

黄天文, 焦飞, 伍志方. 2024. 一种基于迁移学习和长短期记忆神经网络的降水预报方法[J]. 暴雨灾害, 43(1):45–53. HUANG Tianwen, JIAO Fei, WU Zhifang. 2024. A precipitation forecast method based on transfer learning and Long Short Term Memory [J]. *Torrential Rain and Disasters*, 43(1):45–53 (in Chinese). doi:10.12406/byzh.2023–118

一种基于迁移学习和长短期记忆神经网络的降水预报方法

黄天文¹, 焦飞², 伍志方³

(1. 广东省肇庆市气象局, 肇庆 526040; 2. 肇庆学院, 肇庆 526061; 3. 广东省气象台, 广州 510640)

摘要: 为给智能网格强降水预报提供客观参考, 提出了一种基于迁移学习和长短期记忆神经网络(LSTM)的降水预报方法。迁移学习是一种机器学习方法, 可将源域学习到的知识迁移到目标域中以应用; LSTM是一种可以处理序列数据中的长期依赖关系的深度学习模型。基于2009—2022年广东省肇庆市6个国家气象观测站逐小时雨量、气温、气压、相对湿度、风向和风速的观测资料, 以肇庆高要国家气象观测站作为目标域, 其它5个国家气象观测站作为源域, 利用迁移学习方法对目标域有缺失值的观测资料进行订正, 使目标域形成完整的训练样本; 然后, 利用深度学习方法, 对目标域分别建立单变量LSTM日雨量预报模型和多变量LSTM小时雨量预报模型; 最后, 对目标域2022年日雨量与小时雨量进行预报, 与对应实况对比。结果表明: (1) 单变量LSTM预报方法在1—2月、6月、10—12月的24 h晴雨预报准确率在80%以上, 多变量LSTM预报方法在3月、6月、8月、12月的1 h晴雨预报准确率在80%以上。 (2) 单变量LSTM预报方法仅6月能预报出24 h雨量在50 mm以上的强降水, 多变量LSTM预报方法在3月、5月、6—8月能预报出1 h雨量在20 mm以上的强降水, 其中3月和6月的小时雨量预报TS评分高于25%。

关键词: 降水预报; 时间序列; 迁移学习; LSTM; 深度学习

中图分类号: P457.6

文献标志码: A

DOI: 10.12406/byzh.2023–118

A precipitation forecast method based on transfer learning and Long Short Term Memory

HUANG Tianwen¹, JIAO Fei², WU Zhifang³

(1. Zhaoqing Meteorological Bureau, Zhaoqing 526040; 2. Zhaoqing University, Zhaoqing 526061;

3. Guangdong Meteorological Observatory, Guangzhou 510641)

Abstract: A precipitation forecasting method based on transfer learning and Long Short-Term Memory (LSTM) is proposed to provide an objective reference for intelligent grid heavy precipitation forecasting. Transfer learning is a machine learning method that can transfer knowledge learned from the source domain to the target domain for application. LSTM is a deep learning model that can handle long-term dependencies in sequence data and can remember long and short periods. In this study, the hourly observation data (rainfall, temperature, air pressure, relative humidity, wind direction, wind speed) from 2009 to 2022 of 6 meteorological observation stations in Zhaoqing City is used. The Gaoyao National Meteorological Observatory is selected as the target domain and the other 5 national meteorological observatories as the source domain, and the transfer learning method is used to transfer the source domain and correct missing values in the target domain. Then the complete training samples are classified in the target domain. Then, the deep learning methods are applied to establish the univariate LSTM daily rainfall prediction models and the multivariate LSTM hourly rainfall prediction models for the target domain, respectively. The daily and hourly rainfall forecast in the target domain for the year 2022 is compared with the actual observations. The results are as follows: (1) For the daily precipitation forecast of clear rain, the univariate LSTM method from January to February, June, and October to December can achieve an accuracy of over 80%, while for the hourly precipitation forecast of clear rain, the accuracy of the multivariate LSTM precipitation forecast method in March, June, August, and December can be over 80%. (2) The univariate LSTM method can only forecast precipitation with a 24-hour rainfall over 50 mm in June. The multivariate LSTM method can forecast precipitation with a 1-hour rainfall over 20 mm in March, May, and June to August, with the TS score in March and June being higher than 25%.

Key words: precipitation forecast; time series; transfer learning; LSTM; deep learning

收稿日期: 2023–06–12; 定稿日期: 2024–01–10

资助项目: 国家重点研发计划项目(2019YFC1510400); 广东省气象局科学技术研究项目(GRMC2020M21); 广东省肇庆市气象局科技创新团队(C202202); 广东省气象局科技创新团队(GRMCTD202004)

第一作者: 黄天文, 主要从事天气预报及计算机应用研究。E-mail: huangtw_zq@126.com

通信作者: 伍志方, 主要从事灾害性天气预报研究。E-mail: 1147942917@qq.com

引言

强降水常引发洪(内)涝、泥石流、滑坡等灾害,对人民生命财产造成重大损失,降水尤其是强降水预报一直是天气预报的重点和难点。随着经济社会的发展,对降水预报精细化需求越来越高。美国国家大气海洋局已构建较为成熟的定量降水预报(Quantitative Precipitation Forecast, QPF)格点化技术体系(Benjamin et al., 2016)。在我国,数值预报模式的发展主导着整个QPF业务的发展(李泽椿等, 2019),主客观融合、时空降尺度等技术成为当前网格化降水预报业务的主要支撑技术(苏翔等, 2020; 田笑等, 2022)。但根据时空降尺度的小时降水预报由于主要参考天气尺度信息,中尺度信息表现不足,另外受对流天气时空尺度较小、分布较不连续的特点和可预报性差的影响,还不能完全做出类似温度等要素预报的强降水短期预报(郑永光等, 2015; 湛芸等, 2021)。目前广东的小时降水预报仍以监测和短时临近预警为主要手段,短期时效内数值模式对于华南降水预报评分仍较低(伍志方等, 2018; 张华龙等, 2021)。2022年开始广东省各地市级气象台站每天制作空间分辨率5 km、未来3 d逐1 h时间间隔的晴雨和强降水预报等精细化网格预报产品。当前的地市级台站制作智能网格未来24 h逐小时强降水预报主要通过数值模式的降水预报进行主观订正,可供业务参考的客观方法相对较少。2023年上半年广东省级及市级强降水预报TS评分低于10%,因此,智能网格预报中小时强降水预报质量是当前急需解决的问题。

机器学习技术在现代社会的各个学科和领域中已经得到了广泛的应用,基于深度学习的时间序列预测在各个领域的研究得到高度重视(Chen et al., 2019)。许多学者致力于研究深度学习方法在强降水预报技术中的应用(路志英等, 2018; 张亚萍等, 2022; 张勇等, 2023)。利用机器学习方法研究观测、预测数据可以实现有效信息的提取,实现更准确的天气预报(蒋薇等, 2021)。逐时降雨量是一种内联相关性较强的时序数据,其某一时刻的数据与之前时刻的数据之间存在着密切的关系,具有循环神经网络架构的深度学习网络—长短期记忆神经网络(Long Short Term Memory, LSTM)模型被引入到降雨量预测领域(沈皓俊等, 2020; 崔忠捷等, 2022)。但如何基于海量观测和模式预报数据,利用以深度学习为代表的大数据挖掘技术等,提高小时降水预报准确率和精细化程度具有较大挑战性。

传统的机器学习通常需要大量有标记的训练样本,机器学习领域广泛关注的迁移学习(Transfer

Learning)正好解决了这一问题。目前,迁移学习的研究和应用主要集中在识别和分类上(Khamparia et al., 2020; Jiang et al., 2022),对时间序列预测的研究较少,尤其是基于实例的迁移学习的研发和应用。研究基于迁移学习的时间序列预测在现实世界中的应用具有挑战性,本研究参考了迁移学习在气象学中的应用等资料(Chen and Moschitti, 2019; Huang and Jiao, 2020; Li et al., 2021),提出了一种基于迁移学习和LSTM的短期时效的降水预报方法。利用2009—2022年广东省肇庆市6个国家气象观测站(高要、四会、德庆、封开、广宁和怀集)逐小时雨量、气温、气压、相对湿度、风向和风速的观测资料,以高要国家气象观测站(简称高要站)作为目标域,四会、德庆、封开、广宁和怀集5个国家气象观测站(简称其它5个站)作为源域进行研究,利用迁移学习方法对目标域有缺失值的观测资料进行订正,对补充完整的训练样本分别建立单变量LSTM日雨量预报模型和多变量LSTM小时雨量预报模型。选取对应的雨量样本进行训练,对目标域2022年日雨量与小时雨量进行预报,以期对预报员智能网格强降水预报提供客观参考。

1 资料与方法

1.1 资料说明与数据预处理

高要站选取2009—2022年逐小时地面观测数据,包括整点气温、气压、相对湿度、小时极大风向、小时极大风速、小时雨量等要素及1~24 h变温、1~24 h变压、1~24 h湿度变化、1 h风向变化、1 h风速变化、月平均气温距平、月平均气压距平、月平均相对湿度距平等。数据集包含84个特征变量,每小时记录一次观测值,构成 $84 \times 122\ 664$ 的矩阵。其它5个站选取2009—2022年逐小时地面观测数据,包括整点气温、气压、相对湿度、小时极大风向、小时极大风速、小时雨量。以发生降水前的气象要素作为样本,建立以深度学习为框架的降水预报模型。另外,由于需要用过去1 h的多个气象要素来预报现在的降水发生,原始资料采用雨量列错位方法,即当前小时雨量和过去1 h的其它气象要素值错位组合,作为当前小时记录。

过去由于设备故障、网络传输、人工错漏和一些不可抗拒的因素,不少气象台站观测数据存在缺失问题。2009—2022年期间高要站缺失值主要出现在2009—2013年,小时雨量、气温、湿度、气压、风向和风速共存在6 285个缺失值,将导致衍生的相关特征量(n 小时变温、变压、变湿、风的变化等)出现较多的缺失值,从而影响降水预报模型训练,为此,需对原始观测资料中的缺失值进行标注和预处理。高要站和

其它5个站均是肇庆的国家气象观测站,资料容易获取,站与站之间的距离适中,大部分气象要素的变化趋势一致。由于高要站出现缺失值时,其它5个站不曾同时出现缺失值,所以使用其它5个站样本资料迁移补充高要站的缺失资料。

我国天气预报业务中,短时强降水一般指1 h雨量 ≥ 20 mm的降水事件;暴雨是指24 h内累计降水量 ≥ 50 mm的天气事件(孙继松,2017)。降水具有突发性和不连续性,是非常不平稳的时间序列,尤其短时强降水,2009—2022年高要站共有142站次,仅占总样本的0.1%。降水时间序列难以建模、训练和拟合,需要对雨量值进行分级。小时雨量、日雨量分别按各自的标准阈值分为三级,无雨标记为0,有雨且雨量小于标准阈值标记为1,有雨且雨量大于或等于标准阈值标记为2。其中,2009—2022年高要站共有122 712

条小时记录,小时雨量0、1、2等级对应的样本量依次为111 869、10 701、142个。如无特殊说明,以下“雨量”指的是分级后的雨量,预报的是“雨量”的0、1、2等级。

1.2 研究方法

首先采用迁移学习方法,建立目标域(高要站)完整数据集。由于高要站出现缺失值时,其它5个站不曾同时缺失,所以可以将其它5个站逐小时资料对高要站的逐小时资料缺失值进行订正,不同特征变量(小时雨量、气温、气压、相对湿度和风速)根据特点采用不同的订正方法,最终得到包含雨量、气温、气压等多个特征的逐小时资料矩阵。然后,利用深度学习方法,针对补充完整的训练样本分别建立LSTM降水预报模型。本文提出的基于迁移学习的样本订正和基于深度学习的LSTM降水预报方法,流程图如图1所示。

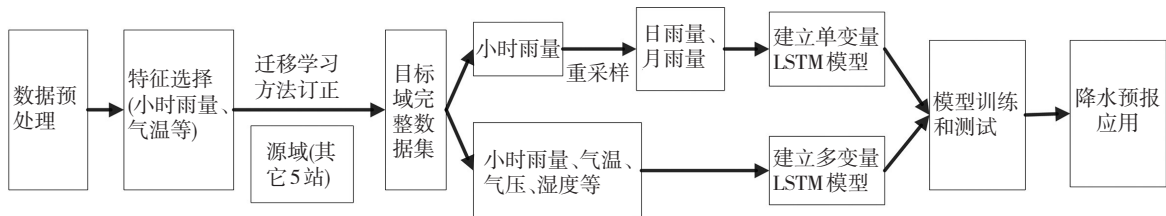


图1 一种迁移学习结合LSTM的降水预报流程图

Fig.1 A flow chart of precipitation forecast based on transfer learning and LSTM

2 基于迁移学习方法的样本订正

迁移学习是一种机器学习方法,有基于实例的迁移学习(样本迁移)、基于特征的迁移学习(特征迁移)、基于参数的迁移学习(模型迁移)和基于关系的迁移学习(关系迁移)。基于实例的迁移学习就是通过一些方法对源域的样本进行权重权衡,根据源域中和目标域样本的相似度来赋予高或低的权重,从而订正目标域存在缺失值的样本。迁移学习需要处理预测对象的目标域和迁移对象的源域之间的关系(闻克宇等,2020)。本文在样本订正时引入了基于实例的迁移学习方法。高要站作为目标域,是被订正的站点,资料的时间序列为目标序列,需要迁移的样本是作为源域(其它5个站)时间序列的同类数据,对于不同特征变量采用不同的样本迁移方法。

2.1 小时雨量补充订正

目标数据可以通过迁移学习的方法被订正,小时雨量可用偏差订正法,即统计不同阈值降水出现的实况频率和预报频率(王丽芳等,2021),根据两个站点之间的相关性使用线性相关分析方法(Cao et al., 2022)。小时雨量序列是离散数据,高要站雨量缺失值不能根据前、后时次的值进行判断、订正,但可以根据其它5个

站同时间点小时雨量与之内在的关联性,将相关性强的站点资料作为源数据迁移过来以订正高要站资料。

皮尔逊(Pearson)相关系数用于度量两个变量之间的线性相关程度,其值介于-1~1之间,其中,-1为完全负相关,0为没有线性相关性,1为完全正相关。线性相关系数的计算如公式(1)所示。

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (1)$$

在公式(1)中, r 表示两个时间序列的相关系数。 X 代表源数据(其它5个站)雨量时间序列, Y 代表目标数据(高要站)雨量时间序列。从高要站与其它5个站之间雨量时间序列的Pearson相关性中,高要站和四会站的相关性最高,作为目标数据的高要雨量和作为源数据的四会雨量具有相似的特征,并且具有较强的相关性,Pearson相关系数为0.701,相关系数在0.01水平上显著(双尾),可使用线性回归模型修订目标数据。

在样本迁移方法中,除了利用相关系数以外,站点之间的距离更为关键。图2直观反映了高要站与其它站点之间的相对位置。

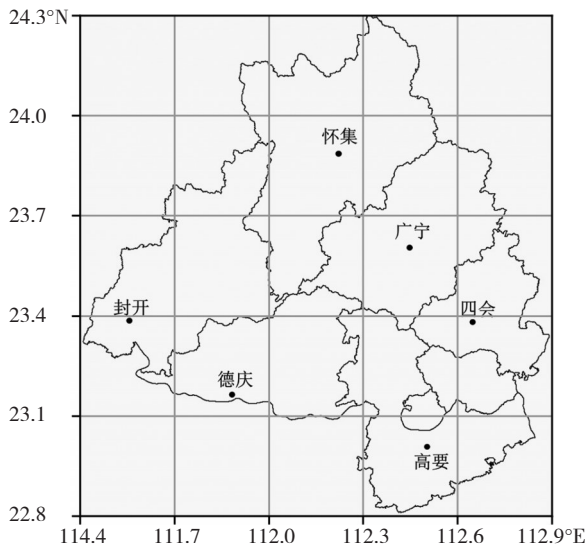


图2 肇庆市国家气象观测站分布图

Fig.2 Locations of Zhaoqing national meteorological observation stations

2.2 气温、气压、相对湿度和风速补充订正

对于气温、气压、相对湿度和风速样本的迁移,由于各自的时间序列的变化比较稳定,可以采用差值订正方法。以气温为例,分别对比高要站与其它5个站,每两站的气温差值变化很小,几乎可以认为是一个常数。设 X 站为源域(资料相对完整),具有 N 年资料; Y 站为目标域(资料存在缺失值),有 n 年资料; $n \leq N$,平均差值记为 d_n ,差值订正如公式(2)所示

$$Y_i = d_n + X_i \quad (2)$$

订正公式的适当性标准,用两站气温的相关系数 r 表示。试验结果表明,气温订正的相关系数大于0.7,效果良好(Huang and Jiao, 2017)。

同样,气压、相对湿度和风速等要素均具有相应差值稳定的特点,可采用差值订正方法进行订正。

3 LSTM建模与应用效果

3.1 LSTM模型

通过迁移学习,高要站获得了完整的训练样本,即可选择合适的模型进行预测。本文采用了神经网络中的LSTM模型(Swathi et al., 2022)。LSTM是由循环神经网络RNN(Recurrent Neural Network)改进而来,它具有强大的记忆功能和序列数据处理能力。与原始RNN的隐藏层相比,LSTM增加了细胞状态,作为一种神经网络,负责计算时间序列中各种观测值之间的相关性(Gul et al., 2021; Yu et al., 2021)。在时间序列预测中,LSTM既可以多变量预测机制又可以单变量预测机制使用。LSTM模型具有长期记忆功能且容易实现,可以处理序列数据中的长期依赖关系,从而捕捉到数据中的复杂模式和规律,LSTM的带门结构可以解决长序列训练中梯度消失和梯度爆炸的问题,更适用于离

散型时间序列的预测(Abbasimehr and Paki, 2022)。参考Hochreiter和Schmidhuber(1997)的研究,绘制LSTM网络结构图,如图3所示。

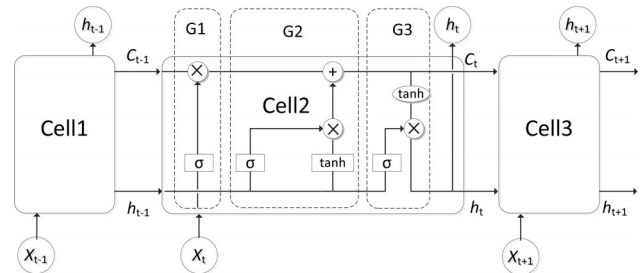


图3 LSTM网络结构图

(Cell代表一个LSTM单元,Cell1、Cell2、Cell3分别表示在上一时刻、当前时刻、下一时刻的一个LSTM单元; C 表示Cell在不同时序时候的状态; X 为输入, h 为输出; $t-1, t$ 和 $t+1$ 分别代表上一个时间、当前时间和下一个时间; $G1, G2$ 和 $G3$ 分别代表忘记门、输入门和输出门;

σ 表示sigmoid函数; \tanh 为激活函数; \times 代表逐点相乘操作,
+代表逐点相加操作)

Fig.3 Structure diagram of LSTM network

(Cell represents an LSTM unit, Cell1, Cell2, and Cell3 represent an LSTM unit at the previous time, current time and next time. C represents the state of the Cell at different time sequences. X is the input, h is the output. $t-1, t$, and $t+1$ represent the previous time, current time, and next time. $G1, G2$, and $G3$ represent forget gate, input gate, and output gate. σ is the sigmoid function. \tanh is the activation function. \times represents the operation of point by point multiplication, $+$ represents the operation of point by point addition)

图3中,Cell具有细胞的状态。忘记门、输入门和输出门三个门可以实现选择性地让信息通过,其中忘记门作用是控制上一个时间的状态 C_{t-1} 中哪些数据应该被“遗忘”,输入门的作用是决定让多少新的信息加入到细胞状态中,输出门的作用是基于细胞状态来确定要输出什么值。 σ 常被用作神经网络的阈值函数,将变量映射到0到1之间; \tanh 让 C_t 中的每一个值都介于-1和1之间。

采用的LSTM输入输出结构是多对一的。分为两个方案:(1)单变量单步,使用一个特征(雨量)训练模型,并使用该模型做预报,利用前 n 天的数据预报第 $n+1$ 天的数据。(2)多变量多步,使用过去的气温、气压、湿度等多个特征的值来预报未来一个时间点的小时雨量,然后把前一个输出也作为特征,预报未来多个时间点的小时雨量。

本方法通过特征方程将原始数据转化为特征,以地面观测站点所观测的数据为基础,根据观测到的数据人工合成“新”的样本,得到84个变量,但要去除噪声,即删除那些模型输入和模型输出不相关的部分变量。最后,分别建立单变量LSTM降水预报模型和多变量LSTM降水预报模型。采用2009—2021年的数据

进行模型训练,其中训练集、验证集比例为8:2,测试集采用2022年的数据,由于降水出现的概率相对整个样本较小,从中固定抽取30个尽可能多的包含降水和强降雨的记录作为测试集。

3.2 单变量LSTM日雨量预报

小时雨量样本由于存在太多0值且突变性强,相对不稳定,加上误差的累积,预测效果极差,而日雨量的变化趋势相对稳定,因此将订正后的小时雨量时间序列进行重采样,转换成日雨量。用日雨量作为研究对象,采用单变量LSTM预报方法。即仅雨量一个变量, t 时刻的预测值取决于 $t-1$ 、 $t-2$ 、……等之前时刻的历史序列,然后将 t 时刻的预测结果放入历史序列,再预测 $t+1$ 时刻的值,迭代重复这个过程预测未来多个时刻的值。深度学习模型的损失函数(loss function)是评价模型性能的重要指标。损失函数用来评估模型预测值与真实值的偏离程度,损失函数越小,一般就代表模型的鲁棒性(Robustness)越好。

经过订正的高要站日雨量时间序列,对整个序列使用单变量LSTM预报方法,2009—2021年单变量LSTM预报日雨量的模型训练中的损失函数曲线如图4所示,训练损失函数曲线快速下降并逐渐趋于零,说明模型拟合度较高。

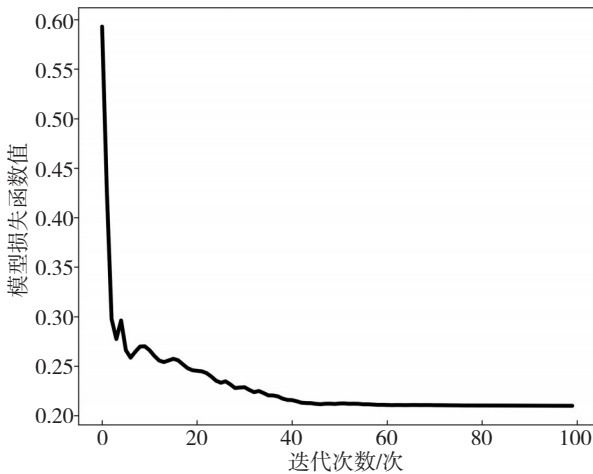


图4 2009—2021年单变量LSTM预报日雨量的模型训练中的损失函数曲线图

Fig.4 Loss function curve of univariate LSTM model training in daily precipitation forecast from 2009 to 2021

图5为2009—2022年逐年日降水预报模型训练、验证和测试效果图。可见整体上变化的趋势拟合效果不错,强降雨命中率较高,虽然也存在漏报,但可以根据实际进行量化分析,降低漏报率。

对逐年各月的日雨量进行预报。以6月份为例,首先筛选出2009—2021年每年该月份的日雨量数据,选取单个变量“日雨量”作为特征进行归一化操作,将数据缩放到-1和1之间。这里使用的是最小-最大归

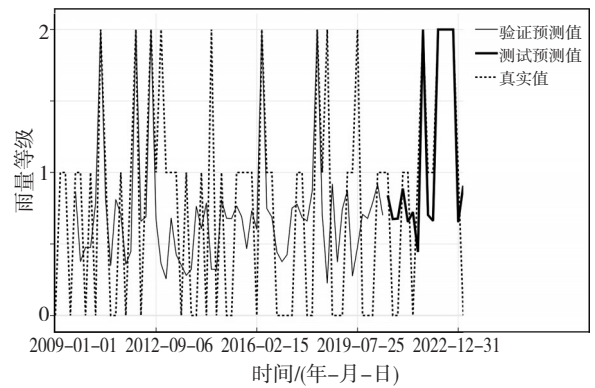


图5 2009—2022年逐年日降水预报模型训练、验证和测试效果图
Fig.5 Model training, validation, and testing effectiveness of daily precipitation forecast every year from 2009 to 2022

一化(Min-Max Scaling)方法:首先计算特征 X 的最大值 Max 和最小值 Min ,然后得到变换后的数值 $\hat{X}_i = (X_i - Min) / (Max - Min)$, \hat{X}_i 缩放到-1和1之间,当 X_i 为最大值时, \hat{X}_i 归一化为1,当 X_i 为最小值时, \hat{X}_i 归一化为-1。然后,按照8:2进行训练集、验证集划分,并设定观察的跨度为20 d,利用前一天的日雨量预测后一天的日雨量。接下来构建模型,输入的维度为1,即只有日雨量一个变量,隐藏层特征维度为32,循环的层数为2。模型训练和预测的原理就是利用前 n 天的数据预测第 $n+1$ 天的数据,即采用多个滑动窗口的数据预测未来一个数据。最后按以下步骤:(1)根据训练窗口的长度获取最新的有效序列;(2)将最新的序列输入到模型中,并预测下一个值;(3)将预测值附加到历史记录中。重复循环上述步骤,即可预测未来多天的日雨量。基于单变量LSTM降水预报模型缺点是误差会积累,需要注意检查误差在可接受的范围。

为了更直观地查看预报准确率,用 N_A 表示预报发生,且实况发生的站点数(命中), N_B 表示预报发生而实况未发生的站点数(空报), N_C 表示预报不发生而实况发生的站点数(漏报), N_D 表示预报不发生,实况未发生的站点数。预报检验标准用晴雨准确率(记为 P_C)、TS评分(记为 T_S)、空报率(记为 R_{FA})和漏报率(记为 P_O),见公式(3)~(6)。

$$P_C = \frac{N_A + N_D}{N_A + N_B + N_C + N_D} \quad (3)$$

$$T_S = \frac{N_A}{N_A + N_B + N_C} \quad (4)$$

$$R_{FA} = \frac{N_B}{N_A + N_B} \quad (5)$$

$$P_O = \frac{N_C}{N_A + N_C} \quad (6)$$

对各月内每日分为无降水(标记为0)和有降水(标记为1或2)两类进行预报。表1为各月日雨量预报

中,有/无降水的结果。其中 N 为样本数。单变量 LSTM 预报方法在 1—2 月、6 月、10—12 月 P_c 在 80% 以上。

表 1 2022 年单变量 LSTM 各月日降水(有/无降水)预报结果

Table 1 Results of daily precipitation(with/without precipitation) forecast based on univariate LSTM method every month in 2022

月份	样本数	准确率/%	TS 评分/%	空报率/%	漏报率/%
1	326	86.5	57.3	19.2	33.7
2	298	88.3	67.0	10.1	27.6
3	315	74.3	56.2	28.3	27.8
4	303	73.3	59.1	28.7	22.5
5	321	67.0	56.7	29.4	25.7
6	318	84.0	78.7	17.2	6.0
7	326	70.6	60.8	31.0	16.3
8	330	64.2	46.1	38.4	35.3
9	316	69.9	37.1	39.1	51.3
10	329	84.8	19.4	50.0	76.0
11	320	84.4	45.7	37.3	37.3
12	324	85.5	43.4	28.0	47.8

对于无降水(标记为 0)、一般降水(标记为 1)和强降水(标记为 2)三种等级的预报,单变量 LSTM 预报方法结果不太理想。由于暴雨在全年的出现几率较小,暴雨预报准确率不考虑 N_D 情况,看暴雨 T_s 即可。由于 1—2 月、10—12 月实况出现强降水的概率极小,表 2 仅列出单变量 LSTM 预报 3—9 月强降水预报结果。

2022 年 6 月份的 318 个测试样本中实际上暴雨出现 9 次,命中(N_A) 4 次,空报(N_B) 1 次,漏报(N_C) 5 次。基于单变量的 LSTM 强降水预报方法仅 6 月份预报出强降水,暴雨 T_s 为 40%,其它月份都不能预报出强降水。

3.3 多变量 LSTM 小时雨量预报

以小时雨量为研究对象,采用多变量 LSTM 预报

表 2 2022 年单变量 LSTM 各月日降水(有/无强降水)预报结果

Table 2 Results of daily precipitation (with/without heavy precipitation) forecast based on univariate LSTM method every month in 2022

月份	TS 评分/%	空报率/%	漏报率/%
3	0.0	-	100.0
4	0.0	-	100.0
5	0.0	-	100.0
6	40.0	20.0	55.6
7	0.0	-	100.0
8	0.0	-	100.0
9	0.0	-	100.0

注:“-”表示当月命中、空报均为 0,当月日降水没有出现暴雨。

方法。除了要预测的作为标签的雨量列,还有其它不同类型的特征值,目标数据筛选出有较高依赖性的一些特征。除了“风向”为分类类型,气温、气压、相对湿度等列均为数值类型。雨量列不仅取决于它以前的值,还取决于其它特征。因此,要预报未来的小时雨量值(目标列),必须考虑包括目标列在内的所有列。

图 6 为 2009—2021 年 12 月小时雨量在多变量 LSTM 预报方法中模型训练的训练损失和验证损失曲线图。可见二者的差距很小,说明该神经网络性能不错。但一般情况下因为网络不可避免地存在一定程度上过拟合,所以训练损失曲线略低于验证损失曲线。

图 6a 中两条曲线震荡剧烈,排除数据集太少和没有经过标准化处理两个原因,适当增大每次训练所抓取的数据样本数量(batch_size)。如图 6b 所示,将 batch_size 从原来的 40 改为 180 后曲线震荡问题得以解决。当数据集中的全部数据样本通过神经网络一次并且返回一次的过程即完成一次训练,称为一个 epoch,经过反复调整各参数,12 月小时雨量预报模型

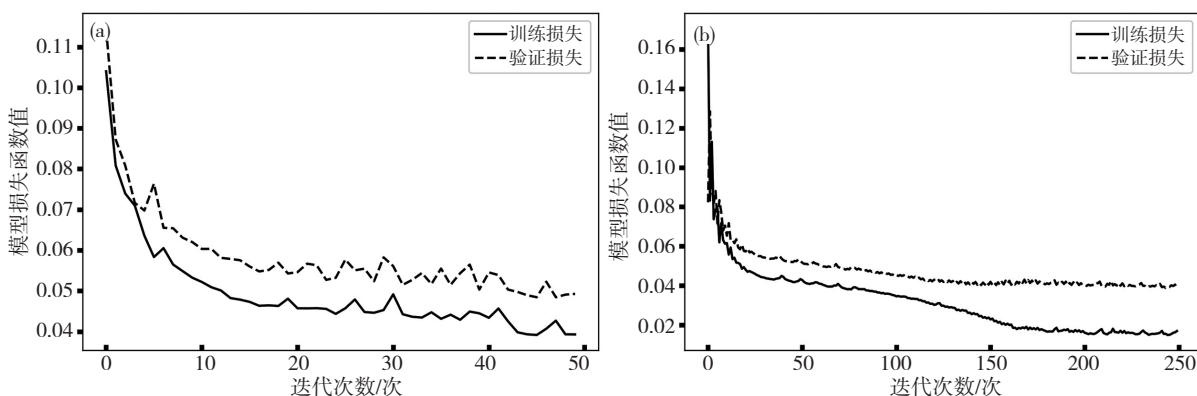


图 6 震荡 (a)和减小震荡后(b)的 2009—2021 年 12 月小时雨量预报模型的训练和验证损失曲线图

Fig.6 Model training and validation loss of hourly precipitation forecast in December from 2009 to 2021 of (a) before and (b) after reduction of oscillation

参数最终选择 epoch 从 50 增大到 250, 图 6b 中训练损失曲线会进一步下降, 验证损失曲线主要是趋于不变, 中间会稍微上升。其它各月模型参数均采用同样办法, 反复试验, 不断调整参数以达到最优预测效果。

训练集用来训练参数, 验证集是在每个 epoch 完成后, 用来验证当前模型的准确率。因为验证集和训练集没有交集, 因此这个准确率是可靠的。多变量 LSTM 降水预报模型训练完成后, 使用的测试集样本数为 30。由于一些月份降水偏少, 而在降水较多的月份, 小时雨量达到强降水也很少连续出现。全部样本中出现强降水的几率极少, 大部分样本为无降水, 样本正负比例极其不平衡, 对比例大的样本造成过拟合, 预报偏向无降水, 这样就大大降低了模型的泛化能力。本试验采用合成样本方法, 测试集长度统一为 30 个, 从最近时间选取有降水或强降水的记录, 按原次序合成新的样本, 使得降水或强降水的记录尽可能达到 1~3 次。

按有/无降水两个级别进行预报, 各月的小时降水预报结果如表 3 所示。多变量 LSTM 预报方法在 3 月、6 月、8 月、12 月 P_c 在 80% 以上。

表 3 2022 年多变量 LSTM 各月小时降水(有/无降水)预报结果
Table 3 Results of hourly precipitation (with/without precipitation) forecast based on multivariate LSTM method every month in 2022

月份	准确率/%	TS 评分/%	空报率/%	漏报率/%
1	70.0	50.0	50.0	0.0
2	53.3	39.1	55.0	25.0
3	86.7	55.6	0.0	44.4
4	60.0	45.5	33.3	41.2
5	70.0	62.5	28.6	16.7
6	80.0	40.0	55.6	20.0
7	70.0	18.2	71.4	66.7
8	96.7	50.0	0.0	50.0
9	70.0	40.0	14.3	57.1
10	73.3	50.0	20.0	42.9
11	60.0	60.0	40.0	0.0
12	93.3	81.8	0.0	18.2

按无降水、一般降水、强降水三个级别进行预报, 多变量 LSTM 预报方法的强降水预报结果优于单变量 LSTM 预报方法预报结果。由于 1—2 月、10—12 月实况出现强降水的概率极小, 表 4 仅列出多变量 LSTM 预报方法对 2022 年 3—9 月强降水预报结果。

多变量 LSTM 预报方法在 3 月、6 月和 8 月暴雨 T_s 较高, 其中 3 月和 6 月 T_s 高于 25%, 5 月、7 月、8 月、9 月的空报率高, R_{FA} 在 90% 以上, 5 月、7 月、9 月的漏报率高, P_0 在 50% 以上, 其中 9 月 P_0 为 100%, 全部漏报。

表 4 2022 年多变量 LSTM 各月小时降水(有/无强降水)预报结果

Table 4 Results of hourly precipitation (with/without heavy precipitation) forecast based on multivariate LSTM method every month in 2022

月份	TS 评分/%	空报率/%	漏报率/%
3	25.0	75.0	0.0
4	-	100.0	-
5	6.7	92.9	50.0
6	36.4	60.0	20.0
7	0.1	90.0	50.0
8	8.3	91.7	0.0
9	0.0	100.0	100.0
3-9	14.1	86.2	30.8

注:“-”表示当月没有出现小时强降水。

4 结论与讨论

本文提出一种基于深度学习的长短期记忆神经网络(LSTM)降雨预报方法, 将迁移学习方法引入到深度学习网络模型中, 得到如下主要结论:

(1) 使用基于实例的迁移学习方法订正目标数据, 合理解决了训练数据集中缺失值的问题, 得到完整的训练样本, 为建立合适的 LSTM 预报模型提供了基础条件。

(2) 对于有/无降水预报, 单变量 LSTM 预报方法在 1—2 月、6 月、10—12 月的日雨量预报效果较好, P_c 在 80% 以上。多变量 LSTM 预报方法在 3 月、6 月、8 月、12 月的小时雨量预报效果较好, P_c 在 80% 以上。

(3) 针对强降水预报, 单变量 LSTM 预报方法不适用于日强降水预报, 仅 6 月份预报出强降水, 暴雨 T_s 为 40%, 其它月份均不能预报出强降水。多变量 LSTM 预报方法可用于小时强降水预报, 3 月、6 月和 8 月小时强降水 T_s 较高, 其中 3 月和 6 月 T_s 高于 25%。

部分月份预报效果不理想, 有的甚至不能预报出强降水。下一步研究除了适当增加邻近区域自动气象站降水观测资料, 还将加入不同高度层相关物理量, 融合多源数据, 形成立体环境变量输入。另外, 将每一时刻 84 个特征变量全部导入模型进行训练, 模型维度过高, 导致训练速度慢, 训练效率低下。后期拟对特征变量进行择优选择, 筛选出对降水具有显著影响的特征变量继续进行研究。针对短期时效内的降水预报时效性问题, 下一步考虑是将雨量列错位 3 h (6 h), 将当前小时雨量和过去 3 h (6 h) 的其它气象要素值进行组合, 尝试提前 3 h (6 h) 预报未来 24 h 逐小时强降水。当然预报时效越长, 所能够预报的精细化程度和准确率相对会越低。后期将尝试研究基于模型的迁

移学习在深度学习中的应用,尝试改进 LSTM 模型,并在智能网格预报平台上进行格点应用,希望提高一般降水和强降水的命中率,为预报员短期预报的一般降水和强降水落区预报提供参考。利用迁移学习对目标域数据存在缺失值进行订正,对补充完整目标域数据集建立 LSTM 降水预报模型方法可适用国内各台站。对不同地理位置的站点进行建模还需要不断调整参数以达到最优。

参考文献(References):

- 湛芸,曹勇,孙健,等.2021.中央气象台精细化网格降水预报技术的发展和思考[J].气象,47(6):655-670. Chen Y, Cao Y, Sun J, et al. 2021. Progress of fine gridded quantitative precipitation forecast technology of national meteorological centre [J]. Meteorological Monthly,47(6): 655-670 (in Chinese). doi:10.7519/j.issn.1000-0526.2021.06.002
- 崔忠捷,卿晓霞,杨森雄.2022.基于 DM-LSTM 的城市降雨径流预测研究[J].中国给水排水,38(19):132-138. Cui Z J, Qing X X, Yang S X. 2022. Prediction of urban rainfall runoff based on DM-LSTM [J]. China Water and Wastewater,38(19):132-138 (in Chinese). doi:10.19853/j.zgjsps.1000-4602.2022.19.021
- 蒋薇,刘芸芸,陈鹏,等.2021.利用深度神经网络和先兆信号的江苏夏季降水客观预测方法[J].气象学报,2021,79(6):1035-1048. Jiang W, Liu Y Y, Chen P, et al. 2021. Prediction of summer precipitation in Jiangsu province based on precursory factors:A deep neural network approach [J]. Acta Meteorologica Sinica,79(6):1035-1048 (in Chinese). doi:10.11676/qxb2021.057
- 李泽椿,湛芸,张夕迪,等.2019.中央气象台暴雨预报业务的发展及思考[J].暴雨灾害,38(5):407-415. Li Z C, Chen Y, Zhang X D, et al. 2019. Development and perspectives on torrential rain forecasting operation in National Meteorological Center [J]. Torrential Rain and Disasters,38(5): 407-415 (in Chinese). doi:10.3969/j.issn.1004-9045.2019.05.002
- 路志英,任一墨,孙晓磊,等.2018.基于深度学习的短时强降水天气识别[J].天津大学学报,51(2):111-119. Lu Z Y, Ren Y M, Sun X L, et al. 2018. Recognition of short-time heavy rainfall based on deep learning [J]. Journal of Tianjin University,51(2):111-119 (in Chinese). doi:10.11784/tdxbz201703106
- 沈皓俊,罗勇,赵宗慈,等.2020.基于 LSTM 网络的中国夏季降水预测研究[J].气候变化研究进展,16(3):263-275. Shen H J, Luo Y, Zhao Z C, et al. 2020. Prediction of summer precipitation in China based on LSTM network [J]. Climate Change Research,16(3):263-275 (in Chinese). doi:10.12006/j.issn.1673-1719.2019.067
- 苏翔,袁慧玲.2020.集合预报统计学后处理技术研究进展[J].气象科技进展,10(2):30-41. Su X, Yuan H L. 2020. The research progress of ensemble statistical postprocessing methods [J]. Advances in Meteorological Science and Technology,10(2):30-41 (in Chinese). doi:10.3969/j.issn.2095-1973.2020.02.005
- 孙继松.2017.短时强降水和暴雨的区别与联系[J].暴雨灾害,36(6):498-506. Sun J S. 2017. Differences and relationship between flash heavy rain and heavy rainfall [J]. Torrential Rain and Disasters,36(6):498-506 (in Chinese). doi:10.3969/j.issn.1004-9045.2017.06.002
- 田笑,余文韬,从靖,等.2022.统计降尺度方法在天津小时降水和气温精细化预报中的应用[J].干旱气象,40(1):135-145. Tian X, Yu W T, Cong J, et al. 2022. Fine prediction of hourly precipitation and air temperature of Tianjin based on statistical downscaling in ECMWF model [J]. Journal of Arid Meteorology,40(1):135-145 (in Chinese). doi:10.11755/j.issn.1006-7639(2022)-01-0135
- 王丽芳,漆梁波,周伟,等.2021.频率匹配技术在暴雨预报中的应用及改进分析[J].暴雨灾害,40(4):352-361. Wang L F, Qi L B, Zhou W, et al. 2021. On the application and improvement of the frequency matching method to rainstorm forecasts [J]. Torrential Rain and Disasters,40(4): 352-361 (in Chinese). doi:10.3969/j.issn.1004-9045.2021.04.003
- 闻克宇,赵国堂,何必胜,等.2020.基于改进迁移学习的高速铁路短期客流时间序列预测方法[J].系统工程,38(3):73-83. Wen K Y, Zhao G T, He B S, et al. 2020. An improved transfer learning based time series prediction method for the high-speed rail short-term volume. Systems Engineering,38(3):73-83 (in Chinese).
- 伍志方,蔡景就,林良勋,等.2018.2017年广州“5·7”暖区特大暴雨的中尺度系统和可预报性[J].气象,44(4):485-499. Wu Z F, Cai J J, Lin L X, et al. 2018. Analysis of Mesoscale Systems and Predictability of the Torrential Rain Process in Guangzhou on 7 May 2017 [J]. Meteorological Monthly,44(4):485-499 (in Chinese). doi:10.7519/j.issn.1000-0526.2018.04.002
- 张华龙,伍志方,肖柳斯,等.2021.基于因子分析的广东省短时强降水预报模型及其业务试验[J].气象学报,79(1):15-30. Zhang H L, Wu Z F, Xiao L S, et al. 2021. A probabilistic forecast model of short-time heavy rainfall in Guangdong province based on factor analysis and its operational experiments [J]. Acta Meteorologica Sinica,79(1):15-30 (in Chinese). doi:10.11676/qxb2020.076
- 张亚萍,张焱,翟丹华,等.2022.深度学习方法在强降水临近预报技术中的应用及其改进思考[J].暴雨灾害,41(5):506-514. Zhang Y P, Zhang Y, Zhai D H, et al. 2022. Thoughts on application and improvement of deep learning in severe precipitation nowcasting technology [J]. Torrential Rain and Disasters,41(5):506-514 (in Chinese). doi:10.12406/byzh.2022-079
- 张勇,刘慧,郑颖菲,等.2023.人工智能模型的分类临近预报产品效果检验与分析[J].沙漠与绿洲气象,17(1):115-121. Zhang Y, Liu H, Zheng Y F, et al. 2023. Verification and analysis of the effectiveness of artificial intelligence models in classifying proximity prediction products [J]. Desert and Oasis Meteorology,17(1):115-121 (in Chinese). doi:10.12057/j.issn.1002-0799.2023.01.017
- 郑永光,周康辉,盛杰,等.2015.强对流天气监测预报预警技术进展[J].应用气象学报,26(6):641-657. Zheng Y G, Zhou K H, Sheng J, et al. 2015. Advances in techniques of monitoring, forecasting and warning of severe convective weather [J]. Journal of Applied Meteorological Science,26(6):641-657 (in Chinese). doi:10.11898/1001-7313.20150601
- Abbasimehr H, Paki R. 2022. Improving time series forecasting using LSTM and attention models [J]. Journal of Ambient Intelligence and Humanized Computing,13:673-691. doi:10.1007/s12652-020-02761-x
- Benjamin S G, Weygandt S S, Brown J M, et al. 2016. A North American hourly assimilation and model forecast cycle: the rapid refresh [J]. Monthly Weather Review,144(4):1669-1694. doi:10.1175/MWR-D-15-0242.1
- Cao W, Li H, Li Q. 2022. A method of thermal error prediction modeling for CNC machine tool spindle system based on linear correlation [J]. The

- International Journal of Advanced Manufacturing Technology,118: 3079–3090. doi:10.1007/s00170-021-08165-1
- Chen L Z, Moschitti A. 2019. Transfer learning for sequence labeling using source model and target data [J]. Proceedings of the AAAI Conference on Artificial Intelligence,33(1):6260–6267. doi:10.1609/aaai.v33i01.33016260
- Chen Y Y, Liu H T, Hsieh H L. 2019. Time series interval forecast using GM(1,1) and NGBM(1,1) models [J]. Soft Computing,23:1541–1555. doi:10.1007/s00500-017-2876-0
- Gul M J J, Firmansyah M H, Rho S, et al. 2021. BI-LSTM-LSTM based time series electricity consumption forecast for south korea [C]//Advances in Artificial Intelligence and Applied Cognitive Computing, 897–902. doi:10.1007/978-3-030-70296-0_71
- Hochreiter S, Schmidhuber J. 1997. Long short-term memory [J]. Neural Computation,9(8):1735–1780. doi:10.1162/neco.1997.9.8.1735
- Huang T W, Jiao F. 2017. Data transfer and extension for mining big meteorological data [C]//Lecture Notes in Computer Science,10361:57–66. doi:10.1007/978-3-319-63309-1_6
- Huang T W, Jiao F. 2020. Study on data transfer in meteorological Forecast of small and medium-sized cities and its application in zhaoqing city [J]. The Computer Journal,63(7):1076–1083. doi:10.1093/comjnl/bxz087
- Jiang L W, Quan H Y, Xie T, et al. 2022. Fish recognition in complex underwater scenes based on targeted sample transfer learning [J]. Multimed Tools and Applications,81(18):25303–25317. doi:10.1007/s11042-022-12525-8
- Khamparia A, Gupta D, de Albuquerque V H C, et al. 2020. Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning [J]. The Journal of Supercomputing,76:8590–8608. doi:10.1007/s11227-020-03159-4
- Li J P, Lo W L, Fu H, et al. 2021. A transfer learning method for meteorological visibility estimation based on feature fusion method [J]. Applied Sciences,11(3):997–1018. doi:10.3390/app11030997
- Swathi T, Kasiviswanath N, Rao A A. 2022. An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis [J]. Applied Intelligence,52:13675–13688. doi:10.1007/s10489-022-03175-2
- Yu Y, Xia X, Lang B, et al. 2021. PT-LSTM: Extending LSTM for efficient processing time attributes in time series prediction [C]. Lecture Notes in Computer Science,12858:450–464. doi:10.1007/978-3-030-85896-4_35

(责任编辑 何明琼)