文章编号:1006—9860(2018)02—0059—08

面向智慧课堂的数据挖掘与学习分析框架及应用

孙曙辉[1]，刘邦奇[1,2]，李 鑫[3]

(1.讯飞教育信息化研究院，安徽 合肥 230088；2.首都师范大学 教育学院，北京 100048；3.科大讯飞大数据研究院，安徽 合肥 230088)

摘要：国内外对教育大数据的研究正从概念、理念层面走向建模分析与应用层面，而应用层面的研究也从教育质量监测统计、教育决策等宏观层面向学校教学、学生学习等微观层面深入发展。该文在教育数据挖掘与学习分析内涵讨论的基础上，结合智慧教学实际提出了智慧课堂的数据模型和体系架构，阐述了“四建模三分析”的教育大数据研究方法论，构建了智慧课堂数据挖掘分析的“整体棋盘”及13个具体研究问题，并提出了智慧课堂数据挖掘分析的四类应用模式，最后基于真实数据探讨了学生主观行为对成绩的影响分析应用案例。

关键词：智慧课堂；教育大数据；数据挖掘；学习分析；师生互动指数

中图分类号：G434 **文献标识码：**A

一、引言

教育大数据是教育过程中产生的或依据教育需求采集到的一切可用于教育发展的数据集合[1]。目前进行教育大数据分析应用正在向教与学聚焦[2]，主要包括教育数据挖掘和学习分析两个方向。教育数据挖掘(Educational Data Mining, EDM)[3]是综合运用统计学、机器学习算法和数据挖掘技术，对教育大数据进行处理和分析，通过建模发现学生学习结果与学习内容、学习资源和教学行为等变量的相互关系，进而预测学生未来的学习趋势。而学习分析(Learning Analysis，LA)[4]则是利用松散耦合的数据收集工具与分析技术，研究分析学生学习参与、学习表现、学习过程的相关数据，运用不同的分析方法和数据模型来解释这些数据，根据解释结果探究过程与情境，为其提供相应的反馈进而促进有效学习。相较而言，教育数据挖掘主要侧重于找出规律，即解决“为什么、是什么”的问题；而学习分析则侧重于应用发现的规律，即落实“如何用”的场景。教育数据挖掘是针对学生进行行为建模与学习趋势预测；而学习分析是利用分析得到的结果指导学习，直接将反馈作用于判别与决策。在实际的教育大数据分析中，我们往往更多的采用归纳性方法来挖掘教育共性规律，采用异常发现来对待个性化需求，并使用演绎性方法来为发现的共性与个性规律寻找适用的应用场景，从而促使有效学习的达成。可见，教育数据挖掘与学习分析为我们应用教育教学领域的大数据规律、开展课堂的教与学应用，提供了完整的思路和方法。

总体上来说，国内外对教育大数据的研究正从概念、理念层面走向建模分析与应用层面，而应用层面的研究也从教育质量监测统计、教育决策等宏观层面向学校教学、学生学习等微观层面深入发展。利用教育数据挖掘分析为受教育者量身定制教育目标、计划、方案、资源，有助于实现“因材施教”，为个性化教学指明方向。近几年国内一些学者结合教学过程的应用开展教育大数据相关研究[5][6]，如从“微课”[7]“慕课”[8]“翻转课堂”[9]等典型应用入手探讨大数据对教育模式转变、教学方式变革的影响等理论研究，针对学习行为数据利用数据挖掘算法和学习分析技术围绕学生进行建模与预测，进行课堂教学的大数据应用研究[10-12]。基于课堂教学行为数据并运用领域知识模型构建技术，研究课堂师生互动、生生互动的实时联系，有助于揭示深层次教学规律，为改进教学和提升教学质量提供依据。本文从微观层面上对目前的热点“智慧课堂”进行大数据挖掘分析专题研究，提出面向智慧课堂的数据挖掘和学习分析框架与应用模型。

二、智慧课堂模式概述

(一)智慧课堂的定义

智慧课堂是基于新一代信息技术打造的智能、高效的课堂[13]，是信息化课堂发展的新形态。目前对教育信息化的研究不断向课堂、向教与学聚焦，智慧课堂成为实施智慧教育的核心载体，也是当前

学校信息化教学改革和企业教育信息化研发推广所关注的焦点。无论是学术界对智慧课堂概念的理解还是实践中智慧课堂的具体构建都没有唯一的模式，比如基于物联网技术的“智能课堂”[14]，基于电子书包应用的“智慧课堂”系统[15]，基于云计算和网络技术应用的“智慧课堂”[16]，等等。

关于智慧课堂的定义总体上有从教育的视角和从信息化的视角两种类型，本研究是从信息化的角度进行探讨。我们曾从信息化的视角系统梳理了当前各种智慧课堂概念或模型，在此基础上对“智慧课堂”提出了一个明确的定义[17][18]：即以建构主义学习理论为依据，利用大数据、云计算、物联网和移动互联网等新一代信息技术打造的，实现课前、课中、课后全过程应用的智能、高效的课堂。基于信息化视角的智慧课堂概念具有鲜明的技术特征：(1)教学决策数据化，即基于智慧课堂教学过程的海量行为数据进行决策分析，在课堂教学中实现了基于数据的教育；(2)学习评价即时化，智慧课堂采取伴随式数据采集与评价，贯穿于课前、课中、课后全过程，进行即时的学习诊断、评价与反馈；(3)交流互动立体化，基于“云网端”平台，实现师生之间、生生之间、教师学生与家长之间，全时空无障碍地立体化沟通、交流；(4)资源推送智能化，依据学生学习行为数据记录和分析，智能化地推送微课、作业等学习资源，满足学生个性化、多样化学习需求；(5)教学呈现可视化，利用学科思维导图、模型图、虚拟现实、增强现实等可视化技术，把本来不可见的“思维”、难以展现的复杂实验过程形象化地呈现出来。

(二)用于研究的智慧课堂平台

科学研究需要真实、具体的数据为基础。基于研究的需要，我们选取了在当前中小学使用较普遍的科大讯飞知名产品“智慧课堂”(以下除非特别说明，智慧课堂均指科大讯飞的智慧课堂产品，简称“智课”)作为研究的支撑平台。该产品以建构主义理论为依据，结合诸如“翻转课堂”“互动课堂”“混合式学习”等先进教学理念，建立“云网端”课堂信息化平台(简称“智课平台”)，帮助师生课前轻松备课、预习，课上移动教学，课后个性学习、辅导。该产品已形成了理论定义、系统组成、教学模式、应用案例的完整体系[19]。

智课平台是由“云”“网”“端”构成的一体化课堂信息化平台[20]。其中云平台主要包含资源管理与服务系统、作业与动态评价系统和微课制作与应用系统等核心应用系统；微云服务器主要实现教室内构建以教室为单元的局域网信息化环境，提供本地网络、存储和计算服务；端应用工具即移动智能终端，是实现智慧课堂教学应用的基本工具。智慧课堂移动端工具基本配备包括教师端、学生端，根据需要也可配备家长端、管理端。教师端工具提供教师课堂教学的基本手段，主要包括PPT制作与投屏、微课制作与发布、互动交流和学习测评等功能，可以进行电子白板式的任意书写、记录与保存，实现任务布置、作业批改、答疑解惑、个别辅导等师生互动。学生端工具包括微课学习、课堂互动交流、作业与动态评价等主要功能，可以进行微课的学习、参与课堂师生、生生互动、完成个性化作业、查看学习成绩等。

三、智慧课堂用户模型及行为数据

(一)智慧课堂“三角用户模型”

对智慧课堂数据挖掘分析，首先要建立智慧课堂的数据模型。从信息系统的视角来分析，智慧课堂教学实际上就是教师、学生借助于信息媒介进行信息交换、传递、接受、互动的信息过程。在智慧课堂教学中，教师与学生是教学信息过程的两个主要参与者，是产生信息、处理和使用信息的主体，是课堂信息系统的活力源泉。通过对智慧课堂信息数据的梳理以及对智慧课堂产品原型的还原，我们可以抽象得到(如图1所示)智慧课堂的“三角用户模型”，用以对智慧课堂用户交互关系进行系统描述。

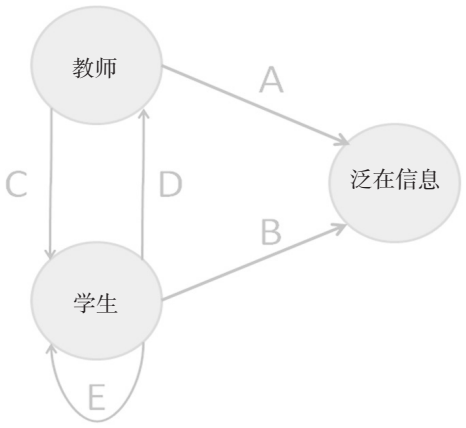

图1 智慧课堂“三角用户模型”

具体来说，智慧课堂信息系统模型构成包含两个主要参与者与一个信息对象。两个主要参与者分别是教师与学生，一个信息对象指的是由师生产生的UGC微课视频资源、各种网络互动信息、文本信息等，我们统称为泛在信息。在“三角用户模型”中，总共有五大类行为，分别是教师生成、读取泛在信息(A)，学生生成、读取泛在信息(B)，教师与学生间的互动(C、D)，学生与学生间的互动(E)。其中教师与学生的互动

C是指由教师主动发起的互动，D是指由学生主动发起的互动。这五大类行为分别对应的具体行为列表及数据中体现的发生频率如表1所示。

表1 智慧课堂用户行为数据及发生频率

行为类别	具体行为	发生频率
A 教师生成、读取泛在信息	教师发布微课视频	中
	教师发布班级通知	中
	教师发布分享圈帖子	低
	教师发布作业	高
	教师举报分享圈帖子	低
B 学生生成、读取泛在信息	学生发布微课视频	低
	学生发布分享圈帖子	低
	学生浏览、点评微课视频	低
	学生浏览、点评分享圈帖子	低
	学生提交作业	高
	学生举报分享圈帖子	低
C 教师对学生的互动行为	教师在微课、分享圈中回复学生	低
	教师回复学生私信	中
	教师批改学生作业	高
D 学生对教师的互动行为	学生购买、观看、点赞、回复、收藏、反馈微课视频	高
	学生回复教师分享圈帖子	低
	学生给教师发布私信	中
E 学生间互动行为	学生互批作业	高
	学生间在微课视频、分享圈中互动	中

(二)智慧课堂互动数据分析

师生互动是智慧课堂的核心标志。由表1可知，师生互动数据是智慧课堂中发生频率较高的教学行为数据。教师与学生的互动主要围绕作业、微课视频、分享圈帖子以及私信三类载体展开。从信息的流向来看，可以分为一对一或一对多，而从信息的私密性来看，可以分为公开信息与私密信息。师生互动数据的具体内涵及潜在价值分析如下：

作业：作业传递是智课平台的基本功能，是由教师发布、学生提交为形态的一对多的公开信息。作业的频次与难度部分反映了教师的教学风格，而学生完成作业的及时性、正确率则是学生学习态度与学业能力的体现。

微课视频：是由个人(老师或学生)录制并供智慧课堂用户课前或课后观看的一种信息载体形式，它是智慧课堂教学内容建设的核心。微课视频在智慧课堂中是由老师、学生共同参与的一对多(一个微课视频只有一个作者、上传者)的公开信息。通过微课发布的频次、包含的知识点可以了解教师上课的进度与状态，借助学生对微课视频的浏览、观看、回复、点赞、收藏以及其中一些付费视频的购买行为，可以进一步衡量教师的影响力与学生对课程、教师的满意程度。学生学习微课的先后顺序则可以用来构建学生学习的知识图谱，并进一步用以比较不同学生之间的学习差异。

帖子与私信：是同一种网络文本信息的两种不同形态，帖子是一对多的公开信息，而私信则是一对一的私密信息。教师与学生，学生与学生间通过帖子、私信相互交流、互动，反映了师生、生生间的联系强弱。基于此类数据，可以构建校园的社交网络地图，进一步分析网络中影响力用户以及基于文本进行校园舆情洞察。

四、智慧课堂建模与挖掘分析体系框架

(一)“四建模三分析”框架

为了系统性地对智慧课堂中的教学行为数据进行数据挖掘与学习分析，我们参考2012年美国教育部发布的《通过教育数据挖掘与学习分析促进教与学》(ET L–EDM LA)的报告[21]，结合学术界部分专家观点，针对性地提出“四建模三分析”的智慧课堂大数据研究方法论。

具体来说，“四建模三分析”主要是基于教育数据挖掘与学习分析技术在教学领域的应用提出的。建模与分析分别是围绕着学习者的内部特征与外部影响两方面展开的。在研究学习者内在行为、经历的基础上针对学习者进行画像，并在此基础上辅以知识领域的建模，全面刻画学习者内在学习特征。而在其外部，通过对学习组件以及环境进行分析，从而得出针对未来的趋势预判分析。“四建模三分析”的基本含义及着力解决的问题如下：

1.行为建模：通过对学生主动发生的学习行为进行学生建模，着重了解学生主观学习行为与学习结果关系、学生学习行为模式共性与差异、师生与生生互动联络拓扑。

2.经历建模：通过学生与教师的互动情况，着重对学生的学习感受进行建模，用以了解学生学习的主观评价以及对授课教师的侧面评价佐证。

3.画像建模：通过对包含互动对象、学习行为、学业结果数据在内的全方位的数据进行建模，对学生进行聚类分组，充分刻画学生的用户画像，以及发掘联络人网络中的有影响力节点。

4.领域建模：通过学生学习路径及其关联的知识点数据，自动对知识图谱进行建模，构建学科领域的知识图谱。

5.组件分析：通过对学习过程中的各种客观行为(组件)进行分析，获得其与学生学业结果的联系。

6.策略分析：通过对教学者教学风格等教学策略进行抽象与归纳分析，获得其对学生学业结果的影响。

7.趋势分析：借助学业结果影响因素的主观、客观、策略等因素的分析，对学业进行趋势预测；与此同时，借助网络文本数据分析对校园舆情进行管窥。

(二)“四建模三分析”的范围界定

上述七种建模与分析角度是目前教育大数据挖掘分析所公认、且成果较为密集的研究领域，并不涵盖课堂教育大数据分析的所有方向。通过“四建模三分析”方法，可以对智慧课堂的研究目标具象化，有助于在研究过程中的聚焦。本研究对智慧课堂数据建模分析研究范围界定如下：

1.对于研究目标不清晰的领域不予涉及。在学生行为建模中，业界提出对学习行为范式进行研究，这依赖于教育学、行为学等交叉学科的理论，有待于与这些领域专家的深入研究和合作，在此基础上进一步使用大数据作佐证，为共同深入该方向研究提供空间，因此本研究暂不涉及。

2.对于对象数据为传统问卷采集方式的研究未涉及。学生经历的建模在传统的教学数据采集手段中使用问卷形式进行，虽然这也是一种有效度、信度的测量方式，但由于在讯飞智慧课堂产品中未有问卷数据的体现，而且我们认为教育大数据分析能够常态化应用的前提是数据采集的常态化，因此对问卷数据采集方式在研究中不予包括。

3.对于个性化学习与自适应学习分析另有研究。个性化学习与自适应学习毋庸置疑是最为前瞻、最具特色、最有价值的课堂教学数据挖掘分析研究方向[22]。鉴于其研究重要性、方法的特殊性、以及内容丰富及相对独立等方面的考量，对个性化学习挖掘分析和自适应学习研究将作为单独领域另行开展研究。

(三)智慧课堂数据挖掘主要算法

“四建模三分析”的落地离不开数据挖掘算法与统计分析技术的应用。基于以上建模和分析的需要，根据我们的研究，智慧课堂数据挖掘使用的常用算法与技术主要包括多元回归分析、分类聚类算法、关联规则挖掘、文本分析挖掘、图构建与挖掘等方面。主要算法与技术如表2所示。

表2 智慧课堂数据挖掘主要算法与技术

挖掘方法	内涵	常用算法和技术举例
多元回归分析	通过对若干因素(自变量)与某一个因素(因变量)的影响进行量化分析，给出自变量可在一定误差约束内与因变量的关系	线性回归、逻辑斯蒂回归、逐步回归、岭回归等
分类聚类算法	在机器学习领域分属于有监督的学习和非监督的学习。其共同特征为对样本进行分组，同组的样本具有相似的特征	分类算法有决策树、朴素贝叶斯、支持向量机、神经网络，聚类算法有K-Means、DBScan、层次聚类等
关联规则挖掘	通过对行为或样本集合进行挖掘，得出行为A与行为B的蕴含式，并给出其支持度与置信度	Apriori、FP-growth等
文本分析挖掘	通过对半结构化与非结构化的文本数据进行挖掘，得出文本情感倾向与主题等	LDA主题模型、其他NLP技术等
图构建与挖掘	通过对清洗后的数据构建图数据结构，并使用图论方法进行研究与挖掘。常用研究任务有影响力节点发现等	Hits、PageRank、PersonalRank等

续表2

五、智慧课堂数据挖掘分析实施方法

(一)构建智慧课堂数据挖掘分析“整体棋盘”

基于“四建模三分析”总体框架以及五大类数据挖掘技术的概述，结合智慧课堂用户模型和数据体系，我们采用棋盘法将研究问题进行具象与细化，形成智慧课堂数据挖掘分析的“整体棋盘”。棋盘的首行列出七大研究方向，首列给出五大数据挖掘技术，在棋盘矩阵中纵横交错的每一个棋盘格子处则是使用某种数据挖掘方法对该类研究方向的具体细化。智慧课堂数据挖掘分析的整体棋盘如表3所示。

表3 智慧课堂数据挖掘分析“整体棋盘”

方向 / 方法	行为建模	经历建模	画像建模	领域建模	组件分析	策略分析	趋势分析
回归预测	学生主观行为对学业的影响研究				客观行为因素对学业的影响研究	师生互动指数分析	学生学业成绩预测
分类聚类		学生智慧课堂忠诚度分析	学生学习成绩分档模型研究				
关联规则	学生群体的行为序列差异研究						
文本挖掘				泛在信息中的知识点提取			校园情感分析、预警
图挖掘	班、校社交网络的构建	教师教学路径可视化	班、校社交网络影响力人物挖掘	自动构建领域知识图谱			

(二)设计研究问题及研究方法

利用教育数据挖掘和分析技术对每一项棋盘格中的研究问题进行建模分析，关键是要对具体的研究问题进行定义，设计基于行为数据的研究对象、方法和策略。根据“整体棋盘”框架，对13个具体研究问题定义如下。

(1)学生主观行为对学业的影响研究：学生在学习过程中的主动参与状态是影响学生学业结果的首要因素。通过对学生主观行为进行梳理并研究其对学业结果影响，有助于找出学业成绩的学生个体

主观行为中的主要成分；(2)客观行为因素对学业的影响研究：学生在学习过程中有不受自身控制的客观行为会影响其学业成绩。研究外界客观行为对学业的影响，有助于删繁就简地找出影响学业中的外界有利因素并加以因势利导；(3)师生互动指数分析：学生学习受教师与同伴的共同影响，研究教师教学策略以及学习伙伴因素对学业的影响，可以进一步印证教育中的有关成熟理论；(4)学生学业成绩预测：基于上述研究中的主、客观行为以及策略等因素对学业影响关系，利用一定时间段内用户综合行为数据对学业成绩进行预测，可预判学习走势，提前干预学习行为；(5)学生智慧课堂忠诚度分析：通过对学生在智慧课堂沉淀数据进行分析，获得用户对该信息化产品的粘性程度指数；(6)学生学习成绩分档模型研究：通过对学生的分学科学业成绩进行聚类分析，得出学生成绩的分档结果，与传统的统计学分位点分档模型进行相互印证，了解学生偏科状况；(7)学生群体的行为序列差异研究：不同学生的学习行为不同，其结果会反映在学业结果上。通过不同群体间行为序列差异的研究，在学生中推广学业优秀学生的学习行为序列，促进有效学习；(8)教师教学路径可视化：通过可视化技术直观显示教师教学行为路径，便于教育管理者进行教学研讨与比较反思；(9)泛在信息中的知识点提取：从半结构化与非结构化的信息中利用文本分析方法自动提取知识点，用于后续知识图谱构建；(10)校园情感分析、预警：利用情感分析技术分析校园内学生发布非结构化文本信息中的正负情感倾向，对个别有负面情绪学生进行预警预报，管窥校园舆情；(11)班、校社交网络的构建：利用师生、生生互动的数据构建班级与校园维度的社交网络，用于校园社交网络挖掘；(12)班、校社交网络影响力人物挖掘：从班级、校园社交网络中发现有影响力的学生，在教学中可利用其影响力，进行教学策略扩散的最大化；(13)自动构建领域知识图谱：学生知识习得的程序遵从一定的顺序，通过行为来自动还原知识图谱网络，并与人工构建知识图谱进行比照、验证。

六、智慧课堂数据挖掘分析应用模式

通过对5大项13个小项研究问题的具体细化，一幅针对智慧课堂数据进行应用研究的全图清晰地呈现在面前。在实践应用中，需要结合具体的专业领域应用需求和应用场景，将研究内容有机地组织起来，形成具体的应用模式。根据智慧课堂全过程、全方位的数据体系及应用需求，这里从课堂互动、学习行为、学习结果、校园社交等重点领域的分析应用入手，构建智慧课堂数据挖掘分析应用的四种基本模式。

(一)课堂互动分析应用模式

课堂互动是智慧课堂的核心特征。学生与教师互动、与资源互动、与平台互动等多向互动，很大程度上体现了学生投入学习的程度，反映了学生主动学习、积极学习的情况。基于学生和教师在智课平台的行为数据建立学生与教师互动、与平台互动的指标体系，同时依据因子分析法计算出互动指标体系的权重，进而建立教师与学生的互动指数、学生对智课平台的粘性程度指数，为设计和改进课堂教学互动提供依据。

(二)学习行为分析应用模式

学习行为数据是反映智慧课堂教学过程的最重要数据。通过从学生主观行为、客观行为、教学策略与学习环境等方面进行可能因素的梳理，利用统计学中相关性分析、显著性检验、因子分析等手段，探寻出影响学业成绩的主要指标。在此基础上通过对学生不同群体的学习行为序列利用关联规则挖掘技术与可视化展现方式进行差异研究，进一步寻找学生个体的学习行为差异，为探究学生学习过程影响因素提供重要手段。

(三)学习结果分析应用模式

学习结果数据是智慧课堂教学成效的基本体现。通过连续多次考试排名建立对学生成绩上升/下降、学习成绩分档模型。对学生考试成绩偏科情况探索，从整体角度分析偏科人数以及偏优和偏弱学科，从个人角度分析学生偏科行为。通过对学生历史考试成绩排名数据以及近期在作业平台上的行为数据进行未来成绩趋势预测。通过学生在智课平台的学习行为来自动还原知识图谱网络，并与人工构建知识图谱进行对比分析，描述学生的知识结构情况。

(四)校园社交分析应用模式

学生校园社交数据是反映学生全面成长、进行校园舆情管窥的重要依据。利用学生与教师、学生与学生互动的数据，分别构建校园维度和班级维度的社交网络。基于建立的班级、校园社交网络从中发现有影响力的教师和学生，在教学过程中，可利用其影响力，进行有效教学干预的最大化。从校园舆情角度来看，通过学习者在智课平台上私信、帖子等所涉及的文本内容，利用自然语言中基于情感词典的文本情感分析法，掌握学习者的情感倾向状态，以便于进行校园舆情的管窥。

七、应用实例：以学生主观行为对成绩的影响分析为例

(一)研究数据来源分析

本研究所使用的数据来源于智慧课堂产品在安徽省某重点中学2014级学生群体中使用的真实数据，涉及35个教学班学生共计1973名，教师98人。由于该年级使用智慧课堂产品两年有余，积累了大量的过程行为数据与学业结果数据，为下面的数据分析提供了大数据的支撑。出于隐私安全考虑，在数据分析时，采用学生匿名编码的形式以保护学生隐私。

在数据采集周期内，共选取了4次全学科考试。这四次考试分别发生在2016年1月20日、4月28日、5月30日和6月12日。经统计，四次考试全部参加的理科学生为1331人，文科学生为496人。本研究分别对理科和文科学生进行了分析，由于篇幅限制，本文中只例举理科学生的主观行为对成绩的影响分析。

(二)行为分析基本框架

分析学生主观行为对成绩的影响主要分为数据收集与处理、模型建立和结果分析三大部分，分析框架如图2所示。在数据收集与处理的过程中，本文选取学生行为指标数据和学生历史成绩数据，在收集数据之后需要对数据进行预处理。基于以上数据，本文使用相关性分析、多元回归分析和因子分析三种方法建立模型分析行为指标对成绩的影响。其中相关性分析和多元回归分析结合了学生的行为指标和历史成绩数据来分析指标之间的相关性并量化行为指标对学生成绩的影响；使用因子分析对多个行为指标进行降维处理，提取出影响学习成绩的因子。在建立模型之后，综合分析不同方法得出的结果，最终得出对学生成绩有显著影响的因素。

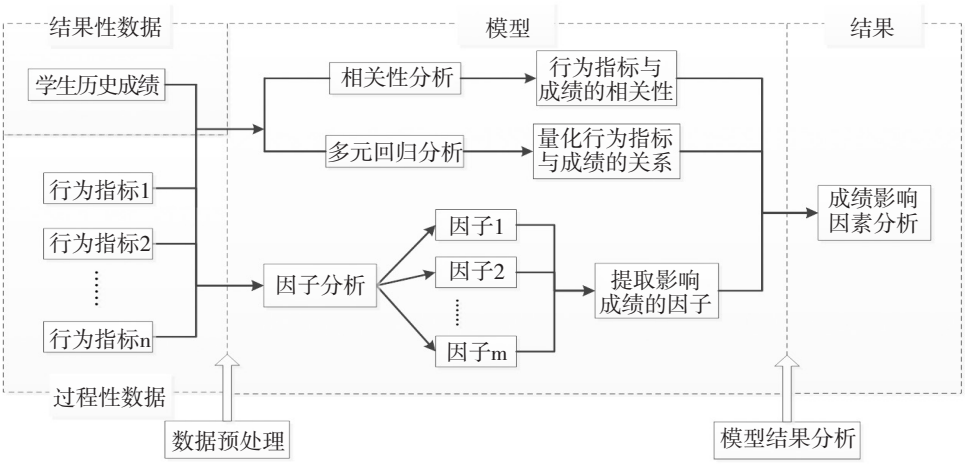

图2 成绩影响因素分析模型详细流程框架

(三)分析过程与结果

1.指标的筛选

初始提取出理科学生主观行为对成绩影响因素指标20个，包括根回复微课次数、非根回复微课次数、递交作业数、错题总数、评级微课次数、评级非微课次数、微课观看个数、微课观看次数、微课收藏次数、通知回复数、微课点赞次数、非微课点赞次数、社区发帖数、回复贴数、帖子中交互人数、访问别人次数、访问别人人数、与学生互批作业数、作业修订数、发给老师私信数。根据实际统计数据对这些指标进预处理，剔除统计量较小的数据，最终得到12个指标，用于下面的建模分析。

2.相关性分析

使用学生在2016年6月考试成绩作为成绩变量与各个行为指标进行相关性分析，从各个指标与成绩之间的相关系数。从相关性分析可以看出：对于理科学生来说，递交作业数、发给老师的私信数以及通知回复数这三个指标与成绩之间的相关性较高。

3.因子分析

利用因子分析法对多个行为指标进行降维处理[23]，即用少量的综合指标来替代多个可观测变量，便于把握主要影响因素。主要包括以下步骤：

首先，要判断数据是否适合做因子分析，采用对数据进行KMO值和Bartlett球形度检验[24]。基于实际数据计算，理科学生的KMO统计量的值分别为0.697，根据评判标准可知，KMO统计量的值大于0.6，适合做因子分析；Bartlett球形度检验的卡方的P值小于0.01的显著性水平，同样显示适合做因子分析。

其次，选取基于主成分分析的提取方式对原始变量进行因子提取。运用社会学统计软件SPSS进行因子分析。通过对原始变量采取主成分分析，依据Kaiser标准(特征根大于1)来提取因子[25]，可以得出特征根大于1的因子有4个。在此基础上使用正交旋转的方式计算4个因子的方差贡献率如下页表4所示。

从下页表4可以看出，这4个主因子的方差贡献率分别为：23.297%，16.873%，14.859%，8.974%。累计方差贡献率为64.003%，通过这4个主因子可以反映原有12个

初始变量的64.003%信息量，因而可作为原有12个变量的主因子。其中因子1的解释率为23.297%，显著高于其他因子，是学生主观行为对成绩影响的首要因子。

表4 基于主成分分析的因子提取(理科)

因子	初始特征值			旋转平方和载入		
	合计	方差贡献率%	累计贡献率%	合计	方差贡献率%	累计贡献率%
1	3.465	28.873	28.873	2.796	23.297	23.297
2	1.893	15.776	44.649	2.025	16.873	40.17
3	1.296	10.798	55.447	1.783	14.859	55.029
4	1.027	8.556	64.003	1.077	8.974	64.003

第三，利用因子载荷阵进行旋转，对4个因子所反映的变量信息进行解释。由于初始提取的公共因子的涵义并不是很清晰，需要对初始的因子载荷阵进行旋转，这样可以使因子载荷矩阵结构更加简化，以方便对每个因子进行解释。按照方差最大正交旋转对初始因子载荷矩阵进行旋转，旋转之后因子载荷矩阵如表5所示。

表5 因子旋转矩阵(理科)

	因子1	因子2	因子3	因子4
根回复微课次数	0.016	0.175	0.736	–0.041
非根回复微课次数	0.11	0.008	0.691	–0.209
递交作业数	0.856	0.229	0.139	0.027
错题总数	0.551	–0.011	0.037	0.04
微课观看个数	0.065	0.98	0.123	0.001
微课观看次数	0.099	0.973	0.156	0.031
微课收藏次数	0.064	0.04	0.551	0.176
微课点赞次数	0.256	0.174	0.579	0.404
非微课点赞次数	–0.054	–0.006	0.011	0.912
与学生互批作业数	0.61	0.033	–0.059	–0.004
发给老师私信数	0.78	0.023	0.172	–0.037
通知回复数	0.824	0.025	0.181	–0.042

从表5可以看出，每个主因子都包含着各自的主要变量(表中每个因子灰度显示的变量)，根据各自变量的含义，对4个主因子进行维度说明。因子1主要反映了学生作业的递交、修改和错题情况，以及对老师通知的回复和私信的发送，故将因子1命名为“自主学习因子”。同理，将因子2命名为“微课观看因子”；因子3命名为“微课交流因子”；因子4命名为“非微课点赞因子”。

从因子分析的结果来看，影响理科生成绩主要的主观因子是自主学习因子、微课观看因子、微课交流因子、非微课点赞因子；其中权重最大的自主学习因子主要包含递交作业数，通知回复数，发给老师私信数，与学生互批作业数，错题总数等；自主学习因子与理科生成绩有着直接关系。

4.多元回归分析

为了拟合成绩与主观行为指标之间的具体关系并确定对成绩的显著性影响指标，这里使用理科学生2016年6月考试成绩作为因变量，并以提取的主观行为指标作为自变量进行多元回归分析。为防止变量间存在共线性现象，应利用方差膨胀因子VIF进行判断。按照统计经验：当0<VIF<10时，不存在多重共线性；当10≤VIF<100时，存在较强的多重共线性；当VIF≥100时，存在严重多重共线性[26]。

通过首轮计算可以看出，微课观看个数和微课观看次数之间存在较强的多重共线性。删除微课观看个数，再进行多元回归分析输出结果如表6所示。

表6 多元回归分析表(理科)

模型	非标准化系数		标准系数	t值	P值	共线性统计量	
	系数	标准误差				容忍度	VIF
(常量)	350.41	0.957		366.32	0.000		
根回复微课次数	–3.375	1.708	–0.060	–1.976	0.048	0.779	1.284
非根回复微课次数	1.707	5.943	0.008	0.287	0.774	0.847	1.181
递交作业数	0.326	0.070	0.223	4.651	0.000	0.312	3.202
错题总数	0.788	2.067	0.011	0.381	0.703	0.851	1.175
微课观看次数	0.015	0.132	0.003	0.112	0.911	0.811	1.233
微课收藏次数	–1.174	1.353	–0.025	–0.868	0.385	0.880	1.137
微课点赞次数	0.465	0.344	0.043	1.350	0.177	0.700	1.429
非微课点赞次数	–4.621	2.789	–0.045	–1.657	0.098	0.959	1.042
与学生互批作业数	–3.658	1.337	–0.087	–2.735	0.006	0.702	1.424
发给老师私信数	1.228	0.883	0.052	1.392	0.164	0.507	1.970
通知回复数	–0.224	0.567	–0.017	–0.396	0.692	0.382	2.621

从输出的VIF值可以看出，删除微课观看个数之后，指标之间不存在多重共线性。

(四)结论

对于理科来说，从回归分析的输出结果可以看出，在0.05的显著性水平下，根回复微课次数、递交作业数、与学生互批作业数这3个变量对成绩的影响是显著的。根回复微课次数、与学生互批作业数的系数分别为–3.375、–3.658，对成绩产生负的影响；递交作业数的系数为0.326，对成绩产生正的影响。影响理科学生成绩的主要指标是根回复微课次数、递交作业数、与学生互批作业数。

参考文献：

[1] 杨现民,唐斯斯,李冀红.发展教育大数据:内涵、价值和挑战[J].现代

远程教育研究,2016,(1):50–61.
[2] 金陵.大数据与信息化教学变革[J].中国电化教育,2013,(10):8–13.
[3][21] Bienkowski M, Feng M, Means B. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief[J]. US Department of Education, Office of Educational Technology, 2012,(1):1–57.
[4] Siemens,G.Learning and Knowledge Analytics –Knewton –the future of education?[EB/OL]. http://www.learninganalytics.net/?p=126,2011–04–14.
[5] 娄晓敏.我国教育大数据建设的现状分析与架构初设[J].中国电化教育,2017,(8):128–131.
[6] 葛道凯,张少刚,魏顺平.教育数据挖掘方法与应用[M].北京:教育科学出版社,2012.12.
[7] 杨永林,张世蓉等.从“慕课”到“小微课”,看大数据在教学中的应用[J].现代教育技术,2014, 24(12):45–51.
[8] 李曼丽,黄振中.MOOCs平台大数据的教育实证[J].科学通报,2015,(5):570–580.
[9] 周丹,陈丽婷.大数据时代背景下基于翻转课堂的教学改革研究[J].教育与职业,2015,(33):90–92.
[10] 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.
[11] 贺超凯,吴蒙.edX平台教育大数据的学习行为分析与预测[J].中国远程教育,2016,(6):54–59.
[12] 刘军.智慧课堂:“互联网+”时代未来学校课堂发展新路向[J].中国电化教育,2017,(7):14–19.
[13] 刘邦奇.智慧课堂:“互联网+”时代的课堂变革[N].江苏教育报,2016–09–21(04).
[14] 王盛之,毛沛勇.基于数字化教学案的智慧课堂互动教学系统实践研究[J].教学月刊(教学管理)中学版,2014,(4):51–55.
[15] 林利尧.中小学智慧课堂建设与应用研究[J].中国现代教育装备,2013,(10):38–39.
[16] 邓光强.“智慧课堂”中的学生个性化学习[J].教育信息技术,2013,(12):11–13.
[17] 孙曙辉,刘邦奇.基于动态学习数据分析的智慧课堂模式[J].中国教育信息化,2015,(22):21–24.
[18] 刘邦奇.“互联网+”时代智慧课堂教学设计与实施策略研究[J].中国电化教育,2016,(10):51–56.
[19] 讯飞皆成[EB/OL].http://www.jevictek.com/www/home/category/llyj,2017–11–11.
[20] 孙曙辉,刘邦奇.智慧课堂[M]. 北京:北京师范大学出版社,2016.
[22] 马相春,钟绍春,徐妲.大数据视角下个性化自适应学习系统支撑模型及实现机制研究[J].中国电化教育,2017,(4):97–102.
[23] 刘亚瑜.中小学教师教学专业化评价指标体系的构建[D].西安:陕西师范大学,2008.
[24] 辛督强,韩国秀.因子分析法在科技期刊综合评价中的应用[J].数理统计与管理,2014,33(1):116–121.
[25] 任雪松,于秀林.多元统计分析.第2版[M].北京:中国统计出版社,2011.
[26] 田水承,薛明月等.基于因子分析法的矿工不安全行为影响因素权重确定[J].矿业安全与环保,2013,(5):113–116.

作者简介:

孙曙辉:执行院长,研究方向为教育信息化、智慧课堂(shsun@iflytek.com)。

刘邦奇:硕士,教授,研究方向为教育技术应用、智慧教育、智慧课堂(lbq–nj@163.com)。

李鑫:博士,研究方向为大数据分析、教育信息化(xli@iflytek.com)。

The System Framework of Data Mining and Learning Analysis for Smart Class

Sun Shuhui[1], Liu Bangqi[1,2], Li Xin[3]

(1. iFlytek Institute of Educational Informatization, Hefei Anhui 230088; 2. School of Education, Capital Normal University, Beijing 100048; 3. iFlytek Big Data Research Institute, Hefei Anhui 230088)

Abstract: Researchers on educational data at home and abroad are moving from the concept, principle level to modeling analysis and application level. The researches in application level are moving from statistics of education quality monitoring, education policy and other macro-level applications to the further development of school teaching, student learning and other micro applications. On the basis of discussion of data mining and analysis of the connotation of education, this paper puts forward data model and system architecture combined with the wisdom teaching practice, elaborates “Four modeling and three analysis” methodology of study of education data, constructs “The whole chessboard” of intelligent data mining analysis of smart class and 15 specific research questions, and puts forward four kinds of application modes of intelligent data mining analysis of smart class. Finally, based on the real data, the influence of students’ subjective behavior on achievement is discussed.

Keywords: Smart Class; Big Data of Education; Data Mining; Learning Analysis; Teacher-Student Interaction Index

收稿日期:2017年11月11日

责任编辑:宋灵青