

结合数据清洗及并联时空神经网络的风电机组功率预测

高革命¹, 叶漫红¹, 刘亚楠¹, 徐鸿琪², 陈凡²

(1. 中国电建集团江西省电力设计院有限公司, 江西 南昌 330096;

2. 南京工程学院 电力工程学院, 江苏 南京 211167)

摘要:针对数据驱动的风电机组超短期功率预测中堆积型异常值难以识别、原始数据特征提取不够充分的问题,提出了一种结合数据清洗及并联时空神经网络的预测方法。首先,提出一种组合数据清洗方法对机组功率数据进行清洗;接着,提出一种并联时空神经网络,分别提取目标机组功率及气象数据的时间特征、相似机组功率数据的空间特征进行融合预测。此外,定义了误差区间准确度指标用于反映不同误差区间下预测结果的准确性,避免了传统误差指标掩盖局部预测误差过大的缺陷。算例分析结果表明,所提方法能够有效识别异常数据,提高了风电机组超短期功率的预测精度。

关键词:风电机组功率;深度学习;数据清洗;特征融合;评价指标

中图分类号:TM28 **文献标识码:**A **DOI:**10.19457/j.1001-2095.dqcd25954

A Wind Turbine Power Prediction Method Combining Data Cleaning and Parallel Spatio-temporal Neural Network

GAO Geming¹, YE Manhong¹, LIU Yanan¹, XU Hongqi², CHEN Fan²

(1. China Electric Power Construction Group Jiangxi Electric Power Design Institute Co., Ltd., Nanchang 330096, Jiangxi, China; 2. School of Electric Power Engineering, Nanjing Institute of Technology, Nanjing 211167, Jiangsu, China)

Abstract: Aiming at the problems of difficult identification of stacked outliers and insufficient extraction of raw data features in the data-driven ultra-short-term power prediction of wind turbines, a prediction method combining data cleaning and parallel spatio-temporal neural network was proposed. First, a combined data cleaning method was proposed to clean the wind turbine power data; then, a parallel spatio-temporal neural network was proposed to extract the temporal features of the power and meteorological data of the target wind turbine, and the spatial features of the power data of similar wind turbines for the fusion prediction, respectively. In addition, a prediction interval accuracy indicator was defined to reflect the accuracy of prediction results under different error intervals, avoiding the drawback of traditional error indicators that obscure large local prediction errors. The analysis results indicate that the proposed method can effectively identify anomalous data and improve the ultra-short-term power prediction accuracy of wind turbines.

Key words: wind turbine power; deep learning; data cleaning; feature fusion; evaluation index

随着风力发电总量在能源系统中比重提高,风电出力的随机性严重威胁了电力系统的安全运行^[1]。实现风电机组超短期功率的准确预测是保障电网安全经济运行的基础。

风电机组功率预测方法按照使用模型可分为物理方法、统计方法、机器学习方法及深度学

习方法^[2]。其中,物理方法计算时间长且依赖高精度的气象数据^[3],统计方法对数据平稳性要求较高^[4-5],机器学习方法处理风电功率的时间相关性能力有限^[6],因此这些方法在风电超短期功率预测中的表现较差。以长短期记忆网络(long short-term memory, LSTM)和卷积神经网络(con-

作者简介:高革命(1967—),男,硕士,正高级工程师,主要研究方向为新能源的勘测设计及研究,Email:519782612@qq.com

通讯作者:徐鸿琪(1999—),男,硕士,主要研究方向为新能源电力系统运行优化,Email:2543879398@qq.com

volutional neural network, CNN)为代表的深度学习模型,能够保留风电序列的长期依赖关系以及空间数据联系,是当前风能预测方法的热点^[7-8]。进一步地,为了避免单一网络数据提取能力的不足,多种组合模型被应用于风电功率预测领域。文献[9]将风电功率数据分解,采用CNN对具有短期相关性的高频子层进行预测,通过LSTM捕捉长期相关性,相较于单一模型获得了更高的预测精度。文献[10]利用自注意力时间卷积网络提取气象数据中的时间特征及变量之间的相关特征,通过LSTM模型建立特征与功率输出之间的复杂联系,实验结果证明了组合模型在提取数据特征及提高风电功率预测精度方面的有效性。文献[11]将多个机组、多维气象特征构造三维矩阵,通过CNN对每个时间断面提取相邻机组的空间信息再进行时序建模。文献[9-11]的研究表明考虑时空特征的组合预测能够更准确地预测风电功率,然而,现有研究通常将网络模型顺序连接,无法充分发挥不同模型对多样化数据的处理性能。

另一方面,数据驱动的模式预测精度依赖于原始数据质量,合理的数据清洗可以提高预测准确性^[12]。文献[13]采用 3σ 准则检测功率数据中的异常值,这种方法假设数据服从正态分布,这与风电功率数据通常呈双峰或多峰分布不相符。此外,这类方法仅从统计规律出发,忽略了功率与风速的耦合关系,因而限制了其对风电功率数据的清洗效果。为进一步提高检测精度,一些文献基于风速-风电机组功率数据的空间分布特征进行风电功率异常数据的识别:文献[14]和文献[15]分别采用聚类算法和离群检测算法将空间上距离正常数据簇较大的数据视作异常数据;文献[16]则基于异常数据易于从正常数据中分离的分布规律,采用孤立森林算法识别多维数据中的异常数据。文献[14-16]提出的方法能够准确判断空间上的离散型数据的异常情况,但是对于密集分布的异常数据却无法识别。为解决这一问题,文献[17-19]基于风速-风电机组功率特性将超出风速对应理论功率阈值的风电功率数据视作异常值,能够对密集分布的异常数据进行识别。其中,文献[17-18]建立了风速和风电机组功率的联合概率分布,根据风电功率散点的后验概率识别风电机组数据中的异常数据;文献[19]对风速-风电机组功率曲线进行了精准刻画,进一步明确了不同区间风速对应的理论功率阈值,使得异常功

率数据的识别更加精确。然而,这类方法依赖于原始数据质量,若原始数据中异常数据占比过高,则会影响建模的准确性。由此可见,现有的基于空间分布特征和基于风电转换模型的风电功率异常数据检测方法均存在一定的局限性。

综上所述,数据驱动的风电机组超短期功率预测方法存在原始数据特征提取不足和风电机组功率异常数据难以识别的问题。为此,本文提出了一种结合数据清洗及并联时空神经网络的风电机组功率预测方法。首先,采用带噪声的密度聚类方法(density-based spatial clustering of applications with noise, DBSCAN)识别分散型异常数据,在此基础上引入风电转换模型清洗堆积型数据,克服了单一的基于空间分布特征或基于风电转换模型的数据清洗方法的不足。接着,对清洗后的数据进行相关性分析,设计了一种融合时序和空间特征的数据集重构方法,构建了一个基于特征融合的并联时空神经网络(parallel spatio-temporal neural network, PSTN)模型。该模型综合考虑了风机功率数据的时序自相关性、与运行特征的互相关性以及相似风机功率数据之间的相关性等特征,实现了计及时空相关性的超短期功率预测。

1 清洗机组功率数据

在风电生产中,由于设备故障或人为限电,采集的原始数据往往包含多种异常数据,无法直接用于预测。实际上,由于风力涡轮机存在推动叶片旋转的最小启动风速以及为确保风机长期运行而采取的保护措施,风力涡轮机的输出特性可用分段函数表示,具体表达式为^[20]

$$P(v) = \begin{cases} 0 & v < v_{in} \\ 0.5\rho AC_p v^3 & v_{in} \leq v < v_n \\ P_n & v_n \leq v < v_{out} \\ 0 & v \geq v_{out} \end{cases} \quad (1)$$

式中: $P(v)$ 为风机的输出功率; P_n 为机组额定功率; v 为风速; v_{in} 、 v_n 、 v_{out} 分别为切入风速、额定风速和切出风速; ρ 为空气密度; A 为叶片扫过的面积; C_p 为风机的风能利用系数。

然而,原始的风电机组功率样本数据中部分数据明显偏离了风机输出特性曲线。这些异常数据根据其分布的位置和密度的不同,可以分为堆积型和分散型两大类,其中堆积型异常功率数据通常由长期因素引起,主要呈水平密集分布;

分散型异常功率数据通常受偶发因素影响,分散且稀疏地分布在散点图的整个区域中。针对两类异常值的检测问题,现有基于数据空间分布的方法能够较准确识别分散型异常数据,但可能将分布较密集的堆积型异常数据误判为正常数据。因此,本文提出一种结合聚类算法和风电转换模型的数据清洗方法,采用DBCSAN识别分散型异常数据,引入风电转换模型对堆积型异常数据进行清洗。同时,为避免传统插值方法无法有效填补大量缺失值的缺陷,对清洗后的风速及功率数据重新建立风电转换模型,对缺失功率数据进行修复。对异常数据检测与修复流程如图1所示。

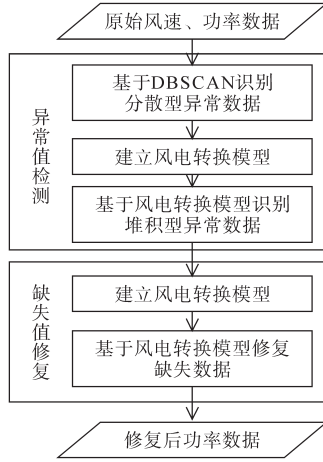


图1 结合聚类算法及风电转换模型的数据清洗

Fig.1 Data cleaning combining clustering algorithm and wind power conversion model

2 重构模型输入数据集

基于深度学习的功率预测模型能有效处理大规模的气象和机组运行数据,挖掘数据间的复杂相关性,显著提高预测精度。然而,现有研究通常只关注目标站点的运行数据,忽略了相似站点运行数据在空间上的一致性,限制了预测精度。因此,本文对原始数据进行相关性分析,重新构建时间和空间特征的输入数据集,构建方法如图2所示:利用待预测机组及相关机组的历史功率数据,采用对称排列方式构建了空间数据集;利用待预测机组的历史功率及运行状态数据构建了时间数据集。

对于时间数据部分,首先需确定最佳时间序列长度的出力信息作为时间特征提取模块的输入。现有研究通常人为设定时间窗口大小,较为主观。本文通过自相关函数 (autocorrelation function, ACF) 分析功率序列与其自身滞后 t 个时间

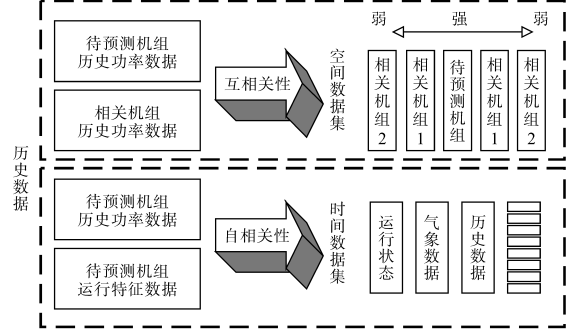


图2 基于相关性分析的数据集重构

Fig.2 Dataset reconstruction based on correlation analysis

步的互相关性,以此确定时间窗口大小。自相关函数公式为

$$ACF(t) = \frac{\sum_{i=0}^{L-t} (p_i - \bar{p})(p_{i+t} - \bar{p})}{\sum_{i=1}^L (p_i - \bar{p})^2} \quad t = 0, 1, \dots, L \quad (2)$$

式中: $ACF(t)$ 为自相关函数, 只与时间间隔 t 有关; $\bar{p}$ 为功率向量 $p = [p_1, p_2, \dots, p_L]^T$ 的均值; L 为原始功率向量长度, 包含一年时间跨度每 10 min 分辨率的采样值。

接着, 通过斯皮尔曼秩相关系数确定功率序列与气象因素及运行状态特征之间的非线性联系大小, 选择合适特征。根据以上信息重构时间数据矩阵为

$$TD = [f_1, f_2, \dots, f_m, p] \quad (3)$$

式中: TD 为时间数据矩阵; $[f_1, f_2, \dots, f_i, \dots, f_m]$ 为机组运行状态 (如机舱温度、机舱偏航角等)、气象数据 (如风速、风向、环境温度等) 的特征向量, 其中 $f_i = [f_{i1}, f_{i2}, \dots, f_{in}]^T, i = 1, 2, \dots, m, m$ 为特征总数; $p = [p_1, p_2, \dots, p_L]^T$ 为功率向量。

空间数据部分, 本文采用互信息作为关联性度量辨识不同风电机组功率序列之间的联系程度。为了更有效地捕捉到空间数据特征, 本文采用一种对称构图方式构建空间数据矩阵。具体来说, 对于目标风电机组 T_0 , 根据互信息指标由大到小的顺序, 选择相关机组为: $T_1, T_2, \dots, T_j$, 则所构空间数据矩阵为

$$SD = [p_{T_j}, \dots, p_{T_1}, p_{T_0}, p_{T_1}, \dots, p_{T_j}] \quad (4)$$

式中: SD 为空间数据矩阵, 包含所选机组的功率向量; $p_{Ti} = [p_{Ti1}, p_{Ti2}, \dots, p_{Tin}]^T$ 为各机组功率向量, $i = 1, 2, \dots, J; J$ 为机组总数; T_i 为机组序号。

通过 CNN 处理构建的空间数据集时, 由于相关性较高的机组位于中间, 卷积过程中它们的特征会得到更多的关注和强化; 边缘较低相关性数

据卷积次数较少,可以减少其对最终特征提取的影响。相比于直接对数据进行卷积,这种通过互信息分析和对称构造空间数据矩阵的做法,能够依据相关系数的大小更有效地利用机组之间的空间相关性。

3 并联时空神经网络

CNN在提取隐藏的空间特征方面具有出色的性能,LSTM能够保留历史信息,具有显著的时序特征提取能力。然而,传统组合模型采用网络模型顺序连接方式,构建成级联深度学习模型(cascaded deep learning, CDL),这种模型结构在一定程度上增强了模型的特征提取深度,但未能充分利用不同模型对多样化数据的处理能力。针对该问题,本文提出了一种基于特征融合的并联时空神经网络模型(PSTN),该模型能够分别提取不同风电机组功率数据间的空间特征与目标风电机组运行数据中的时间特征,通过将这些特征融合以获得风电机组功率预测结果,避免了模型级联导致后续特征依赖前一网络的缺陷,模型结构如图3所示。

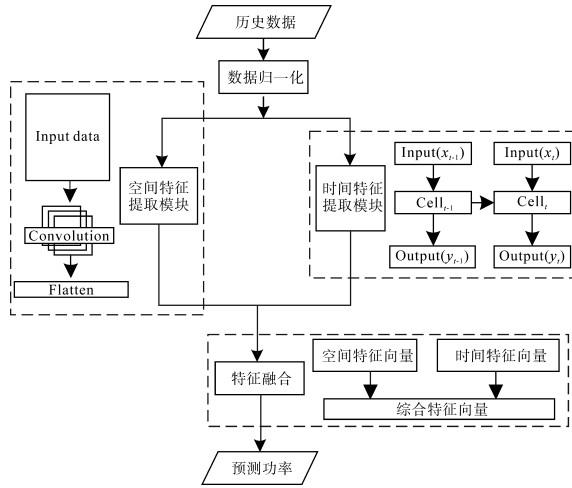


图3 并联时空神经网络

Fig.3 Parallel spatio-temporal neural network

PSTN模型由时间特征提取子模块、空间特征提取子模块以及特征融合模块3个模块组成。其中,时间特征提取模块采用LSTM处理时间序列数据,提取数据在时间维度上的动态模式;空间特征提取模块采用CNN处理相关风电机组的功率数据,捕捉空间维度上的相关性;特征融合模块将时间特征和空间特征向量连接,形成综合特征向量。融合后的特征向量将被馈送到一个全连接神经网络模块,特征权重将根据反向传播

算法自动调整以获得最终的预测结果。

4 算例分析

4.1 算例设置

4.1.1 算例数据

算例数据来自龙源电力集团有限公司提供的某风电场真实数据集^[21]。该数据集通过SCADA系统每10 min采样一次,包括功率数据 $Patv$ 、外部气象特征(如风速 $Wspd$ 、风向 $Wdir$ 、环境温度 $Etmp$)和内部运行状态特征(机舱温度 $Itmp$ 、机舱偏航角 $Ndir$ 、叶片的扭转角 Pab)。算例选择1号风机作为目标风机进行算例分析。

4.1.2 评价指标

本文使用归一化平均绝对误差(normalized mean absolute error, nMAE)、归一化均方根误差(normalized root mean square error, nRMSE)及决定系数(r-square, R^2)衡量预测结果准确性。3种指标的公式为

$$nMAE = \frac{1}{N} \sum_{i=1}^N \left| \frac{\hat{p}_i - p_i}{p_{Cap}} \right| \times 100\% \quad (5)$$

$$nRMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{\hat{p}_i - p_i}{p_{Cap}} \right)^2} \times 100\% \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{p}_i - p_i)^2}{\sum_{i=1}^N (p_i - \bar{p})^2} \quad (7)$$

式中: N 为样本个数; $\hat{p}_i$ 为预测功率值; p_i 为真实功率值; p_{Cap} 为风机容量; $\bar{p}$ 为真实功率向量的均值。

然而,这类指标仅给出了一段数据序列的统计平均结果^[22],无法反映模型在不同误差区间的预测准确度。为此,本文定义了预测误差区间准确度(prediction error interval accuracy, PEIA),用于衡量不同误差区间下准确性,计算式为

$$PEIA = \frac{1}{N} \sum_{i=1}^N C_{flag} \times 100\% \quad (8)$$

$$C_{flag} = \begin{cases} 1 & |\hat{p}_i - p_i| \leq er \\ 0 & |\hat{p}_i - p_i| > er \end{cases} \quad (9)$$

式中: C_{flag} 为预测结果标识,1表示预测结果在误差区间内,视为预测正确; er 为误差区间临界值。

PEIA值越大,则模型在该误差区间内预测准确度越高。在本文算例中,采用1%,5%,10%等3个误差区间来评估模型的预测准确性。

4.2 数据清洗结果分析

由于气象及运行状态各特征序列中错误数

据多由于偶发性元件故障造成,数量较少,因此单独进行处理,本文采用领域知识清洗不符合常理的数据,如风速小于0 m/s的数据,接着采用 3σ 准则清洗较离群的数据。清洗后的缺失数据使用多项式插值进行填补。

原始数据中功率异常值占比较高且难以识别,根据本文所提出的结合聚类算法和风电转换模型的数据清洗结果如图4所示。图4中,I区域的黑色样本为清洗后的正常风速-功率数据样本,其他区域的灰色样本则代表了被识别为异常的样本。III区域的样本功率数据因其在空间分布上的离群性被DBSCAN识别为分散型异常数据;II,IV区域内的样本功率数据则因其不符合正常分布的聚集状态被风电转换模型识别为堆积型异常数据。由图4中可见,清洗后的样本分布已经与理论上的功率分布曲线相近似。

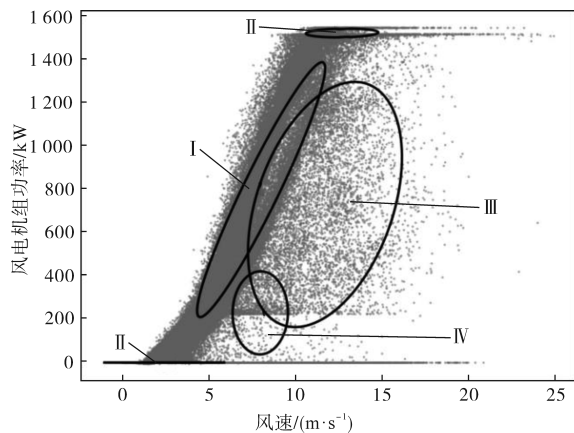


图4 风速-功率样本清洗结果

Fig.4 Results of wind speed-power sample cleaning

将本文所提的方法与常用异常值检测算法对比分析,图5为K均值(K-Means)、DBSCAN两种聚类方法及局部离群因子检测方法(local outlier factor, LOF)、孤立森林算法(isolation forest, IForest)两种异常值检测算法识别结果。其中,黑色样本为清洗后的正常风速-功率数据样本,灰色样本则代表被异常值检测算法识别为异常的样本。

由图5可以明显发现各类检测算法的不足:如图5a所示,K-Means聚类算法假设聚类为球形且需要预设聚类的数量。如果预设的参数不适合数据的实际分布,则识别效果较差;如图5b所示,尽管DBSCAN能够识别分散型功率异常数据,但是对于大多堆积型功率异常数据仍无法做到准确识别;如图5c所示,LOF算法在本研究中表现出了较多的误判情况,它仅能区分出一部分

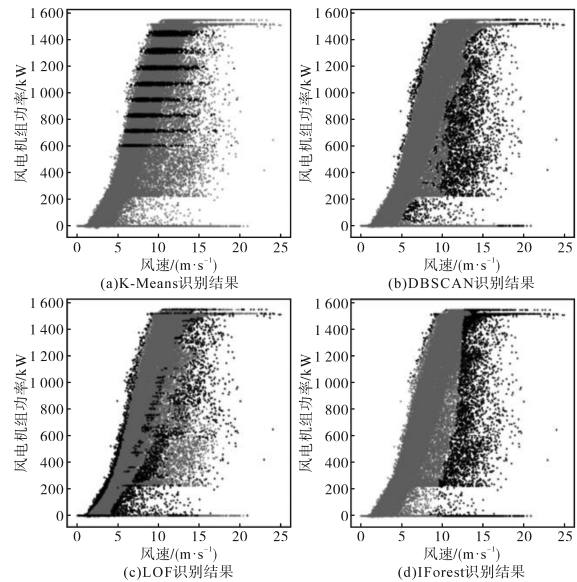


图5 常见异常值检测算法检测结果

Fig.5 Detection results of outlier detection algorithms

密度较低的分散型异常数据;同样,如图5d所示,IForest算法虽然能根据异常功率数据的分布特征来进行识别,但在本算例中,它未能有效区分密集分布的异常功率数据和正常数据。

图4及图5结果表明,尽管传统的异常值检测算法在特定场景下各有其优势,但面对复杂的风电数据特性时,它们的识别效果仍显不足。相比之下,本研究提出的方法有效结合了DBSCAN算法在识别分散型异常数据方面的准确性和风电转换模型在分离堆积型异常数据方面的优越性,克服了单一方法的不足,提高了对两类异常数据的识别能力。

在通过所提方法剔除异常功率数据后,原始数据集出现了大量功率缺失值,无法直接用于功率预测。为了确保修复过程不会对样本的整体分布造成干扰,通过对图4中I区域的正常风速-功率数据样本重新建立风电转换模型,基于该模型对缺失功率数据进行修复。修复结果如图6所示。

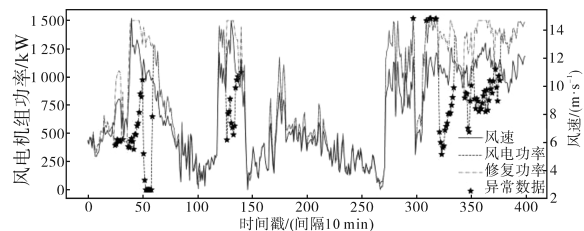


图6 功率序列修复结果

Fig.6 Results of power sequence repairing

图6展示了风电机组输出功率与风速之间的关系。曲线顶部附近的异常值和修复数据之间较为接近,这是因为原发电量数据超出了机组额

定功率,因此将其调整为额定功率值。由图6可见,修复后的风电功率数据的变化趋势与风速高度一致。即使面对连续的异常数据,该方法也能使修复后的风电功率数据保持与风速的对应关系。

4.3 相关性分析

通过对功率数据进行自相关性分析,能够识别出功率数据中潜在的周期性,作为选择时间窗口长度的重要依据。图7展示了功率时间序列在不同时间延迟值下的自相关性。图7中,横轴表示时间延迟,纵轴表示自相关系数的大小,阴影部分表示自相关系数的95%置信区间,如果自相关系数落在置信区间内,则可以认为在该延迟下自相关性不显著,即时间序列在这个时间延迟上的表现更接近随机波动,不受过去值的影响。

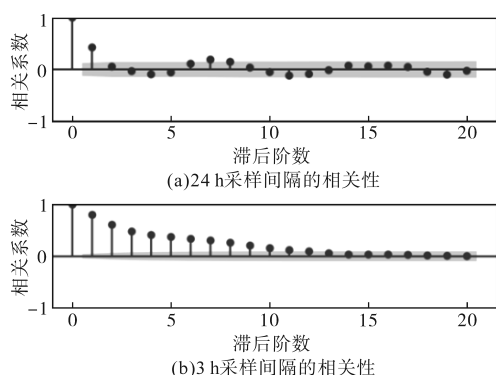


图7 功率序列自相关性分析

Fig.7 Autocorrelation analysis of power series

图7中横坐标表示不同时间间隔,其中图7a中1个时间间隔代表24h,在第2个时间间隔后相关系数落入置信区间,则认为风电机组出力仅与前一天24h出力数据高度相关。为分析风电出力的短期自相关性,以3h为时间间隔做进一步分析,如图7b所示。由图7可见,在滞后30h(即10个3h间隔)后,自相关系数逐渐落入95%置信区间内,表明此后的自相关性不显著。鉴于原始数据以每10min进行采样,研究选择180个时间步作为时间窗口长度,既能捕捉到足够的历史信息对未来出力进行预测,又能避免引入过多的数据冗余。

根据斯皮尔曼秩相关系数对风电机组功率与气象因素及运行状态特征之间的非线性联系进行分析,分析结果如表1所示。

表1 功率与特征的相关性

Tab.1 Correlation between power and features

状态变量	$Patv$	$Wspd$	$Wdir$	$Etmp$	$Itmp$	$Ndir$	Pab
相关系数	1	0.97	0.07	-0.24	-0.08	-0.82	0.42

根据表1中的分析结果,风电机组实际输出功率($Patv$)与各状态变量之间的相关性表现出显著差异。其中,功率与风速($Wspd$)之间的相关系数为0.97,显示出极强的正相关性;相比于风速,风向($Wdir$)与功率之间的相关系数仅为0.07,表明风向可能在模型中的贡献较小;环境温度($Etmp$)和内部温度($Itmp$)对功率输出呈现出轻微的负相关性,分别为-0.24和-0.08,这表明在一定范围内,温度升高对风电机组的输出功率产生了轻微的负面影响;尽管功率与机舱偏航角($Ndir$)之间呈现出显著的负相关,但两者之间的关系仍非常紧密;此外,叶片角度(Pab)与功率的相关系数为0.42,表明叶片角度调整对功率输出有一定影响。基于以上分析,选择的相关特征包括:风速、环境温度、机舱偏航角以及叶片角度。这些特征涵盖了对风电机组功率输出影响最为显著的因素,也包括了与发电效率相关的运行特征。最后,对不同机组之间的功率序列进行互相关性分析,根据互信息指标筛选相关风电机组为:机组45、机组25、机组69、机组23、机组47以及机组70。

4.4 风电功率预测结果分析

4.4.1 模型参数

本文对比了LSTM、级联深度学习(CDL)以及本文提出的并联时空神经网络(PSTN)3类神经网络模型。其中,LSTM包含两层LSTM网络层,每一层含有与输入序列长度相等的神经元个数。CDL由CNN与LSTM顺序连接,CNN部分采用一层二维卷积网络层,卷积核大小为(2,2),通道数为1,步长为(1,1);LSTM部分使用一层LSTM。PSTN由CNN与LSTM并联组成,卷积核大小改为(2,4),其余参数保持一致。

为了提高模型的泛化能力并避免过拟合,3种模型均采用了0.2的dropout率,并在层之间增加了批量归一化和LeakyReLU激活函数。模型的训练采用Adam优化算法,初始学习率设置为0.001,以平均绝对误差作为损失函数。

4.4.2 预测结果分析

为了验证本文提出的数据清洗方法及PSTN模型的先进性,本文对6种实验场景下的预测结果进行了比较分析,6种实验场景为:

实验1:基于LSTM模型,模型输入为包含历史出力、运行状态特征的时间数据集;

实验2:基于CDL模型,模型使用数据集与实验1相同;

实验3:基于 PSTN 模型,模型输入为时间数据集及相关机组出力顺序拼接的空间数据集;

实验4:基于 PSTN 模型,模型输入为时间数据集及相关机组出力按对称构图方式的空间数据集;

实验5:基于 PSTN 模型,采用实验4所用数据集清洗并使用样条插值修复后的数据集作为模型输入;

实验6:基于 PSTN 模型,采用实验4所用数据集清洗并使用风电转换模型修复后的数据集作为模型输入。

6种实验场景下的预测误差结果如表2所示。由表2可见,使用组合模型的实验2~实验6的传统误差结果均低于实验1,这证明了组合模型在处理复杂数据方面的优越性。与实验2相比,实验3和实验4的整体平均误差均有下降,这表明 PSTN 模型能够更充分利用时空数据特征,提高预测精度。同实验3相比,实验4考虑了特征排列的影响,对特征进行了对称构图,进一步提高了对空间特征的利用,从而获得更高的预测精度。相较于传统预测模型,本文提出的 PSTN 模型的统计误差指标 $nMAE$ 和 $nRMSE$ 分别为 7.78% 与 11.79%,相比 LSTM 模型降低了 40.29% 和 28.93%,相比 CDL 模型降低了 33.90% 和 18.86%。比较实验1~实验4的决定系数 R^2 ,可以看出 PSTN 模型获得了更好的拟合效果。

表2 预测结果误差比较

Tab.2 Error comparison of prediction results

实验场景	传统误差指标/%			预测误差区间准确度(PEIA)/%		
	$nMAE$	$nRMSE$	R^2	1% 误差	5% 误差	10% 误差
1	13.03	16.59	74.34	10.28	33.32	46.62
2	11.77	14.53	80.34	6.88	23.09	44.57
3	8.08	12.27	85.96	37.02	53.19	65.24
4	7.78	11.79	87.11	35.19	53.11	68.45
5	11.30	15.56	77.55	17.09	35.91	53.31
6	5.41	8.92	92.62	38.69	63.80	79.88

为直观比较各实验预测结果与实际风电出力之间的差距,图8展示了6种实验场景下的风电预测曲线和实际风电机组功率曲线。由图8可见,大部分实验场景下的预测曲线与实际功率曲线接近。然而,实验2的预测结果在[10,140]的时间戳区间显著偏离实际功率曲线,而在[140,200]的时间戳区间与实际功率曲线较为接近。从传统误差指标来看,实验2的整体预测精度较高,无

法反映局部的预测情况。

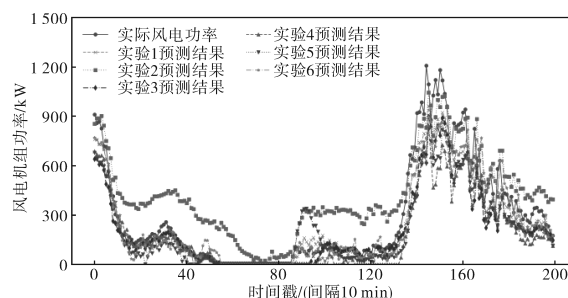


图8 实验预测结果曲线图

Fig.8 Curve chart of the experimental prediction results

传统误差指标仅提供模型预测误差的整体评估,根据本文所定义的预测误差区间准确度(PEIA)指标,能够清晰地观察各模型在不同误差区间的预测准确性。以实验2中的CDL模型为例,尽管在 $nMAE$ 和 $nRMSE$ 等传统指标上,其预测性能与其他模型相近,但在3个误差区间表现均不如其他模型,表现出较差的可靠性,而这一特性在传统误差指标中无法体现。

图9展示了本文各实验的相对误差分布箱型图,其中,纵坐标表示预测功率与真实功率之间的相对误差,横坐标则标识了不同实验。图9中箱体的高度展示了预测误差的集中趋势,须的长度则反映了整体数据分布的范围。

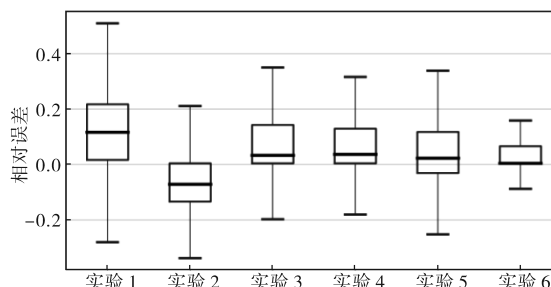


图9 相对误差分布箱型图

Fig.9 Boxplot of relative error distribution

结合图9及表2中PEIA指标可以看出,本文提出的并联时空神经网络模型在1%的误差区间内准确度为38.69%,这表明有38.69%的预测结果与真实值的差距非常小,体现了 PSTN 模型预计结果的高准确性。此外,该模型的中位数与0的距离最接近,意味着大部分预测值都紧密围绕真实值分布,进一步证明了模型的预测性能。

实验1~实验4并未对实验数据集进行数据清洗,因此预测精度有限。为了进一步提高模型的预测精度,实验5采用所提方法剔除了大量异常数据,但由于使用样条插值未能还原风电机组出力的连续性,导致预测精度大幅下降。实验6

在实验5的基础上根据风电转换模型对缺失功率数据进行了修复,使整体误差进一步降低,在图9中展现了最窄的误差区间。

4.5 本文方法的计算效率分析

为进一步说明本文方法在计算效率方面的优越性,本文对所提方法与其他方法在数据清洗和功率预测两个阶段的计算耗时进行了比较。

1)数据清洗阶段:使用算例4.2节中的风速—功率数据样本进行离线训练及在线清洗,不同数据清洗方法的耗时如表3所示。

表3 不同数据清洗方法的耗时

方法	离线训练阶段耗时/s	在线清洗阶段耗时/s
K-Means	1.165 7	0.027 4
LOF	0.945 9	0.911 3
IForest	2.792 8	2.444 2
DBSCAN	5.105 3	0.003 0
本文方法	5.114 3	0.006 0

根据表3给出的不同方法计算用时以及图4和图5给出的清洗结果可见:K-Means, LOF及IForest等传统数据清洗方法耗时相对较少,但是清洗效果难以达到预期;DBSCAN算法离线训练阶段计算用时相对较长,但是在线清洗阶段耗时较少、且清洗效果较好;本文方法在DBSCAN算法的基础上采用了风电转换模型,计算用时有轻微的增长,但是避免了DBSCAN算法无法准确识别堆积型异常数据的缺陷,因而有效提升了数据的清洗效果。

2)功率预测阶段:本文模型与实验1、实验2所用模型在模型训练及预测阶段的用时如表4所示。由表4可见,与其他模型相比,本文提出的PSTN模型训练阶段耗时和预测阶段耗时均较低,具有较高的计算效率。

表4 不同神经网络模型的计算用时

预测模型	训练阶段耗时 $/(s \cdot \text{epoch}^{-1})$	预测阶段耗时 $/(s \cdot \text{epoch}^{-1})$
LSTM	8.630 6	3.200 7
CDL	4.642 0	1.764 9
本文提出的PSTN	4.431 2	1.548 8

综上所述,本文提出的方法仅在数据清洗阶段离线训练时相对需要较多的计算资源;而在线应用时,不论是数据清洗阶段还是功率预测阶段都只需要较少计算耗时,故本文方法具有较高的计算效率。

5 结论

针对风电机组功率预测研究中存在异常值难以识别、原始数据特征提取不足导致精度受限的问题,本文提出了一种结合数据清洗及并联时空神经网络的预测方法。利用某风电场真实数据进行算例分析,得到以下结论:

1)本文提出的结合聚类算法及风电转换模型的数据清洗方法能够有效识别分散型与堆积型异常数据,通过风电机组功率曲线能够较准确修复缺失功率数据,保证数据可靠性。

2)所提出的并联时空神经网络能够有效利用数据时间及空间特征,与传统单一模型及级联组合模型相比,具有更好的特征提取效果和预测准确性。

3)本文提出的预测误差区间准确性指标可以观察模型在不同误差区间下的预测精度,便于从多个角度了解模型的预测性能。

本文采用的数据清洗算法参数基于主观经验选取,需要针对不同数据进行调整。作为一种基于群体智能的优化算法,粒子群算法在参数优化方面具有全局搜索能力强、收敛速度快等优点,未来研究中可以考虑结合该算法对参数进行优化,提高本文方法的适应性和鲁棒性。

参考文献

- [1] 王渝红,史云翔,周旭,等. 基于时间模式注意力机制的BiLSTM多风电机组超短期功率预测[J]. 高电压技术, 2022, 48(5): 1884-1892.
WANG Yuhong, SHI Yunxiang, ZHOU Xu, et al. Ultra-short-term power prediction for BiLSTM multi wind turbines based on temporal pattern attention[J]. High Voltage Engineering, 2022, 48(5): 1884-1892.
- [2] 张浩田,温蜜,李晋国,等. 数据驱动的时间注意力卷积风电功率预测模型[J]. 太阳能学报, 2022, 43(10): 167-176.
ZHANG Haotian, WEN Mi, LI Jinguo, et al. Data driven time attention convolutional wind power prediction model[J]. Acta Energiae Solaris Sinica, 2022, 43(10): 167-176.
- [3] DONG Y, MA S, ZHANG H, et al. Wind power prediction based on multi-class autoregressive moving average model with logistic function[J]. Journal of Modern Power Systems and Clean Energy, 2022, 10(5): 1184-93.
- [4] LI J, SONG Z, WANG X, et al. A novel offshore wind farm typhoon wind speed prediction model based on PSO-Bi-LSTM improved by VMD[J]. Energy, 2022, 251: 123848.
- [5] 曹俊波,周任军,邓学华,等. 考虑优化ARIMA模型差分次数的风功率预测[J]. 电力系统及其自动化学报, 2019, 31

- (1):105-111.
- CAO Junbo, ZHOU Renjun, DENG Xuehua, et al. Wind power forecast considering differential times of optimal ARIMA model[J]. Proceedings of the CSU-EPSA, 2019, 31(1): 105-111.
- [6] 李卓, 叶林, 戴斌华, 等. 基于IDSCNN-AM-LSTM组合神经网络超短期风电功率预测方法[J]. 高电压技术, 2022, 48(6): 2117-2127.
- LI Zhuo, YE Lin, DAI Binhua, et al. Ultra-short-term wind power prediction method based on IDSCNN-AM-LSTM combination neural network[J]. High Voltage Engineering, 2022, 48(6): 2117-2127.
- [7] 冉靖, 张智刚, 梁志峰, 等. 风电场风速和发电功率预测方法综述[J]. 数理统计与管理, 2020, 39(6): 1045-1059.
- RAN Jing, ZHANG Zhigang, LIANG Zhifeng, et al. Review of wind speed and wind power prediction methods[J]. Journal of Applied Statistics and Management, 2020, 39(6): 1045-1059.
- [8] 赵泽妮, 云斯宁, 贾凌云, 等. 基于统计模型的短期风能预测方法研究进展[J]. 太阳能学报, 2022, 43(11): 224-234.
- ZHAO Z N, YUN S N, JIA L Y, et al. Recent progress in short-term forecasting of wind energy based on statistical models[J]. Acta Energaie Solaris Sinica, 2022, 43(11): 224-234.
- [9] HANIFI S, ZARE-BEHTASH H, CAMMARANO A, et al. Off-shore wind power forecasting based on WPD and optimised deep learning methods[J]. Renewable Energy, 2023, 218: 119241.
- [10] XIANG L, LIU J N, YANG X, et al. Ultra-short term wind power prediction applying a novel model named SATCN-LSTM[J]. Energy Conversion and Management, 2022, 252: 115123.
- [11] 陈金富, 朱乔木, 石东源, 等. 利用时空相关性的多位置多步风速预测模型[J]. 中国电机工程学报, 2019, 39(7): 2093-2106.
- CHEN Jinfu, ZHU Qiaomu, SHI Dongyuan, et al. Multi-location and multi-step wind speed forecasting model using spatio-temporal correlation[J]. Proceedings of the CSEE, 2019, 39(7): 2093-2106.
- [12] WANG S, LI B, LI G, et al. Short-term wind power prediction based on multidimensional data cleaning and feature reconfiguration[J]. Applied Energy, 2021, 292: 116851.
- [13] OUYANG T, HUANG H, HE Y, et al. Chaotic wind power time series prediction via switching data-driven modes[J]. Renewable Energy, 2020, 145: 270-281.
- [14] 肖勇, 郑楷洪, 余忠忠, 等. 基于三次指数平滑模型与DBSCAN聚类的电量数据异常检测[J]. 电网技术, 2020, 44(3): 1099-1104.
- XIAO Yong, ZHENG Kaihong, YU Zhongzhong, et al. Anomaly detection of power data based on cubic exponential smoothing model and DBSCAN clustering[J]. Power System Technology, 2020, 44(3): 1099-1104.
- [15] 汪欣, 蔡旭, 李征. 结合交叉局部异常因子和注意力机制的超短期风电功率预测方法[J]. 电力系统保护与控制, 2020, 48(23): 92-99.
- WANG Xin, CAI Xu, LI Zheng. Ultra-short-term wind power prediction method combined with cross-local anomaly factors and attention mechanism[J]. Power System Protection and Control, 2020, 48(23): 92-99.
- [16] GUO H, WANG J, LI Z, et al. A multivariable hybrid prediction system of wind power based on outlier test and innovative multi-objective optimization[J]. Energy, 2022, 239: 122333.
- [17] YE X, LU Z, QIAO Y, et al. Identification and correction of outliers in wind farm time series power data[J]. IEEE Transactions on Power Systems, 2016, 31(6): 4197-205.
- [18] MA Y C, LIU Y Q, YANG Z L, et al. Research on the derated power data identification method of a wind turbine based on a multi-gaussian-discrete joint probability model[J]. Sensors, 2022, 22(22): 8891.
- [19] 娄建楼, 胥佳, 陆恒, 等. 基于功率曲线的风电机组数据清洗算法[J]. 电力系统自动化, 2016, 40(10): 116-21.
- LOU Jianlou, XU Jia, LU Heng, et al. Data cleaning algorithm for wind turbine based on power curve[J]. Automation of Electric Power Systems, 2016, 40(10): 116-121.
- [20] YANG M, SHI C, LIU H. Day-ahead wind power forecasting based on the clustering of equivalent power curves[J]. Energy, 2021, 218: 119515.
- [21] ZHOU J, LU X, XIAO Y, et al. SDWPF: A dataset for spatial dynamic wind power forecasting challenge at KDD Cup 2022 [EB/OL]. arXiv: 2208.04360[2024-05-21]. <https://arxiv.org/abs/2208.04360>.
- [22] 姚琦, 柳玉, 白恺, 等. 风电场功率预测水平的多指标综合评价方法研究[J]. 太阳能学报, 2019, 40(2): 333-340.
- YAO Qi, LIU Yu, BAI Kai, et al. Study of multi-index comprehensive evaluation method for wind farm power prediction level[J]. Acta Energaie Solaris Sinica, 2019, 40(2): 333-340.

收稿日期:2024-05-21

修改稿日期:2024-10-13