引文格式：扈廷锐,高怡凡,陈李,等. 基于决策树的GIS空间分析课程考核分析[J]. 地理空间信息, 2024, 22(8): 129-132. doi:10. 3969/j.issn. 1672-4623.2024.08.029

基于决策树的GIS空间分析课程考核分析

扈廷锐[1]，高怡凡[1]，陈　李[1]，高培超[1*]
（1. 北京师范大学地理科学学部地理数据与应用分析中心，北京 100875）

摘　要：随着教育部发布《普通高等学校本科专业类教学质量国家标准》，科学有效地考核分析对于提高教学质量，并进行教学改革变得尤为重要。常见的课程考核分析方法中，往往无法分析得出各个总评模块对学生成绩的具体影响作用，难以对教学改革产生直接的建议。使用Classifier 4.5算法和Error-Based Pruning 策略对《GIS空间分析》课程的学生成绩进行决策树构建，并根据总体精度和Kappa系数对决策树进行准确性评估。根据决策树生成的分类规则并结合GIS空间分析的课程特性，分析得出学生课程学习中的薄弱环节。建议GIS空间分析课程引导学生更加重视在实践中综合运行理论知识，从而更好地达到课程教学目标。

关键词：教育改革；考核分析；决策树；C4.5算法；EBP策略
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2024）08-0129-04

Assessment Analysis of GIS Spatial Analysis Course Based on Decision Tree

HU Tingrui[1], GAO Yifan[1], CHEN Li[1], GAO Peichao[1]
(1. Center for GeoData and Analysis, Faculty of Geographic Sciences, Beijing Normal University, Beijing 100875, China)

Abstract: With the release of the National Standards for the Quality of Undergraduate Professional Teaching in Ordinary Higher Education by the Ministry of Education, scientific and practical assessment and analysis have become particularly important for improving teaching quality and carrying out teaching reforms. In common course assessment analysis methods, it is often difficult to analyze the specific impact of each overall evaluation module on students' grades, making it difficult to provide direct suggestions for teaching reform. We used the Classifier 4.5 algorithm and Error-based Pruning strategy to construct a decision tree for students' scores in GIS Spatial Analysis, and evaluated the accuracy of decision tree according to the overall accuracy and Kappa coefficient. According to the classification rules generated by the decision tree and GIS Spatial Analysis course characteristics, we analyzed the weak links in students' curriculum learning. The GIS Spatial Analysis course should guide students to pay more attention to the comprehensive operation of theoretical knowledge in practice to achieve the teaching objectives of the study better.

Key words: educational reform, assessment analysis, decision tree, C4.5 algorithm, EBP strategy

随着教育部发布《普通高等学校本科专业类教学质量国家标准》对本科课程的教学质量提出了统一的标准，提高本科教学质量以达到国家标准已是势在必行。其中，课程考核分析已经成为提高教学质量的重要一环，通过科学且有效的考核分析对提高教学质量变得尤为重要[1-3]。

现有的课程考核分析主要是通过试卷成绩的加权分析、K-Means聚类分析、直方图分析、层次分析法、正态分布分析法等方法进行考核分析，并取得了不错的考核分析效果。陈晖[4]等通过加权分析的方法，计算不同考核指标的分值，分别制定出评价银行不同职员的评分标准；李爱军[5]等分别使用t检验和K-Means分析的方法对学科进行统计，得出重点学科的建设成果优于非重点学科的建设成果的结论；艾明耀[6]等通过直方图分析班级学生的成绩呈现正态分布，并对各考核项目进行失分分析，得出考核模式符合预期的结论；黄良珂[7]等通过层次分析法分析测绘类教师的评价指标，从而得出，在“双一流”建设的背景下测绘类教师的绩效评价指标体系；曹强[8]等通过对试卷成绩进行难度、可信度、正态分布等分析，得出该考试较为全面地反映学生对于实验的掌握情况，同时考试时间限制提升了考试质量的结论；李爱民[9]等提出要完善教学方法，综合考虑团队表现、专业知识、等多方面指标，对学生的综合素质进行全面

收稿日期：2023-04-21。
项目来源：国家自然科学基金资助项目（41901316）；北京师范大学2022年青年教师教学发展基金资助项目（2022117）。
第一作者简介：扈廷锐（1999—），硕士研究生，研究方向为高光谱波段选择算法、土地变化模拟，E-mail：hutingrui@mail.bnu.edu.cn。
通信作者：高培超（1991—），博士，副教授，硕士生导师，主要从事土地变化模拟研究工作，E-mail：gaopc@ bnu.edu.cn。

的评价。

本研究旨在使用决策树算法对GIS空间分析课程的学生成绩进行分析，以期制定提高课程教学质量的针对性措施。

1 决策树算法

1.1 决策树算法介绍及选择

决策树算法是一种根据分类属性以树状结构进行分类的算法，主要用于数据挖掘、故障诊断等领域[10-11]。决策树算法具有以下的优点，首先，决策树算法可以快速对影响因素根据重要程度进行排序，利于洞察出最主要的影响因素。其次，决策树可以快速高效地实现构建，无需长时间的构建分析。

决策树算法目前主要有Iterative Dichotomiser 3算法（ID3）、Classifier 4.5算法（C4.5）和Classification And Regression Tree算法（CART）3种。其中，C4.5算法是ID3算法的改进版。C4.5算法相较于ID3算法消除了偏向选择分类属性取值较多的属性进行分类的缺点，增加了连续数据值和缺失值处理的方法[12-13]。同时，C4.5算法相较于CART算法可以处理分类属性具有多个取值的情况[14-15]。所以本文选择C4.5算法作为实验算法。

1.2 C4.5算法

C4.5算法由澳大利亚科学家Quinlan[16]于1993年提出。C4.5算法根据信息增益率最大原则选择分类属性，具体的实现步骤如下所示。假设有数据集 D 具有 n 个分类特征和1个目标特征。

步骤一，计算当前数据集 D 的香农熵。具体计算香农熵的公式如（1）所示。

$$\mathrm{Ent}(D)=-\sum_{i=1}^{k}p_i\log_2 p_i \tag{1}$$

式中，$\mathrm{Ent}(D)$ 为数据 D 中目标属性的信息熵；k 为数据 D 中目标属性的最大类别数；p_i 为数据 D 中目标属性第 i 类占数据 D 中所有类别的比例。

步骤二，计算分类属性 a 的信息增益。具体计算分类属性的信息增益公式如（2）所示。

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{i=1}^{k}\frac{|D^i|}{|D|}\mathrm{Ent}(D^i) \tag{2}$$

式中，$\mathrm{Gain}(D,a)$ 为数据 D 中分类属性 a 的信息增益；$|D|$ 为数据 D 中分类属性 a 的所有样本数。

步骤三，计算分类属性 a 的分裂信息量。具体计算分类属性 a 的分裂信息量的公式如（3）所示。

$$\mathrm{SplitInfo}(D,a)=\sum_{i=1}^{k}\frac{|D^i|}{|D|}\log_2\frac{|D^i|}{|D|} \tag{3}$$

式中，$\mathrm{SplitInfo}(D,a)$ 为数据 D 中分类属性 a 的分裂信息量。

步骤四，计算分类属性 a 的信息增益率，通过信息增益率最大获得区分数据集的最佳分类属性。具体计算分类属性 a 的信息增益率的公式如（4）所示。

$$\mathrm{GainRatio}(D,a)=\frac{\mathrm{Gain}(D,a)}{\mathrm{SplitInfo}(D,a)} \tag{4}$$

式中，$\mathrm{GainRatio}(D,a)$ 为数据 D 中分类属性 a 的信息增益率。

步骤五，重复步骤一至步骤四，直到分类后的子数据集的香农熵为零或者没有分类属性为止。

1.3 剪枝策略及选择

过拟合情况的发生会导致模型只适用于当前数据集，更换其他数据集则会降低模型分类精度。为避免发生过拟合，一般需要对决策树进行剪枝处理，提高模型分泛化能力。

剪枝分成预剪枝（prepruning）和后剪枝（post-pruning）2种，由于在一般情况下后剪枝的泛化能力要优于预剪枝，所以本文选择后剪枝作为实验的剪枝方法。在后剪枝方法主要分成错误率降低剪枝（Reduced-Error Pruning, REP）、悲观错误剪枝（Pesimistic-Error Pruning, PEP）、代价复杂度剪枝（Cost-Complexity Pruning, CCP）和基于错误的剪枝（Error-Based Pruning, EBP）4种剪枝策略。其中，EBP策略是PEP策略的改进版，具有不需要独立剪枝集[17]和减缓过拟合[18]的特点，所以本实验选择EBP策略进行剪枝操作。

1.4 评估指标

采用总体精度（overall accuracy，OA）和Kappa系数2种评估指标对样本分类后的精度进行评估。OA能够判断总体分类结果的准确性。Kappa系数能够改善OA忽略样本数量较小的类别的分类效果的缺点。二者结合使用可以得到更加可靠的分类效果评估结果。

2 决策树的构建

2.1 数据获取与预处理

数据从本校2021—2022学年第二学期开设的GIS空间分析课程中获取。该课程共有55位学生参加，每位学生的各项成绩全部真实有效，不存在数据缺失等问题，同时去除姓名、学号、性别等无关个人信息。

课程总分由出勤、实验作业、期末汇报和期末成绩四部分分别加权组成，其中出勤加权10%，实验作业加权20%，期末汇报加权30%，期末成绩加权40%，如表1所示。

表1　GIS空间分析课程学生成绩

序号	出勤	实验作业	期末汇报	期末成绩	课程总分
1	100.00	89.60	80.50	64.00	77.67
2	80.00	90.70	77.00	69.00	76.84
…	…	…	…	…	…
54	100.00	92.90	67.50	61.00	73.23
55	70.00	7.00	0.00	60.00	32.40

由于各项成绩都为连续型数据，在划分分类属性时存在不便，所以需事先将出勤、实验作业、期末汇报和期末成绩进行数据离散化预处理，即分别进行人为划分等级。其中，出勤成绩90~100分为优秀，80~90分为良好，60~80分为及格，60分以下为不及格；实验作业和期末汇报90~100分为优秀，75~90分为良好，60~75分为及格，60分为不及格；期末成绩85~100分为优秀，75~85分为良好，60~75分为及格，60分为不及格；课程总分90~100分为优秀，80~90分为良好，60~80分为及格，60分为不及格。按照此评价标准，制作成离散化预处理后的学生成绩如表2所示。

表2　GIS预处理后学生成绩

序号	出勤	实验作业	期末汇报	期末成绩	课程总分
1	优秀	良好	良好	及格	及格
2	良好	优秀	良好	及格	及格
…	…	…	…	…	…
54	优秀	优秀	及格	及格	及格
55	良好	不及格	不及格	及格	不及格

为了测试模型的准确性与可靠性，需要保留一定数量的数据作为测试数据。在小规模的数据集下，训练数据与测试数据的比例一般为7∶3[19]，所以将全部55条数据随机分成训练数据39条和测试数据16条，进行决策树的构建与检验。

2.2　决策树构建

C4.5决策树可以通过现有Weka软件进行实现，在Weka软件中C4.5算法被称为J48算法[20]。在Weka软件中输入训练数据，分类方式选择树状分类中的J48算法，并选择EBP策略构建决策树，其中显著性水平α设置为0.5。在构建剪枝决策树的同时生成混淆矩阵，以计算总体精度和Kappa系数。通过总体精度和Kappa系数可以评估决策树模型的准确性和可靠性。

2.3　决策树模型评价

Weka软件具有实现决策树模型检验的功能。在Weka中输入测试样本，进行测试样本的分类。测试样本根据生成的规律自动进行分类操作生成混淆矩阵，以计算总体精度和Kappa系数，从而评估决策树模型的准确性和可靠性。

3　评价与分析

3.1　决策树分类效果评价

通过Weka软件进行C4.5决策树的构建与剪枝操作，剪枝完成后的决策树分类图如图1所示。其中，圆圈内表示该节点的分类属性，方框内表示该叶子节点的类别，方框内括号左边表示该叶子节点全部样本的数目，方框内括号右边表示该叶子节点错分样本的数目。训练数据集的混淆矩阵如表3所示。其中，混淆矩阵的纵轴为真实类别，横轴为分类类别。通过计算训练数据构造的决策树混淆矩阵，可以得出训练数据集的总分类精度达到90%，Kappa系数达到0.81，证明了通过C4.5算法和EBP策略构造的决策树的精度较好，同时构建决策树的模型可靠。

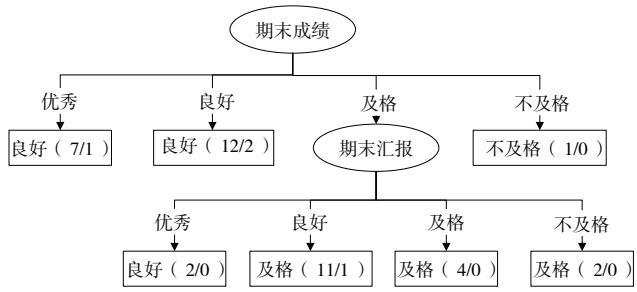


图1　决策树分类图

表3　训练数据集混淆矩阵

	优秀	良好	及格	不及格
优秀	0	1	0	0
良好	0	18	1	0
及格	0	2	16	0
不及格	0	0	0	1

3.2　决策树模型评价

为了评价决策树构建的有效性与合理性，需要对构造好的决策树进行检验。选择测试数据对已建好的模型进行检验，得到测试数据的混淆矩阵，如表4所示。其中，混淆矩阵的纵轴为真实类别，横轴为分类类别，测试数据集的总分类精度达到88%，Kappa系数达到0.77。该结果证明了已建好决策树模型的有效性与合理性。

3.3　课程考核分析

根据构造完成C4.5决策树的分类图，可以总结出从根节点到达叶子节点的目标属性分类规则，生成以下7条不同的规则如表5所示。

表4 测试数据集混淆矩阵

	优秀	良好	及格	不及格
优秀	0	1	0	0
良好	0	8	0	0
及格	0	0	6	0
不及格	0	0	1	0

表5 决策树分类规则

规则名称	分类条件	分类结果
规则1	期末成绩=优秀	课程总分=良好
规则2	期末成绩=良好	课程总分=良好
规则3	期末成绩=及格，同时期末汇报=优秀	课程总分=良好
规则4	期末成绩=及格，同时期末汇报=良好	课程总分=及格
规则5	期末成绩=及格，同时期末汇报=及格	课程总分=及格
规则6	期末成绩=及格，同时期末汇报=不及格	课程总分=及格
规则7	期末成绩=不及格	课程总分=不及格

根据生成的决策树分类图，可以得出决策树选择期末成绩作为决策树根节点时，期末成绩对课程总分影响最大的结论。这和期末成绩在课程总分中占比最大这一事实相符。因此，学生在学习过程中应加强注重对知识点的掌握，同时教师教学过程中应对大纲知识点全覆盖地清晰讲解。

其次，可以得出学生在期末汇报中表现差异大，因此该部分对个人的课程总分影响较大的结论。根据规则3~6，可以看出除了期末汇报为优秀时，课程总分为良好。当期末汇报为良好、及格和不及格时，课程总分都为及格。说明期末汇报汇报是除期末成绩外对课程总分影响最大的因素。其中，1/3的学生并没有取得优秀的成绩。

最后，出勤和实验作业并没有作为分类属性出现在决策树中，这也和出勤与实验作业在课程总分占比最小这一事实相符。通过研究实验数据发现，大部分学生达到了全勤并且实验作业达到了良好以上。在之后的课程中，可以保持现有出勤考核力度。同时，GIS空间分析是一门理论与实践相结合的课程。教师可以考虑在课堂上讲授知识点后进行随堂测验，以加强学生对知识点的掌握。

4 结 论

本文选择C4.5算法对GIS空间分析课程的学生成绩进行决策树构建，并通过测试数据生成混淆矩阵计算测试数据的OA和Kappa系数评估决策树的可靠性。根据构建完成的决策树分类图生成分类规则，从而分析分类规则得出出勤、实验作业、期末汇报和期末成绩对学生成绩的具体影响作用，提出了相关的教学建议，从而更好地达到课程目的与提高教学质量。本文方法可适用于不同课程，具有一定的普适性。

参考文献

[1] 李二俊,蔡庆空.“新工科”背景下《工程测量》课程教学改革与实践[J].地理空间信息,2023,21(3):148-151

[2] 郑春燕,胡华科,张鹏飞.校企协同GIS专业实践教学改革探索[J].地理空间信息,2023,21(4):137-140

[3] 秦永,韩留生,黄洁慧,等.《遥感概论》线上线下混合式教学改革与实践[J].地理空间信息,2023,21(2):161-164

[4] 陈晖,许学军.商业银行柜员绩效考核分析[J].现代商贸工业,2009,21(6):171-172

[5] 李爱军,龚树生,吕璨,等.高等医学院校附属医院临床学科考核评价研究[J].中国医院管理,2012,32(8):38-39

[6] 艾明耀,潘励,张丰,等.卓越工程师能力考核探讨与分析:以“GIS原理课程设计”为例[J].测绘通报,2014(11):123-126

[7] 黄良珂,黄军胜,王浩宇,等.“双一流”建设背景下测绘类教师专业发展绩效评价体系与模型构建[J].测绘通报,2018(7):141-145

[8] 曹强,叶俊,黄华兴,等.医学本科生微创腔镜操作考核成绩及试卷分析[J].医学教育研究与实践,2021,29(6):853-859

[9] 李爱民,陈琛,王金鑫.工科课程测量学思政教学探索[J].地理信息世界,2022,29(6):114-115,119

[10] 罗可,林睦纲,郁东妹.数据挖掘中分类算法综述[J].计算机工程,2005,30(1):3-5

[11] 杨静,张楠男,李建,等.决策树算法的研究与应用[J].计算机技术与发展,2010,20(2):114-116,120

[12] 苗红星,余建坤.基于决策树的ID3算法和C4.5算法的比较[J].现代计算机(专业版),2014,30(15):7-10,14

[13] 王黎黎,刘学军.决策树C4.5算法在成绩分析中的应用[J].河南工程学院学报(自然科学版),2014,26(4):69-73

[14] 庞素琳,巩吉璋.C5.0分类算法及在银行个人信用评级中的应用[J].系统工程理论与实践,2009,29(12):94-104

[15] 陈杰,邬春学.决策树C4.5算法改进与应用[J].软件导刊,2018,17(10):88-92

[16] Quinlan J R. C4.5:Programs for Machine Learning[J]. The Morgan Kaufmann Series in Machine Learning,1993,32(16):235-240

[17] 周美琴,陈诗旭,袁鼎荣,等.一种单位代价收益决策树剪枝算法[J].计算机工程与科学,2016,38(5):1 023-1 030

[18] 宋宇,李鸿禧.基于决策树集成的债券发行企业财务预警研究[J].财会月刊,2020,40(6):45-50

[19] 周志华.机器学习[M].北京:清华大学出版社,2016

[20] 孙宇翼,赵军利,王苗苗,等.基于J48决策树的面向对象方法的土地覆被信息提取[J].国土资源遥感,2016,28(4):156-163