DOI:10.16227/j.cnki.tycs.2021.0078

基于学习分析技术的学习预警系统研究与设计

■ 龚艺[1,2],杨娟[1],纪娟[1,2]

(1.四川广播电视大学,四川 成都 610073;
2.国家开放大学教育信息管理与信息系统研究中心,四川 成都 610073)

[摘要]学习预警是提高远程学习者学习质量的重要手段。文章基于学习分析技术,分析了远程学习者在线学习行为对学习成绩的影响因素,构建了学习成绩预测模型,利用 R 语言工具,以国家开放大学远程学习者在线学习行为和期末成绩作为实验数据,利用朴素贝叶斯、决策树、前馈反向传播神经网络、支持向量机四种数据挖掘中的分类算法,通过学习者在线学习行为数据对学习者学习成绩进行预测,并利用 ROC 曲线等方式对上述分类算法预测结果进行对比分析,结果表明,朴素贝叶斯算法准确度高,是整体性能表现优秀的成绩预测算法。最后依据学习分析技术框架,在学习成绩预测模型的基础上,设计了远程学习预警系统,系统根据远程教育学习者在线学习行为,及时发现存在危机的学习者,为学习预警系统的实现提供了参考依据。

[关键词]学习分析技术 数据分类 学习预警 成绩预测 远程教育

[中图分类号] G647 [文献标识码]A [文章编号]1673-0046(2021)2-0053-04

随着大数据时代的到来,现代远程教育正在发生着深刻的变革。在线学习成为现代远程教育的重要学习形式,在线学习迅速发展的同时也存在诸多问题,如学习质量和效率低下、教师的个性化及适应性教学能力差、在线学习监控管理和评价不及时等[1]。有效地监控学习者学习情况、及时为学习者提供个性化支持服务,有助于提高学习者满意度、保证远程教育学习者学习质量并降低学生流失率。通过在线学习平台中记录的大量学习行为数据挖掘有效的信息,构建学习预警机制,既可为教学管理提供决策参考,又可为学习者提供帮助和指导[2]。学习预警是在线学习监控管理、个性化支持服务的重要环节。

国外的一些机构已经开发了预警系统,例如普渡大学利用学习分析技术分析学生学习情况及特点,从而预测存在学习危机的学生[3]。Desire2Learn 通过分析学习成绩、文档和工具使用以及参与社会化学习程度等数据,辨别“危险”中的学生并进行跟踪干预,提供合适的帮助[4][5]。学习分析技术的发展,为远程教育学习预警系统的构建提供了新的视角,学习分析技术在教育中的应用使人们对学习发生和发展的认证更加明晰,对学习的监控和预警更加直观和便捷[6]。

美国高等教育信息化协会最早将“学习分析”定义为:“使用数据和模型预测学生的收获和具有处理这些信息能力的行为。”[7]学者何克抗[8]将学习分析技术定义为利用各种数据收集和数据分析工具,从海量数据中,通过收集、测量、分析和报告等方式,提取出有价值的各种信息从而为教与学以及教学管理提供辅助决策的技术。学习分析技术注重监测和预测学生的学习成绩,即发现潜在问题,为教学过程提出有针对性的改进策略和教育策略[9]。远程教育学习预警系统,旨在通过提取在线学习平台中学习者产生的大量行为数据,例如学习者的学习次数和频率、学习时长、测评完成情况、使用学习资源的数量、参与学习讨论的次数等。根据学习分析技术,对其上述大量数据进行采集、处理和分析,深入了解学习者在线学习状态,为教学管理者提供数据分析以帮助其了解海量学习者的学习情况,及时发现存在学习“危机”的学习者,以便提供有针对性的学习支持服务、人工干预、个性化教学服务等,为最终构建个性化、智能化的在线学习系统提供实现基础。

本研究根据学习分析技术,构建远程教育学习者预警系统,建立预警系统框架,并解决了预警系统中两个核心问题,数据收集和预测模型构建,为学习预警系统的实现提供了参考。

一、在线学习预警因素分析

远程教育学习者大多为成年学习者,其主要是利用工作之余通过电脑、手机、平板等设备进行在线学习。在线学习平台采集了大量学习者通过各类设备学习的学习行为数据,从这些大量的行为数据中找出能反映学习者在线学习情况,更好预测学习者未来学习成绩的指标变量是进行学习预警的重要环节。

本研究以国家开放大学 Moodle 在线学习平台在线学习行为数据与终结性期末成绩数据作为数据分析的依据,希望利用学习分析技术发现远程学习者在线学习行为数据与学习成绩的关系,找出与学习成绩相关性较大的在线学习行为指标,有助于更准确地预测未来学习成绩,达到学习预警的目的。

基金项目:四川省教育厅 2018 自然科学重点科研项目“基于大数据分析的远程学习者个性化学习系统研究与构建”(项目编号:18ZA0317)

学习者的学习投入状况可以作为预测其学习成就的指标[10]。在网络学习空间中，学习者的学习投入具有时间、空间特性，表现为学习者的参与、专注、规律和交互四个维度[11]。对在线学习平台采集的大量学习行为数据，分别从参与、专注、规律和交互四个维度选择多项指标与终结性期末成绩进行了 Pearson 相关性检验，检验结果如表 1 所示。

表 1 预警因素与成绩的相关性分析

维度	指标	成绩	
		相关系数	*P* 值
参与	平台模块利用频次	0.20	0.004
	出勤天数	0.31	0.000
专注	测验模块利用频次	0.22	0.002
规律	出勤周数	0.34	0.000
交互	论坛发帖数	0.15	0.000
	论坛发主帖数	0.16	0.000

表 1 表明，学习者平台模块利用频次、出勤天数、测验模块利用频次、出勤周数、论坛发帖数、论坛发主帖数与学习者成绩的相关系数均为正值，即呈正相关关系，在显著水平上，*P* 值均小于 0.01，表示为显著相关。通过上述在线学习行为指标与学习成绩相关性分析，本研究在学习预警系统中，采用平台模块利用频次、出勤天数、测验模块利用频次、出勤周数、论坛发帖数、论坛发主帖数上述 6 个指标作为学习预警中预测的指标变量。

二、学习成绩预测建模

获得较为完备的数据后，所要做的就是根据已有的理论和方法对数据进行分析，这是学习分析的核心[12]。通过远程学习者在线学习行为数据和学习者期末学习成绩采集与分析，期望能从数据中发现学习行为与学习成绩之间的关系，建立预测模型，并根据预测模型在未来对学习者的学习行为数据进行分析并预测学习成绩，及时发现学习成绩可能不达标的学习者，从而发出预警信息。这也是学习分析技术中数据分析核心环节。

在数据挖据中，预测的目标是离散值时为分类问题，连续值时为回归问题。本研究将学习成绩值离散为“合格”和“不合格”两个值，对于学习成绩的预测实际上就是数据挖掘中的分类问题。分类问题是根据数据集的特点构造一个分类器，利用分类器对未知类别的样本赋予类别的一种技术[13]。在机器学习、统计学和神经网络等领域，已经有许多经典的分类预测方法。目前，常用的分类算法有朴素贝叶斯网络、决策树、人工神经网络和支持向量机等[14]。

数据分类一般由两步完成，第一步用已知的实例集构建分类器[15]。首先需要将数据集划分出训练集和测试集，通过分类算法对训练集进行训练，得到分类模型，这一步是一个有指导的学习过程。第二步使用构建好的分类器分类未知实例[16]。也就是使用第一部分中得到的分类模型对测试数据集进行预测，如果模型的准确率可以接受，就可以用它来对未知的数据集进行分类。

本研究采用国家开放大学远程学习者在线学习行为采集数据和期末考试成绩数据作为研究样本共计 129432 条，其中学习者 48043 名，共 501 门课程的数据。本研究选取 R 语言作为分类实验工具，R 语言是一款统计分析开源工具，在数据挖掘、统计计算等方面表现优秀。

（一）实验过程

为了更好地验证分类算法，采用十折交叉验证法（10- fold cross- validation），即将数据集平均分成十份，轮流将其中九份作为训练数据，根据数据分类第一步，应用朴素贝叶斯算法、C5.0 决策树算法、前馈反向传播神经网络算法、支持向量机算法四种算法，分别构建学习者学习成绩预测模型。再根据数据分类第二步，结合第一步中的评估模型，用剩下的一份数据集作为测试数据集进行测试，并记录每一次的测试结果。

数据挖掘领域通常利用精确率、召回率和 *F* 值等评价指标进行模型评价。精确率是指被分类模型正确预测的百分率。召回率指真实值被正确识别的百分率[17]。精确率、召回率和 *F* 值的具体计算公式如下：

$$P=TP/(TP+FP) \quad (1)$$

$$R=TP/(TP+FN) \quad (2)$$

$$F=2\times P\times R/(P+R) \quad (3)$$

式中，*P* 表示精确率，*R* 表示召回率，*F* 表示召回率，是精确率和召回率的调和平均。*TP* 表示假设正确识别为此类的样本数目，*FP* 表示为原本不是此类但是被错误地分为此类的样本数目。

在实验中，分别统计了四种算法十次测试中得到的精确率、召回率和 *F* 值并求十次测试结果的均值，得到的结果如表 2 所示。

表 2 学习者成绩预测算法结果

测试项目		朴素贝叶斯	C5.0 决策树	前馈反向传播神经网络	支持向量机
精确度（%）	及格	99.99	99.99	99.85	84.96
	不及格	98.48	0.84	0.24	21.42
召回率（%）	及格	99.73	89.49	84.92	99.98
	不及格	99.97	96.49	0.14	0.01
F 值（%）	及格	99.86	94.45	91.78	91.86
	不及格	99.22	1.67	0.18	0.03

ROC 曲线（receiver operating characteristics curve，接收者操作特征曲线）分析是可视化地评估分类器性能，从而进行模型选择的方法[18]。ROC 曲线是根据不同的二分类方式，以灵敏度为纵坐标，特异度为横坐标绘制的曲线。本研究通过 ROC 曲线进一步评估分类器，四个分类算法的 ROC 曲线如图 1 所示。

（二）预测结果分析

在对分类算法的评测中，精确率和召回率取值范围为[0,1]，数值越接近 1，说明分类算法越好，但是精确率和召回率有可能出现矛盾的情况，即精确率高而召回率低，或者精确率低而召回率高，F 值是两者的加权调和平

均，F 值高时更能说明分类算法的有效性。对上述样本数据的测试统计结果显示，C5.0 决策树、前向反馈传播神经网络和支持向量机三种算法在精确率、召回率和 F 值三者中均出现了趋近于 0 的值，而基于朴素贝叶斯算法的成绩预测结果显示，精确率、召回率和 F 值均接近于 1，是四个分类算法中表现最为优秀的学习者成绩预测算法。

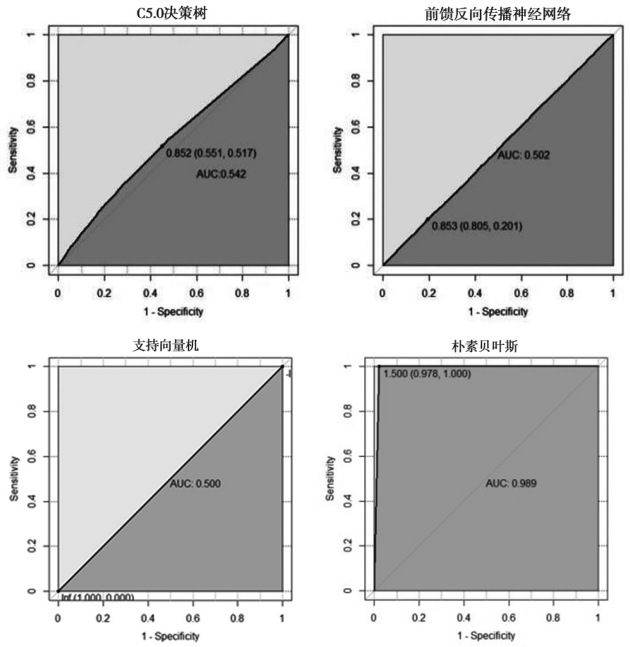

图 1　ROC 曲线

从 ROC 曲线结果来看，根据 ROC 分析思想，计算 AUC(area under the ROC curve, ROC 曲线下方面积)即可获得评价分类器识别能力的量化评价指标，面积越大，分类器的分类能力越好[19]。在图 1 所示的 ROC 曲线结果分析中，C5.0 决策树、前向反馈传播神经网络、支持向量机和朴素贝叶斯算法的 AUC 值分别是 0.542、0.502、0.500 和 0.989，总体来看，朴素贝叶斯算法在预测学习者成绩的 ROC 曲线下面积 AUC 值最大，对系统灵敏度和特异度兼顾性更好，预测算法价值更高。

综合上述分析，朴素贝叶斯算法作为学习预警系统中学习者学习成绩预测的主要算法。

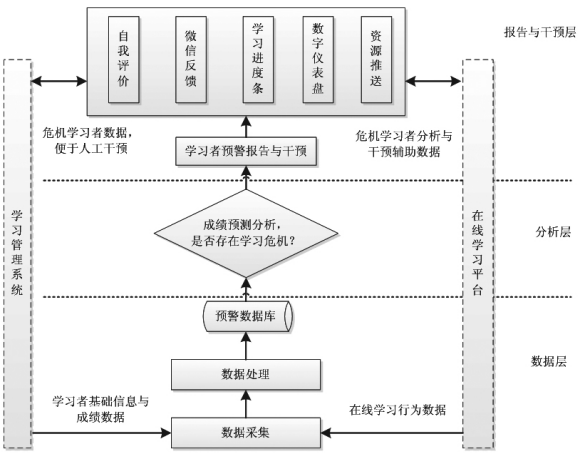

图 2　学习预警系统设计

三、基于学习分析技术的学习预警系统设计

学习分析技术是从教育领域的海量数据中提取隐含的、未知的及有潜在应用价值信息或模式的工具，也是一种决策辅助工具[18]。伊莱亚斯提出学习分析技术下的持续改进环模型，模型包括三个循环改进的过程，分别是数据收集、数据处理包括聚合预测等、在数据处理的基础上将结果进行应用并提出数据收集处理优化措施，从而持续改进模型。我国学者吴永和等人将学习分析技术框架归纳为整个系统分为数据层、分析层、报告层和干预层。其中数据层主要获取学习者相关学习数据，送入系统分析层由分析引擎进行数据分析，之后在仪表盘上按学习者、教学者、管理者三大利益需求产生可视化报告，并在此基础上对学习者进行干预并完成自适应过程[19]。依据本研究中影响学习成绩的在线学习行为因素分析和学习成绩的预测模型的构建，在对已有学习分析技术模型研究的基础上，设计了学习成绩预警系统。学习成绩预警系统设计如图 2 所示。

本研究将学习预警系统设计为三层，包括数据层、分析层和报告干预层，其中数据层负责采集学习者在线学习行为数据和学习者的成绩数据，采集的数据主要采用本研究中线上学习预警因素中 6 个指标以及在期末时作为训练数据用途的学习者期末成绩。采集结束后对数据进行清洗、离散化等数据处理，数据处理结束后转入学习预警数据库，以免对在线学习平台和学习管理系统的数据库运行效率造成影响。

分析层利用采集的学习者在线学习行为数据和成绩数据，根据本研究上述分析中的学习预警模型，主要通过朴素贝叶斯算法，建立预测模型，预测模型建立成功后对需要预测的学习者在线学习行为数据进行预测。为了能够在学期早期及时预警学习者情况，对于预测模型的练习数据采集需要从上一学期中期开始每周采集一次学习者学习行为数据，学期结束后取得学习者的期末成绩数据，分别建立不同周对应的预测模型并保存作为下学期预测使用。下学期中期开始后利用相应的周预测模型进行预测，得出每周的预测数据，通过每周的预测，形成较为动态的分析，有利于帮助管理者、教学者、学习者动态掌握学习者在线学习的情况，进行学习改进。及时将学习者预测结果名单和在线学习行为情况等信息反馈给学习管理系统，便于管理者和教学者及时了解学习者学习情况，对于危机学习者再通过学习者自我评价采集、电话访问、学习进度、学习仪表盘等，及时了解学习者学习情况，并掌握学习者学习中存在的特殊问题给予相应的学习支持服务。针对学习者本人，在线学习平台也会通过微信提醒、学习仪表盘、学习者进度、资源推送等方式，及时告知学习者存在的问题以及需要努力的方向，帮助学习者了解自身的学习情况以及同门课程学习者的平均水平，帮助了解自身存在的不足，促进学习者加强学习，提高学习积极性，并针对自身存在的

不足及时补缺。

在报告与干预层中，通过自我评价、微信反馈、学习进度条、数字仪表盘和资源推送五个方面对学习者进行报告和干预。其中自我评价主要从主观方面了解学习者学习中存在的主观问题因素，以更全面地了解学习者学习情况；微信反馈则通过向学习者发送预警信息、课程学习通知消息、课程任务提醒消息、讨论通知等，与学习者产生紧密互动，以帮助提高学习者学习积极性；学习进度条提示学习者课程学习进度情况，例如课程练习完成比例、视频文本等资源阅读比例、参与论坛讨论活跃度等信息；数字仪表盘则从纵向上提示学习者课程中其他学习者平均学习进度情况，通过对其他学习者平均进度学习情况了解，学习者了解同一门课程的学习者中，自身在各类学习项目中学习进度属于平均水平以上还是以下，如果处于平均水平线以下则警示学习者及时调整自身的学习进度，例如练习量不够，或者是视频文本资源阅读量不足，或者讨论积极性低等，学习者可以根据仪表盘的提示有针对性地进行自我改进，从而达到提高学习者学习质量的目的。

四、讨论

对于远程教育在线学习中存储的海量在线学习行为数据，利用学习分析技术，构建远程学习者学习预警系统。本研究首先研究了学习预警所需要采集的基础数据指标，并在数据收集的基础上构建学习成绩预测模型，通过多种分类算法在大数据样本的实验中进行比较，选择了预测结果表现优秀的朴素贝叶斯算法作为学习者学习成绩的预测算法，具有较高的可行性。最后，在学习分析技术系统框架的基础上，设计了远程学习者学习预警系统的基本框架，为未来远程教育在线学习学习预警系统的构建提供了参考依据。

学习预警的建立，为教育中的学习者、教学者、管理者提供了教学过程中的数据分析和决策支持，在海量学习者中及时发现具有潜在学习“危机”的学习者，通过数据仪表盘、学习进度图等方式为学习者提供及时的预警提示，同时管理者与教学者可以及时发现存在危机的学习者，以便及时给予学习支持服务，提高学习者的学习成功率，同时也有助于降低学习者辍学率。笔者主要解决了学习预警中数据收集和预测算法以及预警系统的设计，在预警后的报告与干预中，尚需要更进一步地研究预警报告和预警干预实施细则，使其能为学习者和教学者提供较为直观清晰的预警报告，并提供较为智能的干预措施。学习分析技术在国内教育领域逐渐受到研究者的热切关注，将学习分析技术应用于远程教育中，从海量数据中发掘出有意义的数据关联，为提高远程教育质量，实现个性化学习提供了有益的视角。

参考文献：

[1]王林丽，叶洋，杨现民.基于大数据的在线学习预警模型设计——“教育大数据研究与实践专栏”之学习预警篇[J].现代教育技术，2016(7):5-11.

[2]肖巍，倪传斌，李锐.国外基于数据挖掘的学习预警研究：回顾与展望[J].中国远程教育，2018(2):70-78.

[3]马红亮，袁莉，郭唯一，等.反省分析技术在教育领域中的应用[J].现代远程教育研究，2014(4):39-46.

[4]姜强，赵蔚，李勇帆，等.基于大数据的学习分析仪表盘研究[J].中国电化教育，2017(1):112-120.

[5]Essa A, Ayad H. Improving Student Success Using Predictive Models and Data Visualisations [J].Research in Learning Technology,2012,(20):58-70.

[6]李香勇，左明章，王志锋.学习分析的研究现状与未来展望——2016年学习分析和知识国际会议述评[J].开放教育研究，2017,23(1):46-55.

[7]Siemens G，Long P.Penetrating the Fog：Analytics in Learning and Education [J].EDUCAUSE Review，2011，46(5):30-32.

[8]何克抗.“学习分析技术”在我国的新发展[J].电化教育研究，2016，(7):5-13.

[9]顾小清，刘妍，胡艺龄.学习分析技术应用：寻求数据支持的学习改进方案[J].开放教育研究，2016,22(5):34-45.

[10]李晓东，蓝艳玉，黄娟.中小学教师远程继续教育学习质量评价问卷的研制及应用[J].远程教育杂志，2011(6):51-58.

[11]张思，刘清堂，雷诗捷，等.网络学习空间中学习者学习投入的研究——网络学习行为的大数据分析[J].中国电化教育，2017(4):24-30,40.

[12]张玮，王楠.学习分析模型比较研究[J].现代教育技术，2015,25(9):19-24.

[13]钱晓东.数据挖掘中的数据分类算法综述[J].图书情报工作，2007,51(3):68-71,108.

[14]陈子健，朱晓亮.基于教育数据挖掘的在线学习者学业成绩预测建模研究 [J]. 中国电化教育，2017(371):75-81,89.

[15]蒋良效.朴素贝叶斯分类器及其改进算法研究[D].武汉：中国地质大学，2009.

[16]吴青，罗儒国.基于在线学习行为的学习成绩预测及教学反思[J].现代教育技术，2017,27(6):18-24.

[17]董元方，李雄飞，李军，等.基于分辨粒度的gROC曲线分析方法[J].软件学报，2013,24(1):109-120.

[18]万柏坤，薛召军，李佳，等.应用ROC曲线优选模式分类算法[J].自然科学进展，2006,16(11):1511-1516.

[19]胡艺龄，顾小清，罗九同，等.教育效益的追问：从学习分析技术的视觉[J].现代远程教育研究，2014(6):41-47.

[20]吴永和，陈丹，马晓玲，等.学习分析：教育信息化的新浪潮[J].远程教育杂志，2013,4(3):11-19.