基于决策树算法的学生学习行为数据分析研究

胡玉兰 胡宁玉 覃晓玲

忻州师范学院计算机系，山西忻州，034000；

摘要：针对线上教学中难以全面掌握学生学习行为及其成绩预测的问题，文章提出了一种基于决策树算法的学习行为数据分析模型，文章首先概述了决策树算法的基本原理、类型及其在数据分析中的应用；其次对学生学习行为数据进行收集与预处理；最后基于决策树算法对学生学习行为数据进行分析，通过采用随机森林算法优化和调整关键参数，将模型预测准确率从 61.1%提升至 86.8%。研究结果表明，该模型能有效预测学生成绩所处层次，为教师开展针对性教学和学生及时调整学习策略提供数据支撑。

关键词：决策树；学生学习行为；随机森林

DOI：10.69979/3029-2735.25.09.050

引言

随着在线教育的快速发展，学生学习行为数据呈现爆发式增长，这些数据包含了大量有价值的教学信息，但传统的数据分析方法难以有效挖掘其中的规律，决策树作为一种直观的数据挖掘方法，在教育领域的应用逐渐受到关注，国内从 2001 年开始在教育数据挖掘方面进行研究，相关成果虽然相对较少，但呈上升趋势，在线教育平台可以记录学生的到课率与预习率，作业完成情况等多维度数据[1][2]。这些数据反映了学生的学习行为特征，利用决策树算法分析这些数据，不仅可以帮助教师掌握学生学习状况，预测学习效果，还能为学生提供学习策略调整的参考，通过对学习行为数据的挖掘分析，可以及时发现学习过程中的问题，为教师调整教学策略提供依据，对提升线上教学质量具有重要意义。

1 决策树算法概述

1.1 决策树算法基本原理

决策树算法是一种树形结构，组成部分主要包含根结点、叶结点和内部结点等。根节点表示整个训练集，决策节点和内部节点表示分类对象的属性，对决策树进行分枝表示为对属性进行选取值，状态结点、叶结点表示分类后的结果[3]。决策树有分类和回归两种模型，用分类的算法来预测离散的变量，用回归的算法预测连续的变量[4]。决策树是数据挖掘研究领域常用算法，采用决策树算法对在线教育研究主要集中于预测学习效果、研究影响学习因素方面[5]。

1.2 决策树算法类型

决策树算法主要且最常用的算法有迭代二分器 3（Iterative Dichotomiser 3，ID3）、C4.5、分类与回归树（Classification And Regression Tree，CART）等算法。ID3 算法是一种多叉树的结构，用来处理离散化的属性，需要独立的测试样本集，把最大信息熵增益作为最优化分的标准。期望信息越大，表示不确定度就越大或者越混乱。ID3 算法不能处理连续的特征值，也没有剪枝策略，容易过拟合，需要进一步改进[6]。熵的计算由公式（1）所示：

$$(X) = -\sum^{n} p(x_i)\log p(x_i) \quad (1)$$

X 表示某个分类，p(x)表示选择该分类概率。C4.5 算法依据 ID3 算法衍生改进，C4.5 算法的优点有：简单易理解，准确率高，处理噪声数据能力强等；属性选择依据是信息增益率,C4.5 算法具有连续属性“离散化”作用，提高了决策树模型有效性。其缺点为在生成树的过程中，算法时间复杂度较大，因为算法计算中含有对数运算[7]。信息增益率计算时需要考虑自身的熵，其计算方法由公式（2）所示：

$$Gini(D) = 1 - \sum_{i=1}^{n} p_k^2 \quad (2)$$

CART 算法是分类回归树,用其算法会生成一棵二叉树。因此，离散特征有 2 个以上可能取值的离散特征取值，不宜采用 CART 算法。树的最优划分用最小剩余方差决定，选择划分属性依据为 Gini 指数：Gini 指数越小，就越优先划分[8]。对数据 D 计算 Gini 指数方式如公式（3）所示：

$$\text{信息增益率} = \frac{\text{信息增益}}{\text{自身的熵}} \quad (3)$$

在公式中，n 表示类别数，Pi 表示数据样本属于类别的概率。通过分析决策树的特点发现，其优点是：（1）决策树具有很强的可解释性，体现在生成过程是可视透明的；（2）决策树预处理少，不需要对数据进行标准化和缩放等处理；（3）决策树能够处理离散和连续变量，分类和回归问题也可处理。其缺点是：（1）决策树在回归方面表现不佳，原因是数据存在太多变化；（2）决策树处理类型多、训练样例相对较少的分类问题中易错；（3）如数据未正确离散化，得出的结果可能会不准确。

2 学生学习行为数据的收集与预处理

2.1 数据收集

通过网上获取某高校的学生线上学习的数据。数据集包括到课率、预习正确率、预习率与成绩，共计 1852 条数据。通过网上获取某高校的学生线上学习的数据。数据集包括到课率、预习正确率、预习率与成绩，共计 1852 条数据。数据来源于该校 2023 年春季学期的在线教学平台，涵盖了计算机科学、数学、物理等多个专业的本科生课程。其中到课率通过学生登录在线课堂系统的记录计算得出，预习正确率基于学生在课前完成的预习测试题目的得分情况，预习率则反映了学生参与课前预习活动的频率。成绩数据包括期末考试成绩和平时成绩两个维度，采用百分制记录。这些数据的采集过程严格遵循教育数据采集规范，确保了数据的真实性和可靠性。

2.2 数据预处理

生成决策树之前，要对数据进行预处理，采用平均值填充方法进行处理。在这些数据集中，数据类型不统一，主要采用合成少数类过采样技术进行过采样处理，通过增加样本个数的方式使得数据类型均衡化。初步分析发现，原始数据中存在的主要问题包括：缺失值主要集中在预习正确率和预习率两个指标上，约占总数据量的 8%；数据分布不均衡体现在成绩分布上，其中优秀（85 分以上）占 15%，良好（70-85 分）占 45%，较差（70 分以下）占 40%。为提高模型的预测效果，采用 SMOTE 算法对少数类进行过采样，使三类样本数量达到相近水平。同时，对异常值采用箱线图方法进行检测和处理，确保数据质量。

3 基于决策树算法的学生学习行为数据分析

3.1 决策树算法的选择与优化

当直接选用经典的决策树算法进行分析时，发现预测正确率约为 61.1%，正确率较低。首先对数据进行检查，发现数据类别不平衡。于是，采用过采样方法使数据类型均衡化。调整好后运行代码发现正确率还未达到理想。查阅资料后决定采用随机森林算法提升正确率。主要的做法为调整随机森林的参数，主要参数有 n_estimators，max_depth，min_samplesplit。其中 n_estimators 主要是控制森林的树木数量，max_depth 表示树的最大深度，min_samplesplit 表示分割所需的最小样本数，对这三者进行取值调整，调整好之后再执行代码，发现预测正确率可以提升至 86.8%。经过反复实验，最终确定最优参数组合为：n_estimators=100，max_depth=8，min_samples_split=5。这组参数在保持模型复杂度适中的同时，有效避免了过拟合现象的出现，使模型具有良好的泛化能力。

3.2 结果分析与讨论

导入的 1852 条数据中存在部分缺失值，因很多算法不支持空值输入，需要在训练数据之前对其填充空值。采用平均值填充缺失值是通过最大概率的可能取值来填补缺失值。数据类别不均衡会影响是否能学习到辨别好坏本质的特征的模型，也会影响决策树模型的鲁棒性。普遍的处理方法主要有：扩大数据集，欠采样和过采样等方法。扩大数据集主要应用于小样本数据，且还能再增加数据的情况；由于数据样本个数较少，且数据样本是直接从网上获取，不易再增加样本数据，故采用过采样方法来解决数据类型不均衡问题。通过特征重要性分析发现，预习正确率是影响学生成绩最重要的因素，其次是到课率，最后是预习率。这一发现与教育实践经验相符，也为教师进行教学干预提供了重要依据。

在随机森林算法提升正确率中，通过查询资料和原因分析等操作，正确率得到提高。在提升预测正确率时，对随机森林相关的参数进行调整和分析，详细说明如下。

（1）对调整随机森林树木的数量进行分析，发现训练和测试的预测正确率相差较大，且预测正确率不会

随着数目的增加而增多。训练的预测正确率始终保持在80%以上，而测试的预测正确率在 65%以下左右徘徊。通过实验发现，最优的树木数量通常在 120-180 之间，超过这个范围后模型性能提升不明显。同时，我们还发现树木数量与预测时间呈线性关系，因此在实际应用中需要在性能和效率之间做出权衡。

（2）研究与分析决策树生成的最大深度，发现测试的预测正确率总体稳定在 50%~65%之间；训练的预测正确率快速上升，正确率高达 85%以上。但总体变化趋势呈现为：随着树的深度不断加深，训练和测试的预测正确率都不断增加。为了更深入地理解树深度对模型性能的影响，我们进行了学习曲线分析。结果显示，当树深度在 5-10 之间时，模型存在欠拟合现象；当深度超过 20 时，则开始出现明显的过拟合。通过记录每一层节点的分裂情况，我们发现大多数有效的特征分裂发生在前 12 层，这为确定最优树深度提供了理论依据。同时，我们还观察到不同特征在不同深度的重要性变化，这有助于理解模型的决策过程。

（3）对分割最小样本数进行研究与分析，最小样本数取值为 2~14 之间，公差为 2。训练的预测正确率随着最小样本数的扩大而减低，呈直线下降趋势。不过在 2~14 取值之间，其预测正确率保持在 90%~75%之间，而测试预测正确率始终在 65%左右徘徊。训练与测试的预测正确率在最小样本数取值中存在较大差距。为了优化这一参数，我们采用了自适应调整策略，根据节点样本的方差来动态调整最小样本数。具体而言，当节点样本方差较大时，增加最小样本数以获得更稳定的分裂；当方差较小时，则允许更细粒度的分裂。这种方法显著提高了模型在不同数据分布情况下的适应性。

4 结束语

本文主要是采用某高校学生线上学习行为数据对所构建的决策树模型进行训练和测试，并通过随机森林算法提升其模型的准确率。但由于学习行为数据偏少，所构建的模型目前主要考虑到课率、习题正确率、预习率 3 个方面，为了完善学习模型的实用性，在后续的研究中，会进一步增加相关的学习行为特征进行分析。

参考文献

[1]杨小娟. 决策树算法在学生课程成绩分析中的应用研究[D]. 昆明：云南师范大学，2021.
[2]普运伟，姜萤，田春瑾，等. 基于 MLP-Bagging 集成分类模型的在线学习行为分析[J]. 云南大学学报(自然科学版)，2024，46(05)：852-861.
[3]胡明明. 决策树算法在学生课程成绩分析中的应用研究[D]. 哈尔滨：哈尔滨师范大学，2019.
[4]罗明挽. 基于决策树算法的学生体质预警系统开发与实现[J]. 电脑编程技巧与维护，2023(02)：12-14.
[5]黎龙珍. 基于决策树算法的在线学习成绩预测[J]. 信息技术与信息化，2021(01)：130-133.
[6]张海燕，刘岩，马丽萌，苑津莎，巨汉基，魏彤珈.决策树算法的比较与应用研究[J].华北电力技术，2017，(06)：42-47.
[7]韩存鸽，叶球孙. 决策树分类算法中 C4.5 算法的研究与改进[J]. 计算机系统应用，2019，28(06)：198-202.
[8]刘学军. 基于最小 Gini 指标的决策树分类算法设计与研究[J]. 软件导刊，2009，8(05)：56-57.
[9]董师师，黄哲学. 随机森林理论浅析[J]. 集成技术，2013，2(01)：1-7.

作者简介：胡玉兰（1985—），女，山西五台，硕士，副教授，研究方向：智能信息处理。
项目来源：山西省高等学校教学改革创新项目（J20241260）；山西省高等学校教学改革创新项目（J20231133）。