计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):339−346 [doi: 10.15888/j.cnki.csa.009039]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合 XGBoost 与 FM 的混合式学习成绩分类预测①

章　刘[1], 陈逸菲[2], 熊　雄[1], 裴梓权[1], 唐乃乔[1]

[1](南京信息工程大学 自动化学院, 南京 210044)
[2](无锡学院 自动化学院, 无锡 214105)
通信作者: 陈逸菲, E-mail: 20201249056@nuist.edu.cn

摘　要: 综合考虑混合式学习成绩分类预测中数据存在不平衡性和稀疏性的特点, 提出了一种 SMOTE-XGBoost-FM 混合式学习成绩分类预测模型. 首先通过 SMOTE 采样均衡数据集; 针对数据稀疏性问题, 使用 XGBoost 对采样后的数据进行特征交叉, 然后对所生成树的叶子节点进行独热编码, 以生成高阶特征数据, 最后将其输入因子分解机 (FM) 进行迭代训练以获最优模型. 实验结果表明, SMOTE-XGBoost-FM 模型在混合式学习成绩分类预测中准确率达到了 92.7%, 相较于单一的 XGBoost、FM 模型分别提升了 5.7% 和 11.7%, 能有效对学生学习情况进行分类预测, 为提高教学效果提供参考.

关键词: 混合式教学; 成绩预测; 机器学习; XGBoost; 因子分解机 (FM)

引用格式: 章刘,陈逸菲,熊雄,裴梓权,唐乃乔.融合 XGBoost 与 FM 的混合式学习成绩分类预测.计算机系统应用,2023,32(4):339–346. http://www.c-s-a.org.cn/1003-3254/9039.html

Blended Learning Grade Classification Prediction Based on XGBoost and FM

ZHANG Liu[1], CHEN Yi-Fei[2], XIONG Xiong[1], PEI Zi-Quan[1], TANG Nai-Qiao[1]

[1](School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](School of Automation, Wuxi College, Wuxi 214105, China)

Abstract: By comprehensively considering the imbalance and sparsity of data in blended learning grade classification and prediction, this study proposes a blended learning grade classification and prediction model, namely, SMOTE-XGBoost-FM. Firstly, an equalization data set is sampled by SMOTE. In order to solve the problem of data sparsity, XGBoost is used to perform feature overlap on the sampled data, and then the leaf nodes of the generated tree are processed by one-hot encoding to generate high-order feature data. Finally, the data are input into a factorization machine (FM) for iterative training to obtain the optimal model. The experimental results show that the SMOTE-XGBoost-FM model achieves an accuracy of 92.7% in blended learning grade classification and prediction, which is 5.7% and 11.7% higher than that of single XGBoost and FM models, respectively. Therefore, it can effectively classify and predict students' learning effects and provide a reference for improving teaching efficiency.

Key words: blended teaching; grade prediction; machine learning; XGBoost; factorization machine (FM)

1　前言

自 2020 年初新冠疫情爆发以来, 在疫情防控常态化的背景下, 线上与线下相结合的混合式教学成为各高校的首选教学方式[1,2]. 如果能够在复杂繁琐的混合式

① 基金项目: 江苏省自然科学青年基金 (BK20210661); 江苏省研究生实践创新计划 (SJCX22_0353); 江苏省高等学校自然科学研究面上项目 (19KJB520044); 南京信息工程大学无锡校区创新实践项目 (WXCX202117)
收稿时间: 2022-08-26; 修改时间: 2022-09-27; 采用时间: 2022-10-27; csa 在线出版时间: 2022-12-23
CNKI 网络首发时间: 2022-12-28

学习数据中提取有用的信息, 结合机器学习算法对学生的学习结果进行早期预测, 可以帮助教师及时对有挂科风险的学生实施干预, 这对于教学工作具有重要意义[3–8].

近年来, 国内外学者在学生成绩分析预测方面进行了大量研究[9–11]. 在预测方法上, Akçapınar 等[12] 采用 KNN 对在线学习中学生期末成绩进行分类预测. 刘博鹏等[13] 采用学生的行为、个人属性和历史成绩等数据通过支持向量机对学生成绩进行预测. 但以上算法较为单一, 在数据量较大情况下模型的预测效果欠佳. 为此, 部分学者采用随机森林[14,15]、CatBoost[16]、XGBoost[17] 等集成树模型提升预测效果. 还有学者采用 Stacking[18–20] 集成策略将多个单一算法进行集成预测, 相比与单一算法预测效果有一定提升. 但在模型训练前需要对稀疏的数据进行大量人工特征工程工作, 较为费时费力. 此外, 由于学生成绩数据存在不平衡性, 类别占比较多的一类预测效果较好, 而类别较差的一类相对较差, 易造成模型的整体预测效果变差[21], 而大部分学生成绩预测模型缺乏对不平衡数据均衡处理.

综合上述研究工作, 虽然目前在成绩预测领域已取得一定进展, 但仍然存在一些问题有待解决: (1) 学生行为特征依赖人工特征工程提取和需要领域专家的参与, 容易忽略一些对分类预测目标造成影响的因素; 如在数据预处理过程中, 需要根据学生的学习特征分布, 主观对部分学习特征进行分箱操作[22], 如果分箱不合理, 可能会造成学生的特征信息丢失; (2) 没有考虑对数据集的学生成绩类别不平衡性对模型的影响. 针对以上问题, 为减小不平衡数据对模型分类预测效果的影响, 并进一步提高对学生成绩分类预测模型的泛化能力, 本文开展了以下工作, 并区别于已有研究.

1) 通过 SMOTE 算法对数据集进行过采样, 以降低不平衡数据对模型的影响.

2) 采用 SHAP (Shapley additive explanation) 模型对影响学生成绩的因素进行分析、特征选择, 增强预测模型的泛化能力.

3) 通过融合 XGBoost 和因子分解机 (FM) 建立学习成绩分类预测模型, 减少传统成绩预测基线模型对人工特征工程的依赖.

2 SMOTE-XGBoost-FM 分类预测模型

2.1 问题定义

学习成绩预测的目的是根据学生历史行为数据预测未来的学习结果, 在混合式学习环境下, 学生历史行为数据主要包括个人基本信息、线上和线下学习数据等. 下面给出问题的定义.

学习行为特征: 行为特征$X_m=\left\{x_m^1,x_m^2,\cdots,x_m^t\right\}$定义为某个学生的所有学习行为数据集合, 其中$x_m^t$为学生 m 的第 t 个教学周的历史行为数据, 为了实现对学生学习结果进行早期预测, 以$x_m^t=\left\{x_{m1}^t,x_{m2}^t,\cdots,x_{mn}^t\right\}$作为样本特征, 其中$x_{mn}^t$为学生 m 的第 t 个教学周的第 n 个行为特征数据.

学习成绩分类预测: 给定学生 m 的第 t 周学习行为特征$x_m^t=\left\{x_{m1}^t,x_{m2}^t,\cdots,x_{mn}^t\right\}$预测$y'_m$. y'_m表示学生的期末学习成绩类别标签, y'_m有 1 和 0 两个状态, 分别表示学生 m 能和未能通过期末检测.

2.2 SMOTE-XGBoost-FM 模型

以混合式学习下学习行为数据为研究对象, 本文结合 SMOTE、XGBoost 和 FM 提出 SMOTE-XGBoost-FM 分类预测模型, 用于均衡样本不平衡类别、提取数据中深度隐藏的信息和提高整体模型的泛化性能. 如图 1 所示, SMOTE-XGBoost-FM 分类预测模型主要分为 4 个核心阶段.

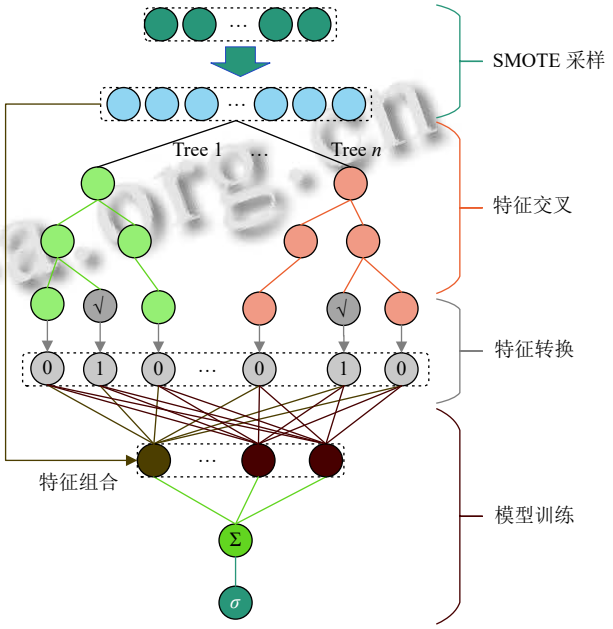

图 1 SMOTE-XGBoost-FM 分类预测模型原理图

1) SMOTE 采样阶段, 通过对原始数据集中学生类别较少的一类进行 SMOTE 过采样, 弥补两个类别间的数量差异均衡数据.

2) 特征交叉阶段, 采用 XGBoost 树结构模型对类

别均衡的特征数据进行训练, 通过树分裂的方式进行特征交叉, 最终生成树结构. XGBoost 对特征数据的训练可以看作是 XGBoost 中每棵决策树对特征的交叉组合, 其中从每棵树的根节点到叶子节点的这条路径可以看作是不同特征之间交叉组合. 叶子节点数可看作为新特征数, 每个样本在所有叶子节点的编码为新的样本特征值.

3) 特征转换阶段, 计算每个样本在每棵树各叶子节点所得到的预测概率值, 将每个样本的预测概率值所属的叶子节点独热编码, 通过以上方式将原始特征重新变换, 获得新稀疏特征矩阵.

4) 模型训练阶段, 将新特征数据输入因子分解机 (FM) 进行训练, 最后通过 Sigmoid 函数将结果分为两类.

2.3 SMOTE-XGBoost-FM 分类预测算法

假设融合模型 SMOTE-XGBoost-FM 的输入训练数据集为$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, 实例样本的特征为$x_m^t=\left\{x_{m1}^t,x_{m2}^t,\cdots,x_{mn}^t\right\}$, 样本特征对应的标签为$y_m$, 样本的特征维度为 n, 记期末考试中学生能通过检测的样本集为D_p, 未能通过检测的样本集为D_f, 其中 p 和 f 分别为多数类别与少数类别样本数量, $p<f$ 且 $p+f=m$, XGBoost 学习器为$\hat{y}(x)$, FM 学习器为$y(x)$. 经过 SMOTE 采样后所产生的新数据集为$D_{\text{new}}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_s,y_s)\}$然后经过 XGBoost 学习器$\xi$的训练, 最终生成 j 个叶子节点, 然后获取各叶子节点对应索引$u_s=\left\{u_{s1},u_{s2},\cdots,u_{sj}\right\}$, 如果叶子节点$u_{sj}$为学生样本$x_m$最终分裂所在位置, 则记叶子节点$u_{sj}$所对应值$q_{sj}=1$, 否则记$q_{mj}=0$. 最终获得新特征集合$Q_s=\left\{q_{s1},q_{s2},\cdots,q_{sj}\right\}$, 将新特征构建新数据集$D'=\{(q_1,y_1),(q_2,y_2),\cdots,(q_s,y_s)\}$, 并将其输入 FM 学习器$\psi$训练预测, 最后得出每个样本最终预测结果$y'_s$. 具体融合算法伪代码如算法 1 所示.

算法 1. 融合算法

输入: 原始数据集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, 样本特征值$x_m=\{x_{m1},x_{m2},\cdots,x_{mn}\}$, XGBoost 学习器为$\xi$, FM 学习器为$\psi$

过程:

```
D_new←SMOTE(D)
for i=1,2,⋯,s do
    u_s←ξ_s(D)
    u_s={u_s1,u_s2,⋯,u_sj}
        if u_sj ← x_s
            q_sj = 1
        else q_sj = 0
    Q_s={q_s1,q_s2,⋯,q_sj}
end for
D′={(q_1,y_1),(q_2,y_2),⋯,(q_s,y_s)}
for i=1,⋯,s do
    y′_s=ψ_s(D′)
end for
```

输出: 每个样本最终预测结果y'_s

SMOTE-XGBoost-FM 算法具体步骤如下.

1) 根据样本类别权重, 通过公式$N=\text{int}\left(\frac{p-f}{f}\right)$对未能通过期末检测的学生样本设置一个采样倍率 N, 其中 N 为整数, 且大于 0. 由采样倍率 N 和少数类别样本数量 f, 计算出未能通过期末检测的样本集 D_f 需要合成的样本数量为 Nf.

2) 对样本集 D_f 中每个样本x_i采用欧式距离计算其到 D_f 中其他所有样本的距离, 得到 k 近邻, 在x_i的 k 近邻中随机选择样本x_j, 按照$x_n=x_i+rand(0,1)\times|x_i-x_j|$计算方式得出新样本$x_n$, $rand(0, 1)$ 表示区间 (0, 1) 的随机数.

3) 根据数据集中未能通过期末检测样本类别的采样倍率 N 重复进行采样过程, 直到满足步骤 1) 中所需要合成的新样本数量. 数据集 D 经过平衡后生成数据集记为$D_{\text{new}}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_s,y_s)\}$, 其中 $s=m+Nf$.

将经过 SMOTE 采样平衡后的新数据集$D_{\text{new}}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_s,y_s)\}$输入 XGBoost 学习器$\hat{y}(x)$中进行迭代训练, 并通过网格搜索调整学习器参数, 迭代训练公式如下:

$$\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i) \tag{1}$$

其中, $\hat{y}_i^{(t)}$表示树结构模型的第 t 次迭代预测结果, 在每一次梯度提升迭代中, 通过损失函数处理残差来修正前一次迭代预测的结果, $f_k(x_i)$ 表示第 k 棵树给第 i 个样本的预测值.

模型每次以式 (2) 为求解最小化目标函数进行迭代:

$$obj^{(t)}=\sum_{i=1}^{n}l\left(y_i\hat{y}_i^{(t)}\right)+\sum_{i=1}^{n}\Omega(f_i) \tag{2}$$

其中, $\sum_{i=1}^{n}l\left(y_i\hat{y}_i^{(t)}\right)$为损失函数部分, 其用来衡量预测值$\hat{y}_i$和目标值$y_i$之间的差异. $\Omega(f_i)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_{j2}$为正则化项, 式中$\gamma$为 L1 正则项, λ为 L2 正则项.

4) 经过步骤 4) 迭代训练后建立树结构, 记树结构的叶子节点个数为 j, 获取各叶子节点对应索引 $u_s=\{u_{s1},u_{s2},\cdots,u_{sj}\}$, 如果叶子节点 u_{sj} 为样本 x_s 最终分裂所在的位置, 则记叶子节点 u_{sj} 所对应值 $q_{sj}=1$, 否则记 $q_{sj}=0$, 最终获得新样本特征集合 $q_s=\{q_{s1},q_{s2},\cdots,q_{sj}\}$.

5) 将步骤 5) 获得的样本特征集 $q_s=\{q_{s1},q_{s2},\cdots,q_{sj}\}$ 构建新的数据集 $D'=\{(q_1,y_1),(q_2,y_2),\cdots,(q_s,y_s)\}$, 将其按一定比例切分为训练集 D'_{train} 和测试集 D'_{test}.

6) 将训练集 D'_{train} 送入因子分解机学习器 $\psi(x)$ 进行训练, $y(x)$ 算法函数如式 (3) 所示:

$$y(x)=\sigma\left(w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}<v_i,v_j>x_ix_j\right) \tag{3}$$

其中, x_i 为样本 x 的第 i 个特征值, n 为样本特征的维度. $w_0\in R$ 为全局偏差, $w_i\in R^n$ 表示第 i 个特征的影响因子, $<v_i,v_j>=\sum_{f=1}^{k}v_{i,f}\cdot v_{j,f}$ 表示两 k 维向量的内积, 隐向量长度 k 为一超参数, $k<<n$. σ 为 Sigmoid 函数, 设置在模型的输出上, 将模型输出分类转换.

7) 将因子分解机学习器 $y(x)$ 通过网格搜索和 5 折交叉验证来调整学习器参数, 使其获最佳效果, 最后得出每个样本最终预测结果 y'_m, 并利用测试集 D'_{test} 验证 SMOTE-XGBoost-FM 模型的泛化能力.

3 实验与结果分析

本文实验环境为 CPU Inter i7-9700, RAM 为 12 GB, 编译软件 PyCharm 2021.3.3, 算法模型使用 Python 实现. 为了展示均衡不平衡数据集的必要性和验证融合 XGBoost 与 FM 算法在学生成绩分类预测中的优越性, 随机将数据集按 7:3 切分为训练集与测试集用来训练与测试, 设置两组对比实验, 对比 logistic regression (LR)、gradient Boosting decision tree (GBDT)、random forest (RF)、XGBoost、FM、XGBoost+FM 在数据集上的准确率、召回率、AUC 值等评价指标, 以及对比未采用 SOMTE 采样与采用 SOMTE 采样的模型效果.

3.1 数据集介绍

本文所研究的数据采集自学校超星学习通平台与教务系统, 共 129 名学生, 一个学期 18 个教学周, 共提取出 16 个学习行为特征, 通过按周进行样本数据提取, 最终共获 2 322 个样本. 数据集中各学生学习行为特征含义如表 1 所示.

表 1 特征含义说明

序号	特征名称	特征含义	类型
1	knowledge_points	知识点任务完成数	int
2	class	班级	int
3	video_progress	课程视频进度	float
4	viewing_duration	学生登录超星平台观看课程视频的总时长	float
5	discussion	讨论次数	int
6	study_counts	学生登录超星平台学习课程内容的总次数	int
7	group_tasks	分组任务得分	int
8	questions	课堂回答问题得分	int
9	chapter_quiz	章节测验得分	float
10	homework	课后作业成绩评分	int
11	sign_in	课堂签到	int
12	interactive	学生参加互动的得分	int
13	accumulate_points	课程积分	int
14	class_test	学生参加随堂测验的评分	float
15	voting	投票数	int
16	final_mark	期末卷面成绩	int

3.2 数据预处理

将原始数据集按周划分并集成新的数据集, 新的数据集中共有 2 322 个样本数据, 通过 Python 中 pandas 库的 describe 函数对新的数据集进行描述性统计, 结果如表 2 所示. 可以发现, 各特征的最小值除了 interactive、sign_in 等几个特征外几乎全为 0, 数据集中各特征数据的 25% 分位中仍然存在大量的 0 值特征, 特征数据的 75% 分位中部分特征依然为 0, 因此可以看出数据较为稀疏. Final_mark 为期末考试卷面成绩, 由于存在缺考同学, 所以存在零分现象, 因此需要缺失值填充. video_progress、questions 等特征由于系统统计问题而存在负值, 对其置 0 处理. 由于本文采用的预测方式为分类预测, 预测问题为识别学生是否存在课程学业风险, 因此将学生的期末成绩分为 failed 与 pass 两类. 经处理后共获得 failed 类别 839 例, pass 类别 1 483 例, 两个类别占比情况如图 2 所示, 从图中可以看出两个类别的分布占比较不平衡, pass 类别占比更大.

3.3 特征选择

特征选择可以减少冗余特征, 提高模型建立的速度, 增强模型泛化能力, 减少模型过拟合问题. 本文采用 SHAP 值来描述和评估特征的重要性, 基于 XGBoost 模型利用 SHAP 值选择最终输入模型训练的特征. SHAP 值可以将每个特征中的所有样本整体可视化[23]. 如图 3 所示, 图中表示特征对每个样本的影响, 每一行是一个特征, 每个点表示一个样本, 横坐标为 SHAP 值, 颜色表示特征值的强弱, 红色和蓝色分别表示高值

和低值. 图 3 中所示, chapter_quiz、discussion 等特征的较低值会增加学生挂科风险的概率, 而 knowledge_points、video_progress、viewing_duration 等特征的较高值会增加学生挂科风险的概率 (可能的原因是个别学生为了获得更好的平时成绩, 将未来需要完成的任务提前刷完, 从而造成个别样本的 knowledge_points、video_progress、viewing_duration 等特征值异常高). 每个特征的平均 SHAP 值表示其特征的重要性, 特征重要性排序如图 4 所示, 从图中可以看出 chapter_quiz 特征最重要, voting 和 questions 特征的重要性几乎为零. 因此, 在模型训练时采用前 13 个特征, 将最后两个 voting 和 questions 特征删除.

表 2 数据集描述信息

Feature	Mean均值	Std标准差	Min最小值	25%分位数	75%分位数	Max最大值
knowledge_points	5.4276	14.4869	0	0	5	43
video_progress	3.4716	10.2968	−29	0	2	56
viewing_duration	24.7501	69.1893	0	0	16	226
study_counts	17.1395	23.2818	0	2	24	210
chapter_quiz	19.1591	18.3864	0	4.85	26.8	100
discussion	16.8243	18.7012	0	0	30	88
homework	64.6251	23.8697	0	52.6	80.675	100
sign_in	96.8372	6.6664	25	96.125	100	100
interactive	26.3343	16.1081	0.7	12.7	38.3	73
group_tasks	45.9134	45.6336	0	0	94	100
accumulate_points	10.2537	5.5217	0	6	14	38
voting	2.4531	2.2513	0	0	4	10
questions	0.0999	0.6244	−2	0	0	11
class_test	4.1146	4.1744	0	0	6	16
final_mark	53.2500	27.7879	0	31.25	75.625	100

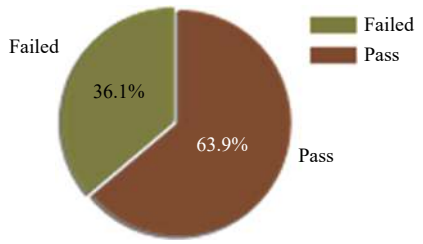

图 2 不同类别占比统计

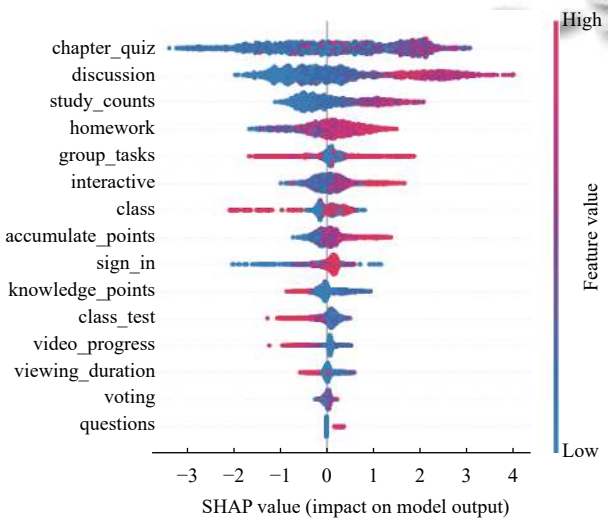

图 3 基于 SHAP 的整体特征可视化

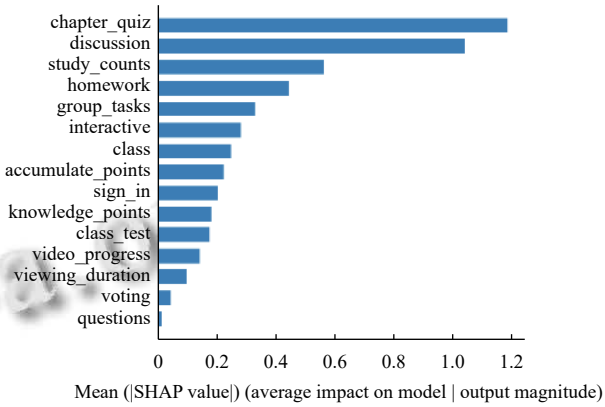

图 4 基于 SHAP 的特征重要性排序

3.4 模型的泛化能力评估指标

预测学生能否通过期末检测, 即将预测结果分为通过 (pass) 和挂科 (failed) 两类, 这是一个二分类预测问题, 其混淆矩阵分析表如表 3 所示, 其中 *TP*、*TN*、*FP* 和 *FN* 分别代表真阳性、真阴性、假阳性和假阴性.

本文选择了准确率 *A* (accuracy)、召回率 *R* (recall)、精确率 *P* (precision)、*F*1 (*F*1 score) 得分和接收器工作特性 (ROC) 曲线作为评估标准. 精确率的通过假阳性来衡量模型的准确性, 即被误判为不能通过期末考试的学生数量, 假阳性越低, 模型的精确率越高. 召回率

则通过假阴性来衡量模型的性能, 即真正挂科学生中预测为挂科学生的数量. 在学习预测的背景下, 真正挂科学生的预测准确性被认为比能通过期末考试的学生更重要. 因此, 就评估模型而言, 召回率比精确率更重要. ROC 曲线为真阳性率和假阳性率的对比图, 用来进一步评估模型的性能. AUC 定义为 ROC 曲线下的面积, 面积值越大, 模型的效果越好. 准确率、召回率、精确率和 $F1$ 得分定义如下:

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$P=\frac{TP}{TP+FP} \tag{5}$$

$$R=\frac{TP}{TP+FN} \tag{6}$$

$$F1=\frac{2P\times R}{P+R} \tag{7}$$

3.5 模型参数设置

选择合适模型参数可以进一步优化模型, 充分发挥模型的性能, 网格搜索法是机器学习中有效的调整模型参数方法. 本文采用 5 折交叉验证验证方式将训练集进一步划分为 5 份, 循环选取其中一份作为最优参数的验证集, 以准确率作为模型预测效果的评价指标, 通过网格搜索法选取各模型的最有参数. 本文中采用的模型最终参数设置如表 4 所示.

表 3 混淆矩阵分析表

混淆矩阵		预测值	
		挂科学生	通过学生
真实值	挂科学生	*TP* (真挂科学生)	*FN* (假通过学生)
	通过学生	*FP* (假挂科学生)	*TN* (真通过学生)

3.6 结果对比分析

采用 SMOTE 算法对原始数据集过采样, 原始数据集的不平衡标签类别经过均衡后生成新数据集, 利用 XGBoost 对新特征数据进行特征交叉, 充分挖掘特征数据信息, 最后采用因子分解机 (FM) 预测学生能否通过期末考试. 融合 XGBoost 与因子分解机 (FM) 的目的是在较少人工特征工程情况下充分挖掘数据集的信息, 提高模型的泛化能力. 为验证所提模型的有效性, 本文设置了相关对比实验, 根据设置的两组对比实验对所有的预测结果进行定性和定量分析比较.

表 4 模型参数设置

算法	参数
LR	Penalty=l2; max_iter=300; solver= L-BFGS; tol=0.0001
GBDT	learning_rate=0.04; loss= deviance; max_depth=6; min_samples_leaf=1; n_estimators=150
RF	Criterion= Gini; learning_rate=0.04; n_estimators=150; min_samples_split=2; min_samples_leaf=1
XGBoost	min_samples_split=2; n_estimators=150; booster=gbtree; binary=logistic; max_depth=6; colsample_bytree=0.5
FM	learning_rate=0.04; max_iter=300
XGBoost+FM	min_samples_split=2; n_estimators=150; booster=gbtree; binary=logistic; max_depth=6; colsample_bytree=0.5; learning_rate=0.04; max_iter=300

(1) 基础模型泛化性能对比分析

图 5 为各模型在未 SMOTE 采样情况下根据预测结果绘制的 ROC 曲线, 从图中可以看出, XGBoost+FM 的 AUC 达到了 95.7%, 预测效果最好. 表 5 为各模型在未 SMOTE 采样与 SMOTE 采样情况下各模型预测的精确率、召回率、$F1$ 得分等结果. 从表 5 中可以看出, 在未 SMOTE 采样的情况下, 各模型的 failed 类别预测的精确率、召回率、$F1$ 得分相对于 pass 类别皆较偏低. 其中, LR 模型预测的 failed 类别精确率、召回率分别为 29.0% 和 41.6%, pass 类别的精确率、召回率分别为 94.5% 和 81.1%, 相差最大. 可以看出样本类别较少的 failed 类别预测效果较差, 数据集的不平衡性对模型的预测效果有一定的影响. 此外, 从表 5 可以看出在 6 个模型中 XGBoost+FM 各项预测效果最优.

(2) 不平衡过采样后各模型对比分析

图 6 为原始数据集经过 SMOTE 采样后根据各模型预测结果绘制的 ROC 曲线, 可以看出, SOMTE+XGBoost+FM 模型的 AUC 在 SMOTE 采样情况下达到了 97.5%, 预测效果最好. 对比图 5 可以看出, 在 SMOTE 采样后 SOMTE+XGBoost+FM 模型的 AUC 提升了 1.8%. 表 6 为各模型在 SMOTE 采样情况下各模型预测的精确率、召回率、$F1$ 得分等结果. 通过对比表 5 和表 6 中的各项结果可以看出, 经过 SMOTE 采样后, 各模型 failed 类别预测的精确率、召回率、$F1$ 得分都

得到了一定的提升, 各模型 failed 类别与 pass 类别预测的精确率、召回率、*F*1 得分之间差异变小, 且各模型整体的预测准确率和 AUC 值等都有一定提升, 其中 SOMTE+XGBoost+FM 的准确率达到了 92.7%, AUC 达到了 97.5%, 效果最好, 相比于 SOMTE+XGBoost 高出 4.3%, 相比于 SOMTE+FM 高出 11.7%.

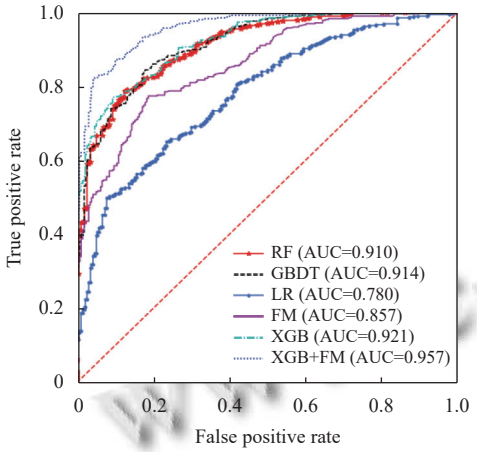

图 5 未 SMOTE 采样的模型 ROC 曲线与 AUC 值

表 5 未 SMOTE 采样各模型结果对比

模型	类别	召回率	精确率	*F*1得分	准确率	AUC
GBDT	Failed	0.808	0.788	0.798	0.859	0.914
	Pass	0.887	0.898	0.892		
LR	Failed	0.740	0.290	0.416	0.714	0.780
	Pass	0.711	0.945	0.811		
RF	Failed	0.827	0.759	0.792	0.859	0.910
	Pass	0.875	0.914	0.894		
XGBoost	Failed	0.833	0.776	0.803	0.867	0.921
	Pass	0.882	0.915	0.899		
FM	Failed	0.708	0.702	0.705	0.793	0.875
	Pass	0.839	0.843	0.841		
XGBoost+FM	Failed	**0.878**	**0.861**	**0.879**	**0.917**	**0.957**
	Pass	**0.926**	**0.947**	**0.937**		

综上所述, 通过 SMOTE 算法对原始数据集进行过采样能够有效均衡数据集中不平衡类别, 对 XGBoost 和 FM 进行融合可以取得较好的分类预测效果, 这是因为 FM 是一个复杂度为线性的模型, 可以很容易地处理大量数据. 但是尽管 FM 存在二阶特征交叉, 却无法进行更高阶特征交叉, 需要通过大量的特征工程才能得到改进. 特征决定了算法模型的预测效果上限, 不同的算法模型只是在逼近该上限的距离上有所不同. 然而, 人工特征工程既耗时又费力, 且最后效果往往不太理想. 因此, 本文通过 XGBoost 进行特征交叉, 生成新的高阶特征, 以弥补人工经验的不足, 充分挖掘特征之间的信息, 最后为避免特征信息丢失, 将新的特征与原始特征一起输入 FM 进行训练. 如图 4, 图 5 以及表 5, 表 6 所示, 相比于单一算法模型以及主流的机器学习模型, 本文所采用的 SMOTE-XGBoost-FM 模型分类预测效果更佳, 进一步验证了其在混合式教学下学生成绩分类预测中的适用性.

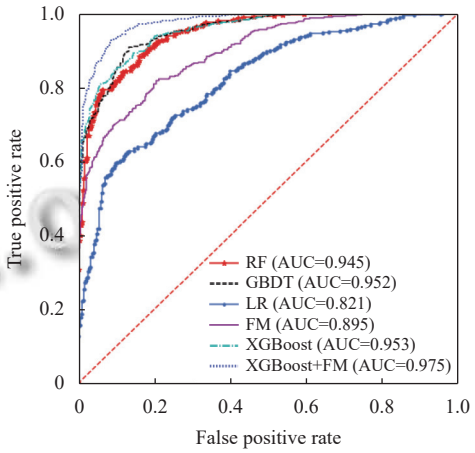

图 6 SMOTE 采样后的模型 ROC 曲线与 AUC 值

表 6 SMOTE 采样各模型结果对比

模型	类别	召回率	精确率	*F*1得分	准确率	AUC
SMOTE+GBDT	Failed	0.847	0.892	0.869	0.865	0.952
	Pass	0.886	0.838	0.861		
SMOTE+LR	Failed	0.708	0.755	0.731	0.722	0.821
	Pass	0.738	0.670	0.713		
SMOTE+RF	Failed	0.848	0.892	0.870	0.866	0.945
	Pass	0.886	0.840	0.863		
SMOTE+XGBoost	Failed	0.849	0.910	0.879	0.874	0.953
	Pass	0.903	0.838	0.870		
SMOTE+FM	Failed	0.790	0.845	0.817	0.810	0.895
	Pass	0.833	0.775	0.803		
SMOTE+XGBoost+FM	Failed	**0.922**	**0.932**	**0.927**	**0.927**	**0.975**
	Pass	**0.932**	**0.921**	**0.926**		

4 总结

本文针对混合式教学下产生的学习行为数据, 基于 SHAP 模型进行特征选择, 采用 SMOTE 算法平衡数据集不平衡类别, 融合 XGBoost 和 FM 模型, 并设置了 6 组对照实验对学生成绩分类预测进行了实证研究. 实验结果表明:

(1) 采用 SHAP 模型特征选择和通过 SMOTE 算法平衡数据集不平衡类别可以有效提高模型的分类预测性能.

(2) 相比与单一算法模型, XGBoost 和 FM 融合模型对于学生成绩分类预测的效果更好, XGBoost+FM

能够充分挖掘变量之间复杂的相互作用和非线性关系, SMOTE-XGBoost-FM 分类预测的准确率和 AUC 值分别为 92.7% 和 97.5%, 取得了最好的预测效果.

本文的研究成果可以帮助教育工作者提前发现未来可能出现学业问题的学生, 以及帮助学生在早期发现自己的不足. 考虑到学生学习行为数据具有时间序列因素, 从学习行为数据中挖掘出时序因素对学习效果的影响是下一步的工作.

参考文献

1 何克抗. 关于我国教育技术学研究现状和教育变革着力点的思考. 电化教育研究, 2018, 39(8): 5–14. [doi: 10.13811/j.cnki.eer.2018.08.001]

2 林健. 工程教育的信息化. 高等工程教育研究, 2022, (1): 1–10.

3 Angeli C, Howard SK, Ma J, *et al*. Data mining in educational technology classroom research: Can it make a contribution? Computers & Education, 2017, 113: 226–242.

4 Nahar K, Shova BI, Ria T, *et al*. Mining educational data to predict students performance: A comparative study of data mining techniques. Education and Information Technologies, 2021, 26(5): 6051–6067. [doi: 10.1007/s10639-021-10575-3]

5 罗明. 教育测评知识图谱的构建及其表示学习. 计算机系统应用, 2019, 28(7): 26–34. [doi: 10.15888/j.cnki.csa.006977]

6 Xu X, Wang JZ, Peng H, *et al*. Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 2019, 98: 166–173. [doi: 10.1016/j.chb.2019.04.015]

7 Cassells L. The effectiveness of early identification of 'At Risk' students in higher education institutions. Assessment & Evaluation in Higher Education, 2018, 43(4): 515–526.

8 Wu BQ, Wu B, Zheng CL. An analysis of the effectiveness of machine learning theory in the evaluation of education and teaching. Wireless Communications & Mobile Computing, 2021, 2021: 4456222.

9 任鸽, 吴猛, 汗古丽·力提甫, 等. 基于改进 Apriori 算法的高校课程预警规则库构建. 计算机系统应用, 2021, 30(7): 290–295. [doi: 10.15888/j.cnki.csa.008040]

10 柴艳妹, 雷陈芳. 基于数据挖掘技术的在线学习行为研究综述. 计算机应用研究, 2018, 35(5): 1287–1293. [doi: 10.3969/j.issn.1001-3695.2018.05.002]

11 郭鹏, 蔡骋. 基于聚类和关联算法的学生成绩挖掘与分析. 计算机工程与应用, 2019, 55(17): 169–179. [doi: 10.3778/j.issn.1002-8331.1902-0223]

12 Akçapınar G, Altun A, Aşkar P. Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education, 2019, 16(1): 40. [doi: 10.1186/s41239-019-0172-z]

13 刘博鹏, 樊铁成, 杨红. 基于数据挖掘技术的学生成绩预警应用研究. 四川大学学报 (自然科学版), 2019, 56(2): 267–272.

14 罗杨洋, 韩锡斌. 基于增量学习算法的混合课程学生成绩预测模型研究. 电化教育研究, 2021, 42(7): 83–90. [doi: 10.13811/j.cnki.eer.2021.07.012]

15 Huang SH, Wei JJ. Student performance prediction in mathematics course based on the random forest and simulated annealing. Scientific Programming, 2022, 2022: 9340434.

16 Ramaswami G, Susnjak T, Mathrani A. On developing generic models for predicting student outcomes in educational data mining. Big Data and Cognitive Computing, 2022, 6(1): 6. [doi: 10.3390/bdcc6010006]

17 Asselman A, Khaldi M, Aammou S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environments, 2021. [doi: 10.1080/10494820.2021.1928235]

18 Yan LJ, Liu YS. An ensemble prediction model for potential student recommendation using machine Learning. Symmetry, 2020, 12(5): 728. [doi: 10.3390/sym12050728]

19 Adejo OW, Connolly T. Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 2018, 10(1): 61–75. [doi: 10.1108/JARHE-09-2017-0113]

20 Pan F, Yuan YC, Song YF. Students' classification model based on stacking algorithm. Journal of Physics: Conference Series, 2020, 1486(3): 032020. [doi: 10.1088/1742-6596/1486/3/032020]

21 Liu XY, Wang ST, Zhang ML. Transfer synthetic over-sampling for class-imbalance learning with limited minority class data. Frontiers of Computer Science, 2019, 13(5): 996–1009. [doi: 10.1007/s11704-018-7182-1]

22 张麒增, 戴翰波. 基于数据预处理技术的学生成绩预测模型研究. 湖北大学学报 (自然科学版), 2019, 41(1): 101–108.

23 Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888, 2018.

(校对责编: 孙君艳)