DOI: 10.19650/j.cnki.cjsi.J2209802

# 基于缺失数据填补的风电齿轮箱状态监测研究*

徐 健,刘长良,王梓齐,赵陆阳
(华北电力大学自动化系 保定 071003)

**摘 要**:风电机组监控和数据采集系统的现场数据普遍存在缺失问题,会对下游状态监测任务产生一定负面影响。为此,提出一种结合注意力机制的掩膜自编码网络,用于填补面板数据样本中的缺失值,增加可用样本数量,提升状态监测结果的准确性与连续性。该方法以降噪自编码网络为整体框架,在编码阶段通过注意力机制对缺失值进行掩膜处理,赋予缺失值更高的权重以强化网络对其关注程度,在解码阶段将缺失值填补后输出完备数据样本。随后,利用长短时记忆网络提取的样本特征对目标变量参数进行预测,依据预测残差实现状态监测。使用某风电齿轮箱运行数据验证,结果表明:提出方法的数据填补偏差相较对比方法至少改善 17.2%;与数据填补前相比,数据填补后样本数量显著增加,使状态监测网络对正常数据的预测残差平均下降 37.4%,对故障数据的检测率提升 6.8%。

**关键词**:缺失数据填补;自编码网络;注意力机制;风电机组;状态监测
**中图分类号**:TH17 TM315 **文献标识码**:A **国家标准学科分类代码**:460

# Research on condition monitoring of wind turbine gearbox based on missing data imputation

Xu Jian, Liu Changliang, Wang Ziqi, Zhao Luyang
(*Department of Automation, North China Electric Power University, Baoding 071003, China*)

**Abstract**: The field data of wind turbine data acquisition and supervisory control system are commonly missing, which have a certain negative influence on the downstream condition monitoring task. To address this issue, a mask autoencoder network with attention mechanism is proposed to impute missing values in panel data samples, increase the number of available samples, and improve the accuracy and continuity of condition monitoring results. The method takes the denoising autoencoder network as the overall framework. In the encoding stage, the missing values are masked by the attention mechanism, and the missing values are given a higher weight to strengthen the attention of the network. In the decoding stage, the complete data samples are output after missing values imputation. Then, the parameter of the target variable is predicted by using the sample features extracted by long short-term memory network, and the condition monitoring is realized according to the prediction residual. This method is evaluated by the operation data of a wind turbine gearbox. Results show that the data imputation bias of the proposed method is at least 17.2% better than that of the comparison method. Compared with before data imputation, the number of samples increased significantly after data imputation, which makes the prediction residual of normal data decreased by 37.4% on average and the detection rate of fault data increased by 6.8%.

**Keywords**: missing data imputation; autoencoder network; attention mechanism; wind turbine; condition monitoring

## 0 引 言

伴随着风电装机容量的不断提升,机组运维问题日益显著。据统计,风电机组的运维成本约占总成本的 5%~23%[1],严重制约风电产业经济效益。齿轮箱作为风电机组传动链的重要设备,维修成本高,停机维护时间长[2],有必要对其进行实时状态监测。近期,基于风电机组监控和数据采集系统(data acquisition and supervisory control, SCADA)数据和正常行为建模方法的状态监测研

收稿日期:2022-05-17 Received Date: 2022-05-17
*基金项目:北京市自然科学基金(4182061)项目资助

究受到广泛关注。文献[3-4]使用神经网络和非参数估计方法对 SCADA 截面数据进行分析,实现了对风电设备的状态监测,但由于未考虑数据的时序性,特征提取不全面。借助滑动窗口法,文献[5-6]将 SCADA 时间序列数据切分为固定时间长度的面板数据样本,随后分别使用长短时记忆网络、时空注意力联合 GRU 网络充分提取样本的时间特征,根据齿轮箱状态变量预测值与真实值的残差实现状态监测,效果优于未考虑数据时序性的方法。

然而现场 SCADA 数据中普遍存在缺失值,造成切分后的面板样本中有数据不齐全的欠完备样本。许多研究选择忽视或删除欠完备样本[7-8],这将导致样本的数量减少、时间连续性差,对下游状态监测任务造成一定负面影响。因此,有必要在数据预处理阶段填补欠完备样本中的缺失值,提升样本数量和信息密度,进而改善下游状态监测任务的结果。

填补欠完备样本的方式有以下两种:1)对原始 SCADA 时间序列数据进行整体填补,但实现大规模数据填补较为困难,且容易产生误差累积效应;2)在数据切分后对欠完备样本进行填补,该方法操作简便,每个样本的填补结果相对独立,适合工程应用与在线状态监测。欠完备样本中,面板数据的缺失情况大致可以分为 4 类:离散缺失、块状缺失、同一时刻数据全部缺失即行缺失、同一变量数据全部缺失即列缺失[9],最常见的情况是数据行缺失。文献[10]采用局部均值替换法填补行缺失数据,以保证状态监测趋势的连续性,该方法较为简单,数据填补精度有较大提升空间。文献[11-12]分别提出数据分类重建和惰性时空系数方法,使用临近风机数据对目标风机行缺失数据进行填补。但是,当临近风机过远或临近风机数据也存在缺失时,上述方法不能完成数据填补任务。

当前对风电机组 SCADA 数据行缺失样本进行填补的研究相对较少,但在其他领域存在多种数据填补方法,是大数据处理的研究热点之一。基于统计学的传统数据填补方法包括均值填充、回归填充、聚类填充、多重填充等方法[13]。这类方法适合处理变量维度低、缺失比例小的数据集。当面对更复杂的情况时,基于深度学习的数据填补方法更具优势。降噪自编码网络具有良好的数据去噪能力,文献[14]将缺失数据视作一种特殊噪声,提出基于降噪自编码网络的多重填补模型,在多个公开数据集上的数据填充效果优于传统方法。注意力机制通过计算注意力权重来度量不同信息的重要性,近年来在多个领域得到广泛应用[15]。文献[16]将注意力机制引入降噪自编码网络,在大比例缺失数据集上取得了良好的填补效果。文献[13-16]的研究对象大多为人工构造的欠完备数据集,重点在于解决数据中部分特征存在缺失的问题,未考虑数据的时序性,不能直接用于风电 SCADA 数据缺失处理。

针对状态监测实际需求,提出一种结合注意力机制的掩膜自编码(masked autoencoder with attention mechanism, MAE-AM)网络,通过增加网络对缺失填补任务的关注程度来提升数据填补效果,在预处理阶段使用单台风机 SCADA 数据解决了面板样本行缺失问题。随后,基于长短时记忆网络的预测残差实现齿轮箱状态监测。以某风场 SCADA 实际数据为例,验证了提出方法的数据填补能力,并将数据填补前后齿轮箱在正常和异常运行状态下的监测结果进行对比。

## 1　面板样本行缺失问题描述

SCADA 系统位于风电场控制中心内,按照固定时间间隔采集反映机组状态的参数,形成多变量时间序列数据集。受自然环境、设备可靠性等因素影响,现场 SCADA 数据集存在大量缺失值;此外,数据集中还有不能被状态监测任务使用的异常值。预处理阶段,会将上述“脏数据”删除,具体情况如下[17]:

1)记录缺失数据。由于正常停机、传感器故障、信息传输异常等原因,数据集中存在标记为“Nan”的缺失值,通常成行出现。

2)非正常运行工况数据。风电机组运行工况受风速波动影响大,实际风速小于切入风速时无法驱动风机正常运行,大于切出风速时机组顺桨停机。当记录风速位于切入切出风速区间外时,对应时刻的整行数据被视为非正常运行工况数据。

3)离群数据。由于弃风限电、传感器异常、测量出现粗大误差等原因,风机正常运行工况下存在部分离群数据。检测离群数据的方法有:拉依达法[18]、四分位法[19]、孤立森林法[20]等。

经过清洗,SCADA 数据集中有许多整行缺失的数据。直接采用滑动窗口法在欠完备数据集上切分样本,滑窗宽度需根据实际情况确定。风电机组状态监测可被视为一个自监督学习问题,每个样本应当包括面板特征数据和目标数据两部分,通过提取面板数据特征对目标数据中的部分值进行预测。设置窗宽为 $L$,步距为 1,沿时间轴使用滑动窗口将 SCADA 数据切分为若干个样本 $\boldsymbol{S}_i$,每个样本均是大小为[$L$, $M$]的数据矩阵,可表示为 $\boldsymbol{S}_i=[x_i, x_{i+1}, \cdots, x_{i+L-2}, y_i]$。以 $L=4$ 为例,SCADA 数据的样本切分过程如图 1 所示。

如图 1 右侧所示,根据特征数据和目标数据的缺失情况,将样本分成完备、目标缺失和欠完备 3 类。完备样本中所有数据均无缺失;目标缺失样本特征数据不一定缺失但目标数据一定缺失;欠完备样本中目标数据完整但面板特征数据部分行存在缺失。

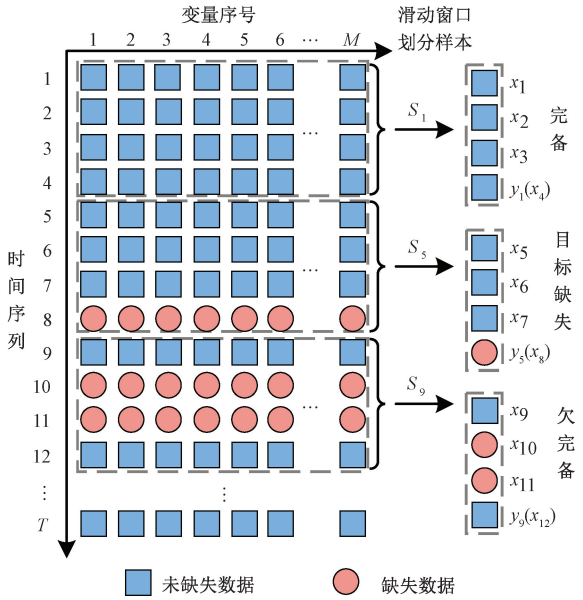


图 1 面板数据样本切分过程

Fig. 1 The process of dividing panel data samples

在状态监测问题中,完备样本适合用于研究;目标缺失样本无法描述机组真实运行状态,不能被使用;欠完备样本的特征数据不完整,以往的研究大多直接将其删去,但小比例缺失的特征数据中仍含有大量真实特征信息,具有潜在价值。若将小比例缺失欠完备样本填补后使用,即可在完备样本的基础上提升状态监测趋势的连续性和同一时间段内的信息密度。此外,数据填补的精度会直接影响状态监测效果,填补的数据要尽可能还原设备的真实状态。

研究一种高精度填补小比例缺失欠完备样本的方法既是工程应用的需要,也是对当前技术水平的挑战,考虑使用先进的深度学习方法解决上述问题。

## 2 提出方法

考虑缺失数据的状态监测方法整体框架包括离线与在线两个阶段。离线阶段使用 SCADA 历史数据训练提出的 MAE-AM 数据填补网络和长短时记忆状态监测网络;在线阶段 SCADA 实时数据经处理后,对原始完备样本和填补生成的完备样本进行状态监测。下面分别对离线、在线阶段进行详细介绍。

### 2.1 离线阶段

1)数据准备

从风电机组 SCADA 数据库中导出历史数据,送入数据预处理单元,依次按照数据清洗、变量选择、min-max 归一化、样本切分 4 个步骤处理原始数据。数据清洗和样本切分过程已在上节中介绍;变量选择过程主要依据工程经验和皮尔逊相关系数[21];对数据进行 min-max 归一化是为消除数据量纲的影响、加快模型训练速度,公式如下:

$$\hat{x}j_i = \frac{x_i^j - x_{\min}^j}{x_{\max}^j - x_{\min}^j} \tag{1}$$

式中:$x_i^j$ 为变量 $j$ 的第 $i$ 个数据;$x_{\min}^j$ 和 $x_{\max}^j$ 为变量 $j$ 的最小值和最大值;$\hat{x}_i^j$ 为归一化后的数据。

预处理后,将目标缺失样本删除,得到若干完备样本和欠完备样本。基于完备样本生成填补网络训练数据,规则如下:首先,分析欠完备样本行缺失情况,确定待填补欠完备样本的最大行缺失数 $R$;随后,剔除完备样本中的目标数据,依次使剩余特征数据随机缺失 1~$R$ 行;重复上一步流程,生成更多训练数据。最终,将生成的行缺失数据作为输入,其对应的完备数据作为目标,训练数据填补网络。

2)MAE-AM 数据填补网络

提出 MAE-AM 网络以降噪自编码网络(denoising autoencoder, DAE)为整体框架,在编码阶段引入注意力机制,依据数据缺失情况计算动态注意力权重,并将其作为掩膜值与数据特征点乘,在解码阶段用样本中的非缺失值替换输出数据,使网络在解码时更关注缺失填补任务。下面分别介绍 DAE 网络、注意力机制和提出的 MAE-AM 网络。

如图 2 所示,DAE 网络是自编码网络的一种变体,通过从加噪数据中恢复原始输入的方式,增强网络消除数据噪声的能力[22]。加噪方法主要有以下两种:在原始数据上添加高斯噪声等函数形式的噪声;随机丢失部分数据,破坏原始输入数据的完整性。DAE 网络的目标是最小化重构数据与输入数据之间的残差,重构误差可用公式表示为:

$$RE = (x_i - \hat{x}_i)^2 \tag{2}$$

式中:$x_i$ 为原始输入数据,$\hat{x}_i$ 为对应的重构数据。

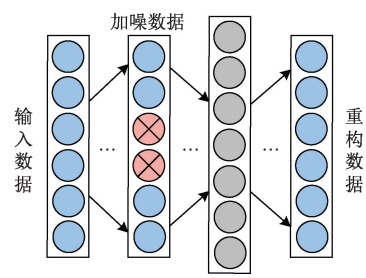


图 2 降噪自编码网络结构

Fig. 2 The structure of the DAE network

图 3 为注意力机制原理图,输入数据通过全连接 Dense 层计算注意力权重,经 Softmax 函数缩放注意力权重至[0, 1]之间,随后将输入数据与注意力权重点乘得

到输出数据[23]。注意力机制能够区分不同信息的重要程度,提升关键信息的权重值。设输入数据为 $\boldsymbol{x}_i$,每组输入均包含 $k$ 个变量,注意力机制计算过程如下。

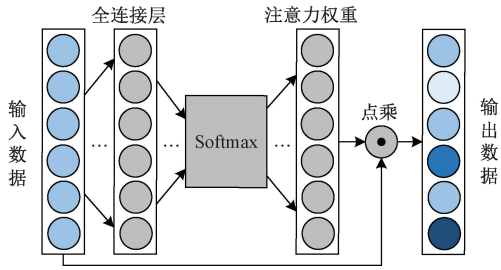


图 3　注意力机制原理

Fig. 3　The principle of attention mechanism

$$\boldsymbol{e}_i = \mathrm{Relu}(\boldsymbol{w}\boldsymbol{x}_i + b) \tag{3}$$

$$a_{i,q} = \mathrm{Softmax}(\boldsymbol{e}_{i,q}) = \frac{\exp(\boldsymbol{e}_{i,q})}{\sum_{q=1}^{k}\exp(\boldsymbol{e}_{i,q})} \quad \sum_{q=1}^{k} a_{i,q} = 1 \tag{4}$$

$$\boldsymbol{A}_i = [a_{i,1}, a_{i,2}, \cdots, a_{i,k}] \tag{5}$$

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{A}_i \odot \boldsymbol{x}_i \tag{6}$$

式中:$\boldsymbol{e}_i$ 为经全连接层计算的注意力权重向量;$\boldsymbol{w}$、$b$ 分别为全连接层的权重和偏置;Relu( ) 为修正线性单元激活函数,用于增加注意力权重之间的差异;$a_{i,q}$ 为 Softmax 函数计算的第 $i$ 组数据第 $q$ 个变量的注意力权重;$A_i$ 为第 $i$ 组数据对应的注意力权重向量;⊙ 为点乘运算符;$\tilde{\boldsymbol{x}}_i$ 为输出数据。

结合 DAE 网络的数据重构特点和注意力机制通过计算权重来区分信息重要性的方式,提出了 MAE-AM 数据填补网络。提出网络以 DAE 网络为主体框架,其任务是将行缺失数据填补成完备数据。训练阶段,行缺失数据由完备数据加噪生成,加噪方式为随机丢失 1~$R$ 行数据。MAE-AM 网络主要包括特征编码、注意力编码、解码重构 3 个部分。

MAE-AM 网络结构如图 4 所示,下面分别对网络的 3 个部分进行介绍:

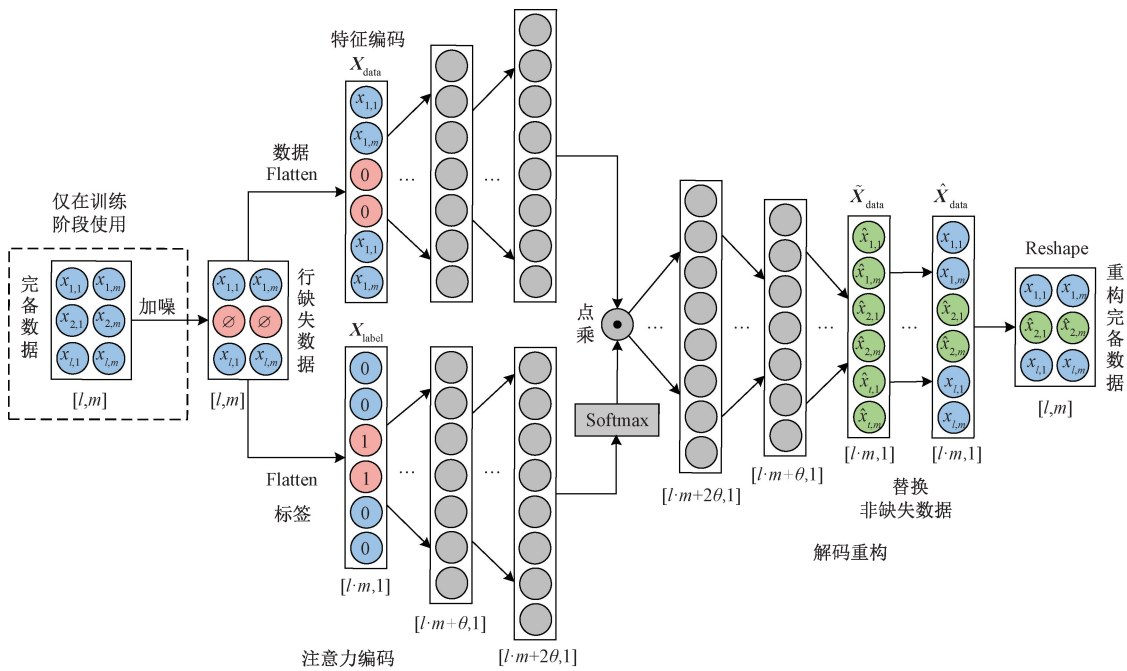


图 4　用于缺失数据填补的 MAE-AM 网络结构

Fig. 4　The network structure of MAE-AM for missing data imputation

(1)特征编码:通过 Flatten 层,将行缺失数据矩阵按行依次拼接为数据向量 $\boldsymbol{X}_{\mathrm{data}}$,并用“0”替换缺失数据“$\varnothing$”;使用两个全连接层提取数据特征,全连接层的神经元个数在 Flatten 层输出神经元个数的基础上,分别增加 $\theta$ 和 $2\theta$ 个。

(2)注意力编码:根据数据缺失情况,将已知数据标记为“0”,标记缺失数据标记为“1”,生成对应的标签矩阵,再经 Flatten 层得到标签向量 $\boldsymbol{X}_{\mathrm{label}}$;采用与特征编码网络结构相同的两个全连接层计算注意力权重,并用 Softmax 函数进行缩放。通过“1/0”标记数据是否缺失的方式,可以提高缺失数据对应权重,使网络在解码时更关注数据填补任务。

(3)解码重构:将数据特征和注意力权重点乘后,使用与编码网络结构对称的两个全连接层,计算求得解码

数据向量 $\tilde{\boldsymbol{X}}_{\text{data}}$；随后用原始数据替换非缺失数据，保留缺失数据的解码值，确保缺失填补过程中不改变非缺失数据的真实值，替换后的数据向量 $\hat{\boldsymbol{X}}_{\text{data}}$ 可由式(7)计算得到；最后经 Reshape 层将数据向量还原为与输入数据维度相同的数据矩阵。

$$\hat{\boldsymbol{X}}_{\text{data}} = \boldsymbol{X}_{\text{data}} + \tilde{\boldsymbol{X}}_{\text{data}} \odot \boldsymbol{X}_{\text{label}} \tag{7}$$

式中：$\boldsymbol{X}_{\text{data}}$ 为原始数据向量；$\boldsymbol{X}_{\text{label}}$ 为标签向量；$\odot$ 为点乘运算符。

搭建的 MAE-AM 网络在训练时，缺失数据的真实值已知，因此可通过最小化重构误差的方式迭代优化网络。此外，为评价网络的数据填补效果，引入均方根误差(root mean squared error，RMSE)、平均绝对百分比误差(mean absolute percentage error，MAPE)两个指标，计算公式如下：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2} \tag{8}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{x}_i - x_i}{x_i}\right| \tag{9}$$

式中：$n$ 为样本数量，$x_i$ 为样本的真实值，$\hat{x}_i$ 为填补网络的输出值。

RMSE 用于计算真实值与填补网络输出值之间的偏差，对数据间较大的偏差更为敏感；MAPE 计算得出真实值与填补网络输出值偏差的绝对百分比，直观体现填补值的偏差比例。RMSE 和 MAPE 均越小越好。

通过上述指标，可以综合评价数据填补模型的效果，选出最佳模型填补欠完备样本中的缺失数据。随后，将初始完备样本和填补后得到的完备样本作为训练数据，训练状态监测网络。

3)长短时记忆状态监测网络

参照文献[5]，使用长短时记忆(long short-term memory，LSTM)网络对风电齿轮箱进行状态监测，网络结构如图 5 所示。网络输入特征数据为 $t$ 时刻及 $t$ 时刻前一段时间的齿轮箱相关状态变量参数；预测目标数据为 $t$+1 时刻的齿轮箱油温参数。

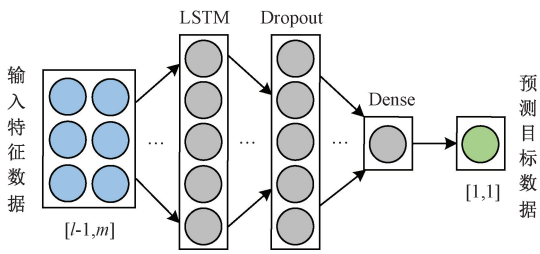


图 5 LSTM 状态监测网络结构

Fig. 5 The structure of LSTM condition monitoring network

图 5 中，LSTM 为包含 16 个记忆单元的单层网络，使用 *Sigmoid* 激活函数；为防止模型过拟合，LSTM 层后接 Dropout 层，参数 rate 设置为 0.2；Dense 层神经元个数为 1，不使用激活函数，输出即为齿轮箱油温预测值。离线阶段，通过对目标数据预测值和真实值的残差进行统计分析，确定状态监测预警阈值，以备在线阶段使用。预测残差为预测值和真实值之间的 RMSE。

受风速影响，风电机组运行状态具有很大随机性，因此将残差按照指数加权滑动平均(exponentially weighted moving-average，EWMA)控制图平滑处理，公式如下[24]：

$$E_t = \lambda E_{t-1} + (1-\lambda)e_t \tag{10}$$

式中：$E_t$ 为 $t$ 时刻 EWMA 控制图的输出，$e_t$ 为 $t$ 时刻 RMSE 残差，加权值 $\lambda$ 取 0.3，$\lambda \in (0,1]$。

EWMA 控制图的上限为：

$$UCL = \mu + K\sigma\sqrt{\frac{\lambda[1-(1-\lambda)^{2t}]}{2-\lambda}} \tag{11}$$

式中：$\mu$、$\sigma$ 为 RMSE 残差的均值和标准差，$K$ 为 $UCL$ 系数，$\lambda$、$t$ 与式(10)一致。根据风电机组实际运行效果，$K$ 取 10，$\lambda$ 取值同上。离线阶段，$UCL$ 随时刻 $t$ 实时变化，将最终得到 $UCL$ 值视为在线阶段的固定预警阈值。

### 2.2 在线阶段

SCADA 实时数据经预处理后，将完备样本和经 MAE-AM 网络填补后的欠完备样本均视为状态监测数据，按照时序依次送入离线阶段训练好的 LSTM 状态监测网络，计算状态变量 RMSE 残差，并使用 EWMA 控制图平滑处理。为确保预警结果真实可靠，当控制图输出连续 10 次超过 $UCL$ 时，发出异常报警信号，提醒运维人员及时关注设备潜在故障。

## 3 案例分析

本文研究对象为河北某风场一台额定功率 1.5 MW 的风电机组，切入切出风速分别为 3 和 25 m/s，SCADA 系统采样间隔为 1 min。运行记录显示，2017 年 11 月 17 日 8:31，齿轮箱发生故障导致机组停运，故障原因是齿轮箱油池温度高于上限值 70℃。从 SCADA 系统历史数据库中导出 2017 年 8 月 1 日 0:00~11 月 17 日 8:31 的风电机组数据，用于缺失填补及状态监测研究。

### 3.1 数据预处理与样本集划分

按照上节所述数据清洗、变量选择、min-max 归一化、样本切分 4 个步骤处理原始数据。数据清洗时，离群数据采用孤立森林法检测，离群值设置为 3%；根据齿轮箱状态监测需要，选择了 7 个状态变量用于研究，如表 1 所示。min-max 归一化后，状态变量随时间序列变化的情况如图 6 所示。

图 6 中的连续空白区域是由于风电机组按计划停机，SCADA 系统未采集数据造成。去除连续空白区

**表 1　齿轮箱相关状态变量**

**Table 1　Gearbox related state variables**

| 变量名称/单位 | 符号 | 取值范围 |
|---|---|---|
| 齿轮箱油池温度/℃ | V1 | [49.98, 70.23] |
| 齿轮箱入口油压/Bar | V2 | [2.56, 3.53] |
| 齿轮箱滤网前油压/Bar | V3 | [3.89, 6.03] |
| 齿轮箱驱动端轴承温度/℃ | V4 | [48.49, 78.8] |
| 齿轮箱非驱动端轴承温度/℃ | V5 | [48.28,74.09] |
| 发电机转速/(r·min$^{-1}$) | V6 | [1065.7, 1924.6] |
| 有功功率/kW | V7 | [0.03, 1549.7] |

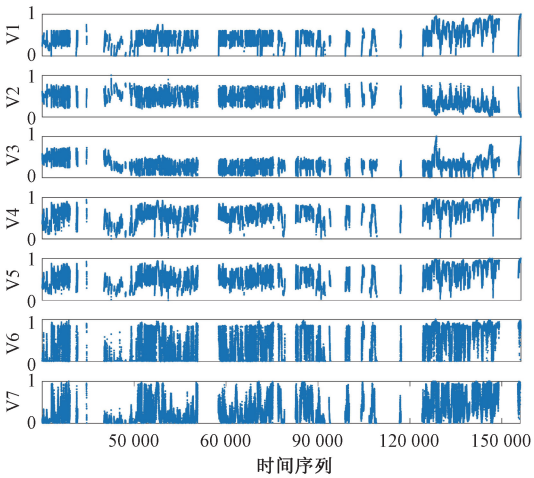


图 6　齿轮箱相关状态变量时序图

Fig. 6　Time series of gearbox related state variables

域后,进行样本切分,计划使用前 10 min 的特征数据对目标数据进行预测,因此设置滑窗宽度为 11。切分后,共得到 121 008 组样本,其中欠完备样本均为行缺失类型,具体情况如下图所示。

如图 7 所示,实例中缺失 3 行及以下的欠完备样本占比高达 76.1%,将其视为待填补的小比例缺失样本,填补后可在完备样本的基础上增加 18.2%的样本数量。随后,以第 30 000 组完备样本对应的时间序号为界将样本划分为离线、在线阶段,如表 2 所示。

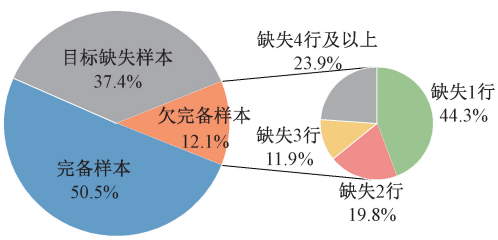


图 7　实例数据样本分类情况

Fig. 7　Classification of the instance data samples

**表 2　样本集的划分与作用**

**Table 2　Division and function of sample set**

| 划分情况 | 样本类型 | 样本数量 | 对 MAE-AM 网络的作用 | 对 LSTM 网络的作用 |
|---|---|---|---|---|
| 离线阶段 | 完备 | 25 000 | 训练 | 训练 |
| | 完备 | 5 000 | 效果评价 | |
| | 待填补欠完备 | 6 045 | 测试 | |
| 在线阶段 | 完备 | 31 104 | - | 测试 |
| | 待填补欠完备 | 5 113 | 测试 | |

离线阶段使用完备样本生成行缺失欠完备样本,用于 MAE-AM 网络的训练及效果评价。待填补欠完备样本在补全后也被纳入 LSTM 状态监测网络的训练集和测试集中。

### 3.2　MAE-AM 网络数据填补效果验证

基于 TensorFlow 深度学习框架,搭建了提出的 MAE-AM 数据填补网络,网络输入为[10, 7]的特征数据矩阵。通过多次试验确定 $\theta = 20$,特征编码和注意力编码网络神经元个数分别为 70-90-110,解码重构网络神经元个数分别为 110-90-70。除解码网络最后一层未使用激活函数外,其余全连接层均使用 *Relu* 激活函数。

网络训练时,选择均方损失函数 MSE,并用学习率为 0.003 的 Adam 算法优化梯度下降过程。输入数据按 70%~30%分割为训练集和验证集,shuffle 设置为 True,打乱训练数据的顺序。采用批处理方法加速训练,batch size = 16。训练周期 epochs = 200,并结合 Early Stopping 策略,若连续 10 个周期验证集损失未降低则停止训练,防止网络过拟合。设置随机种子为 42,选用 TensorFlow 中默认的 Glorot uniform 参数随机初始化方法。

随后,根据待填补样本的缺失情况,生成 MAE-AM 网络的训练和效果评价数据。从离线阶段完备样本中将目标数据剔除,使用剩余特征数据加噪生成行缺失数据,加噪方式为使每组数据随机缺失 1~3 行,最终得到与完备样本数量相同的行缺失数据集。为增强网络的泛化性和鲁棒性,重复上述流程,成倍扩充行缺失数据集。然而,随机行缺失法可能会导致数据集中存在完全一致的样本,浪费计算资源。借助排列组合原理,在每组数据均完全不同的前提下,可以快速生成最多的行缺失数据,组合种类 $P$ 的计算公式如下:

$$P = \sum_{r=1}^{R} C_l^r = \sum_{r=1}^{R} \frac{l!}{r!\ (l-r)!} \tag{12}$$

式中:$l$ 为数据行数,$R$ 为最大行缺失数,$r$ 为行缺失数。实例中 $l$=10,$R$ 分别为 1、2、3 时,$P$ 为 10、55、175。

设置数据的行缺失类型分别为缺失 1 行、缺失 1~2 行、缺失 1~3 行 3 种;以离线阶段完备样本的数量为基

准,采用随机法生成 1 倍、3 倍、5 倍、10 倍容量,组合法生成 $P$ 倍容量的行缺失数据集。MAE-AM 网络的数据填补效果对比情况如表 3 所示。

**表 3 MAE-AM 网络数据填补效果**

**Table 3 Data imputation effectiveness of MAE-AM network**

| 行缺失类型 | 数据集容量/倍 | RMSE 指标 | MAPE 指标/% |
|---|---|---|---|
| 缺失 1 行 | 1 | 0.045 9 | 13.77 |
| | 3 | 0.040 4 | 11.99 |
| | 5 | 0.039 5 | 11.00 |
| | 10 | 0.038 0 | 10.71 |
| | P | **0.034 1** | **8.83** |
| 缺失 1~2 行 | 1 | 0.050 8 | 15.02 |
| | 3 | 0.045 5 | 13.89 |
| | 5 | 0.042 2 | 12.84 |
| | 10 | 0.040 8 | 11.83 |
| | P | **0.036 5** | **9.08** |
| 缺失 1~3 行 | 1 | 0.056 4 | 16.04 |
| | 3 | 0.047 8 | 14.98 |
| | 5 | 0.044 8 | 13.30 |
| | 10 | 0.043 6 | 12.45 |
| | P | **0.040 3** | **9.77** |

如表 3 所示,随着行缺失数据集容量的增加,MAE-AM 网络的数据填补效果有所提升,泛化性和鲁棒性得到增强。当数据行缺失类型变得复杂时,数据填补难度增加,网络填补效果下降。在确定最大行缺失数时,应当综合考虑网络填补效果与数据真实缺失情况。此外,在行缺失类型为缺失 1 行时,对比随机法和组合法生成的 10 倍容量数据集下的缺失填补效果,可以发现在数据集容量相同的情况下,组合法生成的行缺失数据集更具优势。

为进一步验证提出方法的数据填补效果,引入局部均值填补法[6](LMI)、基于链式方程的多重填补法[25](MICE)、基于降噪自编码的多重填补法[14](MIDA)、带掩膜注意力的降噪自编码网络[16](DAEMA)4 种方法进行对比试验。上述方法的参数设置情况均与相应参考文献相同。以缺失 1~3 行的类型为例,采用组合法生成行缺失数据集,数据填补效果对比情况如表 4 所示。

由表 4 可知,提出 MAE-AM 网络的 RMSE、MAPE 指标相较对比方法至少改善了 17.2% 和 29.7%,数据填补精度得到大幅提升。LMI 方法通过局部均值替换填补缺失数据,未考虑数据间的复杂关系;MICE 方法根据当前数据的缺失情况进行多重填补,没有结合数据的历史信息;MIDA 方法经 Dropout 层使输入数据按固定比例随机缺失,与真实数据的行缺失情况不符;DAEMA 网络在数据填补时同样使用了注意力机制,但其数据形式、网络结构、训练方式及应用场景均与 MAE-AM 网络不同。DAEMA 网络的输入为部分特征缺失的数据向量,编码和解码网络均是神经元个数单一的多层特征向量,训练过程中使用 Dropout 层生成伪缺失数据,降低了填补过程中非缺失数据信息量,依据伪缺失数据的填补效果决定模型是否停止训练,旨在解决统计过程中出现的部分特征随机缺失的问题。通过对上述方法进行改进与融合,并考虑现场 SCADA 数据的实际情况,提出的 MAE-AM 网络取得了更好的数据填补效果。

**表 4 不同方法的数据填补效果对比**

**Table 4 Comparison of data imputation effectiveness by different methods**

| 方法 | RMSE 指标 | MAPE 指标/% |
|---|---|---|
| LMI | 0.079 0 | 20.36 |
| MICE | 0.051 6 | 15.97 |
| MIDA | 0.131 5 | 22.01 |
| DAEMA | 0.048 7 | 13.89 |
| MAE-AM | **0.040 3** | **9.77** |

### 3.3 基于缺失数据填补的齿轮箱状态监测

本文使用的 LSTM 状态监测网络同样基于 TensorFlow 深度学习框架。训练过程中,训练周期 epochs 为 50,损失函数、数据分割比例、批处理方式、参数初始化方法均与 MAE-AM 网络相同。

下面,基于 LSTM 网络,对数据填补前后齿轮箱的状态监测结果进行分析。由于目标缺失样本无法计算得出预测残差,缺失 4 行及以上样本训练成本高、填补效果差,不利于下游状态监测任务。因此,实验所用的欠完备样本最大行缺失数为 3。按照离线、在线两个阶段是否使用填补生成的完备样本,分别进行了 3 次状态监测试验,结果如图 8 所示。

由图 8 可知,左侧是重构误差散点图,样本不可用时重构误差为空,不在图中展示;右侧是状态监测结果,带状虚线区域表示因数据缺失造成无法判断齿轮箱运行状态,带状实线区域表示齿轮箱状态异常。以样本序号 17 500 为界,将齿轮箱状态分为正常段和异常段。图 9(a)中,EWMA 控制图的上限 $UCL=0.204\ 3$,由于数据缺失问题导致无法判断部分时刻的齿轮箱状态;图 9(b)中,在离线阶段增加填补生成的完备样本用于 LSTM 网络训练,$UCL$ 降至 0.151 4 且无误报警情况出现,表明训练样本增加后,状态监测网络的学习能力得到提升;图 9(c)中可见,在线阶段增加填补生成的完备样本后,重构误差散点更加丰富,状态监测趋势的连续性得

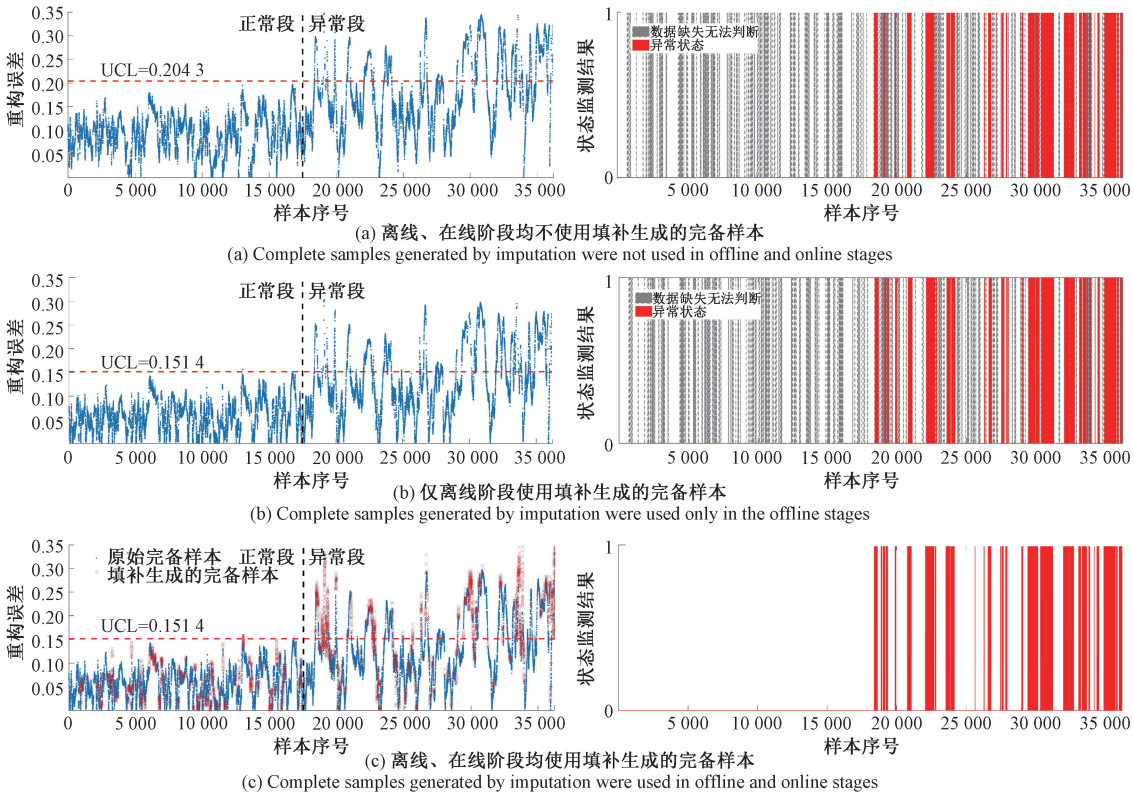


图 8　不同样本集下的风电机组齿轮箱状态监测试验

Fig. 8　Tests of wind turbine gearbox condition monitoring in different sample sets

到增强,同一时间段内的信息密度得到提升,在齿轮箱状态异常段尤为明显。此外,填补生成的完备样本与原始完备样本的重构误差在同一区间内,表明 MAE-AM 网络的数据填补精度满足状态监测需要,达到了预期目标。

为进一步对比不同样本集下齿轮箱的状态监测结果,对正常段和异常段的试验结果进行定量分析,如表 5 所示。其中,异常段的故障检测率为报警次数除以样本数量。

**表 5　齿轮箱状态监测结果定量分析**

**Table 5　Comparison of wind turbine gearbox condition monitoring results**

| 填补生成完备样本的使用阶段 | 正常段 | | 异常段 | | |
|---|---|---|---|---|---|
| | 样本数量 | 重构误差均值 | 样本数量 | 报警次数 | 故障检测率/% |
| 不使用 | 15 530 | 0.092 6 | 15 574 | 4 657 | 29.90 |
| 离线 | 15 530 | 0.057 9 | 15 574 | 5 261 | 33.78 |
| 离线、在线 | 17 500 | 0.057 8 | 18 717 | 6 880 | 36.76 |

如表 5 所示,离线阶段使用填补生成的完备样本时,正常段的重构误差均值降低了 37.4%,异常段报警次数和故障检测率提高,说明样本增加后状态监测网络学习能力得到增强;离线、在线阶段均引入填补生成的完备样本时,正常段重构误差均值保持稳定,异常段故障检测率最终提升至 36.76%,共计提升 6.8%。综上所述,通过 MAE-AM 网络填补小比例缺失欠完备样本扩充数据集,进一步挖掘了 LSTM 网络的潜能,有利于风电机组齿轮箱在线状态监测。

## 4　结　　论

基于面板 SCADA 数据样本的风电机组状态监测研究近期受到广泛关注,然而数据清洗后出现的行缺失欠完备样本通常未被考虑。为提升可用样本数量、强化状态监测能力,基于注意力机制和 DAE 网络,提出 MAE-AM 网络,用于填补小比例缺失欠完备样本。以河北某风场实际数据为例,MAE-AM 网络的填补效果优于 LMI、MICE、MIDA、DAEMA 方法;数据填补后,整体样本数量增加 18.2%,增强了 LSTM 网络的学习能力。在风电机组齿轮箱运行状态正常段和异常段,状态监测结果的连续性和准确性均有显著提升,有利于工程实际。

## 参考文献

[1] REN Z, VERMA A S, LI Y, et al. Offshore wind turbine operations and maintenance: A state-of-the-art

review[J]. Renewable and Sustainable Energy Reviews, 2021, 144: 110886.

[2] 金晓航, 孙毅, 单继宏, 等. 风力发电机组故障诊断与预测技术研究综述[J]. 仪器仪表学报, 2017, 38(5): 1041-1053.
JIN X H, SUN Y, SHAN J H, et al. Fault diagnosis and prognosis for wind turbines: An overview[J]. Chinese Journal of Scientific Instrument, 2017, 38(5): 1041-1053.

[3] YANG Y J, LIU A M, XIN H W, et al. Fault early warning of wind turbine gearbox based on multi-input support vector regression and improved ant lion optimization[J]. Wind Energy, 2021, 24(8): 812-832.

[4] 王梓齐, 刘长良, 刘帅. 基于集成 NSET 和模糊软聚类的风电机组齿轮箱状态监测[J]. 仪器仪表学报, 2019, 40(7): 138-146.
WANG Z Q, LIU CH L, LIU SH. Condition monitoring of wind turbine gearbox based on ensemble nonlinear state estimation technique[J]. Chinese Journal of Scientific Instrument, 2019, 40(7): 138-146.

[5] 何群, 尹飞飞, 武鑫, 等. 基于长短期记忆网络的风电机组齿轮箱故障预测[J]. 计量学报, 2020, 41(10): 1284-1290.
HE Q, YIN F F, WU X, et al. Fault prediction of wind turbine gearbox based on long short-term memory network[J]. Acta Metrologica Sinica, 2020, 41(10): 1284-1290.

[6] SU X, SHAN Y, LI C, et al. Spatial-temporal attention and GRU based interpretable condition monitoring of offshore wind turbine gearboxes[J]. IET Renewable Power Generation, 2022, 16(2): 402-415.

[7] WANG Z Q, LIU CH L, YAN F. Condition monitoring of wind turbine based on incremental learning and multivariate state estimation technique[J]. Renewable Energy, 2022, 184: 343-360.

[8] 王梓齐, 张书瑶, 刘长良. 基于增量式相对熵的风电机组实时状态监测[J]. 电子测量与仪器学报, 2020, 34(12): 125-132.
WANG Z Q, ZHANG SH Y, LIU CH L. Real-time condition monitoring of wind turbine based on incremental relative entropy[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(12): 125-132.

[9] ZHU L, ZHANG X. Time series data-driven online prognosis of wind turbine faults in presence of SCADA data loss[J]. IEEE Transactions on Sustainable Energy, 2020, 12(2): 1289-1300.

[10] KONG Z Q, TANG B P, DENG L, et al. Condition monitoring of wind turbines based on spatio-temporal fusion of SCADA data by convolutional neural networks and gated recurrent units[J]. Renewable Energy, 2020, 146: 760-768.

[11] 刘帅, 刘长良, 甄成刚, 等. 基于群体多维相似性的风机齿轮箱预警策略[J]. 仪器仪表学报, 2018, 39(1): 180-189.
LIU SH, LIU CH L, ZHEN CH G, et al. Fault warning strategy of wind turbines gearbox based on group multidimensional similarity[J]. Chinese Journal of Scientific Instrument, 2018, 39(1): 180-189.

[12] SUN C, CHEN Y, CHENG C. Imputation of missing data from offshore wind farms using spatio-temporal correlation and feature correlation[J]. Energy, 2021, 229: 120777.

[13] 熊中敏, 郭怀宇, 吴月欣. 缺失数据处理方法研究综述[J]. 计算机工程与应用, 2021, 57(14): 27-38.
XIONG ZH M, GUO H Y, WU Y X. Review of missing data processing methods[J]. Computer Engineering and Applications, 2021, 57(14): 27-38.

[14] GONDARA L, WANG K. MIDA: Multiple imputation using denoising autoencoders[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2018: 260-272.

[15] 任欢, 王旭光. 注意力机制综述[J]. 计算机应用, 2021, 41(S1): 1-6.
REN H, WANG X G. Review of attention mechanism[J]. Journal of Computer Applications, 2021, 41(S1): 1-6.

[16] TIHON S, JAVAID M U, FOURURE D, et al. DAEMA: Denoising autoencoder with mask attention[C]. International Conference on Artificial Neural Networks, 2021: 229-240.

[17] 金晓航, 许壮伟, 孙毅, 等. 基于生成对抗网络的风电机组在线状态监测[J]. 仪器仪表学报, 2020, 41(4): 68-76.
JIN X H, XU ZH W, SUN Y, et al. Online condition monitoring of wind turbine based on generative adversarial network[J]. Chinese Journal Scientific Instrument, 2020, 41(4): 68-76.

[18] 刘智慧, 张承瑞, 李瑞珍. 基于机器视觉的光学镜片测量方法[J]. 电子测量技术, 2022, 45(1): 129-133.
LIU ZH H, ZHANG CH R, LI R ZH. Measuring method of optical lens size based on machine vision[J]. Electronic Measurement Technology, 2022, 45(1): 129-133.

[19] LUO ZH H, FANG CH Y, LIU CH L, et al. Method for

cleaning abnormal data of wind turbine power curve based on density clustering and boundary extraction[J]. IEEE Transactions on Sustainable Energy, 2021, 13(2): 1147-1159.

[20] LIN Z, LIU X L, COLLU M. Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks [J]. International Journal of Electrical Power & Energy Systems, 2020, 118: 105835.

[21] QU F M, LIU J H, LIU X Y, et al. A multi-fault detection method with improved triplet loss based on hard sample mining[J]. IEEE Transactions on Sustainable Energy, 2020, 12(1): 127-137.

[22] 王浙超, 曾九孙, 谢磊, 等. 基于去噪自编码器的故障隔离与识别方法[J]. 信息与控制, 2021, 50(6): 641-650.
WANG ZH CH, ZENG J S, XIE L, et al. Fault isolation and identification method based on denoising autoencoder[J]. Information and Control, 2021, 50(6): 641-650.

[23] HE X N, HE ZH K, SONG J K, et al. NAIS: Neural attentive item similarity model for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2354-2366.

[24] 滕伟, 丁显, 史秉帅, 等. 基于 WGAN-GP 的风电机组传动链故障诊断[J]. 电力系统自动化, 2021, 45(22): 167-173.
TENG W, DING X, SHI B SH, et al. Fault diagnosis of wind turbine drivetrain based on wasserstein generative adversarial network-gradient penalty[J]. Automation of Electric Power Systems, 2021, 45(22): 167-173.

[25] ROYSTON P, WHITE I R. Multiple imputation by chained equations (MICE): Implementation in stata[J]. Journal of Statistical Software, 2011, 45: 1-20.

## 作者简介

**徐健**(通信作者),2019 年于华北电力大学获得学士学位,现为华北电力大学硕士研究生,主要研究方向为风电机组状态监测与故障预警。
E-mail: ncepu_zdsxj@163.com

**Xu Jian** (Corresponding author) received his B. Sc. degree from North China Electric Power University in 2019. He is currently a master student at North China Electric Power University. His main research interests include wind turbines condition monitoring and fault warning.

**刘长良**,分别在 1985 和 1990 年于华北电力学院获得学士和硕士学位,2002 年于华北电力大学获得博士学位,现为华北电力大学教授、博士生导师,主要研究方向为风电机组故障诊断,火电机组建模与仿真等。
E-mail: 13603123513@163.com

**Liu Changliang** received his B. Sc. degree and M. Sc. degree both from North China Electric Power College in 1985 and 1990, and received his Ph. D. degree from North China Electric Power University in 2002. He is currently a professor and a Ph. D. advisor at North China Electric Power University. His main research interests include wind turbine fault diagnosis, thermal power unit modeling and simulation, etc.