在线学习成效影响因素分析与研究*

李封 陈默 张昱 赵长宽

东北大学计算机科学与工程学院，沈阳 110819

摘 要 近年来，MOOC教学、直播教学、混合式教学等多种教学模式层出不穷。这些新型的教学模式基本都以在线学习为主，与传统的课堂教学相比，学习者学习环境发生了巨大的改变。因此，有关新教学模式下学习者学习成效影响因素的讨论重新成为了现阶段的研究热点。在此背景下，本文依托调查问卷采集学生的基本信息、学习环境、学习态度、以及在线学习行为等多维度的数据，通过数据分析的方法，尝试找到影响学习者学习成效的主要因素。研究以某高校2020年度某个学期中，理工科专业29个自然班超过800名本科生为研究样本，结合该学期这些学生高级语言程序设计课程的期末考试成绩，利用数据挖掘方法，对学生学习成效的影响因素就行了分析与研究。最终得到在线学习表现类特征对学习者学习成效影响最为显著的结论，同时利用特征选择方法给出各个特征的重要性分数。

关键词：在线学习，学习环境，特征选择，学习行为

Analysis and Research on Factors Influencing the Effectiveness of Online Learning

Feng LI Mo CHENG Yu ZHANG Changkuan ZHAO

College of Computer Science and Engineering of Northeastern University
Shenyang, 110819, china

Abstract— Nowadays, various teaching modes such as MOOC, live teaching, and blending teaching have emerged one after another. These new teaching models are mostly based on online learning, and compared to traditional classroom teaching, the learning environment for learners has undergone significant changes. Therefore, the discussion on the factors affecting the learning effect of learners under the new teaching mode has once again become a research hotspot at this stage. The paper relies on a survey questionnaire to collect data on students' basic information, learning environment, learning attitude, and online learning behavior. Through data analysis methods, it attempts to identify the main factors that affect the learning effect of learners. This study takes over 800 undergraduate students from 29 natural classes in a certain semester of 2020. They are from science and engineering majors in a certain university. Based on the final exam scores in the advanced language programming course of that semester, data mining methods are used to analyze and study the factors affecting their learning effect. The conclusion that online learning performance features have the most significant impact on learner learning outcomes is ultimately obtained, and the importance scores of each feature are given using feature selection methods.

Keywords— Online learning, Learning environment, Feature selection, Learning behavior

1 引言

在政策的指导下，随着互联网技术的迅速发展，学习者获取知识的方式也必然发生了巨大的变化。MOOC 教学、SPOC 教学、混合式教学已不再是新型的教学方式。空间与时间也不再是限制教与学的主要条件。2020 年起，一场突如其来的疫情席卷了全球，更衍生出一种以直播为主的教学方式。在线学习，已经和传统的课堂教学一样，成为学生的学习常态。

从教育技术和教学研究方面来说，线上学习与传统的课堂教学最主要的区别就是学习环境的变化，对学生线上学习成效影响因素分析与研究从理论上讲有助于在线学习环境相关理论的构建以及学习者学习投入度理论的构建。从实用性角度出发，本文以某工科院校 29 个自然班学生为样本，依托调查问卷结果，利用数据挖掘方法对学习者线上成效的影响因素就行了分析，为学习者学习行为的分析和研究工作的开展提供了可靠的数据支撑，更加方便了“教”与“学”方式的调整，以达到提高学习者学习效果的目的。

*基金资助：全国高等院校计算机基础教育研究会计算机基础教育教学研究项目“后疫情时代程序设计类课程中学习者学习行为探究”（编号：2023-AFCEC-186）；全国高等院校计算机基础教育研究会计算机基础教育教学研究项目“面向计算思维培养的《数据科学基础课程》PBL 教学模式研究与实践”（编号：2023-AFCEC-184）。

2 研究意义与研究现状

2.1 研究意义与目的

在我国，在线学习这种特点鲜明的学习模式，作为传统学习方式的一种有效的补充，数年间已经进行了大量的尝试和应用。今天，随着各种各样的情况层出不穷，为了保障学习者的身体健康，同时也为了从更多的维度获取知识。传统的课堂教学往往与直播教学、线上 MOOC 等方式相结合。大学生所需要面临的学习模式包括正常校园教学、居家式在线教学、寝室自主在线学习等多种模式。无论是学习环境、教学组织结构，还是师生之间的关系，又或者是家人与学生之间的关系较之之前的基本教学方式都发生了显著的变化。这些改变无疑给学习者的学习生活带来了不同，同时也为广大的研究人员对学习者学习行为的分析和研究提出了新的挑战。

在不同的学习环境下，找到影响学习者学习成效的主要因素，找到如何更加有效的开展自主学习的方法，已经成为了当今学习型社会中一个比较重要的研究方向，也是未来一段时间内研究者们所关心的主要研究内容。通过对这些主要因素的分析与研究，广大的教育工作者可以及时的调整教学手段、改进教学过程、提高教学效率，从而找到更加有效的教学方式。

2.2 国内外研究现状

对于学习成效的研究，理论上讲就是对不同学习者学习行为与学习成绩的关系进行分析，实际上“学习分析”并不是一个新兴的词语，在 2010 年“学习分析”首次被写进《地平线报告》[5]，这预示着对于学习者的学习行为的研究已经成为了在线教育的主要技术手段。2011 年 Tanya Elias 在其报告中指出[6]，学习分析是一个新兴的领域，就是通过精准的分析来改变教与学。而学习分析与人工智能、网络分析、行为分析密切相连。2012 年 10 月，美国教育部发布了相关报告《Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics》[7]，报告中明确了教育领域的大数据的应用主要在两个方面，其一是教育数据的挖掘，另一方面的应用则是学习分析，很明确的指出了利用数据挖掘的方法进行学习者学习行为的分析是及其重要的研究方向。

对于不同环境下的线上学习的研究，由于其新生事物本身的特点，以及所处的背景环境不同，国外的相关研究比较缺乏。而在国内，大部分学者的研究主要集中多环境下学习的特点的研究、不同环境中学习体验的调查以及学习效果的调查三个方面。例如，在学习特点的研究方面，谢幼如等人指出，在居家学习中其学习环境，教学组织结构，教学设计等都发生了巨大的变化，其学习特点与面对面的课堂教学有着很大的区别[8]。文献[9]同时指出了，如果学习为居家在线学习是在一种完全的物理隔离下进行的教学活动，无论对教师还是学习者都是一种全新的体验，也势必要结合新形势展开新的思考。

而从学习体验的角度，李艳等人以大学生的个体为样本，对学习者的在线学习体验进行了调查研究，得到了在线学习中最大的收获是自学能力的提高，最大的挑战是自我约束能力的不足这一结论[10]。汪卫平和李文则以高校为单位，对处于不同地域的学习者的在线学习体验进行调查研究，得出在我国境内学习者的学习体验由东向西呈下降的趋势。而技术平台，社会性交互与环境因素则是造成这种结果的重要影响因素[11]。

从在线学习效果影响因素的角度，主要包括在线学习平台，学习者和教师的交互以及课程设计等相关的影响因素。例如，在文献[12]中，Van Wart 等人以美国加州州立大学圣贝纳迪诺分校的近 1000 名学习者为研究对象，对学习者在线学习成功的层次因素进行了建模。得出包括教学支持、教学存在感、认知存在感以及在线交互模式等特征均对学习者的在线学习成效有着显著的影响。Naddeo 等人的研究则发现硬件设备、网络连接等因素同样对学习者的在线学习成效有着较明显的影响[13]。

综上所述，现有的有关学习者在线学习影响因素的研究，一部分只是对 MOOC 教学中学生线上行为进行分析，没有真正考虑学生所处的实际学习环境。另一部分以调查问卷的方式对学习者主观感知的定性的分析，并没有研究能够系统的给出学习环境以及学习者学习行为等特征数据与学习成效之间定量的相关关系。本文依托调查问卷的结果，采用数据挖掘的方法，试图找到在不同学习环境下影响大学生在线学习成效的主要因素。

3 研究思路及方案

在确定了研究的意义和目的后，根据“提出问题”、“分析问题”、“解决问题”的解决问题的具体步骤，本文对研究过程进行了整体的规划。

首先“提出问题”，本研究针对当前时代背景下，高校普遍采用的以传统教学为主线，在线学习为支撑的新型的教学方式，试图找到在新的教学方式下，尤其是不同的学习环境中，各种因素对学习者学习成效的影响程度。其次，对于上述问题进行分析，综合考虑国内外的研究现状，本研究拟以发布调查问卷的方式获取学习者的相关特征数据，利用数据挖掘的方法从这些数据中找到与学习者学习成效相关的特征，并定量的给出主要的特征与学习成效之间的相关系数，

为更好的开展新型教学方式提供理论依据。具体如图 1 所示：

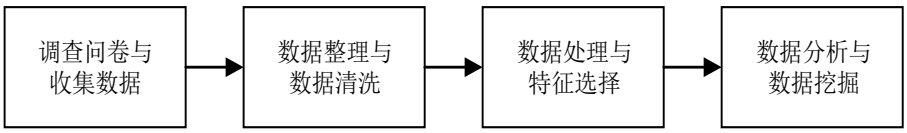

图1 研究思路示意图

4 数据分析与实验

本研究于 2020 年春季学期，面向某 985 高校理工科专业 29 个自然班的程序设计类课程中约 800 名同学发放相关的调查问卷，问卷的具体内容以及详细的研究方案如下：

4.1 研究样本

本研究所采用的数据样本来源于调查问卷，为了使调查问卷的具体内容更有针对性，研究者提前采用了基于小样本的先验随机问卷，问卷以学习者学习的基本情况以及学习者对在线学习的认知情况进行调查，以便为后续的调查问卷的设计提供依据，如表 1 所示。

表 1 先验问卷

序号	题 目
1	你对自己本科的学习生涯有没有规划？
2	你对在线学习有没有了解？
3	你认为在线学习的重要性如何？
4	如果大学的课程改为开放式的在线教学你会怎么看？

根据随机调查的结果显示，大部分的学习者对大学学业的认知不足，同时学生缺乏对在线学习的了解。因此，本研究拟从学习者的基本信息，是否有着相关课程的学习基础，学习者对本专业的满意度，学习者对在线学习的认知情况，在线学习的学习表现以及在线学习时学习者所处的学习环境这 6 个方面设计整理调查问卷的题目。

2020 年春季学期，寒假过后研究样本所在高校的学生处于两种状态，部分同学未能及时返校在家，部分同学未离开学校或提前返校在校。为方便学习者进行学习，该校教师普遍采用腾讯课堂、腾讯会议或者钉钉等直播会议软件进行直播教学，同时使用中国大学 MOOC，智慧树等在线学习平台开展异步在线教学。在此背景下，本研究于 2020 年 4 月上旬通过雨课堂为 2020 年春季学期参加高级语言程序设计课程学习的 29 个自然班共计发放调查问卷 807 份。问卷包含 6 个方面 22 个问题，具体内容如表 2 所示。

表 2 调查问卷内容

调查方向	问卷题目
基本信息	性别
	年龄
	所学专业
相关课程基础	是否学习过相关计算机课程
	是否学习过相关数学课程
	相关课程成绩
本专业认知和满意度	是否会从事本专业相关工作
	对本专业喜爱程度
	对本专业了解程度
对在线学习的认知	是否愿意支付线上学习费用
	觉得线上学习效果如何
	进行线上学习的动力是什么
	MOOC中哪部分最重要
	线上教学中教师的角色是否重要
在线学习表现	线上学习有无迟到早退
	线上学习专注度如何
	观看线上学习有无拖动进度条
	线上提交作业是否会参考其它人答案
	是否积极回答老师线上的提问
学习环境	是否为自己营造适合学习的氛围
	线上学习硬件条件如何
	线上学习周围环境如何

在这些问题中，研究者加入了有关学习环境的讨论，包括进行在线学习时周围环境是否有干扰、网络是否畅通、以及是否有良好的学习氛围。

本研究中，调查问卷的总计发放了 807 份，有效回收问卷数量为 584 份，参与调查者的基本情况如图 2 所示。本文选择其中有效完成度在 50%以上的问卷作为研究样本，总计为 511 份。对于有效样本而言，平均完成时间为 7 分 21 秒，平均完成度为 98%，完成时间为 2020 年 4 月 6 日至 2020 年 4 月 10 日之间。

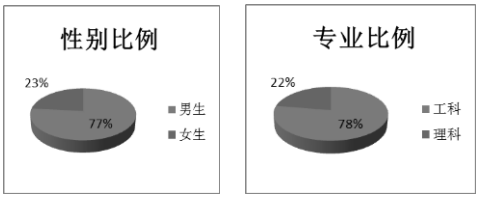

图 2 参与调查者基本情况

4.2 数据与特征的选择

根据上述调查问卷的具体题目，本研究拟将数据特征分为 6 大类，共计 22 个特征，其中涉及到调查者对某种事情综合态度的问题，本文拟采用类似于李克特量表的一种四级量表进行统计，例如“对本专业的喜爱程度”这个问题，其答案选项就设计为“非常喜爱”、“喜爱”、“一般”和“比较厌恶”，分别记为“4-1”的分值来表示。由于网络和编码等原因，原始数据样本中，存在少量属性数值为 NaN，因此需要对此观察矩阵进行清洗和整理，去除异常值。

下一步，本文对分类特征进行预处理，例如“性别”、“所学专业”等。所有属性特征不能简单地表示为数字，因为如果简单描述成数字的话，模型会将特征解释为有序的，但实际上属性特征是无序的，所以本文采用独热编码来表示。这样，具有含有 N 个可能值的特征可以转换为 N 个二进制特征，并且只有其中一个是有效的。例如，“性别”特征则转换为“性别-男”和“性别-女”，如果“性别-男”等于“1”，则“性别-女”等于“0”。

最后，考虑当时所在学期的实际情况，大部分课的授课难度以及测试难度普遍降低，为了更加准确的对学生进行分类，本文拟将学习者的学习成绩分为 2 类，低水平：区间包括 0 至 89 的值；高水平：区间包含 90 至 100 的值。利用“成绩-L”以及“成绩-H”来描述学习水平的分类，并以此为依据对学习者的学习表现进行预测。按照此思路，原始数据样本整理得到一个 511*44 的观察矩阵，共计 510 个样本，43 个属性特征。

本文以此数据为研究对象，对学习者的学习成绩进行预测，展开居家学习下影响学习者学习成效主要因素的研究。

4.3 实验与结果

实验环境的选择，本文的主要实验测试开展在一台购置于 2018 年的个人 pc 机，该电脑的核心硬件设置如下：CPU 为 AMD FX(tm)-8300 Eight-Core Processor，其主频为 3.30GHz；内存为金斯顿 DDR4，双条 16GB；硬盘为西部数据机械硬盘，2T 容量。系统配置，计算机使用的操作系统是旗舰版 Windows7 Service Pack1，系统类型是 64 位操作系统，使用 Matlab2017b 作为主要实验环境。

为实现本文的研究目标，我们展开具体的实验，内容主要包括两个部分：第一部分，使用不同的类别特征对学习者的学习成效进行预测，试图通过分类的准确率找到影响最大的特征类别；第二部分，尝试通过特征选择的方法给出各个属性特征对学习者学习成效影响力的排序。

在分类算法的选择方面，从以目的为驱动的思想出发，研究的主旨是为了查找影响力最大的特征类别，而无需过多纠结分类的准确率的高低，所以本文选择使用常见的支持向量机算法进行分类学习。

支持向量机，其英文为 Support Vector Machine（SVM），是一种有监督的学习方法，将向量映射到一个更高维的空间里，并在此空间中生成一个使得分类数据间隔最大化超平面[14]。假设给定了特征空间中的一个训练数据集，具体如公式 1 所示。

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \quad (1)$$

其中，$y_i \in \{1, -1\}(i = 1, 2, \ldots, n)$，$x_i \in R^n$，$x_i$为输入，$y_i$为$x_i$对应的标签，$n$为样本个数，$(x_i, y_i)$为样本点。当$y_i = 1$时，称$x_i$为正例；当$y_i = -1$时，称$x_i$为负例。那么 SVM 算法的学习目标就是在此特征空间中找到分类超平面，使得该超平面与正反两样本之间的距离越大越好，即实现分类间隔最大化。样本量中等偏小的情况也有较好的效果，有一点泛化能力和鲁棒性。

为提高分类的准确率，实验采用交叉验证的方法，设定了 k 个子集（其中 k=5），每个子集均做一次测试集，其余的作为训练集。交叉验证重复 k 次，每次选择一个子集作为测试集，并将 k 次的平均交叉验证识别正确率作为结果。由此通过计算，可以得到分类的准确率为 72.7%，具体的混淆矩阵如图 3 所示。同时，本文在实验中采用了不同类别的数据特征单独作为分类依据，具体的预测结果如表 3 所示。

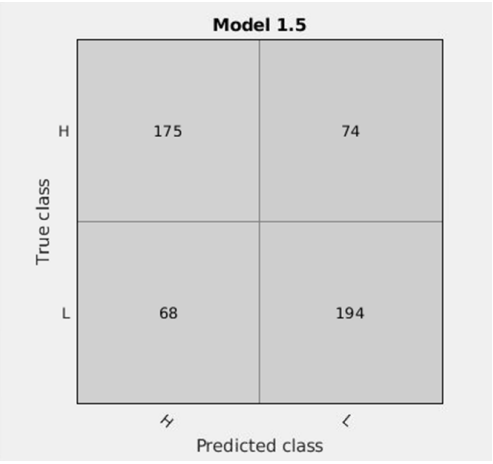

图 3 分类结果混淆矩阵

从分类结果中我们不难看出，数据整体的分类准确率并不是很高，其主要原因可能在于采用调查问卷式的数据采集方法带来的常见问题，用户在回答调查问卷题目时的专注度不够，造成数据偏差，以及调查问卷的题目设置不够准确，特征的区分度不足等原因。

表 3 预测结果

评价 / 特征类别	准确率	精确率	召回率
基本信息	53%	53.8%	25.3%
课程基础	63.2%	64.1%	55.8%
专业认知	49.7%	47.1%	26.1%
学习认知	57.1%	57.6%	42.6%
线上表现	69.5%	69.8%	65.9%
学习环境	58.5%	56.5%	64.3%

此外，表 3 的表明，当采用不同类别的特征对学习者进行分类时，准确率最高的为学习者的“线上表现”类特征，其分类的准确率达到了 69.5%；排在第 2 位的为“课程基础”类特征，其分类的准确率达到了 63.2%。另外“学习环境”和“学习认知”类特征的分类准确率也接近 60%。另外两类特征的分类准确率在 50%左右，实际上对学习者学习成效的分类区分度不是很显著。

为了更进一步查找影响学习者学习成效的具体因素，本文采用卡方检验的方法，计算特征与分类的相关程度。所谓的卡方检验[15]就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，若卡方值越大，两个值之间的偏差程度就越大；反之，二者偏差相对越小；当实际观测值与理论值完全相等时，其卡方值就为 0，表明理论值完全符合。卡方检验应用于特征选择时，如果卡方值为 0，则代表特征与分类完全不相关，换言之，该特征与分类的相关性为 0，不在特征子集内。因此，本文在进行特征选择时，采用如下方法计算卡方值，并以此用来描述特征与分类的关联程度。假设度量特征 t，与分类 c 之间，符合具有一阶自由度的卡方分布，那么 t 与 c 的χ^2值，可以由公式 2 来计算：

$$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (2)$$

在上式子中，χ^2为卡方的值，N 为样本的数量，A 表示含有特征 t 且属于分类 c 的样本数，B 表示包含特征 t 且不属于分类 c 的样本数，C 表示属于分类 c 但是不包含特征 t 的样本数，D 表示既不属于分类 c 也不包含特征 t 的样本数量。当卡方值越大，特征 t 与分类 c 相关性越强。

本文在Matlab中使用fsschi2函数对特征的重要性分数即卡方值进行计算，并记录分数排序后的索引，其中重要性分数排在前七位的特征依次为“线上学习的专注度”、“线上迟到早退情况”、“线上学习参与度”、“相关课程的成绩”、“线上学习硬件条件”、“线上学习学习环境”以及“是否愿意支付线上学习费用”，其重要性分值在5以上，具体分数如图4所示。

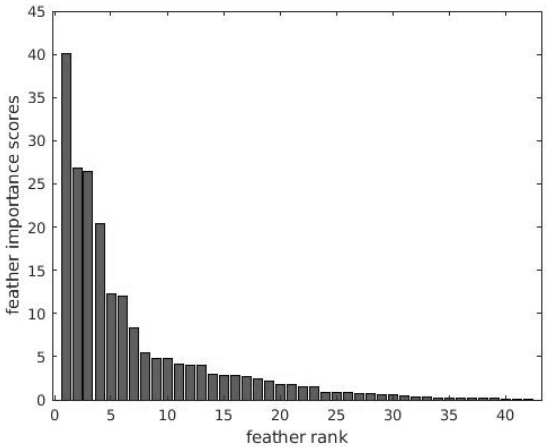

图 4 特征重要性分数示意图

根据实验结果所示，我们不难看出，在以线上学习为主的教学过程中，对学习者学习成效产生较大影响的主要因素可归结为学习者的线上学习态度问题，此外是否有一个良好的学习环境对学习成效也有着一定的影响力，这无疑为广大的教育工作者在新型的教学方式下更好的开展教学活动提供了理论依据和支撑。

5 结束语

本文从各高校出现的新型的教学方式出发，利用发放调查问卷的方法获取了不同环境下学习者在线学习时的相关学习行为数据，给出了一个从不同角度出发的六类的特征分类方法，并利用机器学习的算法计算了采用不同分类特征对学习者学习成效进行分类预测的准确率，得到了“线上表现”类特征的预测准确率最高，也就说明了此类特征是影响大学生居家学习效果最重要的特征类别。同时选择基于卡方检验的方法对数据集中所有特征对学习者学习成效的影响力分值进行计算，并给出了特征的影响力在前七位的分值排序，并对这些特征与学习成效之间的关系进行了简单诠释。由此可见，这种结果表明了线上学习环境下，学生的学习行为中自主性、互动性等因素对学习效果具有显著影响。具体来说，在线上学习环境中，学生能够根据自己的需要选择学习资源，并按照适合自己的方式进行学习，这种个性化学习方式是影响学习效果的重要因素。因此，学生的自主学习能力，包括如

何合理安排学习时间、筛选和分辨信息、运用网络资源进行有效学习，都直接影响学习效果。对于教育工作者而言，关注并优化大学生的“线上表现”，对于提升居家学习效果，提升线上教学成效具有重要意义，这些特征可以作为广大教育工作者在日后展开教学工作的依据，也是教学过程中需要重点关注的方向。

本文的不足，从预测结果来看的其准确率不是很高，造成这样现象的原因可能有三个：第一，用户调查问卷填写时专注度不足，很多答案的选择并没有反应真实情况；第二，调查问卷的题目设置不够准确，造成特征的区分度不足。第三，某些因素可能起到的是间接的影响，例如，“学习环境”可能会对“学习专注度”带来影响。下一步，课题组拟计划采集学习者的在线学习行为数据，并重新研究调查问卷的题目设置，将采集的数据与调查问卷数据相结合，从而提高数据的可信程度以期达到更优的分类效果。

参 考 文 献

[1] 国务院. 关于积极推进“互联网+”行动的指导意见[EB/OL].2015. https://www.gov.cn/zhengce/content/2015-07/04/content_10002.htm

[2] 国务院. 国家教育事业发展“十三五”规划[EB/OL].2017. https://www.gov.cn/zhengce/content/2017-01/19/content_5161341.htm

[3] 教育部. 教育信息化“十三五”规划[EB/OL].2016. http://www.moe.gov.cn/srcsite/A16/s3342/201606/t20160622_269367.html

[4] 教育部. 教育信息化 2.0 行动计划[EB/OL].2018. http://www.moe.gov.cn/srcsite/A16/s3342/201804/t20180425_334188.html

[5] 杨梦柯.高等教育领域技术应用热点综述—基于2010~2017 年地平线报告的分析[J]. 软件导刊(教育技术), 2017,16(9):87-89.

[6] Elias T. Learning Analytics: Definitions, Processes and Potential [EB/OL]. 2011-03. https://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

[7] 徐鹏,王以宁,刘艳华,张海.大数据视角分析学习变革——美国《通过教育数据挖掘和学习分析促进教与学》报告解读及启示[J].远程教育杂志,2013,31(06):11-17.

[8] 谢幼如,邱艺,黄瑜玲,王芹磊.疫情防控期间“停课不停学”在线教学方式的特征、问题与创新.电化教育研究,2020,41(03),20-28.

[9] 刘莉萍,孙杰.自得之教：疫情防控期间在线教学新思维.天津师范大学学报（社会科学版）,2020,05,14-18.

[10] 李艳,陈新亚,陈逸煊,张帆. 疫情期间大学生在线学习调查与启示--以浙江大学竺可桢学院为例.开放教育研究,2020,26(05),60-70.

[11] 汪卫平,李文. 中国大学生在线学习体验的区域差异及影响因素--基于国内 334 所高校调查数据的分析.开放教育研究,2020,26(06),89-99.

[12] Van Wart, Montgomery Medina, Pamel Canelon, et al. Integrating Students’ Perspectives about Online Learning: A Hierarchy of Factors. International Journal of Educational Technology in Higher Education[J], 2020, 17(1), 53.

[13] Naddeo A., Califano R., Fiorillo I.. Identifying factors that influenced wellbeing and learning effectiveness during the sudden transition into elearning due to the COVID-19 lockdown. Work-a Journal of Prevention Assessment and Rehabilitation[J], 2021, 68(1), 45-67.

[14] Huang Y, Zhao L. Review on landslide susceptibility mapping using support vector machines[J]. Catena, 2018, 165:520-529.

[15] 郭雨萌,李国正. 一种多标记数据的过滤式特征选择框架[J].智能系统学报, 2014, 9(03):292-297.