基于先修课程成绩的选修课程推荐研究

涂 辉[1] 许腾飞[2] 丁中燕[3] 张正金[4,5]
TU Hui XU Tengfei DING Zhongyan ZHANG Zhengjin

摘 要 针对高校课程体系复杂化与学生选课效率低下的问题，文章提出了一种基于先修课程成绩的智能化课程资源推荐系统。通过整合学生历史成绩数据，构建学科能力画像，结合课程特征相似性计算与随机森林算法，动态推荐适配课程及学习资源。实验采用某职业技术学院 1 039 条真实数据，通过 SMOTE 过采样与 XGBoost 优化解决类别不平衡问题，在 60 分推荐阈值下实现 93.8% 的准确率与 0.93 的 F_1 分数。结果表明，化学基础（2）与化工制图为关键先修课程，模型能有效缓解数据稀疏性与推荐偏差，为教育智能化管理提供数据驱动的决策支持。

关键词 课程推荐；先修课程；随机森林；教育数据挖掘

doi：10.3969/j.issn.1672-9528.2025.09.014

0 引言

随着硬件算力的极大增强和机器学习、人工智能、深度学习技术的不断发展，教育数据挖掘正广泛应用于教育实践中[1]，高等教育中信息技术的支撑力度也在逐渐增强，其中一个热门的研究分支是预测学习者的学习结果，从而为学习者提供适合的课程建议，这有助于他们在整个学习过程中拥有最佳的学习路径[2-3]。

现代高校课程体系呈现出学科交叉性强、先修逻辑复杂的特点。学生在选课时需综合考虑课程难度、先修知识储备及个人兴趣，而传统选课模式依赖主观经验或导师建议，对学生来说存在三大问题：

（1）信息不对称：学生难以全面了解课程内容与自身能力的匹配度；

（2）路径规划低效：盲目或跟风选课易导致知识断层或重复学习；

（3）资源浪费：存在一部分学生因选课不当需要重修，学生的坚持率很低。

对学校来说，既往的数据沉淀不能有效指导当期学生的选课行为，数据资产未能得到盘活，另一方面，数据分析不到位，不能指导教师有效开展教育教学工作，教学迭代依赖经验而非客观数据。

研究表明，前置课程成绩与后续学习表现呈显著相关性[4]。因此，利用数据挖掘技术构建智能化推荐系统，成为优化教育资源配置的关键路径。本文提出一种基于先修课程成绩的选修课推荐系统，核心目标包括：突破传统协同过滤的数据稀疏性限制，设计轻量级有监督推荐策略，实现了预测学生对目标课程学习能力的评估。

1 相关工作

1.1 教育数据挖掘

教育数据挖掘（EDM）[5]通过分析学生行为日志、成绩数据等，已广泛应用于学习预警、认知状态评估等领域[6]。经典方法包括分类技术、聚类分析、时序建模等。分类技术是一种有监督的机器学习方法，通过对训练数据集中业务数据和结果数据建模，产生一个分类模型，在测试数据集上通过分类模型预测其所属类别，具体包括支持向量机、人工神经网络、决策树等方法，在教育中的应用包括成绩预测、资源推荐等。聚类数据挖掘一般是将学生、资源或教育活动定义到具有相应意义的集群中识别学习行为模式，以发现相似的主体从而分析和指导学生行为及学习风格，形成个性化的学习路径，减少决策的偏差[7]。时序建模通过分析发掘数据集中具有时间顺序的数据，来了解和分析学生的学习路径，刻画学习轨迹，预测学习结果、学业发展趋势等，以在学习过程中为用户提供新的见解[8]。

1. 徐州工业职业技术学院现代教育技术中心 江苏徐州 221140
2. 徐州幼儿师范高等专科学校康管学院 江苏徐州 221018
3. 徐州工业职业技术学院学工处 江苏徐州 221140
4. 巢湖学院计算机与人工智能学院 安徽合肥 23802
5. 澳门科技大学创新工程学院 中国澳门 999078

[基金项目] 2023 年度江苏省教育科学规划课题“本体视域下个性化学习资源推荐平台构建与研究：以药品生产技术为例”（C/2023/02/30）；2024 年江苏省哲学社会科学课题“人工智能背景下高校辅导员工作高质量发展研究”（2024SJSZ0511）

其中教育推荐系统（ERS）是利用数据挖掘、机器学习等技术，对教育系统中的资源、课程、教师等进行智能推荐的系统[9]。这种系统通过分析学生的学习历史、兴趣爱好、能力水平等信息，为学生提供个性化的学习资源和课程建议，以提高学习效果和学习满意度，ERS 对教育供需两端的人员都起到重要作用，对学习者来说 ERS 有助于他们制定个性化的学习内容并为他们取得良好成绩提供动力，对教育资源提供者来说，ERS 有助于他们动态调整教育资源，促进他们的教学实践[10]。其核心功能是帮助用户在海量的学习资源中找到最适合自己的内容，从而提高学习效率和体验。

1.2 课程推荐方法

基于内容的推荐通过分析相关对象的内容特征和用户历史偏好进行推荐，比较依赖内容特征的质量与丰富度[11]，基于课程内容的推荐系统利用课程描述文本如TF-IDF向量化[12]或知识点标签匹配，计算课程特征与用户画像的相似度推荐匹配度高的新课程[13]。

协同过滤（CF）通过分析用户 - 课程交互数据如评分、点击挖掘群体偏好规律，分为两类：

（1）基于用户：找到与目标用户兴趣相似的用户群体，推荐其偏好但目标用户未接触的课程如用户 A 和 B 选课相似，则 B 选的课可推荐给 A。

（2）基于课程：根据课程被共同选择的历史模式，推荐与目标课程相似的其他课程如选修“机器学习”的用户常选“深度学习”，则后者可被推荐。CF 优势在于无需课程内容特征，但需解决冷启动和数据稀疏问题，常用于在线教育平台（如 Coursera）的个性化课程推荐[14]。

混合推荐系统结合协同过滤（CF）、基于内容（CB）及其他方法（如深度学习），通过互补优势提升推荐效果。混合推荐能缓解冷启动（新课程用内容特征）、数据稀疏（行为不足时用知识关联），适用于在线教育平台（如 Coursera）的个性化学习路径规划，兼顾兴趣与学科体系连贯性[15]。

本文针对现有方法对成绩数据利用不足的缺陷，提出以先修成绩特征为核心的推荐框架，以预测学生是否适合选修目标课程如《化学制药工艺技术》。提供可解释的推荐依据（如关键先修课程的影响权重）。通过持续学习新数据动态优化推荐策略。

2 系统设计与方法

该系统将有助于根据前几个学期的表现为即将到来的学期提供个性化的课程推荐。

2.1 数据预处理

数据来源于某职业技术学院 4 个入学年度的学生课程成绩真实数据，包含 1 039 条学生记录的结构化数据，涵盖 5 门前置课程成绩及目标变量，如表 1 所示。目标变量即拟推荐的课程，以化学制药工艺技术为例，5 门先修课程分别是单元操作技术 A1、分析测试技术 B、化学基础（1）、化学基础（2）、化工制图，表 1 中化工制图为选修课程未选修学生则为空值。

表 1 数据示例表

学号	化学基础（1）	单元操作技术 A1	化工制图	化学基础（2）	分析测试技术 B	化学制药工艺技术
8 126	70	76		85	77	74
7 151	60	85	76	60	65	67

表 2 是对表 1 中字段值域的描述。包括（1）数据清洗：删除无关字段（序号、学号）。（2）处理学号脱敏。（3）缺失值处理：“缓考”标记，使用课程均值填充；“旷考”标记，视为 0 分并添加二值标签。由于成绩都是百分制，无需转换即认为都是标准化成绩。

表 2 数据字段说明表

字段名称	数据类型	描述	示例值
化学基础（1）	数值型	0~100 分或特殊标记	70 / 缓考
单元操作技术 A1	数值型	0~100 分	76
化工制图	数值型	0~100 分	76/ 缺考
化学基础（2）	数值型	0~100 分	85/ 空缺
分析测试技术 B	数值型	0~100 分	77/ 空缺
目标课程成绩	数值型	待预测的成绩	78

2.2 推荐算法设计

随机森林[16]是一种基于集成学习的监督学习算法，通过构建多棵决策树并集成其预测结果，显著提升模型的泛化能力与鲁棒性。其核心思想结合了 Bagging（bootstrap aggregating）[17] 与特征随机性两大策略，具体实现如下：

（1）Bagging 自助采样

从原始训练集 D 中有放回地随机抽取 N 个样本，生成 T 个不同的子训练集 $D_1,D_2,\cdots,D_T$（允许样本重复）。每个子集训练一棵决策树，通过多树投票降低单棵树的过拟合风险。

（2）特征随机性

每棵决策树在节点分裂时，仅从所有特征中随机选取 m 个候选特征，通常 $m=M$，M 为总特征数，从中选择最优分裂点，进一步增强模型多样性。

（3）建模

设训练集包含 N 个样本，每个样本有 M 个特征，即先修课程成绩，标签为二分类变量 $y\in\{0,1\}$，0 表示不推荐，1 表示推荐。

（4）单棵决策树的分裂准则

决策树通过最小化不纯度，如基尼指数或信息增益选择分裂特征与阈值。以基尼指数为例：

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

式中：p_k 为节点中第 k 类样本的占比。

分裂时选择使基尼指数下降最大的特征 j 和阈值 s：

$$\Delta\mathrm{Gini}(D)=\mathrm{Gini}(D)-\left(\frac{|D_{\mathrm{left}}|}{|D|}\mathrm{Gini}(D_{\mathrm{left}})+\frac{|D_{\mathrm{right}}|}{|D|}\mathrm{Gini}(D_{\mathrm{right}})\right) \tag{2}$$

（5）集成预测

对于输入样本 x，随机森林中所有 T 棵树的预测结果为 $\{h_1(x), h_2(x),\cdots,h_T(x)\}$，最终分类结果通过多数投票决定：

$$H(x)=\mathrm{mode}\{h_t(x)\},\ t=1,2,\cdots,T \tag{3}$$

输入特征是 5 门先修课程成绩，包括化学基础（1）、化学基础（2）、单元操作技术 A1、化工制图、分析测试技术 B。输出标签是目标课程《化学制药工艺技术》成绩是否高于 60 分，若是，则 y=1 表示推荐；若否，则 y=0 表示不推荐。

（6）参数设置

树的数量（n_estimators）通过网格搜索确定最优值为 200，平衡计算效率与模型性能。最大深度（max_depth）限制为 10 层，防止过拟合。特征随机数（max_features）设置为特征总数的二次方根（即$m=\sqrt{5}\approx 2$），增强特征多样性。

随机森林通过计算特征在所有树中分裂节点的平均不纯度下降量，量化各先修课程对推荐结果的影响权重：

$$\mathrm{Importance}_j=\frac{1}{T}\sum_{1}^{T}\sum_{j}\Delta\mathrm{Gini} \tag{4}$$

本研究结果显示，化学基础（2）与化工制图的特征重要性占比最高，分别为 32% 和 28%，表明这两门课程是目标课程选修成功的关键前置知识，如图 1 所示。

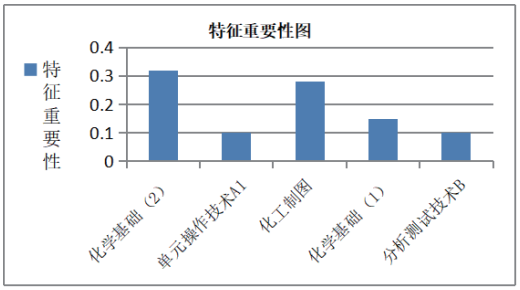

图 1 先修课程对目标课程的重要性图

3 实验与结果分析

3.1 实验设置

数据集划分以此前的 1 039 条真实数据为例，划分训练集 831 条占比 80%，测试集 208 条占比 20%，为确保分布一致性，采用分层抽样，按目标课程成绩分箱。训练数据集各科课程成绩分布如图 2 所示。

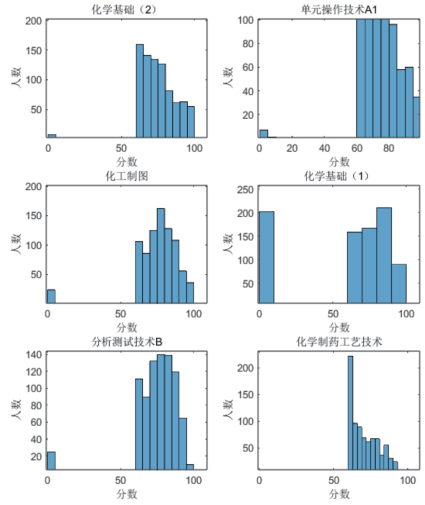

图 2 训练集各科课程成绩分布图

基线模型选择三种基础算法，包括从课程目录中随机选取进行随机推荐，基于 Surprise 库的 SVD 分解的纯协同过滤算法 [18]，基于余弦相似度排序的纯内容推荐 [19]。

评估指标选取准确率（Accuracy）、召回率和 F_1 分数为评判准则，分别表示在测试集上推荐正确的比例、推荐列表中实际选修课程占比、精确率与召回率的调和平均，其值越大表示预测的推荐结果越接近真实结果。

3.2 结果及分析

如表 3 所示，按照成绩高于 60 分进行预测推荐，混合模型较单一模型中最佳的准确率提升 8.6%，召回率和 F_1 分数也相应地提升较多。如果把目标推荐阈值设置为 80 分，则结果性能显著下降，因为以 80 分为界线成绩分布存在极大的类别不平衡问题导致样本失衡。

表 3 模型性能对比

模型	准确率 /%	召回率 /%	F_1 分数
随机推荐	32.1	28.6	0.29
纯 CF	76.3	71.5	0.73
纯 CB	85.2	79.8	0.82
本文模型（60 分阈值）	**93.8**	**88.7**	**0.93**
本文模型（80 分阈值）	88.2	62.1	0.46

3.3 优化措施

引入 SMOTE（synthetic minority over-sampling technique）[20] 过采样，通过合成少数类样本来增加其在数据集中的数量，以达到样本平衡。

采用 XGBoost（extreme gradient boosting）[21] 替代随机森林，通过串行训练多棵决策树，每棵树拟合前一棵树的预测残差（即负梯度方向），逐步降低模型误差。

（1）正则化目标函数

$$\mathrm{Obj}=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{K}\left(\gamma T_k+\frac{1}{2}\lambda\|\omega_k\|^2\right) \tag{5}$$

式中：T_k 表示第 k 棵树的叶子节点数；ω_k 表示叶子权重；γ、λ 表示正则化系数。

（2）二阶泰勒展开优化

对损失函数进行二阶泰勒展开，利用一阶梯度 $\boldsymbol{g}_i$ 和二阶海森矩阵 $\boldsymbol{h}_i$ 加速收敛。

$$\text{Obj} \approx \sum_{i=1}^{n}\left[\boldsymbol{g}_i\omega_q(x_i)+\frac{1}{2}\boldsymbol{h}_i\omega_{q(x_i)}^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \quad (6)$$

（3）贪心建树与分裂准则

$$\text{Gain}=\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}\boldsymbol{g}_i\right)^2}{\sum_{i\in I_L}\boldsymbol{h}_i+\lambda}+\frac{\left(\sum_{i\in I_R}\boldsymbol{g}_i\right)^2}{\sum_{i\in I_R}\boldsymbol{h}_i+\lambda}-\frac{\left(\sum_{i\in I}\boldsymbol{g}_i\right)^2}{\sum_{i\in I}\boldsymbol{h}_i+\lambda}\right]-\lambda \quad (7)$$

式中：I_L、I_R 表示分裂后的左右子节点样本集合。

优化后的 80 分阈值的推荐结果准确率提升至 89.7%，召回率提升至 67%，F_1 分数提升至 0.77，从图 2 可知，化工制图作为重要性第二的先修课程，其 60~100 分段，分布人数呈正态分布特征，均值中心 80 分人数较多，以此为阈值，则左右两侧邻界数据较多且较为接近，因此预测准确率有所下降，可见数据分布对结果存在较大的影响。

3.4 案例研究

学号 7117 的学生推荐结果为推荐选修《化学制药工艺技术》，其中关键影响因素：化学基础（2）成绩 92 分（权重 32%）；化工制图成绩 85 分（权重 28%）。该生推荐选修课程预计分 86 分，实际考试分 88 分，预测准确且结果差异较小。

4 结论

本研究通过随机森林算法构建了基于先修成绩的课程推荐系统，验证了前置课程成绩与后续学习表现的强相关性。实验表明，化学基础（2）与化工制图的成绩对目标课程选修成功率影响显著，优化后的模型在准确率与 F_1 分数上均优于传统协同过滤与内容推荐方法。然而，数据预处理中“缓考”与“旷考”的简化处理可能引入偏差，未来需结合动态学习行为数据（如作业完成率、知识点掌握度）进一步优化画像构建。此外，案例分析的单一性限制了结论的泛化能力，后续可通过多院校数据验证模型迁移性，并探索跨学科课程推荐的可解释性框架。

参考文献：

[1] 陈建校，刘斯琦，左梦雪．人工智能重塑高等教育个性化教学：作用机理与影响效应 [J]. 中国职业技术教育，2025(3): 75-84.

[2] NGUYEN V A, NGUYEN H H, NGUYEN D L, et al. A course recommendation model for students based on learning outcome[J].Education and information technologies, 2021, 26: 5389-5415.

[3] SHAIKH U U, ASIF Z. Persistence and dropout in higher online education: review and categorization of factors[EB/OL].(2022-05-31)[2025-06-23].https://pubmed.ncbi.nlm.nih.gov/35712139/.DOI:10.3389/fpsyg.2022.902070.

[4] SWEENEY M, LESTER J, RANGWALA H,et al. Next-term student performance prediction: a recommender systems approach[EB/OL]. (2016-04-07)[2025-05-04].https://doi.org/10.48550/arXiv.1604.01840.

[5] DUTT A, ISMAIL M A, HERAWAN T. A systematic review on educational data mining[J].IEEE access, 2017, 5:15991-16005.

[6] ROMERO C, VENTURA S.Educational data mining: a review of the state of the art[J]. IEEE transactions on systems, man, and cybernetics, part C (applications and reviews), 2010, 40(6): 601-618.

[7] QUY T L, FRIEGE G, NTOUTSI E. A review of clustering models in educational data science toward fairness-aware learning[J].Educational data science: essentials, approaches, and tendencies, 2023: 43-94.

[8] BLIKSTEIN P, WORSLEY M.Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks[J]. Journal of learning analytics, 2016, 3(2): 220-238.

[9] YAZDI H A, MAHDAVI S J S, YAZDI H A. Dynamic educational recommender system based on improved LSTM neural network[J/OL].Scientific reports,2024[2025-06-12].https://doi.org/10.1080/08839514.2021.2005298.

[10] SILVA F L, SLODKOWSKI B K, SILVA K K A, et al. A systematic literature review on educational recommender systems for teaching and learning: research trends, limitations and opportunities[J]. Education and information technologie, 2023, 28: 3289-3328.

[11] ZHOU H Q. Intelligent personalized content recommendations based on neural networks[J]. International journal of intelligent networks, 2023, 4: 231-239.

[12] LUMINTU I. Content-based recommendation engine using term frequency-inverse document frequency vectorization and cosine similarity: a case study[C/OL] //2023 IEEE 9th Information Technology International Seminar (ITIS). Piscataway:IEEE,2023[2025-06-02].https://ieeexplore.ieee.org/abstract/document/10420137.DOI:10.1109/ITIS59651.2023.10420137.

[13] WANG Y B, GAO S Y, LI W P,et al.Research and application of personalized recommendation based on knowledge graph[C]//Web Information Systems and Applications: 18th International Conference.NewYork:ACM,2021:383 - 390.

[14] 赵秀梅，赵宗昌，袁卫华，等．基于协同过滤的专业学习

（下转第 68 页）

[10]GYGLI M, GRABNER H, RIEMENSCHNEIDER H, et al. Creating summaries from user videos[C]//Computer Vision–ECCV 2014. Piscataway:IEEE,2014: 505-520.

[11]ZHAO B, LI X L, LU X Q. Property-constrained dual learning for video summarization[J]. IEEE transactions on neural networks and learning systems, 2019, 31(10): 3989-4000.

[12]FONTES DE AVILA S E, BRANDÃO LOPES A P, DA LUZ A, et al. VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method[J]. Pattern recognition letters, 2011, 32(1): 56-68.

[13]ZHANG K, CHAO W L, SHA F, et al. Video summarization with long short-term memory[C]//Computer Vision – ECCV 2016. Berlin:Springer,2016: 766-782.

[14]ZHAO B, LI X L, LU X Q. Hierarchical recurrent neural network for video summarization[C]//Proceedings of the 25th ACM international conference on Multimedia. NewYork: ACM, 2017: 863-871.

[15]KHAN H , HUSSAIN T , KHAN S U ,et al.Deep multi-scale pyramidal features network for supervised video summarization[J].Expert systems with application, 2024,237:121288.

[16]ZHONG R, WANG R, ZOU Y, et al. Graph attention networks adjusted Bi-LSTM for video summarization[J]. IEEE signal processing letters, 2021, 28: 663-667.

[17]WU J X, ZHONG S H, JIANG J M, et al. A novel clustering method for static video summarization[J].Mutimedia tools and applications, 2017, 76(7): 9625-9641.

【作者简介】

朱頔（2002—），男，陕西宝鸡人，硕士研究生，研究方向：工业视频摘要。

滕晓宇（1991—），女，山东临沂人，博士，讲师，研究方向：视频摘要、深度学习、工业信息化、可信计算。

王刚（1974—），男，江苏丰县人，博士，教授，研究方向：大数据分析、可信计算、深度学习。

许文丽（1970—），女，山东德州人，博士，副教授，研究方向：大数据分析、数据挖掘、机器学习。

樊懿雯（2003—），女，陕西渭南人，本科，研究方向：工业视频摘要。

何勇（1986—），男，甘肃会宁人，本科，高级工程师，研究方向：工程数字化、工程大数据应用。

张清（1998—），女，山西运城人，硕士，讲师，研究方向：隐私计算、深度学习。

（收稿日期：2025-03-31 修回日期：2025-09-04）

（上接第 63 页）

指导平台建设研究 [J]. 计算机应用与软件 ,2022,39(6):76-81.

[15] 许侃 , 吴鑫卓 , 林原 , 等 . 基于对抗型排序学习的混合推荐算法 [J]. 山西大学学报 (自然科学版),2024,47(3):481-493.

[16] GENUER R, POGGI J M, TULEAU-MALOT C.Variable selection using random forests[J].Pattern recognition letters, 2010, 31(14): 2225-2236.

[17] OTOK B W, MUSA M, PURHADI, et al.Propensity score stratification using bootstrap aggregating classification trees analysis[J]. Heliyon , 2020,6(7): e04288.

[18] MIJANUR RAHMAN M, OAISHI C R,MAHMUD J, et al. Elective course recommendation system using SVD algorithm[C]//Congress on Smart Computing Technologies. Berlin:Springer,2025:461–475.

[19] ABHINAV N, SUJATHA K. Content-based movie recommendation system using cosine similarity measure[C/OL]//Third International Conference on Advances in Physical Sciences and Materials: Icapsm 2022. New York: AIP Publishing, 2022[2025-05-13]. https://doi.org/10.1063/5.0178819.

[20] ELREEDY D, ATIYA A F, KAMALOV F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning[J].Machine learning, 2024, 113: 4903-4923.

[21] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NewYork: ACM, 2016: 785–794.

【作者简介】

涂辉（1987—），男，江苏徐州人，硕士，工程师，研究方向：职业教育、机器学习。

许腾飞（1987—），男，江苏徐州人，硕士，讲师，研究方向：职业教育、人工智能。

丁中燕（1991—），女，江苏盐城人，硕士，助理研究员，研究方向：学生管理、职业教育。

张正金（1984—），男，安徽合肥人，博士，讲师，研究方向：人工智能、因果推理。

（收稿日期：2025-04-16 修回日期：2025-09-10）