文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

基于深度学习的高质量气象数据集研制与算法研究应用

张平
山东省气象数据中心
DOI:10.32629/acair.v3i4.17918

[摘　要] 高质量的气象数据对于精准天气预报、气候研究及应对极端天气事件至关重要。近年来,深度学习的发展为气象数据集的构建与算法优化提供了新机遇。本研究致力于探索基于深度学习的高质量气象数据集构建方法,并分析相关算法在气象预测等领域的应用潜力。通过融合与预处理多源气象数据,应用深度学习技术构建数据集,并开发相应的预测模型。实验结果表明,深度学习方法构建的气象数据集在完整性与准确性上显著提升,相应算法在气象预测任务中展现出更优异的性能与可靠性,为气象研究和业务应用提供了支撑。
[关键词] 深度学习；高质量气象数据集；算法开发；气象预测
中图分类号：S161　文献标识码：A

Research and Application of High-Quality Meteorological Dataset Development and Algorithms Based on Deep Learning

Ping Zhang
Shandong Meteorological Data Center

[Abstract] High-Quality meteorological data is essential for accurate weather forecasting, climate research, and responding to extreme weather events. In recent years, the development of deep learning has provided new opportunities for the construction of meteorological datasets and algorithm optimization. This study is committed to exploring methods for building high-quality meteorological datasets based on deep learning and analyzing the application potential of related algorithms in fields such as meteorological prediction. By fusing and preprocessing multi-source meteorological data, deep learning technology is applied to construct datasets and develop corresponding prediction models. Experimental results show that the meteorological dataset constructed by deep learning methods has significantly improved in integrity and accuracy, and the corresponding algorithms exhibit more excellent performance and reliability in meteorological prediction tasks, providing support for meteorological research and business applications.
[Key words] deep learning; high-quality meteorological dataset; algorithm development; meteorological prediction

引言

气象信息深刻影响人类社会多方面[1]。精准气象信息对生产生活、防灾减灾和国家安全至关重要[2]。气候变化加剧导致极端天气频发,对气象观测精度和预报可靠性要求更高。传统数据采集处理方法在精度、覆盖范围和处理效率上的局限性日益凸显。

深度学习技术在多个领域取得突破,为气象数据处理和预测提供了新途径[3]。它能高效整合处理多源气象数据,构建更高质量数据集；其预测模型能挖掘复杂规律,显著提升降水、气温等要素的预测精度,并增强台风、暴雨等极端天气预警能力。

深度学习在气象领域应用研究快速发展。国内外团队正推动技术研发：通过融合卫星、地面观测等多源数据,运用深度学习提升数据集质量；在气象要素和极端天气预测中,深度学习模型普遍优于传统方法。但该领域仍面临时空异质性、模型可解释性、数据安全等挑战亟待解决。

1 相关理论与技术基础

深度学习基本原理通过多层神经网络自动学习数据表征,从海量数据中提取复杂模式,其核心要素包括神经元、网络架构及训练算法。神经元作为基础计算单元,对加权输入求和后经非线性激活函数输出。不同层级神经元连接形成神经网络,典型结构包括前馈神经网络、卷积神经网络(CNN)、循环神经网络(RNN)及其衍生模型长短期记忆网络(LSTM)和门控循环单元(GRU)。训

练过程中,模型通过最小化损失函数优化网络参数,常用算法包括随机梯度下降(SGD)及其改进算法如Adagrad、Adadelta、RMSProp、Adam等。

气象数据具有时空异质性、多尺度性、高维度及不确定性等特征,主要类型包括地面气象观测数据(温压湿降水风等)、高空探测数据(温湿压风垂直信息)、卫星遥感数据(全球覆盖信息)及数值模式输出数据(含物理过程信息的模拟值)。

深度学习在气象领域的应用日益广泛,涵盖数据处理(如质量控制、插补与融合)、气象要素预测(如降水、气温、风速)、极端天气事件预警(如台风、暴雨预测)及气候模拟(改进物理过程参数化)等方面。尽管取得显著进展,仍面临如何有效融入物理知识、高效处理海量数据、提升模型泛化能力等挑战。

2 高质量气象数据集构建

2.1多源气象数据汇聚与融合

高质量气象数据集的研制依赖多源数据的收集与整合[4]。数据源主要包括全球分布的地面气象观测站,实时采集气温、气压、湿度、降水、风速风向等要素,具有高时间精度和局地代表性；高空探测(探空气球、飞机)获取大气温湿度、气压、风场的垂直结构信息,对理解大气动力热力过程至关重要;气象卫星提供全球云图、海表温度、植被指数等数据,具备大范围覆盖与高频观测优势;数值天气预报模型模拟未来气象状况,输出多种变量预测数据,虽含模型误差但蕴含丰富物理过程信息。针对这些数据在时空分辨率、格式及坐标系统上的差异,需进行有效整合:首先利用工具将非标准格式统一转换为NetCDF等通用格式;其次基于地理坐标,通过重采样和插值技术实现时空配准,统一数据网格;最后构建融合模型,如运用卡尔曼滤波融合地面与高空数据优化垂直结构描述,采用神经网络融合卫星与地面数据提升反演精度,并通过偏差校正整合数值模式输出数据,降低误差形成全面准确的数据集。

2.2基于深度学习的数据清洗与预处理

实践证明,深度学习技术在气象数据清洗与预处理中效果显著。以山东省近些年日照时数数据处理为例,研究团队融合深度学习和传统统计方法,显著改善数据质量。预处理阶段处理逻辑矛盾、重复记录及地理信息缺失,通过分层方案修正错误并统一格式。异常值检测结合自编码器和MRBP分析,准确识别异常值；修复采用局部线性插值或GAN生成替代值。数据增强使用滑动窗口和随机平移扩充样本,Z-score标准化处理。处理后数据缺失率降至1.8%,异常值修正准确率达92.6%,为后续分析奠定基础。

2.3气象数据集构建与评估

气象数据集采用分层存储架构：基础层保存预处理后的多源原始数据(地面观测、高空探测、卫星遥感、数值模式输出)；特征层通过特征工程提取气象预测相关特征并进行筛选组合;标签层按任务需求标注未来要素值或极端天气概率等信息；数据集按时间序列划分为训练集(60%-70%)、验证集(15%-20%)和测试集(15%-20%)以支持模型训练、调优及泛化评估。质量评估涵盖多维度指标：完整性通过时空缺失率衡量；准确性针对观测数据采用RMSE和MAE、针对模式数据结合相关系数分析；一致性通过相对偏差或一致性相关系数检验多源数据吻合度；模型有效性则通过训练收敛效率和验证性能间接评估。

3 深度学习在气象领域的算法研究应用

3.1气象要素预测算法

时空融合模型在气象要素预测中具有显著优势。以山东日照时数短期预测为例,研究者构建了融合BiLSTM与CNN的模型,有效提取了日照时数的时空特征。

模型架构：首先使用CNN提取全省站点日照的空间特征(如沿海-内陆差异、地形影响区域),生成128×128空间特征图；随后按时间序列输入BiLSTM,通过双向学习捕捉日照的日变化规律(如夏季正午峰值、冬季波动)及长期趋势(如雨季/旱季周期)。

训练与验证：数据集按7:3划分为训练集(2012-2018)和测试集(2019-2021),采用Adam优化器最小化MAE损失。结果显示,该模型对未来72小时日照预测的RMSE为0.86小时,较单一LSTM模型降低23.5%,在阴转晴、台风等复杂天气下对突变特征的捕捉更精准。

应用价值:模型预测结果已用于指导农业生产,例如为冬小麦灌浆期提供精准日照预测,辅助农户调整水肥策略,局部地区作物产量波动可减少15%以上。

该案例表明,深度学习技术在气象数据建模与预测算法开发中潜力显著,其核心在于以数据驱动方式挖掘气象要素复杂规律,并结合领域知识提升结果的可靠性与实用性。

3.2极端天气事件预警算法

3.2.1基于深度学习的极端天气特征提取

台风、暴雨、暴雪等极端天气事件强度高、破坏力强,精准提取其特征是实现有效预警的核心[5]。深度学习技术能够从海量气象数据中自动识别极端天气的复杂特征模式,为预警决策提供关键依据。

针对台风预警,可利用卫星云图数据,借助CNN提取其螺旋云带结构、眼区形态等关键空间信息。同时,结合台风移动路径、强度变化等时序数据,运用LSTM模型学习其时间演变规律。融合空间与时间特征,可全面刻画台风特性,服务于台风强度与路径预报。

在暴雨预警中,降水、湿度、气压等数据是重要输入。CNN用于识别降水的空间分布特征(如强降水中心位置、影响范围)；LSTM则分析湿度、气压等要素的时序演变特征,探索其与暴雨发生的关联机制。此外,自编码器可对高维气象数据进行降维处理,提取其中的关键隐含特征,提升特征提取的效率和准确性。

3.2.2极端天气事件识别与预警模型构建

在特征提取的基础上,构建高效的识别与预警模型是实现极端天气预警的关键环节。深度学习的分类与回归模型在此领域应用广泛。

对于极端天气事件的识别,常采用深度学习分类模型。将提

取的特征输入全连接神经网络或支持向量机等分类器,判断是否发生极端天气。例如,在暴雨识别中,模型基于输入的降水特征、相关气象要素特征等,输出暴雨发生的概率,依据预设阈值决定是否发布预警信息。

在预警模型构建中,回归模型常用于预测极端天气的强度、持续时间等关键参数。以台风强度预警为例,将台风特征参数输入LSTM或GRU等回归模型,预测未来的强度等级。同时,结合空间特征模型预测台风的潜在影响范围,为防灾减灾决策提供依据。

此外,可构建端到端的深度学习模型,直接从原始气象数据中学习特征并完成预警任务。此类模型规避了人工设计特征的局限性,能自动发掘数据中潜在的模式与规律,提高预警的准确性和效率。

3.3算法优化与改进

为提升深度学习模型在气象数据上的泛化能力并应对海量数据计算挑战,需采取模型正则化与计算效率优化策略。正则化方面,L1正则化通过添加权重L1范数惩罚实现特征选择和降维,尤其适用于高维气象数据; L2正则化利用权重L2范数约束抑制过拟合; Dropout技术则通过随机屏蔽神经元防止协同适应,在样本有限时显著增强泛化性。计算效率优化则通过精简网络结构(如轻量化CNN/LSTM)降低复杂度,并广泛应用数据并行或模型并行技术,依托GPU/分布式集群加速海量气象数据处理与训练过程,满足实时性业务需求。

4 结论与展望

本研究构建了高质量气象数据集,通过融合多源数据并应用深度学习驱动的清洗、预处理及增强方法,有效提升了数据的完整性、准确性和一致性。在算法层面,循环神经网络(RNN)、卷积神经网络(CNN)及时空融合模型显著优化了气象要素预测的精度与可靠性;同时,基于深度学习的特征提取与预警模型强化了极端天气事件的识别与预警能力。通过模型正则化、计算效率优化和并行计算策略,增强了模型的泛化能力和实际应用价值。

未来研究可从以下方向深化：一是扩展雷达、闪电等多元数据来源,开发更先进的数据融合与增强技术以优化数据集质量;二是加强深度学习模型与气象物理机理的融合,提升模型可解释性和物理一致性,并优化时空融合及极端天气预警模型的时效性;三是探索联邦学习等隐私保护技术,促进数据安全共享与协作;四是推动深度学习在气象业务系统的实际应用,开发高效实用的预测预警系统以提升服务效能。该领域的持续创新有望为气象科研与业务带来突破性进展。

[参考文献]

[1]成日晟,梁佳.内蒙古地区低空经济浪潮下的气象保障之思考[J].内蒙古科技与经济,2025,(10):98-103.

[2]马卓伟,崔琦.强化科技人才支撑促进粮储事业发展——2025年全国粮食和物资储备科技活动周侧记[J].中国粮食经济,2025,(06):28-31.

[3]贺荣.气象技术在水力发电行业的应用综述[J].水电与新能源,2025,39(07):32-35.

[4]余圣琪.公共大模型决策的法治化约束[J].国家检察官学院学报,2025,33(01):160-176.

[5]陈敏,秦小林,李绍涵,等.深度学习应用于强对流天气预测研究综述[J/OL].计算机应用,1-14[2025-07-29].

作者简介：

张平(1976--),女,汉族,山东临沂人,硕士,职称：工程师,研究方向：气象数据与计算机应用。