Performance early warning model based on principal component analysis and multiple logistic regression

Shenghua Zhang

Yangzhou Polytechnic College, Yangzhou, Jiangsu, 225000, China

Abstract

In view of the characteristics of higher vocational college students' relatively weak learning foundation and their curriculum performance being greatly affected by non intellectual factors, this paper proposes a curriculum performance grade early warning model based on principal component analysis and logical regression model. The model collects students' non intelligence factor data such as learning attitude, learning methods, learning efficiency and personality psychology through questionnaires, extracts four principal components by principal component analysis, and then constructs a multiple logistic regression model with the grade of achievement as the dependent variable. An empirical study was carried out among 2024 freshmen in a vocational college. The results show that the model has a high accuracy in identifying grades, which can effectively predict students' final exam grades and provide teachers with early warning and intervention suggestions. In addition, the research also discusses the expansion and application prospects of the model, proposes that more objective measures can be added to improve the recognition accuracy, and develops an intelligent early warning system.

Keywords

Principal component analysis; Multiple logistic regression; Achievement; Early warning model

基于主成分分析和多元 logistic 回归的成绩预警模型

张生华

扬州市职业大学, 中国·江苏扬州 225000

摘要

针对高职院校学生学习基础相对薄弱、课程成绩受非智力因素影响较大的特点, 本文提出了基于主成分分析法和逻辑回归模型的课程成绩等级预警模型。该模型通过调查问卷收集学生的学习态度、学习方法、学习效率和个性心理等非智力因素数据, 利用主成分分析提取四个主成分, 再以成绩等级为因变量构建多元logistic回归模型。在某高职院校2024级新生中进行了实证研究, 结果显示模型识别成绩等级的正确率较高, 能够有效预测学生期末考试成绩等级, 为教师提供早期预警和干预建议。此外, 研究还探讨了模型的扩展与应用前景, 提出可以加入更多客观性度量以提高识别正确率, 并开发智能化预警系统。

关键词

主成分分析; 多元logistic回归; 成绩; 预警模型

1 引言

高等院校教学管理的重心在于对学生课程成绩的管理, 课程成绩不仅反映了学生个人的学习能力和素养, 也是学校教学水平和教学质量的重要体现。大数据时代的到来给高等教育领域带来了巨大的机遇和挑战, 学生的公共基础课、专业必修课以及考试课、考查课、选修课等各种课程产生海量

的成绩数据, 如何充分利用这些数据, 从这些数据中提取出有用的信息以促进教学水平, 提升教学质量, 提高课程考核的通过率, 帮助教师调整教学计划, 优化教学设计, 是大数据时代广大教育工作者必须重视的研究课题^[1]。

对于高等教育学习预警系统, 国内学者研究重点大多聚焦于学业预警, 即对即将毕业的大学生是否修完足够的学分进行预警^[1-4]。而对于学生课程成绩是否及格的预警研究则比较少, 大多使用数据挖掘中的关联规则算法探讨相关课程成绩合格与否的关联性^[5-6]。因此, 如何构建一个既能全面预测学生学习分化又能提供可行改进建议的预警系统仍是一个亟待解决的问题。高职院校的学生学习基础相对薄弱, 课程成绩很大程度上与非智力因素有关, 本研究以主成分分析法和逻辑回归模型为基础, 建立课程成绩等级的预警模型, 根据学生的学习态度、学习方法、学习效率和个性心

【课题项目】扬州市职业大学高教研究课题: 基于机器学习的数字化教学评价机制创新研究的成果(项目编号: 2023GJ07)。

【作者简介】张生华(1977-), 男, 硕士, 从事数据挖掘研究。

理这些非智力因素去预测期末考试成绩等级，对那些处在不及格边缘的学生在学期初尽早进行学习上的督促和帮扶，以便他们顺利通过课程考核。

2 模型设计

本文重点探讨课程期末成绩的等级和学生的学习态度、学习方法等非智力因素的多元 logistic 回归关系，建立课程期末成绩预警模型。

对入学新生发放并回收了调查问卷和入学测试试卷，调查问卷配有 14 个问题，记为 $a_1, a_2, \dots, a_{14}$ ，其中 a_1-a_4 测试学生的学习态度， a_5-a_8 测试学生的学习方法， a_9-a_{11} 测试学生的学习风格和学习效率， $a_{12}-a_{14}$ 测试学生的个性和心理。每一个问题都有 4 个等级，记为 1、2、3、4，等级越高说明该测试项正向得分越高。在考察学生课程成绩这一指标时，生源的多样化导致高考成绩这个变量存在较多的缺失值，极大地影响模型的精确性和有效性，因而将学生的入学测试成绩作为评判指标，记为 a_{15} ，与 $a_1, a_2, \dots, a_{14}$ 一起作为输入变量，使用聚类分析的方法从高到低将学生划分为 A、B、C 三个层次，在教学过程中对各个层次的学生实施不同层次的教学计划，讲授不同层次的教学内容，考核评价采取对应的难度和要求，并记录各层学生的课程期末考试成绩，为了研究的方便，将期末考试成绩划分为 3 个等级：60 分以下为等级 1，60 分（含）到 79 分为等级 2，80 分（含）以上为等级 3，记录为新的变量 grade 并输入数据集。

调查问卷数据集中的输入变量太多，并且具有相关性，直接使用 logistic 回归分析容易造成数据的过拟合，而且对结果的解释过于复杂，实用性差，因此考虑先对数据集进行主成分分析。主成分分析是考察多个变量间相关性的一种多元统计方法，研究如何通过少数几个主成分来揭示多个变量间的内部结构，即从原始变量中导出少数几个主成分，使它们尽可能多地保留原始变量的信息，且彼此间互不相关。利用主成分分析从数据集中提取出对应的四个主成分作为输入变量，以成绩等级作为输出变量构建 logistic 回归模型。

3 模型应用

在某高职院校 2024 级入学新生中随机发放、回收了 400 份调查问卷，采集了他们的数学入学测试成绩数据，使用聚类分析分为 A、B、C 层，分层教学后加入高等数学期末成绩等级数据，对数据集进行信度和效度分析（过程略），

顺利通过检验。

利用 SPSS 软件对输入变量 $a_1, a_2, \dots, a_{14}$ 进行 KMO 和巴特利特检验，从表 1 可以看出，KMO 总体检验系数为 0.865 大于 0.8，巴特利特检验 p 值小于 0.01，说明该数据集结构合理，具有相关关系，适合进行主成分分析^{[7][8]}。

表 1 KMO 和巴特利特检验表

KMO 取样适切性量数		0.865
巴特利特球形度检验	近似卡方	3464.925
	自由度	91
	显著性	.000

对 $a_1, a_2, \dots, a_{14}$ 主成分分析，根据初始特征值大于 1 的条件提取出四个主成分：第一主成分反映的是学生的学习态度，第二主成分反映的是学习方法，第三主成分反映的是学习风格和效率，第四主成分反映的是个性心理，具体来说，第一主成分解释了总方差的 41.84%，主要涵盖了学生学习态度的相关事项，如学习自觉性、作业完成情况等；第二主成分解释了 13.66% 的方差，主要涉及学习方法的合理性，比如是否有系统的学习计划、复习策略等；第三主成分解释了 11.01% 的方差，反映了学习风格和学习效率，包括学习时间管理、注意力集中程度等；第四主成分解释了 10.30% 的方差，主要涉及学生的个性和心理特征，如抗压能力、自信心等。这四个主成分能够解释 76.80% 的数据变异，见表 2。

使用数据挖掘软件 Spss Modeler 依次对 A、B、C 层利用多元 logistic 回归建立成绩等级和四个主成分：态度、方法、效率以及心理的回归方程，采用进入法，以等级 1 作为基准类别，分别得到各层的回归方程以及变量重要性，见表 3。

在模型调试过程中，先后尝试了逐步回归法（向前选择、向后剔除）来优化模型，以避免多重共线性和过拟合问题。逐步回归法可以根据统计显著性自动选择变量，从而提高模型的稳定性和预测准确性。模型结果均显示：对于 A 层、B 层和 C 层，四个主成分对成绩等级影响的重要性排序均为：态度、方法、效率和心理；无论哪一层，对成绩影响最重要的都是学习态度；与 B、C 层相比，A 层学生的学习方法也占了较大比重。由此可见，只要高职院校的学生具有良好、端正的学习态度，再加上正确有效的学习方法，高等数学的成绩等级就不会差。

表 2 总方差解释表

成分	初始特征值			提取载荷平方和			旋转载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	5.857	41.836	41.836	5.857	41.836	41.836	3.075	21.962	21.962
2	1.912	13.657	55.493	1.912	13.657	55.493	2.900	20.712	42.674
3	1.541	11.008	66.501	1.541	11.008	66.501	2.527	18.053	60.727
4	1.442	10.299	76.800	1.442	10.299	76.800	2.250	16.073	76.800

提取方法：主成分分析法

表 3 A、B、C 层多元 Logistic 回归方程

	A	B	C
方程式用于 等级 1	基准类别	基准类别	基准类别
方程式用于 等级 2	74.21* 态度 +49.84* 方法 +42.91* 效率 +9.21* 心理 +48.60	14.31* 态度 +0.32* 方法 +7.58* 效率 +2.99* 心理 +21.45	36.59* 态度 +6.83* 方法 +5.40* 效率 +1.97* 心理 +78.00
方程式用于 等级 3	84.17* 态度 +52.10* 方法 +45.47* 效率 +11.24* 心理 +33.33	18.53* 态度 +2.34* 方法 +10.17* 效率 +4.40* 心理 +13.64	79.94* 态度 +8.21* 方法 +6.27* 效率 +4.84* 心理 +62.24

4 模型检验

将训练得到的 A、B、C 三个层次的多元 logistic 回归模型对相应层次的检验集进行检验，各个层次的检验结果如表 4 所示，可以看出：三个层次的模型识别成绩等级的正确率分别为：95.74%、98.21% 和 98.80%，对于 A 层，多元 logistic 回归模型识别成绩等级 1（不及格）的正确率为 100%，而 B 层和 C 层的回归模型都将一个等级 1 识别成了等级 2，因而识别等级 1 的正确率分别为 $\frac{25}{26} \times 100\% = 96.15\%$ 和 $\frac{12}{13} \times 100\% = 92.31\%$ 。总体上，使用多元 logistic 回归模型识别等级 1 的正确率为 $\frac{41}{43} \times 100\% = 95.35\%$ 。例如：C 层某个学生入学调查问卷 $a_1、a_2、\dots、a_{14}$ 的测试项得分为：2、1、2、1、1、2、1、1、1、1、2、3、3，利用 C 层的多元 logistic 回归预警模型识别出该生高等数学期末成绩为等级 1（不及格）的概率为 0.998，该生在学习态度、学习方法这两个主成分上得分较低，可能是学习态度不够端正或学习方法不当，针对这些问题，教师、班主任、辅导员可以提供具体的干预措施，如安排学习小组、提供个性化学习计划，端正该生的学习态度，寻找对其行之有效的学习方法，提高学习效率，帮助他顺利通过课程考核；A 层某个学生入学调查问卷的测试项得分为：4、3、4、3、3、3、3、3、4、4、3、3、3、3，利用 A 层的预警模型识别该生高等数学期末成绩为等级 3（优秀）的概率为 0.979，因此在教学过程中可以对该生重点关注，作为高数竞赛、数学建模竞赛的后备人员进行重点培养。

表 4 A、B、C 层模型检验分析表

层级	等级 1 正确率	等级 2 正确率	等级 3 正确率	总正确率
A	100.00%	96.82%	92.59%	95.74%
B	96.15%	99.36%	95.00%	98.21%
C	92.31%	100.00%	100.00%	98.80%

在预警模型的使用过程中，根据预警信息，组织学校、家长、教师、学生通力合作，对教学过程进行适当的监控和干预，积极调动受预警学生的学习动力和学习潜能，注重启发、引导、鼓励、表扬和促进，有条件的建立学习管理档案，密切关注他们的学习状态和学习动态，实时记录评价和反馈。

5 模型扩展与应用前景

在多元 logistic 回归模型中再加入章节测试成绩、平时表现打分、在线课程测试成绩等客观性度量，可以进一步提高识别的正确率。这些客观性度量可以通过教学管理系统自动获取，提高数据处理的效率和准确性。基于本文的预警模型，可以开发智能化的成绩预警系统，系统集成数据收集、处理、分析、预警及干预建议等功能。系统可以实时监控学生的学习状态，一旦发现学生有不及格的风险，立即向教师、辅导员及家长发送预警信息，形成家校共育的良好机制。

6 结语

该预警模型综合运用了主成分分析和多元 logistic 回归分析，在分层的 400 名学生中识别效果总体较好，可以有效帮助教师在教学过程中实施因材施教，对“问题学生”重点关注。在模型中再加入一定的客观性度量如章节测试成绩、平时表现打分等以进一步提高识别的正确率，进一步探索模型的鲁棒性、融入更多客观性度量、开发智能化预警系统，以及将模型应用于更广泛的教育领域，为教育质量的提升做出更大贡献，这将是后续的研究工作。

参考文献

- [1] 陈玉芸. 大数据环境下完善学业预警系统研究[J]. 南宁职业技术学院学报, 2018(23):53-56.
- [2] 池振国, 崔灏, 孙宁. 浅析实施学业预警机制对高校学风建设的影响[J]. 天津市教科院学报, 2013(32):23-24.
- [3] 焦春燕. 基于教育大数据的学业预警模型研究[J]. 电子技术, 2017(9):11-13.
- [4] 郑俊玲. 基于KPCA的大学生学业预警模型及其应用[D]. 唐山: 华北理工大学, 2015.
- [5] 张国华. 数据挖掘在独立学院课程预警中的应用与实践[J]. 现代电子技术, 2016(17):136-139.
- [6] 赵子云, 陈明选. 信息化环境下学生成绩预警系统的设计研究[J]. 中国信息技术教育, 2015(23):119-122.
- [7] 张生华. 基于主成分分析和聚类分析的高职高等数学分层教学设计[J]. 扬州职业大学学报, 2019(04):49-52.
- [8] 张生华. 在线课程学习评价的RFM-DS模型[J]. 扬州职业大学学报, 2022(01):50-53.