，以及对学校的归属感；认知投入是一种“思维训练”，包括学习者在学习时使用的认知策略和心理资源的思维或心智。因此，学习风格的标签可以描述为：学习投入：{ 行为；情感；认知 }。

4. 学习结果，即检验学生的知识掌握水平。

学习结果是学生学习过程的直接体现，它真实地反映了学生对知识的理解和掌握程度。对于大数据科学系的学生，一般课程知识可以分为陈述性、应用性和设计性三类知识，通过对学生学习测试结果的分析，可以发现学生在不同的知识维度的掌握情况，深入了解学习者在不同知识维度上的掌握情况，这有助于教师更精确地评估学生的学习状态，并为他们提供针对性的指导和支持。所以，学习结果的标签可以描述为：学习结果：{ 优；良；中；差 }

5. 职业生涯规划

职业生涯画像是一个全面、多维度的个人职业发展蓝图。作为数据科学与大数据技术专业的大学生，希望在毕业时掌握扎实的数据科学基础理论和专业知识，具备数据科学思维能力、大数据应用研究与开发能力，具备复杂工程问题的研究分析能力和工程实践能力，并能够通过继续教育或其他终身学习途径不断拓展自己的能力，在大数据系统的研发、部署与应用等相关领域具有就业竞争力，所以，职业生涯的标签可以描述为：职业生涯：{ 工作；创业；升学 }

（二）学生数据分析

1. 媒介偏好

本研究将媒介偏好分为两个部分，教学模式和学习材料，其中，教学模式分为传统的线下教学模式、线上教学模式和混合式教学模式，学习材料类型分为传统纸质材料包括但不限于参考

书、学习笔记、打印版讲义等和电子材料包括电子书以及互联网查询得出的电子材料。从学生的调研结果看,线下、线上和混合式的偏好比例分别为:20.6%、27.0%和52.4%,更偏重于混合式教学模式。学习材料的偏好分析中,实物材料和电子材料分别占比78.3%和21.7%,出现了明显的差异。

2. 学习风格

根据 Honey 和 Mumford 的学习风格模型,通过对学生的问卷分析,实践型(41.8%)与实用型(28.0%)学习者占据主导地位,而反思型(15.9%)和理论型(14.3%)占比较低,这一分布特征与课程特性、院校定位及教学模式形成三重耦合效应。《UI前端设计》作为典型的理实一体化的项目驱动的应用型科目,课程内容工具性强、迭代性高和即时反馈等特征。实践型学习者“行动优先”的特质与前端开发“所见即所得”的编码特性高度契合,学生通过实时预览界面效果获得正向激励,这种即时反馈机制显著强化了其“编码-测试-优化”的行为模式。实用型学习者的优势则体现在技术选型环节,其“技术落地导向”的思维特征促使学生在框架选择时优先考虑使用代码实现网站效果,而非底层实现原理,这种务实态度与行业快速迭代需求形成共振。

3. 职业生涯规划

对大学生而言,学校生活是连接学习和未来职场的纽带,通过对未来职业的畅想与规划可以帮助学生认清自身兴趣、能力和优势,促使学生思考人生的长远目标,培养责任感和自主管理能力。结合我校对学生的培养目标,将学生的目标职业分为“创业”、“考升学”和“工作”。从学生的数据中也可以看出,70.4%的学生选择毕业后直接进入相关的工作,23.8%的同学希望继续深造,出国或在国内读研究生,仅有5.8%的同学会选择创业。

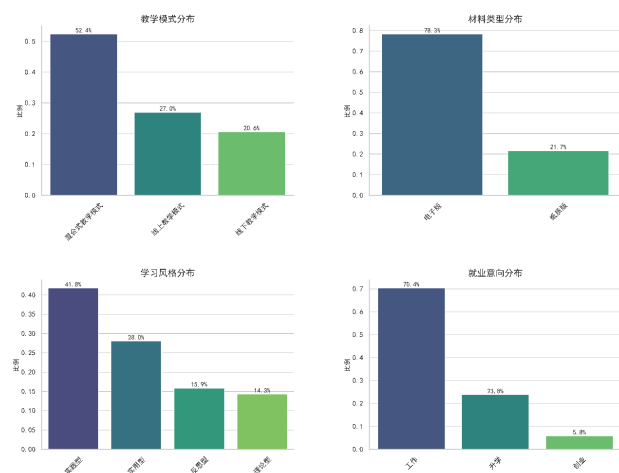


图2 学生基础情况分析

4. 学习投入

本研究基于 Fredricks 等人提出的学习投入三维理论框架,结合教育心理学与认知科学的核心观点,将学习投入解构为行为投入、情感投入与认知投入三大维度,旨在突破传统学业评价的单一性,系统揭示学生学习状态的深层特征。其中行为投入的相关数据通过对平台数据的采集和调查问卷获取,而情感、认知投入

主要通过问卷调查的形式获得数据^[10]。

在行为投入的研究中,选取课程的参与活动包括在“超星学习通”平台部署在线课程视频,共400.1分钟;部署任务点107次;签到16次;测验29次、实验10次等教学活动,在数据处理时,根据学生每项完成率赋值0-4,总最高分为20。对于情感和认知投入,数据来源主要是调查问卷。对于情感投入,主要调研学生对开设课程的兴趣,学习目标的设定,对课程的课程归属感(在学习过程中对本门课程的认识),是否在课程的学习中感受到认同,并在学习过程中大部分的情绪并在此情感状态下学习的主动性,同时还包含学生的社会关系感知,即学生与教师和同学之间在学习过程中的互动和关系,包括合作、交流、分享等。通过对学生的数据聚类分析,得到的聚类中心如表1。群体1是行为投入主导型,行为投入最高情感投入最低(17),这类学生可能更依赖“被动学习”,学习需要来自外部压力,在后续的课程中需要严格要求学习规范,更需要激发内在学习动机。群体2是全面投入型,所有维度均为最高或次高值,情感投入尤其突出,这类学生是学习主动性强的“优等生”群体,既有学习兴趣,又具备高阶思维能力,适合自主探究式学习环境,在课程学习中更需要提供高阶挑战。群体3是行为投入缺陷型,行为投入显著偏低,这类学生更偏向是“聪明但懒散”的类型,具备学习能力(认知)和兴趣(情感),但缺乏行动力,在后续的课程中要强化学习习惯培养,加强学习监督。

表1 聚类中心分布

	群体1	群体2	群体3
情感投入	17.0	18.1	19.2
认知投入	21.0	20.5	19.0
行为投入	19.1	18.9	12.2

5. 学习结果

在学习成绩分布于学习投入的分析中,群体1的成绩呈现较为,平均分82.26分且标准差最小(7.11),成绩集中于76-88分的中间区间。通过规律的学习行为建立的系统性知识框架,使得该群体能够稳定维持中等偏上水平。群体2展现出“双峰分布”的独特成绩特征,其平均成绩最高(83.0分)但标准差最大(8.02),成绩跨度从50分到96分形成两极分化。这种分布形态揭示出高投入群体的潜在分化,多数学生通过多维度的深度投入获得优异成绩(前25%达88分),但存在少数投入与产出失衡的异常个案(最低50分)。群体3成绩呈现“左偏分布”特征,平均分最低(79.24分)但存在个别高分段(93分),中位数(80分)显著低于全面投入群体,从此类别的学生也可以看出,学习投入是学生获得良好成绩的基础,但更高的成绩需要对课程的兴趣与持续不断的探索。

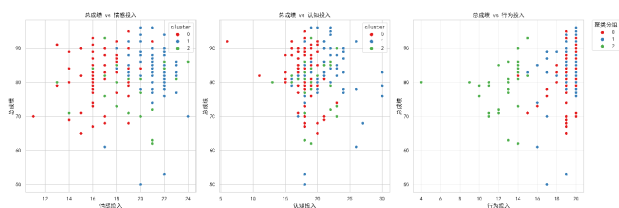


图3 不同聚类中心的学生成绩分布

好显著，实践型学习风格占主导，三类学生群体在行为投入与成绩稳定性方面呈现显著差异。研究成果为精准识别学生特征、制定分层教学策略提供了数据支撑，未来可拓展多源数据融合与动态画像更新机制，进一步提升教育供给的适切性。

四、总结

本研究通过 K-means 聚类算法构建了学生画像，揭示了数据下真实的学生状态。通过分析在本专业的学生中，混合式教学偏

参考文献

- [1] 习近平主持召开中央全面深化改革领导小组第二十二次会议强调：推动改革举措精准对焦协同发力 形成落实新发展理念的体制机制 [N]. 人民日报，2016-03-23
- [2] 柯璟. 基于学生毕业要求达成度评价的用户画像设计与分析 [J]. 教育教学论坛, 2020(45):13-16.
- [3] 冯春燕. 基于学习者画像的个性化教学研究 [D]. 浙江：浙江师范大学, 2019.
- [4] 张娜, 刘晓丹. 学生画像技术在教学中的研究实践与应用 [J]. 计算机时代, 2021(09):70-73.
- [5] 赵国亮, 陈晓军, 李思奇等. 基于数据分析高校学生自画像的初探 [J]. 数字技术与应用, 2017(08):233-234+236.
- [6] 丁凯, 宋林泽. 论高校思想政治教育精准化的机理及实现路径 [J]. 思想理论教育, 2020(06):101-105.
- [7] 赵君, 王丰昌. 基于精准供给的大学生思想政治教育路径论析 [J]. 学校党建与思想教育, 2021,(13):56-59.DOI:10.19865/j.cnki.xxdj.2021.13.013.
- [8] 黄炜, 张治, 胡爱花, 等. 基于“五育融合”的学生数字画像构建与实践分析 [J]. 教育发展研究, 2021,41(18):44-51.
- [9] 余明华, 张治, 祝智庭. 基于可视化学习分析的研究性学习学生画像构建研究 [J]. 中国电化教育, 2020,(12):36-43
- [10] Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59-109.