基于大数据的学生问题干预算法研究与应用

李光荣

南宁职业技术学院

DOI: 10.12238/mef.v4i7.3895

[摘要] 随着大学校园信息化程度的不断提升,校园网络产生和存储了数量庞大的与大学生活动相关的数据。通过采集学生在校期间的学习、生活等多渠道数据,利用大数据技术,采用问题干预算法,及时精准的预判和发现学生可能出现的生活、学习、心理、思想等方面的问题,从而对症下药,有助于提升高校学生管理的质量和效率。

[关键词] 大数据; 学生问题; 干预算法

中图分类号: G717

文献标识码: A

Research and Application of Student Problem Intervention Algorithm Based on Big Data

LI Guangrong

Nanning College for Vocational Technology

[Abstract] With the continuous improvement of university campus informatization, the campus network has generated and stored a huge amount of data related to the activities of university students. By collecting data from multiple channels of students' study and life at school, assiting big data technology, using problem intervention algorithms, timely and accurate prejudice and discover the problems that students may have in life, learning, psychology, and thoughts to improve the quality and efficiency of college student management.

[Key words] big data; student problem; intervention algorithm

近年来,基于互联网的获取数据方式得到了普及,数据的价值也越来越受到重视。基于大数据技术的发展,高校“智慧校园”建设需要充分利用大数据等技术手段,发挥网络与大数据的高效、便捷优势,注重“用数据说话、用数据决策、用数据管理、用数据创新”,并以此为手段提供一种环境全面感知,数据化、网络化、协作型一体化和个性化的教学管理和生活服务。在高校各类工作中,学生管理服务工作中头绪多、难度大,传统的教育管理形式效率低。管理服务工作的质量不仅影响学校人才培养计划的实施,也影响学生的身心健康以及学校的社会声誉和形象,可见学生管理工作的重要性不言而喻。将大数据分析技术应用到高校学生管理工作当中,通过信息化平台和各类信息终端

感知、采集、提取、处理、整合学生在校期间学习、生活、社交的各类原始数据,对学生综合状况进行数学建模,选取设定在各个维度的状态参数,设置阈值,提供预警机制,以期推动精准化指导和个性化服务为特色的学生管理工作。

1 研究要素

利用大数据对学生工作进行分析必须厘清几个要素。首先是数据资源,我们需要大量的数据资源,才能为后期的统计、判断和定性评价提供数据支撑。要对学生在校期间的学习、生活等多渠道数据进行采集,这些数据包括静态数据和动态数据,结构化数据和非结构化数据。其次是对数据进行处理,提升数据质量,规范数据形式,挖掘数据中的价值。数据清洗是提升数据质量的重要

途径,按照数据共享平台的统一数据标准与数据定义范式将不同类型和格式的数据进行转换。为不同的数据源设置不同的数据清洗方式,利用条件校验、过滤、去除重复记录,对空值进行处理,排除无效数据。最后是算法设计,根据所采集的样本数据通过对以往出现的问题和积累因素之间的潜在关系进行数学建模,确立算法,对学生的个体情况进行预判和评估,从而为学生管理工作的开展提供行动指南和重要的参考。通过学生在校状态大数据分析,能比较客观、全面和动态的反映学生群体状态和共性;用大数据分析得到的“学生画像”,能较准确的察觉个体的特性和存在的不足,为个体提供个性化、定制式的精准管理服务,从而提升了管理服务育人的质量和效率。

2 数据源的选型

我们把目前某高校数字校园网的各种动态数据流及静态数据库进行采集,根据其数据类型的差异对数据进行归类。数据可以分为基础类、学习类、生活类三个大类。基础类数据主要是描述学生个体的基础属性和基本情况。例如:学号、姓名、性别、班级、出生年月、家庭情况等。学习类数据主要记录学生的学习情况的相关数据。例如:学分、课时、出勤率、成绩等。生活类数据主要记录学生在校期间的生活情况。例如:消费数据、勤工助学数据、宿舍进出数据、网络社交数据、心理健康数据等。在大数据分析中数据的质量直接影响到分析结果的准确性,因此,在获取到以上三大类数据后对相关数据进行数据清洗是十分有必要的。数据清洗是指通过运用分箱法、聚类法、回归法等方法对数据中冗余、重复、不一致数据,无效值和缺失值数据进行标准化处理的过程。对无效值和缺失值数据一般处理方法是忽略或手工填入,做估算填充、样例删除、变量删除或成对删除处理;对于重复数据,一般通过对比记录关键属性值或聚类计算距离,对相等、相似的记录进行合并清除;对存在语义冲突的不一致数据,可定义完整性约束,进行数据关联分析,发现更正并保持数据一致。

3 学生问题干预算法

在处理过的格式化数据源的基础上,我们需要对所采集的一系列数据开展分析。通过对所采集的数据种类对比分析发现,有一部分数据类别是较为普遍的数据类型,我们称之为一般类数据,例如:学生上网时长、学生出勤情况等这些行为对学生不构成重大影响。还有一部分数据类别是零容忍出现的数据类型,我们称之为关键类数据,例如:学生心理问题等这些异常需要及时干预较为重要。由于学生的预警与否除了与当前的数据相关之外,还与该学生过去的

数据有关,因此这是一个序列预测的问题,考虑利用深度学习中长短期记忆神经网络(LSTM)模型对学生数据进行拟合、预测。

由于关键数据是零容忍数据,因此一旦任何一个关键类数据超出预警值,则表示我们需要对该学生进行适当的干预,否则还需要结合一般类数据进行综合判断。具体算法设计如下:

将清洗后的关键类数据进行处理,如果任何一个关键类数据大于预警值,则表示需要对该学生进行适当的干预,算法结束。

即: $\exists x^{(k)} \geq I$, 则干预,其中 $x^{(k)}$ 为关键类的数据, I 为预警值。

若所有关键类数据都小于预警值,则将关键类数据和一般类数据综合计算来判断。具体算法流程如下:

(1)数据归一化将所有数据进行归一化处理至 $[0, 1]$ 区间,以降低数据分布变化差异太大带来的影响。

$$x^{(k)} = \frac{x^{(k)}}{\max(x) - \min(x)}$$

公式中, $x^{(k)}$ 为每一个学生数据, $\max(x)$ 为所有数据中最大的数据, $\min(x)$ 为所有数据中最小的数据。

(2)数据向量化。将归一化后的学生一般和关键类数据及标签进行向量化处理,形成1个n维数据向量和1个n维的标签向量,同时对标签向量进行One-Hot编码。

向量化后的n维标签数据:

$$x^{(n)} = (x^{(1)}, \dots, x^{(k)}, x^{(k)}), x^{(k)}$$

为每一个学生的数据

向量化后的n维标签数据:

$y^{(n)} = (0, 0, 0, \dots, 1)$, 向量中0代表不需要预警干预, 1代表需要预警干预。

(3)数据拆分。将向量化的数据和标签进行拆分,拆分成训练集和测试集,其中训练集用来模型训练,测试集用来模型验证。拆分比例可以根据实际数据量的大小进行设定和调整。

(4)模型搭建。搭建LSTM模型,其

中,隐藏层有若干个神经元,输出层2个神经元(需要预警和不需要预警),输入变量是向量化的训练数据,损失函数采用平均绝对误差(Mean Absolute Error, MAE),优化算法采用自适应矩估计(adaptive moment estimation, Adam),模型迭代(epochs)若干次。

针对学生的每一项数据,存在一个模型状态值和上一个单元的输出值,每个输出状态值可以用公式表示:

$$h_{t+1} = f(x^{(n)}, c_t, \hat{h}_t)$$

其中, f 为LSTM模型。

为了使得输出状态非线性化,使用 $\tanh(h)$ 函数激活输出值,其中 $\tanh(h)$ 函数为:

$$\tanh(h) = \frac{e^{-h} - e^h}{e^{-h} + e^h}$$

其中, h 为状态输出值,即学生需要干预或不需要干预的One-Hot状态。

(5)模型训练。将训练集数据和标签输入模型中训练,训练过程中,记录训练集和测试集的损失,并在完成训练后绘制损失图,观察模型的状态。

(6)准确率计算。利用模型预测的结果计算准确率。将测试集加载进模型,利用训练好的模型进行预测,将预测结果与测试中的结果进行比对,计算预测结果的准确率。计算方式如下:

$$\text{准确率} = \frac{\text{预测结果正确数}}{\text{测试数据总数}} \times 100\%$$

(7)模型调优。根据计算出来的正确率调整步骤4中神经网络模型的神经元个数、模型迭代次数等参数,同时可以考虑更换损失函数、优化器算法等,以调整LSTM模型内部神经网络的权重系数,提高模型的正确率。

(8)数据预测。模型训练好后,将学生的数据向量化后输入模型,根据预测出来的结果状态值,判断是否需要对学生进行干预处理。

此外,可以利用预测数据与真实数据来反馈给模型,让模型自适应调整内

部参数,以提高模型的准确率。

在实际的学生管理工作中面临的问题多种多样,这就需要发挥能动性改进学校学生管理大数据的个性化算法。我们还可以根据某项数据特征的重要性进行权重加大处理,如果学生管理部门认为某项数据特征较为重要就可以根据其重要性加大该项目的权重。同理,也可以灵活的对特征值进行单项综合分析,例如对贫困生的判定可以辅助筛选真正贫困的学生进行帮扶,对学生就业能力的判定可以针对学生进行个性化的就业指导等。

4 结语

本文旨在通过数据分析对症下药为学生提供个性化管理服务,起到抛砖引玉的作用。学生管理工作纷繁复杂面临的问题也较为繁杂,探索大数据技术在高校学生管理工作中的应用,有助于辅助高校学生管理工作改革提高效率,促进高校管理服务水平

的有效提升。大数据技术在学生管理、学生服务工作中的应用能够改变传统的感性判断、人为决策、被动统计的局面。总之,在校生管理服务大数据一体化平台,为高校治理现代化,尤其是学生管理服务的精准化能够提供有力的决策依据。

基金项目:

中国高校产学研创新本基金——新一代信息技术创新项目《基于大数据分析的学生岗位素质能力模型平台设计》(编号2019ITA01034);广西职业教育第一批专业发展研究基地项目(桂教职教成[2018]37号);广西职业教育新一代信息技术专业群建设研究基地项目。

参考文献

- [1]刘珊珊,尹光华.大数据时代高校学生管理工作信息化建设现状与对策探讨[J].才智,2018(32):120.
- [2]唐婷婷.大数据时代高校学生教育管理工作的个性化研究[J].高教学刊,2021

(07):152-155.

[3]唐丽萍.大数据时代下高校教育管理信息化创新发展路径[J].文化创新比较研究,2018,2(05):121-122.

[4]徐悦.大数据时代高校学生管理工作信息化建设途径探究[J].兰州教育学院学报,2018,34(01):101-102+105.

[5]胥文勋.大数据时代高校学生管理工作信息化建设现状与对策[J].绵阳师范学院学报,2016,35(01):53-56.

[6]李强,王尊博.大数据时代高校学生管理工作的挑战与对策分析[J].江苏科技信息,2016(33):20-21.

[7]徐擎.韩国高校学生工作信息化的借鉴与思考[J].青岛职业技术学院学报,2017,30(02):72-75.

作者简介:

李光荣(1971--),男,汉族,湖南常宁人,本科,高级工程师,研究方向:信息技术与高职教学。