

面向学情诊断的智能学辅模式研究

孙丽娟, 朱亚玲*

兰州工业学院, 甘肃兰州, 中国

*通讯作者

【摘要】在教育数字化背景下, 学习行为数据为精准学业评价和个性化辅导提供了新的支撑。针对传统教学中学情判断不够精细、学习支持缺乏差异化等问题, 本文以 14003 条学生学习行为数据和 16 个特征变量为基础, 经过数据清洗、缺失值处理、标准化转换与特征筛选, 采用决策树和随机森林构建成绩等级分类模型, 并结合 K-means 聚类、相似个体匹配大模型技术, 构建面向不同学习群体的个性化学辅方案设计框架。结果显示, 随机森林模型的平均准确率达到 88.09%, 能够较好完成学生成绩等级判定; 聚类结果可识别出主动型、平衡型等不同学习群体, 为差异化辅导提供依据。研究构建了数据分析支持下的智能学辅框架, 可为教师开展学情诊断、教学干预和辅导方案设计提供参考, 也为普通教学场景中的智能化学辅研究提供了一种可行思路。

【关键词】数据分析; 分类预测; 聚类分析; 个性化学习

【基金项目】2024 兰州工业学院新工科研究与实践项目(编号: 2024-LGYXGK-07); 2025 兰州工业学院教学改革研究项目(编号: LGYJG2025-34)

1.引言

随着教育数字化转型的持续推进, 智慧校园、学习管理平台、在线课程系统和课堂互动软件等信息化教学工具已逐渐融入日常教学。学生在课前预习、课堂互动、课后复习和在线测评等环节中持续产生学习行为数据, 内容涉及学习路径、学习投入、互动频率和阶段性成绩等方面。这些数据为教师把握学生学习状态、改进教学设计和开展个性化辅导提供了新的依据。相比单纯依靠考试成绩和课堂经验进行学情判断, 基于数据的学习分析能够更细致地呈现学生学习过程, 有助于提高学业评价的客观性和针对性。

人工智能技术的发展进一步拓展了学习辅导的实现路径。生成式人工智能(AIGC)和大语言模型在文本生成、学习反馈、资源推荐和智能问答等方面表现出较强适应性, 为个性化学习支持提供了新的技术条件。对于教师而言, 数据分析有助于识别班级整体学习动态和学生个体差异; 对于学生而言, 智能化辅导能够根据其学习状态提供更具针对性的学习建议。因此, 将教育数据分析、机器学习方法与生成式人工智能结合起来, 探索智能学辅模式的构建路径, 具有一定的现实意义。

近年来, 国内外学者围绕教育大数据、个性化学习和人工智能辅助教学开展了较多研究。国内研究认为, 数据驱动的个性化学

习有助于突破传统教学中“一刀切”“同质化”的局限, 使教学更加关注学生学习过程和自主发展[1]; 教育大数据在数据采集、特征提取、学生画像、学习预测和智能干预等方面已形成较为清晰的研究路径[2]; 生成式人工智能与教育大数据的融合, 为个性化学习服务提供了新的技术支撑[3-5]; 在教学应用层面, 已有研究通过教学实验和高校智能辅导系统设计, 验证了 AI 技术在提升学习效率、改善学习体验以及开展分层辅导方面的积极作用[6,7]; 同时, 生成式人工智能也被拓展到沉浸式学习、信息素养教育和思政课程等教学场景中, 为学习情境创设、个性化资源供给和课堂互动方式改进提供了新的思路[8-10]。国外研究则主要集中在智能学习助手、教学型 AI 对话代理和人机交互支持学习等方面, 相关成果为智能学辅系统的功能设计和教学应用提供了借鉴[11-13]。总体来看, 已有研究为智能学辅模式的构建奠定了基础, 但仍存在一定不足: 一是部分研究偏重成绩预测、学习画像或资源推荐等单一环节, 对学情诊断、学生分群和辅导建议之间的衔接关注不够; 二是学生差异分析仍较多依赖成绩、学习时长等显性指标, 对学习行为特征和课堂互动数据的综合利用不足; 三是生成式人工智能若缺少数据分析支撑, 容易生成较为笼统的辅导建议, 难以有效服务于教学决策。

基于此,本文以学生学习行为数据为基础,综合运用决策树、随机森林、K-means聚类、相似个体匹配和AIGC技术,构建数据分析支持下的智能学辅模型。该模型包括学习数据处理、学业成绩预测、学习群体识别和个性化辅导建议设计等环节,旨在提高学情判断的准确性,并为教师开展差异化辅导和教学改进提供参考。

2.相关技术基础与设计框架

2.1 学习行为分析相关理论

学习行为分析是教育数据挖掘的核心研究领域,旨在从学习过程中产生的数据中提取有价值的信息。学习行为数据来源于学习管理系统记录的登录日志、课程访问轨迹、作业提交记录等多元渠道。学习行为分析的理论基础植根于建构主义学习理论和自主学习理论,建构主义强调学习者在知识建构过程中的主动参与,自主学习理论关注学习者的自我调节能力。

2.2 学生画像相关理论

学生画像是对学习者多维特征的综合性描述,通过整合学习行为数据、人口统计信息、认知能力评估等多源数据构建学生的数字化表征。学生画像的构建过程包含数据采集、特征提取、标签生成、画像更新四个阶段。学习基础维度刻画学生的知识储备、能力水平、学习动机等初始状态,学习过程维度记录学生的学习投入、参与行为、策略调整等过程特征,学习结果维度反映学生的成绩表现、能力提升、目标达成等产出指标。

2.3 数据分析方法

数据分析大体可以分为四个层次,即描述性分析、探索性分析、预测性分析和规范性分析。其中描述性分析关注数据的基本特征,通过统计指标和可视化图表呈现数据的分布形态。诊断性分析探究数据背后的原因,通过相关性分析、差异检验等方法揭示变量之间的关联机制。预测性分析基于历史数据推断未来状态,运用回归分析、分类模型等方法生成预测结果。规范性分析提供决策建议,结合预测结果和约束条件生成最优行动方案。在教育数据分析中,Pearson相关系数度量两个连续变量之间的线性关联程度,其计算公式为:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

其中, x_i 和 y_i 分别为两个变量的观测值, $\bar{x}$ 和 $\bar{y}$ 为样本均值, n 为样本数量。差异性检验用于判断不同群体之间是否存在显著差异,方差分析的检验统计量为:

$$F = \frac{MSA}{MSE} = \frac{\sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 / (k-1)}{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 / (N-k)} \quad (2)$$

其中, MSA 为组间均方, MSE 为组内均方, k 为组数, N 为总样本数。

2.4 机器学习方法概述

机器学习方法主要可以分为监督学习、无监督学习和半监督学习三类,在教育领域中各类技术的应用特点大相径庭。分类算法是监督学习的重要组成,它用预测模型把目标变量准确地分层、识别出来。决策树是种典型的分类器,它的构造依据是递归分割的原则,并且用信息增益或者基尼系数这些量化指标来评价特征的选择情况。信息增益体现的是属性选择之后数据集熵值降低的程度,计算公式为:

$$IG(D, A) = H(D) - \sum_{v=1}^V \frac{|D_v|}{|D|} H(D_v) \quad (3)$$

其中, $H(D)$ 为数据集 D 的熵, D_v 为特征 A 取值为 v 的子集。基尼系数度量数据集的纯度,计算公式为:

$$Gini(D) = 1 - \sum_{k=1}^K p_k^2 \quad (4)$$

其中, p_k 为类别 k 在数据集 D 中的比例。随机森林通过集成多棵决策树提升分类性能,每棵树在随机抽样的训练集和随机选取的特征子集上独立生长,最终预测通过投票方式聚合,能够有效抑制单一决策树的过拟合倾向。聚类方法将相似样本归入同一簇, K-Means 算法通过迭代优化簇中心和样本分配实现聚类目标,目标函数为:

$$J = \sum_{j=1}^K \sum_{i \in C_j} \|x_i - u_j\|^2 \quad (5)$$

其中, K 为簇数, C_j 为第 j 个簇的样本集合, u_j 为第 j 个簇的中心。轮廓系数度量聚类效果,计算公式为:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

其中, $a(i)$ 为样本 i 与同簇其他样本的平均距离, $b(i)$ 为样本 i 与最近异簇样本的平均距离。

2.5 设计框架

本文的设计框架如图1所示,分为五个

阶段。首先，整理有关文献资料，搭建起理论分析的框架结构；其次，进行数据收集、预处理及特征提取等工作，并使用统计学的方法对原始的数据做初步的探索性分析；再次，根据学习行为模式与学业表现之间的相关性假设，利用决策树、随机森林等机器学习模型来建立预测模型，并用 K-means 聚类算法将目标群体进一步划分出来；最后，根据分类结果以及个体属性的匹配规则制定出个性化的干预措施，引入人工智能辅助模块提高建议生成过程的效率与准确性。

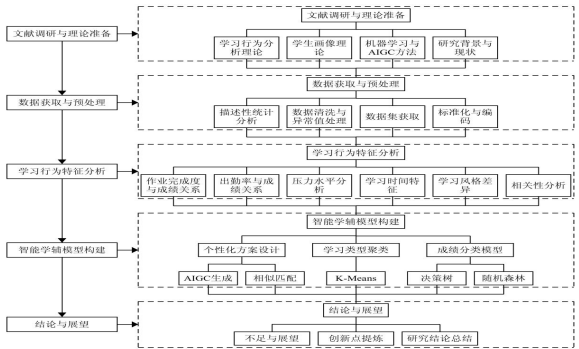


图 1. 框架图

3. 数据集与预处理

3.1 数据集结构

本文使用的数据集来自 kaggle 采集的真实学习行为数据，共包含 14, 003 条有效学生记录与 16 个特征变量，全面覆盖学习投入、课堂参与、任务完成、学习偏好、心理状态等关键维度，所有数据均由系统自动记录、统一采集并标准化存储，保证了数据的真实性、规范性与可用性。依据数据类型，特征变量可划分为连续型与离散型两类。其中，连续型变量包括学习时长、出勤率、作业完成度、考试成绩、年龄、在线课程数量等，用于量化学生学习行为的客观表现；离散型变量包括性别、学习风格、学习动机、压力水平、最终成绩等级等，用于刻画学生属性与分类标签。本文以最终成绩等级为模型预测目标变量，取值 0-3 分别对应优秀、良好、中等、需改善四个等级，样本数量依次为 3832 条、3310 条、3618 条、3243 条，整体分布均衡，有效避免类别不平衡问题，为后续分类模型的训练、测试与评估提供了高质量、高可靠性的数据基础。

3.2 数据预处理

数据清洗是保障模型训练质量的关键环节，本文主要从编码转换、重复记录去除、缺失值检测三个方面开展规范化处理。原始数据采用 GBK 编码存储，直接导入易出现中

文字符乱码、字段解析异常等问题，因此在数据读取阶段通过指定正确编码参数，有效解决跨平台兼容问题，保证数据完整可读。重复记录检测采用完全字段匹配方法，对学习时长、出勤率、学号等全部字段进行一致性比对，精准识别并剔除完全重复的样本，最终保留 14003 条有效记录，样本规模充足，可满足后续机器学习建模与统计分析的大样本要求。缺失值检测结果显示，16 个特征变量均无缺失数据，整体完整性良好，无需进行插值或填充处理。数据清洗前后样本数量对比如图 2 所示。

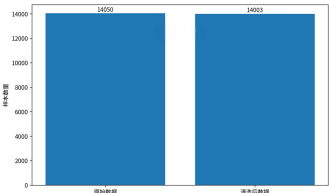


图 2. 数据清洗前后样本数量对比

异常值检测采用四分位距（IQR）方法，判定规则为取值低于 $Q1-1.5\times IQR$ 或高于 $Q3+1.5\times IQR$ 的数据点。学习时长字段存在 45 个异常值，占总样本 0.36%，学习时长超过 36 小时或低于 4 小时。出勤率、作业完成度、考试成绩等字段均未检测到异常值。处理结果的对比如图 3 所示，数据集中不存在缺失值；学习时长字段共识别出 45 个异常样本，主要表现为明显偏低或偏高的录入值。经过边界截断处理后，极端点对整体分布的影响明显减弱。

本文采用 Z-score 标准化方法将连续型变量转换为均值为 0、标准差为 1 的标准正态分布。标准化处理应用于学习时长、出勤率、作业完成度、考试成绩、在线课程数量等连续型变量。Z-score 标准化能够消除不同变量之间量纲和取值范围的差异，使各变量在统一的尺度上参与距离计算和模型训练。离散型变量保持原始数值编码不变，有序等级变量和无序类别变量直接采用原始编码参与分析。

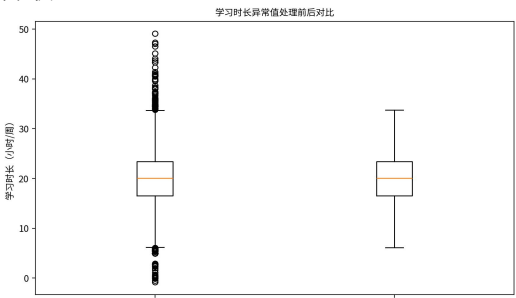


图 3. 学习时长异常值处理前后对比

标准化处理结果对比如图 4 所示, 结果显示, 学习时长 (5-44 小时) 与考试成绩 (40-100 分) 取值范围差异巨大, 直接用于距离计算 (如 K-Means) 会使模型偏向于取

值大的变量。处理后, 所有连续变量的均值均为 0, 标准差均为 1, 处于同一量纲, 公平参与模型计算。

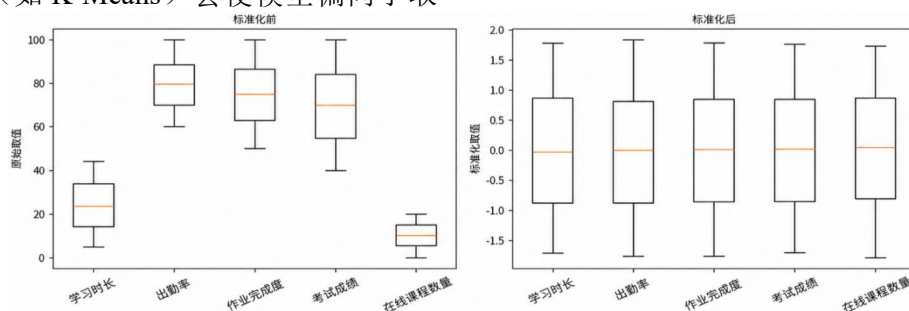


图 4. 标准化处理前后对比

4. 学习行为特征分析

4.1 学习时间特征分析

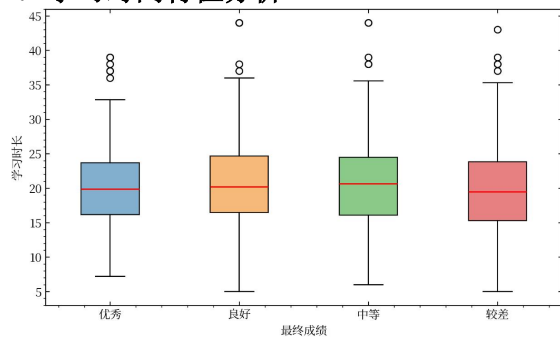


图 5. 各成绩等级学习时长分布箱线图

为分析不同学业表现学生的学习时间投入是否存在显著差异, 本文采用分组统计与方差分析(ANOVA)的方法。根据统计, 数据集中学习时长变量取值范围为 5 至 44 小时, 均值为 20.03 小时。对各成绩等级学生的学习时长进行分组统计方差分析, 结果如图 5 所示, 检验结果 F 统计量为 0.3784, p 值为 0.7686, 未达显著性水平, 表明不同成绩等级学生的学习时长不存在显著差异。这一结果揭示了学习时长与成绩等级之间的非线性关系。陶红与王磊的研究也得出了类似结论, 即单纯的学习时间投入并非学业表现的唯一决定因素, 学习效率和策略更为关键。

4.2 学习风格差异分析

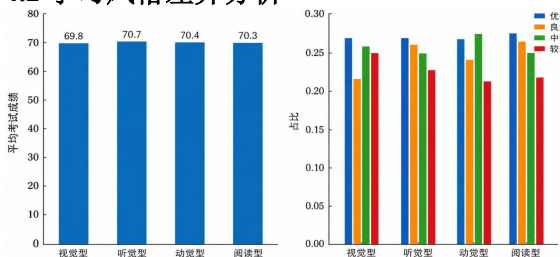


图 6. 学习风格与成绩关系分析

为检验学习风格与成绩等级之间是否存在关联, 本文采用卡方检验进行分析。数据

集中学习风格变量编码为四种类型卡方检验, 学习风格与层级关系分析如图 6 所示, 结果显示 χ^2 统计量为 32.75, p 值为 0.0001, 达到显著性水平, 表明学习风格与成绩等级之间存在统计关联, 但考试成绩均值差异微弱, 暗示学习风格可能通过影响学习路径而非直接决定学习结果。

4.3 各变量相关性分析

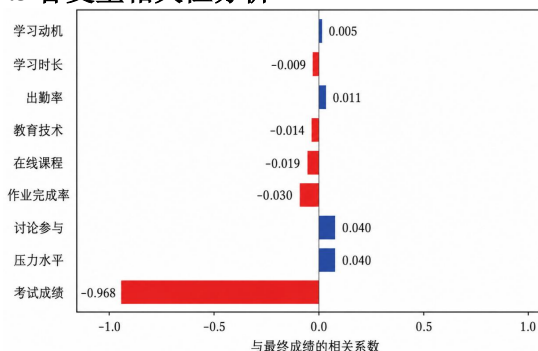


图 7. 各变量与成绩等级相关性对比

对全部特征变量与最终成绩等级进行相关性检验, 系统评估各学习行为特征对学业表现的预测效力。计算各变量与 FinalGrade 的 Pearson 相关系数并进行显著性检验, 识别与成绩等级存在统计关联的特征变量。相关性分析结果如图 7 所示, 结果显示, 考试成绩 ExamScore 与成绩等级的相关系数为 -0.97, p 值小于 0.0001, 呈现强负相关。作业完成度 AssignmentCompletion 与成绩等级的相关系数为 -0.03, p 值为 0.0007, 呈现弱负相关。讨论参与 Discussions 与成绩等级的相关系数为 0.04, p 值小于 0.0001, 呈现弱正相关。压力水平 StressLevel 与成绩等级的相关系数为 0.04, p 值小于 0.0001, 呈现弱正相关。在线课程数量 OnlineCourses 与成绩等级的相关系数为 -0.02, p 值为 0.03, 呈现弱负相关。学习时长 StudyHours、出勤率

Attendance、学习动机 Motivation、教育技术使用 EduTech 等变量与成绩等级的相关系数均未达到显著性水平。

5.智能学辅模型的设计

5.1 学生分类模型设计与群体差异分析

学生成绩等级分类是智能学辅系统的核心功能，通过对学生学习行为特征的分析预测其学业表现等级，为个性化干预提供决策依据。本文将 FinalGrade 作为分类目标，包含优秀（0）、良好（1）、中等（2）、需改进（3）四个类别。根据第 4 部分的相关性分析结果，选取考试成绩 ExamScore、作业完成度 AssignmentCompletion、讨论参与 Discussions、压力水平 StressLevel、在线课程数量 OnlineCourses 等 14 个特征变量作为分类模型的输入，数据集按 8:2 的比例划分为训练集和测试集。

本文采用随机森林算法构建分类模型，随机森林通过集成多棵决策树提升分类性能，每棵树在随机抽样的训练子集和随机选取的特征子集上独立生长，最终预测通过多数投票方式聚合。本文构建包含 100 棵决策树的随机森林模型，最大深度设置为 5 层，特征子集大小设置为特征总数的平方根，模型结构如图 8 所示。

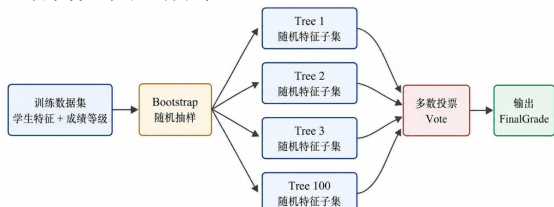


图 8.基于随机森林的分类模型结构图

为检验随机森林算法的分类效果，本文构建了基于决策树的分类模型，采用 Acc、Precision、Recall 和 F1-score 四个模型性能分类评价指标，对基于随机森林和决策树的分类模型性能进行了对比分析。其中，Acc 表示模型整体分类正确的比例，Precision 表示被模型预测为某一类别的样本中实际属于该类别的比例，Recall 表示某一真实类别样本被模型正确识别出的比例，F1-score 是 Precision 和 Recall 的调和平均值，能够综合反映模型的分类性能。分析结果如表 1 所示。

表 1.分类模型性能对比

模型	Acc	F1	Precision	Recall
决策树	52.04%	52.11%	52.69%	52.04%
随机森林	88.09%	88.05%	88.12%	88.09%

表 1 结果显示，随机森林模型在全部指标上均显著优于决策树模型。随机森林的 Acc 为 88.09%，决策树为 52.04%，差距达 36 个百分点；交叉验证结果也进一步验证了随机森林模型的稳定性优势。

本文选取学习时长、出勤率、作业完成度、学习动机、压力水平五个核心特征对不同成绩等级学生的学习行为特征进行统计分析，分析结果如图 9 所示。结果显示，优秀类别学生的作业完成度均值高出较差类别 1.4 个百分点，压力水平均值低于较差类别 0.09，学习时长、出勤率等维度的类别差异微弱。

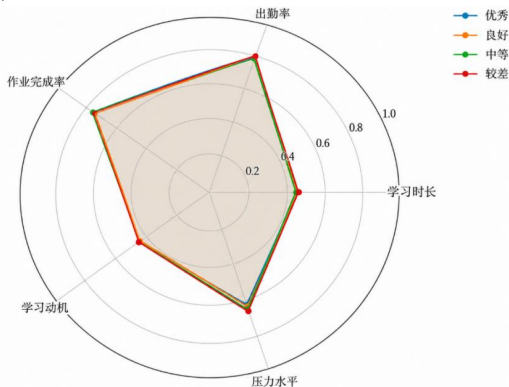


图 9.不同类别学生行为特征雷达图

5.2 学生学习类型聚类分析

学生学习类型聚类旨在根据学习行为特征将学生划分为若干类型群体，识别不同类型学生的学习模式差异，为差异化教学干预提供群体划分依据。本文选取学习时长 StudyHours、出勤率 Attendance、作业完成度 AssignmentCompletion、在线课程数量 OnlineCourses、讨论参与 Discussions、学习动机 Motivation、学习风格 LearningStyle 七个特征作为聚类变量，涵盖学习时间投入、课堂参与、课后执行、资源利用、互动行为、心理驱动、认知偏好等多个维度，能够综合刻画学生的学习行为画像。

本文选择 K-Means 算法构建学生学习聚类模型，K-Means 算法通过迭代优化簇中心和样本分配实现聚类目标，算法将 n 个样本划分为 k 个簇，使簇内样本相似度高而簇间样本相似度低，模型结构如图 10 所示。

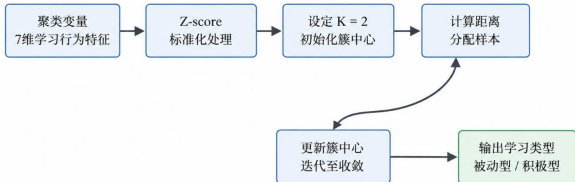


图 10.基于 K-Means 的聚类模型结构图

对 k 取值 2 至 10 进行聚类计算并评估聚类效果, 结果显示 $k=2$ 时轮廓系数为 0.1548, Calinski-Harabasz 指数为 2081.24, Davies-Bouldin 指数为 2.34。 $k=3$ 时轮廓系数下降为 0.1244, Calinski-Harabasz 指数下降为 1684.30。随着 k 增大轮廓系数整体下降, 综合三项评估指标 $k=2$ 为最优聚类数。采用 $k=2$ 的 K-Means 聚类模型对 14003 条学生记录进行聚类, 结果显示簇 0 包含 4, 910 条样本占比 39.38%, 簇 1 包含 7, 559 条样本占比 60.62%, 聚类模型的轮廓系数为 0.1548, 聚类质量达到中等水平。

K-Means 聚类结果如图 11 所示, 通过主成分分析将 7 维特征降至 2 维平面。图中蓝色点代表被动学习者, 红色点代表适度积极学习者, 两类群体在降维空间中形成相对集中的簇, 且边界较为清晰, 验证了聚类效果的有效性。

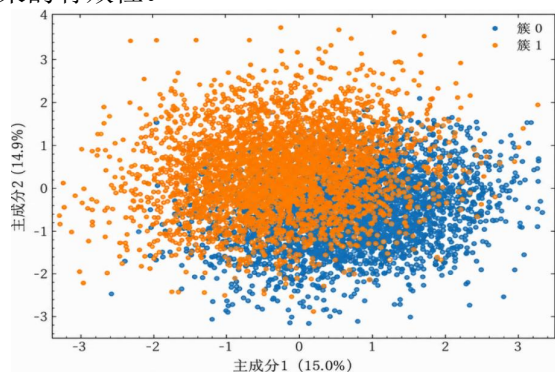


图 11. 聚类结果 PCA 可视化

表 2. 聚类簇特征对比

特征	簇 0 (被动学习者)	簇 1 (适度积极学习者)
样本数	4910 (39.38%)	7559 (60.62%)
学习时长	19.83 小时	20.17 小时
出勤率	80.34%	80.18%
作业完成度	74.49%	74.53%
讨论参与	0.00	1.00
平均考试成绩	71.00 分	69.87 分

对两个学习类型群体的成绩分布进行统计分析, 如表 2 所示, 被动学习者的平均考试成绩为 71.00 分, 适度积极学习者 69.87 分, 被动学习者的优秀等级占比也更高。这一反直觉的结果暗示课堂讨论参与与学业表现之间存在复杂关系, 本研究的聚类结果为差异化教学干预提供了群体划分依据。

5.3 基于大模型的个性化学习方案设计

个性化学习方案生成机制包含学生特征输入、相似学生匹配和建议内容生成三个核

心环节, 结构如图 12 所示。

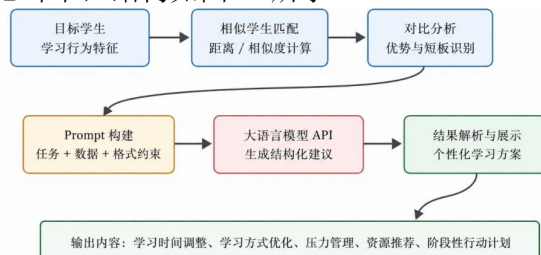


图 12. 基于大模型的个性化方案设计框图

学生特征输入环节接收目标学生的学习行为数据, 包括学习时长、出勤率、作业完成度、学习风格、压力水平等特征变量, 特征数据经过标准化处理后形成特征向量用于后续的相似度计算和群体匹配; 相似学生匹配环节在数据集中检索与目标学生特征相似的学生群体, 计算特征向量之间的欧氏距离或余弦相似度, 选取相似度最高的若干学生作为参照群体, 相似群体的学业表现和学习行为数据为目标学生提供了可借鉴的学习范例; 建议内容生成环节将目标学生的特征数据和相似群体的对比分析结果输入大语言模型, 通过提示工程引导模型生成结构化的学习建议, 建议内容涵盖学习时间调整、学习方式优化、压力管理策略、学习资源推荐等多个方面, 针对目标学生的特征短板和改进方向提供具体指导。

6. 结论

本文构建的智能学辅模型能够较好捕捉学生学习行为特征差异, 为教师开展学情分析、成绩预测和差异化教学干预提供数据支持, 也能够为学生提供更具有针对性的学习改进建议。后续研究将进一步扩展数据来源, 优化模型特征体系, 并结合教师反馈和学生学习效果持续改进个性化推荐机制, 持续提升智能学辅模型的实用价值和推广能力。

参考文献

- [1] 钟绍春, 杨澜, 范佳荣. 数据驱动的个性化学习: 实然问题、应然逻辑与实现路径[J]. 电化教育研究, 2025, 46 (1): 13-19.
- [2] 张雄, 李浩轩, 刘壮, 等. 教育大数据驱动的个性化学习研究综述[J]. 现代远程教育研究, 2024, 36 (2): 45-54.
- [3] 朱永海, 张佳鑫, 韩锡斌. 基于生成式人工智能的个性化学习新形态[J]. 电化教育研究, 2025, 46 (4): 58-64.
- [4] 熊欢. 基于大数据的 AI 智慧教育个性化

- 学习模型构建[J].创意与创新, 2025, 9 (8): 73-76.
- [5] 李敏, 张华.教育大数据与人工智能个性化学习[M].北京: 科学出版社, 2024.
- [6] 陶红.AI技术在个性化教育中的应用与效果研究[J].教育进展, 2024, 14 (8): 100-105.
- [7] 王磊.高校智能辅导系统个性化学习策略研究[D].华东师范大学, 2023.
- [8] 闫寒冰, 杨淑婷, 余淑珍, 陈怡.生成式人工智能赋能沉浸式学习: 机理、模式与应用[J].电化教育研究, 2025, 46 (2): 64-71.
- [9] 孙晓瑜, 孟文杰, 张雪松等.生成式人工智能赋能信息素养个性化教育模式研究[J].中国信息技术教育, 2025, (15): 100-104.
- [10] 周源露.基于 AI 的思政教育个性化教学探索[J].教育研究, 2025, 8 (11): 258-269.
- [11] Sajja R, Sermet Y, Cikmaz M, et al. Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education[J]. Information, 2024, 15(10): 596.
- [12] Looi C K, Jia F. Personalization capabilities of current technology chatbots in a learning environment: An analysis of student-tutor bot interactions[J]. Education and Information Technologies, 2025, 30: 14165-14195.
- [13] Yusuf H, Money A, Daylamani-Zad D. Pedagogical AI conversational agents in higher education: A conceptual framework and survey of the state of the art[J]. Educational Technology Research and Development, 2025, 73: 815-874.