DOI:10.19651/j.cnki.emt.2415878

基于 CEEMDAN-CNN-LSTM 的供热异常数据检测与清洗*

梁晓龙[1] 李金刚[1] 徐平平[1] 马雅楠[2] 孟现阳[2]
(1.国能宁夏供热有限公司 银川 750004;2.西安交通大学热流科学与工程教育部重点实验室 西安 710049)

摘 要:利用供热系统的准确参数,对监测系统状态、识别异常情况具有指导意义。然而大量终端数据,可能存在失真问题,为此本文提出了一种异常数据检测和清洗方法。采用信号模态分解结合深度学习,构建数据检测与清洗模型。首先对由 DeST 获取的供热负荷进行 CEEMDAN 模态分解;其次将产生的本征模函数和剩余量输入 CNN-LSTM 深度学习预测模型,获取高精度结果;最后基于预测值和待清洗数据的偏差,完成检测和数据清洗,提高末端数据准确性。结果表明:本文的 CEEMDAN-CNN-LSTM 组合模型,识别异常数据的准确率和 F1 分数达到:91.36%,86.21%,优于其他 3 种模型;利用预测值替换异常值,保证数据集的完整准确。

关键词:异常检测和清洗;模态分解;深度学习;供热系统
中图分类号:TN98 **文献标识码**:A **国家标准学科分类代码**:520.2060

Heating data detection and cleaning based on CEEMDAN-CNN-LSTM

Liang Xiaolong[1] Li Jingang[1] Xu Pingping[1] Ma Yanan[2] Meng Xianyang[2]
(1. Guoneng Ningxia Heating Co., Ltd., Yinchuan 750004, China;
2. MOE Key Laboratory of Thermo-Fluid Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Using the accurate parameters of the heating system has guiding significance for monitoring system status and identifying abnormal conditions. However, a large amount of terminal data may have distortion problems. To address this, this paper proposed a method for detecting and cleaning abnormal data. Signal modal decomposition combined with deep learning was used to construct a detection and cleaning model. The first step involves conducting CEEMDAN mode decomposition of the heating load obtained by DeST. Subsequently, the intrinsic mode functions and residual quantities generated from the decomposition are input into the CNN-LSTM deep learning prediction model to achieve high-precision prediction results. Finally, based on the deviation between predicted values and data to be cleaned, abnormal detection and data cleaning are completed. The CEEEMDAN-CNN-LSTM combined model in this paper achieves superior accuracy and F1 scores of 91.36% and 86.21%, respectively, outperforming the other three models. Moreover, the predicted values can be used to replace abnormal values, ensuring the integrity and accuracy of the final data set.

Keywords: anomaly detection and cleaning; modal decomposition; deep learning; heating system

0 引 言

供热系统在城市基础设施中占有重要地位,主要负责冬季取暖和居民热水。供热质量与运行状态息息相关。随着我国城镇化发展的推进和能源体系的转型,目前供热系统面临智能化升级的挑战[1]。在“互联网+”智慧能源系统建设的推动下,结合物联网和大数据平台[2],能够有效模拟供热系统的运行状态,实时获取相关参数。利用系统终端数据平台,监测系统状态并进行实时调控,能够有效改善供热质量。然而末端采集到大量供热数据,可能存在缺失、异常及失真等问题,导致供热企业的管网控制和热源调度措施,出现一定延迟和错误[3]。当前研究大多采用基于统计、密度和聚类等传统方法,对异常数据识别的准确度较低,导致供热质量差和用户满意度降低。因此,供热系统终端数

收稿日期:2024-04-18
*基金项目:陕西省自然科学基础研究计划(2024JC-YBMS-257)项目资助

据的异常检测和数据清洗尤为重要,对提高系统效率、确保用户舒适度等具有重要意义[4]。

在集中供热领域,一些企业通常采用传统方式来检测异常情况[5],如定期管道检修和人工经验判别。但需要耗费巨大人力物力,且受管网热惰性影响,不能准确及时地辨别异常数据和系统故障,严重影响供热质量和能耗。随着系统智能化发展,企业多转向采用监视控制与数据采集(supervisory control and data acquisition, SCADA)系统监测和管理供热系统[6-7]。广泛部署采集传输设备,收集、记录和存储大量终端数据[8]。然而在数据测量(设备故障)、传输(通信故障)、控制(时间延迟)过程中,受外界(恶劣天气)和人为(专业知识缺乏、记录误差)等干扰,导致异常情况出现[9]。

改善原始数据集的准确性和可信度势在必行。因此,研究多聚焦异常检测和数据清洗技术[10-11]。通过分析高质量的历史数据,深入研究隐藏规律和系统特征,实现精细管控。在异常数据识别方面,文献提出多种方法解决不同领域异常检测问题,包括基于数据统计、距离、密度、聚类和相似度等[12]。在数据清洗方面,采用机器学习和深度学习等填补空缺和修正数据[13-14]。然而针对供热系统的问题,亟需结合并扩展以上方法改善其末端数据有效性。

夏博等[15]提出综合利用阈值法和 K 均值聚类法,通过人工设定阈值和经验参数,识别不同性质的能耗数据异常值。Sun 等[16]采用专家知识与支持向量机结合的两步训练策略辨别异常类型,但处理大规模异常数据时,可能导致回归模型畸变和拟合性能下降。张童[17]提出基于皮尔逊系数与时序回归分析模型,开展集中供热系统的异常点检测。将检测结果中的异常值,进行高效清洗也是确保数据质量的关键步骤。柳亚楠[18]采用箱线图四分位法和拟合法剔除极端值,并结合 K 均值聚类清洗数据,但簇数选择对结果影响较大。尹薇[19]提出了一种基于自回归模型清洗时间序列技术,采用迭代思想设计自更新系数进行缺失值填充。匡俊搴等[20]提出了基于深度神经网络的快速异常数据清洗算法,采用迭代阈值收缩和深度学习解决清洗难题,结果的完整度和准确性高于传统方法。现有研究尚未充分运用数据挖掘和深度学习等先进技术来检测并清洗供热异常数据,且未提出综合方法。

基于以上问题,本文提出应用信号处理、机器学习和深度学习等方法交叉融合,搭建新模型来高效检测和清洗供暖数据。以某酒店供热系统为例,首先对终端数据进行预处理,包括数据规范化、特征选择等;其次选择自适应白噪声集合经验模态分解(complete ensemble empirical mode decomposition with adaptive noise, CEEMDAN)与卷积神经网络(convolutional neural networks, CNN)结合长短期记忆 (long short-term memory,LSTM)的组合模型进行高精度预测,辨识不符合预期的数据;最后评价模型异常辨识率并清洗数据。完整准确的末端数据,为后续数据分析和应用提供支持,及时优化调整运行方案[21],保障系统安全、经济、节能环保性。

1　模型理论介绍

1.1　信号分解模型

1)经验模态分解

经验模态分解(empirical mode decomposition,EMD)是一种信号处理方法,特别适用于任何非线性、非平稳的信号分解。提取不同频率成分,分解产生不同时间尺度的本征模态函数(intrinsic mode function, IMF)和残余项。基本思路是通过数据局部最大和最小值,获取上下包络的平均值,再用原始信号减去均值信号得出剩余分量,最终求得 IMF,其分解结果如式(1)所示。

$$u(t)=\sum_{i=1}^{N}c_i(t)+res_N(t) \tag{1}$$

式中:$u(t)$为原始供热负荷序列;$c_i(t)$为第 i 个 IMF 分量;$res_N(t)$为残余项。然而实际数据存在噪声干扰,可能出现模态混叠和虚假分量的现象。

2)集合经验模态分解

提出集合经验模态分解(ensemble empirical mode decomposition,EEMD),解决上述问题获得理想子序列,在分解过程中添加高斯白噪声(需要初始化集成平均次数 m 和白噪声幅值 α),实现信号在适当时间尺度自动分布,提高结果可靠性。但经过有限平均计算后仍存在误差。在原始序列 $u_0(t)$ 加入高斯白噪声信号 $n_i(t)$,变为 $u(t)=u_0(t)+n_i(t)$,经过迭代产生 IMF 和残差,如式(2)所示。

$$u(t)=\sum_{i=1}^{N}c_i(t)+res_N(t) \tag{2}$$

式中:$c_i(t)$是第 i 个 IMF 分量,$res_N(t)$为残差项。

3)自适应白噪声集合经验模态分解

为解决 EMD 和 EEMD 的不足,因此本文利用自适应白噪声对时间序列数据分解,使误差在相对较少的平均操作下趋近于零。CEEMDAN 的计算方法如式(3)~(8)所示,分解步骤如图 1 所示。

(1)向原始信号序列 $u(t)$添加标准正态白噪声,ε_0 为信噪比,总共加入了 I 次白噪声;

$$u_1(t)=u(t)+\varepsilon_0\omega_i(t)(i=1,2,\cdots,I) \tag{3}$$

(2)利用 EMD 将 $u_1(t)$分解得到 I 组结果,将分量平均化得到模态分量 IMF_1 和残余项 $res_1(t)$;

$$\overline{IMF_1}=\frac{1}{I}\sum_{i=1}^{I}IMF_1^i \tag{4}$$

$$res_1(t)=u(t)-\overline{IMF_1} \tag{5}$$

(3)向残余信号添加特定噪声后继续 EMD 分解;

$$\overline{IMF_{k+1}}=\frac{1}{I}\sum\nolimits_{i=1}^{I}F_1[res_k(t)+\varepsilon_kF_k(\omega_i(t))] \tag{6}$$

$$res_k(t)=res_{k-1}(t)-\overline{IMF_k} \tag{7}$$

(4)重复执行步骤(3),直至残余信号是单调函数或常

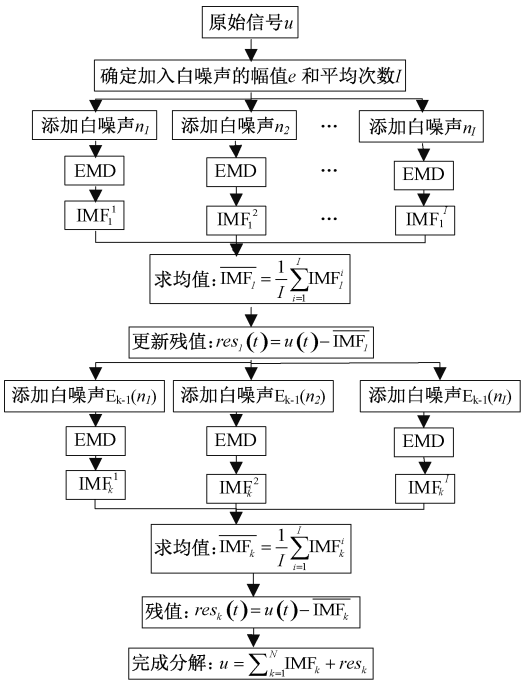

图 1　CEEMDAN 的分解步骤

Fig. 1　Decomposition steps of CEEMDAN

量时完成 EMD 分解。假设获得 N 个本征模态函数 IMF，则可写出原始信号的 CEEMDAN 分解式：

$$u(t)=\sum_{k=1}^{N}IMF_k+res_N(t) \tag{8}$$

EEMD 方法主要通过多次 EMD 分解和集合平均方式降低模态混叠，将添加白噪声后的 M 个信号直接分解，并求对应 IMF 间均值；而 CEEMDAN 方法引入自适应噪声处理，每求完一阶 IMF 分量，又重新给残值加入白噪声，再求此时 IMF 分量均值，并多次迭代。因此在较小的平均次数下，完备性很好，更适用于复杂信号分解和特征提取。

1.2　深度学习算法

随着深度学习技术的快速发展，学者们将数据清洗方法的研究转向深度学习模型，利用神经网络和深度学习方法自动识别修复数据中的异常值[21]。而历史供热负荷与时间序列高度相关，因此本文拟采用时间序列预测模型。其中 LSTM 模型是一种用于处理时间序列的特殊循环神经网络，解决了梯度消失和爆炸问题，提高模型准确可用性。而 CNN 模型可以有效提取输入数据特征，采用权值共享与局部连接的方式，特征提取能力良好。

1）CNN-LSTM 神经网络模型

CNN 包括卷积层、池化层、全连接层等重要结构，首先通过卷积层提取输入序列的特征信息，然后利用池化层精炼提取上一步的特征，最后收集所有数据至全连接层，并供输出使用。因此，选择此网络不仅能有效提取供热序列特征，且能简化网络结构和复杂度。LSTM 基于循环神经网络，特点在于反向传播误差，实现时间序列信息的长期记忆，更好地处理长序列和长距离依赖关系。因此考虑结合 CNN 与 LSTM 模型的优势，充分提取输入序列特征，实现对数据未来的精准预测。

结合输入序列的信息流动，发掘时序特征。通过 3 个门控单元控制，判断需要保留或遗忘的信息，其结构单元如图 2 所示。

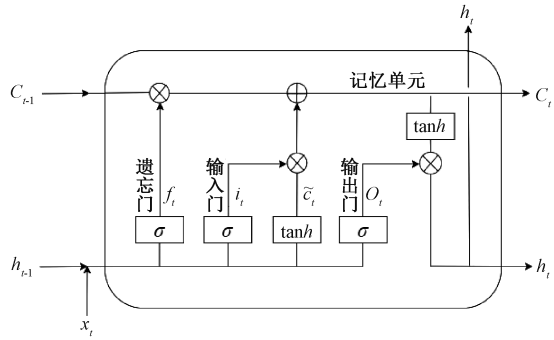

图 2　LSTM 单元结构

Fig. 2　Structure of LSTM unit

其存储、更新、遗忘过程的计算如式(9)～(14)：

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{9}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{10}$$

$$\tilde{c}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C) \tag{11}$$

$$C_t=f_t\otimes C_{t-1}+i_t\otimes\tilde{c}_t \tag{12}$$

$$O_t=\sigma(W_O\cdot[h_{t-1},x_t]+b_O) \tag{13}$$

$$h_t=O_t\times\tanh(c_t) \tag{14}$$

其中，W 和 b 代表权重系数和偏差，f、i、C、O 代表遗忘、输入、输出门和记忆单元；t 表示当前时刻，f_t，i_t，O_t 代表当前遗忘、输入和输出门控状态，$\tilde{c}_t$、C_t 代表当前输入和记忆单元状态，x_t、h_t、h_{t-1} 代表当前输入、输出和上一时刻输出；·表示数量积，⊗表示 Hadamard 积。

2）模型的建立和应用

在本文的供热负荷预测中，利用卷积结合长短期记忆网络，划分训练集和测试集，不断优化模型参数；对比预测和实际供热负荷的偏差，判断数据异常值，完成检测和清洗，具体流程如图 3 所示。

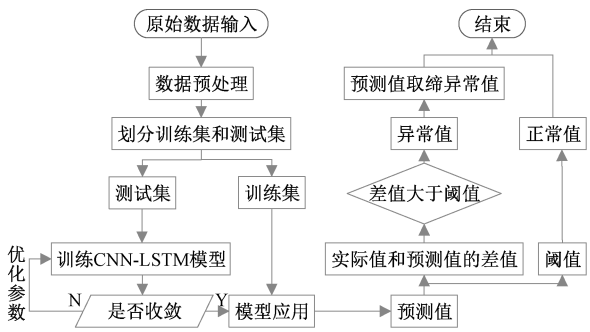

图 3　样本数据清洗模型流程

Fig. 3　Sample data cleaning model process

首先，学习样本数据的时序规律，建立输入与输出的复

杂非线性关系,选择特征向量矩阵训练模型;采用训练集拟合变量间的复杂关系;采用测试集数据验证模型,通过损失函数评价模型预测能力,实现对供热负荷的高精度预测;采用精度满足要求的模型预测值,对比识别异常情况,并替换、填补待检测数据集,保障数据集的完整准确。

1.3　构建组合模型算法

针对不稳定的供热负荷变化,为提高时序数列的预测精度,本文选择 CEEEMDAN-CNN-LSTM 混合模型,作为异常检测和清洗的数据基础。该模型主要由以下关键部分组成:原始序列分解与重构、卷积网络-长短期记忆网络模型、异常识别和评价以及最后的数据清洗环节。首先选择 DeST 建筑负荷模拟软件,通过设置室外气象环境、建筑模型参数、建筑设备系统的控制调节策略、热舒适性要求等条件参数,模拟整个供暖期历史供热负荷。后续进行数据预处理和归一化等,准备好输入数据。由于原始序列具有信息重叠现象,模型选择结合 CEEMDAN 模态分解方法,获取不同波形不同特征的本征模函数和残余项,使非平稳时序数据,转化为多个相对平稳的子序列。划分数据集为训练集和测试集,训练并优化混合模型参数,输出预测结果和图形,通过相关指标评估模型性能。最后,利用精度较高的模型训练结果,识别待清洗数据集的异常值,并进行异常检测结果评价;以预测结果为依据完成数据清洗。模型构建和应用流程如图 4 所示。

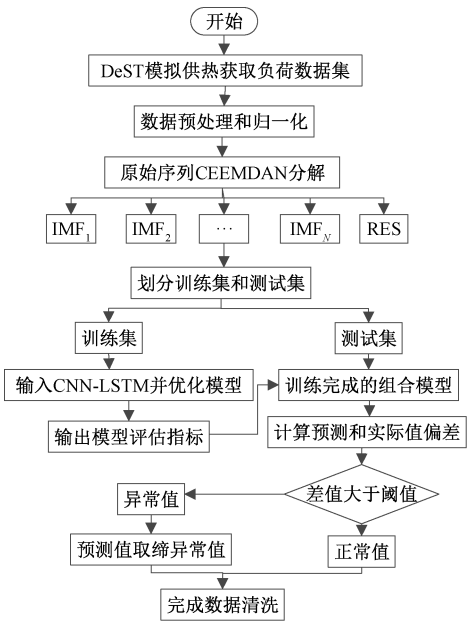

图 4　CEEMDAN-CNN-LSTM 清洗模型流程

Fig. 4　CEEMDAN-CNN-LSTM data cleaning model process

2　CEEMDAN 与 CNN-LSTM 的数据清洗模型验证

2.1　供热数据预处理

供热系统存在时变性、热惯性和时滞性等特点,导致传统控制方法很难对其进行合理有效的管理和调控。发展趋势结合 SCADA 系统和 PLC 技术,实现自动化监控和过程控制。此过程涉及从不同的数据源,以适当的频率,实时采集多种数据类型,整理这些数据,发现存在噪声、缺失、突变等问题。由于缺少实际供热现场参数,采用 DeST 负荷模拟软件,得到热负荷数据集(以当地某酒店为例)。包括整个供暖期的历史运行和环境参数等。对采集数据进行预处理,提高原始序列的可靠性。

1)异常值处理

通常由突发事故造成,表现为某些数据点会与相邻数据产生明显差别,形成阶跃式突变。由于系统热惯性,前后存在关联难以出现突变,首先剔除突变值(零、无穷数等)。

2) 缺失值填补

在剔除突变值后,需要对缺失位置进行填补。传统采用线性差值法、中位值、临近数等。线性差值法使用连接两个已知量的直线,确定两者之间的未知量,以保证历史数据集的完整性。

3) 数据归一化

对于神经网络模型的数据集,不同特征的范围相差较大时易出现不收敛,需要对每个特征变量归一化处理。采用最大-最小值归一化,将特征数据归一化到[0,1],如式(15)所示。训练模型时,使用利用归一化后的数据集,其预测结果也位于 [0,1]内,因此需要对结果进行反归一化,得到最终的预测输出,如式(16)所示。

$$X' = \frac{X - X_{\min}}{X_{\min} - X_{\max}} \tag{15}$$

$$X = X'(X_{\max} - X_{\min}) + X_{\min} \tag{16}$$

其中,X 为原始特征变量,X'为最大-最小值归一化特征变量,$X_{\min}$、$X_{\max}$ 为特征变量最小值、最大值。

2.2　序列分解与重构

分析供热场景热负荷原始数据,针对其非线性、非平稳变化的序列特征,利用模态分解挖掘其中规律,降低预测过程复杂度。以 DeST 某月历史供热负荷为例,时间为 2023 年 2 月 1 日～2 月 28 日,采样间隔为 1 h,共得到 672 组负荷数据。将原始热负荷序列分解,产生多个能够代表其非线性特征的子序列(IMF)与残差序列(RES),便于提高后续预测精度和效率,优化大数据平台的质量。分别采用 EMD、EEMD、CEEMDAN 等模态分解方法,产生不同数量不同特征的本征模态函数和剩余量,其结果如图 5～7 所示。

分析以上结果:EMD 分解可得 6 组 IMF 和 1 组 RES;EEMD 分解可得 8 组 IMF 和 1 组 RES;ICEEMDAN 分解可得 10 组 IMF 和 1 组 RES,算法的信噪比设定为 0.6,额外增添噪声的数量设定为 100。与 EEMD 和 EMD 相比,CEEMDAN 分解产生的分量更多,且分量波形即包含的特征更明显,其分解受噪声的影响较小,在单个 IMF 中不包含不同或相似的序列,不存在模态融合。分解后的 IMF 都

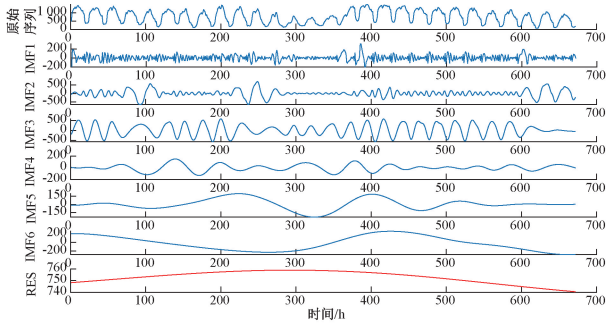

图 5　EMD 模态分解结果

Fig. 5　Results of empirical mode decomposition

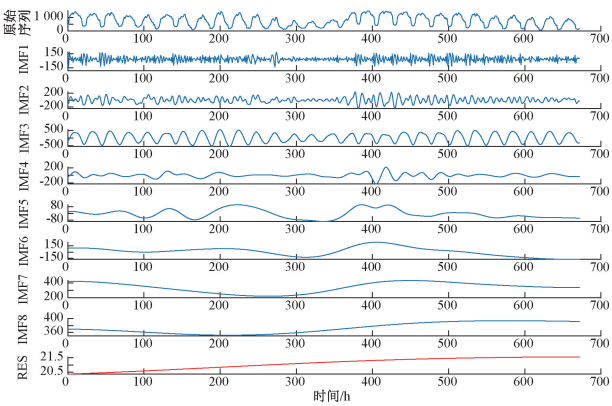

图 6　EEMD 模态分解结果

Fig. 6　Results of ensemble empirical mode decomposition

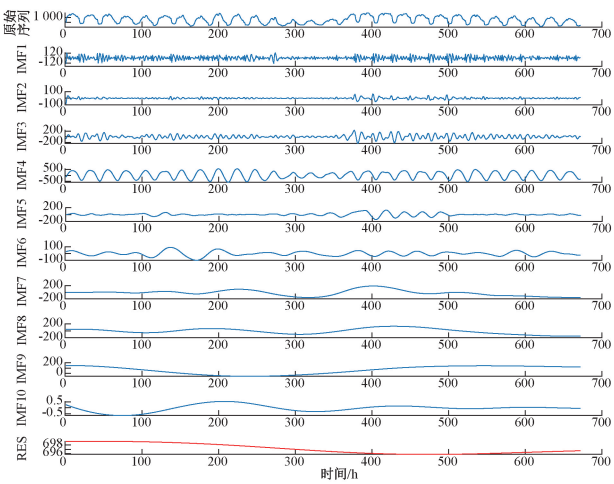

图 7　CEEMDAN 模态分解结果

Fig. 7　Results of complete ensemble empirical mode decomposition with adaptive noise

有清晰稳定的波形特征，RES 值的趋势随时间而减少。结果表明，CEEMDAN 对负荷序列的分解效果优于其他两种分解方法。

2.3　深度学习模型训练

区域供热系统的运行受到多重因素影响。在输入参数的选择方面，采用皮尔逊相关系数分析法，相关系数值越大两者相关性越强。计算多环境因素与热负荷的相关性，计算如式(17)所示。

$$\rho_{U,V}=\frac{\mathrm{cov}(U,V)}{\sigma_U\sigma_V}=\frac{E((U-u_U)(V-u_V))}{\sigma_U\sigma_V} \tag{17}$$

式中：$\rho_{U,V}$ 为变量 U、变量 V 的相关系数，u_U、u_V 为两参数平均值，σ_U、σ_V 为两参数集的方差。

原始数据包括某月的气象数据和供热数据两部分。初始选择 6 个环境因素进行计算，具体结果如表 1 所示。本文输入参数选择高相关性因素，舍弃其余低关联性因素后，综合选择 4 个环境因素。

表 1　热负荷与各环境因素的相关系数计算结果

Table 1　Calculation results of correlation coefficients between thermal load and various environmental factors

环境因素	温度	湿度	风速	空气质量	太阳辐射	气压
ρ	−0.87	−0.53	0.68	0.45	0.26	0.18

整理后实验数据集共包含 672 组样本，每个样本用 7 个维度的参量表示，分别为室外温度、相对湿度、风力等级、空气质量指数以及前一天热负荷、前一小时热负荷和实际热负荷。将以上因素输入神经网络模型进行训练。模型训练前按照一定比例划分训练集、测试集。利用训练集训练模型，使用测试集验证，并结合指标评价模型性能。对于数据驱动方法而言，原始数据的有效是至关重要的。本次训练选取供热负荷序列与 CEEMDAN 的分量和残差，组成输入特征集，建立混合预测模型，其中，前 528 组样本数据为训练集，剩余 144 组为测试集。

1）模型架构和参数

基于 Matlab 平台搭建深度学习模型，选择多层神经网络进行复杂的非线性拟合，为避免过拟合加入正则化层，并增强泛化能力。模型选择两层 CNN 层、三层 LSTM 层，通过 CNN 提取原始数据的特征，利用卷积核与池化层提取时序特征量，后接 LSTM 层提取内部分布特征，最后通过全连接层输出。经过多次优化后确定超参数具体设置。内部架构为：卷积核数 20，大小为 3，隐含层 3 层，隐藏节点数为 128，Dropout 层丢失率为 0.3。

2）模型训练过程

采用 DeST 模拟 2 月 1 日～2 月 28 日的历史负荷和环境参数等，采集逐时的数据，以便于调节热力系统运行。优化器使用 Adam，其特点为简单高效、且占内存少，能够自适应地对学习率不断调整，迭代次数设置为 2 000，学习率为 0.01。

2.4　组合模型训练与分析

1）预测性能验证

构建数据检测和清洗模型的基础，是模型对于输入输出之间良好的拟合能力。模型的预测性能好坏，直接关系到模型异常值辨识以及之后的数据填充效果。利用已训练好的模型对供热负荷进行预测，并与测试集实际数据进行

对比。为了进一步验证本模型的优点和性能,将本文所提组合模型 CEEMDAN-CNN-LSTM 与 LSTM、CNN-LSTM、EEMD-CNN-LSTM 等方法对比,效果如图 8 所示。

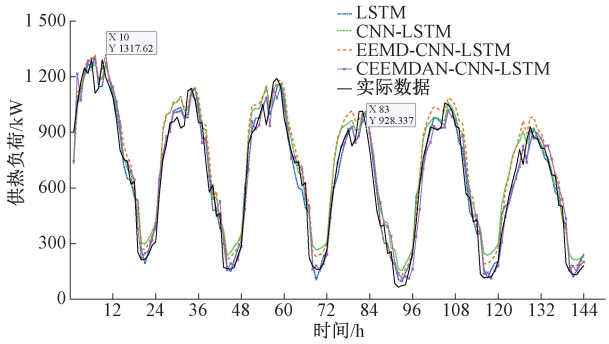

图 8　不同模型对比结果

Fig. 8　Comparison results of different models

2) 评价指标

采用平均绝对误差(MAE)、平均绝对百分比误差(MAPE)和均方根误差(RMSE),如式(18)~(20)。

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | Y_i - y_i | \tag{18}$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{Y_i - y_i}{Y_i}\right| \tag{19}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - y_i)^2} \tag{20}$$

式中:Y_i 为热负荷实际值,y_i 为热负荷预测值,N 为样本总数量。

对比本模型与其他 3 种模型的误差结果,评价指标如表 2 所示。

表 2　预测结果评价指标

Table 2　Prediction result evaluation metrics

评价指标	MAE	RMSE	MAPE/%
LSTM	62.65	81.01	12.50
CNN-LSTM	58.39	62.38	11.55
EEMD-CNN-LSTM	38.65	32.44	8.92
CEEMDAN-CNN-LSTM	17.89	13.14	4.80

对比发现:本模型预测性能较为准确,以平均误差百分比 MAPE 作为损失函数,值为 4.80%,MAE 值为 17.89。3 种指标均表明模型的预测值与实际值的偏差很小,即模型能够有效地学习到输入和输出之间的非线性关系。因此,准确的预测结果,为异常辨识和修正作数据参考。

3　异常检测和清洗效果

3.1　待清洗数据和评价指标

1) 待清洗数据集

调研现场运行情况,结合上述的 672 条数据,以正常历史数据为基础,综合考虑异常现象,包括不符合变化趋势、缺失、无穷数等。人工模拟异常值和正常值的混合数据集,其中前 528 组只包含正常数据,后 144 组为待检测集(异常值占比 20%),辨识并清洗数据集,实际值和异常值具体分布情况和其误差曲线如图 9 所示。

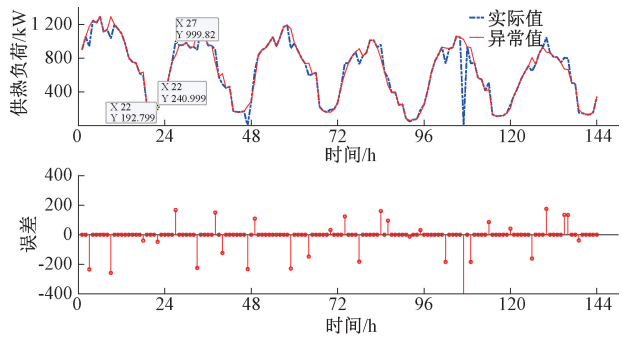

图 9　待清洗数据分布情况

Fig. 9　Distribution of data to be cleaned

2) 异常检测评价

根据上节已训练好的组合模型,计算待清洗数据集中各数据实际值和预测值的偏差大小。依据训练结果选择合适的阈值作为判断依据:偏差大于阈值的判定为异常,反之为正常。模型可以有效地识别出待清洗数据中的缺失、突变等错误值,误判漏判情况极少。为了直观理解模型在异常值检测方面的能力,将异常数据定义为正例,正常数据视为负例,作为异常检测效果评价的基础数据,预测与实际的对应关系如表 3 所示。

表 3　混淆矩阵

Table 3　Confusion matrix

项目	实际异常	实际正常
预测异常	TP	FP
预测正常	FN	TN

采用两种异常值检测评价指标。准确率代表整体预测准确程度,是预测正确的结果占总样本的百分比,但忽视了样本失衡问题,如式(21)所示。因此选取 $F1$ 分数调和,兼顾召回率和精确率。精确率代表预测正常且确为正常的比例;召回率代表预测异常中确为异常的比例,如式(22)所示。

$$\text{Accuracy} = \left(\frac{TP}{TP+FN} + \frac{TN}{FP+TN}\right)/2 \tag{21}$$

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{22}$$

3.2　模型性能验证

为提高数据清洗效率,首要保障异常数据检测的准确性。本文选择模态分解效果更佳的模型,进行真实数据的异常检测和更正。具体清洗过程为:将检测为异常的数据,利用预测模型相应输出值进行替换,提高待清洗数据集的

质量，避免个别偏差较大的数据对系统运行和调控产生错误判断。清洗后及时更新并上传数据至控制终端平台，完成脏数据的剔除和填补工作，为系统准确监测提供数据基础。最终计算得出，TP、FP、FN、TN 值分别为：25、4、4、111。结合准确率和 $F1$ 分数评价异常检测能力，结果如表 4 所示。

表 4　数据检测结果

Table 4　Data detection results

异常评价指标	LSTM	CNN-LSTM	EEMD-CNN-LSTM	CEEMDAN-CNN-LSTM
Accuracy	56.82	73.21	77.83	91.36
$F1$	31.03	37.93	64.28	86.21

综合对比，本文提出 CEEMDAN-CNN-LSTM 组合模型，能够充分拟合输入量与输出复杂关系，对供热系统数据的异常值检测准确率优于其他常见的方法。最终检测结果的准确率高达 91.36%，$F1$ 分数高达 86.21%，既可较好地检测出异常值，又可一定程度上减少将正确值误判为异常的情况。

为了确保数据集的准确和完整，采用此方法纠正异常数据。根据之前对于模型预测性能的验证，其损失函数（平均绝对百分比误差）保持在较小的范围内（低于 5%），表明模型对数据的预测值与实际值之间的差距较小，因此可以采用预测值代替真实值。

4　结　　论

本文基于 DeST 负荷模拟软件，获取某建筑实际供热负荷。通过模态分解方法，将原始负荷序列经过 CEEMDAN 分解得到特征明显的子序列和残余项，显著降低预测难度。基于 Matlab 平台搭建深度学习网络，利用 CNN-LSTM 组合模型，充分拟合输入与输出之间的复杂关系。对比其他单一模型，验证了本模型的准确性，其 MAE、MAPE 值分别为 17.89、4.80%，因此预测结果精度较高。根据偏差大小，对混合数据集进行异常辨识。结合准确率和 $F1$ 分数进行评价：准确率 89.5%，$F1$ 分数 84.2%，模型异常值检测能力准确。利用高精度特点，将异常值替换完成数据清洗。因此，本文组合模型对负荷序列具有良好的预测性能，其变化趋势符合实际分布规律，能够高效完成混合数据集的检测和清洗。

参考文献

[1] 邓伟民. 基于数据驱动的区域供热系统故障诊断研究[D]. 广州：华南理工大学，2021.
DENG W M. Data-driven researches on the fault diagnosis of the district heating system [D]. Guangzhou: South China University of Technology, 2021.

[2] 杨柳林，胡贺骏. 基于改进 GMM 算法的综合能源数据清洗研究[J]. 电子测量技术，2023，46(4)：78-83.
YANG L L, HU H J. Research on comprehensive energy data cleaning based on improved GMM algorithm[J]. Electronic Measurement Technology, 2023, 46(4): 78-83.

[3] 凌继红，邢金城，李昂，等. 基于孤立森林算法的集中供热系统异常数据识别研究[J]. 暖通空调，2023，53(2)：97-102.
LING J H, XING J CH, LI A, et al. Abnormal data detection of central heating system based on isolation forest algorithm [J]. Heating Ventilating & Air Conditioning, 2023, 53(2): 97-102.

[4] WANG Y M, LI Z W, LIU J J, et al. Data-driven analysis and prediction of indoor characteristic temperature in district heating systems[J]. Energy, 2023, 282: 129023.

[5] 马广兴，曲波，常琛，等. 基于 CNN 的供热管道泄漏识别方法研究[J]. 电子测量技术，2022，45(16)：34-41.
MA G X, QU B, CHANG CH, et al. Research on leakage identification method of heating pipeline based on CNN[J]. Electronic Measurement Technology, 2022, 45(16): 34-41.

[6] 王凯，黄丹，梁晓伟，等. 智能电表数据异常在线检测的无监督学习[J]. 电子测量技术，2021，44(8)：125-129.
WANG K, HUANG D, LIANG X W, et al. Unsupervised learning for on-line detection of abnormal data in smart meters[J]. Electronic Measurement Technology, 2021, 44(8): 125-129.

[7] URMENETA J, IZQUIERDO J, LETURIONDO U, et al. A methodology for performance assessment at system level—Identification of operating regimes and anomaly detection in wind turbines[J]. Renewable Energy, 2023, 205: 281-292.

[8] 李泽豪，冯寿廷. 基于模糊 PID 的供暖设备控制系统设计[J]. 电子测量技术，2022，45(23)：7-12.
LI Z H, FENG S Y. Design of heating equipment control system based on fuzzy PID[J]. Electronic Measurement Technology, 2022, 45(23): 7-12.

[9] 梅玉杰，李勇，周王峰，等. 基于机器学习的配电网异常缺失数据动态清洗方法[J]. 电力系统保护与控制，2023，51(7)：158-169.
MEI Y J, LI Y, ZHOU W F, et al. Dynamic data cleaning method of abnormal and missing data in a distribution network based on machine learning[J]. Power System Protection and Control, 2023, 51(7): 158-169.

[10] 吴永斌,张建忠,袁正舸,等. 风电场风功率异常数据识别与清洗研究综述[J]. 电网技术,2023,47(6):2367-2380.
WU Y B, ZHANG J ZH, YUAN ZH X, et al. Review on identification and cleaning of abnormal wind power data for wind farms [J]. Power System Technology, 2023, 47(6): 2367-2380.

[11] 武佳卉,邵振国,杨少华,等. 数据清洗在新能源功率预测中的研究综述和展望[J]. 电气技术,2020,21(11):1-6.
WU J H, SHAO ZH G, YANG SH H, et al. Review and prospect of data cleaning in renewable energy power prediction [J]. Electrical Engineering, 2020, 21(11): 1-6.

[12] 夏延秋,夏和民,冯欣. 一种基于风功率曲线的 SCADA 数据清洗方法研究[J]. 可再生能源,2022,40(11):1499-1504.
XIA Y Q, XIA H M, FENG X. Research on SCADA data cleaning method based on wind power curve[J]. Renewable Energy Resources, 2022, 40 (11): 1499-1504.

[13] NEUMAYER M, STECHER D, GRIMM S, et al. Fault and anomaly detection in district heating substations: A survey on methodology and data sets[J]. Energy, 2023, 276: 127569.

[14] GONG M J, ZHAO Y, SUN J W, et al. Load forecasting of district heating system based on Informer[J]. Energy, 2022, 253: 124179.

[15] 夏博,李春杨,万露露,等. 基于深度学习的风力发电机组故障预警方法研究综述[J]. 科学技术与工程,2023,23(9):3577-3587.
XIA B, LI CH Y, WAN L L, et al. Review of wind turbine fault warning methods based on deep learning[J]. Science Technology and Engineering, 2023, 23(9): 3577-3587.

[16] SUN C H, ZHANG H X, CAO S S, et al. A hierarchical classifying and two-step training strategy for detection and diagnosis of anormal temperature in district heating system[J]. Applied Energy, 2023, 349: 121731.

[17] 张童. 时序数据分析在供热数据异常监测中的研究和应用[D]. 天津:天津科技大学,2019.
ZHANG T. The research and application on anomaly monitoring of the heat addition by the data analysis of time series [D]. Tianjing: Tianjing University of Science & Technology, 2019.

[18] 柳亚楠. 供需不匹配工况源网平衡调控策略研究[D]. 天津:河北工业大学,2022.
LIU Y N. Study on balance control strategy of source-network under condition of mismatch between supply and demand [D]. Tianjing: Hebei University of Technology, 2022.

[19] 尹薇. 时间序列清洗关键技术的研究[D]. 哈尔滨:哈尔滨工业大学,2018.
YIN W. Research on key technology of tmme series cleaning[D]. Harbin: Harbin Institute of Technology, 2018.

[20] 匡俊搴,赵畅,杨柳,等. 一种基于深度学习的异常数据清洗算法[J]. 电子与信息学报,2022,44(2):507-513.
KUANG J Q, ZHAO CH, YANG L, et al. An outlier cleaning algorithm based on deep learning[J]. Journal of Electronics & Information Technology, 2022, 44(2): 507-513.

[21] 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 X G, WANG ZH X, WEANG H J. Turbine data cleaning based on deep LSTM [J]. Thermal Power Generation, 2023, 52(8): 179-187.

作者简介

梁晓龙,工程师,主要研究方向为信息化与智慧供热。
E-mail:12876150@ceic.com

孟现阳(通信作者),博士,教授,主要研究方向为综合能源系统控制与优化。
E-mail:xymeng@mail.xjtu.edu.cn