

基于在线学习行为数据的学习效果模型与影响因素分析

倪 琴, 徐宇辉, 魏廷江, 高 荣

(上海师范大学 信息与机电工程学院, 上海 201418)

摘 要: 为了探究学生在线学习情况与学习效果之间的关系, 采用数据标注的方式解决学生学习行为表示的问题. 以S大学在线教学平台数据为研究对象, 通过数据挖掘技术探寻学生在线学习行为与学习效果之间的关系. 对比多种机器学习算法后, 选定随机森林算法作为学习效果预测模型的基本算法. 研究发现: 最能影响学习效果的因素是文档学习总时长, 最终构建的学习效果预测模型对整体数据集的分类准确率达到84.69%.

关键词: 学习行为表示; 相关性分析; 数据挖掘; 随机森林算法; 分类预测

中图分类号: TP 391.77 **文献标志码:** A **文章编号:** 1000-5137(2022)02-0143-06

Influence factor analysis and learning effect model based on online learning behavior data

NI Qin, XU Yuhui, WEI Tingjiang, GAO Rong

(College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China)

Abstract: In order to explore the relationship between students' online learning situation and learning effect, students' learning behavior representation was solved by adopting data annotation. Based on the data from the online teaching platform of S University, the relationship between students' online learning behavior and learning effects was explored through data mining technology. By comparing a variety of machine learning algorithms, the random forest algorithm was selected as the basic algorithm of the learning effect prediction model. It was found that the total duration of document learning could affect the learning effect greatly. The final learning effect prediction model was constructed which provided an accuracy of 84.69% for the classification of the overall data set.

Key words: learning behavior representation; correlation analysis; data mining; random forest algorithm; classification prediction

收稿日期: 2021-12-29

基金项目: 国家自然科学基金青年基金(6210020445); 上海市自然科学基金(21ZR1446900); 认知智能国家重点实验室智能教育开放课题; 国家新闻出版署重点实验室开放研究课题基金

作者简介: 倪 琴(1989—), 女, 副教授, 主要从事教育大数据分析、数据挖掘方面的研究. E-mail: niqin@shnu.edu.cn

引用格式: 倪琴, 徐宇辉, 魏廷江, 等. 基于在线学习行为数据的学习效果模型与影响因素分析[J]. 上海师范大学学报(自然科学版), 2022, 51(2): 143-148.

Citation format: NI Q, XU Y H, WEI T J, et al. Influence factor analysis and learning effect model based on online learning behavior data [J]. Journal of Shanghai Normal University (Natural Sciences), 2022, 51(2): 143-148.

0 引言

2020年初,在新型冠状病毒引起全国性肺炎疫情的影响下,教育部经多次研究决定2020年春季学期延期开学,提出“停课不停学”的在线学习方式,各大高校均延后开学^[1].一时间在线教学的学习效果引起了广大师生的关注.大学生的学习过程是一个复杂的过程,受到较多因素的影响.通常情况下,教育学者认为学业成绩的高低受学生已有认知水平、学习策略、学习动机以及外部环境等因素的综合影响^[2].

对学生学习行为的研究早期主要以调查问卷和统计分析的方式为主,并且停留在简单分析层面上.GONG^[3]通过对学生的学习行为进行统计与分析,探索师生交互状态与学生学习效率之间的关系.BAO^[4]分析了我国高等院校学生的学习行为特征,以此寻找学生学习行为与高校教学质量的隐含关系.YAO等^[5]从多个角度分析大学生在网络中的学习行为及其学习状态.以上研究均采用调查问卷分析的方法,存在前后侧数据回收统计耗时较长、分析结果较大程度受到问卷信效度制约的问题.

如今,研究者们逐渐转向利用数据挖掘技术对学生学习行为进行建模.ZHAO等^[6]利用多元回归方法寻找影响学生学业成绩的因素指标,并构建干预模型,对教学过程中产生的学习行为进行Logistic回归分析,从而找出存在学习危机的学生,并发起学业警示.LUO等^[7]将机器学习和学生行为分析相结合,通过决策树算法对在线学习行为进行数据分析,搭建“学习行为-效果”模型,在此基础上构建学习行为评估系统.FAKIR等^[8]利用分类、聚类和关联规则挖掘算法对Moodle系统进行数据分析,提取学习行为和模式的特征,创造理想的教育环境.采用机器学习的方法构建相关的学习行为模型具有较强的可解释性.

此外,还可以结合深度学习中的神经网络对学生学习行为进行分析研究.MOU等^[9]为了探究学习行为对学习效果的影响,使用Back Propagation(BP)神经网络探究MOOC课程中的低参与度和高辍学率等特征问题,为提高学习效果提供参考.但是,深度学习天然的黑盒特性,使得神经网络在学习行为分析任务中缺乏可解释性.

因此,本研究选用机器学习的方法对学生学习情况进行分析,并构建学习效果预测模型.从数据标注、数据清洗、数据建模及模型预测4个方面展示了在线学习环境下的学习行为分析过程,为在线教育中学生学习行为数据的表示和在线教学平台的建设提供了参考建议.

1 研究方法

为探究学生学习行为的表示和构建学习效果预测模型,本研究采用文献调查法结合数据挖掘的方法对S大学在线教学平台数据进行分析 and 探索.研究的对象包含一个主体和一个客体,主体是学生,客体是由主体所产生的各种行为与主体基本信息的有机结合.

1.1 学习行为表示方法

学生的学习行为主要包括作业记录、测试记录、讨论记录以及课程的参与情况等.学生与可视化界面之间的互动日志体现了学生的在线学习行为,本研究采用数据标注的方法将大量的日志数据转换成学生的统计型学习行为指标,比如:以进入视频界面到离开视频界面时长作为视频学习总时长,以文档状态为已看的数量除以文档总数所得结果作为文档学习率,以期末测试成绩作为考试得分等.经转换后,得到的17个学习行为指标:任务点查看率、视频完成率、视频学习总时长、文档学习率、文档学习总时长、课程学习时长、资料浏览数、资料浏览时长、作业完成率、作业平均分、测试完成率、测试平均分、发帖数、回帖数、点赞数、访问数和签到数.

在选取日志的过程中,为了保证数据能够代表学生的真实学习记录,除了点赞数、签到数这类瞬时行为之外,其他行为的时长持续时间至少大于30 s,才能被计入统计指标中.另外,保留学生最后一次访问课程的时间作为其学习该门课程的常规学习时间.

此外,为了衡量学生的学习效果,将学生的考试得分划分为优异、良好、合格和不合格4个类别,如

表1所示.

1.2 学习效果预测方法

本节通过对1.1节得到的学生学习行为指标进行梳理,并构建学习效果预测模型,从而找到影响学业成绩的主要学习行为指标.模型的构建过程参照数据挖掘的流程,即数据预处理、特征处理、模型构建和模型评价.

1.2.1 数据预处理

进行数据处理之前,应先对特征数据进行正态分布检验.

本研究的数据检验结果显示大部分数据未呈现正态分布,说明无法直接使用传统的统计检验方法构建模型.为了构建合理的学习效果预测模型,需要对原始数据进行如下预处理:

1) 根据“考试得分”字段滤除成绩为0和空值的样本,即删除没有期末成绩的样本.

2) 将数据集中课程的访问数超过200次的样本作为异常值,并用该学生在其他课程的平均访问数替代该项异常值.

3) 对于其他的学习行为指标,采用拉依达准则^[10]对具有过大误差值的坏值予以剔除.

在教学过程中,具有启发式的讨论式学习可以提高学生学习的主观能动性^[11].经过统计发现,在线教学平台中只有少数学生产生发帖、回帖和点赞等讨论类学习行为,本研究利用层次分析法(AHP)^[12],将主观和客观因素相结合,给讨论类学习行为赋予合理的权重,以解决该特征下有效数据量不足的问题.

1.2.2 特征处理

类别特征,即所属学院、专业、班级等具有能够在有限类别内取值的特征.本研究采用序列编码将类别特征转换成可靠的数值特征,比如:将文本序列“大一、大二、大三、大四”对应转化为数值序列“0,1,2,3”.

与上课时间固定不变的课堂教学不同,在线学习的学生可以根据自己的学习习惯调整相应学习的时间.因此,本研究将每门课程的常规学习时间设置为3个指标:周学习时间、日学习时间及是否为工作日学习时间,具体如下:

1) 周学习时间:一周分为7天,该指标反映学生更偏向于在一周中的哪一天学习.

2) 日学习时间:把一天分为8:00—12:00,13:00—17:00,18:00—22:00和其他时间4个时间片段,该指标衡量学生每天的学习习惯;

3) 是否为工作日学习:把学习时间划分为周末和工作日,该指标衡量学生对工作日和周末学习的偏好程度.

为了衡量学生学习行为及基本信息对学习效果的影响程度,采用Spearman相关系数,根据原始数据的排序位置,求解数据间的相关性,

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (1)$$

其中, r_s 表示Spearman相关系数; $d_i = X - Y$,表示 X, Y 之间的等级差, X 表示某一个学生学习行为或基本信息的指标集合, Y 表示学习效果指标集合; n 表示样本的总量.此外,集合中某个数的等级代表将它所在的一列数按照从小到大排序后所在的位置.

1.2.3 模型构建

由于学习行为数据间不存在线性可分的关系,难以找到合适的回归模型对样本数据进行精确的分数量预测,本研究通过将学习效果划分为4个类别进行预测,将学习效果预测问题转化为分类预测问题.未确定数据分布之前需对以下6种基本的分类算法进行比较,选择最优的算法构建模型:

表1 学习效果划分标准

考试得分	学习效果
90~100分	优异
80~89分	良好
70~79分	合格
0~69分	不合格

1) K近邻(KNN)^[13]:根据距离新样本 s 最近的 k 个样本点的类别来确定新样本 s 所属的类别.

2) 朴素贝叶斯^[14]:假设样本特征均独立的情况下,利用贝叶斯公式计算各类别的概率,选取概率最高的类别作为该样本的类别.

3) 决策树^[15]:从树的根节点开始,将样本数据与决策树中的特征节点进行比较,根据判断结果选择下一层的分支继续比较,最终的叶子节点即为分类结果.

4) 支持向量机(SVM)^[16]:寻求一个能够划分数据集且几何间隔最大的超平面,利用核技巧解决非线性分类问题.

5) 随机森林^[17]:假设单个决策树模型需要训练的特征个数为 n ,决策树每次分裂时,通过比较信息增益、信息增益比及基尼指数3个指标值得到最好的特征进行分裂,循环该分裂过程,直到节点的所有训练样本成为同一类,最终的分类结果由多棵决策树投票得到.

6) Adaboost^[18]:Adaboost算法是集成学习中的一种迭代算法,对同一个训练集训练不同的分类器,即弱分类器,然后将它们整合,形成一个更强的分类器,即最终分类器.

1.2.4 评估方法

本节介绍分类模型的评价指标:准确率 A 、精准率 P 、召回率 R 和F1分数 F .准确率指分类器正确预测的样本数量占被预测样本总数量的比例;精准率指分类器中所有被预测为正的样本中,实际为正的样本的比例,该指标仅代表对正样本结果中的预测准确程度;召回率指分类器中实际为正的样本中同时预测也为正样本的比例;F1分数是平衡召回率和精准率两者之间差异的一个指标,该指标的值越大越好.计算公式如下:

$$A = \frac{T_p + T_N}{T_p + T_N + F_p + F_N}, \quad (2)$$

$$P = \frac{T_p}{T_p + F_p}, \quad (3)$$

$$R = \frac{T_p}{T_p + F_N}, \quad (4)$$

$$F = \frac{2 \times P \times R}{P + R}, \quad (5)$$

其中, T_p 表示被正确预测学习效果的正样本数量; T_N 表示被正确预测学习效果的负样本数量; F_p 表示被错误预测为正样本的数量; F_N 表示被错误预测为负样本的数量.

2 实验设计与结果分析

2.1 数据概览

样本选用S大学学生在线教学平台上的学习行为数据,数据选取的时间跨度为2017年2月19日—2020年7月8日,由日志数据转化得到的统计型学习行为记录共计54 720条,其中,1条样本数据代表1个学生在某1门课程上的学习情况记录.

首先对数据进行统计分析.根据视频完成率的均值(0.32)较低可知,大部分学生未能看完整个视频.浏览时长和文档学习时长的均值分别为13.86 h及37.98 h,明显低于视频学习总时长的均值(144.63 h),可以推测学生在线上学习时更偏向于利用视频进行学习,不包含学生将资料下载后自行学习的时间.讨论模块的发帖、回帖、点赞数的使用均值分别为0.63,1.39及2.63次,均较低,说明该部分学生的使用需求不强烈,导致研究无法统计到可用性强的讨论学习行为指标.作业完成率和测试完成率均值分别为0.42和0.39,可以推测学生进行在线学习时的自觉性不强.

除17个学习行为特征外,本研究还设置了1个由考试得分划分的学习效果标签、6个基本信息类别特征和3个通过常规学习时间生成的学习习惯类别特征,具体的类别数量如表2所示.

表2 类别特征的类别数量

类别特征	类别数量	类别特征	类别数量
学院	18	学年(届)	8
专业	107	日学习时间	4
专业班	825	周学习时间	7
课程	673	是否为工作日学习	2
教学班	1 388	学习效果	4

2.2 学习效果与不同指标的相关性分析

本部分采用Spearman相关系数探究学习行为指标与学习效果之间的相关性.实验结果表明:学生的访问数和资料浏览数、资料浏览时长有极高的正相关性(r 值均为1),由此推测,学生登录网上教学平台,都会查看并下载学习资料;发帖、回帖、点赞的行为和视频学习总时长、测试平均分、作业平均分等具有较高的正相关性(r 值为0.52~0.72),由此可知,平时作业完成情况可以体现学生的学习态度,说明常常参与线上讨论的学生学习态度相对较好.

但是,学习效果与各特征的相关性都较低(r 值低于或等于0.28),其中,文档学习总时长和学习效果存在一定的正相关性($r=0.28$),而学院、专业、专业班、课程、教学班及学年这6个类别特征与学习效果的相关性 r 值均低于0.1,所以用传统的分析方法无法构建学习效果的预测模型,只能用机器学习的方法挖掘特征之间的内在联系和隐含信息.

2.3 学习效果预测模型构建

在构建学习效果预测模型的过程中,各算法参数均设置为sklearn机器学习库中的初始数据,经标准化处理,将训练集和测试集的比例划分为7:3,如表3所示.

表3 分类模型的预测结果

模型	准确率	精准率	召回率	F1分数
KNN	0.476 4	0.510 1	0.471 0	0.489 8
朴素贝叶斯(Gaussian)	0.374 9	0.329 5	0.342 3	0.335 8
朴素贝叶斯(Bernoulli)	0.309 5	0.306 0	0.305 9	0.305 9
决策树(Gini)	0.789 3	0.786 1	0.786 8	0.786 4
决策树(Entropy)	0.765 2	0.762 7	0.762 1	0.762 4
支持向量机	0.725 9	0.725 3	0.720 3	0.722 8
随机森林	0.846 9	0.847 9	0.841 6	0.844 7
Adaboost	0.580 0	0.639 9	0.568 9	0.602 3

从表3可知,随机森林算法的分类效果最佳,准确率、精准率、召回率和F1分数分别为84.69%、84.79%、84.16%和84.47%.由此可知,即使特征变量和目标变量之间的相关性不高,随机森林算法也可以捕捉到特征内部对目标变量的隐含信息,以此提高分类的准确率.因此,建议以随机森林算法作为学习效果预测模型的基本算法.

3 结 论

本研究将学生在在线教学平台中产生的行为日志数据整理成17个有意义的在线学习行为指标,并且将这些指标与学生的学习习惯指标、基本信息指标相结合,利用机器学习探究学生在线学习行为与学习效果之间的关系,通过相关性分析,找出学习行为中最能影响学习效果的指标为文档学习总时长.此外,随机森林算法能够捕捉到各特征内部对学习效果的隐含信息,相比较其他传统机器学习方法,更适用于完成对学习效果预测的工作,最终对数据集的分类准确率达到84.69%.此外,本研究的数据集是S大学在线教学平台中产生的日志数据,数据的选取时间跨越了疫情期和非疫情期,数据真实性得到了保

证.虽然本研究已经得到了较好的实验效果,但是基于日志数据能够生成的学习行为指标远不止研究中提及的17个,未来将更进一步探究学习行为指标的生成,以期为平台的建设和提升线上学习的质量提供更优、更完善的建议.

参考文献:

- [1] XIE Y R, QIU Y, HUANG Y L, et al. Characteristics, problems and innovations of online teaching of “No Suspension of Classes” during the period of epidemic prevention and control [J]. E-Education Research, 2020, 41(3): 20–28.
- [2] LIU S N Y, YANG Z K, LI Q. Education data ethic: the new challenge of education in big data era [J]. Educational Research, 2017(4): 15–20.
- [3] GONG Z W. An Empirical study on the influencing factors of adult students’ online learning behavior [J]. China Electronic Education, 2004(8): 32–35.
- [4] BAO W. A trait analysis of learning behavior of university students in china since the enrolment expansion [J]. Tsinghua Journal of Education, 2009, 30(1): 78–87.
- [5] YAO Q H, WANG J, LI Y B, et al. Investigation and research on the online learning of college students [J]. Electrical Education Research, 2010(7): 57–60, 68.
- [6] ZHAO H Q, JIANG Q, ZHAO W, et al. Empirical research of predictive factors and intervention countermeasures of online learning performance on big data-based learning analytics [J]. E-Education Research, 2017, 38(1): 62–69.
- [7] LUO H, ROBINSON A C, JAE-YOUNG P. Peer grading in a mooc: reliability, validity, and perceived effects [J]. Journal of Asynchronous Learning Networks, 2016, 18(2): 1–14.
- [8] FAKIR M, TOUYA K. Mining students’ learning behavior in Moodle system [J]. Journal of Information Technology Research, 2014, 7(4): 12–26.
- [9] MOU Z J, WU F T. The Exploration of learning outcome prediction indicators and analysis of learning group characteristics for MOOC [J]. Modern Distance Education Research, 2017(3): 58–66, 93.
- [10] ZHANG M, YUAN H. The PauTa criterion and rejecting the abnormal value [J]. Journal of Zhengzhou University (Engineering Science), 1997(1): 87–91.
- [11] XIE X Y, LIU S Y. Discussion-based teaching: the practical application of seminar teaching mode to undergraduate teaching [J]. Theory and Practice of Education, 2013, 33(33): 50–52.
- [12] VAIDYA O S, KUMAR S. Analytic hierarchy process: an overview of applications [J]. European Journal of Operational Research, 2006, 169(1): 1–29.
- [13] GUO G, WANG H, BELL D, et al. KNN model-based approach in classification [C]// OTM Confederated International Conferences on the Move to Meaningful Internet Systems. Berlin: Springer, 2003: 986–996.
- [14] MURPHY K P. Naive Bayes classifiers [J]. University of British Columbia, 2006, 18(60): 1–8.
- [15] SAFAVIAN S R, LANDGREBE D. A survey of decision tree classifier methodology [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3): 660–674.
- [16] SMOLA A J, SCHÖLKOPF B. A tutorial on support vector regression [J]. Statistics and Computing, 2004, 14(3): 199–222.
- [17] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5–32.
- [18] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1): 119–139.

(责任编辑:包震宇,郁慧)