计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(1):167−176 [doi: 10.15888/j.cnki.csa.009310]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

大数据环境下多模态融合的大学生异常行为预警①

王玉标[1,2], 陶八梅[3,4], 李　珩[5], 陶志红[6]

[1](重庆大学 大数据与软件学院, 重庆 401331)
[2](重庆大学 虎溪网络信息中心, 重庆 401331)
[3](重庆大学 管理科学与房地产学院, 重庆 400044)
[4](中国船舶集团海装风电股份有限公司, 重庆 401123)
[5](重庆大学 教师教学发展中心, 重庆 400044)
[6](艾溪湖中学, 南昌 330012)
通信作者: 王玉标, E-mail: wangyubiao@cqu.edu.cn

摘　要: 针对“校园大数据”累积的海量数据呈现出离散性、稀疏性等问题, 如何从基数大、活动广、个性强的校园学生群体中检测出潜在的、有异常行为的学生, 已成为学生异常行为分析亟需解决的问题. 本文提出了一种大数据环境下基于多模态融合的大学生异常行为预警方法 (early warning method for abnormal behavior of college students based on multi-modal fusion in big data environment, EWMAB). 首先, 针对学生行为画像的表征不够丰富, 行为标签存在时效性、动态性等问题, 建立一种基于多模态特征深度学习的跨模态学生行为画像模型; 其次, 针对学生异常行为预测、预警的时效性和后置性问题, 在学生行为画像和学生行为分类预测基础上, 提出了一种基于多模态融合的学生异常行为预警方法, 通过长短期记忆神经网络 (long and short term memory networks, LSTM), 结合学生行为多指标数据和文本信息来解决学生异常行为预警问题; 最后, 本文通过应用实例验证模型以学生学习成绩异常预警为例, 与其他预警算法相比, EWMAB 方法可以提高预警的准确性, 实现学生异常行为预警的时效性和前置性, 从而使学生教育工作更具有针对性、个性化和预测性.

关键词: 教育大数据; 学生行为画像; 多模态融合; 异常行为预警; 分类预测

引用格式: 王玉标,陶八梅,李珩,陶志红.大数据环境下多模态融合的大学生异常行为预警.计算机系统应用,2024,33(1):167–176. http://www.c-s-a.org.cn/1003-3254/9310.html

Early Warning of Abnormal Behavior of College Students Based on Multi-modal Fusion in Big Data Environment

WANG Yu-Biao[1,2], TAO Ba-Mei[3,4], LI Heng[5], TAO Zhi-Hong[6]

[1](School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China)
[2](Huxi Campus of Network Information Center, Chongqing University, Chongqing 401331, China)
[3](School of Management Science and Real Estate, Chongqing University, Chongqing 400044, China)
[4](CSIC Haizhuang Windpower Co. Ltd., Chongqing 401123, China)
[5](Center for Enhancement of Teaching and Learning, Chongqing University, Chongqing 400044, China)
[6](Aixihu Middle School, Nanchang 330012, China)

Abstract: In view of problems such as discreteness and sparsity in the massive data accumulated by “campus big data”, how to detect potential students with abnormal behavior from the campus student groups with a large base, wide activity ranges, and strong personality has become an urgent issue to be solved in the analysis of abnormal behavior of students. This study proposes an early warning method for abnormal behavior of college students based on multi-modal fusion in big data environment (EWMAB). First of all, in view of the insufficient representation of student behavior portraits and

① 基金项目: 重庆市社会科学规划项目 (2021NDYB110); 重庆市科委自然科学基金面上项目 (cstc2021jcyj-msxmX0515)
收稿时间: 2023-04-27; 修改时间: 2023-05-29, 2023-06-14; 采用时间: 2023-07-03; csa 在线出版时间: 2023-11-24
CNKI 网络首发时间: 2023-11-27

the timeliness and dynamics of behavior labels, a cross-modal student behavior portrait model based on multi-modal feature deep learning is established; secondly, for the timeliness and post-alarm of the prediction and early warning of abnormal behavior of students, a multi-modal fusion-based early warning method for student abnormal behaviors is proposed based on the student behavior portrait and student behavior classification prediction. Through the long and short term memory network (LSTM), combined with student behavior multi-index data and text information, the problem of early warning of students' abnormal behaviors is solved; finally, this study uses an example to verify the model and takes the early warning of abnormal academic performance of students as an example. Compared with other early warning algorithms, the EWMAB method can improve the accuracy of early warning and realize the timeliness and pre-alarm of abnormal behaviors of students so that the education of students is more targeted, personalized, and predictable.

Key words: education big data; student's behavior portrait; multi-modal fusion; abnormal behavior early warning; classification prediction

近年来, 以大数据、云计算和人工智能等为代表的新兴信息技术进一步深化了教育改革. 许多学者认为"大数据具有改变人类教育和学习方式的能力, 是推动教育创新发展的重要力量". 2015 年, 国务院印发《促进大数据发展行动纲要》, 为我国大数据产业发展确立了明确的实施路径, 进一步深化大数据在各行业的创新应用, 促进教育大数据在高校的应用. 在大数据环境下研究学生学业、消费、心理等规律, 是大数据技术在教育信息化中的一个研究方向[1].

校园大数据具有复杂、繁多的类型, 是学生行为数据所呈现的明显特点, 各类数据间的关联很难构建. 现有的学生行为画像方法很难精确地刻画学生行为画像, 对于来自不同的模态的数据没有充分利用, 也不能很好地解决学生异常行为分析与预测、预警中普遍存在的数据稀疏、时效性、后置性和准确性等问题. 因此, 如何对学生行为进行深入挖掘, 融合学生行为的多模态信息, 并对学生行为进行分类, 实现对学生异常行为预警的动态性和前置性转化, 给学生异常行为分析与预测、预警带来了挑战. 挑战 1: 构建多模态融合的学生行为画像. 学生行为画像以学生为主体, 通过对学生学习和生活行为等各类校园行为数据进行挖掘, 赋予学生行为画像标签. 随着社交媒体的广泛发展, 学生的网络行为更为复杂、多样, 现有的学生行为画像建模技术中, 经常仅就单一的模态进行处理, 没有充分融合学生的多模态行为信息, 难以全面地刻画学生行为画像的特征. 挑战 2: 学生异常行为预测、预警时存在时效性和后置性问题, 难以早期发现学生潜在的异常行为. 文本信息是对学生行为分析影响的重要维度, 传统的自然语处理技术难以直接对其进行分析. 此外, 其他的学生日常归寝、消费、上课率、借阅等多个维度都是构成学生行为异常场景的重要因素, 传统学生异常行为预测、预警模型的准确性存在较大误差.

由于学生行为数据呈现出离散性、稀疏性等问题, 如何从基数大、活动广、个性强的校园学生群体中检测出潜在的、有异常行为的学生已成为一个急需解决的重要问题, 本文提出了一种大数据环境下基于多模态融合的大学生异常行为预警方法. 首先, 建立了一种基于多模态特征深度学习的跨模态学生行为画像方法, 通过基于多模态特征深度学习的跨模态融合模型, 构建学生行为画像, 其次, 提出了一种基于多模态融合的学生异常行为预警方法, 通过神经因子分解机和长短期记忆神经网络 (long and short term memory network, LSTM), 结合学生行为多指标数据和文本信息来解决学生异常行为预警问题; 最后, 本文通过应用实例验证模型的准确性, 融合学生归寝、学业成绩、校园网络行为、一卡通消费、图书馆借阅、体育锻炼、校内行为轨迹等 70 个维度, 通过实例验证学生异常行为预警的准确性, 实现学生异常行为预警的时效性和前置性.

1 相关工作

1.1 学生行为画像研究

目前, 校园大数据进一步挖掘分析对当前高校管理及培养模式乃至整个教育体系产生深远影响, 建立学生行为画像成为教育大数据应用的重要方式之一[2]. 用户画像是根据用户的数据维度来构建标签化、精准化、个性化、可视化的用户模型. 学生行为画像的思

路主要源于用户画像的广泛应用, 是建立在学生学习、生活行为等校园大数据的基础上, 以学生为主体, 通过对学生个人特征、学习和生活行为等数据进行挖掘, 从学生特征出发, 融合多模态、多维度的学生行为信息, 赋予学生行为画像标签[3]. 文献[4]挖掘教育数据、学生学习态度和学习行为之间的关系, 建立学生画像. 根据 E-learning 资源和学习者行为, 优化 E-learning 平台和改进学习者学习内容. 文献[5]分析非定向在线行为与学业成绩之间相关性, 提出了一个基于用户画像分析非定向行为与学业成绩之间相关性的模型. 基于多元智能构造以及学习和情感策略的学生画像, 对不同的维度、自我调节技能、学习和情感策略如何影响学生的学业成绩进行定量分析[6].

目前, 主流学生行为画像刻画方法一般是基于机器学习、支持向量机、有监督学习等技术. 这些方法从学生数据中提取特征作为学生行为的表示向量, 并利用学生属性标签的数据作为标记数据来训练学生行为画像的预测模型. 但是, 这些方法在学生数据安全、时效性, 融入图像、文本等学生多维信息等方面考虑的不是很充分. 因此, 如何利用不同来源、不同结构和不同模态的学生行为数据进行建模, 将不同形态的多模态特征有效地融合在一起, 考虑到不同模态之间隐性联系, 提高学生行为画像的精度, 是目前教育大数据应用领域亟需解决的问题.

1.2 学生异常行为研究

文献[7]根据在校学生的实时行为数据, 结合问卷调查和人口统计学等相关数据, 构建“学生画像”, 通过跟踪学生日常学习状况, 可以对学生的期末成绩、就业情况提供预警. 通过 Apriori 算法分析图书借阅与学生成绩之间的相关性, 采用基于 BP 神经网络预测算法, 通过课程表现预测借书情况, 构建学生成绩预警模型[8]. 文献[9]考虑学生背景、学生过去的学业成绩和其他属性, 通过线性回归, 采用一种新的机器学习算法, 用于预测学生的学业表现. 文献[10]提出一种新的 GRMF 方法, GRMF 将基于学生和课程的数据构建的两个边图集成到鲁棒低秩矩阵分解的目标中, 学生和课程的学习特征可以从教育情境中得到更多的先验知识, 从而提高成绩预测准确性. 针对高校精准扶贫问题, 以学生行为数据为基础, 结合高校数据的时序性特点, 提出基于深度学习理论的 CW-LSTM 算法进行预测[11]. 文献[12]使用学生的作业数据通过学生的拖延行为预测学生在 Moodle 中的课程成绩, 但用于学生成绩预测的数据没有综合考虑, 可以使用更多的学生活动数据, 例如学生文本数据, 同时, 现有方法在引入深度学习、融合多模态信息、智能识别网络当中一些未知的异常行为还存在一些不足[13,14].

传统的学生异常行为预警在数据处理过程中, 更多地将重点放在对历史数据的分析和挖掘上面, 缺乏动态性、时效性[15]. 同时, 随着学生网络行为的多样化, 文本信息、图像成为学生行为分析的重要维度, 传统的学生异常行为预测、预警模型没有充分利用这些信息, 存在模糊性、不确定性和误差较大的问题. 因此, 考虑学生实时行为数据、文本信息和学生行为相似度, 提高学生行为预测、预警的时效性、动态性、准确性, 由后置性预警向前置性转变, 对学生的全面发展具有很强的指导作用.

2 大学生异常行为预警模型构建

本文采用一种基于深度学习的数据预处理方法来处理高校各个业务系统的数据, 实时、有效地处理多维数据、图像数据, 学生的行为中存在的文本信息; 其次, 在基于深度学习的数据预处理基础上, 融合学生基本信息和学生行为多维特征, 提出基于多模态特征深度学习的跨模态融合模型 (如图 1 所示), 构建精确的学生行为画像; 再次, 考虑不同学生行为特征指标的权重问题, 提出一种基于多模态融合的学生异常行为预警方法, 并进行实例验证.

2.1 数据预处理

数据预处理是指一系列的数据处理过程, 包括数据采集、数据清洗、行为特征提取、数据转换等[3]. 本文采用基于深度学习的数据预处理. VGG16 用于图像数据预处理和学生异常图像识别. 文本信息的数据预处理与图像处理相同, 需要文本向量化. 文本预处理包括 4 个步骤: 读入文本、通过分词创建字典、将每个单词映射到唯一索引 (index)、将文本从单词序列转换为索引序列, 本文使用 Word2Vec 来处理文本向量化.

学生行为数据预处理如图 2 所示, 本文采用并行处理框架来提高数据预处理效率, 采用基于分布式处理框架 (Hadoop) 的无损清洗算法, 它是通过使用并行编程模型作为 MapReduce 来实现, 是流行的并行计算范式之一, 适用于大数据处理中的各种应用, 更好地利用 MapReduce 对图进行并行处理. 学生数据清洗后,

进行学生数据属性识别、数据分类、数据分析, 同时检测每天学生数据是否存在异常.

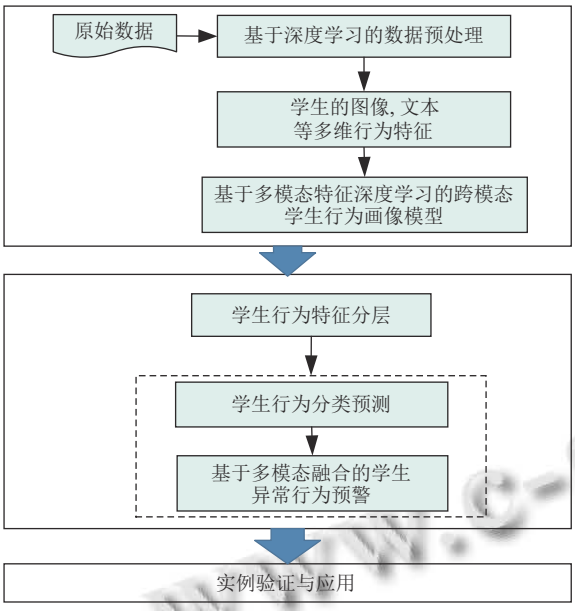

图 1 校园大数据环境下大学生异常行为预警模型

2.2 基于多模态特征深度学习的跨模态学生行为画像模型

本文提出基于多模态特征深度学习的跨模态学生行为画像模型, 在基于深度学习的数据预处理基础上, 提出基于多模态特征深度学习的跨模态融合模型框架, 通过对多元数据获取、文本和图片特征提取及融合方法, 构建精确的学生行为画像. 本项目拟从原始数据出发, 通过对原始数据的统计分析获得客观标签, 客观标签可以数据预处理技术得到.

数据融合是组合来自不同来源数据的过程, 以便使组合后的数据产生比单独使用每个来源得到更多的信息. 在学生行为分析过程中, 充分利用学生文本、图像、语音等多模态数据, 可以更好地分析学生的异常行为. 跨模态融合模型框架如图 3 所示. 多模态数据可以来自不同的学生学习环境, 如学生课堂、学生进出宿舍、学生消费、上网行为、图书借阅等.

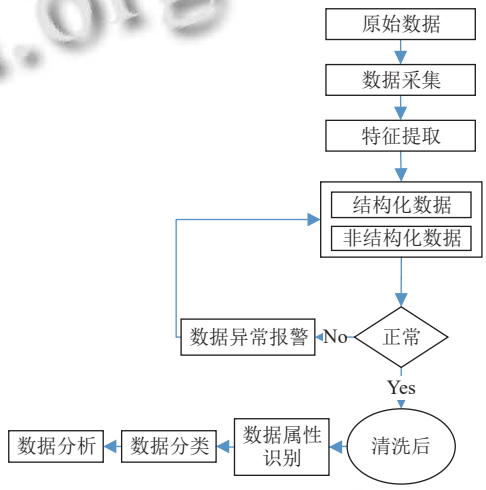

图 2 学生行为数据预处理

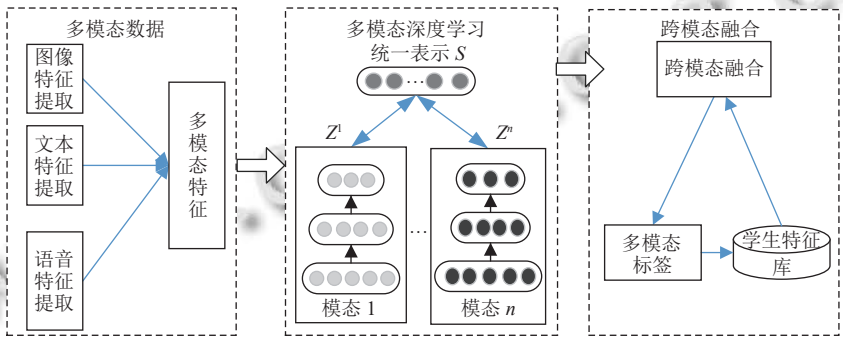

图 3 跨模态融合模型框架

整个模型的学习过程: 首先对各个模态的特征进行深度学习, 得到特征的层次表示. 为了学习输入特征表达, 本文采用堆叠式降噪自编码器来学习输入特征表达; 其次, 每个模态特征由堆叠式去噪自动编码器 Z^i $(i=1,2,\cdots,n)$表示, 对它们之间的关系进行建模, 得到统一表示S; 最后, 将获得的模态表示为Z^i, 作为更高层神经网络的输入, 并对它们之间的关系进行建模和学习. 通过降噪编码器学习到每个模态特征后, 该项目使用受限玻尔兹曼机 (restricted Boltzmann machine, RBM) 来分析多个模态特征之间的关系建模, 以获得联合共享表示层. 同时, 使用长短期记忆网络 (LSTM) 来分析学生行为文本等相关信息.

LSTM 作为一种优化的循环神经网络, 可以捕获序列之间长期和短期相关的特征, 训练输入的原始特

征, 可以较好地融合特征之间的关系. 它是一种特殊的循环神经网络 (recurrent neural network, RNN). 与传统的简单 RNN 相比, 它在简单 RNN 的基础上增加了相关参数并添加了 LSTM 辅助, 解决了处理长序列数据过程中的梯度问题. LSTM 神经网络在处理和预测具有较长时间序列间隔和延迟的数据时具有良好的性能. 因为学生刷卡进出宿舍的数据、图书馆借阅数据、食堂消费数据都是长时间发生的序列, 对时间的依赖性很强, 所以 LSTM 具有良好的表现, 普通的神经网络在处理这些数据时往往表现不佳.

本文在处理不同维度的学生信息时以时间为基准, 使用 LSTM 网络训练依赖时间长的信息. LSTM 网络在序列建模方面具有一定优势, 具有长期记忆功能, LSTM 神经网络如图 4 所示.

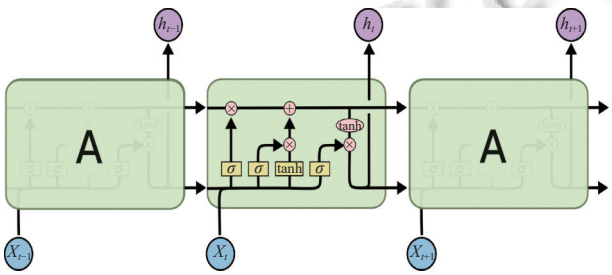

图 4 LSTM 神经网络

2.3 学生行为分类预测

学生“画像”特征库中提取的学生行为特征指标及学生个人基本信息, 构建学生行为特征集合, 学生的特征集合定义为式 (1):

$$C_S = \{C_1, C_2, \cdots, C_i, \cdots, C_m\} \tag{1}$$

学生行为特征从不同程度上代表了学生在校行为表现指标, 不同的特征指标对模型的作用程度不同, 为了建立学生行为特征分层模型, 采用组合权重方法计算学生行为特征指标, 从而得到更为符合实际需求的权重赋值. W_i 表示采用粗糙集属性权重方法计算的客观权重, W_i' 表示采用 AHP 方法计算的主观权重, W_i^* 为综合权重系数, m 为学生特征的个数. 最终建立学生行为特征分层模型如式 (3) 所示, 其中, S_i 表示考虑特征权重后的特征值, $\sum_{i=1}^{m} W_i = \sum_{i=1}^{m} W_i' = 1,\ 0 \leqslant W_i, W_i' \leqslant 1\ (i = 1, 2, \cdots, m)$.

$$W_i^* = \mu W_i + (1-\mu) W_i' \ (0 < \mu < 1) \tag{2}$$

$$S_i = W_i^* C_i \tag{3}$$

本文在利用上述学生行为特征层次模型后, 对 K 近邻非参数回归预测方法进行改进, 提出一种基于 K 近邻非参数回归改进的学生行为分类预测方法 建议的. 首先, 根据学生行为特征的分层模型, 计算目标学生与高年级、同年级的相似度, 在计算两个学生的行为相似度之前, 需要对数据进行标准化处理, 并采用基于深度学习的数据预处理, 具体过程如本文第 2.1 节所述, 以提高分类预测的准确性; 其次, 根据学生行为的相似度, 找到与待预测目标学生 x 最接近或最相似的 K 个高年级学生, 以及 K 个同年级学生; 最后, 本文采用基于云模型的学生行为相似度度量, 然后通过组合权重法计算学生行为特征指标的综合权重[16], 预测学生的行为.

2.4 基于多模态融合的学生异常行为预警

多模态融合是指利用相关手段, 融合分析获得的所有信息, 对信息进行统一评价, 最终得到统一信息的技术. 目前的学生行为预测、预警时考虑学生的文本信息不足, 随着新媒体的快速发展, 学生的网络行为越来越多, 如经常在网络中进行留言、评论, 这涉及学生的文本信息分析. 通常采用基于潜在狄利克雷主题分布获取评论的主题因素无法完全反映与学生行为相关的特征, 并且所使用的词袋语言通常会忽略单词的顺序. 针对这些问题, 提出一种基于多模态融合的学生异常行为预警方法, 通过 LSTM, 通过学生行为分类预测学生行为多指标数据, 结合 LSTM 文本信息来解决学生异常行为预警问题, 进一步改善预测性能.

本文采用的网络结构可以整合多模态数据, 自动提取行为特征. 结合学生行为画像的静态特征, 利用 LSTM 网络对学生涉及的文本行为进行建模, 从而对学生学业异常进行预警, 该模型在真实的数据集上进行了测试. 基于多模态融合的学生异常行为预警如图 5 所示, 由于无法直接使用学生的文本信息, 因此将神经分解机和长短期记忆网络分别用于生成行为预测值和评论生成. 即模型包含 3 个组件: 神经多指标回归预测组件、相关行为评论文本生成组件和学生行为相似度分解组件. 在神经多指标回归预测组件中, $S^*_{U_a,CS_i}$ 表示由神经因子分解机处理后的预测值, 为提高学生行为预测准确性, 首先对＜学生, 行为场景, 指标数据＞进行一位热键编码来构建输入特征向量, $\sum_{i,j}(v_i \odot v_j)x_i x_j$ 表示 Bi_Interaction 层, $\odot$ 表示向量 v_i 和 v_j 的元素积, 然

后使用神经分解回归模型以非线性转换的方式将输入特征向量映射到学生的预测总分. 通过 LSTM 将学生、行为场景和多个指标数据的组合表示转换为一系列表示评论文本的单词$\{W_{U_a,CS_i,0},W_{U_a,CS_i,1},\cdots,W_{U_a,CS_i,n},Z_m\}$, 其中, m表示隐藏层数量, Z_m表示隐藏层, 通过共享神经因子分解机的隐藏层连接预测组件和评论生成组件.

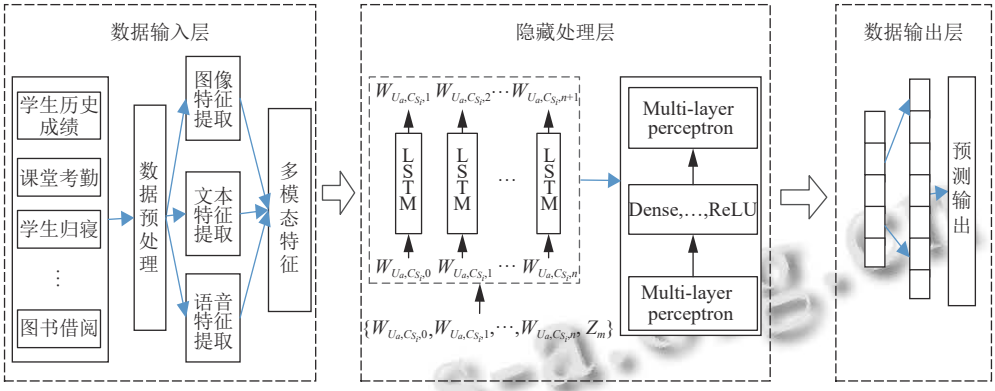

图 5 基于多模态融合的学生异常行为预警

本文将文本特征信息作为输入集成到 EWMAB 模型. 然而, 由于在预测阶段无法直接获得学生行为的文本信息, 并且文本和学生行为多指标数据是两个完全不同的异构数据, 因此不能简单地将它们融合在一起. 传统的基于文本的评论分析和预测一般都是基于词袋模型, 忽略了评论文本中的词的顺序. 为了捕捉词组的序列信息, 本文采用长短期记忆网络 LSTM 进行建模. 假设 V 代表文本中的词汇数, 学生U_a关于场景的文本信息为$W_{U_a,i}=\{W_{U_a,i,0},W_{U_a,i,k},\cdots,W_{U_a,i,n_{U_a,j-1}}\}$, 其中$n_{U_a,i}$是评论的长度, 即评论中的 k+1 个单词, 那么 LSTM 生成的评论内容为顺序的方式. 其中, $1\leqslant k\leqslant n_{U_a,i}-1$, $W_{U_a,i,k}$代表注释中的前 k 个单词, 即第 k 步 (即第 k 个单词) 的状态, ϕ是所有参数集合, $LSTM(\cdot)$是标准 LSTM 单元, $Softmax(\cdot)$ 是 $Softmax$ 激活函数, 隐藏状态转换的是 V 词的生成概率分布.

$$s_k=(LSTM(s_{k-1},W_{U_a,i,k},\phi) \tag{4}$$

$$P(W_{U_a,i,k}|W_{U_a,i<k},\phi)=Softmax_{W_{U_a,i,k}}(s_k,\phi) \tag{5}$$

3 大学生异常行为分析实例

G 大学虎溪校区 15 720 名在校本科生的各种行为数据, 包括学生归寝、学业成绩、校园网络行为、一卡通消费、图书馆借阅、体育锻炼、校内行为轨迹等 70 个维度数据进行行为分析, 形成不同的学生行为画像. 以 2021–2022 年第 1 学期结束时的学生数据作为初始数据, 提取的维度数据有学生基础数据 (包括学号、姓名、学院、GPA、年级、学籍状态、学生类别、学生统一认证号、专业等)、学生归寝指标 (早出、晚归、夜不归寝、24 h 无活动记录等)、食堂就餐消费 (消费时间、消费金额、消费总额、周消费金额、次数、平均消费金额、消费区域、消费明细等)、图书馆借阅数据 (图书馆借阅数、图书馆进出数、借阅时长、借书时间、还书时间、借阅明细、进入时间分布曲线等)、学生上网行为信息 (最长在线时长、游戏时长、平均每天的在线时长、经常在线的时间段、同届学生平均上网时间)、学生课堂考勤数据 (每科到课率、学期到课率汇总、每月到课率的波动曲线等) 等.

3.1 学生行为画像构建

本文通过融合校内各个应用系统的数据, 进行校园大数据分析和挖掘, 分析学生的日常学习和生活行为, 构建精确的“学生行为画像”, 学生行为画像如图 6 所示, 是某位学生在 2022 年春季学期的当前画像, 学生行为画像利用学生的静态属性数据和各类动态行为数据聚类结果的构造学生行为标签, 包括客观标签和主观标签. 本文对现在的各个应用系统数据进行分析与数据预处理 (包括学生基础数据、学生刷脸数据、学生体育锻炼、一卡通消费、图书馆借阅数据、学生上网行为信息、学生课堂考勤数据和学生住宿等数据), 对进行相关进行分析和聚类, 聚类方法采用一种改进的凝聚层次聚类算法进行聚类, 从而学生行为画像, 分析每个学生的特性, 从而有利于该学生异常行为分析.

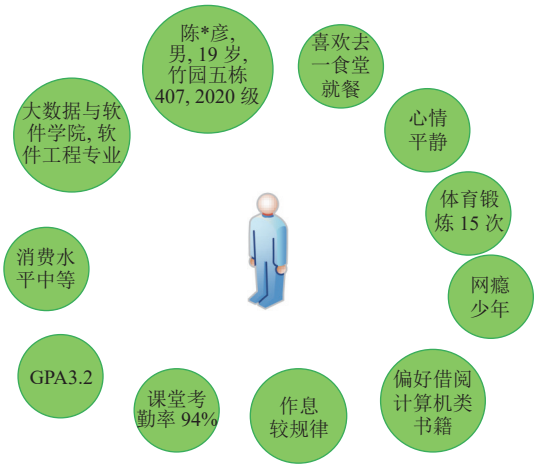

图 6 学生行为画像

如以研究一个 GPA 比较高的学生行为画像为例, 从 2021–2022 年第 1 学期结束时的学生数据数据集中筛选出该学生共计 11 230 条学生数据作为初始数据, 提取的维度数据有学生基础数据、学生归寝指标、食堂就餐消费、图书馆借阅数据、学生上网行为信息、学生课堂考勤数据等. 对该学生进行日、周、月的行为预测, 用数据相关性分析的方法发现其行为规律. 基于学生行为画像的时空行为分析与预测如图 7 所示, 是根据前面学生在不同时间段的行为统计分析, 分析他一天的主要行为路径, 路径 1、4、7 表示学生去食堂就餐, 2 表示去教室, 3 表示学生去图书馆借阅, 5、9 表示学生回寝室休息, 6 表示学生去实验室学习, 8 表示学生体育锻炼. 可以看出该学生在 0:00–7:00 之间发生的行为主要是休息行为, 8:30–12:30 之间发生的行为主要是教室上课、图书馆借阅、自习和饮食行为, 12:30–14:00 主要是中午寝室休息, 14:00–19:00 主要为上课行为、实验室学习、自习行为、饮食行为、体育锻炼等, 在统计时如有较少偏离均值的行为出现, 整体对均值不会产生影响, 从而可以看出学生的日常习惯行为较为规律, 因此, 可以根据这些日常的行为, 对接教务系统的学生课表、学校监控系统、门禁系统等, 预测学生下一天的行为, 如果出现较少的偏差, 如就餐时间, 可以不用处理, 但是如果学生根据课表是在上课时间, 从而出现在校内其他地方, 则可以预警, 相关人员核实学生是否存在逃课行为; 如学生晚上 11:30 后还未回寝室, 监控显示在湖边徘徊或是出现在禁止出入的地方, 则需要让管理人员联系该学生行为是否异常等. 根据学生行为画像, 得到个性化行为习惯与学业成长相关, 为大学生个性化培养提供支撑.

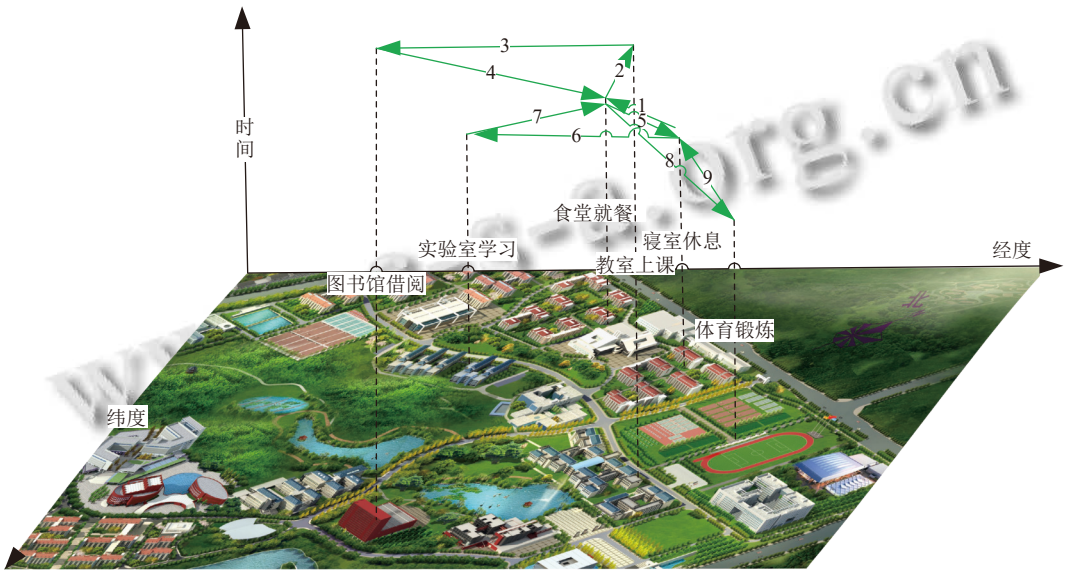

图 7 基于学生行为画像的时空行为分析与预测

3.2 学生行为关系分析

学生安全工作一直是高校学生管理的重要工作. G 大学建设了学生宿舍人脸大数据分析系统, 通过对 2019 年 3 月至 2022 年 3 月间通过对校区 4 500 多万条学生进出宿舍刷脸数据进行分析, 结合学生住宿管理规定, 自动生成晚归、夜不归宿、24 h 无活动记录等数据, 并提供了实时查寝功能, 为在特殊时间节点统计学生在宿情况提供了有效的技术手段, 并将这些信

息自动推送给学院辅导员, 让辅导员教育工作有针对性, 降低学生安全事故的发生, 学生归寝异常分析如图 8 所示.

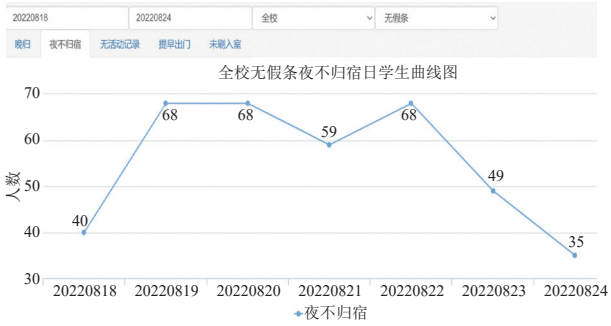

图 8 学生归寝异常分析

3.3 学生行为异常预警

学生上课率、图书借阅、学生归寝、食堂规律就餐、体育锻炼、学生上网等多个维度都是构成学生成绩影响的重要因素. 本文在针对课堂考勤指标、上网时长、学生归寝指标、图书借阅指标、食堂就餐指标、体育锻炼指标等 6 个学习勤奋度指标进行分析, 通过基于 K 近邻非参数回归改进的学生行为分类预测学习勤奋度指标预测值, 最终融合学生行为多指标和文本信息的异常行为预警. 实验以学习行为特征为基础, 学生历史成绩为 2021–2022 年秋季学期之前 GPA 成绩作为初始数据, 能够较好地反映学生的学习基础. 通过学习勤奋度指标, 采用 EWMAB 方法构建学生学业预测模型, 如图 9 所示, 并基于该预测结果对学生行为进行分等级预警.

为了验证本文提出的 EWMAB 方法与其他方法的预测准确性, 本文采用准确率 (*accuracy*)、精确率 (*precision*)、召回率 (*recall*) 和 *F*1-score 这 4 个指标作为评估学生成绩分类预测性能度量. *F*1-score 是 *precision* 和 *recall* 的调和平均数. 正确分类的样本数除以所有的样本数, 通常来说, 准确率越高, 分类器越好. *accuracy* 计算方法如式 (6) 所示, *precision* 计算方法如式 (7) 所示, *recall* 计算方法如式 (8) 所示, *F*1-score 计算方法如式 (9) 所示. 表 1 为混淆矩阵.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$precision = \frac{TP}{TP+FP} \tag{7}$$

$$recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1\text{-score} = 2\times\frac{precision\times recall}{precision+recall} \tag{9}$$

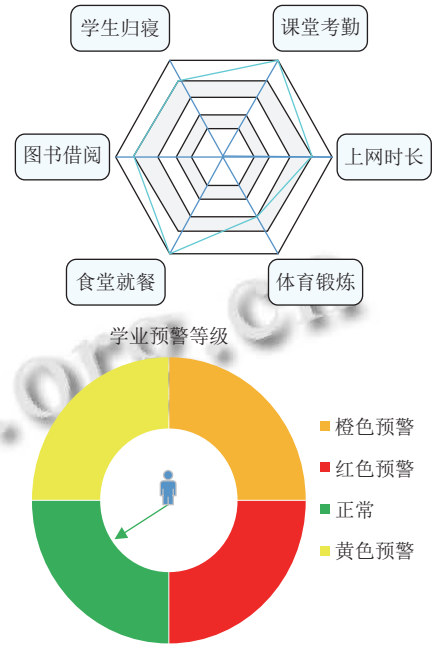

图 9 学生学业异常预警

表 1 混淆矩阵

真实情况	预测情况	
	正例	反例
正例	*TP* (真正例)	*FN* (假反例)
反例	*FP* (假正例)	*TP* (真正例)

在本文中, 为了验证本文的 EWMAB 的准确性, 与其他 5 种方法进行了对比分析, 文献[17]通过构建课程数据的分类器, 分析 Moodle 的数据来预测学生的学习成绩; naive Bayes model 是一种基于贝叶斯定理和特征条件独立假设的分类方法; decision tree (DT), 是一个用于学习成绩预测的决策树; random forest (RF) 是随机森林用于学习成绩预测; support vector regression (SVR), 即找到一个回归平面, 使一个集合的所有数据都最接近该平面的分类方法.

实验以学习行为特征为基础, 15 720 名本科生在 2021–2022 年第 1 学期结束时的 GPA 成绩作为初始数据, 能够较好地反映学生的学习基础. 通过学习勤奋度指标构建学生成绩预测模型, 预测学生的 2021–2022 年第 2 学期 GPA 成绩, 采用 *accuracy*、*precision*、*recall* 和 *F*1-score 这 4 个评估指标来评判成绩预测模型.

实验融合了 6 个学习勤奋度指标进行分析, 对不同预测方法进行 *accuracy*、*precision*、*recall* 和 *F*1-score 评估, 如表 2 所示, 可以看出 6 种不同的成绩预

测方法的全量组合实验, 6 种预测模型指标都表现出较为一致的结果, 从 *accuracy*、*precision* 和 *F*1-score 评估指标中可以看出, EWMAB 最优, SVR 指标性能最差. 本文的 EWMAB 方法考虑了学生的多模态信息, 构建精确的学生行为画像, 在一种基于 K 近邻非参数回归改进的学生行为分类预测基础上, 采用基于 MapReduce 框架的 C4.5 改进算法构建学生行为预警决策树, 分类准确性更高, 在学生成绩异常预测时的 *accuracy* 和 *precision* 较高. 文献[17]是一种改进的分类方法, 它考虑了学生整个课程期间表现在课程成绩预测中的影响, 增加了学生拖延倾向的特征向量, 但是在考虑学生的文本、图像等多模态信息等方面不是很充分, 没有考虑学生非在线学习的其他行为. DT 方法是一种监督学习的分类方法, 能够同时处理数据型和常规型属性, 但是对于连续性的字段等方面比较难预测. Naive Bayes model 和决策树相比, 朴素贝叶斯分类器虽然有较好的数学基础和稳定的分类效果, 但是模型所需估计的参数很少, 对缺失数据不太敏感, 实际分类效果一般, 由于本文数据量较大, 这也是 naive Bayes model 算法分类效果较差的原因之一, 准确率低于决策树方法. RF 是集成学习的一种组合分类算法, 是一种从原数据集中有放回的抽样方法, 分类效果一般. SVR 根据学习勤奋度指标、生活习惯对成绩进行多分类预测时, 可以看出 SVR 的分类效果最差, 而 EWMAB 和文献[17]的总体准确率相对稳定, *accuracy*、*precision*、*recall* 和 *F*1-score 指标较高. 因此, 本文利用学生行为相关数据进行学生成绩预测是可行的.

表 2 不同预测方法在全量组合指标的性能对比

Method	*accuracy*	*precision*	*recall*	*F*1-score
EWMAB	0.8291	0.8495	0.9110	0.8792
文献[17]	0.8011	0.8390	0.9010	0.8689
naive Bayes model	0.7822	0.8032	0.8468	0.8239
DT	0.7928	0.8129	0.8560	0.8336
RT	0.7638	0.7949	0.8375	0.8143
SVR	0.7559	0.7801	0.8178	0.7944

如表 2 所示, EWMAB 的指标性能远大于 naive Bayes model、DT、RF 和 SVR, 略大于文献[17], DT 的 *accuracy*、*precision* 和 *F*1-score 指标略大于 naive Bayes model、RF 和 SVR. 综合来言, 模型性能排序: EWMAB >文献[17]>DT>naive Bayes model>RF>SVR.

4 结论与展望

本文建立了一种基于多模态特征深度学习的跨模态学生行为画像方法, 通过基于多模态特征深度学习的跨模态融合模型, 构建学生行为画像. 重点关注学生异常行为分析与预警, 针对学生行为中存在文本信息的场景, 提出一种基于多模态融合的学生异常行为预警方法, 及时发现有异常行为学生, 有利于高校管理者精确“思政”、个性化教育等, 便于高校管理者及时调整教育模式, 引导学生正确生活和学习, 促进学生身心健康的全面发展, 具有广阔的应用前景. 因此, 校园大数据的大学生异常行为预警体系的构建有利于提高高校教育管理水平, 全面的支撑高校“三全育人”工作的顺利开展. 未来将进一步完善学生隐私数据保护、个性化学习推荐、可视化数据分类展示, 持续探索不同学生异常行为场景预警新方法.

参考文献

1 Jia XG. Research on the role of big data technology in the reform of English teaching in universities. Wireless Communications and Mobile Computing, 2021, 2021: 9510216.

2 Liu ZY, Dong LY, Wu CL. Research on personalized recommendations for students' learning paths based on big data. International Journal of Emerging Technologies in Learning, 2020, 15(8): 40–56. [doi: 10.3991/ijet.v15i08.12245]

3 Li XY, He SW. Research and analysis of student portrait based on campus big data. Proceedings of the 6th IEEE International Conference on Big Data Analytics (ICBDA). Xiamen: IEEE, 2021. 23–27.

4 Liang K, Zhang YY, He YS, *et al*. Online behavior analysis-based student profile for intelligent E-learning. Journal of Electrical and Computer Engineering, 2017, 2017: 9720396.

5 Liang K, Liu JJ, Zhang YY. The effects of non-directional online behavior on students' learning performance: A user profile based analysis method. Future Internet, 2021, 13(8): 199. [doi: 10.3390/fi13080199]

6 Gonzalez-Nucamendi A, Noguez J, Neri L, *et al*. The prediction of academic performance using engineering student's profiles. Computers & Electrical Engineering, 2021, 93: 107288.

7 Mojarad S, Essa A, Mojarad S, *et al*. Data-driven learner profiling based on clustering student behaviors: Learning consistency, pace and effort. International conference on intelligent tutoring systems. Proceedings of the 14th International Conference on Intelligent Tutoring Systems.

Montreal: Springer, 2018. 130–139.

8 Shi CX, Tan Y. A BP neural network-based early warning model for student performance in the context of big data. Journal of Sensors, 2022, 2022: 2958261.

9 Sravani B, Bala MM. Prediction of student performance using linear regression. Proceedings of the 2020 International Conference for Emerging Technology (INCET). Belgaum: IEEE, 2020. 1–5.

10 Zhang YP, Yun Y, Dai H, *et al*. Graphs regularized robust matrix factorization and its application on student grade prediction. Applied Sciences, 2020, 10(5): 1755. [doi: 10.3390/app10051755]

11 聂敏, 张杨, 邓辉, 等. 利用基本信息和行为数据发现高校贫困学生. 电子科技大学学报, 2020, 49(5): 795–800. [doi: 10.12178/1001-0548.2020139]

12 Yang Y, Hooshyar D, Pedaste M, *et al*. Predicting course achievement of university students based on their procrastination behaviour on Moodle. Soft Computing, 2020, 24(24): 18777–18793. [doi: 10.1007/s00500-020-05110-4]

13 Zhang S, Yao LN, Sun A, *et al*. Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 2019, 52(1): 5.

14 Xu X, Wang JZ, Peng H, *et al*. Prediction of academic performance associated with Internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 2019, 98: 166–173. [doi: 10.1016/j.chb.2019.04.015]

15 Nam SJ, Samson P. Integrating students' behavioral signals and academic profiles in early warning system. Proceedings of the 20th International Conference on Artificial Intelligence in Education. Chicago: Springer, 2019. 345–357.

16 Wang YB, Wen JH, Wang XB, *et al*. A cloud service trust evaluation model based on combining weights and gray correlation analysis. Security and Communication Networks, 2019, 2019: 2437062.

17 Quinn RJ, Gray G. Prediction of student academic performance using Moodle data from a further education setting. Irish Journal of Technology Enhanced Learning, 2020, 5(1): 1–19. [doi: 10.22554/ijtel.v5i1.57]

(校对责编: 孙君艳)