

教学研究

大数据环境下精准教育的数学模型与若干问题

李波,王磊,王超

(华中师范大学 数学与统计学学院,湖北 武汉 430079)

摘要:大数据时代背景下,通过教育大数据研究促进教育发展、为教育决策提供支持成为未来教育的发展趋势。大数据环境下的精准教育主要研究如何根据学习者的个性化特征给出相应的教学方案,从而提高教学效果,真正实现“因材施教”。本文对教育大数据的两大主要研究领域,即教育数据挖掘(EDM)和学习分析(LA)进行了简要介绍。同时,本文对当前教育大数据研究领域构建的主要模型进行了总结,并在此基础上构建可研究精准教育中若干问题的一种通用模型——精准教育中的教学评价数学模型,且对该模型研究具体问题的方法进行阐述,着重介绍了以问题为导向的模型构建思路。

关键词:教育大数据;精准教育;教学评价模型;教学方案

中图分类号:O213;G40-051

文献标志码:A

文章编号:2095-3070(2017)04-0032-09

0 引言

随着科学技术的飞速发展,近年来全球产生的数据量呈爆炸式增长,人类已经进入“大数据”时代,大数据的出现将对社会各个领域产生深远影响^[1]。在教育领域,随着教育信息化的不断推进,越来越多的数字化智能系统、智能终端广泛应用于教育实践,教育过程中越来越多的信息被收集起来并以数据的形式呈现,包括学校管理数据、学生行为数据和学生学习数据等等。如何充分分析挖掘这些数据中包含的海量信息,并将其应用于教育实践,为教育决策、教育服务等提供支持,更好地促进教与学,是教育者、管理者 and 研究者密切关注的问题。美国教育部2012年发布了《通过教育数据挖掘和学习分析促进教与学》报告^[2],详细介绍了美国教育大数据应用领域以及应用实施所面临的挑战。我国教育部等五部委2014年在《构建利用信息化手段扩大优质教育资源覆盖面有效机制的实施方案》的通知中指出,到2020年形成与国家教育现代化发展目标相适应的教育信息化体系;信息技术与教育融合发展的水平显著提升;教育信息化对教育改革的支撑与引领作用充分显现^[3]。由此可见,利用大数据促进教育发展已经被提升至国家层面。

教育作为促进人的发展的复杂的社会实践活动,其发展涉及方方面面的因素,教育大数据促进教育发展的相关研究问题亦包括诸多方面,其重要应用之一是促进人们对教育过程的认识,为不同风格、不同认知水平的学习者提供精准服务和个性化学习支持。在教育实践中,学习者往往在文化背景、家庭环境、智力水平、感兴趣的领域等等方面存在差异,对不同教育方案的反应也不尽相同,为尽可能最大限度地促进每一位学习者的发展,真正实现“因材施教”,研究精准教育具有重要意义。在此背景下,精准教育中存在诸多问题尚待研究解决,如精准教学方案的评估和检测、影响精准教学方案的变量选择、最优精准教学方案的设计、多阶段最优精准教学方案的设计和学习者亚组分析等。这里的精准教学方案本质是一种决策函数,它可根据学生特征给出相应的教学策略,也就是说,精准教学方案并不是对所有学生一致地给出某一种教学策略,而是根据学生特征对不同学生给出有针对性的、个性化的教学策略。

收稿日期:2017-10-21

基金项目:湖北省高校教改教研项目(2016086);中央高校基本科研业务费项目(CCNU16A02012)

通讯作者:李波,E-mail:haoyoulibo@mail.ccnu.edu.cn

1 研究背景及现状

国内外教育大数据领域目前应用最多且相互最接近的研究分支是教育数据挖掘(educational data mining, EDM)和学习分析(learning analytics, LA)^[4]。学习分析研究专家 Siemens 和教育数据挖掘研究专家 Baker 详细分析了教育数据挖掘和学习分析的联系和区别^[5],他们认为二者均通过分析数据提高对教育问题的理解及干预计划的制定和选择,从而更好地指导和促进教与学,二者的共同目标是提高大规模教育数据的分析质量,支持教育领域的基础研究和教育实践;不同之处在于二者在侧重点、起源和分析方法方面存在差异。随着教育大数据的发展,教育数据挖掘和学习分析两大研究分支在研究内容上将出现更多的交叉,对应的两大研究主体也将开展更多的交流与合作。

1.1 教育数据挖掘研究现状

教育数据挖掘研究最早可追溯到 20 世纪 80 年代末,研究者开始将数据挖掘技术应用于教育领域;21 世纪以来,教育数据挖掘发展迅速,研究方法和研究成果日益丰富^[6]。教育数据挖掘杂志(Journal of educational data mining, <http://educationaldatamining.org/JEDM/>)于 2009 年创刊。当前,教育数据挖掘的应用主要包括学习者建模(student modeling)、决策支持系统(decision support system)、自适应系统(adaptive systems)、评估(evaluation)、科学探究(scientific inquiry)5 个方面^[7]。

教育数据挖掘建立的模型主要为预测模型(predictive model)和描述模型(descriptive model)两类,预测模型通过已知的数据去预测未知的数据,而描述模型则通过分析数据发现新的模式或结构^[6]。建立模型的数据挖掘技术包括分类、回归、聚类、关联规则和序列模式挖掘、文本挖掘等,具体应用的模型如贝叶斯模型、逻辑回归模型、逐步回归模型、马尔科夫模型等。Şen 等人用逻辑回归模型和神经网络、支持向量机和决策树进行了中等教育中影响学生在测试中成功或失败的因素,并比较了 4 种方法的效果^[8]。Feldman 等人运用朴素贝叶斯分类器,通过跟踪学习者在某益智游戏上的尝试次数、持续时间和最终等级的数据,研究学习者的学习风格^[9]。Zimmermann 等人采用线性回归模型,结合变量选择方法研究通过学习者本科阶段表现指标预测其研究生表现水平的问题^[10]。舒忠梅等人采用逐步回归和神经网络,基于经典的大学生发展理论,对大学生学习成果进行预测评价,并对大学生学习成果影响因素进行分析^[11]。蒋卓轩等人采用逻辑回归和支持向量机对慕课(MOOC)学习者学习行为数据进行分析,并预测学习者的学习结果^[12]。

1.2 学习分析研究现状

学习分析最早的定义源于美国高校教育信息化协会(EDUCAUSE)的“下一代学习挑战”(next generation learning challenges)计划,2011 年关于学习分析的重要国际会议——学习分析和知识国际会议(the international conference on learning analytics and knowledge, LAK)首次举办,迄今已成功举办七届。目前,国际学习分析的研究方向主要集中于学习分析服务框架、分析方法、工具与可视化工具领域;国内主要集中于综述和应用分析^[14]。学习分析主要应用领域包括用户模型、学习者行为与情感分析、预测与评估、结构发现与关系挖掘、趋势分析与活动干预、适应性和个性化以及监督与控制^[15]。

学习分析使用的模型主要包括贝叶斯知识追踪模型、预测模型、学习满意度模型、学习行为模型、学习动机模型、学习投入度模型和学习情感模型等^[16-18]。Slater 等人以贝叶斯知识追踪模型为依据,分析在线智能辅导系统中数学问题的语义,构建主题相关模型以实现自动识别数学问题内容、问题结构和考察知识点^[19]。Brown 等人构建预测模型,分析 566 名学生大学第一学年通识课程的相关数据,预测学生在早期预警系统的潜在分类及其成绩变化(提高或降低),为早期预警系统和高等教育课程设计提供支持^[20]。Kevan 等人提出了慕课满意度模型,将学生满意类型分为 4 方面,即内容满意度、社交满意度、非正式学习满意度和正式学习满意度^[21]。Park 等人提出学生在线行为监测模型,通过分析面授课堂和在线课堂中学生点击流量数据,监测学生行为的实时变化^[22]。舒忠梅等人构建学生投入模型,从大学生学习经历角度,在“生源—学习—成果”的逻辑框架中,分析研究型大学学生在校活动的投入模式和典型相关因素^[23]。

综上所述,当前教育大数据领域模型种类较多且各有侧重,针对的问题也不尽相同,探索建立统一

的、通用的模型与方法,简化学习分析的步骤和过程,是后续研究需要解决的重点问题^[17]。鉴于此,本文期望建立一种可研究教育大数据中多个问题的通用模型,丰富教育大数据的研究方法。本文构建精准教育中的教学评价数学模型,研究精准教学方案的评估和检测、影响精准教学方案的变量选择、最优精准教学方案的设计、多阶段最优精准教学方案的设计和学习者亚组分析等问题。

2 精准教育中的教学评价数学模型

教学效果是教育大数据研究的核心内容,事实上,教育大数据研究的目标正是提高教学效果。换言之,教学效果是衡量教学方案优劣的标准,也是指导教学方案设计的关键因素。现有研究中,教学效果的相关指标有学生学习成绩、学生学习满意度等。对其中的每一个指标,如何充分挖掘数据信息,尽可能准确地评价教学效果,是值得密切关注和深入研究的问题。为此,本文从所研究的问题出发,构建精准教育中的教学评价数学模型。模型建立的思想来源于统计学中的条件期望方法。

具体地,将所关心的教学效果记为 Y ,实际运用的教学策略以 A 表示,学生特征以 X 表示。教学策略 A 取值为自然数,例如 $A=1$ 表示一种新的教学策略,而 $A=0$ 表示原有的教学策略;抑或 $A=1,2,3,4$ 分别表示 4 种不同的教学策略。学生特征 $X=(X_1, X_2, \dots, X_p)$ 为与学生相关的 p 个特征指标构成的向量,这些指标包括学生的各类信息,如性别、年龄等基本信息,学习行为信息等。 (X, A, Y) 构成样本空间,设其分布为 P ,即有

$$(X, A, Y) \sim P,$$

在分布 P 下教学效果 Y 的期望记为 E 。于是教学评价数学模型为

$$E(Y|A, X),$$

其中, $E(a|b)$ 表示统计学中的条件数学期望,即给定 b 条件下 a 的数学期望。该模型的实际意义为对具有特征 X 的学生采用教学策略 A 后,产生的平均教学效果。

上述精准教育中的教学评价模型在理论和实践方面主要有以下优点:

第一,有利于充分挖掘数据信息。相比于大多数模型将学习者作为一个整体对教学效果进行考察,本文构建的教学评价模型针对不同学习者和不同教学策略研究教学效果。例如,对 n 位学习者 S_1, S_2, \dots, S_n ,其对应的教学效果指标为 Y_1, Y_2, \dots, Y_n 。若采用教育大数据研究中常用的回归分析,即使用 $E(Y)$ 评价教学效果,亦即将 n 位学习者视为同一样本中的个体。而本文构建的教学评价模型 $E(Y|A, X)$ 充分考虑 n 位学习者的不同特征及对应的教学策略,分层分类地研究学习者的教学效果,对教学效果的评价更为精细化,更有利于发现个性化规律,从而为精准教育提供支持。

第二,便于研究不同教学策略的教学效果。对于具有某类特征 x_0 的学习者,通过分析比较 $E(Y|A=a_1, X=x_0)$ 和 $E(Y|A=a_2, X=x_0)$,可研究两种教学策略 a_1 和 a_2 的教学效果。同理可研究多种不同教学策略的教学效果。该模型可降低由于学习者差异造成的教学效果干扰,有利于研究不同教学策略的教学效果。

第三,便于研究教学策略对不同学生的教学效果。为研究教学策略 a_0 的教学效果,通过分析比较 $E(Y|A=a_0, X=x_i)$,可研究该教学策略对不同学生产生的教学效果(i 表示不同学生类别)。由于学习者的个性化差异,不同学习者对教学策略的反应也不尽相同。针对不同学习者研究教学策略的教学效果有助于精准教育中精准教学方案的设计。

下文基于精准教育中的教学评价模型对精准教育中的若干问题,即精准教学方案的评估和检测、影响精准教学方案的变量选择、最优精准教学方案的设计、多阶段最优精准教学方案的设计和学习者亚组分析进行分析研究。精准教学方案本质是一种决策函数,它可根据学生的不同特征给出相应的教学策略。若将精准教学方案记为 d ,则 d 为学生特征 X 空间到教学策略 A 空间的映射,即有 $d(X)=A$ 。

2.1 精准教学方案的评估和检测

2.1.1 精准教学方案的评估

根据本文教学评价数学模型,对精准教学方案的评估即估计

$$E(Y|A, X)=E(Y|d(X)).$$

在实际应用中,教学效果 Y 与教学策略 A 以及学生特征 X 的联合分布往往是未知的,因此需运用数理统计知识对 $E(Y|A, X)$ 进行估计。目前已有不少相关研究,例如 Funk 等人提出一种双稳健估计(double robust estimation),该估计量基于对数密度,是一种无偏估计量^[24]。

2.1.2 不同精准教学方案的比较

比较两种个性化教学方案 $d_1(X)$, $d_2(X)$ 是否存在显著性差异,即比较其对应的教学效果 $E(Y|A=d_1(X), X)$ 和 $E(Y|A=d_2(X), X)$ 是否存在显著性差异。不妨记 $d_1(X)=1$, $d_2(X)=0$, 若 $d_1(X)=1$ 为一种新的教学方案,而 $d_2(X)=0$ 为原有的教学方案,则为研究新的教学方案是否有助于显著提升教学效果,可进行假设检验。具体地,记

原假设 $H_0: E(Y|A=0, X) = E(Y|A=1, X)$;

备择假设 $H_1: E(Y|A=1, X) > E(Y|A=0, X)$ 。

应用数理统计知识构造检验统计量,并进行计算。若原假设被拒绝,则认为新教学方案 $d_1(X)$ 的教学效果显著优于原教学方案 $d_2(X)$,应在实践中推广使用新教学方案 $d_1(X)$;否则认为没有足够的理由说明新教学方案和原教学方案的教学效果存在显著差异。

2.2 影响精准教学方案的变量选择

教学过程是一个复杂的过程,教学效果受到诸多因素的影响,因而精准教学方案的设计也与诸多方面有关。研究精准教学方案的影响因素有助于加深对教学过程的认识以及优化教学方案的设计,从而更好地指导教学。随着信息技术的飞速发展和教育信息化的不断推进,获取大量教学过程数据及其他相关数据成为可能。这些数据包括学生基本特征数据,如性别、家庭情况、智力水平、身体素质、心理承受能力和学习能力;学生行为数据,如阅读量、知识结构、作业成绩、消费金额、早起频率;学生情感数据,如情绪状态、孤独指数等。在众多变量中,哪些变量会影响教学效果,进而影响精准教学方案设计呢?这就需要进行变量选择,变量选择就是从众多变量中找出影响精准教学方案设计变量的方法。

本文基于回归分析构建精准教学方案的影响因素模型。若变量个数为 p ,将各变量依次记为 X_1, X_2, \dots, X_p ,则有

$$E(Y|A, X) = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p,$$

其中, $\beta_1, \beta_2, \dots, \beta_p$ 为各变量对应的回归系数。

常用的变量选择方法主要有3类,分别是子集选择法、正则化方法和变量降维方法。下文对这3种方法进行逐一介绍。

2.2.1 子集选择法

子集选择法是通过不断添加或删除变量,直到所得模型满足一定标准的方法,主要包括最优子集法、向前法、向后法和逐步回归法。常用的变量选择标准有 AIC 准则、BIC 准则、 C_p 准则以及交叉验证等。最优子集法最终选择的模型中包含原始变量集合中的部分变量。

2.2.2 正则化方法

正则化方法的思想是在最小二乘法的基础上引入惩罚函数,将某些变量的系数压缩至0或接近0,从而实现变量选择。各变量的系数为

$$\beta = \arg \min_{\beta} |Y - \beta X|^2 + p(\beta, \lambda),$$

其中: $\beta = (\beta_1, \beta_2, \dots, \beta_p)$; $X = (X_1, X_2, \dots, X_p)$; λ 为调整参数,常通过交叉验证的方式确定; $p(\cdot)$ 为惩罚函数。根据惩罚函数的不同,正则化方法具体有岭回归方法^[25]、LASSO 方法^[26]、SCAD 方法^[27] 和弹性网方法^[28]等。

2.2.3 变量降维方法

降维方法顾名思义就是降低变量维数的方法,与子集选择法和正则化方法直接从原始变量集中选择变量不同,降维方法多是通过原始变量进行特定的变换来完成。常用的降维方法主要有主成分回归和偏最小二乘法等。

主成分回归是将主成分分析和回归分析结合起来运用的一种方法。主成分分析通过对原始变量进行线性变换,得到一组线性无关的变量,成为主成分。根据线性变换知识,可得到与原始变量相同个数

的主成分。但事实上,通常选取几个主成分就可提取出原始数据的绝大部分信息。实际应用中通常以方差贡献率为标准确定选取主成分的个数。

偏最小二乘法^[29]通过构建少量与因变量相关性很大的正交潜变量,并将因变量与这些潜变量进行回归分析。具体地,潜变量 Z_k 是自变量 X_1, X_2, \dots, X_p 的线性组合。确定前 k 个潜变量后, Z_{k+1} 是与它们正交且与因变量有最大相关系数的线性组合。据此可由原始变量构建新的潜变量,实现降维。

2.3 最优精准教学方案的设计

最优精准教学方案的设计是教学过程的重要问题。为尽可能提高教学效果,真正实现“因材施教”,必须根据学生特征,对不同学生给出与其特征相适应的教学方案。教学理论中有关于教学策略的研究,教师在教学实践中也会积累一定教学策略,如何在其中找到最优的教学策略,如何提出新的有助于提高教学效果的策略,都是教学过程中需要考虑的重要问题。

若将最优精准教学方案记为 $d^*(X)$, 则有

$$d^*(X) = \arg \max_d E(Y|A=d(X), X),$$

即最优精准教学方案是使得教学效果条件期望最大的教学方案。这是一个优化问题。

寻找最优精准教学方案的一种方法是基于对 $E(Y|A=d(X), X)$ 的估计,在一系列可能的教学方案中搜索使其最大的教学方案。对 $E(Y|A=d(X), X)$ 的估计可采用多种模型,例如线性模型、对数回归模型以及非参数模型和半参数模型等。Zhang 等人提出一种关于 $E(Y|A=d(X), X)$ 的双稳健估计量,并在此基础上找到最优方案 d^* 。估计量的双稳健性保证其不受模型误设的影响,也就是说,即使模型选取有误也能够通过算法对模型进行修正^[30]。

另一种方法是这一寻找最优解的优化问题进行转化。Zhao 等人提出一种基于分类的加权机器学习方法^[31]。文中指出之前的最优方案选择方法更像是寻找某种模型下的最优方案,即满足某种限定条件的最优方案,因而提出了一种直接寻找最优方案的方法——分类加权机器学习算法。延续前文的记号,设样本空间 (X, A, Y) 服从分布 P 。对于某种确定的教学方案 $d(X)$,将使用教学方案 $d(X)$ 对应样本空间的分布记为 P^d ,相应的数学期望记为 E^d 。假设 $A=a \in \{1, -1\}$,假定有 $P(A=a) > 0$,那么 P^d 将关于 P 绝对连续,并且有

$$dP^d/dP = I(A=d(X))/P(A=a),$$

其中, $I(\cdot)$ 为示性函数。从而在给定一种教学方案 $d(X)$ 条件下,其教学效果的期望可表达为

$$E^d(Y) = \int Y dP^d = \int Y \frac{dP^d}{dP} dP = E\left[\frac{I(A=d(X))}{A\pi + (1-A)/2} Y\right],$$

其中, $\pi = P(A=1)$ 。于是问题转化为

$$\begin{aligned} d^* &= \arg \min_d E\left[\frac{I(A \neq d(X))}{A\pi + (1-A)/2} Y\right] = \\ &= \arg \min_d E\left[\frac{Y}{A\pi + (1-A)/2} I(A \neq d(X))\right] = \\ &= \arg \min \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{A_i\pi + (1-A_i)/2} I(A_i \neq d(X_i)). \end{aligned}$$

上式可解释为,如果样本中对某个个体试用策略 A 能够得到较好的教学效果,那么更多地考虑最优决策方案针对这种个体类型给出的策略应当是 A 。Zhao 等人指出这一类优化问题可利用支持向量机(SVM)来解决。

关于寻找最优精准教学方案还有其它方法,如 Fan 等人提出了一种基于和谐函数的学习方法^[32]。

2.4 多阶段最优精准教学方案的设计

教学过程是一个长期的过程,具有其特定的阶段性,每个月、每学期、每学年,甚至 K-12 中的不同时期都是一个阶段。随着时间的推移,教学过程的不断推进,学生的能力水平、自身需求、学习行为等特征也会发生改变。因此,在不同阶段,学生特征不同,精准教学方案也应进行相应的调整。另一方面,教学过程中的各个阶段又相互联系,相互依赖,相互影响,对应的各阶段的教学方案也不是孤立的。因此,研究相对长期的精准教学方案设计应当考虑到教学过程的阶段性,这就需要研究多阶段最优精准教学

方案的设计。此时精准教学方案的设计问题变成一个动态学习的问题。

设研究的教学过程共有 K 个阶段, 则各阶段对应数据为

$$X_0, A_0, X_1, A_1, \dots, X_K, A_K, Y,$$

其中: A_j 为阶段 j 采取的教学策略; X_j 为学生在阶段 j 之前的教学策略 $(A_0, A_1, \dots, A_{j-1})$ 影响下的特征, 即有

$$X_0 \xrightarrow{A_0} X_1 \xrightarrow{A_1} \dots \xrightarrow{A_{K-1}} X_K \xrightarrow{A_K} Y.$$

于是问题转化为如何根据学生的行为变化来调整最优个性化教学方案, 即找到最优动态精准教学策略 $d_0^*(H_0), d_1^*(H_1), \dots, d_K^*(H_K)$, 使得

$$E(Y|A_0=d_0^*(H_0), A_1=d_1^*(H_1), \dots, A_K=d_K^*(H_K))$$

最大化。其中:

$$H_0 = X_0; H_j = [X_0, d_0^*(H_0), d_1^*(H_1), \dots, d_{j-1}^*(H_{j-1}), X_j], j=1, 2, \dots, K.$$

这类动态优化问题常用基于反向递推的动态编程(dynamic programming)方法解决, 主要有 Q-learning^[33], A-learning^[34] 和 Deep A-learning^[35] 方法。

Q-learning(Q 代表“Quality”)基于反向递推的思想, 首先考虑最后阶段即第 K 阶段的优化问题, 即

$$d_K^*(H_{K-1}) = \arg \max_{d_K} E(Y|d_K(H_{K-1})),$$

而后, 向前递推第 $K-1$ 阶段的优化问题, 即

$$Y^{K-1} = E(Y|H_{K-1}, d_K^*),$$

$$d_{K-1}^*(H_{K-2}) = \arg \max_{d_{K-1}} E(Y^{K-1}|d_{K-1}(H_{K-2})).$$

依次递推直至全部 K 阶段最优决策方案完成。最终得到的多阶段最优精准教学方案为

$$d^* = (d_1^*, d_2^*, \dots, d_K^*).$$

Murphy 在其文章中证明了基于这种反向递推原理得到的决策方案可以实现多阶段动态方案的最优选择目标。

Q-learning 方法在应用中需建立模型对 $E(Y|d_{j+1}(H_j)), j=1, 2, \dots, K-1$ 估计, Watkins 提供一种线性模型。A-learning 方法和 Deep A-learning 方法都是在 Q-learning 方法基础上不断改进而衍生出来的方法。

2.5 学习者亚组分析

为促进教学效果的不提高, 教育研究和教育实践中往往需进行教学策略的改革, 研究新的教学策略。当新的教学策略被提出来之后, 一般要进行试验以确定其是否有效, 或者是否对某些学生群体有效。一种新的教学方案未必能够适应所有学生的学情, 但是或许适应于某些学生的学情, 有助于帮助他们提升自我学业成绩。学习者亚组分析研究的就是这一问题, 其目的在于找到某种新的教学方案的目标人群。具体地, 学习者亚组分析研究以下两个子问题。

1) 令 $A=1$ 表示新的教学策略, $A=0$ 表示原有的教学策略。是否存在目标人群 S , 使得当 $X \in S$ 时有

$$E(Y|A=1, X) > E(Y|A=0, X).$$

2) 如果目标人群 S 存在, 如何找到 S 。

可以看到, 如果目标人群 S 存在, 那么对这样一个人群中的学生来说, 使用新的教学方案会比原有的教学方案带来的教学效果要好, 即表示新的教学方案有其适用的学生群体存在。问题 1) 可通过数理统计中的假设检验方法解决, 针对问题 2), Fan 等人提出两种寻找方法, 即极值检验方法和模型选择法。

极值检验法的思想是判定新策略下的最优教学效果是否显著, 对于具有特征 X 的学生, 将实验组即采用了新教学策略的组与采用原有教学策略的组进行对比, 比较其教学效果的最优值, 从而排除其所带来的偶然性因素的影响。具体地, 将采用新的教学策略的学生的最优教学效果记为 $Y_{1, \max}$, 将采用原

有教学策略的学生的最优教学效果记为 $Y_{0,\max}$, 比较前者的分布显著优于后者(如超出 2 倍方差)。若新教学方案下的最优教学效果相比于原有教学方案下的最优教学效果的分布为异常值, 则表明新的教学策略对于具有特征 X 的学生为有效策略, 即分析得到具有特征 X 的学生即为使用新的教学策略的目标人群。

3 总结与展望

本文首先介绍了当前教育大数据应用领域最常用的 2 种分析研究分支和部分研究成果, 即教育数据挖掘和学习分析, 对已有研究中使用的主要模型及其研究现状进行了总结。在此基础上, 本文结合大量参考文献, 运用统计学知识, 构建一种通用的教育大数据研究模型——精准教育中的教学评价数学模型, 对精准教学方案的评估和检测、影响精准教学方案的变量选择、最优精准教学方案的设计、多阶段最优精准教学方案的设计和学习者亚组分析等问题进行研究。

大数据时代背景下, 教育大数据挖掘和学习分析等相关研究正受到越来越多的关注, 教育大数据应用研究方兴未艾。教育大数据研究面临着政策、资源和技术等多方面的机遇与挑战, 深入理论研究、拓展应用领域、多学科协同将是其未来发展的趋势。随着信息技术的飞速发展, 教育大数据研究将更好地促进教育理论和教育实践的健康发展。

参考文献

- [1] UN Global Pulse. Big data for development: Challenges & opportunities [DB/OL]. (2012-05-01). <http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGloabalPulseMay2012.pdf>.
- [2] Bienkowski M, Feng M, Means B. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief[R]. Washington: Office of Educational Technology, U. S. Department of Education, 2012: 1-77.
- [3] 教育部、财政部、国家发展改革委、工业和信息化部、中国人民银行关于印发《构建利用信息化手段扩大优质教育资源覆盖面有效机制的实施方案》的通知 [DB/OL]. (2014-11-16). http://www.moe.gov.cn/srcsite/A16/s3342/201411/t20141124_179124.html.
- [4] Van Barneveld A, Arnold K E, Campbell J P. Analytics in higher education: Establishing a common language[J]. EDUCAUSE Learning Initiative, 2012.
- [5] Siemens G, Baker R S J. Learning analytics and educational data mining: towards communication and collaboration [C]//Proceedings of the 2nd international conference on learning analytics and knowledge. ACM, 2012: 252-254.
- [6] 周庆, 牟超, 杨丹. 教育数据挖掘研究进展综述[J]. 软件学报, 2015, 26(11): 3026-3042.
- [7] Bakhshinategh B, Zaiane O R, ElAtia S, et al. Educational data mining applications and tasks: A survey of the last 10 years[J]. Education and Information Technologies, 2017(22): 1-17.
- [8] Şen B, Uçar E, Delen D. Predicting and analyzing secondary education placement-test scores: A data mining approach [J]. Expert Systems with Applications, 2012, 39(10): 9468-9476.
- [9] Feldman J, Monteserin A, Amandi A. Detecting students' perception style by using games [J]. Computers & Education, 2014, 7(3): 14-22.
- [10] Zimmermann J, Brodersen K H, Heinemann H R, et al. A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance[J]. Journal of Educational Data Mining, 2015(7): 151-167.
- [11] 舒忠梅, 屈琼斐. 基于教育数据挖掘的大学生学习成果分析[J]. 东北大学学报: 社会科学版, 2014, 16(3): 309-314.
- [12] 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, 52(3): 614-628.
- [13] 宗阳, 孙洪涛, 张亨国, 等. MOOCs 学习行为与学习效果的逻辑回归分析[J]. 中国远程教育, 2016, 36(5): 14-22.
- [14] 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): 11-19.
- [15] 王良周, 于卫红. 大数据视角下的学习分析综述[J]. 中国远程教育, 2015(3): 31-37.
- [16] 曹帅, 王以宁, 徐鹏. 学习分析技术的研究现状与未来趋势——基于 2011—2015 年 LAK 会议论文的分析[J]. 中国电化教育, 2016(5): 78-84.
- [17] 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.

- [18]吴永和,李若晨,王浩楠.学习分析研究的现状与未来发展——2017年学习分析与知识国际会议评析[J].开放教育研究,2017,23(5):42-56.
- [19]Slater S, Baker R, Almeda M V, et al. Using correlational topic modeling for automated topic identification in intelligent tutoring systems [C]//Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, British Columbia, Canada, ACM:393-397.
- [20]Brown M G, Demonbrun R M, Lonn S. What and when: the role of course type and timing in students' academic performance [C]//Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, Edinburgh, United Kingdom, ACM:459-468.
- [21]Kevan J M, Menchaca M P, Hoffman E S. Designing MOOCs for success: A student motivation-oriented framework [C]//Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, Edinburgh, United Kingdom, ACM:274-278.
- [22]Park J, Denaro K, Rodriguez F. Detecting changes in student behavior from clickstream data [C]//Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, British Columbia, Canada, ACM:21-30.
- [23]舒忠梅,徐晓东,屈琼斐.基于数据挖掘的学生投入模型与学习分析[J].远程教育杂志,2015(1):39-47.
- [24]Funk M J, Westerich D, Davidian M, et al. Doubly robust estimation of treatment effects[J]. Analysis of Observational Health Care Data Using SAS, 2010:85-103.
- [25]Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1):55-67.
- [26]Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1):267-288.
- [27]Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001(96):1348-1360.
- [28]Zhou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Methodological), 2005(67):301-320.
- [29]Tobias R D. An introduction to partial least squares regression[R/OL]. Technical Report at SAS Institute, <http://support.sas.com/rnd/app/stat/papers/pls.pdf>.
- [30]Zhang B, Tsiatis A A, Laber E B, et al. A robust method for estimating optimal treatment regimes[J]. Biometrics, 2012, 68(4):1010-1018.
- [31]Zhao Y, Zeng D, Rush A J, et al. Estimating individualized treatment rules using outcome weighted learning[J]. Journal of the American Statistical Association, 2012, 107(499):1106-1118.
- [32]Fan C, Lu W, Song R, et al. Concordance-assisted learning for estimating optimal individualized treatment regimes [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2017, 79(5):1565-1582.
- [33]Watkins C J C H, Dayan P. Q-learning[J]. Machine learning, 1992, 8(3/4):279-292.
- [34]Murphy S A. Optimal dynamic treatment regimes[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2003, 65(2):331-366.
- [35]Jiang R, Lu W, Song R, et al. On estimation of optimal treatment regimes for maximizing t -year survival probability [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2017, 79(4):1165-1185.
- [36]Fan A, Song R, Lu W. Change-Plane analysis for subgroup detection and sample size calculation[J]. Journal of the American Statistical Association, 2017:769-778.

Mathematical Model of Precision Education and Several Problems in Big Data Environment

LI Bo, WANG Lei, WANG Chao

(School of Mathematics and Statistics, Central China Normal University, Wuhan, Hubei 430079, China)

Abstract: Under the background of big data era, it is the development trend of future education to promote education and to support education decision-making through research on big data in education. Precision education in big data era mainly studies how to give corresponding teaching plans according to learners' individual characteristics so as to improve teaching

effect and truly realize "teaching students according to their aptitudes". This article briefly introduces two main research methods of big data in education, namely, education data mining (EDM) and learning analysis (LA). This article also summarizes the main models used in current research field of big data in education, and builds a general model that can study several problems in precision education — the mathematical model of teaching evaluation in precision education. The application method of the model in specific problems is presented, which focuses on the construction idea of a problem-oriented model.

Key words: big data in education; precision education; model of teaching evaluation; teaching plan

作者简介

李 波(1977—),男,博士,华中师范大学数学与统计学学院教授,研究方向为教育大数据、应用统计等。

王 磊(1993—),女,在读研究生,研究方向为教育大数据。

王 超(1988—),男,在读研究生,研究方向为教育大数据。

(上接第 23 页)

Dynamic Analysis and Synchronization Control of the Discrete Coupled Complex Network Model with Delay and Non-delay

LI Yin, WEI Ruiying

(School of Mathematics and Statistics, Shaoguan University, Shaoguan, Guangdong 512005, China)

Abstract: Complex network is widely present in everyday life. First, several types of standard network models are given. Second, the optimization control of coupled complex network models with delay and non-delay are designed and implemented by the stability control method. Then, we discuss the projective synchronization problem of coupled complex model via constructing Lyapunov function, and obtain the globally stable conditions of the error system and the effective synchronization controller. Numerical simulation verifies the feasibility and effectiveness of this method.

Key words: the discrete network model with delay; projective control; stability theory

作者简介

李 银(1980—),男,博士,副教授,主要研究方向是非线性动力学。

位瑞英(1981—),女,博士,副教授,主要研究方向是偏微分方程及其应用。