DOI:10.13878/j.cnki.jnuist.20230421001

路凯丽[1]　杨露[1]　李涛[1]

基于集成深度学习模型的空气质量指数预测

摘要

大气污染严重危害居民的出行安全和身体健康,空气质量指数(AQI)是一种用于测量空气质量状况的综合指标,对AQI的预测可以提醒公众空气质量信息,使人们做出更明智的出行决策.通过提前预测空气质量的变化,政府和环保部门可以采取应急措施以减轻空气污染.本文提出基于卷积神经网络和门控循环单元的集成深度学习模型(CNN-GRU)对AQI进行预测.其中,卷积神经网络(CNN)提取污染气体浓度和AQI的时空特征并完成特征映射,门控循环单元(GRU)建模时序关系并高效完成计算.选取2014—2022年北京市和广州市的6种主要污染气体($PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3)日平均质量浓度和AQI进行实例研究,使用CNN-GRU模型对AQI进行预测,与多元宇宙优化的广义回归神经网络模型(MVO-GRNN)、遗传算法优化的BP神经网络模型(GA-BP)对AQI的预测进行对比分析.实验结果表明,本文提出的CNN-GRU模型对AQI的预测误差最小.

关键词

空气质量指数;卷积神经网络;门控循环单元;集成模型

中图分类号 TP183

文献标志码 A

收稿日期 2023-04-21

资助项目 国家社会科学基金青年项目(20CTJ008);全国统计科学研究重点项目(2021LZ28);陕西省自然科学基金项目(2022JQ-042)

作者简介

路凯丽,女,硕士生,研究方向为大数据分析与应用.lu200008@126.com

杨露(通信作者),女,博士,副教授,研究方向为大数据分析与应用.yanglu20497@163.com

1　西安财经大学　统计学院,西安,710100

0　引言

大气污染对国家可持续发展和居民健康带来不利影响.大气污染物浓度增加会显著提升呼吸系统疾病、心血管疾病和肺癌等的发病率和死亡率[1],威胁人们的出行安全和身体健康.除此之外,大气污染物抑制农作物生长,阻碍铁路交通发展,危害巨大.《环境空气质量指数(AQI)技术规定(试行)》(HJ 633—2012)将空气质量指数共分为6级,对应优、良、轻度污染、中度污染、重度污染和严重污染6种空气质量,每种空气质量状况对健康影响不同,分别具有相应的建议和措施以保护公众健康.一般来说,较低的AQI值表示较好的空气质量,较高的AQI值表示较差的空气质量.对AQI进行预测可以帮助人们提前了解大气污染情况,采取相应的健康防护措施.同时,对AQI进行预测可以提前警示大气污染的发生,有助于政府相关管理部门有针对性地制定应急措施和环保措施,例如交通管制、限制工业排放、停工停产等.对AQI的预测还可以用于评估和监控大气污染治理措施的效果,并指导进一步的环境管理和决策制定.

目前,许多学者利用传统的统计方法[2]和机器学习算法[3]对AQI进行预测.焦东方等[4]利用多元回归分析法建立了AQI与$PM_{2.5}$、PM_{10}、SO_2、O_3质量浓度的回归模型,对青岛市每日空气质量指数进行预测.Shishegaran等[5]构建了差分整合移动平均自回归模型(Autoregressive Integrated Moving Average, ARIMA)、主成分回归模型(Principle Component Regression, PCR)、ARIMA-PCR组合模型、ARIMA和基因表达编程(GEP)的组合模型对每日空气质量指数进行预测,并使用平均绝对百分比误差(MAPE)、均方根误差(RMSE)和归一化均方误差(NMSE)等指标来评估和比较模型预测效果.Phruksahiran[6]提出一种将机器学习算法和地理加权预测因子方法(GWP)结合的集成预测方法,包括RF-GWP方法、XG-GWP方法、NN-GWP方法,分别对城市空气质量指数进行预测.但是传统的统计方法无法准确捕捉空气质量中复杂的非线性关系和时序模式,易受异常值和噪声的影响,传统机器学习算法无法自动学习时序特征表示,对特征工程存在依赖性.

卷积神经网络(Convolutional Neural Networks,CNN)可以有效克服上述缺陷.CNN作为深度学习的代表性神经网络模型,可以很好地完成非线性拟合[7]并自动提取特征.孙启森等[8]采用引入注意力机制

的卷积神经网络模型对金融时间序列数据进行预测.袁培森等[9]利用神经网络技术对菊花的原始图像进行逐层特征学习,从而实现对该类农产品花型和品种的高效识别.Shujaat 等[10]构建了基于 CNN 的 pcPromoter-CNN 工具对 DNA 结构中的启动子进行分类和预测.虽然 CNN 具有自动提取特征和权值共享的优势[11],但是其训练过程通常需要较大的计算资源和时间,而门控循环单元(Gated Recurrent Unit,GRU)作为循环神经网络模型(Recurrent Neural Networks,RNN)的变体[12],只使用两个门控开关,具有模型的简单性和计算的高效性.因此,本文构建 CNN-GRU 集成深度学习模型,在处理城市空气质量时间序列数据时可以同时进行特征提取和序列建模,避免了梯度消失和梯度爆炸问题,在小模型参数下有更好的表现能力,对 AQI 的预测效果更好.

1 基本原理

1.1 卷积神经网络

CNN 是在视觉系统结构的启示下开发的一种具有深度结构和卷积运算的前馈神经网络,是深度学习的代表算法之一,擅长非线性数据的特征提取,广泛应用于时序预测[13]、图像分类[14]、故障诊断[15]等方面.CNN 包括输入层、卷积层、池化层、Flatten 层、全连接层和输出层.

输入层是 CNN 的起始层,负责接收、预处理原始输入数据和网络配置.输入数据可以是图片数据、文本数据、表格数据、传感器数据或语音信号数据,输入层接收数据后对其进行预处理,如图像归一化、对语音进行傅里叶变换等.同时,输入层设置输入数据的尺寸、通道数和类型等配置参数,为卷积层提供合适的输入数据.

卷积层是 CNN 的核心层,主要负责特征提取,具有局部连接、参数共享和尺度不变性的特性.卷积层通过卷积核对每个通道的矩阵依次运算来提取输入数据中有用的特征表示,为池化层提供更丰富抽象的输入数据.卷积层采用局部连接方式,可以捕捉输入数据的局部特征.卷积层中的卷积核在输入数据上共享参数,极大地减少 CNN 的参数量,提高了 CNN 的计算效率.卷积操作[16]的数学表示为

$$\boldsymbol{h}^n = f\left(\sum_{i\in \boldsymbol{M}} \boldsymbol{w}_i^n \boldsymbol{x}_i^{n-1} + \boldsymbol{b}^n\right), \tag{1}$$

式中:$\boldsymbol{h}^n$为第 n 层输出;$\boldsymbol{M}$为输入特征矢量;$\boldsymbol{w}_i^n$为第 n 层卷积核;$\boldsymbol{x}_i^{n-1}$为第 n 层输入;$\boldsymbol{b}^n$为第 n 层偏置参数.

池化层是 CNN 中的常用层,负责尺寸缩减、特征保留和特征降维,具有平移不变性.池化层对输入特征映射进行下采样,减小特征映射的尺寸,有利于降低 CNN 计算复杂度,防止过拟合.池化层对局部区域特征进行聚合,通过最大值池化或平均值池化保留输入特征映射中的主要特征信息,降低了特征映射的维度.

Flatten 层在 CNN 中位于池化层和全连接层之间,主要用于将池化层输出的多维数据转换为 1 维向量,以便输入到全连接层进行分类或回归等任务.

全连接层主要负责特征融合、参数学习和非线性关系学习,层中使用激活函数引入非线性关系,增加 CNN 的表达能力.全连接层包含的可训练的权重和偏置参数在训练过程进行优化,使 CNN 具有最优的特征表示.全连接层将高级特征表示转换为最终的分类或回归结果并传入到输出层,输出层负责产生 CNN 的最终预测结果.

1.2 门控循环单元

GRU 是循环神经网络的变体,用于处理序列数据.GRU[17]具有门控机制,由更新门和重置门两个门类组成,可以有效地解决传统 RNN 中的长期依赖和梯度爆炸问题.GRU 相比 RNN 的另一个代表变体长短期记忆递归神经网络(LSTM)[18]结构更简便,计算速度和更新速度更快,其结构原理如图 1 所示.

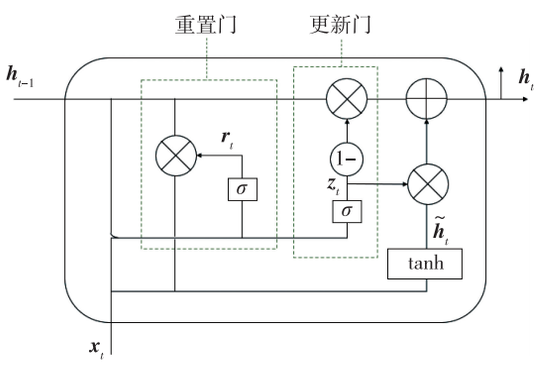

图 1 GRU 结构原理

Fig. 1 Schematic of GRU structure

更新门控制当前时间步的输入是否更新到隐藏状态,让模型自动决定当前时间步的输入对于更新隐藏状态的重要性,从而可以选择性地更新或保留之前的隐藏状态.重置门控制之前隐藏状态和当前输入之间的信息传递,让模型自动决定是否将之前隐藏状态中的信息重置,从而可以选择性地遗忘或保留之前的信息.GRU 通过引入更新门和重置门,可以灵活地控制对输入和隐藏状态的信息更新和遗

忘,从而更好地捕捉长期依赖关系.其数学模型如下:

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_{xz} \cdot \boldsymbol{x}_t + \boldsymbol{W}_{hz} \cdot \boldsymbol{h}_{t-1}), \tag{2}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_{xr} \cdot \boldsymbol{x}_t + \boldsymbol{W}_{hr} \cdot \boldsymbol{h}_{t-1}), \tag{3}$$

$$\boldsymbol{g}_t = \tanh[\boldsymbol{W}_{xh} \cdot \boldsymbol{x}_t + \boldsymbol{r}_t * (\boldsymbol{W}_{hh} \cdot \boldsymbol{h}_{t-1})], \tag{4}$$

$$\boldsymbol{h}_t = \boldsymbol{z}_t * \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) * \boldsymbol{g}_t, \tag{5}$$

式中:z 为更新门,r 为重置门,g 为重置信息,h 为隐藏层,W 为权重,x 为输入信息,t 为时刻,$*$ 表示线性变换.

1.3 CNN-GRU 模型预测流程

CNN-GRU 集成深度学习模型对 AQI 的预测流程包括空气质量数据的输入、CNN 网络的特征提取、GRU 网络的时序计算和 AQI 预测值的输出,预测流程如图 2 所示.

CNN-GRU 集成深度学习模型对 AQI 的预测步骤如下:

1)空气质量数据的输入

输入数据为城市每日的 $PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3 质量浓度值和 AQI 值.由于收集的数据中存在缺失和各指标量纲不同的问题,本文使用插值函数处理缺失值,并对变量进行归一化处理.然后,对数据集划分训练集和测试集后进入 CNN 网络.

2)CNN 网络的特征提取

在 AQI 预测中,CNN 通过卷积操作和非线性激活函数,学习输入数据不同层次的抽象特征表示,从城市空气质量时间序列数据中提取有用的空间模式和局部特征.AQI 数据具有空间分布特征,CNN 可以有效地捕捉空间上的相关性.通过使用卷积层和池化层,CNN 可以识别局部的模式和结构,并对其进行下采样,从而减少参数数量并保留重要的空间信息.城市空气质量数据的空间维度通过池化层后降低,减少了模型的复杂度和计算量,有利于处理大规模空气质量数据并减轻过拟合风险.CNN 通过共享权重的方式,将学到的特征在整个数据集传递,在不同时间步上共享相同的特征提取器,使得模型能够学习到数据的通用特征,并更好地泛化到新的样本上.

3)GRU 网络的时序计算

GRU 通过记忆单元和更新门控制信息的传递和遗忘,使得网络能够有效地处理时间上的相关性.在 AQI 预测任务中,GRU 网络可以利用过去时间步的信息来预测未来时间步的 AQI 值,并且能够自适应地学习序列中的模式和趋势.

4)AQI 预测值的输出

在 CNN-GRU 模型预测 AQI 的过程中,输出层将前面层级中提取到的特征映射到最终的预测结果,得到最终的 AQI 预测值.

2 模型设计

2.1 实验环境

本文的实验环境为 Windows10 64 位操作系统,使用 Anaconda 的 Jupyter Notebook 进行编程,Python 解释器版本为 3.9.12,使用 TensorFlow 和 Keras 搭建神经网络模型,编程语言为 Python,同时使用 MATLAB 构建对照模型.

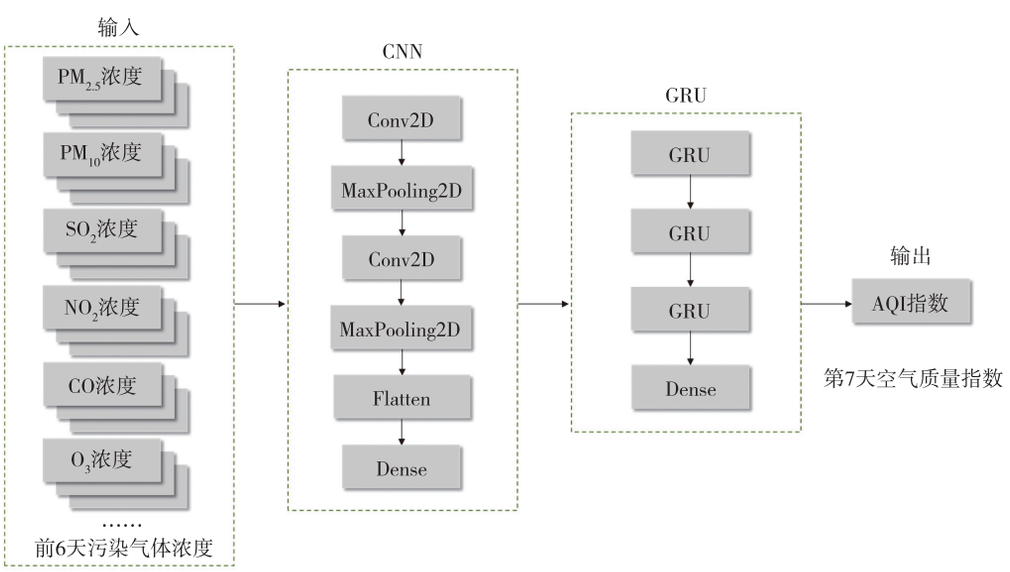

图 2 CNN-GRU 集成模型预测流程

Fig. 2 Processes of AQI prediction by CNN-GRU integrated model

2.2 参数设置

CNN-GRU 模型由输入层、CNN 网络、GRU 网络和输出层组成.输入层设置输入数据的数据类型为 32 位浮点数、数据尺寸为 6×6.采用移动窗口进行平滑训练和预测.CNN 网络设置两层卷积、两层池化、一个 Flatten 层和一个全连接层,参数设置如表 1 所示.第一卷积层中的卷积核尺寸为 2×2,通道数为 1,包含 32 个卷积核,使用 ReLU 激活函数.第二卷积层中的卷积核尺寸为 2×2,通道为 1,包含 64 个卷积核,同样使用 ReLU 激活函数.两层池化均采用最大池化函数进行池化操作.位于池化层和全连接层之间的 Flatten 层将多维输入数据展平为 1 维向量.然后连接隐层节点为 512 的全连接层,并设置 Dropout,防止模型过拟合.CNN 网络中损失函数采用交叉熵,优化器采用 Adam 优化算法.

表 1 CNN 模型参数

Table 1 CNN model parameters

网络层类型	尺寸	设置
卷积层 1	卷积核:2×2	激活函数:ReLU
池化层 1	池化核:2×2	池化方式:最大池化
卷积层 2	卷积核:2×2	激活函数:ReLU
池化层 2	池化核:2×2	池化方式:最大池化

本文使用 Keras 神经网络框架创建 GRU 网络,使用序贯 Sequential 模型对象依次添加 3 个 GRU 层和一个全连接层.其中 GRU 层分别为:一个带有 64 个单元、采用 ReLU 激活函数以及设置 return_sequences=True 的 GRU 层,一个带有 32 个单元、采用 ReLU 激活函数以及设置 return_sequences=True 的 GRU 层,一个带有 32 个单元、采用 ReLU 激活函数以及设置 return_sequences=False 的 GRU 层.设定 return_sequences 参数为 Ture,表示将该层结果作为下一层的输入.全连接层将 GRU 单元的输出映射成一个单一的输出值,即经过模型计算的 AQI 值.损失函数使用均方误差(Mean Squared Error,MSE),优化器选用 Adam,通过 model.compile()函数的调用,将损失函数和优化器配置到模型中,使得模型在训练过程中能够使用这些设置进行参数的更新和优化.调用 summary()函数输出 GRU 模型各层的参数状况,Param 数值如表 2 所示.模型设置权重变量和偏置变量,将多维向量转化为 1 维,输出最终 AQI 预测结果.

表 2 GRU 参数

Table 2 GRU parameters

图层(类型)	输出形状	参数/个
gru_1 (GRU)	(None,1,64)	110 784
gru_2 (GRU)	(None,1,32)	9 312
gru_3 (GRU)	(None,32)	6 240
dense_1 (Dense)	(None,1)	33

2.3 评价指标

本文采用平均绝对误差(MAE)、均方根误差(RMSE)、平均绝对百分比误差(MAPE)作为各模型预测 AQI 效果的评价指标,表达式如下:

$$e_{\mathrm{MAE}}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right|, \tag{6}$$

$$e_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}, \tag{7}$$

$$e_{\mathrm{MAPE}}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|, \tag{8}$$

式中,y_i 为 AQI 真实值,$\hat{y}_i$ 为 AQI 预测值,n 为样本个数.

2.4 超参数调优

当使用深度学习模型完成分类或回归任务时,存在一些超参数,它们控制模型的行为和性能且需要手动设置.超参数调优是通过尝试不同的超参数值来获取最佳超参数配置,优化模型的性能.本文对学习率、Epoch 大小、batch_size、dropout 率和步长进行超参数训练寻优,使用 MAE、RMSE、MAPE 3 个指标进行模型预测性能评估,从中选择 CNN-GRU 集成深度学习模型预测 AQI 最佳的超参数配置.

1)学习率

学习率(Learning Rate)用于控制深度学习模型在训练过程中更新参数的幅度.Learning Rate 的大小对于 CNN-GRU 模型的训练过程和性能起着重要的作用.Learning Rate 过大会导致模型在参数空间中跳过最优解,Learning Rate 过小会导致模型收敛速度慢,模型训练的时间成本高.本文分别对 0.1、0.01、0.001 和 0.000 1 的学习率进行测试,CNN-GRU 模型在不同学习率下损失函数的曲线如图 3 所示.

图 3 中橙色曲线代表训练集损失函数,蓝色代表测试集损失函数.如图 3 所示,当 Learning Rate 为 0.1 和 0.01 时,损失函数曲线均先快速下降后趋于不变,这说明学习率设置不合理.当 Learning Rate 为 0.000 1 时,训练集和测试集的损失函数曲线准确拟合.当 Learning Rate 为 0.001 时,训练集和测试集的

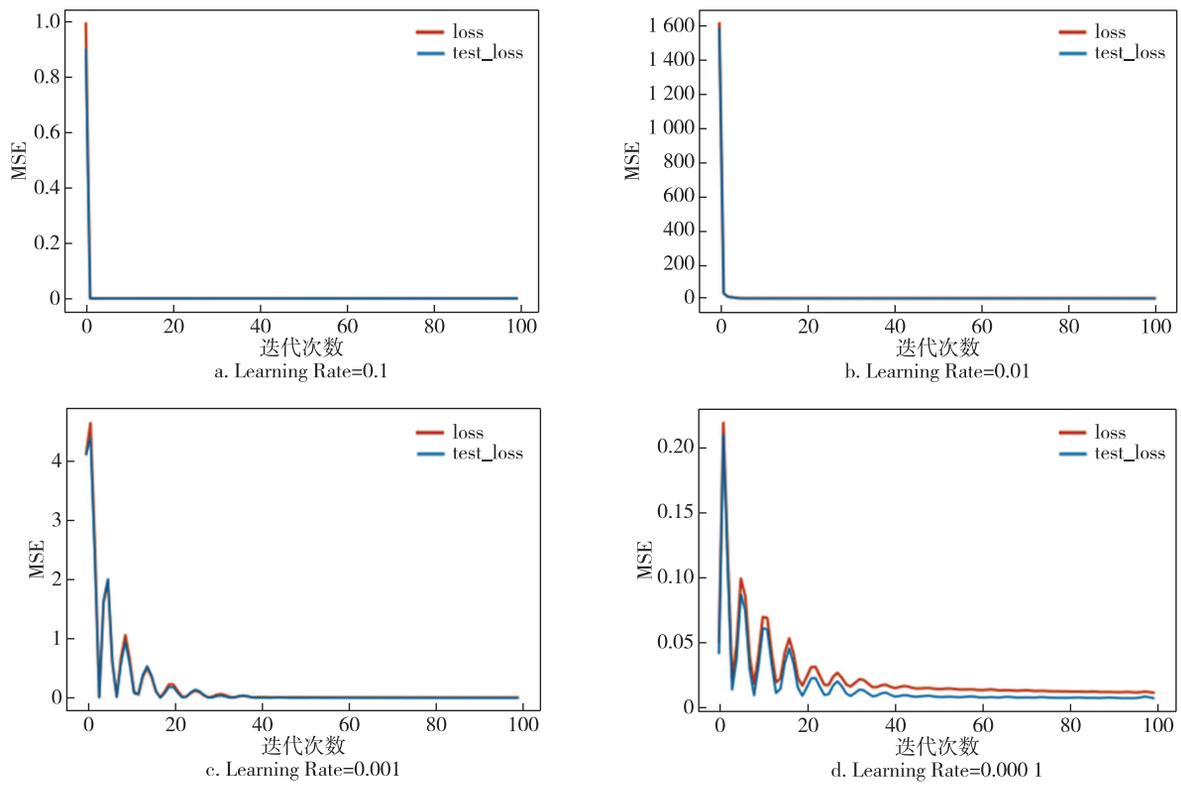

图 3　不同学习率下损失函数曲线

Fig. 3　Loss function curves with different learning rates

损失函数曲线高度拟合,说明 0.001 的学习率最优.分别使用 0.1、0.01、0.001、0.000 1 的学习率进行 CNN-GRU 模型训练,经过 3 次实验,对评价指标求均值得到表 3,验证了学习率为 0.001 时,CNN-GRU 集成模型预测 AQI 值误差最小,因此确定学习率为 0.001.

表 3　不同学习率下各评价指标情况

Table 3　Evaluation indexes under different learning rates

学习率	MAE	RMSE	MAPE/%	训练时长/s
0.1	0.067 9	0.140 3	37.58	54.76
0.01	0.048 6	0.076 8	24.72	54.28
0.001	0.048 1	0.071 5	23.95	51.39
0.000 1	0.049 0	0.080 2	24.84	55.64

2)Epoch

Epoch 是在训练神经网络时,将整个数据集通过神经网络一次的迭代次数.Epoch 过小会导致 CNN-GRU 模型欠拟合,无法充分学习污染气体数据集特征,Epoch 过大会导致 CNN-GRU 模型过拟合,对测试集泛化性能低.本文分别对 50、100、150、200、250、300 的 Epoch 进行测试,CNN-GRU 模型在不同 Epoch 下损失函数的曲线如图 4 所示.

图 4 中蓝色曲线代表训练集损失函数,橙色代表测试集损失函数.如图 4 所示,当 Epoch 为 50、100、150、250、300 时,测试集损失函数曲线均先下降,然后缓慢上升,出现过拟合现象,当 Epoch 为 200 时,测试集损失函数曲线稳定下降,下降趋势与训练集损失函数曲线相似,说明 Epoch 选择 200 最优.分别使用 30、50、100、150、250、300 的 Epoch 进行 CNN-GRU 模型训练,经过 3 次实验,对评价指标求均值得到表 4,验证了 Epoch 为 200 时,CNN-GRU 集成模型预测 AQI 的误差最小,因此确定 Epoch 为 200.

表 4　不同 Epoch 下各评价指标情况

Table 4　Evaluation indexes under different Epochs

Epoch	MAE	RMSE	MAPE/%	训练时长/s
30	0.051 3	0.081 4	26.65	46.77
50	0.049 1	0.087 0	24.21	53.23
100	0.049 0	0.077 9	25.00	76.46
150	0.047 7	0.076 4	24.24	77.71
200	0.047 0	0.074 6	23.58	93.96
250	0.048 3	0.074 9	25.31	110.54
300	0.047 8	0.076 3	23.71	124.48

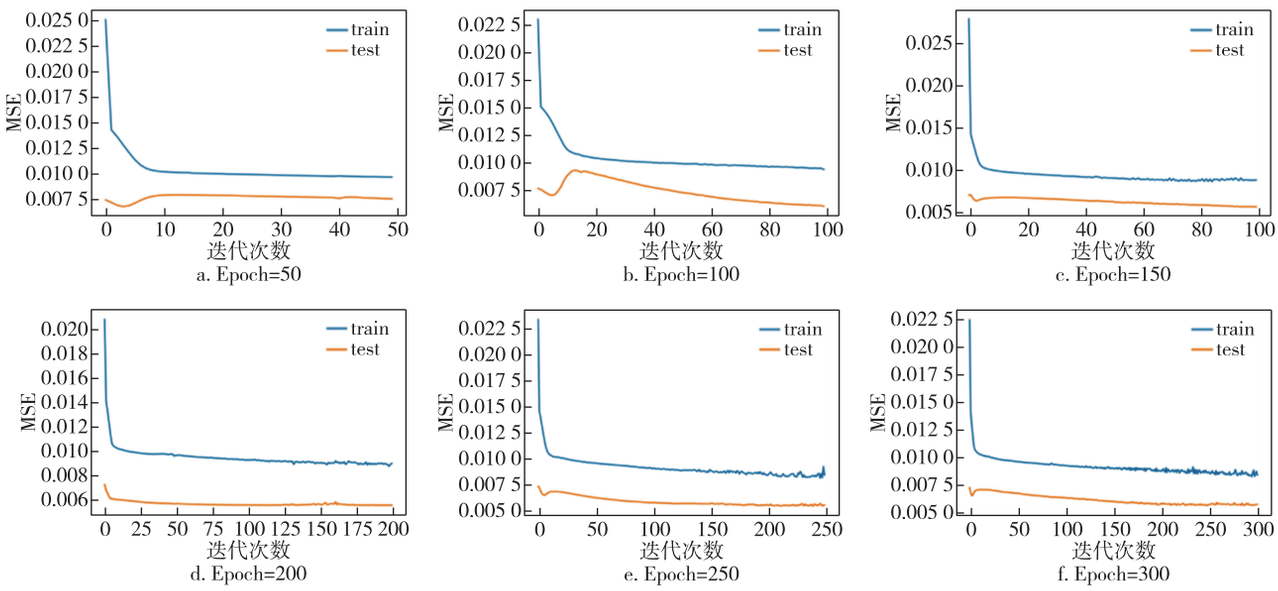

图 4　不同 Epoch 下损失函数曲线

Fig. 4　Loss function curves under different Epochs

3) Batch_size

Batch_size 是 CNN-GRU 集成模型在训练过程中一次处理的样本数量.分别使用 16、32、64 的 Batch_size 进行 CNN-GRU 模型训练,经过 3 次实验,对评价指标求均值得到表 5,验证了 Batch_size 为 32 时,CNN-GRU 集成模型预测 AQI 值误差最小,因此确定 Batch_size 为 32.

表 5　不同 Batch_size 下各评价指标情况

Table 5　Evaluation indexes under different Batch_sizes

Batch_size	MAE	RMSE	MAPE/%
16	0.048 1	0.074 9	24.62
32	0.047 7	0.076 0	24.12
64	0.048 9	0.076 6	24.92

4) Dropout

Dropout 是一种常用于深度学习模型中的正则化技术,通过引入随机性来防止模型过度依赖某些特定神经元,提高模型的泛化能力.在 CNN-GRU 集成模型中,Dropout 应用方式包括:在 CNN 层后面应用 Dropout、在 GRU 层前面应用 Dropout、在全连接层后面应用 Dropout.分别使用 0.7、0.5、0.3、0.2 的 Dropout 率进行测试,结果表明,在 GRU 层前面应用 0.2 的 Dropout,以及在 CNN 层后面应用 0.3 的 Dropout 率情况下,模型预测 AQI 的 MAE 为 0.046 5,RMSE 为 0.074 5,MAPE 为 0.232 2,相比其他超参数配置,此超参数配置最优.

5) 步长

在深度学习任务中,步长(Sequence Length)是指在进行卷积、池化等操作时,在输入数据上移动的时间步数量.步长决定输出的尺寸,过大的步长会导致信息丢失,过小的步长会增大计算量和过拟合风险,因此平衡步长的大小十分重要.本文分别对 1、6、12 的步长进行 3 组测试,求均值得到表 6.实验证明,当步长为 6 时,CNN-GRU 集成模型对 AQI 的预测误差最小,因此确定步长为 6.

表 6　不同步长下各评价指标情况

Table 6　Evaluation indexes under different steps

步长	MAE	RMSE	MAPE/%
1	0.086 4	0.108 9	30.06
6	0.047 1	0.073 7	23.76
12	0.087 2	0.108 4	29.87

3　基于 CNN-GRU 模型的 AQI 预测

3.1　数据收集

北京作为中国的北方城市,属于温带大陆性气候区,以其四季分明、夏季炎热而冬季寒冷的气候特征而闻名.广州位于中国南方,属于亚热带季风气候区,夏季炎热潮湿而冬季相对温暖湿润.通过对两个城市的空气质量进行分析和比较,探究 CNN-GRU 模型在不同气候特征下对空气质量指数的预测效果,为改善和管理城市空气质量提供科学依据.

本文采用2014年1月1日到2022年12月31日北京和广州的6种污染气体的日平均质量浓度数据、平均湿度、平均温度、每日空气质量指数进行实例分析,包含$PM_{2.5}$($\mu g/m^3$)、PM_{10}($\mu g/m^3$)、SO_2($\mu g/m^3$)、NO_2($\mu g/m^3$)、CO(mg/m^3)、O_3($\mu g/m^3$)、平均湿度(%)、平均温度(℃)和AQI值,数据来源于空气质量在线检测分析平台(https://www.aqistudy.cn/).除去缺失值和极端值,2022年北京和广州的AQI数据序列情况如图5所示.

3.2 数据预处理

2014—2022年北京$PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3、AQI分别存在27、67、1、1、1、18、6条缺失值,广州O_3存在12条缺失值.本文采用插值函数interpolate()来填补数据集中的缺失值,取数据框中缺失值的上一条数值和下一条数值的平均值代替原缺失值.

将数据集以7:3的比例划分训练集和测试集,即将前70%的污染气体质量浓度和AQI数据作为训练集,剩下30%的污染气体质量浓度和AQI数据作为测试集.同时,设置特定步长的移动窗口进行平滑移动训练和预测.

为了降低由于量纲不同导致的预测误差,本文采用离差标准化对污染气体质量浓度和AQI数据进行标准化处理,将数据映射到[0,1]之间,公式如下:

$$X' = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})}, \tag{9}$$

式中,X'为归一化后数据,X为原数据,$X_{\min}$为数据最小值,$X_{\max}$为数据最大值.

3.3 特征组合选择与模型对比

为研究特征组合对AQI预测效果的影响,并验证CNN-GRU模型是否为AQI预测的最优集成模型,本文分别在两组特征组合下使用多元宇宙优化的广义回归神经网络模型(MVO-GRNN)、遗传算法优化的BP神经网络模型(GA-BP)和CNN-GRU模型对2014—2022年北京、广州的AQI进行预测,使用MAE、RMSE、MAPE 3种指标评价预测效果.第一组特征包括$PM_{2.5}$、PM_{10}、SO_2、CO、NO_2和O_3,其预测结果如表7所示.第二组特征包括PM_{10}、SO_2、CO、O_3、平均湿度和平均温度,其预测结果如表8所示.

表7 第一组特征下AQI预测结果

Table 7 AQI prediction results under the first group of characteristics

城市	模型	MAE	RMSE	MAPE/%
北京	MVO-GRNN	15.701 2	26.559 6	48.66
	GA-BP	13.946 8	29.050 2	46.03
	CNN-GRU	0.048 1	0.071 5	23.95
广州	MVO-GRNN	8.254 3	9.661 3	59.87
	GA-BP	3.078 4	4.289 5	36.68
	CNN-GRU	0.084 5	0.107 3	29.30

实验结果表明,对于第一组特征组合($PM_{2.5}$、PM_{10}、SO_2、CO、NO_2和O_3),3种模型在AQI预测方面表现出不同的效果,CNN-GRU模型表现最好,显示出较低的误差值,说明该模型在预测AQI时具有较高的准确性.与第一组特征组合相比,第二组特征组合(PM_{10}、SO_2、CO、O_3、平均湿度和平均温度)的预测性能下降.第二组特征下CNN-GRU模型比MVO-

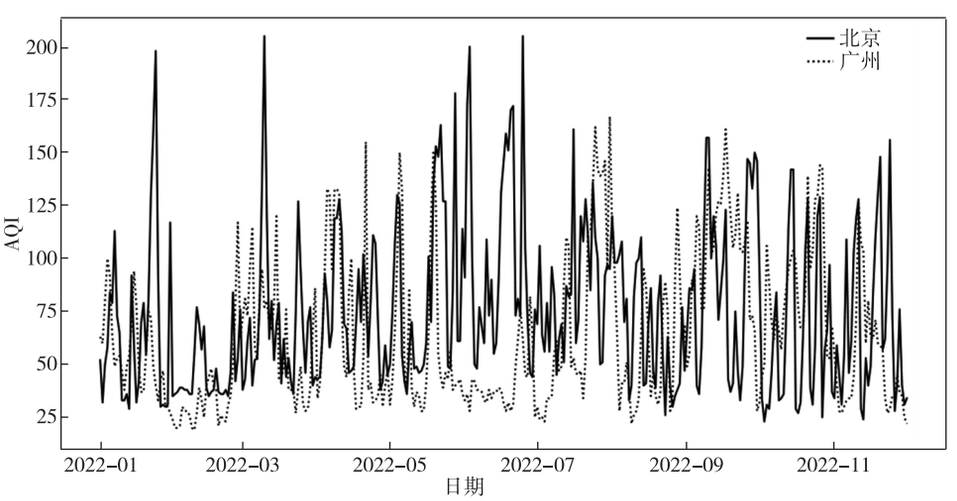

图5 2022年北京、广州的AQI数据序列

Fig. 5 AQI data series of Beijing and Guangzhou in 2022

表 8　第二组特征下 AQI 预测结果

Table 8　AQI prediction results under the second group of characteristics

城市	模型	MAE	RMSE	MAPE/%
北京	MVO-GRNN	47.629 9	53.789 7	150.37
	GA-BP	19.080 7	37.657 1	57.98
	CNN-GRU	0.064 0	0.091 9	33.29
广州	MVO-GRNN	20.302 4	24.702 0	51.24
	GA-BP	7.967 1	12.098 6	43.81
	CNN-GRU	0.121 8	0.145 7	42.68

GRNN 模型、GA-BP 模型表现更好,但相比第一组特征组合显示出相对较高的误差值.

综上所述,在 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2 和 O_3 的特征组合下 CNN-GRU 集成深度学习模型预测 AQI 的误差值最低,更能实现对城市空气质量指数的精准预测.

3.4　AQI 分季节预测情况

为验证 CNN-GRU 模型在分季节和小容量下的预测效果,对 2014—2022 年北京、广州的 AQI 和相关污染物数据以春(3—5 月)夏(6—8 月)秋(9—11 月)冬(12 月—次年 2 月)的划分准则[19]进行季节划分,结果分别如图 6 和图 7 所示.

使用 CNN-GRU 模型对北京的 AQI 预测时,春季、夏季、秋季、冬季预测的 MAE 分别为 0.057 8、0.057 8、0.053 0、0.051 3,RMSE 分别为 0.104 3、0.104 3、0.070 9、0.101 2,MAPE 分别为 0.246 7、0.246 7、0.245 1、0.233 2.相比之下,使用 GA-BP 模型和 MVO-GRNN 模型对相同数据进行分季节预测的结果则有所不同.MVO-GRNN 模型在不同季节下的预测误差值变化较大,不稳定.以 MAE 为例,春季、夏季、秋季、冬季预测的 MAE 分别为 17.632 9、8.809 5、16.690 7、20.453 4.而 GA-BP 模型的预测误差变化更为明显,春季、夏季、秋季、冬季预测的 MAE 分别为 18.515 0、1.976 7、9.592 1、11.506 0.这表明 GA-BP 模型和 MVO-GRNN 模型在不同季节下对 AQI 的预测不如 CNN-GRU 模型稳定.从图 7 可见,对于广州的 AQI 分季节预测,GA-BP 模型和 MVO-GRNN 模型同样表现出不稳定,例如,GA-BP 模型冬季预测的 RMSE 比秋季增高 8.640 5,MVO-GRNN 模型冬季预测的 RMSE 比夏季增高 6.596 5.

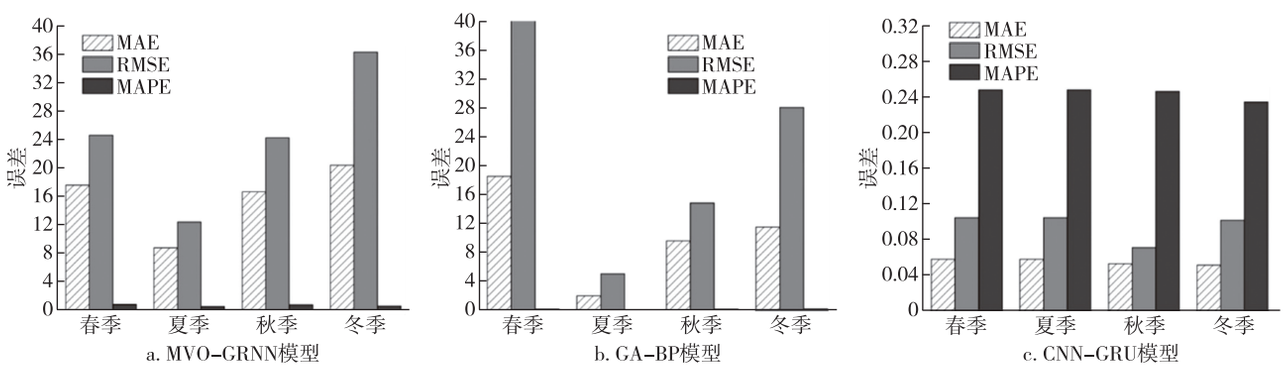

图 6　各模型对北京的 AQI 分季节预测效果对比

Fig. 6　Performance comparison between models for prediction of seasonal AQI in Beijing

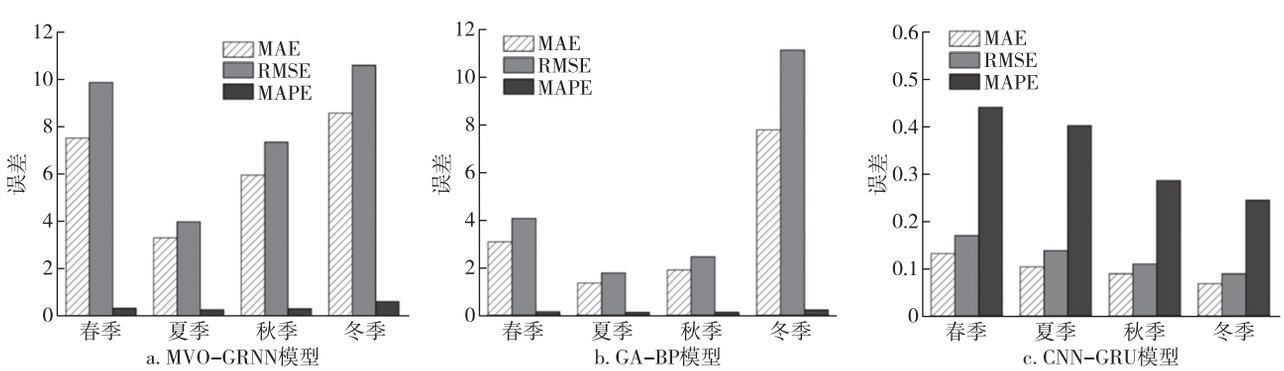

图 7　各模型对广州的 AQI 分季节预测效果对比

Fig. 7　Performance comparison between models for prediction of seasonal AQI in Guangzhou

相比于这两种模型,CNN-GRU 模型对 AQI 预测的误差值最低,各季节预测误差值的差均小于 0.2,表现出较高的稳定性和精确性.

4 结论

针对城市空气质量时间序列具有非线性和复杂性的特点,本文构建了 CNN-GRU 集成深度学习模型,以 2014—2022 年北京和广州的空气质量数据进行实例研究,同时构建多元宇宙优化的广义回归神经网络模型(MVO-GRNN)、遗传算法优化的 BP 神经网络模型(GA-BP)对 AQI 进行对比预测.在进行整体预测后,本文使用以上 3 种模型分别对北京和广州的春季、夏季、秋季、冬季进行预测.通过实验得到以下 4 点结论:

1)在研究特征组合对 AQI 预测效果的影响中,第一组特征组合在 AQI 预测中表现出更好的效果,尤其是在 CNN-GRU 模型下,具有更低的误差值.这表明 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2 和 O_3 这组特征对 AQI 的预测起到了关键作用.

2)在整体预测情况下,与 MVO-GRNN 模型、GA-BP 模型对比,本文提出的 CNN-GRU 集成深度学习模型对 AQI 的预测更精确,验证了 CNN 在处理高维数据和特征提取方面的有效性,以及 GRU 在计算方面的高效性.

3)在分季节预测情况下,与 MVO-GRNN 模型、GA-BP 模型对比,CNN-GRU 集成深度学习模型对 AQI 的预测更稳定,误差变化最小,不会因为季节变化而出现较大的误差波动,验证了 CNN-GRU 模型具有较好的泛化能力和稳定性.

4)对比整体预测情况和分季节预测情况,CNN-GRU 集成深度学习模型在整体预测情况下更精确,验证了 CNN-GRU 集成深度学习模型更适合对空气质量指数进行整体预测,可以避免考虑各季节复杂的特征变化.

参考文献

References

[1] 秦耀辰,谢志祥,李阳.大气污染对居民健康影响研究进展[J].环境科学,2019,40(3):1512-1520
QIN Yaochen,XIE Zhixiang,LI Yang.Review of research on the impacts of atmospheric pollution on the health of residents[J].Environmental Science,2019,40(3):1512-1520

[2] 牟敬锋,赵星,樊静洁,等.基于 ARIMA 模型的深圳市空气质量指数时间序列预测研究[J].环境卫生学杂志,2017,7(2):102-107,117
MOU Jingfeng,ZHAO Xing,FAN Jingjie,et al.Time series prediction of AQI in Shenzhen based on ARIMA model[J].Journal of Environmental Hygiene,2017,7(2):102-107,117

[3] 杨思琪,赵丽华.随机森林算法在城市空气质量预测中的应用[J].统计与决策,2017(20):83-86
YANG Siqi,ZHAO Lihua.Application of random forest algorithm in urban air quality prediction[J].Statistics & Decision,2017(20):83-86

[4] 焦东方,孙志华.空气质量指数回归分析[J].中国海洋大学学报(自然科学版),2018,48(增刊 2):228-234
JIAO Dongfang,SUN Zhihua.Regression analysis of air quality index[J].Periodical of Ocean University of China,2018,48(sup2):228-234

[5] Shishegaran A,Saeedi M,Kumar A,et al.Prediction of air quality in Tehran by developing the nonlinear ensemble model[J].Journal of Cleaner Production,2020,259:120825

[6] Phruksahiran N.Improvement of air quality index prediction using geographically weighted predictor methodology[J].Urban Climate,2021,38:100890

[7] 胡青,龚世才,胡珍.基于改进麻雀搜索算法的空气质量指数预测[J].广西科学,2022,29(4):642-651
HU Qing,GONG Shicai,HU Zhen.Air quality index prediction based on improved sparrow search algorithm[J].Guangxi Sciences,2022,29(4):642-651

[8] 孙启森,张建新,程海阳,等.基于注意力的卷积神经网络金融时序数据预测[J].计算机应用,2022,42(增刊 2):290-295
SUN Qisen,ZHANG Jianxin,CHENG Haiyang,et al.Financial time series data prediction by attention-based convolutional neural network[J].Journal of Computer Applications,2022,42(sup2):290-295

[9] 袁培森,黎薇,任守纲,等.基于卷积神经网络的菊花花型和品种识别[J].农业工程学报,2018,34(5):152-158
YUAN Peisen,LI Wei,REN Shougang,et al.Recognition for flower type and variety of chrysanthemum with convolutional neural network[J].Transactions of the Chinese Society of Agricultural Engineering,2018,34(5):152-158

[10] Shujaat M,Wahab A,Tayara H,et al.pcPromoter-CNN:a CNN-based prediction and classification of promoters[J].Genes,2020,11(12):1529

[11] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(6):1229-1251
ZHOU Feiyan,JIN Linpeng,DONG Jun.Review of convolutional neural network[J].Chinese Journal of Computers,2017,40(6):1229-1251

[12] 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
YANG Li,WU Yuxi,WANG Junli,et al.Research on recurrent neural network[J].Journal of Computer Applications,2018,38(S2):1-6,26

[13] Liu P H,Liu J,Wu K.CNN-FCM:system modeling promotes stability of deep learning in time series prediction

[J].Knowledge-Based Systems,2020,203:106081
[14] 罗建豪,吴建鑫.基于深度卷积特征的细粒度图像分类研究综述[J].自动化学报,2017,43(8):1306-1318
LUO Jianhao,WU Jianxin.A survey on fine-grained image categorization using deep convolutional features[J].Acta Automatica Sinica,2017,43(8):1306-1318
[15] 王艳平,韩晓冰.基于CNN的主动悬架传感器故障诊断[J/OL].控制工程:1-6[2022-11-22].https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C45S0n9fL2suRadTyEVl2pW9UrhTDCdPD64wIMnTVGX3TlB19tACLaj6qrspFdwKw3wbBCaM13H3OmPugJRq6N6D&uniplatform=NZKPT
WANG Yanping,HAN Xiaobing.CNN-based active suspension sensor fault diagnosis [J/OL]. Control Engineering:1-6[2022-11-22].https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C45S0n9fL2suRadTyEVl2pW9UrhTDCdPD64wIMnTVGX3TlB19tACLaj6qrspFdwKw3wbBCaM13H3OmPugJRq6N6D&uniplatform=NZKPT
[16] 蔡薇薇,徐彦伟,颉潭成.基于CNN-LSTM的轴承剩余使用寿命预测[J].机械传动,2022,46(10):17-23
CAI Weiwei, XU Yanwei, XIE Tancheng. Prediction of bearing remaining service life based on CNN-LSTM[J]. Journal of Mechanical Transmission, 2022, 46(10): 17-23
[17] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv e-print, 2014, arXiv:1406.1078
[18] Hochreiter S, Schmidhuber J. Long short-term memory[J].Neural Computation,1997,9(8):1735-1780
[19] 彭豪杰,周杨,胡校飞,等.基于深度学习与随机森林的$PM_{2.5}$浓度预测模型[J].遥感学报,2023,27(2):430-440
PENG Haojie, ZHOU Yang, HU Xiaofei, et al. A $PM_{2.5}$ prediction model based on deep learning and random forest[J]. Journal of Remote Sensing, 2023, 27(2): 430-440

Air quality index prediction based on integrated deep learning model

LU Kaili[1] YANG Lu[1] LI Tao[1]
1 School of Statistics,Xi'an University of Finance and Economics,Xi'an 710100,China

Abstract Air pollution seriously endangers the travel safety and health of residents.As a comprehensive indicator used to measure air quality condition,Air Quality Index (AQI) can alert the public to air quality and enable people to make more informed travel decisions.By predicting the change of air quality in advance,the government and environmental protection departments can take emergency measures to reduce air pollution.Here,we propose an integrated deep learning model based on Convolutional Neural Network and Gated Recurrent Unit (CNN-GRU) for AQI prediction.The CNN is used to extract the spatial and temporal characteristics of air pollutants and AQI and complete the feature mapping,while the GRU to model the temporal relationship and complete the calculation and AQI efficiently.The daily average concentrations of six major air pollutants ($PM_{2.5}$,PM_{10},SO_2,CO,NO_2,O_3) in Beijing and Guangzhou during 2014-2022 are selected for example study,and the AQI is predicted using the CNN-GRU model.The results show that,compared with Multiverse-Optimized Generalized Regression Neural Network model (MVO-GRNN) and Genetic Algorithm-optimized BP neural network model (GA-BP),the proposed CNN-GRU model has the smallest prediction error for AQI.

Key words air quality index (AQI); convolutional neural network (CNN); gated recurrent unit (GRU); integrated model