基于大数据的高职学生行为分析

程光胜

（宁夏财经职业技术学院 宁夏 银川 750021）

摘要：为改变依靠说教和事后分析找对策的传统学生管理方式，基于大数据的视角，构建了校园学生行为数据的分析模型，并通过大数据处理框架 Hadoop 和数据分析语言 R 设计了学生行为分析平台架构，基于此架构，实现了学生行为特征的分析，为学生管理的科学化、全面化和个性化提供了技术服务，为学校的高层决策提供智力支持。

关键词：互联网 + 教育；教育大数据；学生行为分析；Hadoop 框架；R 语言

中图分类号：G718 **文献标识码**：A **文章编号**：1672-5727(2020)08-0076-05

2015 年 8 月，国务院发布了《促进大数据发展行动纲要》，指出加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。2019 年 10 月，教育部办公厅发布的《关于推荐遴选“基于教学改革、融合信息技术的新型教与学模式” 实验区的通知》明确要求，通过云计算、区块链技术等采集学习过程中的生成性行为数据，开展教学分析与过程性评价，提升课堂教学和育人的有效性，从而促进学生个性化全面发展的成长路径。可以看到，大数据已经成为国家实施创新驱动发展战略的内在需要和必然选择，在教育领域也已成为促进学生全面发展和个性化发展的重要技术支撑[1]。

随着“互联网+教育”的深入推进，教育信息化得到快速发展，在很大程度上支撑和引领着教育现代化的发展。伴随着新一代信息技术的催生和应用，在有力推动教育理念更新、模式变革和体系重构的同时，也产生了记录学生学习和生活的海量数据。通过这些大数据，一方面，能够发现学生在校期间的成长变化规律，捕获学生不同行为背后的特征及相关性；另一方面，可以实现更加科学化、全面化和个性化的校园管理和服务，进而引导学校形成健康科学的学生培养模式和教学生活管理方式，同时也为学校发展决策提供科学依据。

一、大数据与学生行为分析

麦肯锡认为，大数据是一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合。因此，大数据代表了存储和处理海量数据的一种技术，所以，大数据既是一种资源或工具，更是一场革命。这场革命需要大数据技术才能使得大数据发挥其重要的价值，以此来提供更强的决策力、洞察力和发现力。

在教育领域，随着近年来教育信息化的大力发展，产生了大量的教育数据，即教育大数据。教育大数据记录了教育教学活动中所产生的各种数据，以及学生在校期间的各种行为数据，这些数据在某一个时间点上是静态离散的，反映了学生的学习状态、成果，但在特定的时段内，这些数据是动态连续的，反映了学习及生活活动的行为轨迹。因此，基于教育大数据的

作者简介：程光胜（1981—），男，硕士，宁夏财经职业技术学院讲师，研究方向为职业教育、数据科学、软件工程。

基金项目：2019 年度宁夏财经职业技术学院重点立项课题“‘互联网+教育’下基于大数据的高职学生行为分析研究”（项目编号：NCYHLW201929）

分析和挖掘，对于学生的全面发展和个性化培养，具有非常重要的意义和价值。

行为分析源自心理学，由美国心理学家亨特提出，他指出应该尽量避免应用带有心灵色彩的术语，力求通过外在行为来描述、解释、预测和控制有机体。在校园里，传统的学生行为管理主要通过学校的政策文件和学生行为手册来规范和约束，这种管理方式通常是固定的、模式化的、路径依赖式的，与科学化的学生管理相比，还有很大的差距。基于教育信息化下的教育大数据，真实、准确地记录了学生的学习和生活行为，通过对这些数据的挖掘和分析，可以充分了解学生的“个性”，进而制定和实施个性化的管理方案，实现对学生个性化的人文关怀，为学生提供个性化的过程预警。

学生在校期间所产生的行为数据，既有课堂学习、参加考试的结构化数据，也有上网购物、浏览页面的非结构化数据，以及通过校园一卡通进入学校餐厅、超市、图书馆等场所消费及学习的其他数据，这些数据记录了学生个体层面在校的行为轨迹[2]。为了方便后续的分析，这里对行为数据进行进一步的梳理，构建了行为数据分析模型，如图 1 所示。学生行为数据通过学生基本数据、课堂学习、课外学习、学生成绩、校园生活、校园活动六个方面来获取，这六个方面形成了学生行为数据分析的六个维度，通过这六个维度可以全方位刻画学生的学习和生活概况，也可以分析不同维度下的指标关系，还可以进行深层次的数据挖掘。

二、学生行为分析平台的架构

(一)数据处理框架和分析语言的选择

目前，对于大数据的处理框架主要包括 Hadoop、Spark 和 Storm，这些框架具有特定的优势和不同的应用场景：Hadoop 通过 HDFS 实现海量数据的存储，通过 MapReduce 进行分布式计算，非常适合处理批量离线数据，同时，针对不同的需求有不同的支持工具，所以，Hadoop 目前是一个庞大的 Hadoop 生态系统；Spark 是专门为大规模数据处理而设计的快速通用的计算引擎，基于内存计算，支持分布式数据集上的迭代作业，支持交互式计算和复杂算法，现已成为一个高速发展且应用广泛的生态系统；Storm 是一个分布式实时大数据处理系统，高性能、可扩展、高容错。综合比较来看，Hadoop 擅长批处理、吞吐量大、做全量数据的离线分析；Spark 适合构建大型的、低延迟的数据分析应用程序；Storm 在实时方面具有先天的优势，但单位时间内的吞吐量要小于 Hadoop。分析学生的行为数据，对实时性没有特定要求，故不考虑基于流式的 Storm 计算框架，而 Spark 本身也没有提供分布式文件系统，所以，本文最终选择 Hadoop 作为学生行为数据的处理框架。这是因为 Hadoop 提供了成熟的海量数据存储方案，还有大量可供选择的第三方存储工具，比如 HBase、Hive 等，以及基于 HDFS 读写数据吞吐量大，进行离线分析不会影响到正常的业务系统的运行。

对于数据分析编程语言，目前应用最多的是 Python 和 R。R 是一种自由、开源的语言和操作环境，在统计分析、绘图和统计编程上具有先天的优势，目前大量应用在统计分析、数据挖掘、机器学习、生物信息、金融分析等领域，拥有大量的支持包，能够调用 C、C++、Fortran、Java 等其他编程语言；Python 是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言，目前在数据分析、机器学习、矩阵计算、科学数据

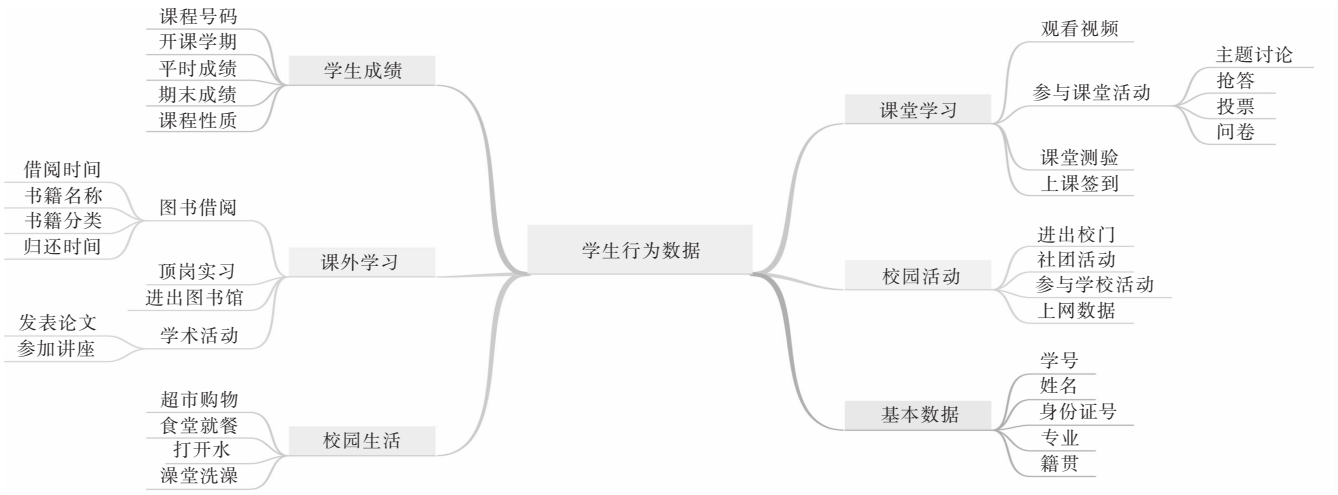

图 1 学生在校行为数据分析模型

可视化、数字图像处理、Web 应用、网络爬虫、系统运维等方面都具有特定的优势，拥有大量的第三方库，已形成了较为庞大的生态系统。由于本文的研究目前在实验分析阶段，所以选择 R 作为数据分析语言。

(二)平台架构设计

基于大数据处理框架 Hadoop 和数据分析编程语言 R，学生行为分析平台的架构设计如图 2 所示。

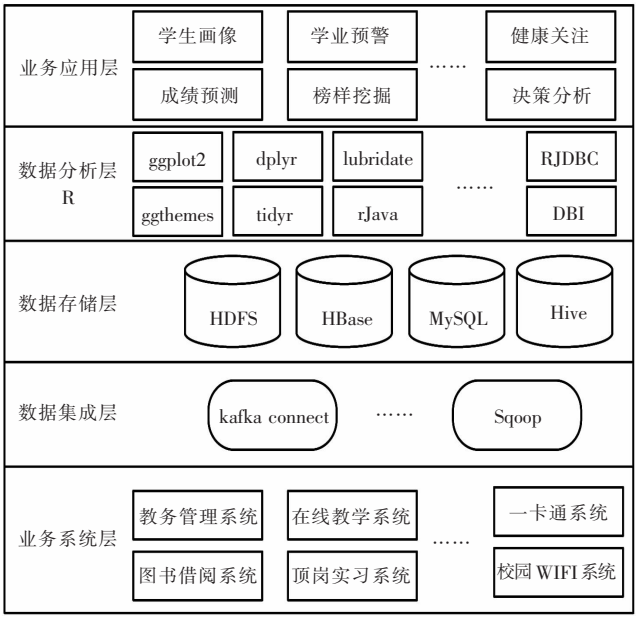

图 2 行为分析平台架构

从图 2 可以看出，整个架构设计分为五层，从下到上分别是：业务系统层、数据集成层、数据存储层、数据分析层和业务应用层。业务系统层是目前支持校园各项业务运作的系统，比如用于教务管理的教务管理系统、用于开展线上教学的在线教学系统、用于学生购物消费的一卡通系统、用于进出图书馆以及图书借阅的图书借阅系统、用户跟踪学生定岗实习的定岗实习系统等。系统集成层可以通过 kafka connect、Sqoop 等将业务系统中产生的数据统一集成到数据存储层中。数据存储层通过 Hadoop 提供的 HDFS 实现海量的数据管理，同时基于 HDFS 可以支持 HBase 以及 Hive 等具体的数据应用环境。数据分析层根据业务需求实现对数据的各种分析，包括基本的统计分析、关联分析、相关及回归分析、分类及聚类分析、时间序列分析等。业务应用层实现对学生科学化、精细化及个性化的管理，比如学生画像、心理及身体健康关注、成绩预测、学生管理等。除此之外，还有保证整个系统稳定健康运行的安全机制，包括安全管理、运维监控、调度管理、质量管理等。

三、学生行为特征分析

(一)数据预处理

在进行具体分析前，需要对不同业务领域获取到的数据进行有效清洗和预处理，具体包括：剔除无效数据、数据压缩(归约处理)、生成新的度量指标、数据拆分、数据变换(比如取对数)、归一化处理等。例如，针对学生一卡通的消费数据，由于学生可以通过一卡通进行就餐消费、学校超市购物消费等，所以数据量非常庞大，在单独进行一卡通的消费数据分析时，可以将这些数据按时间进行拆分处理，这样可以对不同数据段的消费单独处理，然后再对处理结果进行合并[3]。

为了对学生的行为特征进行综合全面的分析，需要将不同数据源(业务系统)获取到的数据进行连接操作，形成一个大的数据表，连接的字段是学生的学号。通过这种连接操作，一方面可以过滤掉对分析目标没有贡献的无效数据，另一方面也可以适当降低总数据量的大小，最后将连接形成的数据集(表)存放到 Hive 数据仓库中。

Hive 是一种底层封装了 Hadoop 的数据仓库处理工具，可以将结构化的数据文件映射为一张数据库表，并提供完整的 SQL 查询功能，也就是说通过 Hive 可以在 HDFS 上使用 SQL 语句执行 MapReduce 的计算任务。而 R 可以通过 JDBC 来连接 Hive，具体操作需要在 R 中安装 rJava、DBI、RJDBC 等包，同时还需要将集群中 Hive JDBC 的驱动包拷贝出来放在 R 程序所在的本地。R 连接 Hive 成功后，就可以对 Hive 中的数据进行各种分析和图表展示。

(二)学生画像

为了对学生的总体信息和行为特征进行全方位的展示，或对特定时间段内的行为进行具体描述，就需要通过学生行为特征数据对其画像。有效地构建学生画像，有利于精准剖析学生成长过程和特点，有助于提供个性化的培养和指导。由于学生在校的行为是多元化的，所以学生画像也是多元化的。学生画像，一方面是对学生信息和行为的汇总输出；另一方面也是对学生行为特征数据背后的深层次挖掘。

在实现学生画像前，需要构建有效的画像模型，即画像设计。可以在学生行为数据分析模型的基础上，将画像抽取为五个维度，即：总体信息概览、课程学习成绩、就餐消费展示、图书馆进出及借阅统计、校

园影像。每一维度下，又有具体的输出设计，比如，总体信息概览有学习、就餐消费、购物消费等的频次描述，针对这些频次数据可以给学生打上不同的标签，比如是否学霸、消费水平高低等；课程学习成绩分年级（一年级、二年级、三年级）和课程类型（公共基础、专业基础、专业核心、专业选修）等，以饼图、直方图、折线图、雷达图对比展示，以说明在不同时间节点上成绩的波动情况；校园影像根据学生在校期间参加社团、参加各种比赛、超市消费、进出澡堂等数据，展示学生的兴趣特长、购物偏好、卫生健康等。在此基础上，对学生行为数据进行结构化和标签化处理，提取出精准的学生特征标识，然后通过文字描述、图表展示等不同的载体方式呈现出来，这里主要使用到了 R 中的描述统计分析方法，对用户的行为数据进行高度概括，同时图表的输出使用到了 R 中的 ggplot2 包。

（三）日常行为与学业表现相关性分析

在校园里，不同学生有不同的行为方式，优秀学生和成绩不理想学生在日常的行为中同样存在不同的差异，对这些差异可以根据学生的行为数据来进行分析。比如，线上教学环境下，学生是否关注课程通知、学习视频观看长度、参与课堂互动等行为是否与课程的学习成绩有关；线下学生出入图书馆的次数、早中晚餐的就餐规律等是否与学业成绩有关。由于学生的行为数据在总体上可分为离散型变量和连续性变量，所以在度量不同类型的行为变量与学业关系时，其计算指标也不相同。比如，在 R 中，对于连续性行为数据，可以通过 cor()函数中的 pearson 系数和 spearman 系数来分析相关性，而对于离散型行为变量，可以通过箱线图来刻画和描述，因为箱线图能够从平均水平（中位数）、波动程度（箱子高低）以及异常值等方面来对分类变量进行直观比较。

通过分析发现，在连续性变量中，学生课堂学习得分、进出图书馆的次数与学习成绩之间存在明显的正相关性，相关系数高达 0.75；而与购物次数、在线学习时长相关性不大。对于此结果，可以这样理解，学业成绩优异的学生有自己学习和思考的方式，这些学生可能大部分时间花费在了图书馆看书和思考问题上面了。度量离散型变量和学业成绩之间的关系，可以通过箱线图来直观地进行展示，从结果来看，就餐时间的规律性、借阅书籍的类型与学业成绩有较大关系，而消费水平与学业成绩关系不大。

以上只是从变量的角度度量了学生行为相关指标与学业之间的相关关系。由于影响学生学业的因素可能有很多，这些单个因素的影响可能是微不足道的，但是这些因素联合起来就会对学业起到非常突出的影响。所以，为了更准确建立学业与日常行为之间的数量关系，本文选取众多的行为指标数据，采用随机森林算法构建学生行为与学业表现的预测模型。与其他机器学习模型相比，随机森林能够充分发挥决策树的分类优势，并有效避免了单个决策树容易产生过拟合的缺陷，同时对多元共线性不敏感，结果对缺失数据和非平衡的数据比较稳健。使用随机森林来构建学生行为与学业之间的关系，可以很好地发挥随机森林在预测多达几千个解释变量方面的作用优势。

在 R 中，随机森林的支持包为 randomForest，其核心函数为 importance()、MDSplot()、rfImpute()、treesize()和 randomForest()。在预测学业时，将学业输出分为优秀、良好、中等、合格和不合格五个等次，在建模过程中，需要多次调用 set.seed()设置随机数生成器初始值，而且每次设置不一样，这意味着每次随机抽样的结果会不相同。从结果输出可以看出，模型包含 500 棵决策树，总的预测误差为 33.76%，并且能够针对优秀、良好、中等、合格和不合格五个等次输出预测正确和错误的样本数以及预测误判率，同时可以通过 importance()函数查看到对模型影响重要的变量。具体有：就餐率（早餐、中餐）、图书馆进出次数、借阅图书数量、参加专业技能比赛等。事实上，从预测误差上来看，这个输出不是特别理想，需要进行优化，优化时，需要改变 randomForest()函数的默认参数，同时针对决策树分支节点变量个数，以及决策树数量，采用逐一增加变量的方法不断测试，最终构建的最优化模型，节点变量个数为 2，决策树数量为 450，此时总体误判率为 21.94%，与之前输出相比，预测准确率得到明显提升。

（四）情感分析

学生在校期间，受某些事件的影响，个人情感会产生波动，如果是负面的，在一个特定的时间段内如果得不到调节和改变，就会影响到学习以及身心健康。在目前“互联网+教育”环境下，通过学生的网络行为可以提取出与情感相关的信息，对其进行深入分析，可以检测到学生基于情感的异常行为，从而进行及时的干预和引导。

情感分析主要针对文本,核心是利用计算技术对文本的主客观性、情绪等进行挖掘和分析,从而对文本的情感倾向做出分类判断。目前,情感分析的方法主要包括情感词典法和机器学习法。对于情感词典法,主要是通过分词技术提取出待分析文本中的关键词,然后与情感词典中的词通过相似度计算判断情感倾向。所以,该方法的关键是分词技术和情感词典的构建和扩充,比如,“好难过啊”如果拆分为“好”和“难过”,就会失去句子的本意。对于机器学习法,需要标注文本语料(有人工标注和机器自动标注之分),然后运用 SVM、NBM 等机器学习分类算法进行情感分析,最后得到情感的分类结果。相比较而言,通过机器学习进行情感分析,具有一定的优势,这是因为机器学习算法在识别准确性上有很大的提升空间,比如参数优化、算法可选空间大等,同时在 R 中均有很好的支持。当然,在移动互联网环境下,随时随地都会诞生很多网络新词,比如,“我太难了”“蓝瘦香菇”,而这些网络新词最受大学生的喜爱,所以,在进行情感分析时,网络新词应该受到极大的关注[4]。

从学生的行为数据中,可以获取学生最近借阅的图书、使用 PC 或手机终端发布的学习和生活方面的主题信息,从而提取这些文本数据,作为学生情感分析的输入。接下来,通过 R 中的 tm 包、RTextTools 包、text2vec 包、glmnet 包等实现情感分析,选择知网情感词典 HowNet 获得情感信息特征,然后对学生文本数据进行分词、修剪,并消除低频词、增加文字信息量 n-gram、实现 TFIDF、构建 DTM、生成情感模型等,在此基础上完成情感的预测和识别。需要说明的是,text2vec 是一个 R 包,为文本分析和自然语言处理提供了一个简单高效的 API 框架,通过该包可以实现文本向量化、主题模型分析、Word2Vec 的“升级版 GloVe 词嵌入表达”、相似性度量等功能。

学生是教育的主体,对学生科学化、全面化和个性化的管理是教育管理者一直追求的目标。本文基于大数据视角,对学生在校期间的行为数据进行了梳理,获得了行为数据模型,在此基础上,比较分析了大数据处理框架和数据分析编程语言,构建了学生行为分析平台架构,然后使用 R 提供的函数和功能模块,对学生进行行为特征分析。通过对学生行为数据的深入挖掘和分析,一方面,可以为学生管理提供科学化指导;另一方面,可以对学生的学习行为起到预警,对学生的负面情感及时进行疏导,同时,也为学校的科学管理提供有价值的决策信息。

参考文献:

[1]邓逢光,张子石.基于大数据的学生校园行为分析预警管理平台建构研究[J].中国电化教育,2017(11):60-64.

[2]李有增,曾浩.基于学生行为分析模型的高校智慧校园教育大数据应用研究[J].中国电化教育,2018(7):33-38.

[3]胡茜茜.基于学生个人大数据的行为特征分析[D].武汉:华中师范大学,2019.

[4]刘丽,岳亚伟.面向高校学生微博的跨粒度情感分析[J].计算机应用研究,2019(6):1618-1622.

(责任编辑:张宇平)

Behavior Analysis of Higher Vocational Students Based on Big Data

CHENG Guang-sheng

(*Ningxia Vocational College of Finance and Economics, Yinchuan Ningxia 750021,China*)

Abstract:To change the traditional way of student management, which depends on preaching and post-analysis to find countermeasures, the analytic model of campus student behavior data is constructed based on the perspective of big data, and the platform architecture of student behavior analysis platform is designed through the Hadoop of big data processing framework and R of data analysis language. On this basis, students´ behavior characteristics are analyzed, providing technical services for the scientific, comprehensive and individualized management of students, and intellectual support for the high-level decision-making of the university.

Key words:Internet+ education; educational big data; student behavior analysis; Hadoop framework; R language