MOOCs学习行为与学习效果的逻辑回归分析*

□ 宗 阳 孙洪涛 张亨国 郑勤华 陈 丽

【摘 要】

目前，MOOCs（大规模开放在线课程）在世界范围内迅猛发展，但是随之而来的是对MOOCs学习质量和高辍学率等现象的质疑。现有MOOCs平台大都对学习者在线学习行为有较为详细的记录。如何对学习行为数据进行分析、建模和解读是大数据时代教育研究的热点和难点所在。逻辑回归方法作为一种成熟的机器学习方法可以有效地建立学习行为和学习效果之间的模型。本研究总结了在线学习领域逻辑回归研究的流程，在此基础上，从MOOCs在线学习过程出发构建了学习行为指标，并应用逻辑回归对MOOCs学习数据进行分析，就学习者在线学习行为对学习成绩的影响展开了探索。研究检验了逻辑回归对于在线学习效果研究的价值，发现了课程注册时滞、登录课程次数、提交作业测试次数、习题保存次数的均值和视频观看完成度等指标与成绩的相关性。研究发现：在该课程中提交作业测试可以作为MOOCs学习成绩预测的关键指标，所构建的逻辑回归模型预测准确率达到98%。

【关键词】 MOOCs；逻辑回归；在线学习行为；学习效果

【中图分类号】 G420 【文献标识码】 A 【文章编号】1009—458 x (2016)05—0014—09

DOI:10.13541/j.cnki.chinade.20160527.002

一、引 言

近年来，MOOCs的快速发展使其教学效果受到越来越多的关注。MOOCs学习的规模效应，使得在MOOCs中难以开展个性化的教学。有研究对目前部分高校主流MOOCs平台的数据进行统计表明，虽然课程完成率有达到40%的情况，但大部分课程完成率不到10%（Jordan, K., 2013）。大规模的学习者注册似乎意味着大规模的辍学率和未通过率，提高MOOCs学习者的学习效果是当前包括MOOCs在内的在线教育面临的重大难题。大量研究通过对不同环境中在线学习行为与学习效果的关系进行实证研究，发现学习者的在线行为对学习效果有着重要影响（F. Kizilcec, 2013; 姜蔺, 韩锡斌, 2013; 宏梅, 2008; 吕媛, 2004）。

已有关于在线行为与学习效果的关系研究中，在线行为特征的获取大多是基于学习者学习过程中的单一维度或某几项维度，如学习者的注册时间，什么时间与何种课程资源交互，如何交互以及交互的程度，练习次数、成绩、错误率、错误的内容和学习成绩等（Macfadyen, 2010; RaMesh, 2013; Balakrishnan, 2013; 蒋卓轩, 2014）。关于学习行为与学习效果之间关系的研究，由于在线学习过程的复杂性，相关研究所得出的结论也不尽相同。例如王萍（2015）的研究表明学习者观看视频数和学习章节数等参与行为与学习成绩没有直接关系，而贾积有等（2014）的研究表明观看视频次数、观看网页次数、浏览和下载讲义次数等学习行为与学习成绩呈显著相关。

MOOCs学习者的学习效果受到诸多因素的影响。为了全面深入地探索学习效果的影响因素，研究者需要对反映整体学习过程的大量数据进行挖掘和分析。本研究从MOOCs学习者整体学习过程出发，构

* 本文系北京师范大学自主科研基金项目“学习者在线学习状态分析与可视化工具研发”课题成果，获得中央高校基本科研业务费专项资金资助。

建MOOCs学习者学习行为分析框架，通过对一门实际MOOC中行为数据进行分析，应用逻辑回归方法分析MOOCs学习者的影响因素。

二、在线学习领域的逻辑回归研究

（一）逻辑回归及相关研究

逻辑回归（Logistic Regression, LR）是数据挖掘和机器学习的常见方法之一，属于有监督的学习方法。它根据一个或多个连续型或离散型自变量来分析和预测离散型因变量的广义线性回归。逻辑回归的因变量通常为类别等离散变量。二元逻辑回归是最常用形式之一，其因变量只包含两个类别值。在线学习分析中，常常会遇到一些表示研究对象状态的离散变量，例如学习者参与课程后能否取得好的学习效果，获得相应证书，考试能否及格得到相应学分等。在在线学习环境中，学习者的学习行为体现在在线学习的各个方面，可以使用二元逻辑回归方法分析学习行为对学习效果的影响。

在在线学习领域中，国外已经有很多研究通过建立逻辑回归模型来对学习者的学习表现等进行分析和预测。例如Harrell Ⅱ和Bower（2011）选取了学习者的三个特征（听觉学习风格、计算机技能和成绩平均积点），通过逻辑回归分析确定模型，预测基于学习社区的学习者是否会辍学；San等人（2013）通过智能引导系统搜集学习者在初中阶段的学习过程中所表现的学习投入及情感特征来预测其是否能上大学；Park和Choi（2009）从个体特征、家庭社会因素和心理三方面探究影响成人学习者在线课程辍学率的因素，并预测学习者成功的可能性。

与国外相比，国内关于在线学习领域应用逻辑回归的相关研究较少。较有代表性的研究为蒋卓轩等（2014）运用逻辑回归等方法，通过看视频次数、提交测验次数、记录密度、论坛发帖次数、论坛看帖次数和注册时间距离开课日期的天数6个行为特征，对学习者的最后学习成果进行了预测。

文献研究发现，在线教育领域中的逻辑回归的相关研究大多针对标志学习过程或结果的某一重要变量作为因变量（如是否辍学、是否考试及格等），分析各种自变量（如特征变量、行为变量和心理学变量等）与因变量的相关关系，最终实现分类和预测的目的。

（二）在线学习领域逻辑回归研究的流程

在借鉴上述研究的基础上，本研究对在线学习领域逻辑回归研究的一般流程进行了梳理。

1. 变量选择

变量选择是逻辑回归的第一个步骤。变量要满足自变量与因变量的密切相关，以及各个自变量之间相互独立的两个条件。为保证变量选择的合理和有效，需要对变量进行完整的预处理和相关性分析。

从原始数据集抽取变量时，需要对指标进行数据预处理。通过预处理对变量的缺失值和异常值等进行处理，剔除不符合要求的数据。为了有效地建立逻辑回归模型，需要对变量进行相关分析，应尽可能地将显著相关的自变量选入建模过程。例如，Park和Choi（2009）选取了性别、年龄、教育程度、家庭支持、组织支持、学习者满意度与课程关联度等变量研究成人学习者在线课程辍学率，并利用相关分析法进一步分析了研究选择的变量，发现性别、年龄和教育程度等人口学特征与学习者的辍学与否在统计学意义上并不相关，因此剔除了这3个变量，将家庭支持、组织支持、学习者满意度与课程关联度这4个显著相关的自变量放入逻辑回归的模型中。

2. 逻辑回归建模

确定进入逻辑回归的变量后，需要将数据样本按一定的比例随机分成训练集和验证集，每次实验用训练集训练参数，用验证集验证预测精度。例如Harrell Ⅱ和Bower（2011）在225份有效样本中随机选择了116条数据（51.6%），用于初步的逐步逻辑回归，余下的109条数据用于验证模型的预测精度。

逻辑回归建模具体可细分为向前引入法、向后剔除法和逐步回归法，三种方法各有优劣。San等人（2013）选用初中生的知识量、习题正确率、投入度、粗心、无聊、困惑、开小差等9个特征变量采用向后剔除的方法进行逻辑回归建模预测学习者是否能升入大学。

就在线教育逻辑回归研究而言，对模型的教育意义进行分析是关键环节。模型计算是一个客观过程，但其初步结果不一定具有合理的教育学解释。当模型和远程教育既有研究和实践有差异时，需要

对数据和建模过程进行反复分析和验证，甚至引入其他数据分析与挖掘方法辅助分析，才能确定最终的结论。

3. 模型应用效果评价

研究建立的逻辑回归模型可以通过一系列指标进行评价。常见的评价指标包括正确率、错误率、灵敏性和特效性等，以及一些综合性判断指标，如ROC曲线、KS值和Lift值等。最直观有效的评价指标是模型的预测准确率。此外，ROC曲线通过曲线下的面积（AUC分数）来表征模型准确度。面积越大的模型对应的模型准确度越高。如Harrell Ⅱ和Bower（2011）通过ROC曲线下面的面积（AUC分数为0.617）进一步评估3个变量模型的有效性，表明所选取的变量能够有效预测学习者在线社区活动的持久性。

三、MOOCs学习行为分析指标

为了分析学习行为与学习效果之间的关系，需要对MOOCs的学习过程进行解析。在现有研究中，学习行为大多基于“注册—听课—课堂随测—作业—讨论—考试—结业—证书”（孙立会, 2014）的基本流程。贾积有等（2014）通过Coursera平台上6门MOOCs中学习者的在线行为数据，分析学习行为数据及其与成绩的关系。结果表明：成绩与开始学习时间呈显著负相关，与在线时间、观看视频次数、观看网页次数、浏览和下载讲义次数、平时测验成绩之和、论坛参与程度（发帖、回帖）6个指标呈显著正相关。王萍（2015）基于edX平台开放数据对学习者进行研究，选取注册课程时间、最后登录时间、课程交互次数、课程访问天数、播放视频次数、学习章节数和论坛发帖数来探索中外MOOCs学习者的学习行为和特征。研究发现，获得证书的学习者一般浏览了较多的课程章节内容，但在视频观看上，成绩较高的学习者也没有显著的视频访问增加行为。蒋卓轩等（2014）通过分析挖掘北京大学在Coursera平台上6门MOOCs中学习者的在线行为数据，选择了观看视频次数、提交测验次数、记录密度、论坛发帖次数、论坛看帖次数、注册时间距离开课日期的天数6个与学习成绩有影响且课程共有的特征对学习者的成绩进行预测，得到了较高的预测准确率。

通过对已有研究的分析发现，其选取的学习行为在完整性方面存在一定不足，可以建立更为完整的学习行为分析指标，表征在线学习过程。在已有研究的基础上，结合学习者实际在线学习过程，将MOOCs学习者在线学习过程归纳为学前准备、登录平台、资源学习、交流讨论和作业考核等阶段，并根据数据情况构建了18个MOOCs学习行为指标，具体指标及指标编码如表1所示。

表1 MOOCs学习行为指标

在线学习过程	分析维度	行为指标	指标编码
一、准备与登录	1.1 学前准备	1.1.1 浏览课程详情页次数	1L_IV
		1.1.2 注册课程时滞	2L_TL
	1.2 出勤	1.2.1 登录课程次数	3L_LC
二、资源学习	2.1 视频以外其他资源学习情况	2.1.1 访问课程其他学习资源的次数	4R_WV
	2.2 视频资源学习情况	2.2.1 每次登录观看视频时长	5R_TS
		2.2.2 视频观看完成度	6R_VF
		2.2.3 视频观看密度	7R_VD
	2.3 视频资源学习坚持情况	2.3.1 重复观看视频的次数	8R_VR
		2.3.2 视频重复观看程度	9R_RF
		2.3.3 提交作业测试后，观看所对应视频资源的次数	10R_AV
三、论坛交互	3.1 交互参与情况	3.1.1 论坛发帖数	11F_PC
		3.1.2 论坛回帖数	12F_RC
		3.1.3 论坛浏览次数	13F_VC
四、作业测试	4.1 学习任务完成量	4.1.1 提交作业测试次数	14T_QC
		4.1.2 习题保存次数的均值	15T_TS
	4.2 完成学习任务的积极性	4.2.1 提交作业测试与发布时间差	16T_DA
		4.2.2 作业测试提交密度	17T_AD
		4.2.3 提交作业测试时间间隔	18T_DC

（一）准备与登录

MOOCs学习者登录过程可以分为注册课程前和注册课程后两个过程，对应到二级分析维度上主要包括学前准备和出勤两个方面。学前准备通过学习者注册课程前浏览课程详细页次数和注册课程时滞两个指标来测量表征；出勤通过学习者登录平台课程的次数来表征。

（二）资源学习

视频是xMOOCs中最为重要的学习资源，观看视频是此类MOOCs最为重要的学习方式（郑勤华, 2015）。xMOOCs中视频以外学习资源较少，可以汇总成一类。因此，资源学习将资源分为视频和视频以外的学习资源两种。通过视频资源学习情况、视频资源学习坚持情况和视频以外其他资源学习情况三个维度进行学习行为分析。视频资源学习情况进一步细分

为每次登录观看视频时长、视频观看完成度和视频观看密度三个指标来表征。其中，视频完成度（R_VF）体现学习者课程资源学习完成情况，计算方式为学习者观看学习视频总时长除以课程所有学习视频总时长；视频观看密度（R_VD）体现了学习者课程资源学习的集中程度，计算方式为视频观看次数除以最后一次看视频与首次观看视频的时间差。视频资源学习坚持情况用重复观看视频的次数、视频重复观看程度和提交作业测试后观看所对应视频资源的次数三个指标进行表征。其中，视频重复观看程度（R_RF）体现学习者资源重复学习的程度，计算方式为重复观看的视频数除以视频观看数。

（三）论坛交互

在MOOCs学习的过程中，学习者可以根据需要与教师和其他学习者在论坛中进行交互。针对该过程选取了学习者在论坛参与的交互情况，用学习者的论坛发帖数、论坛回帖数和论坛浏览次数三个指标进行表征。

（四）作业测试

在MOOCs中评价手段相对比较单一，对学习者的学习评价大都通过作业和测试进行。本研究选取了学习任务完成量和完成积极性两个维度进行分析。用提交作业测试次数和习题保存次数的均值来表征学习任务的完成量；用提交作业测试与发布时间差、作业测试提交密度和提交作业测试时间间隔来表征完成学习任务的积极性。其中，作业测试提交密度（T_AD）体现学习者进行作业测试的集中程度，计算方式为作业提交次数除以最后一次提交作业测试与首次提交作业测试的时间差。

四、案例研究

本研究选取365大学平台上一门MOOC中的学习行为数据，采用逻辑回归方法对学习者学习成绩进行分析。该MOOC开课时间为2015年10月01日，结束时间为2016年01月16日。在开课期间共有512人参与学习。该MOOC共有42个教学视频，课程评价采用章节作业、测试和期末测试的形式，共有12个课后单元作业测试和一个期末考试测试。该案例中用学习者课程成绩表征学习者的学习效果，将成绩合格与否作为因变量，MOOCs学习行为的18个指标作为自变量，按照在线学习领域逻辑回归研究流程对学习者学习效果进行预测。

（一）变量选择

1. 数据预处理

本研究18个预设学习行为指标涉及次数、时间间隔、时长、均值、比率等多类数据。绝大部分测量指标需要通过算法对相关数据表原始数据进行计算后获得数据。首先，根据预设指标意义和原始数据库表结构编写获取指标数据的算法，然后根据算法编写SQL函数及存储过程获取学习者样本在18个指标上的数据值。将指标数据首先进行缺失值和异常值处理，剔除缺失样例和缺失数据较多的指标变量。MOOC中论坛交互较低是较为普遍的现象，本研究的MOOC论坛中仅有9条帖子，经过分析发现，论坛帖子内容均是关于考试的评论与咨询，与学习者学习效果无关，因此，将论坛交互的三个指标论坛发帖数（F_PC）、论坛回帖数（F_RC）和论坛浏览次数（F_VC）剔除。在变量缺失值和0值统计分析中发现，在案例MOOC中访问课程其他学习资源的次数（R_WV）、重复观看视频次数（R_VR）、视频重复观看程度（R_RF）、提交作业测试后观看对应视频资源的次数（R_AV）和提交作业测试时间间隔（T_DC）5个指标的缺失值和0值所占比例均超过了72%。因此，将上述5个指标变量剔除，剩下指标变量进入相关性分析步骤。

2. 相关性分析

将预处理后剩余的10个指标与学习者成绩使用SPSS Statistics 20进行Pearson相关分析，结果如表2所示。可以看出，浏览课程详情页面次数（L_IV）和视频观看密度（R_VD）两个指标与学习成绩没有显著相关性；每次登录观看视频时长（R_TS）和作业测试提交密度（T_AD）虽然与学习成绩在0.01水平上显著相关，但是相关系数均<0.2，即基本与学习成绩无关；在剩下的6个指标中，从相关系数可以看出，提交测试作业次数（T_QC）以及提交习题与发布习题时间差的均值（T_DA）和习题保存次数的均值（T_TS）之间均值在0.01水平上显著相关，并且相关性为.772和.629，因此T_QC与T_DA和T_TS之间存在较强的共线性。研究发现，提交作业测试次数（T_QC）与成绩之间的相关性为.971，而本研究MOOC学习

者成绩最后是由测试作业成绩按一定比例权重累加给定，学习者提交作业测试次数和学习成绩高相关性的现象表明学习者只要提交了作业测试就会有好的成绩，这可能与该课程考核形式相对简单有关。基于共线性关系，本研究决定剔除提交测试作业次数（T_QC）指标。综上所述，共有5个指标，即课程注册时滞（L_TL）、登录课程次数（L_LC）、提交习题测试时间与发布时间差的均值（T_DA）、习题保存次数的均值（T_TS）和视频观看完成度（R_VF）进入初步逻辑回归建模过程。

（二）逻辑回归建模

1. 初步模型构建

本研究采用逻辑回归分析中的二元逻辑回归模型，探讨在MOOCs中学习者学习合格的发生概率。假设P为学习者学习合格的发生概率，其取值范围为[0，1]，(1−P)即为不合格的概率。$P/(1-P)$为学习合格逻辑回归发生比，对其取自然对数为$\ln[P/(1-P)]$。

假设自变量为X_1，X_2，…，X_k，因变量为P，则逻辑线性回归函数方程可表示为：

$$\ln[P/(1-P)]= B_0+B_1*X_1+B_2*X_2+\cdots+B_k*X_k \quad (1)$$

上式中$B_i(i=0,1,2,\cdots,k)$为逻辑回归系数。

根据(1)式，可得学习者学习合格发生概率为：

$$P= e^{(B0+B1*X1+B2*X2+\cdots+Bk*Xk)}/[1+e^{(B0+B1*X1+B2*X2+\cdots+Bk*Xk)}] \quad (2)$$

利用(2)式便可以计算MOOCs中学习者学习合格的发生概率。

在本案例中，将学习者成绩转化为合格（分数>=60）和不合格（分数<60）二元因变量，上述5个指标作为自变量，进行逻辑回归的结果如表3所示。

可以看出自变量与因变量之间具有较强的关联强度，Hosmer和Lemeshow检验结果达到显著，说明该模型适配度较差。5个指标中只有登录课程次数（L_LC）可以预测解释学习者成绩合格与否。该模型预测分类正确率结果如表4所示。

可以看出，该模型预测学习者学习合格的正确率达到99.5%，预测学习者学习不合格的正确率达到75.9%，整体预测正确率为93%。

上述模型预测准确率较高，但是模型适配度不佳。通过进一步分析发现，视频观看完成度（R_VF）与学习成绩合格的相关性系数最高为1.144，但是却没有达到显著水平，并且习题保存次数的均值（T_TS）与学习成绩出现负相关系数为−.085。本研究案例课程学习资源基本上全部为视频，但是视频观看完成度（R_VF）与学习成绩未达到显著相关。为了深入分析这一现象，笔者对相关数据进行了进一步分析。

表2 相关性分析结果

		grade	L_IV	L_TL	L_LC	R_TS	R_VD	T_QC	T_TS	T_DA	T_AD	R_VF
grade	Pearson 相关性	1	.032	−.294**	.482**	.190**	−.066	.971**	.591**	.731**	.172**	.340**
	显著性(双侧)		.470	.000	.000	.000	.138	.000	.000	.000	.000	.000
	N	512	512	512	512	512	512	512	512	512	512	512
L_IV	Pearson 相关性	.032	1	.106*	−.017	.091*	−.013	.009	−.038	.056	.189**	.046
	显著性(双侧)	.470		.016	.694	.040	.771	.841	.391	.205	.000	.296
	N	512	512	512	512	512	512	512	512	512	512	512
L_TL	Pearson 相关性	−.294**	.106*	1	−.394**	−.225**	.087	−.284**	−.310**	.047	.118**	−.335**
	显著性(双侧)	.000	.016		.000	.000	.050	.000	.000	.290	.008	.000
	N	512	512	512	512	512	512	512	512	512	512	512
L_LC	Pearson 相关性	.482**	−.017	−.394**	1	.075	−.095*	.488**	.484**	.274**	−.193**	.470**
	显著性(双侧)	.000	.694	.000		.089	.032	.000	.000	.000	.000	.000
	N	512	512	512	512	512	512	512	512	512	512	512
R_TS	Pearson 相关性	.190**	.091*	−.225**	.075	1	−.111*	.219**	.141**	.080	.022	.732**
	显著性(双侧)	.000	.040	.000	.089		.012	.000	.001	.072	.621	.000
	N	512	512	512	512	512	512	512	512	512	512	512
R_VD	Pearson 相关性	−.066	−.013	.087	−.095*	−.111*	1	−.080	−.115**	−.040	−.021	−.130**
	显著性(双侧)	.138	.771	.050	.032	.012		.069	.009	.368	.641	.003
	N	512	512	512	512	512	512	512	512	512	512	512
*.在.01水平(双侧)上显著相关												
**.在0.05水平(双侧)上显著相关												

表3 整体模型的适配度检验及个别参数显著性的检验摘要表

	B	S.E,	Wals	df	Sig.	Exp (B)	关联强度
L_TL	.000	.000	16.690	1	.000	1.000	Cox & Snell R 方=.523 Nagelkerke R 方=.756
L_LC	.103	.039	7.120	1	.008	1.109	
R_VF	1.144	.870	1.729	1	.189	3.138	
T_TS	−.085	.197	.185	1	.667	.919	
T_DA	.000	.000	38.593	1	.000	1.000	
常量	−1.247	.643	3.760	1	.052	.287	
整体模型适配度检验	卡方=378.901*** Hosmer 和 Lemeshow 检验卡方=22.854，Sig.=0.004 ***P<0.001						

表4 预测分类正确率交叉表

	已观测		已预测		
			pass		百分比校正
			0	1	
步骤 1	pass	0	107	34	75.9
		1	2	369	99.5
	总计百分比				93.0
a. 切割值为 .500					

2. 聚类分析基础上的模型构建

基于上述推测，利用SPSS Modeler 14.2将学习者的视频观看完成度（R_VF）和提交测试作业次数（T_QC）与学习成绩进行K-means聚类，当K=3时达到较好的聚类效果，平均轮廓=0.8，聚类结果如图1所示。

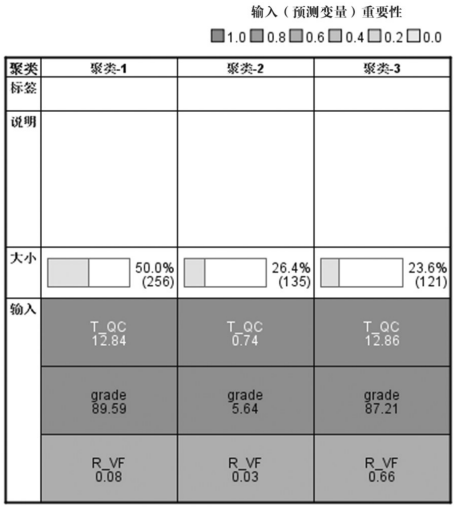

图1 聚类结果

从聚类结果可以看出，聚类-1这个类别占据50%，平均成绩89.59，处在较高的水平；而视频完成度平均为0.08，处于较低的水平。这说明有一半的学习者未观看视频，仅提交了作业，并取得了较好的成绩。本研究按照上述分类结果将学习者样本分为两组，一组是基本不观看视频直接提交测试作业取得高分的学习者共256个（称为异常组），剩余部分样本共256个（称为正常组）。在上一步的基础上剔除与所有学习者学习成绩显著相关提交测试作业次数（T_QC）指标后重新对异常组和正常组分别进行逻辑回归，结果如表5和表6所示。

正常组逻辑回归预测学习者学习成绩合格与否正确率如表7所示。

通过对比上述两组逻辑回归结果可以发现，异常组没有行为指标可以预测解释成绩合格与否，自变量和因变量之间的关联强度非常低，这进一步验证了本研究的推测，该组学习者可能仅为拿到学分，采取只提交测试作业而不观看视频资源的学习方式。正常组逻辑回归发现视频观看完成度（R_VF）和登录课程次数（L_LC）可以有效预测解释学习成绩合格与否，自变量与因变量之间的关联强度很高，模型适配度较好，模型预测学习者学习合格的正确率达到98.3%，预测学习者学习不合格的正确率达到97.8%，整体预测正确率为98%。

3. 逻辑回归方程

通过初步模型构建，依据表3和公式(1)、公式(2)可得逻辑回归线性方程模型为

$$\ln[P/(1-P)]=0.103*L\_LC-1.247 \quad (3)$$

学习者学习成绩合格发生概率方程模型为

$$P=e^{(0.103*L\_LC-1.247)}/[1+e^{(0.103*L\_LC-1.247)}] \quad (4)$$

剔除异常学习者样本后，可以应用视频观看完成度（R_VF）和登录课程次数（L_LC）两个指标学习成绩合格与否进行有效预测解释，根据表6和公式（1）、公式（2），可得逻辑回归线性方程模型为

$$\ln[P/(1-P)]=0.167*L\_LC+10.404*R\_VF-7.848 \quad (5)$$

学习者学习成绩合格发生概率方程模型为

$$P=e^{(0.167*L\_LC+10.404*R\_VF-7.848)}/[1+e^{(0.167*L\_LC+10.404*R\_VF-7.848)}] \quad (6)$$

（三）模型应用效果评价

本研究采用ROC曲线和AUC值来对案例MOOC形成的预测模型进行效果评价。对于MOOCs中所有学习者的学习成绩合格预测模型，预测结果ROC曲线如图2所示，曲线下的面积（AUC）值为0.943，可以看出该模型虽然整体适配度不佳，但仍

表5 异常组模型的适配度检验及参数显著性表

	B	S.E,	Wals	df	Sig.	Exp (B)	关联强度
L_TL	.000	.000	.728	1	.394	1.000	Cox & Snell R 方=.006 Nagelkerke R 方=.039
L_LC	.024	.084	.082	1	.775	1.024	
T_DA	−3.224	5.109	.398	1	.528	.040	
T_TS	−.277	.609	.207	1	.649	.758	
R_VF	.000	.000	.095	1	.758	1.000	
常量	5.434	2.991	3.300	1	.069	229.072	
整体模型适配度检验	卡方=1.487 Hosmer 和 Lemeshow 检验卡方=8.449，Sig.=0.391						

表6 正常组逻辑回归结果

	B	S.E,	Wals	df	Sig.	Exp (B)	关联强度
L_TL	.000	.000	3.158	1	.076	1.000	Cox & Snell R 方=.708 Nagelkerke R 方=.945
L_LC	.167	.076	4.794	1	.029	1.182	
T_DA	.000	.000	2.022	1	.155	1.000	
T_TS	.918	.489	3.525	1	.060	2.505	
R_VF	10.404	2.364	19.366	1	.000	32978.867	
常量	−7.848	2.283	11.823	1	.001	.000	
整体模型适配度检验	卡方=315.033*** Hosmer 和 Lemeshow 检验卡方=7.141，Sig.=0.521 ***P<0.001						

表7 预测分类表

	已观测		已预测		
			pass		百分比校正
			0	1	
步骤 1	pass	0	134	3	97.8
		1	2	117	98.3
	总计百分比				98.0
a. 切割值为 .500					

然具有非常好的预测效果。对于正常组学习者学习成绩合格预测模型，预测结果ROC曲线如图3所示，曲线下的面积（AUC）值为0.994，非常接近1，说明该模型几乎是完美的预测模型。结合表3和表6所有学习者和正常组学习者学习成绩合格逻辑回归结果可以看出，正常组学习成绩预测模型在关联强度、整体模型适配度和ROC曲线AUC值上比所有学习者预测模型均有较为明显的提升，并且模型预测准确率由93%提高到98%。

五、讨论与总结

通过案例研究，一方面发现了在线学习行为与学习成绩之间的关联性，另一方面也验证了逻辑回归方法在远程教育中的实践价值。

（一）在线学习行为与学习成绩显著相关

研究发现了多个学习行为指标与学习效果显著相关，包括课程注册时滞（L_TL，相关性−.294）、登录课程次数（L_LC，相关性.482）、提交作业测试次数（T_QC，相关性.971）、习题保存次数的均值（T_TS，相关性.591）和视频观看完成度（R_VF，相关性.340）。上述5个与学习成绩显著相关的指标变量分布于在线学习的准备与登录、资源学习和作业测试三个维度，说明MOOCs学习者学习成绩与在线学习行为的密切关系。

（二）提交作业测试是预测MOOCs学习成绩的关键指标

在所选的MOOC中，学习行为指标中与学习者成绩相关性最高的指标是提交作业测试次数（T_QC），相关性为.971。在获得学分的学习者中，提交作业测试次数与学习成绩显著正相关。这一现象一方面反映了作业和测试在评价中的有效性，在作业测试方面投入更多精力的学生获得了更好的成绩；另一方面，可能与当前MOOCs作业测试设计相对简单有关，一定程度上反映了当前MOOCs中形成性评价机制的问题。同时，不排除有学习者通过多次试错获得答案的可能性。

（三）应用逻辑回归可较有效地预测学习效果

本研究中应用逻辑回归的方法对所有学习者和正常组学习者学习效果进行预测，均取得了较好的预测效果。两个预测模型预测正确率均在93%以上，模型应用ROC曲线下的面积（AUC）均高于0.9，表明应用逻辑回归方法基于学习行为对学习效

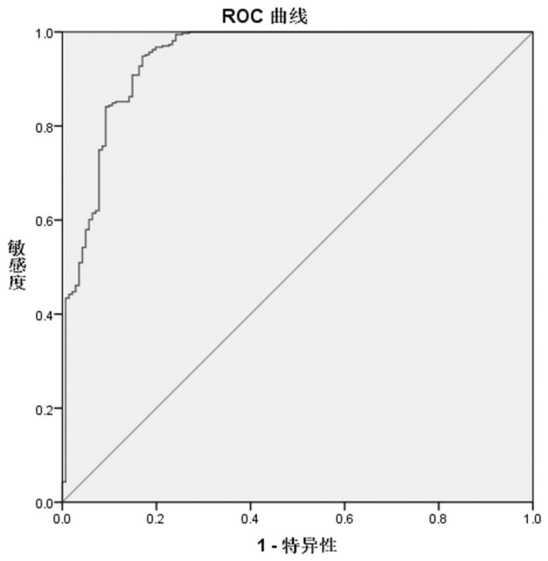

图2 所有学习者预测模型ROC曲线

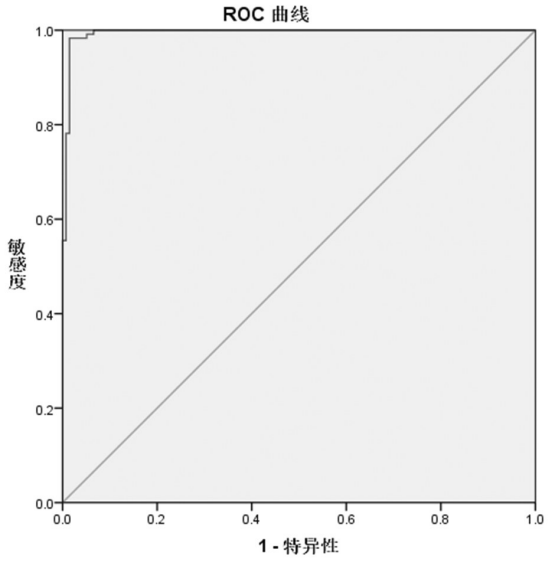

图3 正常组学习者预测模型ROC曲线

果进行预测有着重要的实践价值，可以通过逻辑回归方法分析学习行为且较为有效地预测MOOCs学习效果。

（四）案例课程存在两种典型学习模式

通过相关分析发现，学习者视频观看完成度（R_VF）与成绩相关性为.340，而在初步逻辑回归建模过程中出现与学习成绩相关但不显著，并且习题保存次数均值（T_TS）出现负相关的异常情况。进而通过聚类研究发现，有50%的学习者视频观看完成度很低，但是提交了作业测试，并取得较高的学习成绩。在对异常组和正常组学习者分别进行逻辑回归中发现了行为指标对这两类学习者的学习效果预测有着很大差异。正常组学习成绩预测模型具有很高的准确率，但异常组没有发现能有效解释预测学习成绩的指标。

这说明了案例课程中存在两种典型模式：一类学习者通过正常的学习流程，先进行资源学习，进而完成作业和测试；另一类学习者则不通过资源学习直接提交作业测试，获得成绩。后者的成因有待深入研究，可能存在其他学习方式替代了在线学习过程。学习模式的差异直接导致了学习行为的差异。这表明在应用逻辑回归方法建模前，通过无监督机器方法对学习者进行类别划分将对模型的有效性起到重要作用。

综上所述，逻辑回归方法作为一种有监督的机器学习方法在学习分析领域有着重要意义，通过逻辑回归可以较为有效地预测MOOCs学习者的学习效果。学习行为是预测在线学习效果的重要依据。但在模型构建过程中，需要将有效的在线学习行为甄别出来，以此为依据构建的模型才更为真实、可信。为了实现这个目标，研究者往往需要将多种数据分析与挖掘方法综合应用，并通过在线教育专家对分析结果不断进行深入解读。领域专家的知识与数据分析挖掘方法的有机整合是在线教育领域中基于数据研究的质量保障。

[**参考文献**]

[1] BALAKRISHNAN G. Predicting student retention in massive open online courses using hidden markov models, UCB/EECS-2013-109 [R/OL].Berkeley:University of California, Berkeley, 2013.

[2] Ivan L. Harrell II & Beverly L. Bower (2011) Student Characteristics That Predict Persistence in Community College Online Courses, American Journal of Distance Education, 25:3, 178- 191, DOI: 10.1080/08923647.2011.590107

[3] Jordan, K. MOOC Completion Rates: The Data[EB/OL]. http://www.katyjordan.com/MOOCproject.html,2013-09-22.

[4] MACFAYDEN L P,DAWSON S.Minging LMS data to develop an "Early Waring" system for educators:a proof of concept[J]. Computers & Education,2010, 54(2):588-599.

[5] Park, J.-H., & Choi, H. J. (2009). Factors Influencing Adult Learners' Decision to Drop Out or Persist in Online Learning.Educational Technology & Society, 12 (4), 207-217.

[6] RAMESH A, GOLDWASSER D, HUANG B, et al,. Modeling learner engagement in MOOCs using probabilistic soft logic[C/OL]. //

NIPS Workshop on Data Driven Educatio, 2013[2014-06-0]. http://lytics. stanford.edu/datadriveneducation/papers/

[7] San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. Proceedings of the 6th International Conference on Educational Data Mining, 177-184. ramshtal. pdf.

[8] 宏梅,刘满贵,杨隽. 学习行为与学习效果的相关调查之研究——网络多媒体教学模式下学习者英语自主学习调查与研究[J]. 大学英语(学术版),2008,(02):145-148.

[9] 贾积有,缪静敏,汪琼. MOOC学习行为及效果的大数据分析——以北大6门MOOC为例[J]. 工业和信息化教育,2014,(09):23-29.

[10] 姜蔺,韩锡斌,程建钢. MOOCs学习者特征及学习效果分析研究[J]. 中国电化教育,2013,(11):54-59、65.

[11] 蒋卓轩,张岩,李晓明. 基于MOOC数据的学习行为分析与预测[J]. 计算机研究与发展,2015,(03):614-628.

[12] 吕媛,易银沙,邓昶,易尚辉. 网络行为对学习者学习成绩和心理健康状况的影响[J]. 中国学校卫生,2004,(02):250-251.

[13] 孙立会. 开放教育基本特征的变迁——兼议MOOC之本源性问题[J]. 远程教育杂志,2014,(02):30-38.

[14] 田娜,陈明选. 网络教学平台学习者学习行为聚类分析[J]. 中国远程教育,2014,(11):38-41.

[15] 王萍. 大规模在线开放课程的新发展与应用:从cMOOC到xMOOC[J]. 现代远程教育研究,2013,(03):13-19.

[16] 薛薇. SPSS统计分析方法及应用[M]. 北京:电子工业出版社,2013.

[17] 郑勤华,李秋菊,陈丽. 中国MOOCs教学模式调查研究[J]. 开放教育研究,2015,(06):71-79.

收稿日期:2016-03-16

定稿日期:2016-04-13

作者简介:宗阳,在读硕士;张亨国,在读硕士;陈丽,博士,教授,博士生导师。北京师范大学远程教育研究中心(100875)。

孙洪涛,博士,高级工程师,中央民族大学现代教育技术部(100081)。

郑勤华,博士,副教授,北京师范大学教育学部(100875)。

责任编辑　韩世梅

Possibilities and challenges of augmented learning

Rebecca Ferguson

Digital technologies are becoming cheaper, more powerful and more widely used in daily life. At the same time, opportunities are increasing for making use of them to augment learning by extending learners' interactions with and perceptions of their environment. Augmented learning can make use of augmented reality and virtual reality, as well as a range of technologies that extend human awareness. This paper introduces some of the possibilities opened up by augmented learning and examines one area in which they are currently being employed: the use of virtual realities and tools to augment formal learning. It considers the elements of social presence that are employed when augmenting learning in this way, and discusses different approaches to augmentation.

Keywords: augmented learning; augmented reality; social presence; virtual field trips; virtual reality; virtual tools

A logistic regression analysis of learning behaviors and learning outcomes in MOOCs

Yang Zong, Hongtao Sun, Hengguo Zhang, Qinhua Zheng and Li Chen

The rapid growth of Massive Open Online Courses (MOOC) on a global scale has brought with them doubts about learning quality and concerns about their high attrition rates. Generally speaking, MOOC platforms keep a detailed record of learners' online learning behaviors. Analysis, modeling and interpretation of these data is top on the educational research agenda in the era of big data.

Logistic regression analysis is used in this study to explore the impact of online learning behaviors on learning outcomes. Findings from the study indicate that learning outcomes are correlated with delay in course registration, how many times that learners log in and submit their assignments/quizzes, the average number of times they save their exercises, and the extent to which video clips are watched. It is also found that assignment/quiz submissions can serve as a key index in predicting MOOC learning outcomes and that the resultant logistic regression model has 98% prediction accuracy.

Keywords: Massive Open Online Course (MOOC); logistic regression; online learning behavior; learning outcome

Failure to follow learning rules in blended teaching: possible causes and recommended solutions

Hang Shu, Fan Wang and Lu Yuan

Proper learning rules are key to ensuring the implementation of blended learning, the success of which is directly affected by what rules are formulated and how students accept and respond to them. Findings from this study indicate that students showed strong resistance to the preset learning rules, which led to their perfunctory participation in online discussion. To be specific, in order to pass the course, students were found to contribute a large number of posts which involved no meaningful interaction as expected in online discussion. Causes of this phenomenon are identified, including information overload, imbalance of cognitive surplus, subjective evaluation mechanism, and unhealthy competition. Based on the findings, the article suggests that teachers make timely informed adjustments to their