

一种卷积自编码深度学习的空气污染 多站点联合预测模型

张 波^{1,3}, 陆云杰¹, 秦东明^{2,4}, 邹国建¹

(1. 上海师范大学信息与机电工程学院, 上海 200234; 2. 同济大学电子与信息工程学院, 上海 201804;
3. 上海智能教育大数据工程技术研究中心, 上海 200234; 4. 中科三清科技有限公司, 北京 100089)

摘要: 城市空气污染因空间扩散特性呈现出区域内的浓度高关联性. 因此如何通过多个空气污染监测站的时空数据预测特定目标地点的污染情况, 以解决站点分布不匀的问题, 是一个重要的研究工作. 本文结合空气污染物因素特性和气象因素的多维度影响, 提出了一个利用区域内多站点空间监测数据实现特定目标站点的空气污染物浓度预测模型. 该模型通过多层卷积神经网络(Convolutional Neural Network, CNN)实现城市多站点污染物浓度与气象数据之间的维度关联特征及空间关联特征学习, 进而利用基于多层长短记忆网络(Long Short-Term Memory, LSTM)的自编码网络实现多站点浓度的时序关联特征分析. 实验通过真实数据集验证, 所提出的预测模型获得了高于传统机器学习污染物浓度预测模型的预测准确度, 且在多个城市数据集上验证了模型的泛化能力.

关键词: 深度学习; 空气污染; 时空数据; 多站点联合预测; 卷积神经网络; 长短期记忆网络

中图分类号: TP183; TP399

文献标识码: A

文章编号: 0372-2112(2022)06-1410-18

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.12263/DZXB.20200258

A Multi-Site Joint Air Pollution Prediction Model Based on Convolutional Auto-Encoder Deep Learning

ZHANG Bo^{1,3}, LU Yun-jie¹, QIN Dong-ming^{2,4}, ZOU Guo-jian¹

(1. College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China;

2. College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China;

3. Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai 200234, China;

4. 3Clear Science & Technology Co. Ltd, Beijing 100193, China)

Abstract: Due to the nature of spatial diffusion, urban air pollution presents a high correlation with regional concentration feature. Therefore, how to use spatio-temporal related data from urban multiple air pollution detection sites to predict air pollution concentration of a special target location is an important research effort for solving the problem of uneven site distribution. Concerning the multi-dimensional impacts of the air pollutant factors' features and the influence of meteorological factors, we propose an air pollutant concentration prediction model, which uses multi-site spatial detection data within a region to predict the concentration of the target station in this paper. This model is able to learn dimensional correlation characteristics and spatial correlation characteristics from multi-site pollutant concentration and meteorological data in the urban area through the multi-layer convolutional neural network, and then analyze the time-series correlation characteristics of multi-site concentration by utilizing the multi-layer auto-encoder network based long short-term memory network. The experimental results show that our proposed model obtains better performances than traditional machine learning based models under a real-world dataset, and meanwhile, the generalization performance of the proposed model has been examined based on multiple urban air pollution database.

Key words: deep learning; air pollution; spatio-temporal data; multiple sites joint prediction; convolutional neural network; long short-term memory network

1 引言

城市空气污染问题日趋严重,已经给人们的身体健康^[1]以及日常生活造成了严重的影响^[2],相关环境部门以及研究人员对空气质量问题的关注度越来越高^[3].随着信息技术的飞速发展,空气污染预测问题也迎来了全新的研究思维.利用大数据思维与深度学习技术的结合对空气污染进行有效的数据分析,进而做出准确的预测是当前环境科学和计算机科学交叉学科研究的前沿热点问题^[4-6].

目前,城市空气污染预测主要通过部署多个污染监测站点进行数据监测,然后采用综合数据分析等方法开展.一般来说,数据分析预测,如概率模型法、机器学习方法等,都在这类问题中有广泛的应用,这些研究方法各有特点,如朴素贝叶斯^[7]、BP(Back-Propagation)神经网络^[8]等都能在一定规模的数据集下取得比较好的预测效果.但是这些方法还有一些不足,比如,模型结构及获取的特征较单一,其计算模式适应特定城市条件而缺乏泛化能力;同时,由于站点分布地理位置不均,站点数据之间的地理空间关联特征未获得挖掘,无法充分提取多站点的污染物及气象大数据之间的时序以及空间关联性的问题,使得预测不够精确.因此传统的预测方法仍存在各自的瓶颈有待突破^[9-11].

近年来,深度学习方法在各个领域都获得了突破,比如图像识别^[12]、自然语言处理^[13]、生物工程^[14]以及时空结构的特征学习与分析^[15]等领域,基于深度学习的城市空气污染物浓度预测也获得了相应的关注.通过对大量数据的有效训练,深度学习可以很好地提取数据之间的时间与空间关联性,这是传统预测方法所不具备的.

卷积神经网络(Convolutional Neural Network, CNN)^[16,17]已经在图像领域取得了很大的成就,证明其在处理空间数据方面具有极其强大的功能.因此,在处理分布不均匀站点间的污染物数据间的空间相关性时,CNN能够获取多站点间的空间信息,然而仅仅使用CNN获取到的空间信息对于解决长时间的时序预测问题是不够的.长短期记忆网络(Long Short-Term Memory, LSTM)^[18]在时间序列数据处理方面具有优异的性能,而污染物数据也是以时间序列数据形式呈现的,所以大多数污染物预测问题使用LSTM进行预测^[19-36].但是单独使用LSTM预测时考虑到时序数据间的长时间依赖关系,且其结构仅能进行单点的预测,例如根据输入的历史24 h数据,单独地预测未来25 h、26 h的数据,不能根据上一时刻的预测输出作为下一时刻的输入,做到连续的预测.秦东明等人^[21]提出的自编码网络空气污染预测模型正是基于LSTM的,该模型的编码解码部分均由多层的LSTM堆叠构成,并且该模型可以解决

LSTM做长时间预测时的缺点,但是该文章仅根据单城市的综合历史污染物数据来预测未来单城市的污染物浓度,没有运用在复杂的分布不均匀的多站点间的污染物浓度预测中.现有的应用在城市空气污染预测的深度学习模型虽然能够取得一定的预测效果,但是都面临着以下几个问题:

(1)站点地理分布不均,无法提取数据间深层次的空间以及时间关联关系,从而难以实现特定地点预测水平的提升;

(2)不能同时融合时空特征,有效地连续预测未来一段时间内的污染物情况;

(3)由于模型结构以空间或时间为主,多维度数据的利用能力不足,提取数据内部关联特征能力单一,导致模型的泛化能力不足.

针对传统机器学习方法的不足,本文提出了CAE-Learning(Learning net based on CNN and Auto-Encoder)模型.此前的研究大多聚焦于以单城市内的综合污染物数据来进行单城市的污染物浓度预测^[21,33,34],并未考虑城市内多个分布不均匀站点之间的时空关系特征来进行联合预测.针对该问题,本文模型提取到城市内多个分布不均匀站点之间的污染物浓度及气象数据在空间及时间上的关联性,实现了多站点间的联合预测.CAE-Learning从模型的串行角度考虑污染物和气象数据时空特性的耦合关系.本文在模型的构建过程中,从多站点污染物浓度和气象数据的特性角度出发,选择合适的网络作为构建CAE-Learning模型的重要组成部分,充分考虑了多个城市污染数据的时空关联特征,从数据和模型的角度出发进行预测.从数据的角度,多个城市之间的数据在空间上存在着相互的关联特性,因此在污染物浓度预测的过程中应该充分考虑空间关联特征^[31];在时间维度上,污染物具有动态变化的过程^[20],且目标任务是预测未来一段时间内的污染物浓度;从模型的角度,CNN在提取空间关联数据特征上具有巨大的优势,因此使用CNN作为模型的底层来提取空间关联特征^[37].LSTM的端到端模型的优势在于能处理长时间序列预测任务,因此使用基于LSTM的端到端模型作为长时间污染物浓度序列预测的生成器,实现污染物浓度的精准预测^[21].本文提出的新型CAE-Learning污染物浓度预测模型,其使用串行的连接方式来提取多城市污染物浓度和气象数据的时空关联特征.本文使用CAE-Learning网络结构来做多城市污染物浓度和气象数据的时空特征提取器,实现目标城市未来一段时间内污染物浓度的精准预测.

根据数据的特性,构建CNN和基于LSTM端到端模型的CAE-Learning.首先,针对多站点的污染物浓度和气象数据的空间特征,模型采用的是多维卷积CNN作

为空间关联特征提取器。多维卷积 CNN 可以提取输入数据的空间关联特征^[37], 针对具有空间特性的环境污染数据进行更深度的提取, 产生高维的语义特征信息。提取后的信息作为后续的端到端模型的输入。而基于 LSTM 的端到端模型架构, 其 Encoder 以及 Decoder 部分均由 LSTM 组成。Encoder 部分结合了时间以及空间的特征, 即每一个时间点, Encoder 部分的 LSTM 均能提取到数据间的空间关联特性, 在整个输入时间序列维度上来说, 提取到的是时间和空间两个方面的特征。Decoder 部分根据 Encoder 部分的输出和上一时刻的输入迭代完成长时间污染物浓度序列预测任务, 即 Decoder 部分每一时刻的预测值均进行了相互的强关联^[38]。该多站点联合预测模型可以根据过去 72 h 的多站点的污染物浓度以及气象数据来预测未来 24 h 的特定站点的污染物浓度, 实现了对未来长时间内的污染物浓度的连续精准性预测。在测试集上的实验结果表明, 提出的 CAE-Learning 模型在不同城市的预测上均可以获得较高的精确度, 模型具有较高的泛化性。

CAE-Learning 模型在处理空气污染预测问题时具有如下贡献:

(1) 能有效地将单城市多站点间的污染物浓度和气象数据进行结合, 多维卷积的 CNN 能够联合多站点间的污染物浓度和气象数据特征做到深层的空间相关性提取, 能够从环境污染大数据的空间关联特征角度去进一步提升预测模型的精确度, 并且模型的卷积部分, 采用了全卷积方法, 去除了池化层带来的大幅度特征损耗问题, 充分地提取污染物与气象数据的空间特征^[39]。

(2) 引入了端到端的编码预测模型, 轻量化了基于 LSTM 的自编码模型的复杂度, 能够充分提取多站点间空气污染物和气象数据的时间关联性, 降低了模型的过拟合问题, 避免了梯度消失和梯度爆炸问题, 能够从时序数据的角度进一步提升预测模型的精确度。

(3) 本文使用的是新型预测模型, 且综合了时空域特征问题, 解决传统模型的特征提取深度不足以及特征关联度不强等问题。模型可以将预测的污染物浓度结果进行前后关联, 对未来一段时间内污染物浓度连续性预测的精确度有了很大的提升。

2 相关工作

空气污染浓度预测方法可以分为传统的非深度学习方法和基于深度学习的方法。

2.1 非深度学习的方法

非深度学习的方法用于空气污染浓度预测, 包括基于经验模型、基于概率模型、基于传统机器学习的预测。

(1) 基于经验模型的预测, 根据相关数据通过归纳参数和变量之间的关系得出相应的数学关系式。如经

验统计方法^[22]、回归方程法^[23]。但是在进行环境空气污染物浓度预报时, 经验模型往往大量地引入历史观测数据, 有很大的局限性, 不符合实际情况。

(2) 基于概率模型的预测, 以概率统计规律为基础, 结合统计学或数学的一些方法建模。例如, 用决策树模型来预测大气污染物^[24]; 以高斯预测概率密度函数的形式产生概率预报^[25]; 利用贝叶斯算法研究不同来源的污染物对其预测浓度的影响^[26]; 用隐马尔可夫模型进行污染物浓度预测^[27]。

(3) 基于传统机器学习的预测, 最早应用于环境空气污染预测领域的智能算法。在传统的机器学习预测中, BP 神经网络经常被用来做预测^[28, 29], 该方法能够在小规模的数据集上取得有效的预测结果, 然而大规模的空气污染浓度及气象数据之间具有时间依赖以及空间相关性, 而 BP 神经网络无法挖掘数据间的这些深层次联系。比起深度学习, 传统机器学习不能挖掘数据中深层次的联系而无法建立更精确的预测模型。

2.2 基于深度学习的方法

深度学习的方法^[30]能够通过合适的训练方法对样本数据进行一系列的训练, 并反向调整网络参数, 最后得到具有深层次的网络结构的机器学习过程。由于传统预测方法的不足, 近年来学术界开始尝试采用深度学习方法进行城市环境空气污染预测的工作^[31]。而深度学习虽然是机器学习的一种, 但是比起传统机器学习方法, 深度学习的方法在预测时能够充分地提取数据间的关联特征, 并在此基础上建立更为精准的预测模型。目前国内外研究者已经使用深度学习的方法开发出多种空气质量预测模型, 如使用深度集成模型^[32]进行空气质量的预测, 使用扩散卷积神经网络^[33]进行精确的空气质量预测, 使用自编码网络^[21]做空气污染物浓度变化的预测, 采用深度学习进行空气质量的插值、预测、特征分析^[34], 采用新型的时空长短期网络^[35]进行空气污染预测。

(1) 深度集成模型利用历史空气质量及气象数据以及不同的天气模式划分不同的区域, 对每个区域采用深度 LSTM 学习数据间的长短期依赖关系, 再对每个区域得到的预测结果进行集成, 得到最终的预测结果。

(2) 扩散卷积神经网络在进行空气质量预测时, 利用邻域特征来表示空间相关性, 并构建一个图, 表示监测站点之间的相似程度。该模型根据过去数小时的空气质量、气象数据及地理环境的图像, 来预测单城市的特定污染物的浓度变化。

(3) 自编码网络预测空气污染物浓度变化时, 根据城市内综合的历史空气污染物浓度数据, 利用自编码网络的特性, 使用编码器对已知时间范围内的历史污染物数据进行学习, 输出一段包含历史数据特征的隐

藏向量;进而使用解码器利用这段隐藏向量预测未来短期时间的污染物浓度,达到对城市内综合污染物浓度变化预测的目的。

(4)利用深度学习方法解决了空气质量的插值、预测、特征分析这三个问题,并且在实现过程中,在深度学习网络的不同层次中嵌入半监督学习,来提高插值、预测的性能,最后实验表明该模型可以在单城市的数据集里面解决上述问题。

(5)新型时空长短网络结合了卷积神经网络及长短期记忆神经网络,在使用该网络对城市内污染物浓度进行预测时,对污染物数据进行时空特征提取后,再加入气象数据及气溶胶数据,帮助模型更好地预测空气污染物的变化。

以上基于深度学习的研究方法大多着力于解决单城市内的综合污染物数据来进行单城市的污染物浓度预测^[21,33,34],并没有考虑到城市内多个分布不均匀站点之间的空间及时间关联性对城市内单个站点的污染物浓度变化的影响,也并未做到对未来一段时间内的空气污染物浓度变化情况的连续性预测。相比而言,本文提出的 CAE-Learning 模型,由 CNN 及基于 LSTM 的自编码网络构成,突出城市内多个区域的空气污染物及气象数据联合对特定位置预测的能力,其特点在于:实现了轻量化端到端的编码预测模型^[21];避免池化层的大幅度特征损耗问题,充分地提取空间关联特征。

3 CAE-Learning 空气污染预测模型

本文提出的 CAE-Learning 预测模型融合了 CNN 及基于多层 LSTM 的自编码网络,可以有效地对城市内多站点数据间的时间、空间关联特征都做到充分的提取。由于城市内多个站点的污染物以及气象数据在空间上具有空间关联性,并且 CNN 由于其模型结构,相比于其他深度学习、神经网络模型而言,其空间信息提取能力要更强,所以 CNN 在提取空间关联信息上具有很强的优势^[35]。而污染物浓度在时间维度上是具有前后动态性的,是随着时间进行变化的过程。同时 LSTM 神经网络很适合于提取长时间序列的特征^[39],相较于其他模型,该模型的时序信息提取能力更强、更好。所以根据污染物浓度的特性以及 LSTM 的特点,采用 LSTM 作为时序信息提取器。而为了得到未来 24 h 的污染物浓度的预测值,采用了端到端的模型架构,并且 Encoder 以及 Decoder 部分均采取 LSTM 网络,进行有效的时序信息的提取。相较于在之前的时空污染物浓度预测任务中取得最好预测效果的串行耦合的 CNN-LSTM^[37]而言,本文在提取空间信息时也采用 CNN,但是在后续的时序信息获取时,为了能够更精确地根据历史多个小时多站点组成的污染物浓度数据来预测未来 24 h 的特

定污染物浓度预测值,采用了基于 LSTM 的端到端预测模型,获得了更好的预测效果。本节先介绍时空预测问题的建模,再介绍 CAE-Learning 模型的实现及训练过程。

3.1 时空预测问题建模

本文的多站点联合预测问题主要是解决时空融合的时间序列预测问题,根据城市内多站点间空气污染物浓度及相关气象数据来对特定目标站点的污染物浓度进行预测。如图 1 所示,图 1 左侧部分为多个相邻站点及目标站点之间的联合预测示意图,其中每个站点的数据特征都由污染物浓度及气象因子构成。由多个站点及其特征组成一个二维的矩阵,根据输入的具有时空特性的二维矩阵得到输出的一个一维向量,即所要得到的时间序列预测。

具体时空预测问题建模的数据流表示为:给定多个站点 $S = \{s_1, s_2, \dots, s_i, \dots, s_n\}$, 其中 s_i 为目标站点,其余站点为相邻站点, n 为站点总数。例如站点 s_1 包含污染物浓度及气象数据的时间序列数据 $R_{s_1} = \{r_1, r_2, \dots, r_i, \dots, r_k\}$, 其中, r_i 为待预测的目标污染物 $PM_{2.5}$, w 为每个站点包含的污染物浓度及气象数据特征数。在当前多站点数据集下,数据的输入格式、输入时间间隔以及未来污染物浓度预测的时间序列的长度为:给定一个时间点 t , 将 t 之前的 D 小时内的数据作为历史输入数据,数据时间间隔为 1 h, 令 $T_1 = \{t-D, t-D+1, \dots, t\}$ 为 t 之前的 D 小时内输入到预测模型中的数据序列,用于预测之后 M 个小时 $T_2 = \{t+1, t+2, \dots, t+M\}$ 的目标污染物 r_i 的浓度,预测数据的时间间隔也为 1, 且每个时间点的输入数据为一个 $n \times w$ 的二维矩阵 I_i 。图 2 表示由城市内多站点间的时间序列数据组成的二维矩阵数据流,作为模型的输入数据形式。

输入数据时间间隔为 1, 将 t 之前的 D 小时内的数据表示为 $I = \{I_{t-D}, I_{t-D+1}, \dots, I_t\}$, 其中 I 的维度为 $D \times (n \times w)$, 对于 M 小时后预测的目标污染物的浓度序列时间间隔也为 1, t 之后的预测序列数据表示为 $P = \{P_{t+1}, P_{t+2}, \dots, P_{t+M}\}$, 其中 P 的维度为 $(M \times 1)$, 将 CAE-Learning 模型表示为目标函数 f_{CAE} , 可得数学模型表达式为

$$f_{CAE}(I = \{I_{t-D}, I_{t-D+1}, \dots, I_t\}) = \{P_{t+1}, P_{t+2}, \dots, P_{t+M}\} \quad (1)$$

3.2 CAE-Learning 预测模型

模型从空间和时间两个维度出发,设计了一种以 CNN 作为底层以提取空间关联特征,基于多层 LSTM 的自编码网络作为中间层以提取时间序列特征,全连接层作为顶层以产生最终预测结果的三层架构的预测模

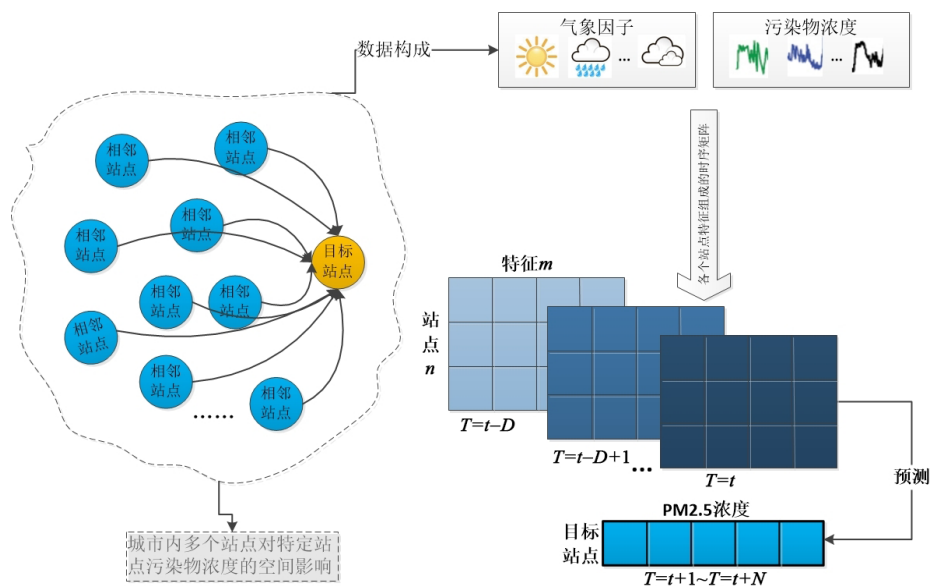


图1 多站点的时空关系图

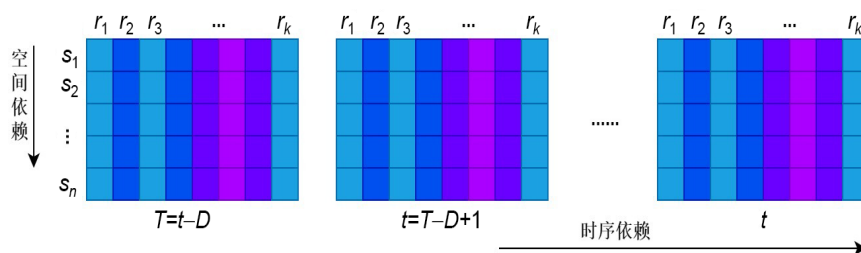


图2 输入模型的数据形式

型 CAE-Learning. CAE-Learning 的整体模型框架如图 3 所示. 首先模型的输入为时间序列 D 的城市内多个站点的空气污染物浓度数据和气象数据, 然后数据进入具有卷积层和池化层的 CNN 中进行空间关联性特征提取, 将提取过的特征数据输入到自编码网络的编码器

中进行时间关联特征提取, 并最终将自编码网络的解码器的每个时刻隐藏状态送入到全连接层产生一个一维的预测结果, 即需要的 $PM_{2.5}$ 的预测结果. 下面分别介绍 CNN、自编码网络两个模型的预训练过程以及最终 CAE-Learning 的全局训练过程.

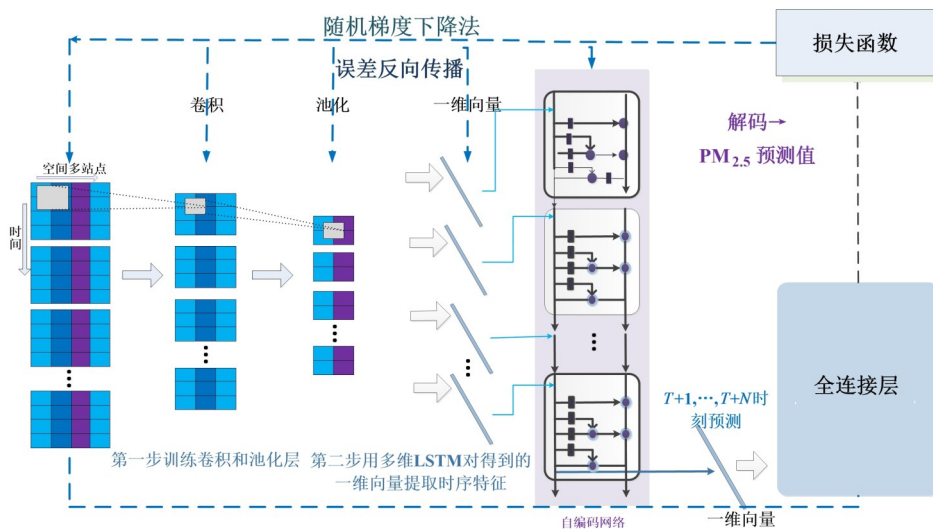


图3 CAE-Learning 预测模型框架

(1) CNN 的预训练过程

在对 CNN 训练之前需要做的是对数据的预处理工作:首先对数据进行空值填充后再归一化,然后将输入污染物浓度数据和气象数据转化成 CNN 可接收的具有时间序列的二维矩阵,然后再输入到 CNN 中进行空间特征的提取.本阶段的预训练是用前一时刻的污染物浓度和气象数据作为 CNN 的输入,再利用全连接层接收卷积神经网络的输出并产生下一时刻的目标污染物浓度预测值.令 η 为当前卷积神经网络模块正在训练的层数, m 代表最终提取到的特征图,卷积层的上一层输出的特征图由该卷积层的卷积核 k 进行特征提取并学习, f 为 Relu 激活函数,通过激活函数对卷积的结果进行非线性变换而得到输出的特征图, i, j 均为特征图下标, M 为特征图的通道数,即

$$m_j^\eta = f \left(\sum_{i \in M} m_i^{\eta-1} * k_{ij}^\eta + b_j^\eta \right) \quad (2)$$

特征图经过 CNN 中的卷积层卷积过后,得到 N 个特征图作为池化层的输入,本模型中的池化层通过平均池化方法对 N 个特征图进行降维,将输出 N 个缩小后的特征图,过程如下:

$$m_j^\eta = f \left(\beta_j^\eta \text{down} \left(m_j^{\eta-1} \right) + b_j^\eta \right) \quad (3)$$

其中, β 和 b 分别作为输出图像的相乘性质和加性偏置, $\text{down}(\cdot)$ 表示下采样函数, $\text{down}(\cdot)$ 即为平均池化方法,将这 N 个特征图展开成 N 个一维向量,再经过全连接层的解码,最后得到输出的污染物浓度值.

这一阶段输入的二维矩阵为 t 时刻之前 D 小时的多站点的污染物浓度和气象数据,以均方根误差衡量预测的准确性.预训练过程中采用误差反向传播算法,将池化层作为考虑的因素并基于所有值更新卷积层的权重,优化网络预测性能,减少预测值和观测值之间的误差.

这一阶段的训练的主要作用就是将输入的二维矩阵进行压缩,同时深层次挖掘数据间的空间关联特征.当网络符合期望后,停止第一阶段网络的训练,将本阶段训练好的 CNN 权重参数迁移到本节提出的 CAE-Learning 模型当中,然后进入第二阶段的训练.

(2) 自编码网络的预训练过程

为了有效利用现有污染物浓度和气象数据来提取数据间的时序特征,并能够预测未来一段时间内的污染物浓度,本文采用了一个基于多层 LSTM 的自编码网络,其解码与编码部分都由多层的 LSTM 组成,LSTM 所具有的解决长时间依赖问题的能力和避免梯度消失问题的优点,可以应用这些优点来解决时序预测问题.自编码网络的编码器部分用来提取输入序列的时间关联特征,解码器部分则将每个时刻的隐藏状态输入到全连接层产生最终的预测结果.

如图 4 所示,自编码网络先由 LSTM 构成的编码器部分来提取出城市空气污染物浓度和气象数据的时序特征,实现对历史污染物的浓度和气象数据的编码,然后形成一个具有时序特征的隐藏向量 C ,隐藏向量作为解码器端的输入,进一步做时序预测.

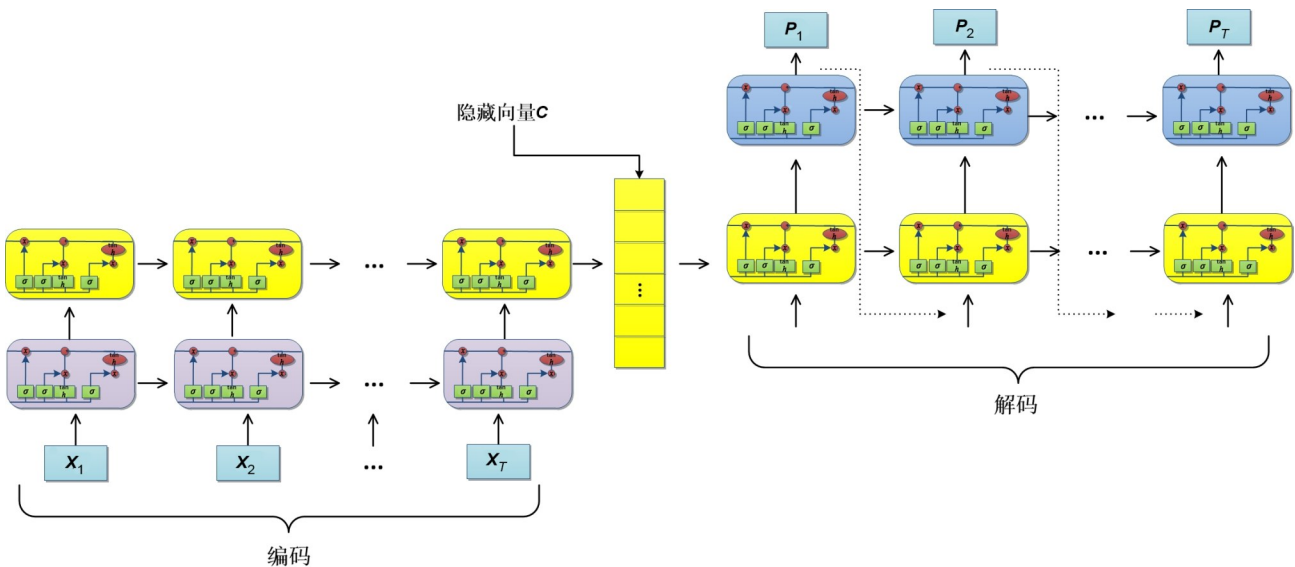


图 4 自编码网络模型结构图

(a) 编码器部分的设计

假设给定输入序列 $X = (x_1, \dots, x_t, \dots, x_T)$, 则隐藏向量 C 为

$$h_t = f(x_t, h_{t-1}); C = \rho(\{h_1, \dots, h_T\}) \quad (4)$$

$$\begin{cases} i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \\ f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \\ C_t = f_t * C_{t-1} + i_t * \tan h(W_c[h_{t-1}, x_t] + b_c) \\ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \\ i_t = o_t \circ \tan h(C_t) \end{cases} \quad (5)$$

其中, x_t 为 t 时刻的输入值; h_{t-1} 为上一个时刻长短期记忆网络输出的隐藏状态; f 为 LSTM 函数; h_t 为 t 时刻的隐藏状态; ρ 为隐藏状态计算函数, 可由式(5)得出; 向量 C 为 LSTM 中的最后时刻单元的输出状态. 式(5)中, i, f, o 分别表示 LSTM 中的输入门、遗忘门和输出门; C_t 表示神经元的状态信息; x 表示网络输入; W 表示网络参数; b 表示偏置量; h 表示隐藏状态; σ 表示 sigmoid 函数, 输出 0~1 的值, 表示让多少信息通过, 1 表示让所有信息都通过. 由式(4)得到的隐藏向量 C 会作为解码器的输入. 同样由多层 LSTM 组成的解码器会将最后一层的 LSTM 的隐藏状态作为整个编码器的最终输出状态, 此时编码器的任务结束.

(b) 解码器部分的设计

解码器主要功能是结合语境向量 C 和当前时刻的输入数据预测下一时刻的污染物浓度.

解码器的主要计算方法如下:

$$\begin{cases} h_t, s_t = f(h_{t-1}, s_{t-1}, C) \\ p_t = Wh_t + b \end{cases} \quad (6)$$

其中, f 为 LSTM 函数, 其函数的实现形式如式(5)所示; s_t 是当前时刻 t 的隐藏状态; h_{t-1} 为上一时刻的输出值; s_{t-1} 为 $t-1$ 时刻的隐藏状态; C 为编码器输出的语境向量; h_t 为 t 时刻输出; p_t 为 t 时刻污染物浓度的预测输出; W 和 b 为模型参数.

在自编码网络中, 编码器和解码器使用同类型 LSTM 结构, 一个用来编码输入序列, 另一个用来解码输出序列. 其中, 编码器和解码器中的 LSTM 层数是可以调节的.

(c) 自编码网络的预训练过程

(I) 在自编码网络的训练过程中, 首先对编码器的第一层 LSTM 参数进行随机初始化, 利用自编码网络可以通过无监督的预训练过程学习到数据之间的关联特征的优点, 进行参数调优. 自编码网络可以使得每层隐

藏层的输入与隐藏层的输出最大概率的相同, 从而减少输入的数据信息在网络层中传播而产生的信息损耗, 自编码网络中每个网络隐藏层 LSTM 的预训练过程如式(7)所示, 预训练的目标是输出数据特征 x' 等于输入特征 x . 在预训练过程中, 选择 L2 范数的平方来表示误差函数, 用式(8)计算损失函数大小, 并作为衡量无监督训练过程的信息损耗的指标. 计算如下:

$$\begin{cases} X = \sigma_a(W_a * x + b_a) \\ x' = \sigma_s(W_s * X + b_s) \end{cases} \quad (7)$$

$$\text{loss}(w, b) = \frac{1}{n} \sum_{i=1}^n \|x^i - x'^i\|_2^2 + \alpha R(W, b) \quad (8)$$

$$R(W, b) = \|W_a\| + \|W_s\| \quad (9)$$

其中, x 是自编码网络的每个隐藏层 LSTM 的输入(即输入数据单元某时刻的环境空气污染物浓度和气象数据); X 是隐藏层的输出; $W_a \in \mathbf{R}^{u \times v}$; $b_a \in \mathbf{R}^v$; σ_a 是隐藏层的激活函数 sigmoid($u \times v$ 表示的是权重矩阵维度); x' 是隐藏层的输出, $W_s \in \mathbf{R}^{u \times v}$, $b_s \in \mathbf{R}^v$, σ_s 是隐藏层的激活函数 sigmoid. 这样在训练过程中, X 就可以看作是 x 的一种特征表示, 同时 x' 表示为隐藏层的输出经过解码运算的结果, 可最大程度逼近隐藏层的输入特征 x . 在式(8)中, $R(W, b)$ 是权值衰减的正则化项, 由式(8)可知, α 为正则化系数, x^i 和 x'^i 分别表示当前预训练层的输出和输入, n 表示输入数据 x 的特征维度, $\text{loss}(W, b)$ 为损失函数.

编码层中的后续 LSTM 层的输入即为上一层 LSTM 层的输出隐藏特征向量 X , 且都用(I)中相同的训练方式进行预训练.

(II) 自编码网络的全局微调过程

全局参数优化过程中, 本节将已预训练好的编码器参数迁移到当前的自编码网络中进行参数微调. 微调的过程可以解释为, 将历史的污染物浓度和气象数据输入到编码器中进行时序特征提取, 自编码网络将编码器的最终的隐藏状态作为解码器的输入并产生污染物浓度预测值. 全局微调的过程中, 使用式(11)计算的均方根误差作为损失函数计算预测值与观测值之间的误差, 通过随机梯度下降算法更新全局权重, 直至模型收敛.

(3) CAE-Learning 的全局训练过程

考虑到深层次的神经网络在训练时容易产生过拟合问题, 本文在 CNN 中加入了 dropout 方法来避免过拟合, 并且在整个模型中使用随机梯度下降法, 通过误差反向传播的方式计算误差函数来对网络全部权重和偏置值的梯度进行更新, 直至模型的性能符合期望, 同时避免网络对训练数据的过度学习.

为解决深度网络在训练时易出现的过拟合问题,使用EN(Elastic Net)算法进行正则化约束,使目标函数在训练的fine-tuning阶段达到最小,EN的优势也在文献[36]中被实验证实,故选用式(10)作为目标函数:

$$E(\phi) = \sqrt{\frac{\sum_{i=1}^N (O_i - P_i)^2}{N}} + \lambda [\zeta \phi^T \phi + (1 - \zeta) |\phi|] \quad (10)$$

网络的目标函数设置为均方根误差和正则项之和.式(10)等号右侧的前半部分为均方根误差, O_i 是目标污染物的观测值(观察值), P_i 是目标污染物的预测值, N 为预测的时间段污染物浓度时间序列长度;后半部分中引入EN算法进行正则化约束, λ 为一个非负超参数, ϕ 包含了整个训练过程参数,包括卷积层的权值矩阵、池化层的权值矩阵、自编码网络内部神经元之间传递信息的权值矩阵、全连接层的权值矩阵, ζ 为控制L1,L2惩罚大小使用的比例参数, $\zeta \in (0, 1)$.

4 实验及结果

4.1 实验数据集和评价标准

本文将单城市多个站点的污染物浓度和气象数据作为实验对象,环境空气污染物 $PM_{2.5}$ 浓度为目标污染物.本节实验目标城市为上海,其有9个城市环境空气质量监测站点,分别为十五厂、虹口、上师大、杨浦四漂、青浦淀山湖、静安、浦东川沙、浦东新区和浦东张江,各个站点在地图上的位置如图5所示.图5展示了上海市各个空气质量监测站点对特定目标站点(徐汇上海师范大学站)的多站点联合预测示意图.

数据的时间跨度从2014年5月13日到2018年3月

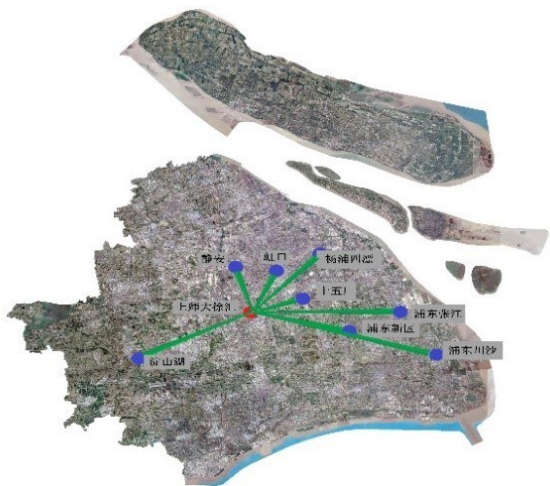


图5 上海市多站点联合预测示意图

24日,其中包括环境空气污染物数据以及气象数据,监测时间间隔为一个小时.其中训练数据有275 221条,测试数据有17 830条,数据特征包括:时间、 $PM_{2.5}$ 、AQI、 PM_{10} 、 SO_2 、 NO_2 、 O_3 、CO、气温、大气压、风向、风速、云量、气象条件、相对湿度和累积降雨量.通过利用计算空值或异常值前后非空数据的平均值来填充数据集中的空值和异常值的方法来完成对实验数据集中空值和异常值的预处理工作,然后将数据归一化到[0,1]之间并整理成模型可接收的标准化数据格式.

在每个站点的16个特征属性中,AQI、 PM_{10} 、 SO_2 、 NO_2 、 O_3 、CO为相关污染物浓度特征数据,气温、大气压、风向、风速、云量、气象条件、相对湿度和累积降雨量为气象特征数据, $PM_{2.5}$ 为目标污染物浓度,模型输入数据是过去72h内的上海市内所有9个监测站点各个特征属性值的每小时平均值,输出是特定一个站点(徐汇上海师范大学站)未来24h的 $PM_{2.5}$ 浓度的每小时平均值序列.在训练及测试时,模型每次获取训练数据集的时间长度为96h,每次移动时间窗口为1,直到训练集或测试集中的数据读取完,完成一次数据集的训练或测试.连续输入72h的数据包含上海市内所有监测站点的污染物浓度和气象数据的2维矩阵,未来24h的污染物浓度观测值在训练阶段用于调整模型参数,在测试阶段用于评估模型的预测性能.

选择利用城市内多个站点而不是一个站点的数据来预测目标站点的目标污染物浓度可提升预测目标站点污染物 $PM_{2.5}$ 浓度的精确度^[20].实验中,本文提出的预测模型和其他的对比模型均会在同一个测试集上进行10次的预测实验,其预测结果的均方根误差的最终值为10次实验误差的平均值,最大迭代次数为100次.关于数据集的其他详细说明可见表1.

表1 数据集详细参数

参数名称	参数值
数据集时间跨度	2014.5.13—2018.3.24
训练集	2014.5.13—2017.12.31
测试集	2018.1.1—2018.3.24
预测未来时长	24 h
历史数据时长	72 h
矩阵大小(站点数*特征数)	[9,16]
最大迭代次数(Epoch)	100

实验采用均方根误差(Root-Mean-Square Error, RMSE)、平均绝对误差(Mean Absolute Error, MAE)和相关系数(Corr)作为衡量预测精确度,RMSE、MAE和Corr

的计算式如下:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (O_i - P_i)^2}{N}} \quad (11)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |O_i - P_i| \quad (12)$$

$$\text{Corr} = \frac{\text{Cov}(O, P)}{\sqrt{\text{Var}[O] * \text{Var}[P]}} \quad (13)$$

其中, O_i 是目标污染物的观测值(观察值); P_i 是目标污染物的预测值; N 表示测试集的大小; $\text{Cov}(O, P)$ 为观测值和预测值的协方差; $\text{Var}[O]$ 和 $\text{Var}[P]$ 分别是观测值和预测值的方差. RMSE 和 MAE 的值越小, Corr 的值越大, 证明预测的准确度越高.

4.2 多站点联合预测效果对比

以下分析本文提出的 CAE-Learning 融合神经网络和对比模型在测试数据集上不同迭代次数下的拟合趋势, 以及为了对比 CAE-Learning 性能, 选用了 3 种经典模型(BP 神经网络、RNN 和 LSTM)作对比. 同时为了验证 CNN 和基于 LSTM 的自编码网络串行耦合的有效性, 加入了 CNN 单模型、CNN 与 RNN 串行耦合的 CNN-RNN 模型、CNN 与 LSTM 串行耦合的 CNN-LSTM^[37] 模

型, 以及并行耦合的 CNN+LSTM 模型作为对比实验. 每个模型在测试的过程中使用的数据集和 CAE-Learning 所用相同, 且都选择了相同迭代次数的拟合趋势进行比较, 分别是 10 代、30 代、50 代、70 代、90 代和 100 代, 每个模型的训练都是预测值向着真实值不断趋近的过程.

本文在主要考虑空间特征的时候, 使用 CNN 作为空间关联特征提取方法, 得出的实验结果是 RMSE 为 27.068, 仅使用 LSTM 作时间序列预测, 主要考虑时间关联特征时, 实验结果是 RMSE 为 9.668, 综合比较 CNN、RNN、LSTM 以及 CAE-Learning 模型, 当综合时空特征的时候, 实验效果要比仅考虑空间特征或时间特征的模型效果要好.

图 6 即为 CAE-Learning 模型在根据过去 72 h 的多站点污染物数据及气象数据对未来 24 h 内的 $\text{PM}_{2.5}$ 的浓度作出的预测情况的拟合曲线. 从图 6 可以看出, CAE-Learning 模型可以达到很好的拟合效果, 只在第 10 代的时候误差较大, 但是接下来的拟合情况都有明显提升, 在整体预测趋势以及精确度方面都达到了很好的预测效果, 能够使预测值无限接近于真实值, RMSE 平均值低至 8.880 (下文会有表格详细描述).

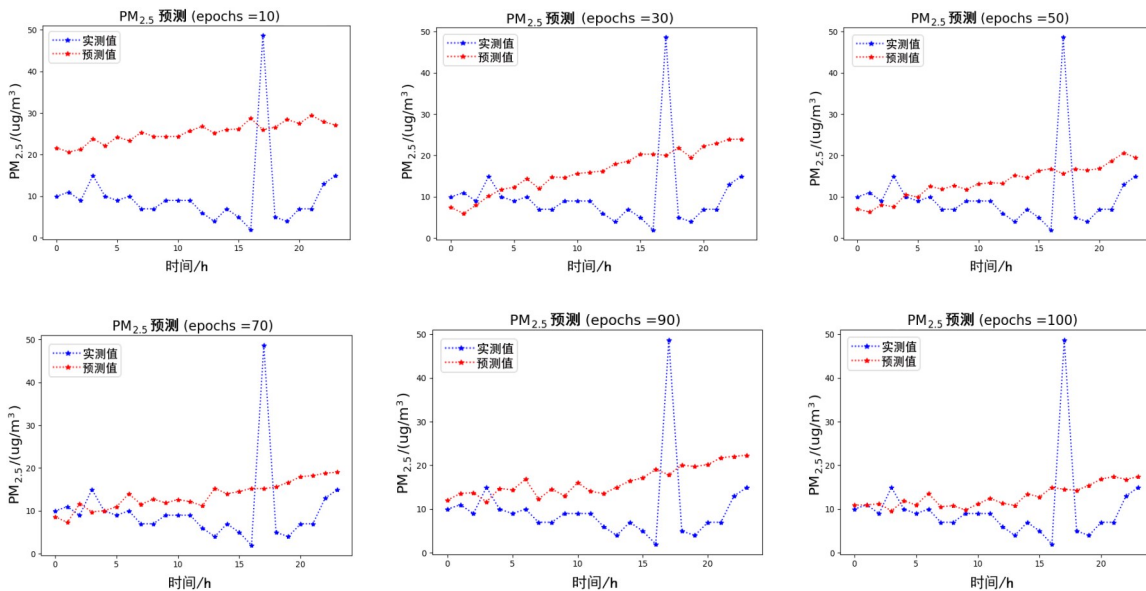


图 6 CAE-Learning 的多站点联合预测结果

图 7 为 CNN-LSTM 模型的拟合结果图, 从图中可以看出, 该模型在 50 次迭代之前精确度比较低, 50 代之后才慢慢有较好的拟合趋势. 作为 CAE-Learning 的对比实验, 在提取空间信息时采用相同的 CNN 结构的情况下, 由于基于多层 LSTM 组成的自编码网络能够

根据编码器中提取到的上一时刻的信息, 输入到解码器中作下一时刻的预测, 其对时间序列的提取效果较单独的 LSTM 模型会有较大的提升. 从图 6、图 7 的实验结果对比中也可以发现, 自编码网络在整体的预测趋势及精确度上要优于 LSTM, 且预测稳定性较高.

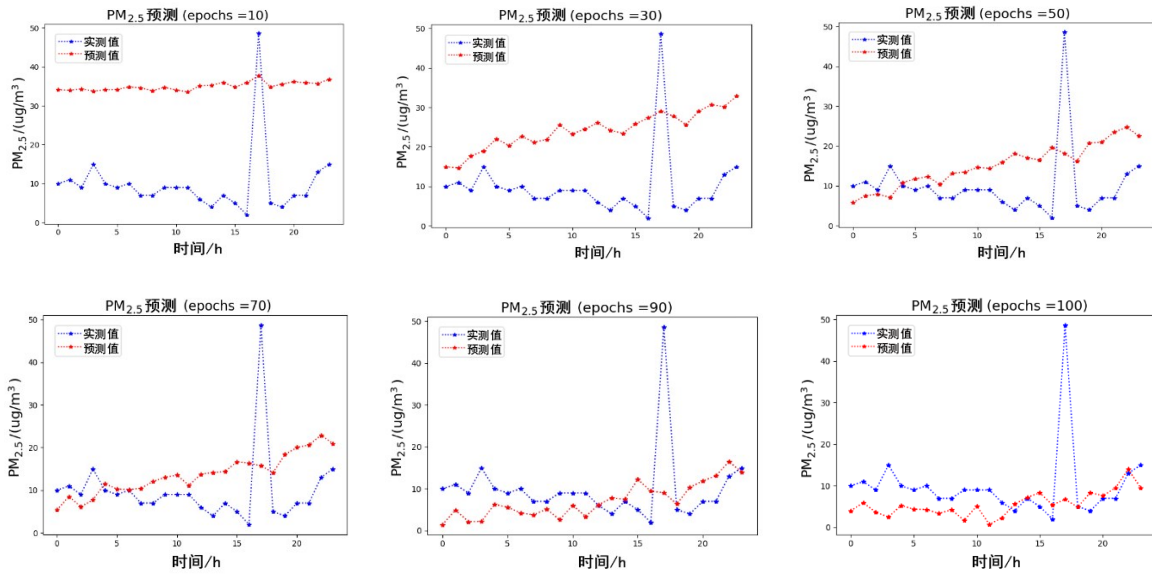


图7 CNN-LSTM的多站点联合预测结果

从而可以比较得出,基于LSTM的自编码网络比传统的LSTM模型在较长时间内的时序预测方面性能更好。

对比模型CNN-RNN在相同的训练数据集下,与CAE-Learning模型以及CNN-LSTM模型比较而言,拟合的效果稍显逊色。从表2的评价指标来看,CNN-RNN的RMSE值为15.710,CNN-LSTM的RMSE值为9.173,CAE-Learning的RMSE值为8.880,从中可以得出,在空间信息提取器均为CNN的情况下,采用端到端模型作为后续的时序提取模型会取得更好的预测效果。

相较于CNN-LSTM以及CNN-RNN等串行耦合的预测模型而言,CNN+LSTM模型采用了并行的模型融合方式对时空数据进行耦合,如图8所示。即在训练过程中,同时对CNN以及LSTM进行训练,以并行的耦合方式同时获取多站点数据之间的时间、空间特性,对CNN以及LSTM得到的预测结果进行综合,得到两个模型的并行的耦合结果的融合。由于并行的CNN+LSTM模型在时空信息提取的过程当中,CNN与LSTM两个模型会同时提取到数据集中的时间及空间信息,最终进行融合的时候,会一定程度上造成信息的冗余。结合实验结果来看,CNN+LSTM模型的预测值与真实值之间的拟合程度随着模型迭代次数的增加反而呈现降低的趋势,最终的RMSE值为22.902,与CAE-Learning以及CNN-LSTM、CNN-RNN等串行的时空信息提取模型相比较而言,预测的结果也相对较差。

CNN单模型、LSTM单模型、RNN单模型以及BP神经网络模型的实验效果均与CAE-Learning的预测结果

存在一定的差距。其中,CNN单模型能够更好地获取空间信息,体现在实验数据上的是CNN取得的相关性系数比较高;而LSTM单模型能够较好地获取时序信息,迭代次数越多,可以得到更好的预测效果,但是由于其自身无法捕获长时间数据依赖,最终预测效果不如本文提出CAE-Learning模型;BP神经网络的最终预测效果在3个网络中最差。

为了更直观地展示本文提出的CAE-Learning模型的优势,每个模型训练1次收敛的时间(用训练时间表示)以及在测试集上最终的RMSE、MAE和相关系数Corr(计算方法参考式(11)、式(12)和式(13))在表2中列出。

Corr的值表示预测值和观测值之间的关联度,Corr的值越高,表示两者关联度越高,模型的性能越好。当Corr的值达到1时,表示预测值和观测值完全相关。实验结果分析如下。

(1)将LSTM与其他单模型进行对比,由于LSTM在处理时间序列预测任务时,比RNN、CNN、BP网络具有更好的适用性,所以LSTM在实验结果上也要比其他模型好,从表2中也可以获得与模型机理一样的实验结果:LSTM的RMSE的值最小可达9.688,Corr的值最大可达到0.958,优于其他的单模型。

(2)将CNN和其他的单模型进行对比,CNN具有提取污染物与气象数据的相关性的能力,虽然其在时序预测效果上与其他单模型有一定差距,但是其在预测污染物变化的趋势上具有一定的优势。从图8以及表2中也可以发现这一特点,CNN的相关系数最大可以达到0.980,这表明在考虑了污染物和气象数据的空间相关性后,可以更准确模拟出污染物浓度的变化趋势。而

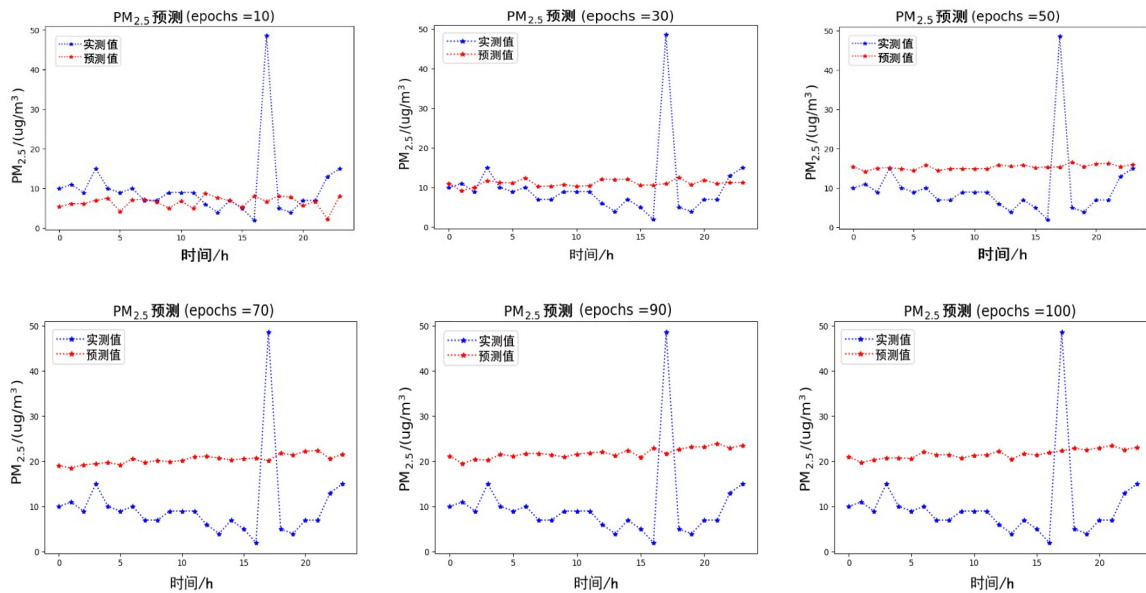


图8 CNN+LSTM并行的多站点联合预测结果

表2 每个模型的RMSE、MAE、相关系数和训练时间

模型	RMSE	MAE	相关系数	训练时间
BP	15.857	10.284	0.875	181 s
RNN	15.670	9.398	0.905	197 s
LSTM	9.688	7.139	0.958	304 s
CNN	27.068	21.048	0.980	871 s
CNN+LSTM	22.902	19.033	0.897	980 s
CNN-RNN	15.710	12.932	0.951	1219 s
CNN-LSTM	9.173	7.108	0.975	1342 s
CAE-Learning	8.880	7.001	0.980	2249 s

CNN的RMSE及MAE值均比其他单模型高,即比其他单模型在时序预测上效果要差。

(3) CAE-Learning与CNN-LSTM、CNN-RNN以及并行的CNN+LSTM模型进行比较,由于CNN-LSTM可以由CNN提取污染物数据及气象数据间的空间关联性,再由LSTM来提取时序关系,因此可以达到比其他单模型以及组合模型都要好的预测效果。从表2及图6、图7可以看出,CNN-LSTM的预测效果仅次于CAE-Learning,其RMSE最低可以达到9.173。CAE-Learning由于其结合了基于多层LSTM的自编码网络和多维卷积CNN的优势进行污染物浓度预测任务,能够充分挖掘数据间的时空相关性,从表2及图6可以看出,与其他模型相比较,该模型在 $PM_{2.5}$ 预测方面具有最好的预测效果及性能, RMSE的值最小可以达到8.880, Corr值最高可达到0.980。

4.3 CAE-Learning的泛化性

为了验证模型的可迁移性以及泛化性能,选取杭州、苏州、重庆、北京作为验证CAE-Learning模型泛化性能的城市,做城市内多站点污染物浓度的联合预测,并与在上一节中预测效果仅次于CAE-Learning的CNN-LSTM进行比较,证明CAE-Learning的泛化效果。选取的城市的每个监测站点的数据特征和时间跨度与上海市监测站完全相同。

(1) 杭州市实验结果

实验将杭州和睦小学作为目标站点,预测该站点的 $PM_{2.5}$ 浓度变化,其余的滨江、西溪、千岛湖、下沙、卧龙桥、浙江农大、朝晖五区、临平镇、城厢镇、云栖,这10个站点作为相关站点,做时空预测。实验结果如图9、图10及表3所示。

(2) 苏州市实验结果

实验将苏州相城区作为目标站点,预测该站点的 $PM_{2.5}$ 浓度变化,其余的上方山、南门、彩香、轧钢厂、吴中区、苏州新区、苏州工业园区,这7个站点作为相关站点,做时空预测。实验结果如图11、图12和表4所示。

(3) 重庆市实验结果

实验将南坪作为目标站点,预测该站点的 $PM_{2.5}$ 浓度变化,其余的缙云山、高家花园、天生、两路、虎溪、唐家沱、茶园、白市驿、解放碑、杨家坪、空港、新山村、礼嘉、蔡家、鱼新街、南泉,这16个站点作为相关站点,做时空预测。实验结果如图13、图14和表5所示。

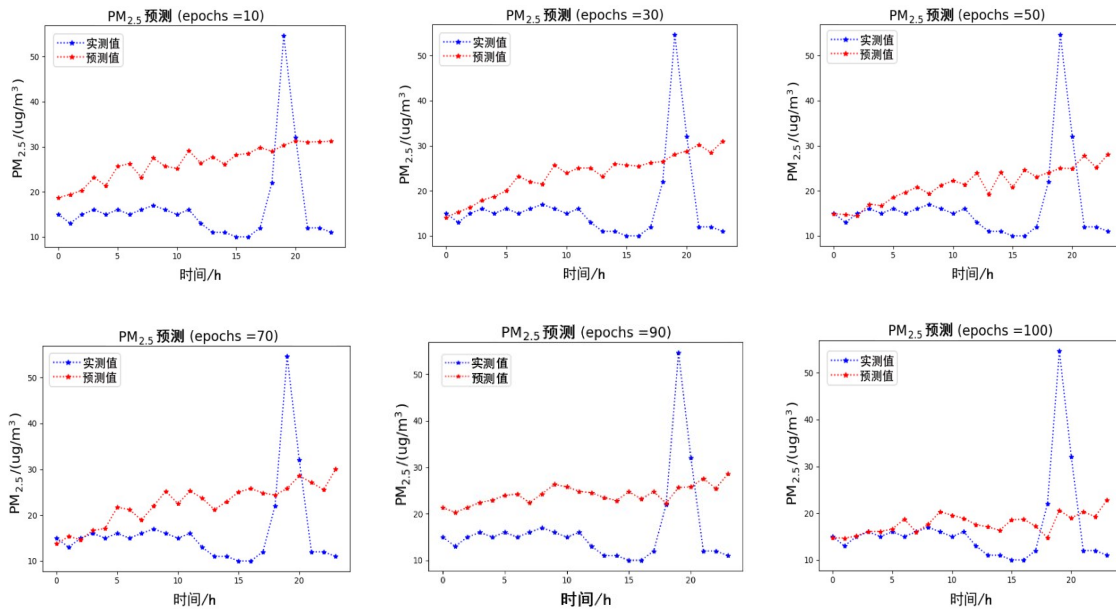


图9 CAE-Learning的多站点联合预测结果

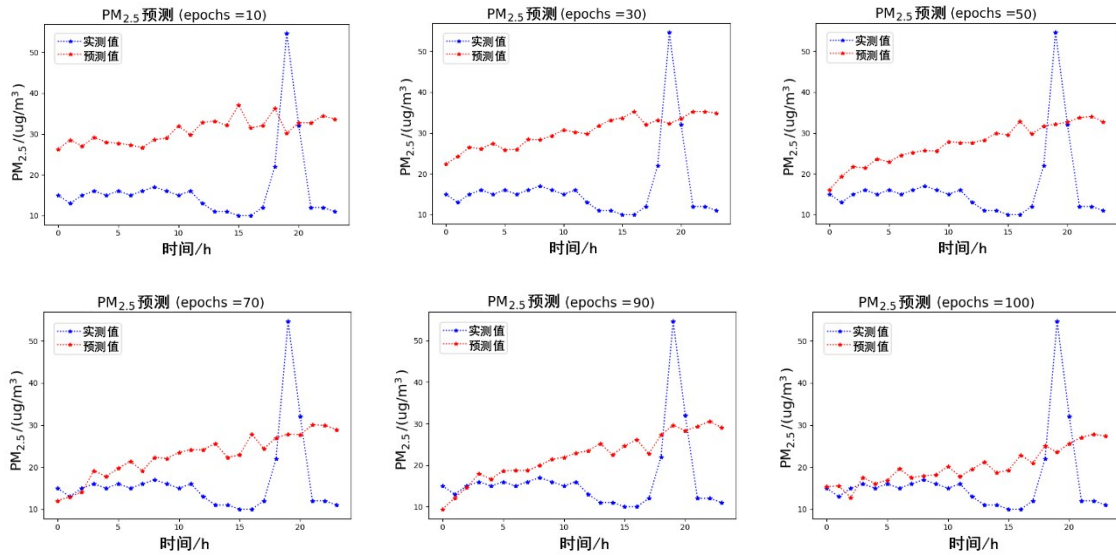


图10 CNN-LSTM的多站点联合预测结果

表3 模型在杭州市数据集下的RMSE、MAE、相关系数和训练时间

模型	RMSE	MAE	相关系数	训练时间
CNN-LSTM	23.153	21.909	0.899	1576 s
CAE-Learning	18.249	16.535	0.945	2785 s

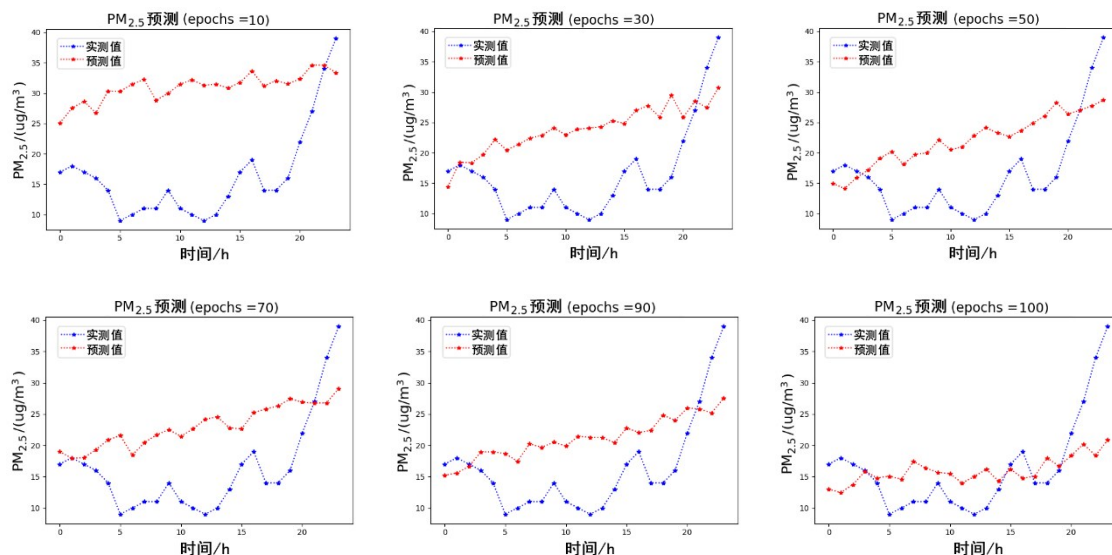


图 11 CAE-Learning 的多站点联合预测结果

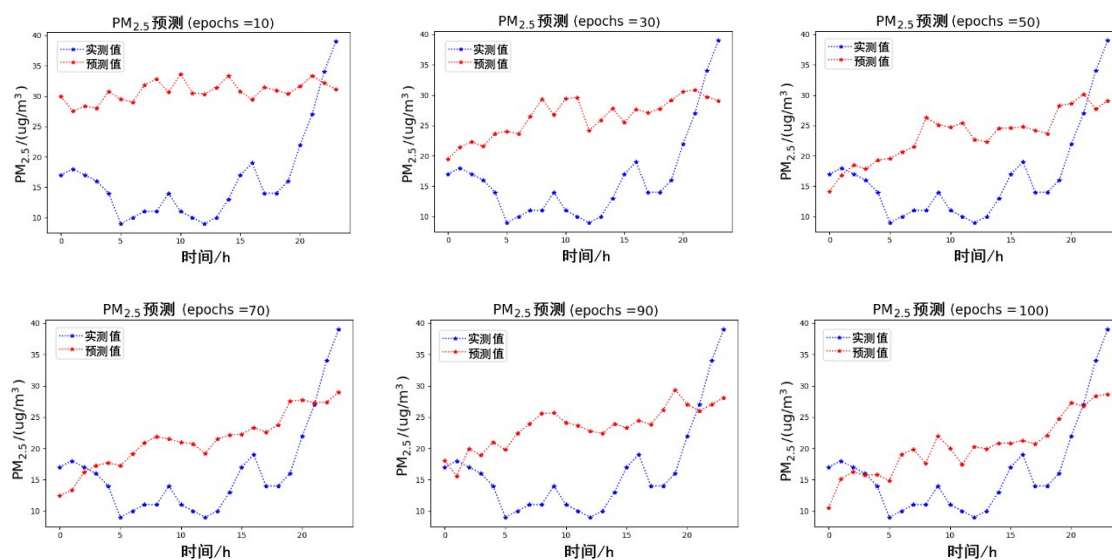


图 12 CNN-LSTM 的多站点联合预测结果

表 4 模型在苏州市数据集下的 RMSE、MAE、相关系数和训练时间

模型	RMSE	MAE	相关系数	训练时间
CNN-LSTM	24.529	20.693	0.893	1657 s
CAE-Learning	18.249	16.535	0.945	2269 s

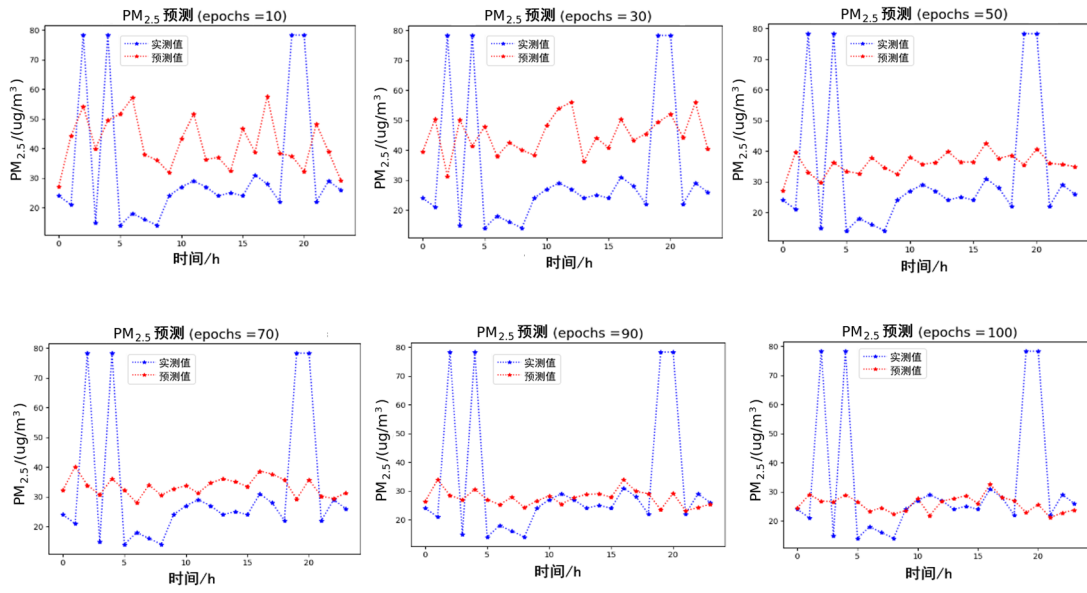


图 13 CAE-Learning 的多站点联合预测结果

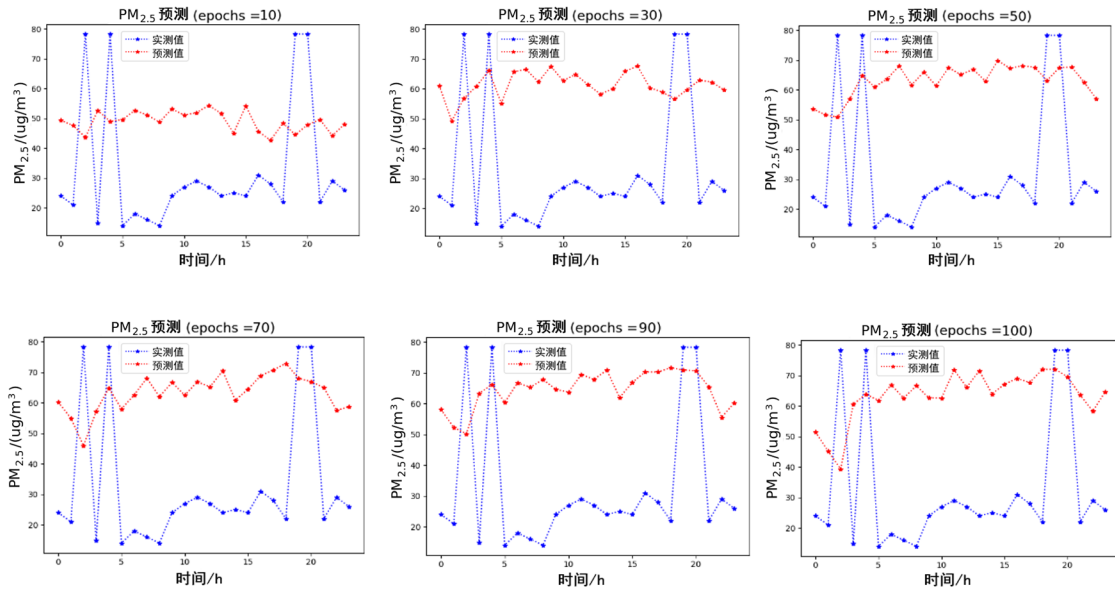


图 14 CNN-LSTM 的多站点联合预测结果

表 5 模型在重庆市数据集下的 RMSE、MAE、相关系数和训练时间

模型	RMSE	MAE	相关系数	训练时间
CNN-LSTM	37.881	28.744	0.407	1867 s
CAE-Learning	18.735	15.662	0.923	2487 s

(4)北京市实验结果

实验将顺义新城作为目标站点,预测该站点的PM_{2.5}浓度变化,其余的万寿西宫、定陵、东四、天坛、农

展馆、官园、海淀区万柳、怀柔镇、昌平镇、奥体中心、古城,这11个站点作为相关站点,做时空预测.实验结果如图15、图16和表6所示.

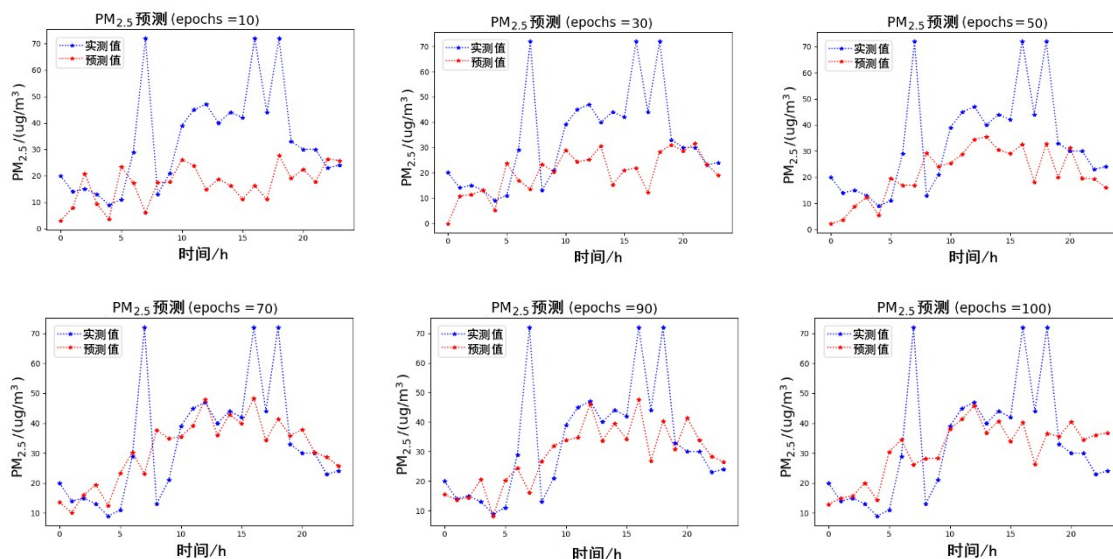


图15 CAE-Learning的多站点联合预测结果

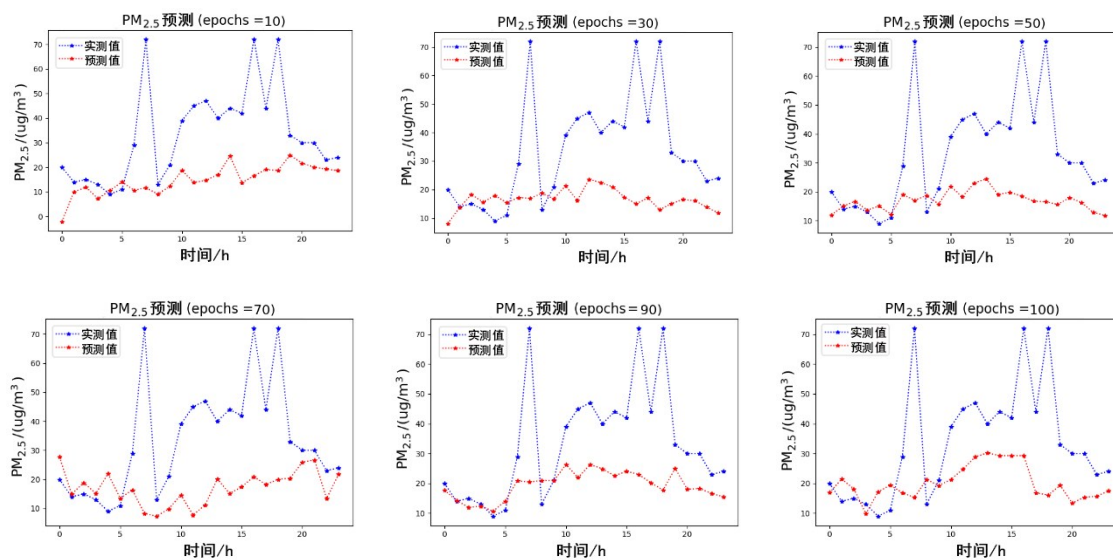


图16 CNN-LSTM的多站点联合预测结果

表6 模型在北京市数据集下的RMSE、MAE、相关系数和训练时间

模型	RMSE	MAE	相关系数	训练时间
CNN-LSTM	36.158	24.504	0.601	1578 s
CAE-Learning	18.727	15.260	0.903	2278 s

通过 CAE-Learning 及 CNN-LSTM 两个模型在杭州、苏州、重庆、北京这4个城市的空气污染浓度预测实验上的对比可以发现,本文提出的 CAE-Learning 模型迁移到其他城市做多站点联合预测时,还是可以充分

发挥其模型的内在优势的,能深度提取多站点间的时空关系,对长时间序列预测问题能够做到有效的预测.在迭代到100代的时候,不同城市多站点间的实验结果都表明:CAE-Learning 比 CNN-LSTM 对真实值的拟合程

度更高,预测效果更好.从表3、表6中列出的评价指标 RMSE、MAE 和相关系数 Corr 的比较中也验证了 CAE-Learning 的预测效果更好.

5 总结

本文利用城市内多站点的监测数据,提出了基于 CNN 和自编码网络的时空融合模型(CAE-Learning)来联合预测目标站点目标污染物浓度的方法. CNN 是模型的底部,用于提取空间特征,卷积和池化后获得数据之间的相关性.基于 LSTM 的自编码网络是模型的顶部,用于提取输入的时间序列特征.

CAE-Learning 模型对应的预测任务是用单城市多个监测站点的污染物浓度数据和气象数据作为模型的初始特征输入,来预测特定目标站点未来 N 小时内的污染物浓度.从基于真实数据的实验证明,对于时空融合的时间序列预测问题,CAE-Learning 和其他对比模型相比有很好的预测性能,在上海市的测试数据集上获得了不错的预测效果后,在杭州、苏州、重庆、北京这 4 个城市也有着很好的表现,具有较好的泛化能力.与传统模型相比,CAE-Learning 不仅考虑了空间关联性——城市内多站点之间污染物浓度的相互影响,而且考虑了污染物浓度的时间关联性——污染物浓度前后时间段内的相互影响,从而使实验的预测效果有了明显的提升.

该模型相对于传统机器学习方法和单一经典网络有较好的性能提升,已经多次在国家级区域空气污染监测预报任务中作为实际辅助模型之一得到应用,体现出较好的应用效果和价值.但由于数据方面的限制,该模型在不同地形、不同气候带、不同城市群特征、多模态数据等细化环境下的性能状态仍有待检验.因此,未来的研究工作将集中在多种类、多模态气象及污染物数据间的维度分析、语义提取、关联性分析及多个预测模型的融合方面,进一步提升最终预测的精确度.

参考文献

- [1] SCHRAUFNAGEL D E, BALMES J R, COWL C T, et al. Air pollution and noncommunicable diseases: A review by the Forum of International Respiratory Societies' Environmental Committee, Part 2: Air pollution and organ systems [J]. *Chest*, 2019, 155(2): 417-426.
- [2] ZOU B, YOU J, LIN Y, et al. Air pollution intervention and life-saving effect in China[J]. *Environment International*, 2019, 125: 529-541.
- [3] CHANG X, WANG S, ZHAO B, et al. Contributions of inter-city and regional transport to PM_{2.5} concentrations in the Beijing-Tianjin-Hebei region and its implications on regional joint air pollution control[J]. *Science of the Total Environment*, 2019, 660: 1191-1200.
- [4] LEONG W C, KELANI R O, AHMAD Z. Prediction of air pollution index(API) using support vector machine(SVM) [J]. *Journal of Environmental Chemical Engineering*, 2020, 8 (3): 103208.
- [5] MACIĄG P S, KASABOV N, KRYSZKIEWICZ M, et al. Air pollution prediction with clustering-based ensemble of evolving spiking neural networks and a case study for London area[J]. *Environmental Modelling & Software*, 2019, 118: 262-280.
- [6] CABANEROS S M S, CALAUTIT J K, HUGHES B R. A review of artificial neural network models for ambient air pollution prediction[J]. *Environmental Modelling & Software*, 2019, 119: 285-304.
- [7] CORANI G, SCANAGATTA M. Air pollution prediction via multi-label classification[J]. *Environmental Modelling & Software*, 2016, 80: 259-264.
- [8] LI A, XU X. A new PM_{2.5} air pollution forecasting model based on data mining and BP neural network model[C]//Proceedings of the 2018 3rd International Conference on Communications, Information Management and Network Security(CIMNS 2018). Berlin: Atlantis Press, 2018: doi:10.2991/cimns-18.2018.25.
- [9] FANG R, POUYANFAR S, YANG Y, et al. Computational health informatics in the big data age: A survey[J]. *ACM Computing Surveys*, 2016, 49(1): 1-36.
- [10] WAX A, ASECIO R, CARTER D R. Thinking big about big data[J]. *Industrial and Organizational Psychology*, 2015, 8(4): 545-550.
- [11] 秦喜文, 刘媛媛, 王新民, 等. 基于整体经验模态分解和支持向量回归的北京市 PM_{2.5} 预测[J]. *吉林大学学报(地球科学版)*, 2016, 46(2): 563-568.
- [12] QIN X W, LIU Y Y, WANG X M, et al. PM_{2.5} prediction of Beijing city based on ensemble empirical mode decomposition and support vector regression[J]. *Journal of Jilin University(Earth Science Edition)*, 2016, 46(2): 563-568. (in Chinese)
- [13] 桑海峰, 王传正, 吕应宇, 等. 基于多信息流动卷积神经网络的行人再识别[J]. *电子学报*, 2019, 47(2): 351-357.
- [14] SANG H F, WANG C Z, LV Y Y, et al. Person re-identification based on multi-information flow convolutional neural network[J]. *Acta Electronica Sinica*, 2019, 47(2): 351-357. (in Chinese)
- [15] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[EB/

- OL]. (2018)[2020]. <https://arxiv.org/abs/1810.04805v2>.
- [14] LEE J, YOON W, KIM S, et al. BioBERT: Pre-trained biomedical language representation model for biomedical text mining [EB/OL]. (2019)[2020]. <https://arxiv.org/abs/1901.08746v4>.
- [15] 廖勇, 花远肖, 姚海梅, 等. 高速移动环境下基于深度学习的信道估计方法[J]. 电子学报, 2019, 47(8): 1701-1707.
- LIAO Y, HUA Y X, YAO H M, et al. Channel estimation method based on deep learning in high-speed mobile environments[J]. *Acta Electronica Sinica*, 2019, 47(8): 1701-1707. (in Chinese)
- [16] GONG Z, ZHONG P, YU Y, et al. A CNN with multiscale convolution and diversified metric for hyperspectral image classification[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(6): 3599-3618.
- [17] 刘启超, 肖亮, 刘芳, 等. SSCDenseNet: 一种空-谱卷积稠密网络的高光谱图像分类算法[J]. 电子学报, 2020, 48(4): 751-762.
- LIU Q C, XIAO L, LIU F, et al. SSCDenseNet: A spectral-spatial convolutional dense network for hyperspectral image classification[J]. *Acta Electronica Sinica*, 2020, 48(4): 751-762. (in Chinese)
- [18] SÜLO I, KESKIN S R, DOGAN G, et al. Energy efficient smart buildings: LSTM neural networks for time series prediction[C]//2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML). New York: IEEE, 2019: 18-22.
- [19] ZHAO J, DENG F, CAI Y, et al. Long short-term memory-Fully connected(LSTM-FC) neural network for PM2.5 concentration prediction[J]. *Chemosphere*, 2019, 220: 486-492.
- [20] ZHAO P, ZETTSU K. Convolution recurrent neural networks based dynamic transboundary air pollution prediction[C]//2019 IEEE 4th International Conference on Big Data Analytics(ICBDA). New York: IEEE, 2019: 410-413.
- [21] 秦东明, 丁志军, 金玉鹏, 等. 基于自编码网络的空气污染物浓度预测[J]. 同济大学学报(自然科学版), 2019, 47(5): 681-687.
- QIN D M, DING Z J, JIN Y P, et al. An air pollutant prediction model based on auto-encoder network[J]. *Journal of Tongji University(Natural Science)*, 2019, 47(5): 681-687. (in Chinese)
- [22] FRANCK U, ODEH S, WIEDENSOHLER A, et al. The effect of particle size on cardiovascular disorders—the smaller the worse[J]. *Science of the Total Environment*, 2011, 409(20): 4217-4221.
- [23] CHEMEL C, FISHER B E A, KONG X, et al. Application of chemical transport model CMAQ to policy decisions regarding PM2.5 in the UK[J]. *Atmospheric Environment*, 2014, 82(1): 410-417.
- [24] WANG Y, KONG T. Air quality predictive modeling based on an improved decision tree in a weather-smart grid [J]. *IEEE Access*, 2019, 7: 172892-172901.
- [25] GNEITING T, RAFTERY A E, WESTVELD A H, et al. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation[J]. *Monthly Weather Review*, 2010, 133(5): 1098-1118.
- [26] BALACHANDRAN S, CHANG H H, PACHON J E, et al. Bayesian-based ensemble source apportionment of PM2.5[J]. *Environmental Science & Technology*, 2013, 47(23): 13511-13518.
- [27] SARVI F, NADALI A, KHODADOST M, et al. Application of poisson hidden Markov model to predict number of PM2.5 exceedance days in tehran during 2016-2017[J]. *Avicenna Journal of Environmental Health Engineering*, 2017, 4(1): 58031.
- [28] WANG M, YUAN Z, ZHANG X, et al. Construction of air quality evaluation system based on FCM algorithm and BP neural network[J]. *Agricultural Biotechnology*, 2018, 7(5): 274-276.
- [29] CHEN Y. Prediction algorithm of PM2.5 mass concentration based on adaptive BP neural network[J]. *Computing*, 2018, 100(8): 825-838.
- [30] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. *Nature*, 2015, 521(7553): 436-444.
- [31] LE V D, BUI T C, CHA S K. Spatiotemporal deep learning model for citywide air pollution interpolation and prediction[C]//2020 IEEE International Conference on Big Data and Smart Computing(BigComp). Busan: IEEE, 2020: 19534862.
- [32] WANG J, SONG G. A deep spatial-temporal ensemble model for air quality prediction[J]. *Neurocomputing*, 2018, 314: 198-206.
- [33] LIN Y, MAGO N, GAO Y, et al. Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning[C]//Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2018: 359-368.
- [34] QI Z, WANG T, SONG G, et al. Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2018, 30(12): 2285-2297.
- [35] WEN C C, LIU S F, YAO X J, et al. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction[J]. *Science of The Total Environment*,

2019, 654(1): 1091-1099.

- [36] ONG B T, SUGIURA K, ZETTSU K. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5[J]. Neural Computing and Applications, 2016, 27(6): 1553-1566.
- [37] HUANG C J, KUO P H. A deep CNN-LSTM model for particulate matter(PM2.5) forecasting in smart cities[J]. Sensors, 2018, 18(7): 2220.
- [38] KAREVAN Z, SUYKENS J A K. Transductive LSTM for time-series prediction: An application to weather forecasting[J]. Neural Networks, 2020, 125: 1-9.
- [39] 常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络[J]. 自动化学报, 2016, 42(9): 1300-1312.
- CHANG L, DENG X M, ZHOU M Q, et al. Convolutional neural networks in image understanding[J]. Acta Automatica Sinica, 2016, 42(9): 1300-1312. (in Chinese)

作者简介



张 波 男, 1979年11月出生, 江苏常州人. 同济大学计算机应用技术专业工学博士. 现为上海师范大学信息与机电工程学院院长、教授. 主要研究方向为新一代人工智能与深度学习技术等.

E-mail: zhangbo@shnu.edu.cn



陆云杰 男, 1995年6月出生, 江苏苏州人. 上海师范大学信息与机电工程学院计算机应用技术工学硕士. 主要研究方向为机器学习.

E-mail: 1000459475@smail.shnu.edu.cn



秦东明 男, 1981年4月出生, 河南社旗人. 同济大学电子与信息工程学院工学博士. 主要研究方向为高性能计算、人工智能、深度学习等.



邹国建 男, 1993年11月出生, 安徽滁州人. 上海师范大学信息与机电工程学院计算机应用技术专业工学硕士. 主要研究方向为深度学习、知识图谱.