DOI: 10.19666/j.rlfd.202210213

基于深度长短记忆网络的汽轮机数据清洗

许小刚[1,2,3]，王志香[1]，王惠杰[1,2,3]
（1.华北电力大学动力工程系，河北 保定 071003；
2.华北电力大学河北省低碳高效发电技术重点实验室，河北 保定 071003；
3.华北电力大学保定市低碳高效发电技术重点实验室，河北 保定 071003）

［摘 要］汽轮机运行过程会产生多样且大量数据。为适应大数据驱动及仿真建模对高质量数据的要求，高效的数据清洗十分必要。利用长短记忆层对于时序数据出色的非线性拟合能力搭建了汽轮机半监督数据清洗模型。模型选取机组的 3 个边界条件作为输入，对待清洗数据进行预测，根据预测值与实际值的残差完成异常值剔除，之后选用模型的预测值进行数据填充，保证数据的完整性。利用模型对某电厂 650 MW 机组进行数据清洗，并且为克服样本失衡给清洗模型指标选取带来的问题，对准确率进行了改进并将其作为清洗效果的衡量指标。结果表明：深度长短记忆网络的数据清洗模型改进准确率高于其他 3 种常见清洗方法，可有效识别数据是否异常，且可利用预测值进行数据填充，保证清洗前后数据量一致。

［关 键 词］长短记忆网络；深度学习；数据清洗；异常值；汽轮机

［引用本文格式］许小刚，王志香，王惠杰. 基于深度长短记忆网络的汽轮机数据清洗[J]. 热力发电，2023，52(8)：179-187.
XU Xiaogang, WANG Zhixiang, WANG Huijie. Turbine data cleaning based on deep LSTM[J]. Thermal Power Generation, 2023, 52(8): 179-187.

Turbine data cleaning based on deep LSTM

XU Xiaogang[1,2,3], WANG Zhixiang[1], WANG Huijie[1,2,3]
(1. Department of Power Engineering, North China Electric Power University, Baoding 071003, China;
2. Key Laboratory of Low Carbon and Efficient Power Generation Technology of Hebei, North China Electric Power University, Baoding 071003, China;
3. Baoding Key Laboratory of Low Carbon and Efficient Power Generation Technology, North China Electric Power University, Baoding 071003, China)

Abstract: A large amount of data is generated during steam turbine operation. In order to meet the requirements of high quality data driven by big data and simulation modeling, efficient data cleaning is very necessary. The semi-supervised data cleaning model of steam turbine is built by using the excellent nonlinear fitting ability of long and short memory layer for time series data. The model selects three boundary conditions of the unit as input to predict the cleaning data. Outliers are eliminated according to the residual difference between the predicted value and the actual value. Then, the predicted value of the model is used to fill the data to ensure the integrity of the data. The model is used to clean the data of a 650 MW unit in a power plant. To overcome the problems caused by sample imbalance in the selection of cleaning model indicators, the accuracy rate is improved and taken as the measurement index of cleaning effect. The results show that, the improved accuracy of the data cleaning model of the deep long and short memory network is higher than that of the other three common cleaning methods, which can effectively identify whether the data is abnormal, and can use the predicted value to fill the data to ensure the consistency of data before and after cleaning.

Key words: long and short memory networks; deep learning; data cleaning; outliers; steam turbine

汽轮机作为锅炉与发电机的中间设备，将锅炉产生的蒸汽转换为机械能从而带动发电机输出电能。汽轮机运行是一个不间断进行的过程，每一时刻的运行都与之前息息相关，期间产生的数据量不

修 回 日 期：2022-10-18 网络首发日期：2022-11-01
基 金 项 目：中央高校基本科研业务费专项资金资助（2019MS094）
Supported by：Fundamental Research Funds for the Central Universities (2019MS094)
第一作者简介：许小刚（1979），男，博士，高级工程师，主要研究方向为故障诊断、系统寻优，xxg@ncepu.edu.cn。

会发生突变。常见的汽轮机“脏数据”类型主要包含数据缺失以及不符合运行数据时序性和其他原因造成的数据与运行不符所带来的数据异常。

1）缺失数据 在火电厂监测数据中，缺失值是指某条记录的属性字段值被标记为 NaN 的数据，出现缺失值的原因主要是数据采集传感器短时异常或数据传输链受阻等因素导致数据未被写入[1]。

2）异常数据 随着火电厂检测水平的日益提高和大数据分析对于数据量需求巨大，传感器测点遍布机组的每一个部分，不停地将机组运行过程中的各项数据进行传输、保存，而这些数据的准确性会受到设备老化或故障，测量精度不足、信息传输故障、信号干扰等一系列问题的影响，最终保存进 DCS 的数据会被污染，产生异常数据[2]。

在发电趋于自动化、智能化的大背景下，随着大数据驱动，深度学习在故障诊断、工况划分、仿真模拟等方面的应用，对于所使用的数据质量要求更为严格。完成“脏数据”的清洗，可提高数据的质量，从而提升以数据为前提的诊断、模拟、预测等各项工作的准确性，对于“智慧电厂”的实现具有深远意义。

对原始数据中“脏数据”的识别与修复是数据质量分析中的一项主要研究工作。目前常见数据清洗方法包括基于统计的 3σ 准则、箱型图以及基于机器学习的聚类法、局部异常因子、孤立森林和深度学习法。杨茂等针对 3σ 准则的模型参数 μ 和 σ 提出了改进建立类 3σ 准则，对于光伏功率的异常数据进行识别。但由于该方法是针对正态分布的数据，其识别的准确率并不高，且对于其他分布类型的数据适用性较差[3]。何高清等利用箱型图对于轴承内径尺寸进行了异常数据的识别，相较于 3σ 准则而言箱型图的普适性更强，适用于大多的数据分布，但其仅仅只能识别简单的离群点[4]。许璟琳等利用 k 近邻距离对于医院的用电能耗离群点进行了检测，但针对文中提到的 3 种异常值，仅可以检测其中的局部异常[5]。陈洪涛等提出了一种基于 k-means 聚类算法的线损异常辨别方法，根据分析线损的大小决定对数据进行几次聚类，正确率相对较高，但在其异常数据检测过程中线损大小、聚类类别、阈值等多处需要人为分析确定，检测结果受人为影响较大[6]。贺玉海等采用 k-Medoids 算法与具有噪声基于密度聚类（DBSCAN）算法的组合聚类算法对于交通流数据进行清洗工作，但其对于聚类中心距离大以及高维度数据聚类效果较差[7]。石玉亮等通过增大可达距离，降低局部可达密度，提高了异常帧与正常帧之间的区分度对局部异常因子进行了改进，运用改进的局部异常因子法对多维数据的异常值进行识别检测[8]。Wang B 等通过对大数据集进行聚类选取新的数据集，之后选用局部离群因子（LOF）算法对异常值进行识别，K-LOF 降低了大数据量的计算复杂度，但随之而来的缺点就是数据的误检以及数据量的缩减[9]。侯振英通过将孤立森林和局部异常因子 2 个算法的结果映射到同一空间，最终确定异常值[10]。

聚类、局部异常因子以及孤立森林都是将数据的异常值简单定义为离群点，而对于其他与正常值差异不大的异常数据其检测性能会下降。吴磊等研究表明，长短记忆网络具有良好的预测性能，在电力预测等方面应用较多，且预测性能较好[11]。吴飘利用深度卷积的超强特征学习能力以及残差损失对 2 个公开数据集合进行了异常数据检测，结果表明基于深度学习的检测结果优于传统的 PCA 和 KNN[12]。随着深度学习在各行各业的深入发展，越来越多的学者将数据清洗的研究转到深度学习。就汽轮机而言，多数学者将研究的重点放在故障检测分析上，而对于数据清洗工作大多还停留在阈值分割、聚类等较为传统的方法上，忽略了高质量数据对于数据挖掘工作的重要性。

本文首先利用深度长短记忆网络（deep long short-term memory，DLSTM）搭建半监督数据清洗模型，对汽轮机运行数据进行清洗工作。长短记忆（long short-term memory，LSTM）可以很好地学习到汽轮机运行数据的时序性；而半监督清洗模型利用各项输入进行待清洗数据的预测，利用预测值与真实值之间的残差作为异常值检测的阈值，并且之后可以利用预测值进行异常值剔除之后的数据填充，保证数据的完整性。之后，利用搭建好的 DLSTM 模型准确高效地完成了对某电厂 650 MW 机组汽轮机运行数据清洗。

1 基于 DLSTM 的数据清洗模型

1.1 LSTM 算法

LSTM 算法是基于循环神经网络（recurrent neural network，RNN）的一种变体智能算法，其在 RNN 的基础上增加了 3 个门结构，从左至右依次是遗忘门、输入门、输出门。相比普通的 RNN（图 1），

LSTM（图 2）由于能够学习长期的依赖关系，在长时间序列数据的处理上表现优异[13-15]。

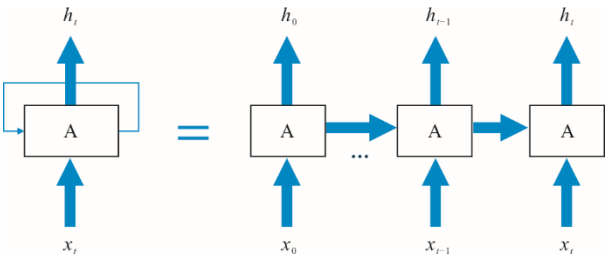


图 1 RNN 展开结构
Fig.1 RNN unfolding structure

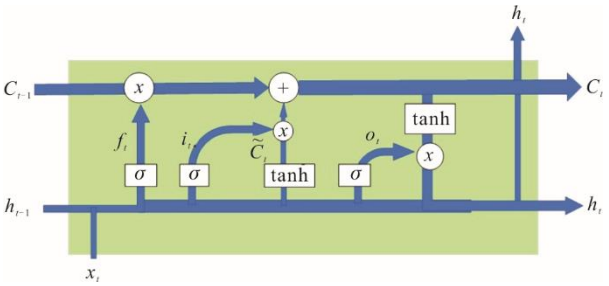


图 2 LSTM
Fig.2 LSTM

1）遗忘门　决定 C_{t-1} 信息的取舍。其输入为前一个细胞的输出 h_{t-1} 以及当前的输入 x_t，该门的输出 f_t 为一个 0~1 的数字，表示对于 C_{t-1} 信息的保留与取舍：

$$f_t=\sigma(W_{\mathrm{f}}[h_{t-1},x_t]+b_{\mathrm{f}}) \tag{1}$$

其中：W_{f} 为遗忘门的激活函数；b_{f} 为偏置。

2）输入门　通过候选记忆细胞 C_t，决定输入有多少进入当前的 C_t 中，可以有效减少无关信息的输入：

$$i_t=\sigma(W_{\mathrm{i}}[h_{t-1},x_t]+b_{\mathrm{i}}) \tag{2}$$

其中：W_{i} 为输入的激活函数；b_{i} 为偏置。

3）输出门　决定当前的输出 o_t：

$$\mathrm{o}_t=\sigma(W_{\mathrm{o}}[h_{t-1},x_t]+b_{\mathrm{o}}) \tag{3}$$

长短记忆的记忆细胞更新公式为：

$$C_t=\tanh(W_{\mathrm{c}}[h_{t-1},x_t]+b_{\mathrm{c}}) \tag{4}$$

$$C_t=C_{t-1}\otimes f_t+i_t\otimes C_t \tag{5}$$

其中：W_{o} 为输出门的激活函数；b_{o} 为偏置；W_{c} 为记忆细胞更新时的激活函数；b_{c} 为相应偏置；C_t 为候选记忆细胞。

隐藏状态 h_t 结合输出门输出以及当前记忆细胞输出：

$$h_t=o_t*\tanh(C_t) \tag{6}$$

LSTM 将 RNN 中的“全乘”变为了“乘加结合”可在一定程度上解决梯度爆炸[16]。

1.2 DLSTM 模型搭建

深度学习（deep learning）是神经网络的一部分，可以根据多层神经网络逐步进行多样化深层次特征提取以及进行复杂的非线性拟合。

根据汽轮机运行数据所具有的时序特性以及皮尔逊系数选取了 3 个边界条件作为输入量，模型构建了 3 层 LSTM 层，可以充分拟合输入与输出以及数据前后时序之间的关系。同时，为了避免过拟合的风险，增强模型的泛化能力，在每层 LSTM 之后加入了正则化层。对于神经网络单元，按照一定的概率将其暂时从网络中丢弃，可以简化网络，提高模型的训练时效性[17]。最后，加入 2 个全连接层将之前提取学习到的特征进行关联，并且将结果映射到输出。利用训练集训练深度长短记忆模型，选取损失函数为平均绝对误差 δ_{MAE}（式(7)）。δ_{MAE} 表示预测值与实际值差值的平均，直观反映模型预测能力高低，同样也为之后设置判别异常值所需的阈值提供了便利条件。根据训练结束所产生的动态差值设置阈值：据训练数据集训练结束的预测值与实际值之间差值的绝对值关系设定合适的判别异常阈值 s。模型的优化器使用 Adam。该优化器简单高效并且内存占用少，可以自然地实现步长退火过程，即自适应地调整学习率[18]。

$$\delta_{\mathrm{MAE}}=\frac{1}{m}\sum_{i=1}^{m}|a-e| \tag{7}$$

其中：a 为实际值；e 为异常值。

DLSTM 模型内部架构见表 1。

表 1 DLSTM 模型架构
Tab.1 The DLSTM model architecture

Layer (type)	Output Shape	Param
lstm_1 (LSTM)	(None, 200, 20)	2 080
dropout_1 (Dropout)	(None, 200, 20)	0
lstm_2 (LSTM)	(None, 200, 20)	3 280
dropout_2 (Dropout)	(None, 200, 20)	0
lstm_3 (LSTM)	(None, 20)	3 280
dropout_3 (Dropout)	(None, 20)	0
dense_1 (Dense)	(None, 2)	42
dense_2(Dense)	(None, 1)	3
Total params: 8 685 Trainable params: 8 685 Non-trainable params: 0		

1.3 多维时序 DLSTM 数据清洗

LSTM 可以根据输入的多维度数据进行相关数据预测，充分挖掘利用数据之间复杂的非线性关系，以期根据时序数据的特点以及输入量之间的关系很好地预测输出量[19]。本文所构建的 DLSTM 数据清洗模型的清洗原理是利用机组边界条件作为多维时序数据输入，对汽轮机待清洗数据进行预测训练；之后根据训练好的模型对测试集数据进行预测，依据预测值与实际值之间的残差关系判别数据是否异常。模型清洗结构如图 3 所示。

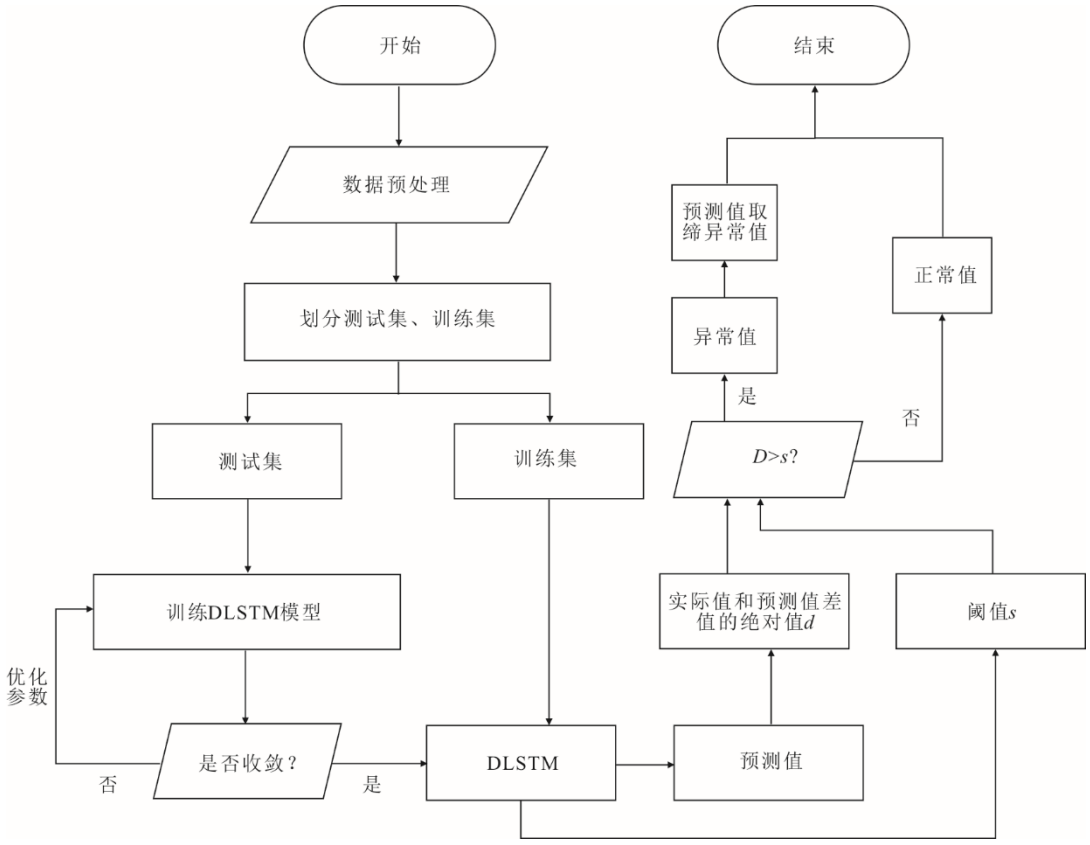


图 3 DLSTM 数据清洗模型流程
Fig.3 Flowchart of the DLSTM data cleaning model

2 模型验证

2.1 汽轮机数据以及输入量的选取

本文采用国能黄金埠发电有限公司超临界 650 MW 机组 2020 年 1 月 17 日—2 月 28 日的历史运行数据作为数据集。海量数据会大大降低数据清洗的时效性，且会带入大量的随机干扰信息。但汽轮机组的运行并不是一个快变工况，机组会在短时间内维持一定的运行状态，选取 432 s 所得的数据既可以维持机组运行数据的规律性，采集到每个时间段的数据，又提高了数据清洗的效率，满足数据清洗的要求。因此，本文样本选取频率为 432 s/条，数据共包含 8 600 条运行数据。

对于缺失数据，当前在数据分析中对应的处理方法多为删除缺失值和修复缺失值 2 种。为了在保证数据高质量的前提下保留数据的完整性，本文将缺失数据填充为 0，然后利用模型进行缺失数据的修复工作。

汽轮机组的整体运行情况与运行边界条件具有较强的依变关系。选取边界条件作为输入量使模型具有更普遍的适应性，并且可以避免数据清洗效果受到其他多项输入数据质量的影响。汽轮机组的边界条件包括可控边界条件（主蒸汽压力、主蒸汽温度、减温水量、再热蒸汽温度、循环水流量）以及不可控边界条件（负荷、环境温度、大气压力、煤质）。

皮尔逊系数（式(8)）可以度量 2 个变量之间的相关性。系数取值介于−1~1，正负值仅表示变量之间呈现的正负相关性。皮尔逊系数的绝对值越大，二者相关性越强。

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left[\left(X-u_X\right)\left(Y-u_Y\right)\right]}{\sigma_X\sigma_Y} \tag{8}$$

其中：u_X、u_Y和σ_X、σ_Y分别为变量 X、Y 的均值和标准值。

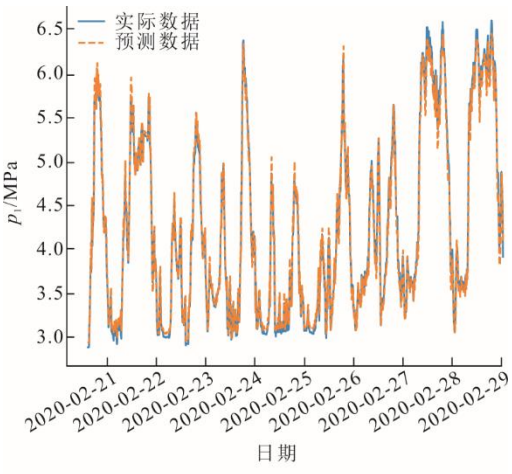
计算待清洗数据与各边界条件的皮尔逊相关系数，结果见表 2。选取与待清洗数据（也即输出量）相关性较强的前 3 个机组边界条件（负荷、主蒸汽压力、主蒸汽温度）作为输入量，待清洗数据一段抽汽（一抽）压力（p_1）、三段抽汽（三抽）压力（p_3）、主凝结水流量（D_{c_sj}，简称 D_j）作为输出量。

表 2 待清洗数据与各边界条件的皮尔逊相关系数
Tab.2 Pearson correlation coefficients between the data to be cleaned and each boundary condition

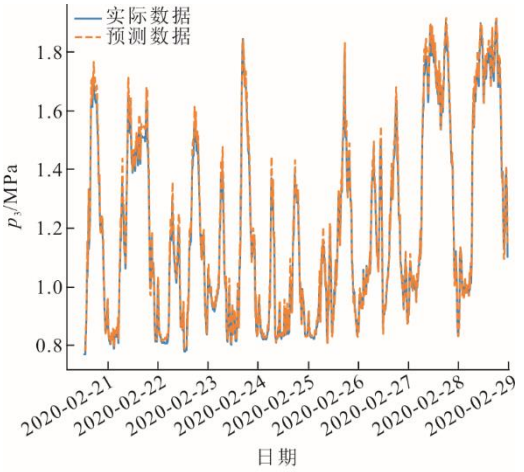
边界条件	待清洗数据		
	p_1	p_3	D_j
负荷	0.999	0.999	0.995
主蒸汽温度	−0.403	−0.404	−0.401
主蒸汽压力	0.920	0.926	0.922
减温水量	0.202	0.201	0.200
大气压力	−0.021	−0.024	−0.021
再热蒸汽温度	−0.035	−0.036	−0.035
循环水入口温度	0.110	0.114	0.108

2.2 模型预测性能验证

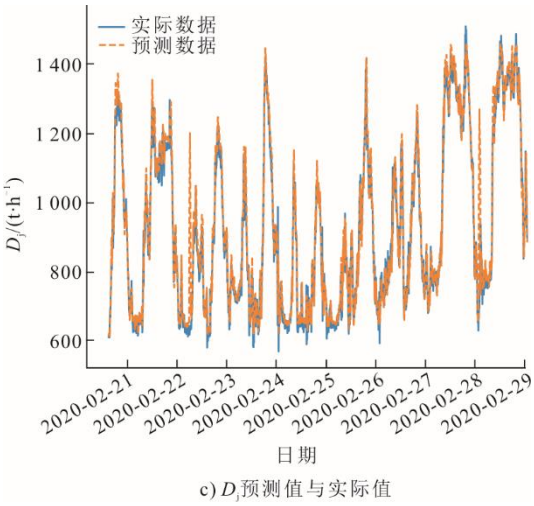
DLSTM 数据清洗模型的构建以模型对于输入输出之间良好的拟合能力作为基础。模型预测性能好坏直接关系到模型异常值剔除以及之后的数据填充。利用已训练好的模型对一抽压力、三抽压力、主凝结水流量进行预测，并且与实际正确的运行数据进行对比，以验证模型的拟合预测性能。图 4 为模型预测值与实际值。由图 4 可知，模型具有较为准确的预测性能，其以 δ_{MAE} 作为损失函数的损失可以稳定在 0.023 以下，误差百分比（式(9)）在 0.30%~0.45%，均表明模型的预测值与实际值的差值很小。



a) p_1预测值与实际值



b) p_3预测值与实际值



c) D_j预测值与实际值

图 4 模型预测值与实际值
Fig.4 The model predicted values and the actual values

$$P_{error}=\left|\frac{a-p}{a}\right|\times 100\% \tag{9}$$

其中：P_{error}为误差百分比；a、p 分别为实际值和预测值。

由图 4 还可知，随着时间变化，预测趋势与实际数据分布趋势完全相符，可以说明模型很好地学习到了数据的时序性分布规律。良好的模型预测能力为之后准确的数据清洗构建了基础。

2.3 待清洗数据构成

以该机组历史运行分析后得到的正常数据为基础，人工模拟异常数据进行清洗验证。8 600 条数据中：前 6 920 条为训练集，只包含正常的数据；后 1 680 条为测试集，是异常值和正常值的混合数据。表 3 为人工模拟异常数据集情况。由表 3 可知，汽轮机 3 个测试集异常值占比分别为 5.06%、3.69%、5.65%，与机组实际运行数据异常、正常样本所存在的样本失衡现象吻合，具有实际应用性。

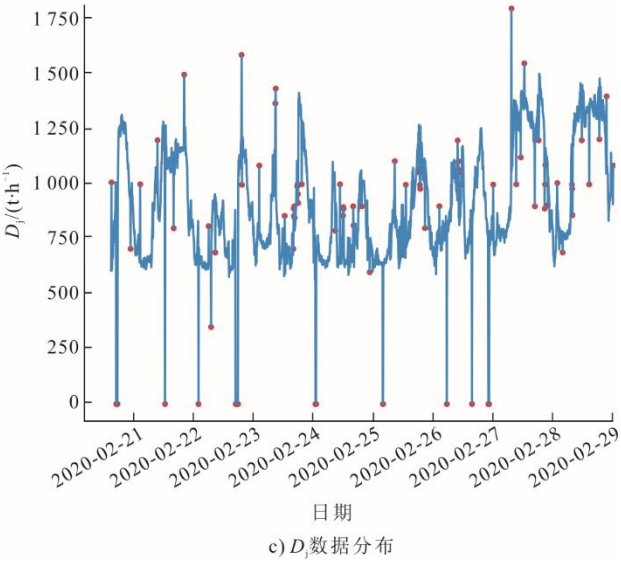
表 3 人工模拟异常数据集情况
Tab.3 Manual simulation of abnormal data sets

待清洗数据类型	缺失/个	时序异常/个	其他异常/个	正常数据/个	异常比/%
一抽压力	15	10	60	1 595	5.06
三抽压力	20	12	30	1 618	3.69
凝结水流量	26	19	50	1 585	5.65

图 5 为 3 个待清洗数据的原始数据分布。图 5 中，红色点为异常数据，0 为缺失值。由图 5 可知，异常数据分布较为复杂，无法使用简单阈值或者统计学方法完成“脏数据”的检测工作。

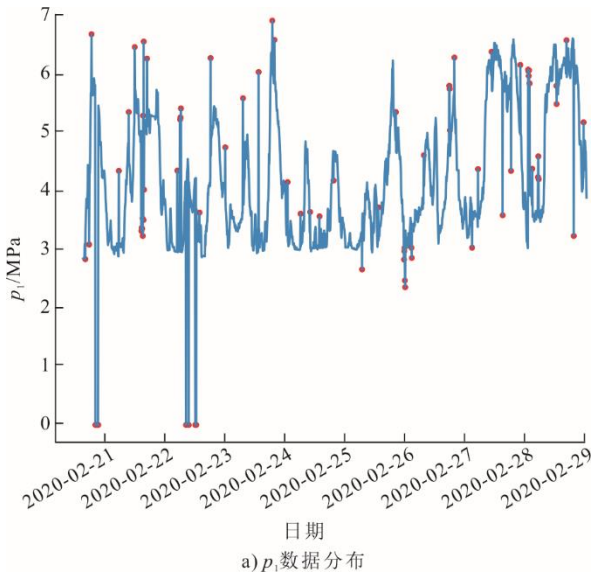
2.4 深度长短记忆实现汽轮机数据清洗

用前 6 920 条数据进行训练得到的模型对待清洗数据进行数据清洗。当待处理数据输入模型中，各数据的预测值与实际值之间的残差如图 6 所示。

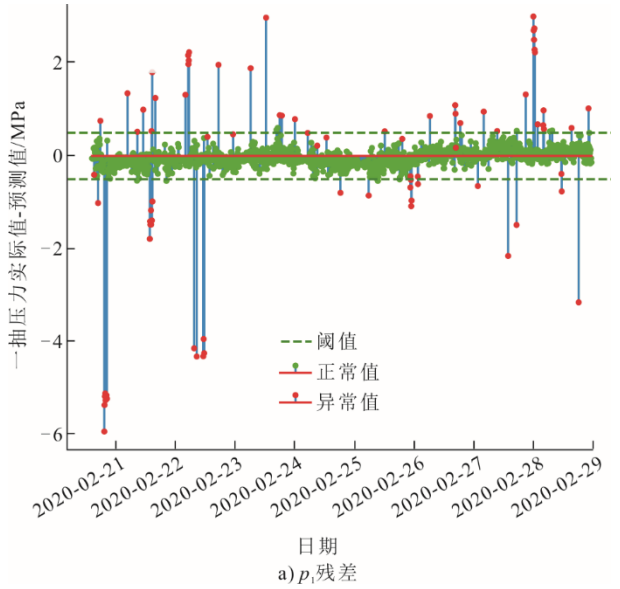


c) D_3数据分布

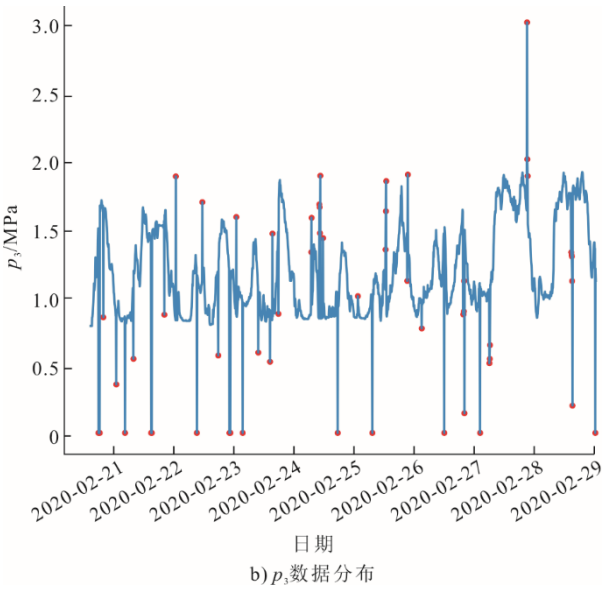
图 5 待清洗数据分布
Fig.5 Data distribution of the data to be cleaned



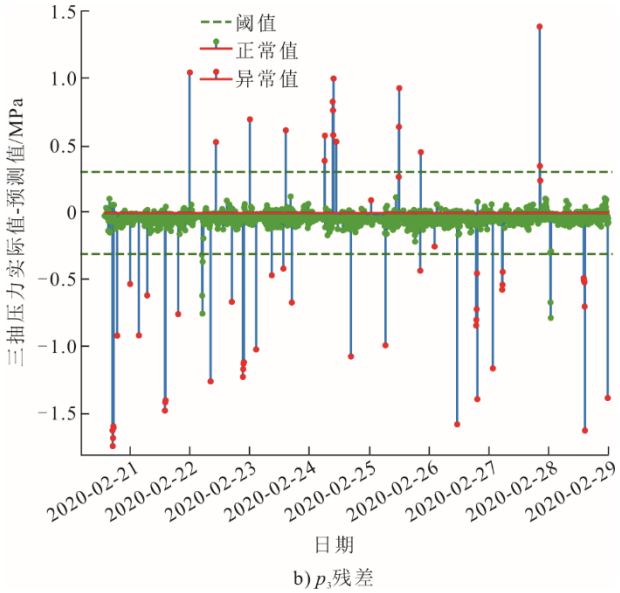
a) p_1数据分布



a) p_1残差



b) p_3数据分布



b) p_3残差

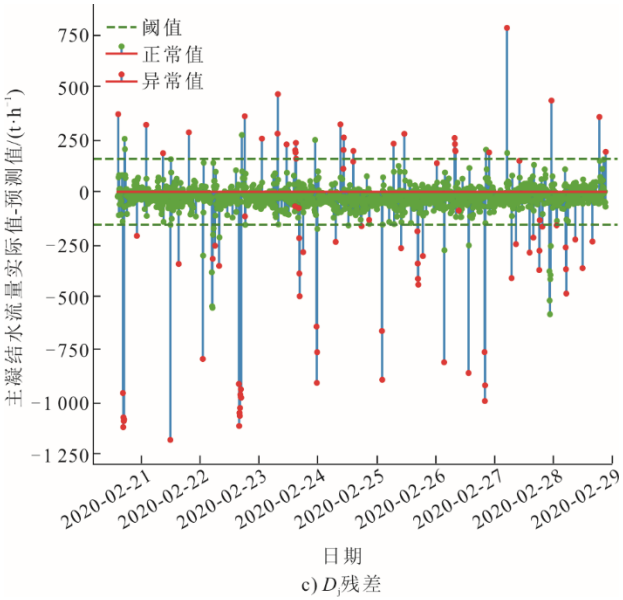


图 6 DLSTM 清洗数据残差
Fig.6 The DLSTM cleaning data residuals

图 6 中，红色为实际异常值所对应残差，绿色为实际正常值所对应残差。据之前训练所得模型选取合适阈值 s。将所有残差中预测值与实际值之间的残差小于阈值 s 的数据判别为正常值，阈值外的判别为异常值。由图 6 可知，模型可以检测出绝大多数的异常值，误判、漏判值较少。

异常值检测的评价指标一般选取 $F1$ 分数、召回率以及准确率[20]。由于异常值检测中样本失衡严重，准确率计算（式(10)）受到大量正确值的影响，各模型准确度中异常值检测能力会遭受正常值样本的湮没，不能仅凭此时的准确率作为衡量指标。对此，提出了改进准确率公式（式(11)），将准确率的计算与样本分布比例进行结合，可以避免样本失衡带来的问题。为了直观理解模型对于异常值的检测能力，将异常数据视为正例，正常数据视为负例，预测与实际所对应的关系见表 4。召回率表示实际异常数据中被检测为异常所占的比例，但召回率忽略了数据清洗中既要尽可能多地检测出真正的异常值，也要避免正确数据被误检，所以引入了 $F1$ 分数（式(12)）来调和。

表 4 混淆矩阵
Tab.4 The confusion matrix

项目	实际异常	实际正常
预测异常	TP	FP
预测正常	FN	TN

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{10}$$

$$\text{Accuracy}=\left(\frac{\text{TP}}{\text{TP+FN}}+\frac{\text{TN}}{\text{FP+TN}}\right)\div 2 \tag{11}$$

$$F1=\frac{2\text{TP}}{2\text{TP+FP+FN}} \tag{12}$$

式(12)可以同时兼顾召回率与精确率（预测异常中确为异常的比例），而改进准确度则是考虑了模型对于正常值以及异常值判断的准确性能。最终得到数据清洗结果见表 5。

表 5 数据清洗结果
Tab.5 The data cleansing results

数据	清洗模型	$F1$ 分数/%	改进准确率/%
一抽压力 p_1	箱线图	30.00	58.82
	k-means 聚类	53.59	73.27
	LOF	80.25	88.20
	深度长短记忆	86.42	92.60
三抽压力 p_3	箱线图	3.17	50.81
	k-means 聚类	90.27	91.13
	LOF	69.29	84.83
	深度长短记忆	92.06	96.59
主凝结水流量 D_j	箱线图	44.26	64.21
	k-means 聚类	59.89	76.98
	LOF	68.48	93.98
	深度长短记忆	85.26	94.60

根据表 5 的评价指标可知，本文所提出的深度长短记忆模型对于汽轮机运行数据的异常值检测准确率都优于其他常见的检测方法。DLSTM 数据清洗模型可以保证较高的 $F1$ 分数，即相较于其他方法既可以较好地检测出其中的异常值，又可以在一定程度上减少将正确值误判为异常情况的发生。

异常数据识别剔除后会出现数据缺失的情况。为了保证数据的完整性，需要进行数据的填充工作。常见的数据填充法主要有 k 近邻[21]、相似度度量分析[22-23]、均值、深度学习[24]以及预测相关的方法。最近邻以及均值填充方法简洁，但其准确率不高；预测性方法[25]可以充分考虑特征值对于输出之间的关系，准确进行缺失数据的填充工作。

根据之前对于模型预测性能的验证，其损失函数在 0.021，且误差百分比在 0.30%~0.45%，两者均保持在较小的范围内，表明模型对于数据的预测值与实际值差距很小，可以用预测值代替真实值。因此，本文 DLSTM 数据清洗模型对于待清洗数据具有良好的预测性能，可以结合输入量以及之前数据的时序信息进行高效数据拟合填充。图 7 为 DLSTM 模型数据清洗前后对比。由图 7 可知，DLSTM 数据清洗模型不仅对于待清洗数据集拥有较好的异

常值检测能力，而且可以完成对于异常值剔除后的数据填充工作，填充完整的数据集变化趋势符合实际的分布规律。

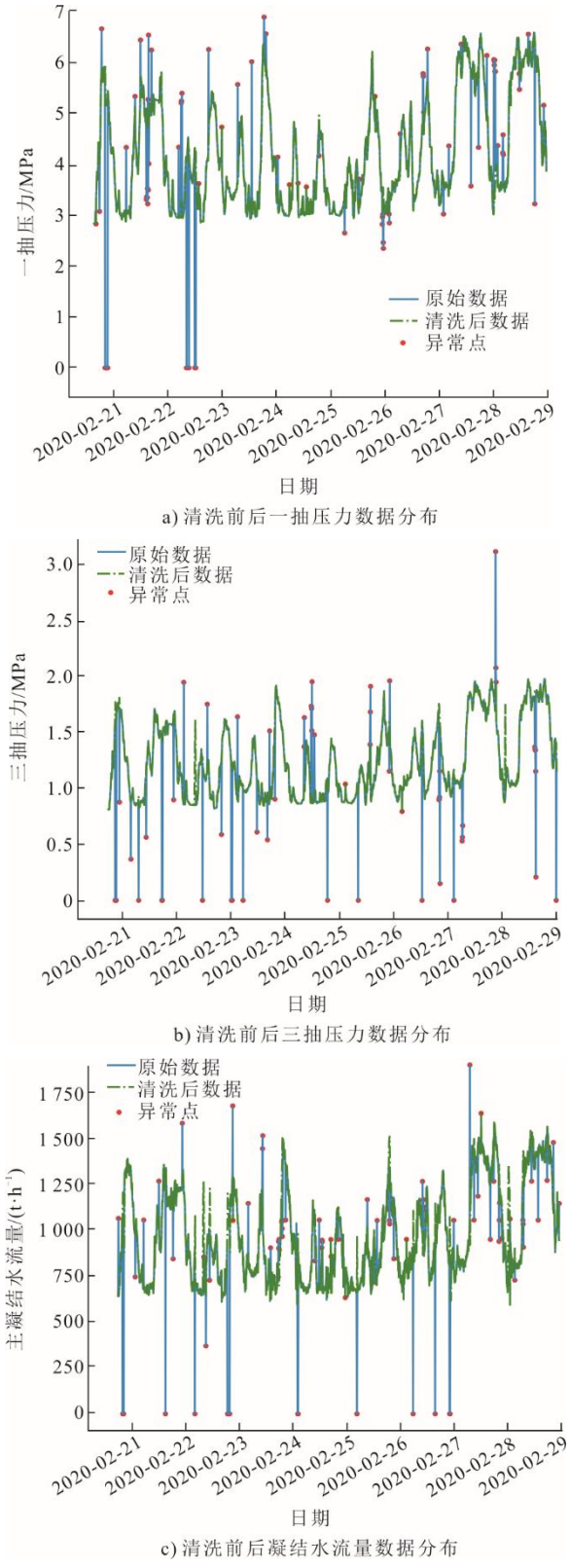


图 7 DLSTM 模型数据清洗前后对比

Fig.7 Comparison before and after the DLSTM model data cleaning

3 结　论

1）构建深度长短记忆数据清洗模型，模型可拟合待清洗数据与输入数据之间的非线性关系，进行预测值的输出，之后根据模型预测值与实际值的残差大小进行异常值的检测。

2）对于评价标准准确率进行了改进，以避免样本失衡所导致的准确率对于异常值检测性能不敏感的问题。

3）本文构建的模型对于汽轮机的异常数据具有较好的检测能力。利用模型对于汽轮机一抽压力、三抽压力以及主凝结水流量进行异常值识别，其改进准确率分别高达 92.60%、96.59%、94.60%。并且与常见的箱型图、*k*-means 聚类以及局部异常因子法进行比较，该模型的异常值识别能力更加准确。

4）DLSTM 数据清洗模型预测误差百分比可以稳定在 0.30%~0.45%，可完成缺失数据以及剔除异常数据后的数据填充工作，保证数据的完整性，从而保证了数据清洗的高效性和完整性。

［参 考 文 献］

［1］ 李琳，董博，郑玉巧. 大型风力机异常功率数据清洗方法[J]. 兰州理工大学学报，2022，48(3)：65-70.
LI Lin, DONG Bo, ZHENG Yuqiao. Abnormal power data cleaning method of large wind turbine[J]. Journal of Lanzhou University of Technology, 2022, 48(3): 65-70.

［2］ 代杰杰，宋辉，杨祎，等. 基于栈式降噪自编码器的输变电设备状态数据清洗方法[J]. 电力系统自动化，2017，41(12)：224-230.
DAI Jiejie, SONG Hui, YANG Yi, et al. State data cleaning method of power transmission and transformation equipment based on stacked noise reduction autoencoder[J]. Power System Automation, 2017, 41(12): 224-230.

［3］ 杨茂，孟玲建，李大勇，等. 基于类 3σ 准则的光伏功率异常数据识别[J]. 可再生能源，2018，36(10)：1443-1448.
YANG Mao, MENG Lingjian, LI Dayong, et al. Identification of photovoltaic power anomaly data based on Class 3σ criterion[J]. Renewable Energy, 2018, 36(10): 1443-1448.

［4］ 何高清，肖健. 轴承尺寸检测数据的异常值检测与数据处理研究[J]. 机电工程，2021，38(2)：198-203.
HE Gaoqing, XIAO Jian. Study on outlier detection and data processing of bearing size detection data[J]. Mechanical and Electrical Engineering, 2021, 38(2): 198-203.

［5］ 许璟琳，彭阳，余芳强. 基于 *k*-means 聚类和离群点检测算法的医院建筑节能诊断方法[J]. 计算机应用，2021，41(增刊 1)：288-292.
XU Jinglin, PENG Yang, YU Fangqiang. Diagnosis method of hospital building energy saving based on *k*-means clustering and outlier detection algorithm[J]. Computer Applications, 2021, 41(Suppl.1): 288-292.

[6] 陈洪涛, 蔡慧, 李熊, 等. 基于 *k*-means 聚类算法的低压台区线损异常辨别方法[J]. 南方电网技术, 2019, 13(2): 2-6.
CHEN Hongtao, CAI Hui, LI Xiong, et al. Identification method of line loss anomaly in low-voltage station area based on *k*-means clustering algorithm[J]. China Southern Power Grid Technology, 2019, 13(2): 2-6.

[7] 贺玉海, 周庆琨, 程埝晟, 等. 基于改进 *K*-Medoids 的组合聚类算法及异常值检测研究[J]. 大连理工大学学报, 2022, 62(4): 403-410.
HE Yuhai, ZHOU Qingkun, CHENG Kunsheng, et al. Combinatorial clustering algorithm and outlier detection based on improved *K*-Medoids[J]. Journal of Dalian University of Technology, 2022, 62(4): 403-410.

[8] 石玉亮, 王呈. 基于 Pearson-LOF 算法的梯联网数据采集端异常帧检测[J]. 控制工程, 2022, 29(8): 1457-1463.
SHI Yuliang, WANG Cheng. Anomaly frame detection at ladder network data acquisition end based on Pearson-LOF algorithm[J]. Control Engineering, 2022, 29(8): 1457-1463.

[9] WANG B Y, LUO X Y, ZHANG S M. An improved outlier detection algorithm K-LOF based on density[J]. Computing, Performance and Communication Systems, 2017, 2(1): 1-7.

[10] 侯振英. 基于 Isolation Forest 的城市道路交通异常检测[D]. 北京: 北京工业大学, 2017: 1.
HOU Zhenying. Detection of urban road traffic anomalies based on imaging forest[D]. Beijing: Beijing University of Technology, 2017: 1.

[11] 吴磊, 康英伟. 基于改进粒子群优化长短时记忆神经网络的脱硫系统 SO_2 预测模型[J]. 热力发电, 2021, 50(12): 66-73.
WU Lei, KANG Yingwei. SO_2 prediction model of desulfurization system based on improved particle swarm optimization long and short-term memory neural network[J]. Thermal Power Generation, 2021, 50(12): 66-73.

[12] 吴飘. 基于深度学习的时序数据异常检测算法研究[D]. 大连: 大连理工大学, 2021: 1.
WU Piao. Research on anomaly detection algorithm of time series data based on deep learning[D]. Dalian: Dalian University of Technology, 2021: 1.

[13] 刘云鹏, 王权, 许自强, 等. 基于多层架构的油中溶解气体数据清洗与异常识别方法研究[J]. 华北电力大学学报(自然科学版), 2022, 49(1): 81-89.
LIU Yunpeng, WANG Quan, XU Ziqiang, et al. Research on data cleaning and anomaly identification method of dissolved gas in oil based on multilayer architecture[J]. Journal of North China Electric Power University (Natural Science Edition), 2022, 49(1): 81-89.

[14] 刘建伟, 宋志妍. 循环神经网络研究综述[J]. 控制与决策, 2022, 37(11): 2753-2768.
LIU Jianwei, SONG Zhiyan. A review of recurrent neural network research[J]. Control and Decision, 2022, 37(11): 2753-2768.

[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[16] 朱乔木, 李弘毅, 王子琪, 等. 基于长短期记忆网络的风电场发电功率超短期预测[J]. 电网技术, 2017, 41(12): 3797-3802.
ZHU Qiaomu, LI Hongyi, WANG Ziqi, et al. Ultra-short-term prediction of wind farm power generation power based on long-term short-term memory network[J]. Power Grid Technology, 2017, 41(12): 3797-3802.

[17] 宋旭东, 朱大杰, 杨杰, 等. 一种基于 L2 正则化迁移学习的变负载工况条件下故障诊断方法[J]. 大连交通大学学报, 2022, 43(2): 106-109.
SONG Xudong, ZHU Dajie, YANG Jie, et al. A fault diagnosis method under variable load conditions based on L2 regularization transfer learning[J]. Journal of Dalian Jiaotong University, 2022, 43(2): 106-109.

[18] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//International Conference on Learning Representations. Ithaca, NYarXiv.org, 2014.

[19] 杨锡运, 赵泽宇, 杨岩, 等. 基于时空信息组合的分布式光伏功率预测方法研究[J]. 热力发电, 2022, 51(8): 64-72.
YANG Xiyun, ZHAO Zeyu, YANG Yan, et al. Research on distributed photovoltaic power prediction method based on spatio-temporal information combination[J]. Thermal Power Generation, 2022, 51(8): 64-72.

[20] 贺之豪. 数据驱动的汽轮机组性能诊断研究[D]. 北京: 华北电力大学, 2019: 1.
HE Zhihao. Data-driven turbine unit performance diagnosis study[D]. Beijing: North China Electric Power University, 2019: 1.

[21] 徐鸿艳, 孙云山, 秦琦琳, 等. 缺失数据插补方法性能比较分析[J]. 软件工程, 2021, 24(11): 11-14.
XU Hongyan, SUN Yunshan, QIN Qilin, et al. Comparative analysis of performance of interpolation method for missing data[J]. Software Engineering, 2021, 24(11): 11-14.

[22] 纪德洋, 金锋, 冬雷, 等. 基于皮尔逊相关系数的光伏电站数据修复[J]. 中国电机工程学报, 2022, 42(4): 1514-1523.
JI Deyang, JIN Feng, DONG Lei, et al. Data repair of photovoltaic power plant based on Pearson correlation coefficient[J]. Proceedings of the CSEE, 2022, 42(4): 1514-1523.

[23] 蔡文斌, 程晓磊, 王鹏, 等. 基于 DBSCAN 二次聚类的配电网负荷缺失数据修补[J]. 电气技术, 2021, 22(12): 27-33.
CAI Wenbin, CHENG Xiaolei, WANG Peng, et al. Patching of missing data of distribution network load based on DBSCAN secondary clustering[J]. Electrical Technology, 2021, 22(12): 27-33.

[24] 张晟斐, 李天梅, 胡昌华, 等. 基于深度卷积生成对抗网络的缺失数据生成方法及其在剩余寿命预测中的应用[J]. 航空学报, 2022, 43(8): 441-455.
ZHANG Shengfei, LI Tianmei, HU Changhua, et al. Missing data generation method based on deep convolution generation adversarial network and its application in remaining life prediction[J]. Journal of Aeronautics, 2022, 43(8): 441-455.

[25] 郑欣彤, 边婷婷, 张德强, 等. ARIMA 和 LSTM 方法长时间温度观测数据缺失值插补的比较[J]. 计算机应用, 2022, 42(增刊 1): 130-135.
ZHENG Xintong, BIAN Tingting, ZHANG Deqiang, et al. Comparison of missing value interpolation in long-term temperature observation data of ARIMA and LSTM methods[J]. Computer Applications, 2022, 42(Suppl.1): 130-135.

（责任编辑　刘永强）