模型驱动的教育大数据挖掘促进教与学

——访美国犹他州立大学米米·雷克教授

本刊特邀记者　余　亮　杨秋燕　赵　楠

[编者按]　米米·雷克(Mimi Recker)是犹他州立大学教育与人力服务学院教授,兼任美国国家科学基金资助咨询委员会委员,还先后担任《学习科学杂志》(Journal of the Learning Sciences)、《教学媒体国际期刊》(International Journal of Instructional Media)、《教育技术研究与发展》(Educational Technology Research and Development)、《美国教育研究杂志》(American Education Research Journal)等编委,研究方向包括交互式学习环境的设计和评估,新媒体、信息和通信技术的教育应用以及创客教育等,目前致力于探索学校及公共图书馆创客发展和在线学习分析技术研究,曾主持八项美国国家科学基金项目,在《学习科学杂志》《教学科学》(Instructional Science)、《教育技术研究与发展》等期刊发表学术论文 30 多篇,2010 年获美国教育传播与技术协会(AECT)杰出发展奖提名,2013 年获犹他州立大学年度优秀导师称号。

[关键词]　教育大数据;数据挖掘;模型;学习分析

[中图分类号] G431　[文献标识码] A　[文章编号] 1007-2179(2018)01-0004-06

记者:雷克教授,您好,感谢您接受我们的采访。教育大数据现在发展迅速,您能谈谈教育大数据挖掘在美国的发展状况吗?

雷克教授:自 20 世纪 80 年代以来,教育研究人员一直在分析由学生与计算机系统交互而产生的日志文件(早期形式的"大数据")。其中最突出的是研发智能导学系统(intelligent tutoring system)(包括研究学习过程),为这些数据源开发新的分析方法。

随着网络教育环境的发展,特别是大规模在线开放课程的出现,计算机系统搜集的大数据数量呈爆发式增长。与传统的教育数据集不同,这些数据集量大(上万名学生)、多维(每个学生的不同方面)、时间序列短(数据采集时间间隔短)、时间长(数据搜集周期长)和深度(具有理论意涵),对常规教育数据研究方法提出了挑战。研究人员越来越多地转向新兴的数据挖掘和统计技术,如文本挖掘、马尔可夫模型和多项式逻辑回归等。这些数据挖掘方法有助于发现和解决教育中的问题,如提高学生学习成效,帮助教师改进教学方法和课程资料,辅助管理人员有效决策等。教育数据挖掘主要应用于可视化(visualization)、学生建模(student modeling)、预测学生表现(predicting student performance)、推荐系统(recommender system)以及自适应系统(adaptive system)。为了推动教育数据挖掘研究,2008 年教育数据挖掘研究人员联合成立了国际教育数据挖掘协

[收稿日期]2017-11-10　[修回日期]2017-12-25　[DOI 编码]10.13966/j.cnki.kfjyyj.2018.01.001
[基金项目]2016 年重庆市研究生教育教学改革研究项目"面向多形态课程的数字资源特征及共享机制——以免师硕士在线课程为例"(YJG20163068);中央高校基本科研业务费专项资金创新团队项目"智慧学习环境研究"(XDJK2014A002)。
[作者简介]余亮,博士,副教授,西南大学计算机与信息科学学院(yuliang@swu.edu.cn);杨秋燕,硕士研究生,西南大学计算机与信息科学学院(qzyy770520@email.swu.edu.cn);赵楠,硕士研究生,西南大学计算机与信息科学学院(zhao690911@email.swu.edu.cn)。

会,并于当年举办了第一次研讨会,次年创办了《教育数据挖掘》期刊。2012 年,美国教育部发布了《通过教育数据挖掘和学习分析促进教与学》蓝皮书。

记者:教育大数据挖掘的基本过程是怎样的?建模是教育大数据分析的关键环节,教育大数据挖掘有哪些模型?常使用哪些方法?

雷克教授:教育数据挖掘通常采用数据科学技术,尝试从教育环境收集的大规模数据中抽取有用的模型和信息。这种研究方式往往没有任何学习与教学理论为基础,是令人担忧的。没有理论,所得到的模型就难以解释。因此,我们认为教育数据挖掘必须基于学习或教学理论,并将其作为模型开发的关键部分。

通常,教育大数据挖掘过程遵循迭代策略,包括:1)从学习系统收集并存储数据;2)根据需求整理数据,对信息进行常规化和标准化处理。这有时也被称为"数据清洗",通常占整体工作量的 30% 以上;3)采用机器学习或统计技术对数据进行挖掘和分析;4)报告和可视化呈现结果,应用于学习环境以改进绩效。

这些数据通常是多层次(hierarchical)、强时效(time sensitive)、有序(sequenced)和情境化(contextual)的。例如,击键(keystroke)数据是学习者整体学习序列的一小部分。当学习者在某一学校、年级、课程学习某一主题,数据序列捕获学习主题的关键顺序,数据情境帮助研究者和相关人员解释结果并理解模型的局限。

教育数据挖掘旨在探讨和建立可集成到在线学习系统的模型。第一种是学习者模型,它捕获学习者已有的知识、学习动机、先前经验、个体数据和行为,以用于个性化内容推荐,提供即时帮助和建议,并使学习者保持积极的学习状态;第二种是领域知识模型,用以表征将要学习的内容、范围及顺序。第三种是教学模型,它规定了教学序列或理论(如基于问题的学习)。总之,这些模型通过计算建模改进教学过程,扩展学习与教学的科学知识。

教育数据挖掘常用方法有五种:预测(prediction)、聚类(clustering)、相关性挖掘(relationship mining)、萃取数据支持判断(distillation of data for human judgement)和利用模式探索(discovery with model)。下面我简要介绍一下。预测指从预测变量的某种组合中建立模型,预测结果变量,如学生成绩。许多研究使用预测方法找出学生成绩的重要预测因子和非预测因子。聚类是将数据集自然分为较小子集的数据点(data point)。用于聚类的分组对象可以是学生、课程或者学习内容,以学生最为常见。基于模型的聚类方法越来越多,尤以 K-means 法最常用。相关性挖掘是发现数据集中变量之间的关系,并将其编码,作为规则供后续使用,它包含关联规则挖掘(association rule mining)和序列模式挖掘(sequential pattern mining)。相关性挖掘的教学应用包括探索学生成绩与课程序列间的关系,以及哪种教学方法能有效支持学习。萃取数据支持判断指描绘数据,包括数据可视化,使人们能快速识别其中的模式并理解其特征。它包括热图(heat map)、图形和散点图等。利用模式探索指利用已验证的模型(预测、聚类或者知识工程方法等得到的)为基础进行分析。例如,以心理测量模型和机器学习模型为基础,利用模式探索研究学生行为与学生个体特征或情境变量间的关系。

记者: 教育大数据挖掘如何有效支持教与学?

雷克教授:教育数据挖掘使用数据挖掘技术开发学习绩效的预测模型。这一系统通常由以下组件组成:1)内容管理组件,呈现个性化内容、问题和评价信息,以支持学生学习;2)学生学习数据存储库,自动存储学生活动时所记录的带有时间戳(time-stamped)的行为;3)预测模型,即将学生个体特征数据(来自外部学生信息系统)和学生学习活动数据库的数据相结合,预测学生未来学习行为或学习成绩;4)报告系统,即将预测模型的输出结果报送教师、管理员和家长等;5)干预或学习分析系统,使教师、管理员或系统开发人员能为学习者提供帮助。

许多应用实例显示,这些技术可以帮助识别那些不及格或有辍学风险的学生,进而确定干预措施。这些技术可以用作"推荐引擎",向学生推荐特定的课程或课程序列,以及可能适合他们的教学材料。总之,数据挖掘技术开发和应用模型要回答以下问题:学生何时应该进入下一学习主题或课程?什么时候需要干预?什么时候感到无聊或分心?什么时候他们有失败或不能完成课程的风险?在教学数据

挖掘的过程中,这些模型也可用于分析教学绩效(即教学分析),识别有效的教学模式并为低效率教学活动提供更有效的教学策略。

记者:您和您的团队多年一直从事 CANVAS 平台的教育大数据挖掘研究,目前取得了哪些进展?

雷克教授:越来越多的高等教育机构和学校正在使用学习管理系统支持教学。学习管理系统具有开展教和学活动的功能,例如访问教学内容、分发测验、评估教学以及支持师生间的在线交流和协作。近年来,随着大学生修读在线课程的增加,美国几乎所有大学都在使用学习管理系统。

学习管理系统提供了丰富的教育情境,包括不同的教学模式(在线、混合以及面对面)、各类学科(文科、理科和工科)以及不同类型的学生(全日制或非全日制)。它会自动存储教师和学生的所有在线交互数据,如学生查看任务发布、提交作业、下载附件等。这些隐形而又短暂的数据点,再加之大数据和教育数据挖掘技术的支持,能为研究高等教育教与学提供了前所未有的机会。

我们的研究旨在探索利用一系列分析技术有意义地应用于 Canvas[①]所搜集的复杂数据集。该数据集包含美国某中等规模公立大学四年所开设的 3300 门课程中所有师生在线活动的匿名数据。当学生和教师访问 Canvas 时,系统通过云存储方式保存所有浏览页面和互动日志。

Canvas 数据由一系列功能版块下的用户活动记录组成,包括教师和学生的观点和参与度,如作业、小测验、会议、讨论等。这些功能可归为四类:管理(如名单视图)、评价(如测验)、内容(如内容浏览)和参与度(如参加在线讨论)。

研究过程以建立好的知识发现与数据挖掘框架为指导。该框架由三个阶段组成:1)数据清理、预处理、选择和转换;2)数据挖掘、模型构建和模型选择;3)模型评估、解释和呈现。知识发现和数据挖掘的目的是把庞大而粗糙的数据集转换为紧凑和实用的数据集,以探索其中的模式。

第一阶段是转换数据,包括采用数据挖掘方法对数据进行清理(删除不需要的项目)、转换和改进(从现有数据中计算新属性)、集成和优化等,将原始数据处理成适当的形式,如排除交互较少的课程。此外,许多变量的分布都有偏差,特别是学生的最终成绩(超过 30% 的成绩是 A,低成绩组比例较低),违背了参数统计的基本假设。因此,我们采用各种转换策略,例如转换为 Z 分数[②],或根据分析目的将最终成绩分成等级,最后使用皮尔逊相关性检测与结果变量高相关的变量,然后去除或合并彼此高度相关的预测因子,避免多重共线性(Multicollinearity)[③]。

第二阶段是选择和建构模型。我们开发了流程和建模方法,应用预测、多项式逻辑回归以及聚类法,聚类使用期望最大化(expectation-maximization)和层次聚类分析(hierarchical clustering analysis)数据挖掘方法,并通过热图分析学生和教师的交互模式及其与学生学习成果间的关系。使用这些不同的模型考察微观的课程层面和宏观的跨课程层面中学生和教师的活动模式,研究模型结果与学生学习成果特别是学生最终成绩间的关系。使用的分析工具包括 SPSS、Weka、Tableau 和 R studio。

第三阶段是数据后处理(如模型评估、解释和呈现),包括比较模型、检查适配度以及将结果和数据之间的关系可视化,以发现新的见解。多项式逻辑回归结果如下:第一个多项式逻辑回归模型使用面对面学生的数据(N = 19,162),以"最低"成绩等级类别作为参照组。结果表明,九个学习管理系统功能是学生最终成绩的重要预测因子,"完成作业的数量"变量有最大且正向的比值比(odds ratio)。第二个多项式逻辑回归模型基于在线学生数据(N = 5279),再次使用"最低"成绩等级作为参考类别。结果表明,九个学习管理系统功能同样是学生最终成绩的重要预测因子,但"测验的数量" 变量有最大且正向的比值比。宏观层面的聚类结果如下:期望最大化算法基于教师和学生使用 Canvas 的功能输出三个群集。对群集行为和结果的研究表明,师生在 Canvas 活动水平越高,学生的最终成绩越高。微观层面的聚类结果如下:为了调查学生微观层面的活动模式及其与最终成绩的关系,我们选择同一位教师在面对面(N = 33)和在线(N = 36)两种模式下教授同一门课程。为了在课程模式中可视化地分析学生在学习管理系统的应用模式,我们构建了聚类图,将层次聚类分析与热图结合,并把所有学生数据都转换为 Z 分数以使方差标准化。使用层次聚类分析对行(学生)和列(Canvas 功能)进行聚类。我们还使用分析

比率或区间尺度数据(interval scale data)时最常用的欧几里得距离度量法(Euclidean distance measure)。对于聚类算法,我们选择平均链接(average linkage)聚类法,把两个集群间的距离定义为两个集群所有成员间的平均距离。结果显示了两种课程模式间的差异,即不同的模式侧重不同的学习管理系统功能,且学生使用的功能与其最终成绩相关。研究还发现,集群不是简单呈现整体平均数,而是提供了一幅丰富的情境画面,描绘了学生间的交互模式,以及交互模式如何与其他学生及最终成绩的关联。

记者: 您的研究遇到哪些挑战?有什么经验?

雷克教授:分析大型学习管理系统数据集有几个挑战。例如,很大的精力被用于数据预处理,特别是数据清理。而且,随着数据分析由简单趋向复杂,知识发现和数据挖掘过程是迭代的。另外,因错误的数据假设使得无法执行某些分析技术时,我们需要重新进行数据清理。数据预处理和建模过程需要执行多任务(如数据转换等)。建模方法需要综合利用统计学技术和机器学习技术,且要分析不同层次的数据。其他挑战还包括处理缺失样本(missing cases)和数据偏移(常见于 Internet 日志文件)。

我们感受最深的是,软件工具捕获的数据通常受底层技术驱动,而不是受教育问题驱动。由此,教育大数据挖掘着眼于解决重要的教育问题时才能彰显其价值。尽管可以依靠算法对大量数据进行计算机提取,但数据并非越多越好。研究人员需要对研究问题有明确的认识,必要时还须回溯到知识发现和数据挖掘阶段,密切关注不同分析技术的结果。基于学习管理系统的海量数据,我们应用的多种分析技术及得出的结论将有助于教育数据挖掘领域的发展。

记者:您刚才提到"教育数据挖掘研究如果没有任何学习与教学理论为基础,这是令人担忧的,没有理论,所得到的模型就难以解释。"为什么这么说?应该如何理解它?

雷克教授:教育数据挖掘领域发展迅速,主要来自于计算能力、数据存储和数据挖掘算法等技术的推动,这些技术也一直处于教育数据挖掘研究的前沿,却淡化了教育或学习理论的应用。此外,教育数据挖掘研究常着眼于学习环境中易于获取的数据(如内容浏览次数),而这些数据与具有教学价值的结构类型数据(如学习动机)存在很大差距。

针对这一差距,应用大数据分析工具的数据挖掘研究人员和正在解决重要教育问题的学习科学研究人员建立双向联系至关重要。学习科学研究人员感兴趣的结构类型数据通常是复杂的,并以特定理论为基础(如行为主义)。他们需要有效地交流和解释这些结构类型,并参与选择、建议和修改用于测量结构类型的指标。这将使数据挖掘研究人员能够应用算法开发有效的学习分析技术,并设计出有助于学习科学研究的信息可视化方法。

这一差距也表明教育中使用大数据的典型观念:认为"大数据"等同于"好数据",只要拥有更多的数据就会带来新的发现。这不一定是对的。前面已提到,利用大数据需要进行复杂而冗长的数据清理,更大量的数据只能增加工作量。其次,更多的数据不会改变信噪比(signal-to-noise ratio),如果数据集已有大量噪声,更多的数据只会增加噪声。

为便于开展教育研究,数据集具有更多的教育相关特性(数据集的列)比拥有更多数据量(数据集中的行)更重要。虽然传统的教育研究方法可能比较耗时费力,但它们经由精心设计、测试和修订,数据有利于捕捉其教育意义,并具有较高的信噪比。因此,教育数据挖掘研究应该检视分析过程,以最大化数据信号,减少数据集的维度,并保持计算的易处理性。

最后是避免错误发现,包括负误识(false negative)和正误识(false positive),二者都是数据挖掘算法没有正确分类的结果,负误识是将不及格学生归为及格类型,正误识是将及格学生归为不及格类型。例如,预测学生是否处于不及格风险状态这一难题。一般观念认为,大多数学生都会通过课程,所以不及格是低概率事件,负误识可能将有不及格风险的学生没有归为不及格风险状态,因此不会收到有效的干预。在正误识情况下,一个不太可能失败的学生被归类为不及格风险状态,这可能会对学生造成心理伤害。教育理论可以帮助减少这类错误。

记者: 您谈教育数据挖掘时,时常提到学习分析这个领域,这两个领域有什么联系?

雷克教授:与教育大数据应用相关的两个研究领域分别是教育数据挖掘和学习分析。这两个领域

有很大的相似之处,没有明确的区别。然而,它们正在沿着不同轨迹发展。

一般而言,教育数据挖掘采用新兴的机器学习和统计技术,对教育中收集的特定类型的大数据进行挖掘,从而产生新的洞见,服务于教育。收集的数据包括学生在教学系统中的行为(如浏览课程、提交作业、参加测试等)、线上互动(如论坛发帖、实时聊天、相互点赞等)以及一些不直接属于教学过程的行为(如收发邮件、发送短信、语音留言等)。教育数据挖掘重点关注开发新算法和模型,以便更好地了解学生、教师和学习发生的环境。

学习分析研究人员很大一部分是从教育数据挖掘领域分化出来的,他们关注利用数据分析支持学习理论发展、教育技术设计以及学习数据应用的伦理问题。2011 年,他们成立学习分析研究协会,同年举办首届学习分析与知识国际会议,并于 2013 年创办《学习分析杂志》期刊。与教育数据挖掘相比,学习分析测量、收集、分析和报告有关学习者及其学习环境的数据,目的在于认识和优化学习及其发生的环境,还以教育数据挖掘的结果来评价、预测和改进学生的表现,进而解决教学问题。

记者:教育大数据挖掘的未来发展是什么?

雷克教授:计算机科学、人工智能和统计建模的研究正在以前所未有的速度发展,并将继续引领教育数据挖掘的发展。

自然语言处理是其中快速发展的领域之一,其软件和硬件的应用通过神经网络在理解、描述和翻译自然语言方面取得了长足的进步。

另一个发展较快的领域是个性化学习,它采取以学生为中心而不是以变量为中心的方法。以往的教育研究主要分析众多属性的聚合数据,探究关键变量的特征和趋势,然后分析学生及其在不同组的差异。随着学生行为实时海量数据的搜集,采用以学生为中心的方法,研究学生交互模式随着时间的推移变得可行了。这种方法可以突显具有共同数据模式的学生类别,从而将这些学生分为一组。然后,这些类别可用于支持个性化体验,推送信息,并为教育技术与学生互动提供适应性服务。

在另一项关键工作中,研究人员使用细粒度学生数据检验学习理论,检测不同类型的教学实践和课程设计要素的有效性。例如,由于学生数量庞大,教育技术研发人员可以对大量学生进行快速测试,以反复改进在线学习系统,更好地为学生、教师和管理人员服务。尤其是,开发人员可以进行实验,随机分配学生接受不同的教学或学习方法,为他们部署不同的教育技术版本:A 版或 B 版,以快速产生数据,分析学习问题,并持续改进教学方法。

记者:您对中国开展教育大数据挖掘研究有哪些建议?

雷克教授:技术的飞速发展正推动教育领域发生前所未有的变革。鉴于这一现实,我的建议是:明确需要解决哪些教育问题,了解复杂的伦理和隐私问题,认识到教育数据挖掘是一种社会技术方法,其中,人力与组织问题和技术问题同等重要。

第一,教育机构要确定最迫切需要解决的问题。还是所有问题数据挖掘都能提供解决方案,因此,必须确定教育数据挖掘能够支持解决的紧迫问题,而且,不能迷失在数据、技术或者复杂的输出结果中,应通过基于证据的决策,为教育目标的实现作出贡献。

第二,如上所述,利用学生先前数据可以预测和分析其未来发展并提供干预措施。在此过程中,研究和其他人员应合理地利用数据挖掘模型,而不要基于对学生的需求和潜质做出错误推断。例如,一名学生可能被算法错误地贴上“不及格风险”的标签,这可能会产生严重的负面影响。更糟糕的是,算法可能只是盲目地根据父母的教育水平和高中成绩平均绩点等惯常指标描述学生特征。这样,一直以来的不公平现象就会简单地传播开来。因此,研究和管理人员有责任调查用于对学生、教师和教学做出重要决定的任何预测模型的有效性。他们必须了解并能够解释模型预测背后的证据以及学习管理系统下一步的行为。最后,他们应该知道教育数据挖掘可能有不容忽视的隐私问题,因为需要收集和存储大量的学生个人信息。另一个重要问题是持续完善用于数据聚合和分解的分析工具的匿名化,以保护个人隐私,并确保改进教育数据的应用。

第三,要认识到教育数据挖掘是社会技术系统的一小部分,正如人类决策和行动举措只是任何成功解决方案的一部分一样。总而言之,许多因素影

响教育大数据挖掘在教育中的成功应用,而且组织和人力方面的因素比技术因素影响大。这些因素包括:1)领导者及其组织必须致力于以证据为基础的决策;2)熟练掌握新兴的教育数据挖掘技术的工作人员;3)能够有效收集、挖掘、分析数据的技术平台;4)可以获取教育数据挖掘所要求的高质量的数据;5)可负担得起的技术平台,并展现出良好的投资回报(return on investment)。

[注释]

①Canvas是一种学习管理系统,具有生态系统的特性,方便教师和学生开展在线教学与学习活动。目前广泛应用于世界各地3000多所高校、学区以及教育单位。

②Z分数也叫标准分数(standard score),是一个数与平均数的差再除以标准差。

③多重共线性是指线性回归模型中的解释变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确。

[参考文献]

[1]Canvalytics Research Group(2016). Exploring learning management system interaction data: Combining data-driven and theory-driven approaches[A]. Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016) [C]. New York: ACM: 324-329.

[2]Recker, M., & Lee, J.(2016) Analyzing learner and instructor interactions within learning management systems: A review of approaches[A]. Spector, J. M., Lockee, B., and Childress, M. (Eds.), Learning, design, and technology. An International Compendium of Theory, Research, Practice, and Policy[C]. New York: Springer:1-23.

[3]Recker, M., Walker, A., & Lawless, K.(2003). What do you recommend? Implementation and analyses of collaborative filtering of Web resources for education[J]. Instructional Science, 31(4/5): 229-316.

[4]Recker, M., Yuan, M., & Ye, L.(2014). Crowd teaching: Supporting teaching as designing in collective intelligence communities [J]. International Review of Research in Open & Distance Learning, 15(4):138-160.

[5]Xu, B., & Recker, M.(2012). Teaching analytics: A clustering and triangulation study of digital library user data[J]. Educational Technology & Society Journal, 15(3):103-115.

[6]Yuan, M., & Recker, M.(2014). Characterizing user behaviors and resulting products in an online educational community: A comparison between novices and elders[J]. Journal of Learning Analytics, (3):150-153.

(编辑:徐辉富)

Model Driven Educational Big Data Mining for Enhancing Teaching and Learning: An Interview with Dr. Mimi Recker from Utah State University

YU Liang, YANG Qiuyan & ZHAO Nan

(*College of Computer and Information Science, Southwest University, Chongqing* 400715, *China*)

Abstract: *Mimi Recker is a professor of Emma Eccles Jones College of Education & Human Services at Utah State University and a former head of the Department of Educational Technology and Learning Science. She is currently a member of National Science Foundation Grant Advisory Board. Also, she is an editor of Educational Technology Research and Development and American Education Research Journal and has been an editor for Journal of the Learning Sciences and International Journal of Instructional Media. Her research interests cover the design and evaluation of the interactive learning environment, educational uses of new media, information, and communication technologies, and maker education. At present, she conducts research about maker education in schools and public libraries and educational big data mining. She has been PI for eight National Science Foundation Grants of USA and published more than* 30 *papers in the prestigious journals of educational technology, such as Journal of the learning sciences, Instructional Science, and Educational Technology Research and Development. She was nominated for the Outstanding Development Award of the American Association for Educational Communication and Technology (AECT) in* 2010 *and awarded the title of outstanding mentor at Utah State University in* 2013.

Key words: *educational big data; data mining; model; learning analytics*