文章编号:1005-9679(2018)01-0119-07

校园大数据文献综述

吴胜男
(上海外国语大学 国际工商管理学院,上海　200083)

摘　要: 本文旨在对于高校基于一卡通的大数据分析尝试进行梳理和总结。首先,本文介绍了目前高校基于一卡通数据进行的大数据应用的主要类别,然后对于目前的部分应用及文献从目的、使用的模型和算法等方面进行总结、阐释和综述,最后,对这些文献和研究现状进行了评述。

关键词: 大数据;高校;一卡通;文献综述
中图分类号: G 353.11　　**文献标志码:** A

A Review on Campus Big Data Analysis

WU Shengnan
(School of Business and Management, Shanghai International Studies University, Shanghai 200083, China)

Abstract: This paper aims at summarizing and reviewing papers and researches about campus big data analysis. In this paper, firstly papers and researches about campus big data analysis is categorized and then analyzed and summarized. Finally comments of these literatures are put forward.

Key words: big data;campus;smart card;review

1　高校大数据分析的主要类型

高校与商业界和公共管理界具有天然的不同。商业界的构成主要为各类企业及从业人员,企业以长期利润为其终极目标。又由于在经营活动中产生的海量数据的潜在收益迫使业界无法忽视对潜在信息的挖掘,因此企业中的大数据分析尝试也建立在有助于企业的长存和利润的获取的基础之上,共同的目标有助于获得相关业务部门的高度配合,而务实的目的(经济回报)往往催生出实用性较高的成果。丰厚的研发资金和自身业务积累的大量数据又使得成果的转化成为可能,从而形成一种良好的生态循环。而公共管理部门的特征在于以政府为核心,集合各个社会部门的力量,运用政治经济及文化的一系列手段,提升政府的绩效和公共服务品质。其以政府为中心的特征决定着不必过于在乎前期庞大的开发费用,加上在社会公共服务和管理业务中,由于涉及方面的广阔和服务人数的巨大,必然产生大量种类各异的数据,为大数据分析的发展提供了良好的机遇。而高校一端联系着社会,一端联系着公共服务。作为社会公共服务机构的重要一环,高校人口中学生和教职工占据绝大多数,因此它的核心在于培育更加优秀的学生,为教职工团队在科研和个人发展方面提供支持和良好的管理,而不是经济利润的最大化,对于新技术的转化和花费比较大的科研项目,都需要政府或者企业的资助。从高校的业务来看,由于业务种类和涉及的结构都相对简单,产生的数据不可避免地体量更轻、结构更加单一,因此从前期投入来源和数据的结构来看,都不具有明显优势。而在高校的大数据分析尝试中,信息部门通常独担重任,虽然在高校信息化的过程中发挥着日益重要的作用,信息部门在日常管理和运作中并不占据核心地位,因此在数据完整性和数据治理上比商业界和公共管理界略逊一筹,难度更高。

高校的服务性质决定着其进行大数据分析的目的和侧重点也不一样,更强调是否能为学生和教职工更好地服务,人文关怀性和科研性更重,不过于强调经济效益。因此,目前高校的大数据分析尝试中,大多与学生的经济状况和行为模式挖掘有关,一则由于数据来源相对容易,二则由于与高校尝试大数

收稿日期:2017-12-05
基金项目:第二届"上海外国语大学青年教师创新团队":大数据环境下的多语种 Web 文本挖掘(QJTD14ZY001)。
作者简介:吴胜男(1994—),女,四川广安人,硕士研究生,研究方向:数据挖掘、管理信息系统和知识管理。E-mail: wushengnan0101@163.com.

据分析的初衷最为相符。而由于学生和教职工的活动范围主要集中在校内，校园一卡通的使用率很高，其中既包含消费行为所产生的经济数据，又包含日常活动，如图书馆打卡、借阅等产生的应用数据，且数据格式整齐、质量较高，为高校大数据分析提供了良好的数据基础，成为一种常用的数据来源。

从主题的角度分类，目前基于高校一卡通数据的大数据分析主要在以下三个方面：

对学生经济状况的挖掘。主要目的在于通过学生的一卡通消费数据以及学生的行为数据推测学生的经济情况和在校学习状况，找出符合特定经济标准的学生。一般应用于高校贫困生的认定额资助工作以及奖学金的评定。它的出现是为了避免在此类评定中评定人主观因素对于资助精确度与效果的影响，以及避免评定过程中的违规操作。

对学生学业状况的挖掘。主要目的在于通过学生以往的行为数据和学业成果相关数据，如绩点、科研成就、就业情况等，挖掘出二者的相关关系，从而预测其他学生的学业成就情况，并且对于有学业风险的学生进行干预。这有利于发现在教育中应该重点关注的学生中潜在的科研人才、市场精英以及需要帮助的学生，提高高校的教学质量。

对后勤服务及管理状况的挖掘。这一类的研究应用通常与“智慧校园”主题相关，也与“物联网”概念相似。此类研究通常关注高校后勤服务产生的数据，如寝室门禁数据、食堂就餐人数和开水消耗数量等，找出这些数据的规律，可提高后勤服务质量，减少浪费缩减成本。

2　数据来源

按照主题相关性和典型性，本文共选取 26 篇文献进行分析和梳理，其中期刊文章 18 篇，硕士毕业论文 4 篇，会议论文集收录的文章 4 篇。由于本文主要梳理大数据分析在中国高校的应用，因此基本不涉及国外的研究，文章来源为中文论文最权威的来源——中国知网，检索方法为按照三类主题在中国知网上分别进行高级模式下的跨库检索，跨库选择为期刊、国内会议、国际会议、硕士论文、博士论文和报纸。

按照三类主题，本应当为每一类主题应用不同的关键词进行检索，然而实际操作后发现即使主题相同，不同的学者在措辞及题目选取上会有一定范围的差异，比如用大数据方法发现贫困生的研究可以叫做“学生经济状况挖掘”也可以叫做“贫困生发现”，因此为每一类研究各指定一组关键词进行检索非常容易发生漏误的状况，因此本文在检索中采用先使用“一卡通”和“数据挖掘”为关键词，检索出绝大部分符合要求的文献，再根据三类主题的特点，以“数据挖掘”“成绩预测”为关键字进行检索，在检索出的文献中选取和高校相关并且基于一卡通数据的文章，补充出遗漏的第二类主题相关文章。同理，以“智慧校园”为关键字进行检索补充出遗漏的第三类主题相关文章。

以“一卡通”“数据挖掘”为关键字进行检索，共有文献 29 篇，时间跨度为 11 年。按照时间进行排列，2016—2014 年每年各 5 篇，2013—2011 年每年各 1 篇，2010—2008 年每年各 3 篇，2007 年成果为 1 篇，2003 年 1 篇，学科主要集中在计算机软件和计算机应用、高等教育、互联网技术和无线电电子学。研究层次主要集中在自科下的工程技术。其中 SCI，CSSCI，EI 来源期刊文章数量为 1，按照主题的相关性，共选取 21 篇进行分析。

以“数据挖掘”“成绩预测”为关键字进行检索，共有文献 44 篇，时间跨度为 13 年，大多和高校领域的研究不相关。其中，SCI，CSSCI，EI 来源期刊文章数量为 4，与高校领域的研究均不相关。按照主题相关性和典型性，共选取 3 篇作为补充。

以“高校”“智慧校园”为关键字进行检索，共有文献 466 篇，时间跨度为 6 年。其中，SCI，CSSCI，EI 来源期刊文章数量为 29，按照主题相关性和典型性，共选取 2 篇作为补充。

本文所选取 26 篇文章的时间跨度和涉及领域如图 1、2 所示。

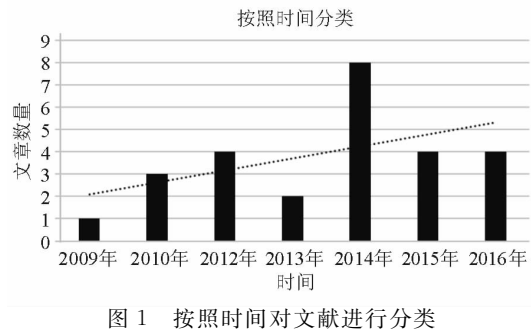

图 1　按照时间对文献进行分类

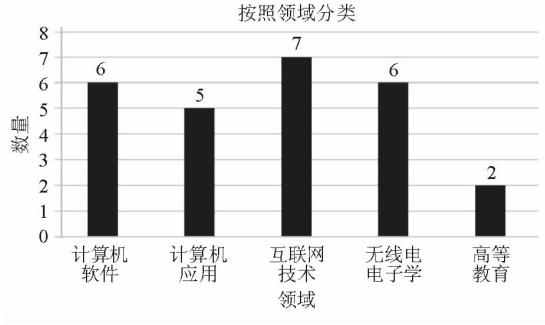

图 2　按照领域对文献进行分类

由于在检索时使用了先查找出所有符合“一卡通”“数据挖掘”关键字的文章，再检索遗漏补充的策略，跨库选择也较为全面，并且在第一次检索出的文章中除掉了几篇和高校应用场景无关的文章后全部选中进行分析，因此这 26 篇文章可以客观反映出这一领域的大数据应用的发展状况。从统计图可知，文献数量具有明显的随着时间而增长的趋势，呈现

出蓬勃发展之态。从来源上分析，硕士论文占有一定比例，证明越来越多的年轻研究人员开始对这一领域感兴趣，显示了一定的发展潜力。这些研究主要属于传统的计算机技术及应用领域，学科交叉领域的成果较少，学科之间的渗透不够深入，还有巨大的研究空间和研究价值。在期刊来源的文章中，SCI，CSSCI，EI 来源期刊文章所占比例极低，此领域的研究还不够成熟，仍然处于起步阶段。

3 研究现状

大数据分析成为一个研究热点以后，许多研究者也将目光投向了校园平台数据的大数据分析。校园平台数据可分为师生日常活动产生的数据和教务数据两类，而前者载体多为一卡通，因此逐渐形成了以一卡通数据和教务数据为主、后勤数据为辅、着重研究学生经济状况和学业状况的特点。以下，将按照前文所分的三个类别分别进行梳理。

在对学生经济状况的挖掘中，因为消费金额是消费能力的直接体现，所以大多从学生的一卡通消费金额入手。而一卡通所记录的消费数据并不是学生全部的消费数据，更多的消费发生在校外，因此在此类分析中，为了提高准确率，往往还要加上对学生行为的分析。

徐剑[1]对学生的消费水平进行了聚类，旨在探寻学生的消费模式与成绩之间的相关关系，为学校有关部门的科学决策提供一定的依据。他重点研究了关联规则挖掘算法，并将经典算法发展为基于布尔稀疏矩阵的算法，提高了效率。首先，使用 K-MEANS 算法，将每人每月的食堂消费额分成高、中、低三档（聚类时，每一档的初始值为最高消费、平均消费和最低消费）。再将每个月吃早餐的次数按照十五次为界，划分为生活规律和不生活规律两类。最后，把学生成绩按照不及格、及格、中等、良好、优秀分成五档，并且将学生的这三个指标转成高维矩阵，运行 Apriori 算法，寻找其中的相关关系。其先聚类、后进行相关分析的处理思路及模型对之后的研究产生了较为深远的影响，后人在其基础上改进，产生了新的模型。

同年，陈建兵[2]也对贫困生的消费状况进行了挖掘，主要研究算法，与徐剑直接寻找相关关系不同，研究从贫困生和非贫困生的消费状况差异的角度进行。在对数据进行了简单的规约之后，使用 Apriori 算法和 FP-Growth 算法进行相关分析，用 SQL 语句统计了每个学生每顿饭的平均价格和平均次数，然后把学校贫困生的数据也做相同处理，两个结果对比，找出偏离点。

张佳[3]主要研究了鉴别贫困生和对热水需求量最大的时间段，以及分析了学校内商铺的营业状况。他借鉴了徐剑对于消费金额进行聚类的思想，具体操作中使用不同的算法。在贫困生分析中，使用 Microsoft ID3 智能决策树算法，将学生的消费额分为高、中、低三个档次，并且使用相同的算法，将学生在食堂的消费次数分为高、中、低三个档次，再在消费额处于低档的同学中除去消费次数也处于低档的同学，剩下的自动设立为贫困生备选名单。该模型对徐剑的模型进一步发展，考虑了消费的频次和稳定性，本质上是计算平均消费金额。而在热水消费分析和商户的营业状况中，使用 Excel 对消费数据进行加总和统计，得出消费热水多的时段和商户的营业状况。该文章虽使用了经典的数据挖掘算法，但由于模型的限制，主要实现的功能仍然集中在传统数据分析领域。

张林红、刘红梅[4]在徐剑的基础上，改善了原文章的模型，用早餐时间、早餐次数、早餐时间标准差三个维度共同度量生活习惯的规律与否，提高了在相关分析中的准确性和模型的实用性。

费小丹、董新科、张晖[5]沿用了张佳统计消费频率的思想并且进一步验证和发展了张佳的模型。因为校内消费的价格相比校外更低，作者首先提出了贫困生在校内的消费频率更高的假设。作者将所有学生的数据按照消费总金额、总次数、每次平均金额、日均总金额进行聚类，发现贫困程度和消费次数成正比，和消费总金额、日均消费总金额、每次平均金额成反比。因此，提出了贫困指数的公式，帮助贫困生的认定工作。作者的核心思维是均次消费金额的比较，有一定的参考价值，然而模型假设和简单的平均消费统计基本相同。

董新科、张晖[6]研究的重点在于挖掘算法的比较。该研究比较和分析了几种聚类算法在一卡通数据分析中的作用，主要目的在于得出最适合挖掘该类数据的算法。测试任务为对每个学生的消费记录按照消费总额、消费次数、人均消费金额、次均消费金额进行聚类，评价标准为简单易用性和有效性。最后测试结果为 K-MEANS 最适合对校园卡数据进行聚类，为其后的校园数据挖掘采用合适的方法提供了一定指导和依据。

薛黎明、栾维新、李志淮、樊铁成[7]分别从时间、地点、消费金额等几个维度对校园卡的消费数据进行了分析，旨在挖掘出学生消费的高峰时间、地点和消费层次。研究中大量应用聚类算法，不存在具体的模型，是以聚类算法代替单纯统计，将经典数据挖掘算法引入传统数据分析的一次尝试。按照时间、对消费记录进行等距离离散化，统计出了几个消费时间峰值。再加入学生性别一项进行交叉分析，分析性别对消费的影响。按照消费地点对消费记录进行统计，可得每个消费地点的记录总数，分析出每个地点受欢迎的程度。加入学生类别一项进行交叉分析，可得每类学生对每个消费地点的偏好程度。按照消费金额对消费记录进行分析，首先用等距离散化和 kohonen 神

经网络聚类，然后使用决策树算法进行分析，得出每个类别的消费者所属的消费金额层次，是使用神经网络进行校园数据分析的一次探索性实践。

姜楠、许维胜[8]基本沿用了徐剑的思路，但从提高聚类准确度的角度改善了其模型。主要用迭代、选取方差之和最小的一组的方式寻找 K-MEANS 最佳聚类中心值，优化了聚类的结果。然后，按照食堂平均消费金额、超市消费金额、用卡次数和常去地点对学生的消费模式进行聚类。并用同样的方法对学生以奖学金为代表的学习成绩模式进行聚类，对以图书馆借书次数为代表的学习习惯模式进行聚类。最后，采用基于稀疏矩阵的 Apriori 算法进行相关分析，探讨其中的相关关系，有利于从学生的生活习惯中探究影响学业成就的因素。

樊搏、姜玉国[9]继承徐剑对消费金额聚类的思路，运用 K-MEANS 算法以及支持向量机的算法，将学生的食堂消费金额分为五个档：贫困、较差、中等、较好、优越，自动区分出贫困生。该模型优势在于两种算法可互相检验分类结果，但文章主要使用食堂消费数据，可以纳入更多的学生消费数据，进一步提升模型准确度。

樊搏、吕艳芝[10]在前人主要依据消费金额的模型基础上，从心理学入手，将圈存数量和早餐次数纳入贫困生挖掘模型。该研究介绍了目前高校信息化建设和贫困生认定工作的现状，并且在该作者《基于数据挖掘的贫困生认定辅助系统设计》一文的基础上得到了进一步的发展。该文章把仅从食堂消费平均金额的额度判断标准发展为消费状况、早餐状况和圈存状况的多方面的模型。作者认为，贫困生的食堂平均消费更低，并且大多更加勤勉，因此，早餐时间更早也更加规律(就餐时间稳定)，并且由于心理的不安全感，圈存时一般小额多次。因此这几个方面可以辅助学校进行贫困生的识别工作，帮助校方进行科学决策。

Chu Gu 等[11]极大地发展了现有的学生经济状况挖掘模型，打破了过于依赖食堂消费金额的现状，利用高校一卡通数据和校园平台上的其他数据建立起了综合的模型，对贫困生进行自动的识别。模型的搭建主要围绕一卡通数据、校园网络的使用情况和学生在校园内的轨迹画像三个方面进行。模型的建立上，在一卡通数据方面主要考察消费额度变化、消费行为的规律性和消费的沉默期；在校园网络的使用方面，主要考察上网时长与费用、上网周期与站点记录、上网流量序列；校园内轨迹画像方面，主要针对学生的行为模式挖掘和时空感知属性。本文的基本假设在于行为模式相似的学生往往具有相似的经济条件，因此获得相似的奖学金额度。文章创立 dis-HARD 学习算法，计算学生特征与奖学金的相关性(路径越短则特征越相合，相关性越高)并对比了其他数据挖掘算法(SVM，MKL，multi-label LSI，TODMIS)，证明了 HARD 学习算法的优越性，该算法在电子科技大学已经有了成功的应用。

高校作为服务于全社会的教育部门，十分重视学生的学业发展状况，学业发展状况最为直观的指标是学生的课程成绩。在对学生学业状况的挖掘中，主要集中在对于学生成绩的预测，其中途径之一是用以往的成绩预测未来的成绩情况或排名情况，途径之二是用学生的日常学习及生活相关数据预测其成绩情况或排名情况，二者联合使用的情况也存在。由于学生的成绩是具有趋势性的，以往排名靠前的学生极有可能此次也排名靠前，因此第一种途径过于显然，第二种途径就成为了研究的热点。从所使用的模型来看，与学生经济状况的挖掘呈现出的明显脉络和继承发展关系不同，模型和方法多样化、关注点和切入角度各不相同是其特点。

武彤、王秀坤[12]旨在预测单门课程学生的通过状况，并发现影响学生通过单门考试的因素有哪些. 应用 C4.5 决策树算法，通过学生性别、对基础知识的掌握程度和上机时间来预测学生考试的通过率。共抓取了五百组学生数据，其中三分之二作为训练集，三分之一作为检验集。最后结果显示，预测的准确率高达 87.5%。该研究是采用自动化的方法对学生成绩进行预测的一次尝试，为以后的研究提供了一定程度的参考。

罗永国[13]使用改进的遗传算法与 BP 神经网络，从学生到课率、历年排名状况、平时作业成绩、小测验成绩几个方面来构建模型，预测学生最近一学期的期末排名，并和其最近一学期的期末排名对比。共抽取了 5000 名学生的数据，其中 4500 组用作训练集，500 组作为测试集。预测的效果十分优秀，误差不超过 5%。这是从学生学习状态来对学业成就进行预测的实践之一，为之后的研究提供了参考。

刘志妩[14]应用 C4.5 决策树算法，用学生所有科目的成绩数据构建决策树，找出其中的关键节点，以此来探究学生各科成绩之间相互依赖、相互影响的状况。从思维上来说，以决策树为工具对学生的学习状况做出预测，将决策树应用到传统的相关关系分析里，有利于对其后的研究在方法上的创新提供参考。

黄建明[15]与刘志妩[14]的思路相通，切入角度较为相似，但把同期的数据扩展为不同时期的数据，因此得以研究先导对后续的影响。选取了五届学生七门主要课程的成绩，离散化后，通过贝叶斯网络构建出贝叶斯图，通过节点之间的链接来显示相关关系的有无，通过权值来显示概率和强弱，从而挖掘出七门课程中先导课程对后续课程的影响及程度，也能在已知一部分成绩的情况下预测其他科目的成绩，是对贝叶斯网络应用于学业表现上的一次应用。但它关注的重点在于课程之间的互相影响，而不是预测学生全面的学业表现。

吕红胤、连德富、聂敏、夏虎、周涛[16]利用校园平台上的一卡通数据对高校学生的学业成就进行预测。模型的设计较为合理,围绕努力程度和学生生活的规律性进行,参考了徐剑对于生活规律性的度量,并从就餐的规律性扩展到了全面的生活的规律性。具体而言,努力程度由自习次数和上课次数反映,生活规律性由出入宿舍的规律性、就餐的规律性、洗澡洗衣服的规律性、购物的规律性反映。研究证明努力程度和学生的成绩呈正相关,并且一个学生的成绩往往与其朋友的学习成绩相关。本文采集学生六个学期内的上述数据,前五个学期的数据作为训练集,提取关联规则,第六学期的成绩作为预测内容,研究证明该模型的预测效率达百分之九十以上。该研究成为了该团队其后推出的一系列校园大数据分析的先导。文中表示,为了保护学生的隐私,所以把成绩换成排名,并进行归一化。归一化只是为了取消各个学院之间评分标准和课程的差异造成的成绩差异,没有这一步就无法进行客观的比较。成绩和排名的转换脱敏效果十分有限。大数据背景下,数据的脱敏一直都是一个问题,因为数据量的庞大,互相对照会使得脱敏失效。

谢星宇和张颖璐[17]从自动分类的角度切入,将涉及心理学、教育学和管理学的成绩预测问题转化为纯粹的算法问题。将学生前两个学期的成绩数据、一卡通数据以及图书馆借阅数据作为训练集,挖掘出其中的相关关系。此研究的主要贡献在于改进了 TrAdaboost 算法,并用改进后的 TrAdaboost 算法对学生第三学期的成绩作出预测。

蔡兴雨等[18]利用问卷的方法收集数据,然后利用粗糙集理论的属性约减算法和属性提取算法挖掘出影响高校学生成绩的关键因素以及这些因素和学生成绩之间的依赖关系,有利于改善教师的教学方法及学生的学习方法,提高学生成绩。数据约减后,一共保留了十四个项目。提取其中的有效规则后发现,学生的成绩与主观的学习态度以及客观的家庭环境都有关系。意外发现女性学生的成绩普遍高于男性学生,同时还发现母亲的职业比父亲的职业对于孩子成绩的影响更大。该研究的独特之处在于并不预先设立一个预测模型,而是围绕数据,进行开放式的探究活动使得研究不局限于初始的假设,可以挖掘出让人意想不到的结论,比如本研究的意外发现。

李彤彤等[19]认为学习干预对于学生的发展十分重要,然而学界对于此的关注并不太多。作者围绕干预引擎,从学习者状态识别、干预策略匹配计算、干预策略实施、干预效果分析四个方面搭建了自己的学习干预模型。学习者的状态识别主要包括学习风格、学习进度、学习互动水平与学业成就四个方面。数据来源分为量表采集和线上教育平台数据。主要的分析方法为聚类,首先建立干预库,然后根据学生状况的不同,经过计算给予干预库中最优的方法,干预效果由系统和教育者共同追踪。这是校园平台大数据在学习干预中的一次重要尝试,可以纳入具体的实践方法,使文章具有更多的实践意义。

高校在承担教学任务以外,大量的后勤工作也不容小视,后勤服务效率与质量的提高可以极大地提高高校的整体服务质量。近年来,通过数据分析和数据挖掘的方法对高校后勤及各项事务进行分析,以期对其进行流程的改造及重组的研究兴起。这类分析的主题集中在学校的设施服务情况和后勤服务,如食堂和澡堂等。由于此类研究常和实际的需求及实践活动(建设统一化的平台,改善经营绩效)联系在一起,实用性高于科研性,因此挖掘的深度并没有特别深入,在数据分析的方法上主要使用统计方法,但是应用大数据的分析方法也已经成为了新的趋势。

张兵兵等[20]采用了 sql server 中自带的数据挖掘算法,主要采用了 Microsoft 决策树算法和 Microsoft 关联规则算法,分析了学生的哪些特征和丢卡次数密切相关,最后得出结论学院是最强的因素。研究生院和国际教育学院的学生最爱丢卡。从得出的结论中分析其原因,有可能模型的设计中未能直接定义到影响丢卡次数的主要因素,而学院又和这个因素有强烈的相关关系(比如同一个学院的学生表现出相似的行为模式,而这个行为模式和丢卡与否相关)。

许华虎等通过决策树算法分析一卡通中记载的学生体育锻炼数据,为学生的体育锻炼情况做出分级。作者从常规的分类方法中受到启发,选取学生的活动强度、性别、年龄、体质因素,应用 ID3 算法生成决策树,以此对学生的体育锻炼情况做出分级评价。这是使用大数据分析方法进行校园事务分析的一项有意义的尝试,尽管思路与常规的分类比较相似,但在使用的方法上有一定程度的创新。

许彩娥等旨在建立一个以校园一卡通为介质的校园综合门禁管理平台,切入的角度为数据治理,为其后的数据集中分析打下了基础。针对目前高校门禁系统存在的介质不统一,流水数据分散,认证数据重复存放的问题,设计出了一套以校园卡为唯一介质,流水数据集中存放、统一管理、刷卡即回传数据的综合门禁管理平台,克服了目前存在的问题,对门禁数据的统一管理、提升校园后勤服务质量和后期对于门禁数据的分析具有重要意义。

石飞飞设计并且实现了一个智慧校园挖掘平台,尝试了三类数据挖掘和分析:对后勤服务的挖掘、对于学生的挖掘以及对于教务信息的挖掘。对后勤服务的挖掘中,主要使用了直方图和散点图,并未涉及经典的数据挖掘算法。在对学生信息的挖掘中,主要使用 C4.5 算法,将学生按照消费金额和消费次数聚类。然后,利用 Apriori 算法分析学生的在网时长、消费水平、图书馆借阅、出勤信息、门禁情况和成

绩是否有相关关系。在对教务信息的挖掘中,使用K-MEANS算法对各类数据进行聚类。该研究是一次比较综合的校园平台数据挖掘实践,方法上大数据分析方法和传统的统计方法兼具。在对学生信息的挖掘中,模型综合性较强,具有一定的参考价值。

陈锋通过一卡通记载的学生就餐记录统计出了学生集中就餐的峰值时间段以及峰值就餐人数,为食堂提供了安全及营业时间方面的建议。同时,作者还按照正常行课时间与节假日时间,分别统计了学生的消费金额、用餐时间和刷卡消费次数,根据学生在不同时间的消费行为模式,为校内商户提供经营的建议。在数据分析方法上主要为简单的统计分析,属于传统数据分析领域。

马秀麟、衷克定、刘立超用量化的手段判断学生的评教数据是否具有有效性。首先,使用克朗巴哈阿尔法系数法,判断学生的评教数据是否具有内部一致性(是否具有信度),然后使用学生每年的评教数据与督教的评教数据进行相关性分析,看是否具有一致性,从而证明学生的评教数据是否具有有效性。最后,对于那些评教分数比较低的老师,用相关性分析来分析到底哪一方面对于评教得分最有影响,从而对教师的教学工作提出建议。本研究是分析评教结果的主要相关因素,用量化的方法代替了人为的主观评估,结果更加客观。

金培莉、王晓震通过实例探寻了校园卡数据对于学校决策支持帮助的可能。作者应用了食堂就餐数据分析、教师就餐补贴分析、热水洗浴分析三个实例。就方法而言,属于简单的数据加总和平均数分类,处于传统的统计分析领域。

在后勤及其他事务的数据挖掘与分析中,从数据的分析方法来看,大数据的分析方法已经成为主流,特别是分类方法。K-means 和 C4.5 已经成为最常用的方法,但仍有一部分研究使用平均数等简单的统计方法。大数据方法并不天然比统计方法更加高级,它们有各自不同的应用场景,然而在上述文章中,使用统计分析方法的研究很大一部分并没有充分发掘出数据的潜在价值,而是简单的数据加总和分类,而使用大数据分析方法的研究中也存在着模型效果不佳等问题。从目前的状况看来,这类研究并不成熟,还有进一步研究的必要和空间。

4 文献评述

通过对以上三类主题的文章进行分析和梳理可以发现,学者们应用大数据的分析方法对于校园平台上的数据进行分析尝试,并且随着时间的推移,模型呈现出越来越成熟、方法也越来越智能化的趋势,为校园事务的决策提供了支持,也为后续的分析尝试提供了重要的参考和宝贵的经验。然而,由于领域的不成熟和客观条件的限制,仍然能发现以下三个问题:

分析的数据种类和来源过于单一。以上大部分分析的模型都严重依赖学生的一卡通消费数据,在对学生的学习行为进行分析时,又严重依赖图书馆的自习和借阅数据。总体而言,数据种类较少、来源单一。大数据分析的魅力之一在于利用多元异构化的数据建立全面的模型,从行为入手,达到准确的分析和预测效果。数据来源的单一性直接导致对被分析对象的行为掌握不全面,因此影响分析和预测的准确度。目前,对于一卡通消费数据和图书馆自习及借阅数据的依赖有其客观原因:进行大数据分析的基础是数据的可得性。由于一卡通涉及消费、账户安全问题,通常受到校方的高度重视,要求进行统一的信息化管理,而图书馆借阅每日庞大的流水数据量也促使校方对于该业务迅速进行电子化和信息化,一卡通消费数据与图书馆借阅数据通常是校园数据中数据治理程度最好、质量最高、取得最为容易的,因此最便于进行数据分析的研究。而其他方面的数据,若要取得并进行研究,还依赖于整个校园事务进行信息化和信息治理的程度,而这个程度通常低于前两项数据的程度,给研究造成一定不便。因此在设计模型时,会倾向于对于其他质量不高的数据进行避免,因此形成了这样的依赖现状。而在前面章节的梳理中,也可以发现,对于学生经济状况的挖掘、贫困生的发现的研究数量比其他两类稍多,其中的原因之一也在于数据的来源。一卡通消费数据可以作为可以获得的、反映学生经济变量的重要指标。随着数据治理的开展,这样的情况会有所改善。

模型单薄。从以上章节的分析梳理中可以发现,对于学生经济状况的挖掘,无论采用什么方法进行,核心思想基本在于统计学生食堂就餐次数和总消费,如关于贫困生发现的研究,即筛选出消费总金额低而消费次数高的学生。这样的模型本质在于筛选出平均单次消费金额低的学生,符合贫困生的消费模式。然而,符合这一模型的,除了贫困生以外,还有在校外就餐,仅在学校购买一些小点心的学生,甚至还有一些处在节食减肥阶段的学生。后两类学生的行为并不是经济状况导致的,对模型的准确率造成很大的影响。仅从消费数据上考虑的单薄模型并不利于识别的准确率,综合学生的行为模式一起分析,建立更加全面的模型会有更好的效果。在对学生学业状况的挖掘和其他校园事务的挖掘中,建立的模型和对结果的分析通常只局限在自己研究的小问题内,学科之间的交叉不够深入,也限制了模型的准确度和对研究结果的进一步解读。早在 1984 年,Astin 就在院系影响力理论的基础上提出了经典的 IEO 模型。他认为,学生的学业成就受投入和院系环境的双重影响。其中,成就部分不单指学生的学习成绩,个性及价值观也包含在内。而学生的投入包含学生入学前的经历、家庭背景等。院系环

境还包括院系氛围和文化、教学设施及风格等等。而 Astin 后来的一系列文章又深化了该理论，使其成为教育学领域的基础。把学业成就简单等同于成绩必然对分析结果的准确性和实际应用产生影响。

方法和模型之间不匹配。从目前研究所使用的分析方法来看，大数据的分析方法逐渐成为主流趋势，最常使用的为聚类和相关分析，传统的统计分析方法正逐渐被大数据分析方法取代。然而从目前的文献看来，大数据的分析方法并没有发挥出其优势，只是作为传统统计方法的替代，如在学生经济状况挖掘中普遍存在的用聚类方法统计平均消费的做法，而如果应用统计方法按照平均数过滤，效果相当而成本远远更小。传统的统计分析和现在的大数据分析方法并没有优劣之分，只有各自更加适应的应用场景，没有必要在不适合的地方，特别是传统统计方法已经有成熟和便利的处理方式的场景下盲目使用大数据分析方法。大数据的分析方法对于研究特殊性，而不是共性有着非常独到的优势，而目前这种趋势在一定程度上浪费了数据中信息的丰富性。

根据以上三个问题，可以针对性地进行改进。首先，对校园平台数据进行数据治理。从以上分析可知，数据来源和数据质量对研究的可行性和质量有着重要的影响。良好的数据治理程度、丰富的数据来源和数据的可得性、便利的数据提取接口是进行数据分析的基础。进行数据治理后，除已经大量使用的一卡通消费数据和图书馆借阅数据以外，其他学生的行为数据也能够进行分析，有利于模型的综合化，改进对现有数据的过度依赖，提高研究质量。许彩娥团队的数据治理尝试就是一个示例。

而针对模型单薄的问题，可以考虑引入行为模式分析，建立更加综合性的分析模型，分析模型的搭建不应该被研究主题的领域完全限制。大数据的魅力在于从杂乱中寻找相关关系，而这种相关关系在很多时候都是出乎意料的，一个经济问题的表现也是方方面面的，而不局限于经济领域。因此，在搭建模型的时候要全面考虑，从行为模式入手，综合分析。并且，各学科的合作会使得模型的搭建更加合理。

针对方法和模型不适配的问题，统计分析方法在提取共性方面已经十分成熟，在前期数据的清洗和特征的合并方面具有十分重要的作用，可以为后期的大数据分析打下良好的基础。二者可以考虑联合使用，它们并不是互相排斥的，不需要为了追求高技术含量而在统计方法可以处理的场景使用大数据分析方法。通常，分析效果不佳都与常见问题的解决有着重要的关联，因此要注意细节，而不是只要使用了最新的方法就能达到最好的效果。

针对目前的研究现状和存在的问题，在接下来的研究中，可以考虑这样的研究思路：在数据的清洗和特征合并等前期工作中使用传统的统计方法，提高效率。在模型的构建中引入行为模式的分析，可借鉴商业上已经应用成熟的用户画像系统，从行为模式入手，建立更加综合的模型。最后，在分析结果的解读中，可联系其心理学的依据，进行更加全面和更有深度的解释，使得研究具有更加明确的现实意义。

参考文献：

[1] 徐剑. 基于一卡通数据的消费行为与成绩的关联性研究分析[D]. 南昌：南昌大学，2010.

[2] 陈建兵. 利用校园一卡通数据优化高校贫困生认定系统[D]. 成都：电子科技大学，2012.

[3] 张佳. 数据挖掘技术在校园一卡通系统中的应用研究[D]. 苏州：苏州大学，2013.

[4] 张林红，刘红梅. 基于一卡通数据分析的学生早餐习惯与成绩关联规则挖掘[J]. 阜阳师范学院学报(自然科学版)，2014(4)：92-95+105.

[5] 费小丹，董新科，张晖. 基于校园一卡通消费数据的高校贫困生分析[J]. 电脑知识与技术，2014(20)：4934-4936.

[6] 董新科，张晖. 基于校园一卡通消费数据的几种聚类算法的分析比较[J]. 计算机系统应用，2014(1)：158-161+183.

[7] 薛黎明，栾维新，李志淮，等. 数据挖掘在校园一卡通消费数据分析中的应用[A]. 中国高等教育学会教育信息化分会. 中国高等教育学会教育信息化分会第十二次学术年会论文集[C]. 中国高等教育学会教育信息化分会，2014：8.

[8] 姜楠，许维胜. 基于校园一卡通数据的学生消费及学习行为分析[J]. 微型电脑应用，2015(2)：35-38.

[9] 樊搏，姜玉国. 基于数据挖掘的贫困生认定辅助系统设计[J]. 软件导刊，2015(12)：134-135.

[10] 樊搏，吕艳芝. 基于一卡通数据中心的贫困生辅助认定分析[J]. 科教文汇(上旬刊)，2015(11)：122-123.

[11] GUAN C，LU X J，LI X L，et al. Discovery of college students in financial hardship. 16th IEEE International Conference on Data Mining (ICDM 2016)

[12] 武彤，王秀坤. 决策树算法在学生成绩预测分析中的应用[J]. 微计算机信息，2010(3)：209-211.

[13] 罗永国. 基于改进的遗传算法的学生成绩预测模型[J]. 科技通报，2012(10)：223-225.

[14] 刘志妩. 基于决策树算法的学生成绩的预测分析[J]. 计算机应用与软件，2012(11)：312-314+330.

[15] 黄建明. 贝叶斯网络在学生成绩预测中的应用[J]. 计算机科学，2012(S3)：280-282.

[16] 吕红胤，连德富，聂敏，等. 大数据引领教育未来：从成绩预测谈起[J]. 大数据，2015(4)：118-121.

[17] 谢星宇，张颖璐. 基于改进的 TrAdaboost 算法的学生成绩排名预测[J]. 计算机与现代化，2016(2)：122-126.

[18] 蔡兴雨，徐怡，程智炜. 基于粗糙集理论的影响高校学生成绩因素研究[J]. 计算机技术与发展，2016(11)：1-5.

[19] 李彤彤，黄洛颖，邹蕊，等. 基于教育大数据的学习干预模型构建[J]. 中国电化教育，2016(6)：16-20.

[20] 张兵兵，王建，张建威，等. 数据挖掘在校园一卡通系统中的应用初探[J]. 数理医药学杂志，2009(5)：572-575.