文章编号: 1001-4543(2020)03-0213-07 DOI: 10.19570/j.cnki.jsspu.2020.03.006

学业导师制提升课程成绩的机器学习评价方法

罗加美, 牛 森, 安俊宇, 薛建新

(上海第二工业大学 计算机与信息工程学院, 上海 201209)

摘 要: 学业导师制是实施和完善学分制的一种辅助制度, 可以有效地提高学生的综合素质、创新精神和实践能力。为了量化地分析引入学业导师制度对学生课程成绩的影响程度, 基于传统的机器学习模型, 提出一种基于多元线性回归模型的机器学习评价方法。该方法主要分为数据的预处理、数据的特征筛选、模型的训练、交叉验证以及成绩预测等 5 个阶段。最后, 根据学生的成绩数据进行实验分析, 对比学业导师制实施前后学生成绩的变化情况, 验证了学业导师制能够有效提升专业课程的及格率和优良率。

关键词: 学业导师制; 线性回归; 机器学习; 成绩预测

中图分类号: TP391 **文献标志码**: A

An Evaluation Method of Machine Learning for Academic Performance and Tutorial System

LUO Jiamei, NIU Sen, AN Junyu, XUE Jianxin

(School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China)

Abstract: The tutorial system is an auxiliary system for implementing and perfecting the credit system, which can effectively improve the comprehensive quality, innovative spirit and practical ability of students. In order to analyze the influence of the tutorial system on student course performance, this paper proposes an evaluation method of machine learning based on multiple linear regression models. The method is mainly divided into five stages: data preprocessing, data feature selection, model training, cross-validation, and score prediction. Finally, an experimental analysis was conducted on the course score. The comparison results proved that the tutorial system can effectively improve the pass rate and excellent rate of the course score.

Keywords: tutorial system; linear regression; machine learning; score prediction

0 引言

2018 年 10 月, 教育部发布的《深化本科教育教学改革相关意见》提出要建立健全本科生学业导师制度, 让符合条件的教师帮助学生制订更具个性化的培养方案和学业生涯规划。本科生学业导师制度, 是对学生进行生涯规划教育和引导 [1]。

目前, 在全员协同育人理念下, 更加突显学业导师制度意义重大。已有众多的高等院校开展了学业导师制度, 在专业的个性化培养学生方面有了显著提升。当前, 针对学业导师制度的研究主要集中在思想政治理论探讨上, 樊奔 [2] 结合自身实践, 针对目前高校本科生导师制度的人才培养效果, 剖析了制度实施过程中存在的主要问题和困难, 并提出了自己的观点和理念。匙芳廷等 [3] 指出本科生导师制对大学生创新教育和实践能力培养方面的优点以及如何加强本科生导师制度建设的方法和手段; 宋怀

收稿日期: 2020-06-30

通信作者: 牛 森 (1984–), 男, 河南济源人, 讲师, 博士, 主要研究方向为服务计算、数据挖掘等。E-mail: niusen@sspu.edu.cn

基金项目: 上海高校青年教师培养资助计划 (ZZEGD20018) 资助

涛等 [4] 提出构建突出专业价值的本科生导师制度培养模式, 将本科生导师培养与专业的特点和学生的兴趣相结合, 培养学生的职业精神; 闫冬春等 [5] 提出了学业导师指导下的 “三位驱动” 生成实习模式, 包括就业、考研和毕业论文的三位模式。然而, 当前的研究还缺少确定的、可以量化的指标来衡量学业导师制度对学生成绩带来的积极影响。

2016 年上海第二工业大学开始出台学业导师制度的相关实施办法, 至今已经产生了充足的研究数据。对于学生成绩而言, 学业导师制度对其应当具有积极作用, 并可能对不同的学生群体产生不同的作用程度, 如某学生对 “数据结构与算法” 课程兴趣颇深且获得了好成绩, 那么他更大概率会对学业导师在该方面的指导更专注, 从而在下学期的 “算法设计与分析” 课程中取得比无学业导师更高的成绩。在此基础上, 通过分析引入学业导师制度前后智能科学与技术专业的某班级成绩变化, 提出了基于多元线性回归的机器学习评价方法, 对学生的成绩进行预测分析, 从而对学业导师制度的评价进行精确的量化。

通过多元线性回归算法, 将某几门专业基础课程成绩作为特征数据, 训练模型预测此后开展的专业课程成绩, 其中特征中使用的课程发生时间早于标签所使用的课程成绩。因此, 将未引入导师制度的成绩作为训练集, 引入导师制度的成绩作为测试集。这里假设每门课的打分制度是一致的, 又因为选入特征的课程为更早开展的课程, 受导师制度影响小于或远小于标签所使用的课程, 故训练集和测试集的特征基本是分布一致的。而标签之间的分布区别主要受导师制度的影响, 因此使用训练集得到的模型对测试集进行预测, 预测值与真实值的差距即为导师制度的影响程度。最后, 通过真实的班级成绩数据, 可以证明学业导师制度在我校实施以来所取得的成效显著。

1 多元线性回归模型

多元线性回归是线性回归重要的组成部分, 被广泛应用在众多的科学研究中 [6-8]。一般多元线性回归模型的基本形式 [9-10] 为:

$$y = w_1x_1 + w_2x_2 + \cdots + w_px_p + b \tag{1}$$

式中: y 为因变量; x_i 为自变量, $i \in (1, 2, \cdots, p)$, 是数据的 p 维属性特征; w_i 为回归系数, $i \in (1, 2, \cdots, p)$, 其决定了因变量和自变量间的线性关系; b 为偏移误差项。

对于一个具有 n 组数据, p 个维度特征的现实问题, 此多元线性回归模型可以写成如下的矩阵形式:

$$\bar{\boldsymbol{y}} = \boldsymbol{X}W + \boldsymbol{B} \tag{2}$$

式中:

$$\bar{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$\boldsymbol{W} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_p \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

在多元线性回归模型求解过程中, 利用最小二乘法对 $\boldsymbol{W}$ 和 $\boldsymbol{B}$ 进行参数的估计。若将 w 和 b 合并形成向量 $\hat{\boldsymbol{w}} = (w; b)$, 则矩阵 $\boldsymbol{X}$ 就变成了最后一个元素恒为 1 的 $n \times (p+1)$ 大小的矩阵, 即为:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & 1 \\ x_{21} & x_{22} & \cdots & x_{2p} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & 1 \end{bmatrix} = \begin{bmatrix} x_1^{\mathrm{T}} & 1 \\ x_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ x_n^{\mathrm{T}} & 1 \end{bmatrix}$$

通过线性回归模型的参数估计, 分别对 w 求导, 可得的如下公式:

$$\frac{\partial \boldsymbol{E}_{\hat{w}}}{\partial \hat{\boldsymbol{w}}} = 2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) \tag{3}$$

当 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 为正定矩阵时, 估计参数求解表达式为:

$$\hat{\boldsymbol{w}}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}X^{\mathrm{T}}\boldsymbol{y} \tag{4}$$

从而, 最终可得到多元回归模型如下:

$$f(\hat{\boldsymbol{x}}_i) = \hat{\boldsymbol{x}}_i^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{5}$$

2　基于多元线性回归的机器学习评价方法

为了分析学业导师制度对学生成绩的影响程度, 结合机器学习模型, 提出了基于多元线性回归的分析评价方法, 如图 1 所示。

整个评价方法流程分为 5 个阶段, 包括数据预处理、特征筛选、模型训练、交叉验证和预测分析等。技术流程如下:

(1) 数据预处理。由于部分课程学生缺考, 原始数据中存在缺失值的情况。在此数据预处理阶段, 主要利用计算平均值的方式来补全缺失值。计算公式为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{6}$$

式中: $\bar{x}$ 为缺失值; x_i 为数据中同一属性值相同的完全变量。

(2) 特征筛选。选出某一门课程, 将其与预测的目标课程计算 Pearson 相关系数, 得到对目标影响较大的课程作为特征数据集 D。r 的计算公式如下:

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} \tag{7}$$

式中: x_i, y_i 分别代表 X 和 Y 数据的第 i 条记录; N 表示记录的个数。

(3) 模型训练。利用 2014、2015 年未引入学业导师的学生课程数据, 指定 3 门专业课程作为目标, 通过特征筛选, 分别生成模型的训练集和测试集。然后, 根据多元线性回归模型的理论进行参数估计, 得到回归模型。

(4) 交叉验证。在交叉验证阶段, 为了保证模型训练、参数评估的准确性, 采用 5 折交叉验证方法, 从而优化模型的参数。

(5) 预测分析。在训练好的回归模型基础上, 对引入本科生学业导师制度的学生课程成绩进行预测, 进而去比较预测值和真实值之间的变化情况, 从而评估学业导师制度的重要性程度。

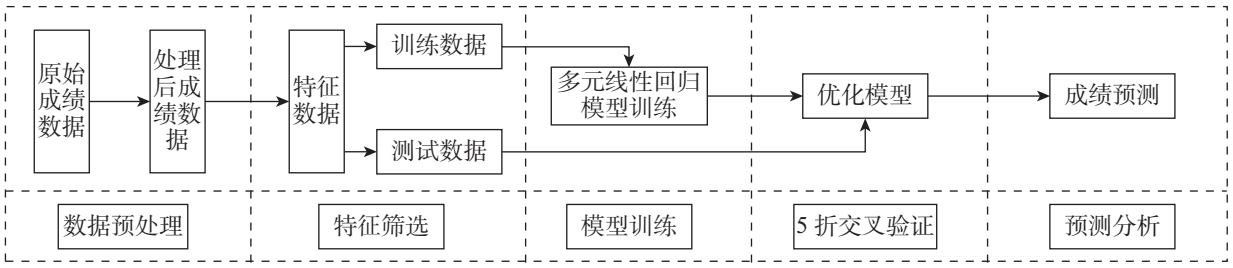

图 1　基于多元线性回归的机器学习评价方法流程

Fig. 1　The processes of evaluation method of machine learning based on multiple linear regression

3　实验结果与分析

3.1　实验数据集描述

使用了 2014—2017 年我校引入本科生学业导师制度前后某班级学生的各科成绩数据, 其中 2014、2015 年为未引入学业导师制度的成绩数据, 2016、2017 年为引入学业导师制度后的成绩数据。原始数据中包含了 174 人, 共 36 门课程的成绩。在特征筛选阶段, 选择了 7 门主要课程为目标, 其中 2014、2015 年为训练集共 79 条数据, 2016、2017 年为验证集共 95 条数据。由于整体数据量较小, 采用 5 折交叉验证来保证模型训练的有效性, 并利用平均绝对误差和均方根误差来衡量模型训练的误差。整个实验过程利用 Python3.7 语言进行编程, 在 anaconda 集成环境中进行开发实现。

3.2　实验设置

将 2014、2015 年的数据作为未引入导师制度时的训练集, 同时将 2016、2017 年的数据作为引入导师制度后的测试集, 共使用 3 组特征和标签进行实验分析。根据提出的机器学习分析框架, 经过数据预处理阶段, 在特征筛选阶段, 通过每个学生的课程成绩, 根据培养计划中课程的联系, 利用 Pearson 相关系数计算得出 “数据结构与算法” “离散数学” “算法设计与分析” “概率论与数理统计” “模式识别” “人工智能” 和 “计算机组成原理” 等 7 门课程之间的相关系数, 如表 1 所示。

根据表 1 中的课程相关系数, 选择相似度阈值 $\geqslant 0.4$ 为条件, 进行特征数据的筛选。在此基础上, 在模型训练和模型验证中作如下的实验设置:

表 1 课程间的相关系数表

Tab. 1 The correlation coefficient between courses

课程名称	数据结构与算法	离散数学	算法设计与分析	概率论与数理统计	模式识别	人工智能	计算机组成原理
数据结构与算法	1.000	0.326	0.427	0.359	0.393	0.417	0.259
离散数学	0.326	1.000	0.429	0.436	0.497	0.556	0.365
算法设计与分析	0.427	0.429	1.000	0.438	0.344	0.559	0.393
概率论与数理统计	0.359	0.436	0.438	1.000	0.499	0.492	0.464
模式识别	0.393	0.497	0.344	0.499	1.000	0.591	0.479
人工智能	0.417	0.556	0.559	0.492	0.591	1.000	0.412
计算机组成原理	0.259	0.365	0.393	0.464	0.479	0.412	1.000

(1) 利用 “数据结构与算法” “离散数学” 和 “概率论与数理统计” 3 门课的成绩作为特征, 预测 “算法设计与分析” 的课程成绩。

(2) 利用 “数据结构与算法” “算法设计与分析” “离散数学” “概率论与数理统计” 和 “计算机组成原理” 5 门课的成绩作为特征, 预测 “人工智能” 的课程成绩。

(3) 利用 “人工智能” “离散数学” “计算机组成原理” 和 “概率论与数理统计” 4 门课的成绩作为特征, 预测 “模式识别” 的课程成绩。

3.3 实验结果与对比分析

根据 3.2 中的 3 个实验设置, 在模型训练阶段, 根据多元线性回归模型训练的过程, 参数估计、误差分析、预测值与真实值之间对比结果分别如表 2∼4 所示。

在表 2 中, 根据实验设置 (1) 中的特征数据, 分别得出了多元线性回归方程中的系数分别为 0.205、0.201 和 0.44, 偏移项为 5.525。依据估计参数可计算出模型训练的平均绝对误差和均方根误差分别为 4.924 和 6.508。同时, 可计算出每位同学在此基础上 “算法设计与分析” 课程的成绩分布。根据预测出的成绩可计算出课程预测的及格率和优良率分别为 86.7% 和 2.2%。通过预测值和真实值之间的对比, 可以发现在引入学业导师制后, “算法设计与分析” 课程在及格率和优良率上分别提高了 10% 和 22.2%。

在表 3 中, 根据实验设置 (2) 中选择的特征数据, 可得出多元线性回归模型中的估计参数系数分别为 0.255、0.218、0.248、0.308 和 −0.041, 偏移项值为 3.555。依据估计参数可计算出模型训练的平均绝对误差和均方根误差分别为 5.358 和 6.568。同时, 可计算出 “人工智能” 课程的学生成绩分布。根据预测的成绩分布, 得到预测的及格率和优良率分别为 96.6% 和 25.6%。通过与真实值对比, 可以发现在引入学业导师制度后, “人工智能” 课程学生在优良率上提升了 13.3%。

在表 4 中, 根据实验设置 (3) 要求的特征数据, 可训练出多元线性回归模型中的系数参数分别为 0.424、0.279、0.06 和 0.039, 偏移项值为 14.035。根据估计的模型参数, 可得出模型训练的平均绝对误差和均方误差分别为 4.108 和 5.066。同时, 可以计算出 “模式识别” 课程学生的成绩分布, 得出预测的及格率和优良率分别为 94.4% 和 20%。通过与真实的值进行比较, 发现在引入学业导师制度后, “ 模式识别” 课程学生的及格率和优良率分别提高了 3.4% 和 6.6%。

表 2 “算法设计与分析” 课程预测参数评估和值对比表

Tab. 2 Parameters estimation and value comparison on Algorithms Design Techniques and Analysis

课程特征	数据结构与算法	离散数学	概率论与数理统计
估计系数参数	0.205	0.201	0.44
估计偏移项		5.525	
平均绝对误差	4.924	均方根误差	6.508
预测及格率	0.867	真实及格率	0.967
预测优良率	0.022	真实优良率	0.244

表 3　“人工智能” 课程预测参数评估和值对比表

Tab. 3　Parameters estimation and value comparison on Artificial Intelligence

课程特征	数据结构与算法	算法设计与分析	概率论与数理统计	离散数学	计算机组成原理
估计系数参数	0.255	0.218	0.248	0.308	−0.041
估计偏移项			3.555		
平均绝对误差	5.358		均方根误差	6.568	
预测及格率	0.966		真实及格率	0.956	
预测优良率	0.256		真实优良率	0.389	

表 4　“模式识别” 课程预测参数评估和值对比表

Tab. 4　Parameters estimation and value comparison on Pattern Recognition

课程特征	人工智能	离散数学	计算机组成原理	概率论与数理统计
估计系数参数	0.424	0.279	0.060	0.039
估计偏移项			14.035	
平均绝对误差	4.108		均方根误差	5.066
预测及格率	0.944		真实及格率	0.978
预测优良率	0.2		真实优良率	0.266

为了进一步分析学业导师制度的积极影响, 下面以 “算法设计与分析” 课程为例, 分析各个成绩阶段分数的变化情况, 分别如图 2∼5 所示。

在图 2∼5 中, 左图为课程训练集的成绩分布, 右图为课程测试集的成绩分布。在图 2∼4 中, 及格学生的分数、成绩在 $x \geqslant 80$ 分, $70 \leqslant x < 80$ 分间的成绩平均提高了 0.82 和 3.07 分、11.08 和 15.62 分、2.69 和 3.93 分。由此可以发现在学业导师制度影响下, 学生的 “算法设计与分析” 课程的总体成绩都有所提升。在图 5 中, 成绩在 $60 \leqslant x < 70$ 分的学生成绩提升了 −2.86 分和 −3.96 分。直观地发现, 成绩 $60 \leqslant x < 70$ 分的学生分数在引入学业导师制度后, 这部分学生的分数有所下降, 但通过进一步分析, 可知道学业导师制引入后, 该门课程的整体及格率提升了 10%。这部分学生大多是学业导师制度实施前不及格的部分。进一步也证实了学业导师制度对学生课程成绩影响的积极作用。

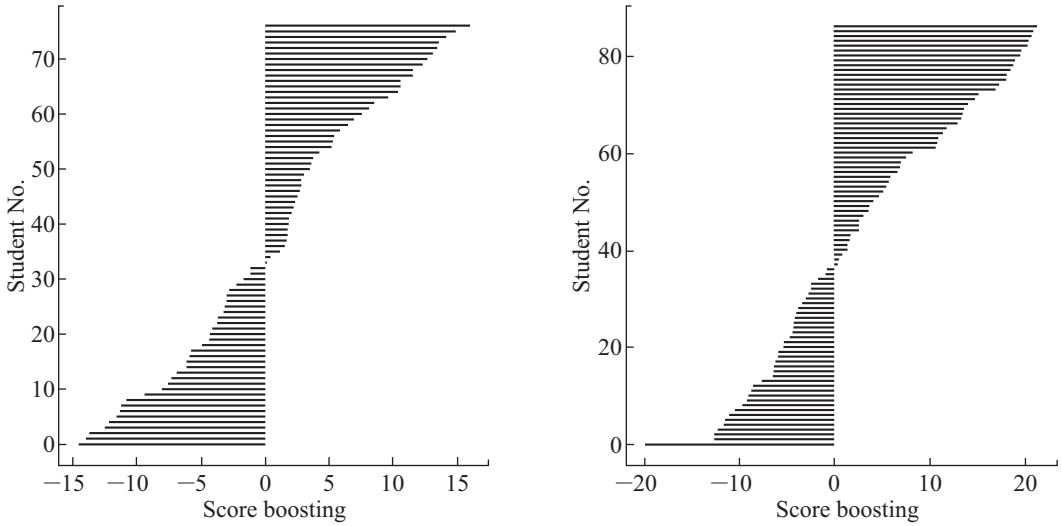

图 2　“算法设计与分析” 成绩及格学生的分数提升分布

Fig. 2　The distribution of increased scores of students passed on Algorithms Design Techniques and Analysis

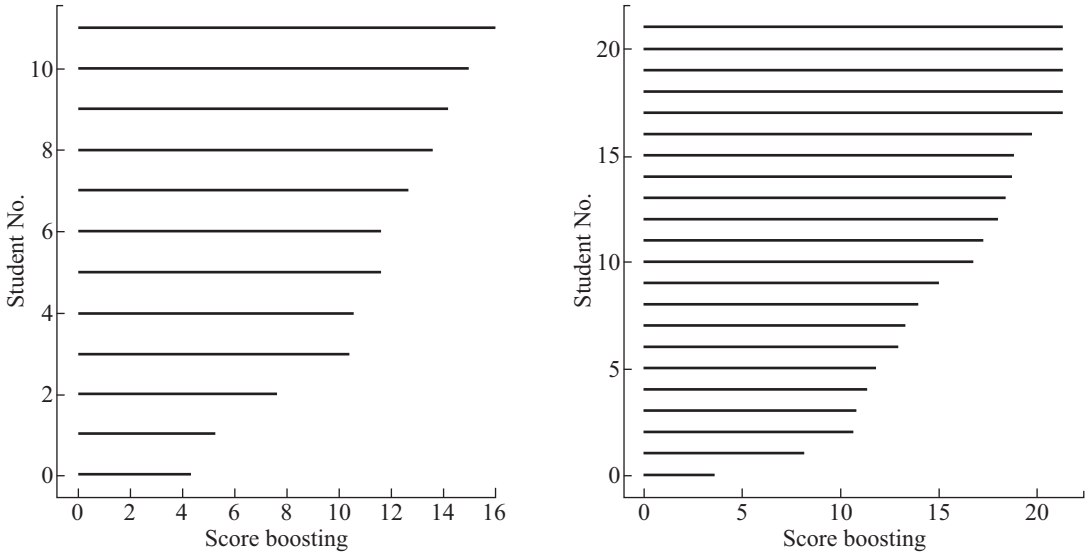

图 3 “算法设计与分析” 成绩 ⩾ 80 分的学生分数提升分布

Fig. 3 The distribution of increased scores of students‘ scores ⩾ 80 on Algorithms Design Techniques and Analysis

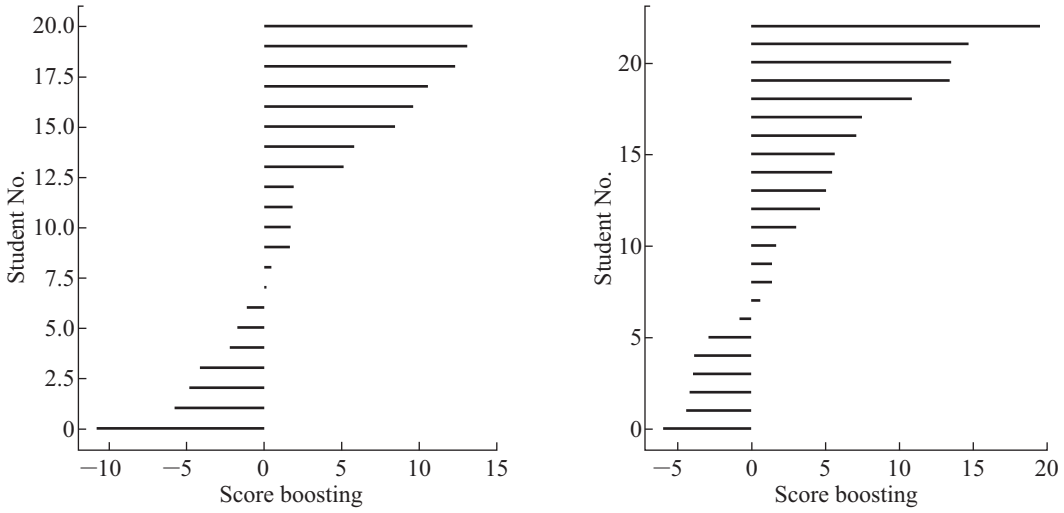

图 4 “算法设计与分析” 成绩 ⩾ 70 分且 < 80 分的学生分数提升分布

Fig. 4 The increased distribution of scores of students‘scores ⩾ 70 & < 80 on Algorithms Design Techniques and Analysis

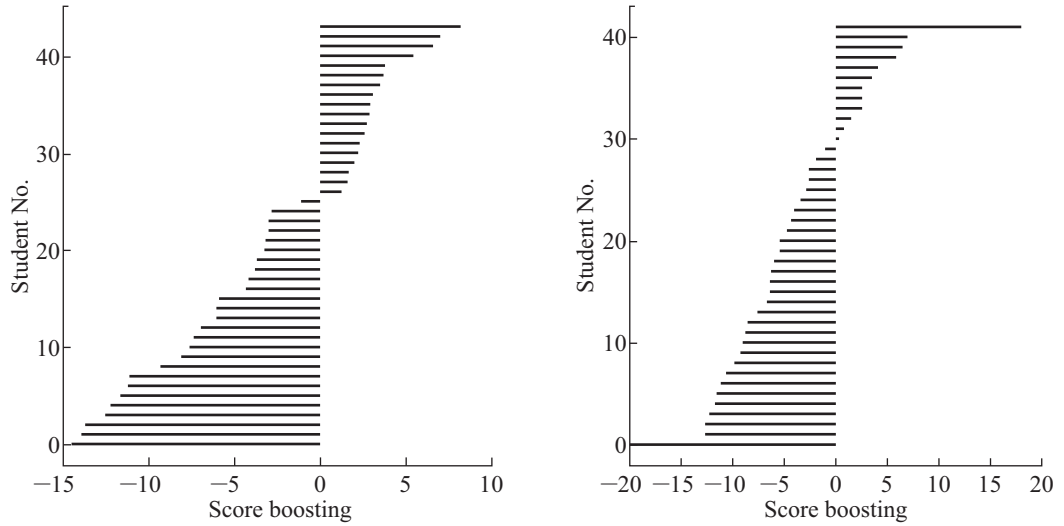

图 5 “算法设计与分析” 成绩 ⩾ 60 分且 < 70 分学生的分数提升分布

Fig. 5 The distribution of increased scores of students scores ⩾ 60 & < 70 on Algorithms Design Techniques and Analysis

4 结　论

本科生学业导师制度让学生在导师的影响下获得更具个性化的培养方案和学业生涯规划。在实施学业导师制度前后学生的课程成绩数据基础上, 提出了基于多元线性回归的机器学习分析框架。通过实验分析和对比, 发现在学业导师制度实施后, 无论在课程的及格率、优良率和整体分数上, 都有了很大的提高。可见学业导师制度可以极大提高学生学习的积极主动性。

目前的研究与分析更偏向于实行学业导师制度的计算机相关专业学生或工科学生, 且局限于对学习成绩的影响。今后可以进一步分析其他专业方向的学生, 研究该制度在其他学科上的影响程度, 并进一步涵盖学生的竞赛、升学、就业等全方位影响。

参考文献:

[1] 魏志荣. 本科生导师制: 历史、现状与未来 [J]. 山东高等教育, 2015, 3(10): 62-67.

[2] 樊奔. 本科生导师制的实践与思考 [J]. 教育现代化, 2016, 3(15): 27-29.

[3] 匙芳廷, 易发成, 王烈林, 等. 本科生导师制对大学生创新教育与实践能力培养研究 [J]. 吉林省教育学院学报, 2018, 34(4): 130-132.

[4] 宋怀涛, 马瑞, 李森, 等. 本科生导师制专业价值培养探析 [J]. 教育教学论坛, 2019(17): 78-79.

[5] 闫冬春, 程显好, 王凯, 等. 学业导师指导下的 “三位驱动” 生产实习模式的探索与实践 [J]. 大学教育, 2020(2): 81-83.

[6] MAXWELL S E. Sample size and multiple regression analysis [J]. Psychological Methods, 2000, 5(4): 434-458.

[7] GKIOULEKAS I, PAPAGEORGIOU L G. Piecewise regression analysis through information criteria using mathematical programming [J]. Expert Systems with Applications, 2019, 121: 362-372.

[8] LLOYD-JONES L R, ZENG J, SIDORENKO J, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics [J]. Nature Communications, 2019, 10(1): 1-11.

[9] DEFRIES J C, FULKER D W. Multiple regression analysis of twin data [J]. Behavior Genetics, 1985, 15(5): 467-473.

[10] HO J, PLEWA C, LU V N. Examining strategic orientation complementarity using multiple regression analysis and fuzzy set QCA [J]. Journal of Business Research, 2016, 69(6): 2199-2205.