

适用于学生成绩预测的学习数据特征综述

张 峰, 陈静静

(山东科技大学计算机科学与工程学院, 山东 青岛 266590)

✉ zhangfeng@sdust.edu.cn; cjj202083060002@163.com



摘 要:当前,线上线下的学习过程中,学生的学习活动产生了多种类型的学习数据,基于这些数据,采用机器学习技术能够训练学生成绩的预测模型。文章首先介绍基于机器学习技术构建学生成绩预测模型的基本流程;其次系统地梳理基于机器学习技术的成绩预测工作中使用的学习数据特征,并介绍基于非过程性数据特征和过程性数据特征的成绩预测研究工作;最后分析和总结使用两类学习特征数据进行成绩预测的不足,并对未来研究工作做展望。

关键词:非过程性特征;过程性特征;教育数据挖掘;成绩预测

中图分类号:TP391 **文献标志码:**A

Overview of Learning Data Features Applicable to Student Performance Prediction

ZHANG Feng, CHEN Jingjing

(School of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

✉ zhangfeng@sdust.edu.cn; cjj202083060002@163.com

Abstract: Currently, in the process of online and offline learning, students' learning activities generate various types of learning data. Based on these data, machine learning technology can be used to train prediction models for students' grades. Firstly, the basic process of building student performance prediction model based on machine learning technique is introduced. Then, the learning data features used in performance prediction based on machine learning technique is systematically reviewed, and the research work of performance prediction based on non-process data features and process data features is introduced. Finally, the shortcomings of using the two types of learning data features to predict performance are summarized, and the prospects for future research work is presented.

Key words: non-process features; process features; educational data mining; performance prediction

0 引言(Introduction)

教育数据挖掘^[1](Educational Data Mining, EDM)是当前教学技术领域研究的热点,旨在从海量的教育数据中发现隐藏信息。目前,已有大量工作通过分析学生学习数据的非过程性特征,如教育背景、上课出勤率等,利用决策树^[2]、朴素贝叶斯分类法^[3]、线性回归^[4]和神经网络^[5]等机器学习方法建立成绩预测模型。然而,现有方法大多忽略了学生在学习过程中产生的一些过程性特征数据,也包含有价值的学生知识水平信息,

如学习过程中视频观看的频次等数据。这些数据能够反映学生真实的学习情况,可以作为预测学生学习成绩的依据^[6]。例如,对于一组被学生循环观看的视频,可以认为学生学习视频内容时难度较大,也反映学生在学习部分视频时投入精力较多^[7]。目前,很少有综述性研究从成绩预测所使用的学习特征的层面对已有工作进行梳理、分析和总结。因此,本文根据不同的学习特征对现有学生成绩预测的研究工作进行分类总结。

收稿日期:2023-02-20

基金项目:山东科技大学青年教师教学拔尖人才培养项目(BJ20200505);山东省教育科学“十四五”规划课题(2021YB028);山东科技大学优秀教学团队建设计划

资助项目(JXTD20180503)

1 基于机器学习技术的成绩预测流程 (Performance prediction process based on machine learning technique)

基于机器学习技术的学生成绩预测模型构建的基本流程如图1所示,主要包括数据采集、数据提取与预处理、构建成绩预测模型三个步骤。首先,通过在线学习平台获取与学生学习相关的过程性特征、非过程性特征数据;其次,采用各种学习方法对数据进行分析,确定可用于预测的数据,并对数据进行预处理,包括对数据进行过滤和筛选,确保采集的数据与学习高度相关联,此外要进行必要的格式转换、数据归一化,以及处理大量杂乱的数据等;最后,提取数据特征,采取多种机器学习算法包括逻辑回归、决策树、朴素贝叶斯、支持向量机、反向传播和随机森林建立成绩预测模型。通过对各种模型进行对比,选择最佳的模型对学生成绩进行预测,为学生学习及教师指导提供意见。

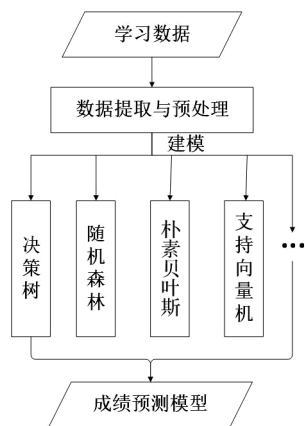


图1 基于机器学习技术的成绩预测模型构建流程

Fig. 1 Construction process of performance prediction model based on machine learning technique

2 基于学习特征的成绩预测 (Performance prediction based on learning data feature)

基于学生的教育数据进行学生成绩预测,是教育数据挖掘领域的研究热点。如何利用这些教育数据,将其转换为有用的信息和知识,为教育决策、教学优化服务提供帮助,已经成为教育数据挖掘的研究目标,也是教育工作者尤其是教育决策者重点关注的内容^[8]。研究分析学生学习的相关数据,并以此为据,为后续的建模工作提供充足的数据资源,是学生成绩预测任务的基本前提。本文基于现有数据特征,从非过程性特征、过程性特征两个方面对预测成绩的研究工作进行总结。

2.1 基于非过程性特征的成绩预测

使用非过程性特征进行成绩预测,主要包括个人教育背景、先验知识、认知能力、学习成绩、学习时间、学习态度等^[9]。本小节将非过程性特征分为静态数据和动态非过程性数据,并分别从这两个方面梳理和总结主要工作。

2.1.1 静态数据

使用静态数据(即很长时间内不会变化的数据),主要包括学生个人教育背景、学校类型、学校排名、家长受教育程度、社

会关系等预测成绩^[10-11]。例如,ABU AMRA等^[12]首先在线下收集学生的学习行为特征(基本信息、教育背景、父母受教育程度、社会关系),然后进行数据分析以及预处理,采取多种机器学习算法预测学生的成绩。ACHARYA等^[13]收集整理了500名学生的信息,包括性别、出生日期、所学专业、所在城市、父母是否工作等,分别利用最近邻算法和朴素贝叶斯算法预测学生的学习成绩,并进行成绩预测模型的对比,选取最佳模型预测学生的成绩。PANDEY等^[14]基于524名学生的性别、年龄、毕业学校等18个特征信息,利用决策树算法预测学生完成学位的评价等级,从而帮助教师了解学生的学习状况。SAARELA等^[15]利用硕士研究生之前的最终评级和他们在本科阶段的成绩作为特征,用统计方法预测硕士研究生在校的表现。DÍAZ MUJICA等^[16]使用学生情感和情绪变量(对大学课程的满意度、对学业成绩的看法、自己压力、自尊和疲劳等情感因素)对学生表现进行早期预测。NGUYEN等^[17]通过收集越南1117名学生的数据,包括个人情绪智力(Emotional Intelligence, EI)、生活满意度、职业、情绪、学习目标、毅力、学习环境、社交技能等,使用常见的数据分析方法分析以上数据对学生学业成绩的影响。

2.1.2 动态非过程性数据

使用动态非过程性特征(即指学生在线学习过程中产生的一些非过程性数据),包括在线学习时间、访问量、平均学习成绩等预测成绩^[6]。例如,MARBOUTI等^[19]利用大学一年级学生学习计算机课程前5周的行为数据,包括家庭作业分数、测验成绩、期中考试成绩等,基于多种机器学习算法(逻辑回归、决策树、朴素贝叶斯和随机森林等)创建该课程的预测模型。HANA^[20]收集了马萨里克大学信息学院的学生提供的与学习课程相关的学分数、平均成绩等,使用支持向量机、随机森林、基于规则的分类器、IB1和朴素贝叶斯算法预测学生的学业表现。HUSSAIN等^[21]通过数字电子教育与设计套件(Digital Electronics Education and Design Suite, DEEDS)系统采集到学生在数字设计课程中各课程学习的平均时间、活动总数、平均空闲时间、平均点击时间和与学习相关的活动总数,使用人工神经网络、支持向量机、逻辑回归、朴素贝叶斯和决策数等机器学习算法,预测每节课的学生成绩。JIMÉNEZ等^[22]利用静态数据(学习者性别、居住地、以前学校类型)和动态数据(学生在每门课程中的学习成绩)为特征,提出了多目标时间优化模型并运用逻辑回归、随机森林等机器学习算法对学习者的辍学预测。XU等^[23]通过收集小规模在线课程(Small Private Online Course, SPOC)平台2017年秋季和2018年春季学生学习计算机网络课程的学习行为,包括点击流、每次活动花费的时间、作业活动取得的成绩等,使用多元回归模型预测学生的成绩,该成绩预测工作具有实用性,为学习分析和个性化教学提供了启示。宋丹等^[24]在教学多源数据采集与处理的基础上,以“操作系统”课程为例,基于已授专业课、单元测试(可以理解为平时测验或作业成绩)、移动端教学数据和虚拟仿真实验数据等多个维度的数据进行课程成绩预测。XU等^[25]将学生在线学习的时间特征、学习成绩和在线课程平台论坛内学生互动相关的大量属性构成训练集,提出一种新的半监督回归算法,用于预测远程在线课程中本科生的最终成绩。

HASSAN等^[26]通过收集开放式大学的学习数据集中的学生论坛互动的学习特征,使用深度长期短期记忆模型预测学生的早期辍学情况。HUNG等^[27]结合静态数据(包括学生的基本信息、学生开学前的历史成绩等)和动态数据类型(访问课程材料的频率、发布在线讨论的频率等),采用深度学习和集成学习作为主要算法进行绩效成绩预测。

2.2 基于过程性特征的成绩预测

使用过程性特征预测成绩,现有工作大部分都是通过数据挖掘、学习分析等技术构建行为分析模型,以此深度挖掘学习者在线学习行为特点,但很少将其作为特征应用于成绩预测领域。例如,MUKALA等^[28]采用模糊挖掘算法,提取学生在大规模开放在线课程(Massive Open Online Course, MOOC)视频中的学习行为模式,发现学业成绩差的学生行为结构差且不可预测,而学业成绩优秀的学生的学习行为序列总体相似,这类有规律的行为是影响学习效果的重要因素。CRISTOBAL等^[29]根据开源学习管理系统(Modular Object-Oriented Dynamic Learning Environment, MOODLE)课程的数据,使用聚类算法将学生分为两组,然后使用启发式挖掘算法分别构建学生学习过程模型,从中发现成绩优秀的学生在测验中非常活跃,而成绩差的学生重视理论学习而忽略了实践,并且成绩差的学生模型比成绩优秀的学生的模型结构要小得多,说明成绩差的学生学习时间少。JUHAŇÁK等^[30]采用以过程为中心的方法分析管理系统MOODLE中学生的学习行为,特别是基于测验的学习活动中使用启发式挖掘算法挖掘学生学习行为的依赖图,并且每个活动以及路径都补充了给定活动或路径的执行频率,从过程模型中揭示有关学生测验行为的事实,例如大部分学生通常不会自己提交测验供模型打分,但当设置的时间限制达到时,系统会自动提交学生测验。BAO等^[31]提出了一种基于行为过程相似的学生成绩预测方法,考虑过程性特征预测成绩,具体包括活动相似度、流相似度和时间相似度,它主要基于在线评测系统(Online Judge, OJ)中积累的学生答题行为日志,以学生不同题目的出现频次及答题时间延迟信息为出发点,挖掘不同成绩等级学习者的公共答题行为过程模型,计算学生个体答题行为与公共答题行为过程之间的4种相似性度量值作为输入特征,运用分类学习算法建立成绩预测模型。

3 存在的不足以及未来工作 (Existing deficiencies and future work)

根据使用非过程性特征和过程性特征两类数据用于学生成绩预测领域的分析和研究,发现其存在的不足,并对未来工作进行展望。

3.1 存在的不足

第一,使用非过程性特征进行成绩预测的方法中,考虑了多种学习数据,借助积累的学习数据预测学生未来的学习成绩,既可以帮助教师了解学生的学习状态,还能方便他们及时发现教学过程中存在的问题,以便及时调整教学方法;但不足之处在于难以动态掌握学生的学习状态,忽略了学习过程中隐藏地对成绩预测有效的信息,预测模型整体的准确率还有待提高。第二,使用过程性特征进行成绩预测的方法中,虽然使用了学习行为过程相关的特征数据,但是不足之处在于很少有工作将其用于成绩预测,即使在成绩预测时考虑了过程性特

征,也只是面向某种类型的学习行为,没有综合考虑多种类型的学习行为,从而影响了成绩预测的准确性。

3.2 未来工作

未来可从学习数据和成绩预测技术方面进行改进。第一,未来需要进一步挖掘学生在学习过程中隐含的有价值的信息,可以从整个视频学习过程的角度,考虑学习过程中更细粒度的视频学习行为,例如某个视频的观看时长,或者观看视频时的拖动、暂停等动作,验证是否可以获得更好的预测效果。第二,基于深度学习的成绩预测是当前教育数据挖掘领域的研究热点,下一步将视频学习行为特征融入深度学习中。

4 结论 (Conclusion)

随着教育信息化的发展,在线学习平台中积累了大量的学习行为数据,为教育数据挖掘的研究提供了数据支撑。作为教育数据挖掘中的研究热点,基于机器学习技术的成绩预测能够及时反馈学生的学习状态,而学习行为数据是成绩预测的基础。本文从基于机器学习技术的学生成绩预测使用的学习数据特征的角度,总结了使用非过程性特征和过程性特征进行成绩预测的相关工作,指出现有工作存在的不足,希望能够为使用学习数据进行成绩预测方面的研究构建一个较为完整的全景图。在后续的研究中,可以进一步深入挖掘学习过程中产生的各类对成绩预测有效的数据特征,为构建更加准确的成绩预测模型奠定基础。

参考文献 (References)

- [1] 周庆,牟超,杨丹. 教育数据挖掘研究进展综述[J]. 软件学报, 2015, 26(11): 3026-3042.
- [2] POLYZOU A, KARYPIS G. Feature extraction for next-term prediction of poor student performance[J]. IEEE Transactions on Learning Technologies, 2019, 12(2): 237-248.
- [3] LÓPEZ-ZAMBRANO J, LARA TORRALBO J A, ROMERO C. Early prediction of student learning performance through data mining: a systematic review[J]. Psicothema, 2021, 33(3): 456-465.
- [4] JOVANOVIĆ J, SAQR M, JOKSIMOVIĆ S, et al. Students matter the most in learning analytics: the effects of internal and instructional conditions in predicting academic success[J]. Computers and Education, 2021, 172: 104251.
- [5] MONLLAÓ OLIVÉ D, HUYNH D Q, REYNOLDS M, et al. A quest for a one-size-fits-all neural network: early prediction of students at risk in online courses[J]. IEEE Transactions on Learning Technologies, 2019, 12(2): 171-183.
- [6] CHU Y W, TENORIO E, CRUZ L, et al. Click-based student performance prediction: a clustering guided meta-learning approach[C]//USAMA F, ZHU X Q, CHEN Y X, et al. 2021 IEEE International Conference on Big Data. Piscataway: IEEE, 2022: 1389-1398.
- [7] ZHANG F, LIU D, LIU C. MOOC video personalized classification based on cluster analysis and process mining[J]. Sustainability, 2020, 12(7): 3066.

- [8] 孟卓,袁梅宇. 教育数据挖掘发展现状及研究规律的分析[J]. 教育导刊,2015(2):29-33.
- [9] CARTER A S, HUNDHAUSEN C D, ADESOPE O. The normalized programming state model: predicting student performance in computing courses based on programming behavior[C]//BRIAN D, JUDY S, QUINTIN C, et al. Proceedings of the 11th annual International Conference on International Computing Education Research. New York: ACM, 2015:141-150.
- [10] BAHRTIDINOV B, SÁNCHEZ E. Predicting academic success by using context variables and probabilistic classification[J]. Expert Systems, 2017, 34(4):e12195.
- [11] TANI K, GILBEY A. Predicting academic success for business and computing students[J]. International Journal of Information and Communication Technology Education, 2016, 12(4):15-24.
- [12] ABU AMRA I A, MAGHARI A Y A. Students performance prediction using KNN and Naïve Bayesian[C]//ASHHRAF Y A, IBRAHIM E, ALI A H, et al. 2017 8th International Conference on Information Technology. Piscataway: IEEE, 2017:909-913.
- [13] ACHARYA A, SINHA D. Early prediction of students performance using machine learning techniques[J]. International Journal of Computer Applications, 2014, 107(1):37-43.
- [14] PANDEY M, KUMAR SHARMA V. A decision tree algorithm pertaining to the student performance analysis and prediction[J]. International Journal of Computer Applications, 2013, 61(13):1-5.
- [15] SAARELA M, KÄRKKÄINEN T. Analysing student performance using sparse data of core bachelor courses[J]. Journal of Educational Data Mining, 2015, 7(1):3-32.
- [16] DÍAZ MUJICA A, PÉREZ VILLALOBOS M V, BERNARDO GUTIÉRREZ A B, et al. Affective and cognitive variables involved in structural prediction of university dropout[J]. Psicothema, 2019, 31(4):429-436.
- [17] NGUYEN K T, DUONG T M, TRAN N Y, et al. The impact of emotional intelligence on performance: a closer look at individual and environmental factors[J]. The Journal of Asian Finance, Economics and Business, 2020, 7(1):183-193.
- [18] MIGUÉIS V L, FREITAS A, GARCIA P J V, et al. Early segmentation of students according to their academic performance: a predictive modelling approach[J]. Decision Support Systems, 2018, 115:36-51.
- [19] MARBOUTI F, DIFES-DUX H A, MADHAVAN K. Models for early prediction of at-risk students in a course using standards-based grading[J]. Computers & Education, 2016, 103:1-15.
- [20] HANA B. A comparative analysis of techniques for predicting student performance[C]//TIFFANY B, CHI M, FENG M Y, et al. Proceedings of the 9th International Conference on Educational Data Mining. North Carolina, United States: IEDMS, 2016:306-311.
- [21] HUSSAIN M, ZHU W H, ZHANG W, et al. Using machine learning to predict student difficulties from learning session data[J]. Artificial Intelligence Review, 2019, 52:381-407.
- [22] JIMÉNEZ F, PAOLETTI A, SÁNCHEZ G, et al. Predicting the risk of academic dropout with temporal multi-objective optimization[J]. IEEE Transactions on Learning Technologies, 2019, 12(2):225-236.
- [23] XU Z J, YUAN H, LIU Q S. Student performance prediction based on blended learning[J]. IEEE Transactions on Education, 2021, 64(1):66-73.
- [24] 宋丹, 刘洞波, 丰霞. 基于多源数据分析的课程成绩预测与课程预警研究[J]. 高等工程教育研究, 2020(1):189-194.
- [25] XU J, MOON K H, SCHAAR M. A machine learning approach for tracking and predicting student performance in degree programs[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(5):742-753.
- [26] HASSAN S U, WAHEED H, ALJOHANI N R, et al. Virtual learning environment to predict withdrawal by leveraging deep learning[J]. International Journal of Intelligent Systems, 2019, 34(8):1935-1952.
- [27] HUNG J L, SHELTON B E, YANG J, et al. Improving predictive modeling for at-risk student identification: a multistage approach[J]. IEEE Transactions on Learning Technologies, 2019, 12(2):148-157.
- [28] MUKALA P, BUIJS J C, LEEMANS M, et al. Learning analytics on coursera event data: a process mining approach[C]//MILAN P, MAHIN A, ABDULAZIZ M, et al. Proceedings of the 5th International Symposium on Data-driven Process Discovery and Analysis. Vienna, Austria: CEUR, 2015, 1527:18-32.
- [29] CRISTOBAL R, REBECA C, ALEJANDRO B, et al. Educational process mining: a tutorial and case study using Moodle data sets[J]. Data Mining and Learning Analytics Applications in Educational Research, 2016, 1(3):1-28.
- [30] JUHAŇÁK L, ZOUNEK J, ROHLÍKOVÁ L. Using process mining to analyze students' quiz-taking behavior patterns in a learning management system[J]. Computers in Human Behavior, 2019, 92:496-506.
- [31] BAO Y X, LU F M, WANG Y X, et al. Student performance prediction based on behavior process similarity[J]. Chinese Journal of Electronics, 2020, 29(6):1110-1118.

作者简介:

张峰(1981-),男,博士,副教授。研究领域:教育数据挖掘,业务流程管理。

陈静静(1997-),女,硕士生。研究领域:教育数据挖掘。