DOI:10.16652/j.issn.1004-373x.2020.18.015

大数据驱动的学习分析技术研究进展

胡金蓉[1]，邹茂扬[1]，文　武[1]，周子龙[2]
（1.成都信息工程大学，四川 成都　610225；2.昊华化工科技集团股份有限公司，四川 成都　610225）

摘　要： 为了促进学习分析在高等院校的实施，该文对大数据驱动的学习分析技术及其应用进行了综述。首先，对分类、回归、聚类、潜在知识评估、文本和语音挖掘、社会网络分析、序列模式挖掘等学习分析技术进行分析总结；然后，根据它所解决的问题进行分类，从分析学习行为、评估学习结果和支持学生的个性化学习三方面详细阐述了学习分析技术的应用职能；最后，总结了学习分析的挑战和展望。在多学科协同下，学习分析融入了多方面的新技术来研究学习发生的内在机制和深层次的原因，揭示学习规律，从而为学习者提供更优化的学习环境。

关键词： 学习分析技术；大数据驱动；学习行为；个性化学习；数据挖掘；多学科协同
中图分类号： TN919-34　　**文献标识码：** A　　**文章编号：** 1004-373X(2020)18-0054-05

Research process of big-data-driven learning analysis technology

HU Jinrong[1], ZOU Maoyang[1], WEN Wu[1], ZHOU Zilong[2]
(1. Chengdu University of Information Technology, Chengdu 610225, China; 2. Haohua Chemical Science & Technology Corp., Ltd., Chengdu 610225, China)

Abstract: The big-data-driven learning analysis technology and its application are reviewed to promote the implementation of learning analytics in higher education institutions. The learning analysis technologies such as classification and regression, clustering, potential knowledge assessment, text and speech mining, social network analytics and sequence pattern mining are analyzed and summarized. The learning analytics technologies are classified according to what they can solve, and then the application functions of the learning analytics technology is elaborated in the aspects of analyzing learning behavior, evaluating learning results and supporting students′ personalized learning. The challenges and prospects of learning analytics are summarized. In the multidisciplinary collaboration, the learning analytics integrates various new technologies to study the internal mechanism and deep-seated reasons of learning occurrence, and reveal the learning pattern, so as to provide learners with a more optimized learning environment.

Keywords: learning analytics technology; big data driven; learning behavior; personalized learning; data mining; multidisciplinary collaboration

0　引　言

随着大数据时代的来临，“数据驱动学校、分析变革教育”成为了教育创新变革的战略之一。2012年，美国联邦政府教育部技术办公室发布《通过教育数据挖掘和学习分析改进教与学：问题简介》[1]，指出教育数据分析的两个重要的因素：教育数据挖掘(Educational Data Mining，EDM)和学习分析(Learning Analytics，LA)。国际教育数据挖掘协会对EDM的定义为：“教育数据挖掘是一门新兴学科，利用开发方法来探索教育数据，以便更好地了解学生以及他们的学习环境”[2]。学习分析研究协会对LA定义为：测量、收集、分析和报告有关学习者及其背景的数据，以了解和优化学习及学习环境[3]。二者的研究领域有不少交叉的地方，但也存在差异，“EDM主要关注典型数据挖掘技术的应用，以支持教师和学生分析学习的过程。除了数据挖掘技术之外，LA还包括其他方法，如统计和可视化工具或社会网络分析(Social Network Analysis，SNA)技术，并将它们付诸实践，以研究它们在改善教学和学习方面的实际效果”[4]。

学习分析是利用技术为教学服务，它有助于将教育

收稿日期：2020-01-08　　**修回日期：**2020-03-12
基金项目：国家自然科学基金项目(61806029)；国家自然科学基金项目(61602390)；四川省高等教育人才培养质量和教学改革项目(365)；2018—2020成都信息工程大学高等教育人才培养质量和教学改革项目(JY2018019)；2019—2021年成都信息工程大学第一阶段本科教学工程项目(BKJX2019108)

管理模式转向动态化管理，将教学由事后干预转为事前预防。学习分析的重要性可以从美国新媒体联盟与北京师范大学智慧学习研究院合作的《2016 新媒体联盟中国基础教育技术展望：地平线项目区域报告》[5]中看出：基于大数据的学习分析技术将在未来2~3年成为极具影响力的教育技术。对于学习分析这个研究热点，该研究关注各种为教学服务的学习分析技术和其应用，以期待能够为大数据驱动的教育改革提供参考。

1　学习分析的主流算法

当前的学习分析技术主要是建立在数据挖掘方法的基础上，学习分析技术的主流算法的类别如图1所示，包括6个方面。这些技术可以做到分析学生的学习状态、学习能力，跟踪学生的知识掌握情况，对掉队的学生预警，为各种类别的学生推荐不同的资源和学习路径等。

图1　学习分析主流算法

1）分类与回归。分类和回归是对2种或2种以上变量间相互依赖的定量关系进行建模，包括连续变量的预测和离散变量的预测。在学习分析中常用到的算法包括线性回归、逻辑回归、决策树、随机森林、朴素贝叶斯分类器、支持向量机和人工神经网络。它在教育中的应用场景比较多，根据学生的行为对学生分类，可以分析影响学习效果的强因素，分析学习特征，预测学生的学习效果，对离群点预警，还可以为学生提供不同的学习资源分析依据。

2）聚类。聚类是将物理或抽象对象的集合分成由类似的对象组成的多个类的过程，在学习分析中常用到的算法有K-means及其各种变种、自组织映射网络。它在教育中的应用场景和分类类似。

3）潜在知识评估。潜在知识评估是追踪对知识点掌握的情况。应用比较多的算法有贝叶斯知识追踪模型，目前最新的算法有循环神经网络、双向长短时记忆网络。它在教育中的应用场景是可视化学生的知识掌握变化情况，预估学习者完成课程的可能性。

4）文本、语音挖掘。它是当前人工智能的研究热点。它运用各种机器学习、深度学习的方法，挖掘文本、语音的深层次概念。主流算法有循环神经网络、双向长短时记忆网络、双向长短时记忆循环神经网络、深度全序列卷积神经网络。尤其是深度全序列卷积神经网络，它是科大讯飞提出的一种全新的语音识别框架。文本、语音挖掘在教育中的应用场景是按照内容对文档、语音进行识别分类，提取有用的知识和信息，辅助对教学过程进行分析。目前它们在学习分析中的应用还不多，多数是在辅助前文的分类和聚类算法。

5）社会网络分析。社会网络分析是对网络中各种关系进行客观的定量分析，在学习分析中常用到的算法有PCA主成分分析算法。它在教育中的应用场景是通过个体和群体的互动行为，分析团队交互对学习者认知的影响，确定每个成员在合作项目中的贡献。

6）序列模式挖掘。序列模式挖掘是挖掘相对时间或其他模式出现频率高的模式。在学习分析中常用到的算法有AprioriAll算法、GSP算法、FreeSpan算法、PrefixSpan算法，它在教育中常被用于个性化学习，分析出来的关联性关系用于为学生可能感兴趣的内容提供建议，推荐个性化学习资源，优化学生的个性化学习路径。

2　学习分析技术的应用研究

研究者应用上述的学习分析算法进行建模和分析，其最终目的都是改善教与学的效果。本文按照学习分析技术能够解决问题划分，将学习分析技术的应用分为两个方面：第一是分析学习行为评估学习结果；第二是支持学生的个性化学习。研究者们在以上两个领域，尤其是第一个领域的应用研究，已经取得了可喜的成绩，以下对两个领域的应用研究进行详细的阐述。

2.1　分析学习行为评估学习结果

当前，学习分析技术应用最多的是对学习过程进行统计分析，评估学习结果。通过可视化信息提供学习行为、学习内容、学习活动以及学习共同体相关信息，激发学生在学习分析过程中得到的自我认知，进而触发学生的学习反思。

对于可视化学习过程以及统计结果，目前各国比较流行使用各种学习分析仪表盘（Learning Analytics Dashboard，LAD），实现了知识生成与教育数据挖掘结果可视化，能够支持学生自我认知、自我评价、自我激励和社会意识及优化未来的智慧学习环境[6]。除了LAD，Dietzuhler等人通过监控学生的表现，潜在地发现课程中需要改进的地方[7]。Olmos等人将学习经验数据可视化，提高了学生的成功率，降低了留级率[8]。

在可视化学习过程及统计结果的基础上，研究者通过学习行为与学习结果的关联分析，使学习分析技术有了进一步的应用，主要的研究集中在3个方面：

1）研究对学习效果的预测，对离群点发出预警；

2）研究影响学习成绩的强因素；

3）研究学习特征。

这些研究能够进一步揭示因果关系，评估学习结果，引导学生进行自我的意义建构，同时为老师调整教学，及时干预提供数据支持。

2.1.1 预测学习结果，对离群点预警

预测学习结果的方法比较多。He等人使用回归分析来分析学生的在线学习行为及其在课程中的表现[9]，他们分析了学生的参与和登录频率，以及他们向教师提交的聊天消息和问题的数量，用于预测学生的最终成绩。Kizilcec团队和Giesbers团队研究了互动与学生最终成绩之间的关系，用于预测他们的表现。Kizilcec研究的是学生在实时视频工具中的互动[10]，Giesbers研究的是学生在同步工具中的互动[11]。针对数据集不充分的情况，Natek等人研究了对小型学生数据集的数据挖掘，预测成功学生的百分比[12]。

学校一般特别关注高危学生，为了降低辍学率和留级率，在预测结果的基础上，对老师和学生发出预警，提醒老师在初期进行干预。早期的学习预警研究大多基于思辨或依据期末成绩等指标进行，存在指标主观单一、时间滞后等不足。近年来兴起的数据挖掘技术则为学习预警研究提供了方法上新的切入点[13]。赵慧琼等人利用多元回归分析法判定影响学生学习绩效的预警因素，在此基础上构建了干预模型，将其应用于教学实践中，并结合问卷调查和访谈法，对该模型在学习活动、知识习得等方面的有效性进行了验证[14]。Marbouti等人调查了120名美国本科生的课前、课中、课后数据，使用7种方法对高危学生进行预警，包括了逻辑回归、支持向量机、K-邻近、决策树和朴素贝叶斯模型等。其中，朴素贝叶斯分类器模型和集合模型（支持向量机，K-近邻和朴素贝叶斯分类器的结合）在七种测试建模方法中预测结果最佳[15]。

2.1.2 影响学习效果的强因素研究

在哪些环境下哪些行为与预测结果的相关性大？是否有与学生的成功率有关的特定学生特征？这是学习行为和学习结果关联性分析中的研究热点。一般而言，研究数据来源于学习管理系统和在线学习平台。

对于学习管理系统，Smith等人使用登录频率、参与度、学生进度和作业成绩来预测课程结果，识别有风险的学生[16]。Yu等人采用多元线性回归分析来确定哪些因素影响学生的学业成绩[17]。韩国女子大学的84名本科生参加了这项研究，预测因子有在线时间、同伴交互数量、教师交互量、总登录频率、下载次数与学习时间间隔规律等。研究结果表明，在学习管理系统中的总学习时间、与同伴的互动、学习管理系统中学习间隔的规律性以及下载次数这4个因素被确定为影响学生学习成绩的重要因素。Wong等人对WebCT课程管理统计工具进行分析，论证了学生参与在线资源的程度与其整体学业成绩之间的正相关关系，更多地使用在线教学资源对学生的学习成绩产生了积极的影响。他所采用的预测指标有登录次数、每一章节平均的学习时间、查看和下载资源数量、在线测试完成次数及时间、论坛参与频率[18]。针对在线学习平台，李曼丽等人使用Tobit和Logit两个定量分析模型，分别对MOOC学习者的课程参与和完成情况进行深入分析，研究影响学习者完成课程的因素，包括课程注册时间，学习者在课程讨论区和Wiki的各种表现等[19]。Pardo等人研究的特别之处是将学习动机、学习策略的问卷数据与在线活动的可观察性指标相结合。对145名学生的案例研究表明，一起考虑这两种方法的因素，可以更好地解释学生期末成绩的变化[20]。除了个人的学习行为，在最近几年，针对社会网络交互行为的研究悄然兴起，令影响学习效果的因素研究更加全面。

有研究发现在高效的协作学习中，学习者需要综合多维度的感知信息来了解协作过程，以此提升其协作积极性，进而提高小组协作质量[21]。He等人使用数据挖掘和文本挖掘技术，分析由实时视频流系统自动记录的在线问题和聊天消息，发现了在线问题（学生-教师互动）和在线聊天消息（学生-学生互动）之间，学生参与模式的差异和相似之处，确定了学生在线参与的学科差异，以及学生提出的在线问题数量与学生的最终成绩之间的相关性[22]。Tirado等人通过内容分析来评估知识构建过程的质量，使用参与者之间的响应关系来分析网络结构，证实了在线协作学习中，小组成员之间的参与、交互对协作学习有促进作用[23]。Lin等人研发了一个自适应的学习系统，将团队协作过程作为一个重点可视化地呈现出来，为学习者提供小组成员交互图，在某大学中计算机课程中使用，增强了学习者的自我调节，促进了学习者的协作参与，提升了团队学习效果[24]。

2.1.3 对学习特征的研究

Jovanovic等人应用K-means模型来预测学生的表现，并根据认知风格及其整体表现，提供能够更好地适应学习的方式[25]。Feldman等人通过提取学生与益智游戏的互动信息，训练朴素贝叶斯分类器来检测学生的感知风格[26]。通过使用游戏成功预测了感知风格，准确率为85%。张琪等人研究了“互联网+”混合学习场景，基于5类人格分类，利用多元线性回归构建相应的预测模型，分析学习行为指标与不同人格特质群体学习结果之间的相关性[27]。

2.2 支持学生的个性化学习

因材施教是一个持久的教育命题,学习分析技术的发展为学生的个性化学习提供了技术保障。无论对于学校的物理班级,还是网络学习的虚拟班级,学习分析技术可以用来创建一个定制好的学习环境,在这个环境中,可以为学生提供个性化的学习路径来优化学生的表现[28]。姜强等人基于AprioriAll算法,挖掘分析相同或相近学习偏好、知识水平的同一簇群体学习行为轨迹,并以学习者特征与学习对象媒体类型、理解等级、难度级别的匹配计算为基础,生成精准的个性化学习路径[29]。立陶宛维尔纽斯大学Eugenijus Kurilovas采用基于协作和信息素的蚁群优化方法,可以选择静态和动态学习单元中的学习路径[30]。刘淇等人依托应用平台——科大讯飞在线教育系统"智学网",根据基于认知诊断的个性化学习资源推荐方法,给出针对教师的教学建议[31]。个性化学习适应学习者的需求和增强学习者的动机,令学习活动更加有效。

3 学习分析面临的挑战与展望

学习分析是一门交叉学科,由于教育大数据和人工智能的推动,最近几年得到了快速发展,形成了不少成果。目前,它面临的挑战有多模态数据采集与处理、深度学习的深度融合、实证研究不够等。这些当前提出的热点问题同时也是今后学习分析研究的方向。

3.1 多模态数据和学习分析互操作性

多模态数据包括生理层数据、心理层数据、行为层数据和基本信息数据。目前,研究者的数据多来源于学习管理系统和在线学习平台,采集最多的是学习行为轨迹和学习结果数据,其次还有文本、音频、视频数据。随着虚拟技术和射频技术的发展,采集生理层的数据越来越方便,包括眼动、面部表情等,这使学习分析的研究范围扩大了,也令分析结果更加准确。但是,多模态的数据也带来了学习分析互操作性(Learning Analytics Interoperability,LAI)的困难。近年来,一些学习分析互操作规范化解决方案不断涌现,如ADL的XAPI和IMS的Caliper等[32],但是面对多模态数据的数据结构以及语义的差异,自动化的处理技术仍需改进。

3.2 与深度学习深度融合

模拟人脑学习的深度学习在图像识别、语音识别、自然语言、计算机视觉等领域取得了巨大的成功。深度学习在教育大数据处理中的深度融合,是机遇也是挑战。它在学习分析上的应用有2个发展方向:第一是运用深度学习在特征提取方面的优势,进行学习者特征建模;第二是深度学习技术自动挖掘实体,自动抽取实体间的语义关系,更加精确和智能地构建多学科知识图谱。

3.3 实证研究和在学校的应用

目前,学习分析的成果大多在研究层面,在高校中有部分应用,如2014年英国诺丁汉特伦特大学在全校推出了学习仪表板;重庆大学的研究人员利用教育数据挖掘技术,绘制学生在课堂内相互交流的信息等。但是,它在高校的应用并没有进入常态化,并且在学校的实际应用还面临数据隐私的问题。要形成从理论体系、技术研究到实证的研究体系,学习分析还有很长的路要走。

4 结 语

本文从技术角度和应用角度分析了学习分析这个新兴领域的发展,总结了它的挑战和展望。总体而言,学习分析令决策者、学生、教师都受益,学习分析的价值和潜在影响得到了教育界的肯定。随着学习内容和技术手段的多样化,学习分析变得更加复杂,在多学科协同配合下,学习分析将融入多方面的新技术(如虚拟现实、深度学习等)来研究学习发生的内在机制和深层次的原因,揭示学习规律,从而为学习者提供更优化的学习环境。

注:本文通讯作者为邹茂扬。

参 考 文 献

[1] BIENKOWSKI, M, FENG M, MEANS B, et al. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief [R]. Washington: U. S. Department of Education, Office of Educational Technology, 2012.

[2] REDA A, JON R. Educational data mining [J]. Computer science, 2018(7): 17-19.

[3] PAPAMITSIOU Z, ECONOMIDES A A. Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence [J]. Journal of educational technology & society, 2014, 17(4): 49-64.

[4] CHATTI M A, DYCKHOFF A L, SCHROEDER U, et al. A reference model for learning analytics [J]. International journal of technology enhanced learning, 2012, 4(5/6): 318.

[5] JOHNSON L, LIU D, HUANG R, et al. 2016 NMC technology outlook: chinese K-12 education [R]. Austin: The New Media Consortium, 2016.

[6] 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.

[7] DIETZUHLER B, HURN J E. Using learning analytics to predict (and improve) student success: a faculty perspective [J]. Journal of interactive online learning, 2013, 12(1): 17-26.

[8] OLMOS M, CORRIN L. Learning analytics: a case study of the process of design of visualizations [J]. Journal of asynchronous learning network, 2012, 16(3): 39-49.

[9] HE W, YEN C J. Using data mining for predicting relationships between online question theme and final grade [J]. Journal of educational technology & society, 2012, 15(3): 77-88.

[10] KIZILCEC R F, PIECH C, SCHNEIDER E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses [C]// Proceedings of the third international conference on learning analytics and knowledge. New York: ACM, 2013: 170-179.

[11] GIESBERS B, RIENTIES B, TEMPELAAR D, et al. Investigating the relations between motivation, tool use, participation, and performance in an e-learning course using web-videoconferencing [J]. Computers in human behavior, 2013, 29(1): 285-292.

[12] NATEK S, ZWILLING M. Student data mining solution-knowledge management system related to higher education institutions [J]. Expert systems with applications, 2014, 41(14): 6400-6407.

[13] 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.

[14] 赵慧琼,姜强,赵蔚,等.基于大数据学习分析的在线学习绩效预警因素及干预对策的实证研究[J].电化教育研究,2017,38(1):62-69.

[15] MARBOUTI F, DIEFES-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.

[16] SMITH V C, LANGE A, HUSTON D R. Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses [J]. Journal of asynchronous learning network, 2012, 16(3): 51-61.

[17] YU T, JO I H. Educational technology approach toward learning analytics: relationship between student online behavior and learning performance in higher education [C]// Proceedings of the Fourth International Conference on Learning Analytics and Knowledge. New York: ACM, 2014: 269-270.

[18] WONG L. Student engagement with online resources and its impact on learning outcomes [C]// Proceedings of the 2013 International Conference on Information Science and Technology Applications. Macau: ICISTA, 2013: 129-146.

[19] 李曼丽,徐舜平,孙梦嫽,等.MOOC学习者课程学习行为分析:以“电路原理”课程为例[J].开放教育研究,2015(2):63-69.

[20] PARDO A, HAN F, ELLIS R A. Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance [J]. IEEE transactions on learning technologies, 2017, 10(1): 82-92.

[21] 李艳燕,张媛,苏友,等.群体感知视角下学习分析工具对协作学习表现的影响[J].现代远程教育研究,2019(1):104-114.

[22] WU He. Examining students' online interaction in a live video streaming environment using data mining and text mining [J]. Computers in human behavior, 2013, 29(1): 90-102.

[23] TIRADO R, HERNANDO A, IGNACIO A J. The effect of centralization and cohesion on the social construction of knowledge in discussion forums [J]. Interactive learning environments, 2015, 23(3): 293-316.

[24] LIN J W, LAI Y C, LAI Y C, et al. Fostering self-regulated learning in a blended environment using group awareness and peer assistance as external scaffolds [J]. Journal of computer assisted learning, 2016, 32(1): 77-93.

[25] JOVANOVIC M, VUKICEVIC M, MILOVANOVIC M, et al. Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study [J]. International journal of computational intelligence systems, 2012, 5(3): 597-610.

[26] FELDMAN J, MONTESERIN A, AMANDI A. Detecting students' perception style by using games [J]. Computers & education, 2014, 71: 14-22.

[27] 张琪,王红梅,庄鲁,等.学习分析视角下的个性化预测研究[J].中国远程教育,2019(4):38-45.

[28] LIN C F, YEH Y C, HUNG Y H, et al. Data mining for providing a personalized learning path in creativity: an application of decision trees [J]. Computers & education, 2013, 68(4): 199-210.

[29] 姜强,赵蔚,李松,等.大数据背景下的精准个性化学习路径挖掘研究:基于AprioriAll的群体行为分析[J].电化教育研究,2018(2):45-52.

[30] KURILOVAS E, ZILINSKIENE I, DAGIENE V. Recommending suitable learning paths according to learners' preferences: Experimental research results [J]. Computers in human behavior, 2015, 51: 945-951.

[31] 刘淇,陈恩红,朱天宇,等.面向在线智慧学习的教育数据挖掘技术研究[J].模式识别与人工智能,2018(1):77-90.

[32] 郑隆威,冯园园,顾小清,等.学习分析:连接数字化学习经历与教育评价:访国际学习分析研究专家戴维·吉布森教授[J].开放教育研究,2016,22(4):4-10.

作者简介:胡金蓉(1983—),女,四川南充人,工学博士,副教授,研究方向为人工智能教育应用、机器学习。
邹茂扬(1974—),女,四川泸州人,工学硕士,副教授,研究方向为数据挖掘、深度学习、教育大数据。
文 武(1979—),男,四川蓬溪人,工学博士,讲师,研究方向为机器学习、数据挖掘。
周子龙(1969—),男,四川成都人,高级工程师,研究方向为数据挖掘。