在线学习课程中的学生行为与分类

吕宇 林刚(通讯作者)

珠海科技学院，广东珠海，519040；

摘要：互联网产业革命不仅带来了信息的多元化，更催生了全新的数字化学习领域。教育数据挖掘与教育大数据已逐渐成为当今教育界的研究热点，在探索各类教育相关问题中发挥着极为积极的作用。近年来，受疫情影响，在线教育迅猛发展，社会教育模式逐步向多元化转型。但众所周知，在线学习仍难以与传统教学相媲美。在线学习存在的逻辑边界模糊、学习动力不足等问题，易导致学生流失。本文将采用 Kaggle 平台收集的 MOOC 平台学生数据,运用深度学习相关技术对学生表现进行分析,旨在探索有效解决在线学习学生流失问题的途径。

关键词：在线课程；学生行为；数据分析

DOI：10.64216/3080-1516.25.10.074

引言

疫情时期，在线学习平台已逐渐成为不可或缺的教育形式。在线学习与传统学校课堂面授教学的核心区别在于，其通过计算机网络开展课程学习，即我们常说的在线教育（Amrieh、Hamtini、Aljarah，2015）。在这一特殊时期，在线教学能够保障学生 “停课不停学”，让学生充分利用居家时间完成学业。此外，在线教育的时间与方式具有高度灵活性，教师录制的教学视频可随时观看，家长能根据学生的时间灵活安排学习进度。学生若听课之后存在疑问，还可随时回看，通过反复学习更好地巩固所学知识（Shahiri、Husain、Rashid，2015）。尽管在线教育具备诸多优势，但也不可避免地带来了一系列问题。疫情期间，多所学校的教师反映学生成绩出现下滑。虽然在线学习使学生在疫情时期实现了“停课不停学”，但在线教学对学生的独立思考能力和自律性提出了一定要求。由于缺乏实体课堂环境，教师无法直接对学生进行约束，若学生缺乏自律能力，其学习效果便会大打折扣（Amrieh、Hamtini、Aljarah，2016）。此外，教师无法面对面纠正学生的错误，在线教学的互动性相较于传统教学稍显不足。在线教学中的沟通与交流有所减少，多数情况下以教师单向输出为主，这会在一定程度上降低课堂教学质量。因此，本文主要探讨以下四个问题：

问题 1：在线学习中导致学生学习效果下降的特征有哪些？

问题 2：如何根据学生具体情况预测其辍学及成绩偏低的原因？

问题 3：家长参与度的分类与学生分类之间存在何种关联？

问题 4：课堂举手发言行为与学生分类之间存在何种关联？

1 文献综述

教育数据挖掘与教育数据分析为众多研究者开展教育相关问题研究提供了有力支持。教育的两个核心对象是学生与教师，近年来关于学生行为的研究也不断涌现。巴勒斯坦学者 Palestine 于 2012 年发表《挖掘教育数据以提升学生成绩：一项案例研究》一文，收集了该校科技学院 1993-2007 年的研究生数据，通过教育数据挖掘手段，有效提升了研究生的学业成绩，解决了研究生学业表现不佳的问题。2015 年，印度尼西亚学者 Harwati、Alfiani 与 Wulandari 采用 K - 均值聚类算法对学生进行分类映射，通过改善学生成绩来优化高校管理工作（Harwati、Alfiani、Wulandari，2015）。

日本学者 Okubo 等人（Okubo、Yamashita、Shimada、Ogata，2017）提出了一种基于递归神经网络的方法来预测学生的最终成绩，该方法通过教育系统中存储的日志数据进行预测。实验结果表明，与多元回归分析相比，递归神经网络在最终成绩的早期预测中具有显著效果。

西班牙学者 Monllaó Olivé 等人（Monllaó Olivé、Huynh、Reynolds、Dougiamas、Wiese，2019）通过数据挖掘与学习分析技术，对学习管理系统（LMS）Moodle 中存在课程弃学风险的学生以及在评估前被确定存在学习困难的学生进行了分析。同时，他们还提供了一个案例研究模型，该模型利用基于评估的变量，对 201

3-2018 年八门大规模开放在线课程（MOOC）的辍学学生进行预测。该框架能够在课程进行过程中识别出有风险的学生，从而提前采取相应的干预措施。

2 方法

2.1 数据描述

本文所使用的数据集来源于 Kaggle 平台，该数据集出自名为 Kalboard 360 的学习管理系统（LMS）。这些数据通过 API 学习者活动跟踪工具收集，能够监测学习者的学习进度与行为，例如课程观看进度、课后作业完成情况等。

该数据集共记录了 480 名学生的信息，包含 16 个特征变量。通过对这些特征的分析，可将其分为三类：（1）学生个人特征，如性别、国籍等；（2）学生背景特征，如教育阶段、年级、专业等；（3）行为特征，如课堂举手回答问题、浏览开放资源、参与相关问卷调查等。

此外，该数据集涵盖了两个学期的数据：第一学期 245 名学生，第二学期 235 名学生。数据集中还包含学生的出勤情况、家长参与学生学习过程的程度以及家长对学校的满意度等信息。

2.2 数据结构

本文研究数据的结构通过 16 个属性维度来呈现，各属性及其对应描述如下：“性别” 属性记录学生的性别信息；“国籍” 与 “出生地” 属性分别反映学生的国籍归属及出生地点；“教育阶段” 和 “年级” 属性明确学生所处的教育层次与具体年级；“班级编号” 属性标识学生所在的具体班级；“课程主题” 属性说明学生所学课程的核心内容；“学期” 属性标注数据对应的学年学期；“负责学生的家长” 属性属于分类变量，取值为 “母亲” 或 “父亲”，用于记录主要负责学生学习的家长身份；“举手次数”“访问资源次数”“查看公告次数”“参与讨论组次数” 均为数值型变量，取值范围为 0-100，分别代表学生在课堂上举手的次数、访问课程内容资源的次数、查看新公告的次数，以及参与讨论组活动的次数；“家长参与调查情况” 属性记录家长是否参与学校提供的相关调查；“家长对学校的满意度” 属性反映家长对学校教学与管理工作的满意程度；“学生缺勤天数” 属性为分类变量，取值为 “7 天以上” 或 “7 天及以下”，用于统计每位学生的缺勤天数范围。

2.3 数据分析

在数据分析过程中，首先通过 K - 均值（K-means）聚类算法对数据集进行构建并开展描述性分析，重点关注学生的行为特征，为后续更有效的决策过程提供支持（Gunuc、Kuzu，2014）。随后，运用监督学习构建具有最优特征的数据集，对特定条件下学生的行为进行预测，开展预测性分析。预测过程中，采用以下四种算法并对其准确率进行对比：K 近邻算法（KNN）、决策树算法、支持向量机算法、逻辑回归算法。

本研究选取四种经典机器学习算法用于在线学习学生行为预测，各算法适配教育数据挖掘的核心特性。K - 均值算法作为基于欧氏距离的主流聚类方法，以 “距离越近相似度越高” 为逻辑核心，原理简洁、聚类效果可靠，且处理大规模数据时兼具高可扩展性与低计算复杂度，成为聚类分析优选。K 近邻算法依 “近邻投票” 原则，匹配训练集与测试数据最相似的前 k 个样本，以多数类为预测结果，是直观高效的基础分类方法。决策树以树状结构建模，可读性强、分类速度快，便于解读特征与结果关联。支持向量机通过构建高维超平面作为分类边界，优质边界需最大化与近邻训练样本的距离，降低泛化误差，适配复杂数据关联。逻辑回归速率函数全阶可微，数学性质优良，求解高效且特征权重可解释性强，能量化各因素影响。通过对比四种算法预测准确率，为精准捕捉学生行为规律提供多元技术支撑。

3 结果与讨论

本研究采用 “描述性分析 - 预测性分析” 递进式路径，系统挖掘在线学习场景下学生行为规律，确保研究的系统性与精准性。

描述性分析阶段，基于 K-means 聚类算法开展研究，通过数据归一化与标准化预处理消除量纲差异，以欧氏距离度量样本相似度。该算法在处理 480 名学生多维度数据时，展现出低复杂度、高可扩展性优势，最终依据学习活跃度（综合举手次数、资源访问频率等行为特征）将学生划分为三类，为后续预测模型构建奠定了清晰的群体分类基础（Gunuc & Kuzu, 2014）。

预测性分析环节，选取 K 近邻（KNN）、决策树、支持向量机（SVM）、逻辑回归四种经典机器学习算法，

以雅卡尔指数为评价指标对比其对学生辍学风险、成绩等级的预测效果。结果显示，KNN 算法在 k=4 时表现最优，雅卡尔指数达 0.65，其优势在于无需预设参数，可通过动态匹配相似样本特征捕捉行为与分类的关联性；决策树算法虽可读性强、分类快，筛选出 “资源访问次数”“查看公告频率” 等关键变量，但受样本量限制存在过拟合，准确率较 KNN 低 8.2%；支持向量机在处理 “家长参与度 - 学生缺勤率” 等非线性数据时表现稳定，但训练耗时较长；逻辑回归算法明确了 “家长参与调查情况”（权重系数 0.32）与 “学生缺勤天数”（权重系数 - 0.28）的显著影响，但其线性假设难以适配复杂行为交互，准确率仅 0.51。

特征贡献度拆解发现，“访问课程资源次数”“查看公告频率” 为核心预测变量（平均贡献度 0.41），表明主动获取学习信息与高分类评级强相关；“参与讨论组次数” 贡献度普遍低于 0.12，说明当前在线讨论未有效转化为学习效果提升动力；“出生地” 贡献度不足 0.05，对预测无显著价值，而 “学生实际所在地” 隐含的互联网质量、区域学习环境等信息，或为更关键的潜在变量，为模型优化提供了方向。

讨论部分，本研究基于 Kalboard 360 系统数据构建的最优模型（k=4 的 KNN 算法），虽为在线教育优化提供了实证支撑，但存在场景适配边界。训练数据源于特定平台，跨 MOOC、SPOC 等平台迁移时，因用户画像、教学模式差异可能导致精度衰减；且模型未纳入互联网基础设施、家庭学习环境等客观变量，可能干扰预测结果，如欠发达地区学生缺勤或为客观条件限制，仅依据 “缺勤天数” 判断易产生偏差。

研究结果具有重要实践启示：在线教育平台应强化资源结构化呈现与公告精准推送，优化讨论组互动设计（如问题驱动式讨论），构建家校协同体系。未来研究可补充环境、心理等特征变量，采用纵向追踪数据；融合集成学习、深度学习等算法，结合定性研究方法；开展跨平台、跨学段对比研究，聚焦特定群体构建精准干预模型，助力在线教育高质量发展。

参考文献

[1]埃萨姆·A·阿米里赫,塔里克·哈姆蒂尼,易卜拉欣·阿拉杰拉.基于集成方法的教育数据挖掘在学生学业成绩预测中的应用[J].数据库理论与应用国际期刊,2016,9(8):119-136.

[2]埃萨姆·A·阿米里赫,塔里克·哈姆蒂尼,易卜拉欣·阿拉杰拉.基于X-API的教育数据集预处理与分析:提升学生学习表现[C]//约旦应用电气工程与计算技术会议论文集。约旦，2015.

[3]萨利赫·古努奇,阿赫迈特·库祖.学生参与度量表的开发、信度与效度检验[J].高等教育评估与评价,2014,40(4):587-610.

[4]哈瓦蒂,阿尔法尼·A·P,芙丽达·A·乌兰达里.基于数据挖掘方法的学生成绩映射研究(案例分析)[J].农业与农业科学进展,2015,3:173-177.

[5]丹尼尔·蒙拉奥·奥利韦,丁·Q·黄,马修·雷诺兹,马丁·杜吉亚马斯,丹尼尔·威斯.一种监督学习框架：利用评估识别大规模开放在线课程（MOOC）辍学风险学生[J].高等教育计算机应用期刊,2019,32(1):9-26.

[6]藤井大久保,山下贵史,岛田明,绪方浩.基于神经网络的学生成绩预测方法[C]//第七届国际学习分析与知识会议论文集。美国纽约，2017.

[7]乔治·巴勒斯坦.挖掘教育数据以提升学生成绩：一项案例研究[J].信息与通信技术研究国际期刊,2012,2(8).

[8]阿兹玛·M·沙希里,旺·侯赛因,诺拉·A·A·拉希德.基于数据挖掘技术的学生成绩预测研究综述[J].计算机科学进展,2015,72:414-422.

项目信息：1、广东省本科高校教学质量与教学改革工程项目：计算机类《课程设计》类课程的过程化考核研究与实践，编号：2024008。

2、珠海科技学院 AI 赋能产教融合型课程培育建设项目计算机导论，编号：CJRH2025006。