计算机科学
COMPUTER SCIENCE

在线课堂学习者互动状态识别方法

饶怡, 袁博川, 袁玉波

引用本文

饶怡, 袁博川, 袁玉波. 在线课堂学习者互动状态识别方法[J]. 计算机科学, 2024, 51(11A): 231200133-9.

RAO Yi, YUAN Bochuan, YUAN Yubo. Recognition Method of Online Classroom Interaction Based on Learner State [J]. Computer Science, 2024, 51(11A): 231200133-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200133

在线课堂学习者互动状态识别方法

饶　怡[1]　袁博川[1]　袁玉波[1,2]
1 华东理工大学信息与科学工程学院　上海 200237
2 上海大数据与互联网受众工程技术研究中心　上海 200072
(ryecust@163.com)

摘　要　随着人工智能在教育领域的广泛应用,在线课堂已成为当今极为便捷高效的新教育模式。然而,如何有效管理学习者的课堂学习状态成为一项重要的教育管理难题。鉴于此,提出一种在线课堂学习者互动状态识别方法。首先,将在线课堂数据源分为视频数据和音频数据,基于视频数据构建了包括学习者上肢姿态特征、表情特征以及面部特征等多维度的互动状态特征,基于音频数据构建了学习者的课堂应答状态特征。其次,通过特征选择算法筛选出的关键特征,构建二分类模型,采用贝叶斯优化实现对学生课堂互动状态的精确识别。最后,设计了一个在线课堂总体情况评估模型,为教师提供全面的课堂评估结果,优化教学策略。在自建的在线课堂视频数据集上,该单名学习者课中互动状态识别算法的准确率能够达到93%以上。

关键词:人工智能;教育管理;面部特征;姿态特征;表情识别;课堂评估
中图分类号　TP399

Recognition Method of Online Classroom Interaction Based on Learner State

RAO Yi[1],YUAN Bochuan[1] and YUAN Yubo[1,2]
1 School of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China
2 Shanghai Engineering Research Center of Big Data & Internet Audience,Shanghai 200072,China

Abstract　With the widespread application of artificial intelligence in the field of education,online classrooms have become a highly convenient and efficient mode of modern education. However,effectively managing the learning status of students in the classroom has become an important challenge in education management. In light of this,method for recognizing learner interaction states in online classrooms is proposed. First,the online classroom data source is divided into video data and audio data. Based on video data,a multidimensional set of interaction state features is constructed,including learner upper body posture features,facial expression features,and facial features. Based on audio data,classroom response state features of the learners are built. Next,using a feature selection algorithm to select key features,a binary classification model is constructed to achieve precise recognition of students' classroom interaction states using Bayesian optimization. Finally,an overall classroom assessment model is designed to provide comprehensive classroom assessment results for teachers and optimize teaching strategies. The accuracy of the learner interaction state recognition algorithm in this single-student classroom exceeds 93%,as validated on a self-constructed online classroom video data set.

Keywords　Artificial intelligence,Education management,Facial feature,Posture feature,Facial expression recognition,Assessment of class

1　引言

近年来,在信息化的大背景下,人工智能与教育行业正在进行深度融合,在线课堂以其灵活便捷的特点逐渐走入人们视野。在线课堂在时间和空间上的灵活性特点能够突破地理限制,为更多的学习者提供学习机会,大大提高了教学资源的利用率;并且,上网课也成为了保障基础教育的必备手段。如今,在线课堂在教学中占据日益重要的地位,智慧课堂也成为教育领域研究的重点[1]。

在传统的面对面教学模式中,老师可以直观地观察学生的学习状态和参与程度,从而做出及时的调整和引导。然而,由于时空分离的特点,在线课堂环境中老师对学生学习时的干预作用减弱,不利于老师对学生学习进程的把握,当学生注意力不集中时,老师可能没办法及时引导,课中缺乏反馈机制,老师和学生之间的交互性较差[2]。由于缺乏合适的系统监控手段,在线教育的教育质量低下,依靠数字化技术手段和有效的系统监控学习者的学习过程并向教师提供反馈是提升教育效果的关键。图1显示了某状态检测系统的界面,该系统旨在帮助教育者实时监测学生在在线课堂中的学习状态。其中,课中状态识别算法在系统中扮演了关键角色,通过课中状态识别算法,系统可以自动判定不同时间下不同学生的状态,例如专注度、情感、理解度等。因此,算法的准确性和

基金项目:上海市工程技术中心项目(18DZ2252300)
This work was supported by the Shanghai Engineering Research Technology Center Project(18DZ2252300).
通信作者:袁玉波(ybyuan@ecust.edu.cn)

实时性直接决定了该系统的有效性和可用性。然而，现有的学习者状态识别方法往往未能充分利用到在线课堂中不同模态的数据，仅仅对某一维度的数据进行单一的决策，而提高算法的准确度和性能恰恰需要对课中数据进行更深入的探索。

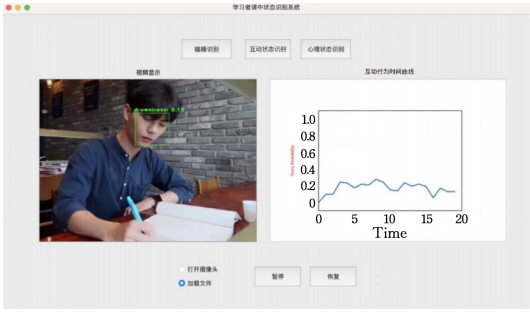


图 1 课中状态界面示意图

Fig. 1 Schematic diagram classroom state interface

因此，为解决这一问题，本文提出一种新的方法，通过深入挖掘学习者的课中视频和音频数据，设计并构建出一系列有关学生课中互动状态的多维度特征。这些特征涵盖了学生的表情、语音、姿态等多方面的信息，全面反映了学生在课堂中的实际状态。这一过程的实现，充分利用了智慧教育平台提供的丰富数据资源，以及对学生学习行为的深入理解和分析。然后，引入特征选择算法，从众多特征中选出对学生学习状态评估最具影响力的关键特征，并基于筛选出的关键特征构建在线课堂学习者互动状态识别的二分类模型。最后，进一步构建了在线课堂总体情况评估模型，该模型为通过统计每名学习者的课中状态识别结果得到一系列度量课堂互动程度的关键性指标。

该方法能够帮助教师在网络课堂上更有效地管理学生，提升教学质量。其直接提取和学习状态相关的高层次特征，相比深度学习方法，模型的可解释性更强，教师和学习者可以得到评估结果的具体依据，以便进行针对性改进。

为了全面验证本文所提出的课堂互动状态检测算法的有效性，构建了一套真实课堂场景下包含不同学习者互动行为的视频数据集，用于模型的训练和测试，并且利用该模型设计了一套基于在线课堂教学视频的课堂状态实时评价系统，从而对提升教学质量发挥支持作用。本文的主要贡献和创新点如下：

(1)建立了一个新的学习者互动状态识别模型，利用真实课堂场景下的教学视频库进行实验，证明模型的可行性。

(2)给出了表情、上肢姿态、面部、语音等多模态特征模型，并通过特征选择算法筛选关键特征。通过分类器训练及模型优化手段，实现对学生课堂状态的精确识别。

(3)构建了在线课堂总体互动情况评估模型，实现了对整个班级的全面评估。这种评估模式不仅包括单个学生的学习状态，还包括整个班级的总体学习情况，提供了更为全面和深入的课堂状态评估，有助于教师更有效地调整和优化教学策略。

2 相关工作

早期的课中状态评估方法主要依赖人工，教师或者教育研究者需要观察和记录学生的行为和表情，这种方法虽然直观，但是评价相对主观，时效性也不高[3]，且耗费大量的人力。为了提高评价的客观性，有研究者尝试使用脑电图和心电图[4-5]等方式来监测学习者的生理状态，从而间接推断他们的课中状态。Zaletelj 等[6]提出利用 Kinect One 传感器所获的数据计算出一组关于学习者的面部和身体属性特征来评估学习者的课中互动状态，构建了一套自动评估系统，实验结果显示其能够达到 75.3% 的精度。这种方法可以提供比人工更准确和细致的数据，但是需要昂贵的硬件设备作为技术支持，且有一定的侵入性，其应用范围相对有限[7]。随着人工智能和计算机视觉技术的发展，课堂状态识别开始转向非侵入式的方法，并且近年来智慧教育平台功能的日益丰富，可以为我们提供大量的在线课堂数据[8]，包括学习者的视频和音频记录，以及文本聊天记录等等，通过分析和挖掘这些数据，能够深入理解和评估学生的课堂状态，从而更好地指导教学。一些研究者使用深度学习等算法对学生在课堂上的图像进行特征提取和识别，从而评估他们的状态。Zhao 等[9]从行为检测以及表情识别来判断学习者的兴奋状态，在行为检测方面，结合了 CPM(Convolutional Pose Machines)和 CMU(Carnegie Mellon University)OpenPose 关键点检测技术。Chen 等[10]运用迁移学习技术将在 ImageNet 数据集上训练过的经典深度网络模型 VGG16 应用到学生课堂行为识别中，通过学习者的课堂行为评估其上课的专注情况。Chen[11]利用语音识别技术定位课堂发言片段的起始位置，通过语音性别识别技术和语音情感识别技术对课堂发言的性别和情感交互进行识别和记录，多维度地对课堂情绪、课堂模式以及课堂互动结构等进行了量化分析，探究了多种课堂交互行为。以上方法虽然在准确率上有所提升，但是仍有不足。一方面，尽管深度神经网络可以实现特征的自动提取而不需要人工干预，但在缺少公开的在线课堂数据集，即训练样本数量不够庞大的情况下，模型在泛化方面会存在一定的局限性。另一方面，这些方法往往从单一角度出发，比如只针对表情或学习者行为进行识别，然而真正的学习过程是复杂的，单一维度的数据在提供学习过程信息时存在局限性，无法充分涵盖学生学习过程中复杂的心理路径，从而限制了学习分析的准确性。为了准确理解学习过程，必须采用更全面的数据收集分析方法。

因此，开发融合多模态数据的课堂状态评估方案，成为实现更全面、深入学习图景的关键[12]。多模态学习分析技术(MMLA)的广泛应用，为理解学生心理状态与学习活动的关系提供了新的视角[13]。已有多项研究证明了在多模态学习状态分析中整合不同数据源的重要性。Ochoa 等[14]成功从工作会话的多模态记录中提取简单特征，包含写字速度、数字和数学术语的发音等，这些特征能够高度成功地区分数学问题解决过程中的专家和非专家。本项研究尽管应用场景相对专一，但为类似方法在更广泛应用中的有效性提供了初步证据。Maldonado 等[15]利用与学习材料和数字评估的交互数据以及自我报告的数据来建立学习者成绩的预测模型，结果表明平均错误率为 15%，最佳情况错误率为 11.3%。Oviatt 等[16]探讨了使用多模态数据如语音、书写、凝视、手势等来评估学习者的心理状态，并对各模态的特征进行了多层次分析，

对学习成果的预测可以在个性化学习中应用,但是对具体的特征提取技术没有详细说明,并且对多模态的特征数据的同步缺乏深入探索。Giannakos 等[17]建立了基于生理反应的多模态数据流模型,收集了点击流数据,尽管该研究在捕捉和预测学习者行为方面取得了进展,但其在多模态数据范围和选择上存在一定局限性,主要关注了诸如击键、脑电图、眼动追踪等传统生理和行为数据,而忽略了面部表情和身体姿态等关键的非言语交流特征,缺乏这些客观的非言语特征限制了学习状态分析的全面性和深度。

迄今为止,还没有研究试图利用在线课堂学生生成的多模态数据对教学中的课中学生互动状态进行自动化识别,也没有基于视频学习环境深入探索除生理特征外,其他可利用的重要特征与学生的课中学习状态之间的相关性。因此,本文提出了一种新的多模态分析模型,结合深度学习技术,提取和融合了视觉和听觉的多模态特征,实现了特征的有效对齐,并且验证了所构建特征的有效性,这些特征随后被用于训练互动状态分类模型。通过这种方法,模型能够全面捕获学生在学习过程中的微妙变化。实验结果表明,与当前的基于单模态的课中互动状态分析方法相比,本文提出的多模态分析模型在准确性和可靠性方面有显著提升。

3　在线课堂学习者互动状态识别方法

3.1　基本框架

为了更全面地捕捉在线课堂中学生的互动状态,本文提出了一种基于多维在线课堂情景数据的互动状态检测方法,算法流程如图 2 所示。该算法不仅考虑了学生的姿态、表情、面部和语音特征,而且最终能够综合所有学生的互动状态,构建出全局的在线课堂互动状态识别模型。

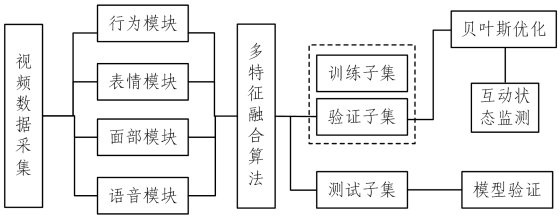


图 2　课堂互动状态识别算法的技术路线图

Fig. 2　Technology roadmap of classroom interaction state recognition

首先,对输入的视频及音频数据进行处理,利用 OpenPose[18]框架提取学生的上肢关键点,并基于这些关键点构建出上肢姿态角度特征,以捕捉学生的身体语言。对于面部表情特征,采用 Mini_XCEPTION[19]模型进行提取,从而识别出学生的情绪状态。同时,对学生的视线方向、眉间距、眼睛闭合度等面部特征进行了分析,进一步丰富了面部特征的维度。最后,使用说话人识别技术计算出学生的语音特征,以更好地理解学生的语音交流情况。

其次,利用 LightGBM[20]算法对上述提取出的 16 维特征进行了融合,结合贝叶斯优化策略对 XGBoost[21]和 CatBoost[22]等模型进行优化,使得模型在 SSHI-ED 数据集上表现最佳。

通过以上步骤,构建了一个在线课堂学习者互动状态识别模型。该模型不仅能够深入挖掘学生的多种特征,而且还能够充分利用这些特征之间的关联,从而实现更为准确和全面的在线课堂状态评估。

3.2　特征抽取

3.2.1　书写特征抽取

在传统课堂教学中,学习者的身体语言,尤其是上肢姿态,常常被视为其课堂参与度和互动情况的直观指标。姿态,如举手、记笔记或趴桌子等,不仅反映了学习者的参与程度,同时也关系到教学质量和课堂活跃度的提升。但现有的学习状态分析中,仍然缺乏对这些姿态的识别和解释。为了弥补这一空缺,本文构建了描述书写姿态的特征集合,利用 OpenPose 框架提取人体姿态 18 个关键点,分别对应人体上可以自由活动的关节,包括颈部、左右肩部、肘部等位置。

在线课堂场景中,由于摄像头位置的局限性,只能获取到关于学习者上肢姿态的有效信息,包括头部、颈部、肩部以及肘部等,因此本文重点关注人体上肢关键点。对需要关注的关键点进行编号,本文所提取的主要关键点编号以及相应位置如表 1 所列。

表 1　主要关键点编号以及相应位置

Table 1　Main key point numbers and their corresponding positions

编号	0	1	2	3	4	5	6	7
位置	鼻尖	颈	右肩	右肘	右腕	左肩	左肘	左腕

根据对学习者常见课堂行为动作的分析,当学生处于互动状态时可能会有频繁翻书或频繁写字的动作,观察双肩夹角、双肘夹角以及大拇指和食指的夹角及其变化角可以判断学生是否进行了上述动作。本文构建了 5 个书写特征,如表 2 所列。

表 2　书写特征构造公式

Table 2　Writing feature construction formula

书写特征	计算公式
α_1	$\frac{1}{2}\times\left[\arccos\left(\frac{F_{24}^2-F_{23}^2-F_{34}^2}{-2\times F_{23}\times F_{34}}\right)+\arccos\left(\frac{F_{57}^2-F_{56}^2-F_{67}^2}{-2\times F_{56}\times F_{67}}\right)\right]$
α_2	$\frac{1}{2}\times\left[\arccos\left(\frac{F_{13}^2-F_{12}^2-F_{23}^2}{-2\times F_{12}\times F_{23}}\right)+\arccos\left(\frac{F_{05}^2-F_{01}^2-F_{15}^2}{-2\times F_{01}\times F_{15}}\right)\right]$
α_3	$\arccos\left(\frac{F_{05}^2-F_{01}^2-F_{15}^2}{-2\times F_{01}\times F_{15}}\right)$
α_4	$\arccos\left(\frac{L_{08}^2-L_{06}^2-L_{78}^2}{-2\times L_{06}\times L_{78}}\right)$
α_5	$\arccos\left(\frac{L_{08}^2-L_{06}^2-L_{78}^2}{-2\times L_{06}\times L_{78}}\right)$

为了监控学习者在一段时间内的课堂行为变化情况,本文统计了一段时间内各姿态角的变化值及平均变化值,能够更加准确地判断学习者是否处于记笔记、趴桌子等状态。部分计算方法如式(1)、式(2)所示。

$$\beta_1=\frac{\partial\alpha_3}{\partial t} \tag{1}$$

$$\beta_2=\frac{\partial(\alpha_1+\alpha_2+\alpha_4+\alpha_5)}{\partial t} \tag{2}$$

根据学习者的上肢关键点信息构建上肢姿态角特征,并通过滑动窗口计算在一段时间内姿态角的变化情况,从而达到精确评估学习者的课中互动状态的目的。学习者书写特征提取算法的伪代码如算法 1 所示。

算法 1　Extracting Writing Feature Vectors via Sliding Window

Require: Video sequence $Video_i$

Ensure: Writing feature vector F

```
1. function EXTRACTFEATURES(Video_i, N)
2.     Initialize sliding window W of size N
3.     for k ←1 to K do
4.         keypoints ← EXTRACT(Video_i[k])
5.         angles ← CALCULATE(keypoints)
6.         Appendangles to feature vector F
7.         if LENGTH(F) > N then
8.             Remove the first element from F
9.         end if
10.        if LENGTH(F)==N then
11.            Δ←AVERAGECHANGERATE(F,N)
12.            return Δ
13.        end if
14.    end for
15. end function
```

3.2.2 表情特征抽取

人的表情在日常交流中扮演着重要的角色，能够极大程度地反映出一个人的内心状态。在线课堂场景中，学习者的表情和他们的课堂互动状态密切相关。通过分析学习者的表情，有助于评估学习者以及他人的参与、互动程度。

为抽取表情特征，本文搭建了表情识别模型，参考了CNN的主流框架Mini_XCEPTION。Mini_XCEPTION是一个基于深度可分离卷积[23]的架构，包含4个残差模块[24]，每个残差模块后都有一个批标准化层和激活函数ReLU。最后一个卷积层采用全局平均池化层和Softmax激活函数进行预测。本文提出的架构结合了残差模块和深度可分离卷积的优势，减少了参数个数，同时在FER-2013数据集的情感分类任务中实现了66%的准确率，保持了实时性。模型得出的7类表情的分类结果包括喜悦、愤怒、惊喜、恐惧、厌恶、悲伤和中立。对于全连接层处理提取出的特征，通过Softmax函数将其转变为概率形式。然而，在实际的课堂环境中，学生的表情往往不会如此丰富和复杂。因此，为了更贴合在线课堂场景，将表情主要分为积极、中立和消极3类，积极的表情包括喜悦和惊喜，中立表情自成一类，愤怒、恐惧、厌恶和悲伤为消极类。同时，为了平衡不同情绪类别的影响，根据每个类别内包含的表情数量来分配权重，以确保每一种基本表情具有公平的代表性，若一个类别中有更多的表情，每种表情的权重就会相应减少。具体而言，将总权重1平均分配给积极、中立和消极类，然后根据每类情绪包含的表情数量，进一步平均细分权重。最后，表情特征维度由加权后的表情概率得分构成。课堂教学视频中学生表情识别算法的伪代码如算法2所示。

算法2 Facial Feature Analysis for Emotion Recognition

```
1. X ← detect()
   / * Detect student's face and obtain facial dataX * /
2. F ← [F_0, F_1, F_2, F_3, F_4]
   / * Extract a set of featuresF from X * /
   / * Attention weight allocation and optimization * /
3. for i ← 0 to 4 do
4. z[i] ← denseLayer(F[i])
   / * Process featureF[i] using a dense layer * /
5. q[i] ← softmax(W_2 · ReLU(W_1 · z[i])
   / * Transformz[i] and apply softmax to get attention weight q[i] * /
6. end for
7. F_m ← (∑(q_i F_i)) / (∑(q_i))
   / * Compute the fused feature F_m * /
8. Y ← softmaxLayer(F_m)
   / * Apply softmax to convertF_m into probabilities * /
9. Weights ← [W_1, W_2, W_3, W_4, W_5, W_6, W_7]
   / * Define weights for seven emotion categories * /
10. for i ← 0 to 7 do
11.    Y[i] ← Y[i] · Weights[i]
       / * Weight each emotion category's probability * /
12. end for
13. Scores ← [∑(Y[0:2]), ∑(Y[3:4]), ∑(Y[5:7])]
    / * Aggregate scores for emotion groups * /
14. return Scores
```

3.2.3 面部特征提取

面部特征，如视线方向和眉毛的皱褶，往往可以直观地反映学习者的注意力集中和情绪状态。例如，视线的定向和持续性能揭示学习者对课程内容的关注度，而眉间的距离变化可能与理解上的困难或不满情绪相关。为抽取面部特征，本文运用双边滤波、腐蚀和阈值处理等方法来分割瞳孔位置，并计算图像矩以确定瞳孔质心的位置。基于这些质心位置，我们能够判断学习者的视线方向。此外，通过面部关键点定位计算出眉间距离，从而得到反映学习者情绪状态的关键数值指标，技术路线如图3所示。

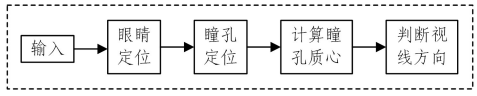


图3 视线追踪方法的技术路线图

Fig. 3 Technical roadmap of gaze tracking method

瞳孔质心位置的表达式如下所示：

$$P_{XC}=\frac{m_{10}}{m_{00}} \tag{3}$$

$$P_{YC}=\frac{m_{01}}{m_{00}} \tag{4}$$

其中，(P_{XC}, P_{YC})表示瞳孔的质心位置，m_{ij}表示图像的阶矩。最后通过质心横坐标位置与眼睛轮廓的宽度计算出眼睛追踪指标，表达式如下所示：

$$Gztde=\frac{P_{XC}}{Ex\times 2} \tag{5}$$

通过P_{XC}和眼睛轮廓的宽度的比值计算出视线方向指标Gztde，$Ex\times 2$为眼睛的轮廓宽度。

此外，根据面部关键特征点的变化可以判断眉间距的变化，当学习者皱眉时，可以推测出该名学习者可能处于疑惑或者不耐烦状态。所以本文将眉间距的变化数据作为衡量学习者课堂互动状态的标准之一，计算公式如下：

$$b_w=\frac{1}{W_f}\left[\frac{1}{n}\sum_{i=17}^{21}(P_{x_i+5}-P_{x_i})\right] \tag{6}$$

$$b_h=\frac{1}{W_f}\left[\frac{1}{n}\sum_{i=17}^{21}\frac{1}{2}(P_{y_i}+P_{y_i+5}-2f_t)\right] \tag{7}$$

其中，b_w表示眉毛的距离占比；b_h表示眉毛的高度占比；P_{x_i}表示眉间关键点的横坐标；P_{y_i}表示眉间关键点的纵坐标；n表示所用到的眉间关键点数；W_f表示人脸识别框的宽度；f_t表示人脸识别框的top坐标。

课堂教学视频中学生面部特征提取算法的伪代码如算法3所示。

算法 3　Keypoint Detection and Feature Extraction for Eye and Mouth in Video Frames

Require: Video sequence V

Ensure: Keypoint coordinates and feature ratios

1. for each frame f_k in V do
2. $D_k, B_k \leftarrow$ Detect(f_k)
   / * Detect eye and mouth keypoints * /
3. $G_k \leftarrow$ GazeDetector(D_k)
   / * Extract pupil centroid and compute gaze direction * /
4. $F_k \leftarrow$ EyebrowDetector(B_k)
   / * Compute frowning degree based on eyebrow distance * /
5. frown_ratio. append(F_k)
6. gaze_ratio. append(G_k)
7. end for
8. return { froun_ratio, gaze_ratio}

3.2.4　语音特征提取

在课堂互动中，学生的语音反馈对评估互动程度至关重要。通过对学生语音的实时监测和分析，能够更准确地理解学生的参与度和反馈情况，说话的频率、持续时间等信息可以反映学生的活跃度；且与视觉特征相比，语音特征因不易受视角或遮挡影响而更稳定、可靠。为抽取语音特征，本研究通过语音分割和说话人识别技术，基于高斯混合模型（Gaussian Mixture Model，GMM）进行语音活动检测，剔除空白部分并通过滑动窗口和贝叶斯信息准则识别说话人变化，最后利用梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients）特征和预训练的 GMM 模型识别说话人身份，有效补充其他模态信息。该方法的技术路线如图 4 所示。

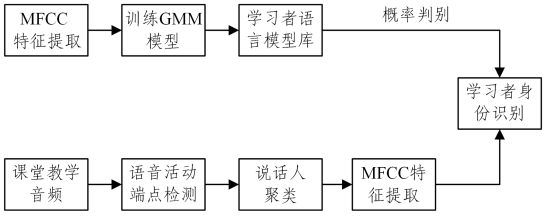


图 4　学习者身份识别方法的技术路线图

Fig. 4　Technical roadmap of learner identity recognition method

由于尚无公开的大型在线课堂的语音数据集，本文提出的算法对训练数据量的要求相对较低，且相对于复杂的声学模型和深度学习方法，基于贝叶斯准则和高斯混合模型的算法通常具有较低的计算复杂度，因此该方法可以在资源有限的设备上运行。学习者应答状态检测算法的伪代码如算法 4 所示。

算法 4　Speaker Change Detection in Speech Segments

Require: Audio signal A

Ensure: Speaker change responses {response}

1. speech_segment ← GMM_speech_detection(A)
   / * Detect speech activity * /
2. window_size ← 1 000
   / * Set the size for the sliding window * /
3. overlap ← 500
   / * Set the overlap size for the window * /
4. speaker_change ← False
   / * Initialize spcaker change status * /
5. for k ← 1 to K do
6. windows ← sliding_windows(window_size, overlap)
7. distances ← calculate_distances(windows)
8. speaker_change ← BIC(distances)
   / * Detect speaker change using BIC critcrion * /
9. if speaker_change then
10. response[k] ← identify(gmm_feature)
    / * Identify new speaker * /
11. end if
12. end for
13. return {response}

3.3　模型构建

3.3.1　BO-CatBoost 算法

为了优化模型的计算效率，本文采用了数值化的方法来整合多模态特征，降低了计算复杂性。本文选择 CatBoost 算法对多模态学习状态特征进行分析并进行学习者互动状态预测。CatBoost 是一种结合了 GBDT[25] 和分类特征的算法，对于各种特征类型，尤其是数值化特征，具有出色的处理能力和鲁棒性。相比于传统的 GBDT 算法，在处理类别特征、防止过拟合、训练速度、特征组合以及模型可解释性等方面都有了显著的改进。其中，对类别特征的处理方式是利用样本标签的平均值进行表示。

$$x_{ik}=\frac{\sum_{j=1}^{n}[x_{jk}=x_{ik}]\cdot Y_j}{\sum_{j=1}^{n}[x_{jk}=x_{ik}]} \tag{8}$$

CatBoost 在处理类别特征时采用了一种随机序列生成的方法，以避免过拟合的问题。它通过生成一个随机序列，并利用前面遍历到的记录来计算每个特征的数值。在这个过程中，结合了样本标签的均值、先验值和先验值的权重进行计算，以有效地处理类别特征。

$$x_{\sigma_p,k}=\frac{\sum_{j=1}^{p-1}[x_{\sigma_j,k}=x_{\sigma_p,k}]\cdot Y_{\sigma_j}+a\cdot P}{\sum_{j=1}^{p-1}[x_{\sigma_j,k}=x_{\sigma_p,k}]+a} \tag{9}$$

在机器学习领域，超参数优化是提升模型性能的关键步骤。超参数，即模型训练前需设定的参数，对学习过程和最终性能有决定性影响。因此，针对特定数据集或任务找到合适的超参数组合至关重要，这个过程通常需要训练多个不同参数组合的模型，而后评估找到最优性能的模型。

贝叶斯优化（Bayesian Optimization，BO）是一种用于超参数优化的迭代算法。与传统的网格搜索和随机搜索方法相比，BO 能够更高效地搜索参数空间，避免了维度灾难的问题。其利用历史评估结果，不断更新概率模型，减少了迭代次数，提高了搜索效率。利用 3.2 节提出的多维度特征抽取算法构建多特征向量，采用特征选择算法筛选对学习者互动状态识别贡献程度较高的关键特征，构建课中互动状态数据集，利用训练集数据进行 BO-CatBoost 模型的超参数寻优过程，迭代获得最优超参数组合，进一步提升模型的预测性能。算法伪代码如算法 5 所示。

算法 5　Classroom Interaction State Identification

Require: Video frame sequence **V**

Ensure: Interaction state labels **M**

1. **E** ← ExtractExpressionFeatures(**V**)
2. **P** ← ExtractPostureFeatures(**V**)
3. **F** ← ExtractFacialFeatures(**V**)
4. **V** ← ExtractResponseFrequency(**V**)

5. **S** ← Concatenate(**E**, **P**, **F**, **V**)
6. **S**′ ← XGBoostOptimize(**S**)
7. Data ← Concatenate(**S**′, **y**)
8. Train, Test ← CrossValidationSplit(Data)
9. OptimizeParameters(Models, Train)
10. pred ← Predict(Models, Test)
11. **M** ← Classify(BO-CatBoost, **S**′)

3.3.2 在线课堂教学质量评估模型

本文基于学习者课中互动状态识别算法构建了在线课堂智能教学评估模型，依靠对单名学习者的课中互动状态识别，对班级内的学习者的课中状态进行全面的评估。系统流程如图 5 所示。首要步骤是基于坐标基准切割在线课堂教学视频数据得到单名学生视图，对这些分割的视频片段进行音频提取，从而得到每个学生的视频和音频数据。将这些信息分为 4 个维度进行特征提取，包括学习者的姿态特征、表情特征、面部特征，以及音频特征。其次，经过在线课堂学习者互动状态识别模型输出得到每名学习者的课中互动状态结果。为了得到全班的互动状态概览，将所有学生的状态结果进行统计，从而分析出班级总体的互动教学情况，帮助教师实时地调整教学方式，实现教育智慧化。

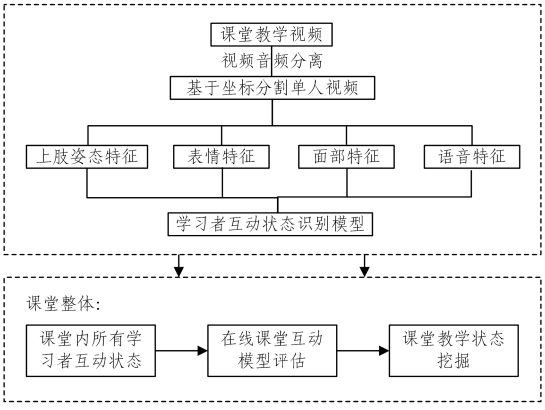


图 5 课堂教学评估模型框架

Fig. 5 Framework of classroom teaching evaluation model

在评估模型中，考虑到学习者的状态并非瞬时变化，而是具有持续性，在实际的在线教学过程中，由于网络延迟、图像质量等原因，有些帧的信息可能会丢失或识别不准确，使用滑动窗口方法评估学习者的课堂互动状态可以在一定程度上抵消这些噪声和误差，提高状态识别的准确性。本文选取了长度为 150 帧的滑动窗口，在该时间范围内若学习者在某帧内处于互动状态，则将这个窗口内的学习者标记为互动状态，以此统计在线课堂学习者互动频率区间分布及互动得分时间序列变化指标。计算方式如下式所示。

$$F_i=\frac{1}{m}\times\sum_{i=1}^{m}(b_i \text{ for } b_i \text{ in } T_i),T_i=\{t_1,t_2,\cdots,t_m\} \tag{10}$$

其中，T_i 为滑动窗口序列，m 表示课堂内划分的滑动窗口总数，b_i 为第 i 个窗口内学习者的互动状态值。

另外，通过记录学习者互动状态平均得分随时间变化的情况，构建一个时间序列，展示课堂互动情况在不同时间段的趋势，如式(11)所示。

$$E_i=\frac{150\times\sum_{i=1}^{n}b_i}{n},i\in[1,n] \tag{11}$$

其中，n 为该课堂视频总帧数。

4 实验

4.1 数据集

SSHI-ED 数据集为网络上以及真实课堂场景下，经过师生同意后录制的视频组成的共 20 段视频，每段视频时长大约为 30 min。每个视频都代表一个真实的在线课堂场景，包含多名学习者的视频图像数据和课堂音频数据，涵盖了学习者认真听讲、记笔记、玩手机、左顾右盼等情景。为了便于后续分析，根据不同在线课堂平台的特点，对视频图像数据进行了基于坐标的切割，以获得单人视频数据。相应的音频数据被保留，以备后续分析使用。经人为筛选后保留关键帧共 14 926 张，并按照 70%和 30%的比例划分训练集和测试集。课中互动状态标签已经在数据集中逐帧标出，在标注过程中进行了多次验证，以保证标注结果的可靠性。数据集部分示例如图 6 所示，每个示例由视频片段中的关键帧组成。

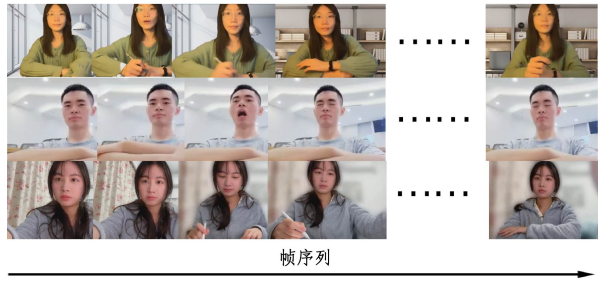


图 6 SSHI-ED 视频数据集的样本帧示例

Fig. 6 Sample frame example of SSHI-ED video dataset

4.2 实验结果与分析

实验的环境是：Windows 操作系统，8 个 CPU 核心，1 颗 Nvidia RTX3060LP GPU 核，16GB 的 RAM，在 Python3. 8 的环境中编程实现。

本文采用准确率(Accuracy)和 F1 值来评估模型的效果。

准确率指的是在预测结果中预测正确的数量与总数量的比例，即所有样本中预测正确所占比例：

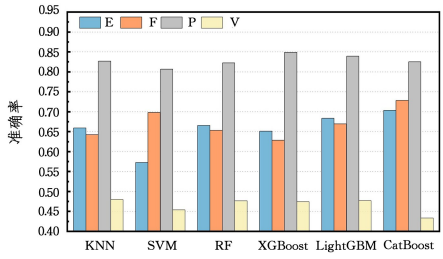
$$Accuracy=\frac{TP+TN}{TP+TN+FN+FP} \tag{12}$$

其中，TP 为所有类别的实例数中被正确分类的数量；TN 为被正确分类为属于与所考虑类别不同的类别的实例数；FP 为所有类别的实例数中被错误分类的数量；FN 为被错误分类为属于与所考虑类别不同的类别的实例数。

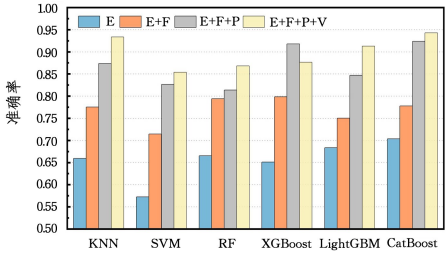
F1 值结合了精确率与召回率两个指标，F1 值越高说明实验效果更好。

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{13}$$

本文分别对不同特征维度在不同模型中的评估指标进行比较，将所提取到的互动状态特征维度输入到 KNN，SVM，RF，XGBoost，LightGBM 和 CatBoost 模型中。图 7(a)展示了单维度特征分别输入到不同模型中的表现，其中，基于姿态特征 P 的模型准确率普遍高于其他特征维度，说明学习者姿态特征和课中互动状态联系更为紧密。图 7(b)展示了在表情特征 E 的基础上依次新增的特征维度在各个评估指标上的结果都有一定的提升效果，即在表情特征的基础之上依次加入面部特征、书写特征、语音特征时，评估指标都呈梯度上升，验证了本文多特征的有效性。



(a)基于单特征维度的模型准确率对比

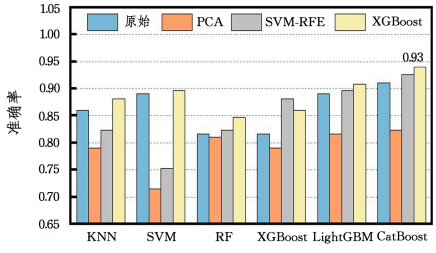


(b)基于不同特征维度组合的模型准确率对比

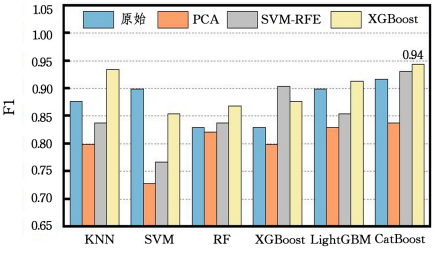
图 7　不同特征维度的评估结果

Fig. 7　Evaluation results of different feature dimensions

图 8(a)和图 8(b)分别为不同特征选择算法对不同模型的准确率和 F1 值的对比。图中显示，相较于其他模型而言，利用 XGBoost 特征选择后的 CatBoost 表现效果最好。F1 分数总体高于准确率，说明相较于非互动状态的检测，预测互动状态正确的数量更多。此外，由特征重要性的排序结果可得，姿态维度特征占据了关键特征的最高比例，其次是表情维度特征、语音维度特征和面部特征，这说明学习者的互动状态与姿态特征、表情特征联系更为紧密，而语音特征的相对较低排序则可能是由于其受限于视频中的时长和数据质量所致。并且，当特征数量递减到 8 维时，评估的准确率达到最高，因此综合考虑特征选择结果以及分类器的表现后，最终选择这 8 个关键特征作为输入，用于训练分类器。



(a)特征选择后不同模型的准确率对比示意图



(b)特征选择后不同模型的 F1 对比示意图

图 8　特征选择后不同模型的评估指标对比

Fig. 8　Comparison of evaluation indexes of different models after feature selection

参数对模型最后的结果有重要影响。因此，想要效果达到最佳，还需要对各个模型进行参数调优。本文所选择的参数是影响模型效果最重要的 5 个参数，图 9 显示了基于贝叶斯优化策略的模型参数迭代图。

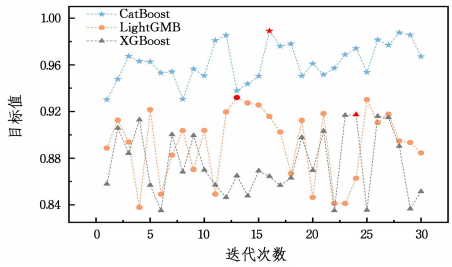


图 9　基于贝叶斯优化的模型参数迭代图

Fig. 9　Iterative graph of model parameters based on Bayesian optimization

从图 9 中可以看出，CatBoost 中的最优参数组合在第 17 个 epoch，XGBoost 中的最优参数组合在第 23 个 epoch，LightGBM 中的最优参数组合在 12 个 epoch。表 3 列出了具体参数的名称、查找范围以及最优参数值等。

表 3　基于贝叶斯优化模型的最优参数选择

Table 3　Optimal parameter selection based on Bayesian optimization model

参数	查找范围	最优参数值
learning_rate	$(1\times10^{-6},1\times10^{-2})$	0.008652
depth	(2,15)	8.391
l2_leaf_reg	(0,5)	0.08059
n_estimators	(30,2000)	701.4
learning_rate	$(1\times10^{-6},1\times10^{-2})$	0.00753

为了验证模型在不同课堂场景下的效果，将部分测试集依据光线以及有无遮挡进行了划分。表 4 列出了在不同情景下的模型评估结果。在光线方面，比较了光线良好和光线较差的情况对模型的影响。结果显示，在光线良好的情况下，模型的准确率明显高于光线较差的情况，表明模型在充足光线下能够更好地捕捉关键特征，从而提高准确率。另外比较了脸部或身体是否有遮挡对模型性能的影响，在无遮挡情况下，模型的准确率明显高于有遮挡的情况，表明模型在特征提取完整的情况下表现更佳。

表 4　不同情境下互动状态评估算法的结果对比

Table 4　Results comparison of interactive state evaluation algorithms in different scenarios

(%)

场景	准确率	*F*1	平均值
光线较好	93.7	95.6	94.6
光线较暗	85.8	86.2	86.0
有遮挡	87.6	88.7	87.1
无遮挡	92.4	94.2	93.3
平均值	90.3	91.2	90.2

为了充分验证所提出的基于多维度课堂情景数据的在线课堂学习者互动状态识别算法，将其与几种不同方法进行了对比，包括基于表情的学习状态评估算法、基于行为的学习状态评估算法以及基于两者的学习状态评估算法。将预测结果和实际标注结果进行比较，部分对比结果列于表 5 中。结果显示，本文的 BO-CatBoost 算法在识别准确率方面具有明显优势。此外，一些基于表情和行为的算法的结果低于单独基于表情或行为的算法的结果。推测这是由于有限的信息融合导致的不稳定性。本文算法充分利用多种信息源，通过结合表情、行为、语音和面部特

征，能够更好地捕捉学习者的多模态交互行为。

表 5 各课堂互动状态评估算法的结果对比

Table 5 Comparison of classroom interaction evaluation results among different algorithms

(%)

评估算法	视频编号				
	1	3	7	9	10
基于表情的状态评估	80.4	78.2	83.2	79.2	75.8
基于行为的状态评估	87.3	88.2	85.7	82.8	79.6
基于表情和行为的状态评估	82.8	86.3	89.3	83.5	79.8
本文算法	91.7	90.1	92.9	93.6	88.5

4.3 课堂样例展示

课堂互动状态往往被认为是评估学习者参与程度和学习效果的重要指标。为了更全面地了解课堂互动的情况，本章选取了 SSHI-ED 数据集中的一节包含 21 名学习者的在线课堂，时长为 30 min。该时长较为典型，可以较好地代表一节课的时间跨度。在该时间范围内，捕捉和识别学习者的互动行为和状态，通过分析揭示学习者互动的整体水平、个体差异及时间变化，部分视频样例如图 10 所示。在图中会显示学习者上肢姿态及面部区域特征等各项指标，当学习者处于互动状态时，系统根据学习者兴奋状态识别算法将其标注为互动状态，1 代表学习者处于互动状态，0 代表学习者处于非互动状态。

(a)识别样例 1

(b)识别样例 2

(c)识别样例 3

图 10 提出的模型在课堂场景中的识别效果

Fig. 10 Recognition effects of the proposed model in classroom scenes

获取到每名学习者随时间变化的互动状态值后，从课堂开始的时间点开始，以每 5 min 为间隔进行了互动状态的分析。在每个时间点，平均互动分值是一个综合考量，反映了课堂整体互动情况的指标。较高的平均互动分值表示课堂中的学生在互动方面表现较好，如表 6 所列。

表 6 课堂互动状态随时间的变化

Table 6 Changes in classroom interaction state over time

(%)

时间	互动平均值	互动人数	非互动人数
00:00	0.82	7	14
00:05	0.78	8	13
00:10	0.76	6	15
00:15	0.83	9	12
00:20	0.72	5	16
00:25	0.85	10	11
00:30	0.81	7	14

通过表 6 能够直观地了解课堂中的互动情况是如何随着时间的推移而变化的。这些指标为进一步分析学生在不同时间段的互动情况，为探讨教学效果以及优化教学策略提供了有力的数据支持。

结束语 本文针对在线课堂场景下的学习者互动状态识别任务，提出了一种基于多特征融合的学习者课中互动状态识别模型。该模型结合了在线课堂音视频数据，基于姿态、表情、面部及语音 4 个模块构建了 16 维互动状态特征，实现了在线课堂情景下的学习者互动状态识别。在自建的真实课堂数据集上的实验结果表明，对比基于学习者表情或行为的识别模型，本文提出的互动状态识别模型取得了较高的准确率。此外，本文基于提出的在线课堂互动状态识别模型，设计并实现了基于学习者互动状态识别的在线教学质量评估系统，实现了课堂状态的整体评估，助力智慧教育的管理和质量的提升。

本研究面临的主要挑战之一是在线课堂数据集的公开可用性不足。为应对这一问题，本文选择了建立自有的数据集，并将研究范围限定在大学生课堂环境中。这一选择虽然使得研究能够在受控条件下进行，但也带来了样本量有限和被试选择局限性的问题。当前数据集可能导致结果存在一定的取样偏差，其在不同教育阶段的应用范围和有效性仍有所欠缺。考虑到中小学生课堂行为及表现的特殊性和差异性，未来的研究计划中包括将中小学生的课堂数据纳入验证范围以增强模型和方法的泛化性。此外，本研究多模态数据的分析过程不是实时的，因此学习者无法获得多模态数据的教育支持，以推进学习技术并增强学生的学习体验。向学生提供实时反馈可以帮助学生立即改正不正确的行为或能花更多的时间重新思考自己的意见。另一方面，虽然本研究能初步识别课堂互动状态，但没有对互动状态进行等级上的划分，未来也会在这一方面进行进一步的研究。

参 考 文 献

[1] ZHANG W. In the Blink of an Eye: 70 Years of Monumental Changes in Chinese Educational Informatization[J]. China Education Network, 2019(10): 7-10.

[2] SENSMEIERJ, ANDERSON C. ANI CONNECTION: Technology Informatics Guiding Education Reform Moves Into Phase III: Implementation[J]. Cin Computers Informatics Nursing, 2009, 27(4): 265-266.

[3] WEI Y, QIN D, HU J, et al. Student Classroom Behavior Recognition Based on Deep Learning[J]. Modern Educational Technology, 2019, 29(7): 5-7.

[4] MASSOZ Q, LANGOHR T, FRANCOIS C, et al. The ULg multimodality drowsiness database (called DROZY) and examples of use[C]// Winter Conference on Applications of Computer Vision. Lake Placid: IEEE, 2016: 1-7.

[5] SVENSSON U. Blink behaviour based drowsiness detection: method development and validation[J]. Journal of Swedish National Road and Transport Research Institute, 2024, 362A(1): 37-39.

[6] ZALETELJ J, KOIR A. Predicting students' attention in the classroom from Kinect facial and body features[J]. Eurasip Journal on Image & Video Processing, 2017, 2017(1): 80-82.

[7] SHINODA K, YOSHII M, YAMAGUCHI H, et al. Daytime

Sleepiness Level Prediction Using Respiratory Information [C]// Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: IJCAI, 2019: 5967-5974.

[8] CAI Y W. Optimization Research on Wisdom Tree and Tencent Classroom under Computer Artificial Intelligence Algorithm [C]// International Conference on Data Analytics, Computing and Artificial Intelligence. Zakopane: ICDACAI, 2022: 334-337.

[9] ZHAO Y, YAN H, WANG Z. The Advisable Technology of Key-Point Detection and Expression Recognition for an Intelligent Class System [J]. Journal of Physics: Conference Series, 2019, 1187: 052011.

[10] CHEN G, JI J, HUANG C. Student Classroom behavior Recognition based on OpenPose and Deep Learning[C]// 7th International Conference on Intelligent Computing and Signal Processing. Xi'an: IEEE, 2022: 576-579.

[11] CHEN Y. Multidimensional Classroom Interaction Analysis Based on Speech Recognition [D]. Wuahn: Huazhong Normal University, 2021.

[12] PENG S, NAGAO K. Recognition of Students' Mental States in Discussion Based on Multimodal Data and its Application to Educational Support[J]. IEEE Access, 2021, PP(99): 1-1.

[13] EMILY D, TRENT K, DAVE D. Rethinking classroom observation. Journal of Educational Leadership, 2014, 71(8): 24-29.

[14] OCHOA X, CHILUIZA K, MÉNDEZ, et al. Expertise estimation based on simple multimodal features[C]// ACM on International Conference on Multimodal Interaction. ACM, 2013.

[15] MALDONADO-MAHAUAD J, PÉREZ-SANAGUSTÍN M, MORENO-MARCOS P M, et al. Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning [C] // Lifelong Technology-Enhanced Learning: 13th European Conference on Technology Enhanced Learning (ECTEL 2018). Leeds, UK, Springer International Publishing, 2018: 355-369.

[16] OVIATT S, GRAFSGAARD J, CHEN L, et al. Multimodal learning analytics: assessing learners' mental state during the process of learning[M]// The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2. Association for Computing Machinery and Morgan & Claypool, 2018: 331-374.

[17] GIANNAKOS M N, SHARMA K, PAPPAS I O, et al. Multimodal data as a means to understand the learning experience [J]. International Journal of Information Management, 2019, 48: 108-119.

[18] BADAVE H, KUBER M. Evaluation of Person Recognition Accuracy based on OpenPose parameters [C] // 5th International Conference on Intelligent Computing and Control Systems. Madurai: ICICCS, 2021: 635-640.

[19] ARRIAGA O, VALDENEGRO-TORO M, PLÖGER P. Real-time Convolutional Neural Networks for Emotion and Gender Classification[J]. arXiv: 1710. 07557, 2017.

[20] XU N, LI S, WU X, et al. An APT Malware Classification Method Based on Adaboost Feature Selection and LightGBM[C]// Sixth International Conference on Data Science in Cyberspace. Shenzhen: IEEE, 2021: 635-639.

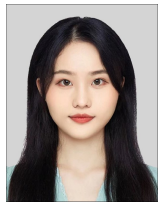
[21] ZHOU Y, SONG X, ZHOU M. Supply Chain Fraud Prediction Based on XGBoost Method[C]// 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering. Nanchang: IEEE, 2021: 539-542.

[22] ZHANG X, WU G X. Text Classification Method of Dongba Classics Based on CatBoost Algorithm [C] // The 8th International Symposium on Test Automation & Instrumentation. Online: ISTAI, 2020: 133-139.

[23] HOWARD A G, ZHU M, CHEN B et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [J]. arXiv: 1704. 04861V1, 2017.

[24] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770-778.

[25] CHEN B, LIN R, ZOU H. A Short Term Load Periodic Prediction Model Based on GBDT[C]// 18th International Conference on Communication Technology. Chongqing: IEEE, 2018: 1402-1406.



RAO Yi, born in 1999, postgraduate. Her main research interests include big data analysis and data mining.

YUAN Yubo, born in 1976, Ph.D, associate pprofessor. His main research interests include artificial intelligence, data science, big data analysis, data quality assessment and data mining.