

基于大数据算法的高校学生学情数据研究

李 欣

成都锦城学院计算机与软件学院 四川 成都 611731

【摘要】随着信息技术的发展,结合大数据和人工智能技术,智慧智能教育是数据时代的发展趋势和应用热点^[1],而实现个性化教育的前提和保障是形成丰富完备且有指导意义的学生画像标签,进而充分了解每名学生的长短板和学习习惯,针对性制定培养方案,从而因材施教实现千人千面的个性化教育。本文基于成都锦城智慧校园信息化数据,结合成绩管理、图书借阅、一卡通消费、学生基本信息等维度构建画像指标,通过 Pearson 相关性、回归分析算法,探索分析相关度,进而构建了锦城学院学情大数据学生画像标签库。

【关键词】智慧教育;大数据;学情画像;Pearson 相关性

1 引言

现代社会已进入大数据时代,基于用户画像的智能化个性化信息服务深刻地改变了人们的生活,研究用户画像对社会的发展具有重要作用,大数据构成了用户画像研究的数据基础,随着计算机和信息网络技术的发展,用户画像研究和实践不断发展,并在企业中得到了非常充分的应用和推广。

目前,也有一些学校尝试将用户画像的思想引入高等教育,其中图书情报与智慧图书馆是学生画像研究的重要领域,另外四川电子科技大学做了学生画像的应用拓展,将大学生学习行为和自习室打水频次相结合,挖掘学生学习规律。对于教育行业来说,管理信息化的支撑更是十分重要。科技部新一代人工智能发展研究中心和罗兰贝格管理咨询公司联合发布《智能教育创新应用发展报告》指出,智能教育当前已经从教育辅助向价值创造阶段过渡。如何将智能教育引入高校并进行学生画像^[2]并进行个性化教育是新时代教育工作要关注的重点

和核心。

因此,本文深入研究大数据学生画像技术在我校人才培养中的应用,首先整合多个系统中的学生数据(包括学生基本信息、教务管理、成绩管理、图书借阅、一卡通消费等系统),通过分析每名学生的高考成绩、学期平均成绩、图书借阅习惯、消费习惯等对学生进行多维度画像,来刻画学生的状态、定位、及学习习惯,了解学生的长板,为个性化教育提供数据支撑,切实发挥大数据技术对改进人才培养目标、课程体系等培养环节的推动作用,为人才培养模式的构建提供重要数据基础。

2 模型构建整体框架

锦城学生大数据学情画像的数据来源于锦城智慧校园共享数据包括成绩管理、图书借阅、一卡通消费、学生基本信息等数据,研究的对象为全校 2018 级学生,探索的周期为 2018-2019 学年历史性学情数据。学生画像模型构建的整体框架如下图所示

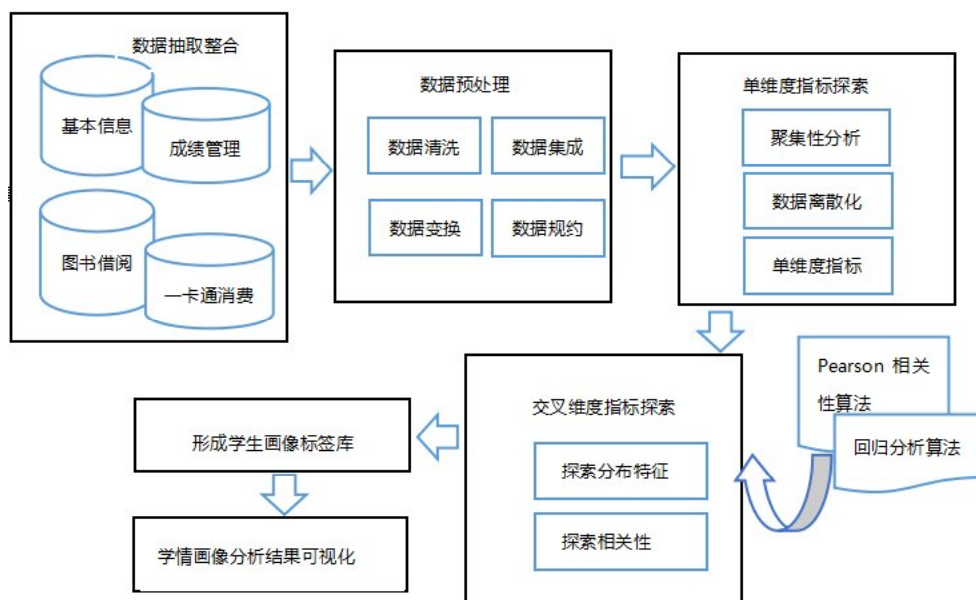


图 1 学生画像模型构建的整体框架

整体实施过程为:在学校智慧校园系统中根据数据挖掘需要进行数据的抽取整合,目前抽取整合的数据源

包括包括成绩管理、图书借阅、一卡通消费、学生基本信息等数据,经过数据的预处理(数据清洗集成变换规

约等)形成可通分析建模使用的干净数据集,经过单维度的探索分析了解数据整体分布状态结构,通过交叉维度指标的相关性分析,寻找挖掘维度间的相关性,最终形成学生画像标签库并实现分析结果的可视化。

3 模型构建步骤

3.1 特征维度

本文加工的特征维度为锦城 18 级学生 18-19 学年的成绩分层、入学成绩分层、图书借阅等级、一卡通消费等级等,及各维度的统计指标,维度分层的依据特征大数据的分布特点,结合连续型数据正太分布的 3σ 原理及分箱技术离散化得到,部分画像特征维度和离散化结果如下:

客观属性维度:学生自然属性特征、层次、学院、专业及客观属性统计数据

加工特征维度:18-19 学年成绩分层(优秀、良好、中、一般、不及格);入学成绩分层(结合不同地域省份及文理科的不同分为 5 个层级);图书借阅等级分层(结合图书管理系统年度借阅记录及图书科目的不同分为高频借阅、中频借阅、低频借阅、惰性借阅);一卡通消费等级分层(结合可采集年度一卡通消费终端分为高消费、中偏上消费、一般消费、可能异常四个等级)

3.2 具体探索分析流程

(1) 成绩维度。首先,对成绩进行单维度数据画像,数据来源为智慧校园数据库 2018-2019 学年 18 级全校大一新生各科成绩,分析其分散分布情况及数据整体的偏度、峰度、正态性等,从宏观上把握整体的成绩学情数据。其次,对成绩和学生层次、课程类型(通识课程、专业基础课程、实验实践类课程、专业综合课程等)进行维度透视分析,探索不同层次学生成绩等级分布情况,掌握不同层次学生分数段痛点,了解学生不同课程类型的学习偏好,精准掌握学生能力长板及水平欠缺短板;最后通过将 18 级统招本科生入学成绩和入学一年(18-19 学年)的成绩进行关联分析,探索入学后的学生成绩和高考学生成绩是否具有线性相关性,即是否存在高考成绩越高,该段优+良学生的占比越高,而不及格学生的占比越低,并利用皮尔森算法度量相关性的强弱,进而关注两级异常群体,即入学成绩较高,入学一年成绩反而下降,这部分同学我们应该给予更多的关注,让优秀的生源最大化的发挥其价值和能动性,不要在大学生涯开始之初掉队,相反对于入学成绩较低,入学一年后成绩优异明显的学生,教育工作者在鼓励学生的同时可总结成长方案,便于成功培养路径的推广。

(2) 图书借阅维度。图书借阅是大学生活中一个重要的学情数据收集分析环节,本文的对图书借阅的分析维度主要从三个方面。首先,是单维度的图书借阅等级全校的分布状态,从数据层面了解全校的学生图书借阅活跃度,并结合学院、专业、年级等维度进行交叉透视分析,探索由于学院、专业、年级不同所带来借阅频度差异;其次,分析不同时段不同借阅等级学生的借阅规律和时间偏好;最后,对图书借阅等级与借阅图书门类的交叉分析,挖掘学生兴趣爱好和学习痛点。

(3) 一卡通消费维度。一卡通消费数据作为学情数据补充记录了学生除学习以外的部分生活习惯数据,

从对这部分数据的分析探索中,我们能够宏观掌握学生的消费等级分布及不同学院、不同专业、不同层次的消费差异,进而,结合消费的时间、消费类型,分析大学生的消费习惯,挖掘业余时间的分配及行为习惯。

(4) 交叉维度探索。交叉维度分析主要将上述三个单维度的特征进行交叉融合,深度挖掘学习成绩与图书借阅习惯,学习成绩与一卡通消费的内在联系及规律,学习有价值的规则。首先,利用单维度探索时的结构化特征数据,分析学习成绩与图书借阅习惯的相关性,从清洗的数据整体训练结果上看,学习成绩与图书借阅成正相关,采用回归分析、Pearson 相关性等算法抽取专家样本计算,相关度为:0.618。从数据看出,成绩优异的学生大都具有良好的借阅习惯和学习自主性,而成绩一般和不及格的学生,从趋势上高频借阅的比例明显降低,惰性借阅的比例明显增高,大学期间培养良好的学习习惯很重要,借阅习惯是其中的一个方面,应给予关注。其次,进行学习成绩与一卡通消费的相关性探索,从我校的整体数据上看,学习成绩与消费水平成弱负相关,采用回归分析、Pearson 相关性算法,相关度为-0.488。从成绩分层的角度来看,优秀+良学生高消费比例较低,大多数为一般消费,消费积极性不高;不及格学生高消费比例高于全校平均水平。一卡通维度与学习成绩的相关性分析是一种探索性尝试,从消费层面,刻画大学生人群的整体趋势画像。

3.3 学情画像数据可视化应用

本文从学习成绩、图书借阅维度、一卡通消费维度及融合交叉维度四个层面构建学情数据画像标签体系,该体系的特征工程借助了锦城大数据实验室科研平台进行周期任务调度,使得画像标签具有很好真实性和时效性,标签的可视化呈现借助了锦城大数据 BI 平台,使得教学管理活动中的不同角色用户能够实时访问可视化画像分析终端。该学情数据画像标签体系,有利于教育管理者宏观掌握学生成绩、图书借阅、生活消费、特征关联趋势等学情状态,为教育管理者提供决策依据及个性化学习路径和培养方案的科学数据支撑;有助于教师及时洞察学生的长短板,因材施教,开展千人千面的个性化教育;有益于学生,精准了解自己学习状态及自身特征在校、院、系、班中的定位,见贤思齐,取长补短,针对性的提升和进步。

结束语和展望

本文基于锦城智慧校园共享数据包括成绩管理、图书借阅、一卡通消费、学生基本信息等数据,以全校 2018 级学生 2018-2019 学年的数据为基础构建学生画像,形成标签库,并借助锦城大数据实验室平台进行周期性特征加工及可视化分析结果呈现,用数据来刻画学生肖像辅助教育管理者了解掌握学生的状态、定位、学习习惯等。后期数据挖掘的维度会考虑加入门禁系统、学习通线上学情数据、第四课堂综合素质等数据,并对现有维度进一步下钻挖掘,丰富画像指标,归纳学习路径,并结合深度学习推荐系统^[3]等算法训练数据,从而在充分学习个体学习者的差异后,精准推荐适合的学习方案,从而实现智慧教育。

【参考文献】

[1] 贾积有. 人工智能与教育. 职业与教育学术年会. 2018. 12. 04

牟智佳. 学习者数据肖像支撑下的个性化学习路径破解_学习计算的价值赋予[J]. Journal Of Distance Education, 2016, (6):11-19.

[2] 梁婷婷, 李丽琴. 基于深度学习的资源个性化推荐算法及模型设计[J]. 智能计算机与应用, 2018, (6):108-110