在线学习行为分析和成绩预测方法

安梦蕾[1] 韩 蒙[1] 任 绒[1] 柯程虎[1] 常永明[1]
AN Menglei HAN Meng REN Rong KE Chenghu CHANG Yongming

摘 要 随着在线课堂在高校课程中的应用越来越广泛，如何对大学生线上课程的学习效果进行有效的跟踪和指导，成为高校教育的难题。采用基于注意力机制的 CNN-LSTM 预测模型，将学生学习行为数据经过相关性分析和归一化处理后输入模型中，并利用注意力机制区分信息的重要性程度，进而提高模型的准确性和可靠性。同时，采用基于 Canopy 的改进 K-means 聚类算法，将学生学习行为特征及成绩预测值利用 Canopy 算法进行预聚类，将预聚类结果结合 K-means 进行聚类。实验结果表明，所提出的成绩预测模型使学生成绩预测准确率达到了 95.69%。

关键词 学习行为分析；成绩预测；LSTM；注意力机制

doi：10.3969/j.issn.1672-9528.2024.09.004

0 引言

在当今世界，信息技术如洪流般汹涌，尤其在人工智能这一领域，其发展之迅猛令人瞩目。教育领域正在经历着前所未有的变革，各种新型的教育模式如在线教育、远程教育和智慧教育等应运而生，极大地丰富了教育的实践形式。在线课堂等新型教育形式的出现为公共服务模式带来了新的活力。同时，社会各界也被鼓励参与到“互联网 + 公共服务”的实践中，在线教育成为当前教育工作者和互联网领域人士重点研究内容 [1]。

在当前的“人工智能 + 教育”背景下，深入研究影响学生成绩的理论和方法，探索学生学习的通用模式，开发出一套精确预测学生成绩的算法或平台，已成为主要的研究任务 [2]。

近年来，许多学者通过一些复杂的算法进行学习行为的数据分析和学习成绩预测。利用这种方法不仅可以使学生可以更好地了解自己在学习过程中的表现，及时调节学习策略，提高学习成效 [3]，还可以帮助教师评估学生学习效率，改进教学方法，并且对学生提出更加有效的干预措施和警告。此外，利用数据挖掘技术分析结果，用于不断优化改进 MOOC 等在线学习平台的功能，提高在线教育平台利用效果，提升在线学者和教师的满意率 [4]。

1 研究现状

1.1 学习行为分析

学习行为分析方向有理论研究和实际应用两个方向，理论研究方面 Malcolm[5] 提出的学习分析框架涵盖数据收集、分析、学习者、利益关联方及干预措施五大组成部分。学生能够基于自身学习数据进行自我调整与提升，而教学管理者则可据此完善教学决策，进而提升教学管理水平 [6]。实际应用方面，Fan J 等人 [7] 结合线上教学产生的多类不同数据，设计了一种基于多头注意力机制的视频推荐算法，旨在通过算法系统可以提供更加个性化的视频推荐内容。

1.2 成绩预测研究分析

在成绩预警方向上也同样分为理论研究和实际应用两个方向。理论研究方面，Gou J 等人 [8] 通过对公共在线教育平台的学生学习行为和成绩进行深入研究，探究影响最终成绩的组成元素，并据此构建了成绩预测模型。实际应用层面，Qiu L 等人 [9] 提出了一种基于卷积神经网络的辍学预测模型，该模型将特征提取和分类任务整合到一个统一的框架中。Mueen 等人 [10] 通过把学生过去学期的学习成绩以及曾经在论坛发表评论情况和阅读论坛内容情况输入到朴素贝叶斯算法 [11]、决策树算法来预测学生的最终学习成绩。

教育方面使用大数据算法进行成绩预测和评估学习状态，与传统算法相比具有明显优势 [12]。当前学习分析和学业预警方法虽然取得了一些进展，但仍面临一些挑战。例如在算法选择上，多数研究采用回归分析方法，少数文献采用了神经网络算法，缺乏对多种算法的比较。而且，在预测学习成绩之后，仅有少部分能够进一步分析学习者的群体行为特征，并据此提出有效的干预策略。

因此，本文旨在通过机器学习和大数据分析等人工智能技术，改进学习分析和学业预警方法。首先，利用学生学习行为数据建立预测模型，并通过比较多种算法，找出最有效

1. 西安文理学院信息工程学院 陕西西安 710065

[基金项目] 基于智能识别的大学生学习状态干预机制研究（编号：205230004）

的模型，用于预测学生未来的学习表现[13-14]；然后，运用聚类分析方法，对学生数据进行深入挖掘[15]，并与预测出的学业成绩进行关联分析。通过这种方式，可以识别出学生的群体性学习特征。根据这些特征，可以为教学反馈和个性化辅导提供更有效的干预措施。

2 算法模型

2.1 预测算法模型

2.1.1 总体模型

本文的预测算法流程图如图 1 所示。首先，利用卷积神经网络（CNN）来提取深层学生行为特征。卷积操作可以提取数据空间特征，并且能够利用最大池化选择出学习行为特征中最显著的特征，保留住重要信息。然后利用这些特征作为输入传递给长短期记忆网络（LSTM），用于学生成绩的预测。LSTM 能够捕捉到学生行为特征的时序依赖关系，并有效预测学生成绩[16]。最后，将 LSTM 的输出作为注意力机制的输出，通过该注意力机制，对不同的学生行为特征以及时间序列分配注意力权重，从而更加准确地预测学生成绩。

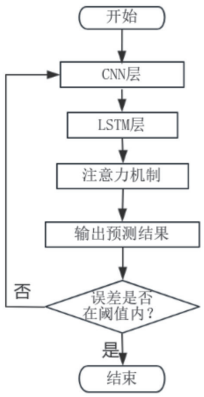

图 1 预测算法流程图

2.1.2 CNN

CNN 层利用 CNN 提取数据中不同特征之间的空间结构，样本数据进入 CNN 层中会依次进行卷积操作、池化操作。针对本文数据集特点，模型使用一维卷积结构，即按照时域方向进行卷积。卷积层公式为：

$$C = f(X \otimes W + B) \tag{1}$$

式中：C 为经过卷积核后的特征向量，X 属性特征，$\otimes$ 为卷积运算；W 为卷积核的权重向量；B 为偏移量；$f(\cdot)$ 为非线性激活函数。特征数据经过上述卷积操作之后，进入最大池化层，压缩数据维度，并且自动筛选出影响学生成绩的学生行为显著特征。

2.1.3 LSTM

LSTM 的单元结构见图 2。长短时记忆网络（long short-term memory，LSTM）模型是一种特殊的 RNN（循环神经网络），它可以有效地处理长序列数据，并解决了传统 RNN 存在的梯度消失和梯度爆炸问题。LSTM 网络包含一个单元状态和三个门限结构，分别是输入门（input gate）、遗忘门（forget gate）和输出门（output gate）。这些门可以控制信息在网络中的流动，从而实现对长序列数据的处理。

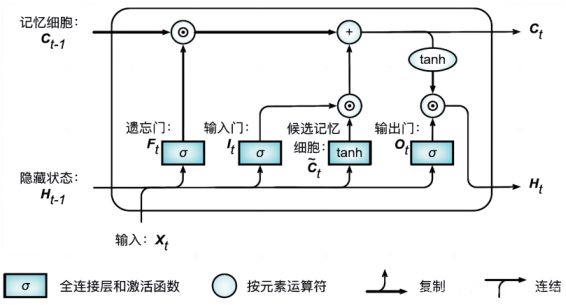

图 2 LSTM 的单元结构

2.1.4 注意力机制

为了提高预测精度，在预测模型中引入了注意力机制，强调重要因素的影响。注意力层结构见图 3，它将 LSTM 层的输出作为输入，并通过计算序列中不同时间步之间的关系来加权重要的信息。

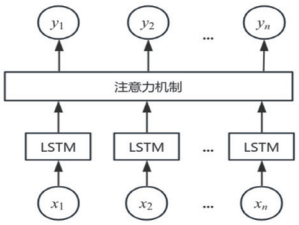

图 3 注意力层结构

$$s_i = \tanh(\mathrm{w}h_i + b_i) \tag{2}$$

$$\partial_i = \mathrm{Softmax}(s_i) = \frac{\exp(s_i)}{\sum_{t=1}^{N}\exp(s_t)} \tag{3}$$

$$c = \sum_{i=1}^{N}\partial_i h_i \tag{4}$$

式中：h_i 为第 i 组子变量 LSTM 特征向量输出；Softmax 函数将权重进行归一化处理，分别计算不同特征在 t 时刻隐藏状态的注意力得分；c 为所有隐藏状态及其相应注意权重的加权和结果。

在经过注意力加权后的数据进入全连接层，最终在输出层给出预测结果。使用 Sigmoid 激活函数。Sigmoid 函数将输入映射到 0 和 1 之间的一个值，表示样本属于正类的概率。

3 基于改进 K-means 的学生聚类分析

3.1 结合 Canopy 算法的 K-means 聚类

由于传统 K-means 聚类算法[17]是随机选取 K 值和初始聚类中心，而 K-means 算法的结果取决于初始聚类中心的选择，不同的初始中心可能会导致不同的最终聚类结果，因此影响聚类的效果及稳定性。为了解决这一问题，本文采用结合 Canopy 算法的 K-means 聚类[18]。使用 Canopy 算法进行预聚类，得到初始聚类中心和最优聚类数目。这样的好处是可以避免 K-means 聚类算法中随机初始化带来的聚类结果不

稳定，提高聚类的可靠性和准确度。

3.1.1 本文聚类的算法流程

本文使用的聚类算法流程图见图 4，整体算法步骤分为以下几步。（1）将原始样本进行数据清洗以及标准化处理。（2）利用皮尔逊系数进行相关性分析，得到与成绩相关性较强的特征数据。（3）对在线学习行为数据使用自编码器降维，得到输入的低维特征表示。（4）使用 Canopy 确定初始聚类中心与 K 值。（5）使用 K-means 进行数据集的聚类，并对聚类结果进行性能评估。

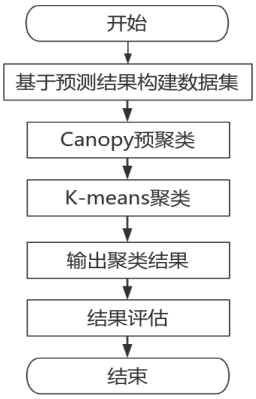

图 4 本文聚类算法流程图

4 实验结果

4.1 数据集来源与预处理

本研究使用的数据来源于某高校学生信息管理系统大一至大四部分学生在线学习数据，通过问卷调查方式获得的课程评价数据，以及学生上学期智育成绩排名。将以上数据进行整合，数据集分为学生基本信息、学生平台互动信息、学生学习背景信息、学生课程评价信息四类属性，每个所包含的特征见表 1。

表 1 学习行为数据

属性类别	特征	内容
学生基本信息	学号	—
	性别	男性 / 女性
	年级	大一至大四
	出生地	精确到省份
平台互动信息	在线回答问题次数	0 ～ 1000，整数
	观看教学视频次数	0 ～ 1000，整数
	评论次数	0 ～ 1000，整数
	登录系统次数	0 ～ 1000，整数
	签到次数	0 ～ 1000，整数
学生背景信息	截至目前请假及旷课次数	0 ～ 10，整数
	测验平均成绩	0 ～ 100，整数
课程评价信息	是否填写评价	是、否
	评价本次课程等级	0 ～ 10 分，分值越高越满意

为了保证数据一致性，在实验开始之前进行数据清洗和剔除。把从平台导出的数据集中存在重复信息的数据、存在明显错误的数据、由于学生转专业或休学等原因导致不完整的数据直接进行处理或剔除。

采集到的原始数据共 1659 条记录，其中 156 条记录中存在缺失值，剔除含缺失值的数据，最后保留 1503 条有效记录。

数据清洗之后，为避免数据取值范围的差异对后续实验的干扰，首先对数据集中的一些非数值型数据进行特征编码哑变量处理，将其转换为数值型特征，再对数值型特征使用式（5）进行 StandardScaler 标准化，标准化公式为：

$$X' = \frac{X - \mu}{\sigma} \tag{5}$$

式中：X 为清洗后的学习数据；X' 为标准化后的负荷数据；μ、σ 分别为样本数据的均值、标准差。

4.2 相关性分析

为进一步明确样本多维数据特征中数据特征与目标变量之间是否存在相关性，本文在数据集使用皮尔逊相关系数确定变量间相互关系。借助 SPSS 工具，对不同属性与学业成绩之间的相关性进行评估，利用皮尔逊相关系数的大小对它们进行排序。属性评估方法的排序结果见表 2。第 1 列是根据属性的相关系数进行排序，第 2 列则是对应属性的名称。根据表 3 相关系数的数值大小，可以判断前 9 项是影响学业成绩的主要因素，将其作为预测建模的自变量。

表 2 皮尔逊相关性系数表

排序	属性
0.577**	在线回答问题次数
0.532**	观看教学视频次数
0.527**	截至目前请假及旷课次数
0.520**	评论次数
0.498**	登录系统次数
0.435**	测验平均成绩
0.308**	评价本次课程等级
0.281*	签到次数
0.256*	是否填写评价
0.109	年级
0.081	学院
0.072	出生地
0.025	学号

（注：* 表示显著性水平小于 0.05，** 表示显著性水平小于 0.01）

4.3 在线学习行为统计

数据集为 16 周学生学习数据，为了利用数据集的时序特征，将数据集以周为频率，统计各个属性特征的频率，用于后续成绩预测特征描述见表 3。在预测成绩时，将最终成绩分为两类：成绩合格（大于等于 60 分）、成绩不合格（小于 60 分）。

表 3 预测模型数据集

特征序号	属性
0	学号
1	周期
2	每周在线回答问题次数
3	每周观看教学视频次数

表 3（续）

特征序号	属性
4	每周请假及旷课次数
5	每周评论次数
6	每周登录系统次数
7	每周测验平均成绩
8	每周签到次数

4.4 成绩预测

成绩预测算法见图 5。数据经过预处理和皮尔逊相关系数分析，通过一维卷积神经网络 (CNN) 提取学生行为数据的空间特征，然后经过长短期记忆网络（LSTM）捕捉行为数据之间的时序依赖关系，利用注意力机制动态分配权重，预测学生成绩。将训练集的数据输入模型进行训练，CNN 层卷积核大小设为 2，池化层中 stride=1，LSTM 分为三层其后连接 Flatten 层，Flatten 层起到加强学习数据特征的作用，损失函数为 Sigmod 函数。训练次数为 100 次梯度下降的方式采用 adam 训练算法。

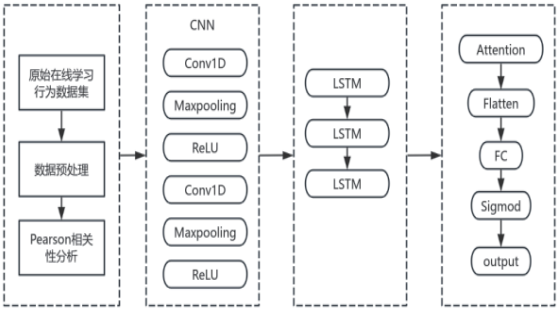

图 5 预测算法流程图

实验中预选数据样本共 1503 条，按照 8:2 的原则随机分割选取训练集与测试集。本文实验环境和参数见表 4。

表 4 实验环境和参数

环境	参数
计算机处理器	Intel(R) Core(TM) i7
运行内存	16 GB
编程语言	Python
开发框架	PyTorch
GPU	NVIDIA
数据集	600×9（总计 600 个学生数据样本，根据特征筛选成果，将筛选后的特征与考试成绩输入模型网络，每一个样本维数为 9）

为了探究本文方法的性能，将 SVM 二分类模型、单层 LSTM 分类模型、LSTM 加入注意力机制二分类模型与本文算法在精确率和召回率指标上进行对比。从表 5 中可以看出，在综合考虑评估指标的基础上，本研究设计的网络模型算法展现出最佳的性能。本文使用的预测算法的预测精确率达到 95.69%，精确率为 82.16%，优于传统机器学习算法和其他算法。这表明本文的模型在预测学生学习成绩方面具有较高的精度和召回率，为识别和帮助那些面临学业挑战的学生提供了有效的工具和支持。

表 5 预测结果

model	准确率 /%	召回率 /%
SVM	71.85	71.59
LSTM	74.36	73.17
LSTM+ 注意力	87.91	79.24
本文	95.69	82.16

通过结合 CNN、LSTM 和注意力机制，本文的模型能够充分利用学生行为数据的特征，实现更准确的学生成绩预测。这种方法可以有效地解决手动特征提取和模型融合的问题，提高预测的准确性和性能。

4.5 聚类分析

首先基于前文处理好的数据集及成绩预测结果构建新的数据集，共 1503 条 10 列数据；接着使用 Canopy 确定初始聚类中心与 K 值，作为下一步的初始参数；最后使用 K-means 进行聚类，并对聚类结果进行性能评估。

当聚类中心为 4 时，使用传统 K-means 聚类方法、结合 Canopy 的 K-means 算法、本文的聚类方法在戴维森堡丁指数（DBI）、SSE 进行对比。

结果如表 6 所示。这两个评价指标上本文算法都表现更好，说明它在聚类质量上都优于传统的 K-means 算法以及基于 Canopy 的 K-means 算法。因此，在本文使用的特定数据集上，使用本文算法进行聚类会得到更好的结果。聚类结果见图 6。

表 6 聚类结果对比

	K-means(K=4)	Canopy-Kmeans(K=4)	本文
DBI	1.952	0.886	0.849
SSE	309.641	258.322	147.449

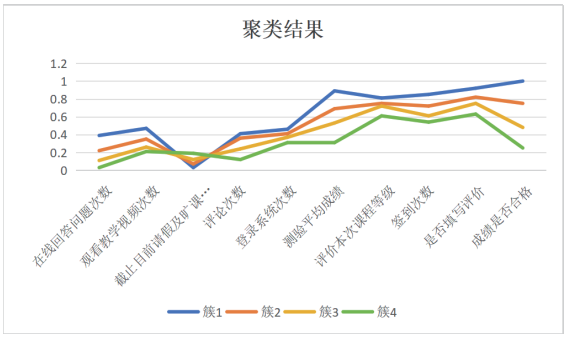

图 6 实验结果

簇 1 代表表现出色型学习者，占学生总数的 22.32%。相较于其他学生类型，这些学生在学习行为上表现得更加优异。簇 2 代表合格型学习者，占学生总数的 57.24%，占比较大，

说明有更多的学生符合这类特征。簇 3 所代表的学生类型可以被描述为普通学习者，占比约为 12.41%。相对于其他类型，这些学生在学习行为上呈现出一般水平，缺乏学习动力和积极性。簇 4 为风险学习者，占有效学习者的 8.03%。这个学生群体在各项学习行为上都表现最差，与平均水平存在较大差距。

5 总结与展望

实验结果表明，本文算法能够较准确地预测学生成绩，预测精度能够达到 95.69%，与其他预测方法相比，本文预测效果更好。基于聚类分析的结果，设计了具有针对性的教学指导策略，旨在对学生进行前瞻性的精确干预。

在未来的研究中，将着眼于综合考虑多个来源的相关因素，以提高成绩预测模型的可靠性和准确性。构建更为细致和全面的学习者画像，从而更好地理解学生成绩背后的复杂因素，提高成绩预测模型的精度和稳定性。

参考文献：

[1] 邢西深 , 李军 . “互联网 +”时代在线教育发展的新思路 [J]. 中国电化教育 , 2021(5):57-62.

[2]DURAIRAJ M, VIJITHA C.Educational data mining for prediction of studentperformance using clustering algorithms[J].International journal of computer scienceand information technologies,2014,5(4):5987-5991.

[3]ROMERO C, VENTURA S. Educational data mining and learning analytics:an updatedsurvey[J]. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2020, 10(3): 1-21.

[4] 陈子健 , 朱晓亮 . 基于教育数据挖掘的在线学习者学业成绩预测建模研究 [J]. 中国电化教育 ,2017(12):75-81+89.

[5]MALCOLM B.Learning analytics: the coming third wave[EB/OL].(2011-04-15)[2024-04-20].http://net.educause.edu/ir/library/pdf/ELIB1101.pdf.

[6]PENG H, MA S, SPECTOR J M.Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment[J].Smart learning environments, 2019, 6(1):1-14.

[7]FAN J, JIANG Y, LIU Y, et al.Interpretable MOOC recommendation: a multi-attentionnetwork for personalized learning behavior analysis[J]. Internet research, 2022, 32(2): 588-605.

[8]GOU J, QIN Y, LUO Y, et al.Prediction of learning performance in online course based onlinear regression model: IEEE ITAIC(ISSN: 2693-2865)[C]//2022 IEEE 10th JointInternational Information Technology and Artificial Intelligence Conference.Piscataway:IEEE,2022, 10:1979-1983.

[9]QIU L, LIU Y, HU Q, et al.Student dropout prediction in massive open online courses byconvolutional neural networks[J].Soft computing,2019,23(20):10287-10301.

[10]MUEEN A, ZAFAR B, MANZOOR U.Modeling and predicting students' academic performanceusing data mining techniques[J]. International journal of modern education and computerscience, 2016,11(11):36-42.

[11]SNOEK J, LAROCHELLE H, ADAMS R P. Practical bayesian optimization of machine learning algorithms[C]// International Conference on Neural Information Processing Systems.Lake Tahoe: Curran Associates Inc.,2012:2951-2959.

[12] 胡艺龄 , 顾小清 , 赵春 . 在线学习行为分析建模及挖掘 [J]. 开放教育研究 , 2014, 20(2):102-110.

[13]ZHAO C, MI C.A study on the differences of attitude learning and learning behavior sequences for university computer MOOC+ SPOC Course[C]//2020 15th International Conference on Computer Science & Education (ICCSE).Piscataway:IEEE,2020:365-369.

[14]ŠARIĆ-GRGIĆ I, GRUBIŠIĆ A, ŠERIĆ L, et al.Student clustering based on learningbehavior data in the intelligent tutoring system[M]//Research Anthology on RemoteTeaching and Learning and the Future of Online Education.Hershey:IGI Global, 2023: 785-803.

[15] 王全民 , 张书军 . 基于 Canopy-K-means 算法的高校贫困生预测的研究 [J]. 计算机与数字工程 ,2020,48(12):3012-3016+3041.

[16] 杨丽 , 吴雨茜 , 王俊丽 , 等 . 循环神经网络研究综述 [J]. 计算机应用 , 2018,38(S2):1-6+26.

[17] 李召鑫 , 孟祥印 , 肖世德 , 等 . 基于 Flink 框架的 K-means 算法优化及并行计算策略 [J]. 计算机与数字工程 , 2023, 51(10): 2231-2235.

[18] 张珂嘉 , 黄树成 . 一种改进的 K-means 入侵检测算法 [J]. 计算机与数字工程 , 2021,49(10):1963-1966+2047.

【作者简介】

安梦蕾（1996—），通信作者（email: 411483985@qq.com），女，河北衡水人，硕士，助教，研究方向：图像处理。

韩蒙（1997—），女，陕西临潼人，硕士，助教，研究方向：生命教育。

任绒（1995—），女，陕西渭南人，硕士，助教，研究方向：高校思想政治教育。

柯程虎（1985—），男，陕西西安人，博士，讲师，研究方向：无线光通信。

常永明（1983—），男，陕西西安人，博士，讲师，研究方向：人工智能。

（收稿日期：2024-06-06）