

DOI: 10. 19650/j.cnki.cjsi.J1804464

基于数据分类重建的风电机组故障预警方法^{*}

刘 帅¹, 刘长良^{1,2}, 甄成刚¹

(1. 华北电力大学 控制与计算机工程学院 北京 102206;

2. 新能源电力系统国家重点实验室(华北电力大学) 北京 102206)

摘 要:为预警风电机组潜在故障、增强机组出力安全性,基于监控和数据采集(SCADA)系统,提出一种异常数据重建的风电机组故障预警方法。首先,充分利用同风场风机 SCADA 数据,分别重建输入类与输出类目标机组数据,克服了部分数据信息丢失、数据异常问题;其次,使用提取的代表性数据建立故障预警模型,所得预警模型更贴近机组运行动态特性;最后,采用改进的衰退指标预警潜在故障,直观展示机组阶段性衰退程度。案例研究中使用某风电场 SCADA 故障数据,并使用 3 种标准确定所提策略参数设定值,结果表明可至少提前 3 周预警风电机组齿轮箱潜在故障,验证了所提故障预警方法的时效性。

关键词: 故障预警方法;梯度提升回归树;高斯混合模型;数据分类重建;衰退指标

中图分类号: TH17 **文献标识码:** A **国家标准学科分类代码:** 480. 6040

Fault warning method for wind turbine based on classified data reconstruction

Liu Shuai¹, Liu Changliang^{1,2}, Zhen Chenggang¹

(1. School of Control and Computer Engineering, North China Electric Power University, Changping District, Beijing 102206, China;

2. State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources,

North China Electric Power University, Changping District, Beijing 102206, China)

Abstract: In order to warn the potential failure of wind turbines and enhance the safety of the unit output, a wind turbine fault warning method with abnormal data reconstruction is proposed based on the supervisory control and data acquisition (SCADA) system. Firstly, the SCADA data of the wind turbines in the same wind farm are fully utilized to reconstruct the two types of target unit data of the input and output, respectively, which overcomes the problems of partial data missing and data anomaly. Secondly, a fault warning model is established using the extracted representative data, which is closer to describe the dynamic characteristics of the unit. Thirdly, the improved deterioration degree is adopted to warn the potential failures and intuitively show the phased deterioration of the unit. In the case study, the SCADA fault data of a certain wind farm were used, and the parameter settings of the proposed strategy were determined with three criteria. The results show that the proposed method can predict the potential fault of the gearbox of the wind turbine at least 3 weeks ahead, which verifies the timeliness of the proposed early warning method.

Keywords: fault warning method; gradient boosted regression tree; Gaussian mixed model; classified data reconstruction; deterioration degree

0 引 言

近年来,风力发电项目部署增长较快,风电装机容量在世界范围内呈明显上升趋势。根据我国“十三五”期间规划,风电将新增装机容量 80 GW 以上。然而,风电

项目抢装潮过后的“风电机组出质保”困局却没有得到足够关注,据保守测算,截至 2016 年末我国至少有 85 GW 风机超出质保期。随着大量的风电机组质保时间到期,国内风电故障事故也将迎来爆发期。加之由于风电机组严苛的运行环境,风电机组主要机械部件,如轴承,发电机,叶片和齿轮箱^[1-2]等,更应该注意检测潜在

收稿日期:2018-11-23 Received Date:2018-11-23

^{*} 基金项目:北京市自然科学基金(4182061)、中央高校基本科研业务费专项资金(2018ZD05,9163116001)项目资助

故障^[3]。

风电机组的故障研究工作属于当前热点,研究方法的数据来源主要有两种:1) 监控和数据采集(supervisory control and data acquisition, SCADA)数据;2) 振动数据。此处以数据源作为分类分析目前风电机组故障研究现状。

基于 SCADA 数据的方法,Qiu 等^[4]提出了一种基于热物理的风电机组传动系故障诊断方法,从热物理学角度揭示了风电机组运行时发生的变化;Dao 等^[5]提出了一种基于 SCADA 数据协整分析的新方法,可以有效地分析非线性数据趋势;Borchersen 等^[6]使用扩展卡尔曼滤波器分析了 43 个风机的 3 年历史数据并发现了 15/16 类故障;Cao 等^[7]开发的无迹卡尔曼滤波器模型为风力发电机故障检测提供了一种有效途径。以上故障诊断方法尝试使用多种建模方法及数据处理方法预警机组潜在故障,但在数据预处理方面仍需深入研究。

基于采集的振动数据,在风电机组进入某种预定义的情况时,Mitra 等^[8]通过加速度计收集振动信号,然后使用标准偏差识别出有缺陷的轴承。经验模态分解(empirical mode decomposition, EMD)^[9]具有将信号分解为不同模态的能力,Li 等^[10]提出了改进的变分模态分解(variational mode decomposition, VMD),以识别轮对轴承故障;Hang 等^[11]提出了一种基于多类模糊支持向量机(fuzzy support vector machine, FSVM)分类器的风力机故障诊断方法。由于振动数据采样频率较高,其信号所携带的故障特征较为丰富,所以诊断结果较为准确,但此类方法往往受制于信号采集设备的部署难题。

组方法,Bangalore 等^[12]提出了一个利用各种来源数据的框架,并为基于状态的维护计划提出了比例风险模型。Xin 等^[13]使用非线性状态估计(nonlinear state estimation technique, NSET)方法实现了变速箱、发电机和主轴承的状态监测和故障诊断。这类方法可以将两类信号分别处理合并得出机组诊断结果,但如何有效结合两种信号不同时间尺度问题尚未有充分研究成果。

风电机组的故障监测数据源使用 SCADA 的突出特点为易于获取,风场都会部署相关软硬件。但 SCADA 原始数据经过传感器采集、线路传输、系统处理、硬盘存储等多个环节后,会出现噪声数据较多、存储条目不全等现象。所以,基于 SCADA 数据研究一种可克服通讯异常、数据缺失的故障预警方法既是工程应用的需要,也是对当前技术水平的挑战。

针对此问题,本文提出一种基于异常数据重建的风力发电机故障预警方法。通过应用动态时间规整(dynamic time warping, DTW)和梯度增强回归树(gradient boosting regression tree, GBRT)算法分别重构了不同类别数据,研究了基于高斯混合模型(Gaussian

mixed model, GMM)的预警模型,并提出了冗余的、改进的衰退指标,用于指示风机的健康状况。

1 数据处理

1.1 SCADA 数据说明

该数据来自于福建省某风电场,其配备维斯塔斯集团 2 MW 风力发电机,切入风速为 4 m/s,切出风速为 25 m/s。由于风场所在地理位置良好,月平均风资源大于 5 m/s。共收集了风机的 20 多个属性参数,每个风机有近 23 000 个数据项。数据中心通过远程传感器传输每 10 min 记录一次 SCADA 数据。表 1 为所收集的部分有效 SCADA 数据。

表 1 收集到的有效 SCADA 数据						
Table 1 The collected effective SCADA data						
时间	2016/1/1	2016/1/1	2016/6/23	2016/6/23	...	
	0:00	0:10	...	0:00	0:10	...
风速	16.26	16.54	...	6.04	5.59	...
风向	39.31	41.02	...	229.04	228.27	...
环境温度	8.57	8.5	...	26.05	26	...
发电机转速	1 680.62	1 678.81	...	1 296.05	1 182.84	...
叶轮转速	14.9	14.9	...	11.6	10.52	...
齿轮箱轴承温度	62	62	...	54	53.61	...
有功功率	2004.16	1989.04	...	309.48	248.94	...
无功功率	0.03	0.03	...	0.01	0	...

1.2 数据分类重建过程

图 1 所示为收集的风电机组 SCADA 风速数据,部分异常 SCADA 数据在数据中心中记录为错误代码-902,图 1 中小于 0 值处即为错误代码替代处(图中横线相交位置)。

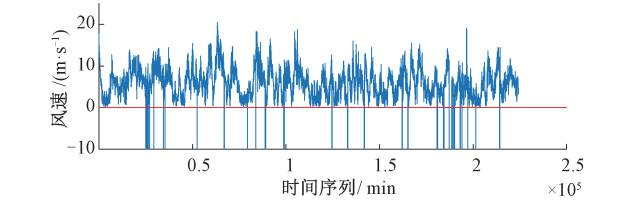


图 1 采集的 SCADA 风速数据

Fig.1 The collected SCADA wind speed data

将 SCADA 数据按属性分为两类,输入类(包括风速、环境温度等)和输出类(包括齿轮箱轴承温度、发电机转速、有功功率等)。输入类为风电机组的外部输入参数,输出类为风电机组的内部状态参数。由于异常数据值被记录为-902,并且气象数据(离地面 70 m 以上)难以获

取,需要两个步骤来重建完整的 SCADA 数据。首先,通过规整风电场中的其他风机的正常时间序列段来重建输入类参数(不局限于同时刻数据);输出类是使用缺失数据的前一段正常数据训练模型进行回归重建。

1) 输入类重构

DTW 算法^[14]通过最小化成本函数,一个数据点可以匹配多个数据点,可以通过 DTW 测量不同长度的数据序列。DTW 具有大量的应用场景,如网络行为识别^[15]、脑数据分析^[16]等。

发生齿轮箱故障的风电机组为 14 号风机,同时,也从同风场收集了 16 台周边机组的 SCADA 数据。应用全部 16 台风力机重构 14 号风力机的风速,采用加权动态时间规整(weighted dynamic time warping, WDTW)算法得到输入类参数。

加权 DTW 算法计算如下:

$$WS_{No.14} = \sum_{i=1}^N \frac{1}{\sum_{j=1}^N \frac{\log(\text{dist}(\text{seg}_{No.14}, \text{seg}_{No.i}))}{\log(\text{dist}(\text{seg}_{No.14}, \text{seg}_{No.j}))}} WS_{No.i} \quad (1)$$

式中: WS 为风电机组风速; dist 为 DTW 原始算法得出的时间序列段之间距离; seg 为风电机组缺失数据前的正常数据段; N 为有实际正常风速的风机数量。

2) 输出类重建

输出类参数通过集成学习回归方法重建。GBRT 算法^[17]的核心是将几个弱回归器组合成一个强回归器,找到最小化每个步骤的损失函数的解决方案。GBRT 算法的应用范围广泛,如预测地热能通量^[18],发现废物产生模式^[19],基于模型的个性化推荐^[20],预测短期电能^[21]等。使用 GBRT 处理在远程传输过程中缺少某些输出类 SCADA 数据的情况,其输入为已重构的机组输入类参数,输出为输出类重建参数。

GBRT 算法的流程为 $\{(x_i, y_i)\}_{i=1}^n$ 表示一组训练样本,其中 y 是目标向量, x 是输入向量。设 $L(y, F(x))$ 是损失函数,其中 $F(x)$ 是 x 映射至 y 的模型函数。假设 $F_0(x) = c$, c 为常数。则循环计算如下步骤,直至迭代之 M 次结束。

(1) 计算残差的负梯度:

$$g_i = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x_i) = F_{m-1}(x_i)} \quad (2)$$

其中, $i = 1, \dots, n$ 。

(2) 然后根据训练集和残差的负梯度生成第 L 个回归树 T_m 。

(3) 计算新回归树的权重因子:

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma T_m(x_i)) \quad (3)$$

(4) 更新树模型:

$$F_0(x) = F_0(x) + \gamma_m T_m(x_i) \quad (4)$$

2 故障预警方法

所提风电机组故障预警方法由数据分类重建、模型训练和故障预警 3 部分组成,是一种在线离线通用的方法,图 2 所示为故障预警方法整体框架。将历史 SCADA 数据经数据重建,以训练数据生成 GMM 预警模型,使用测试数据或在线完整 SCADA 数据实现故障预警。预警过程中,整条完整 SCADA 数据直接经 GMM 预警模型判断是否存在潜在设备故障,而异常值需要先行重建数据后,才能与完整 SCADA 数据一样完成故障预警流程。

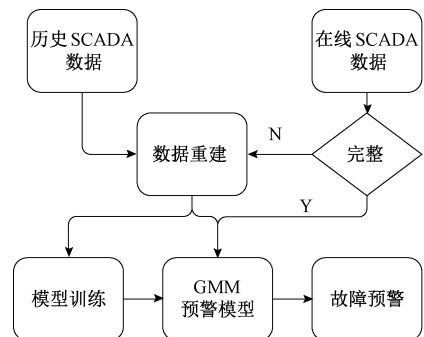


图2 所提故障预警方法结构

Fig.2 The structure diagram of the proposed fault warning method

2.1 数据分类重建

由于 SCADA 数据在风电公司的数据中心远程存储过程中,可能会出现一些连接或传感器错误,这是导致收集到的 SCADA 数据残缺的主要原因。数据分类重建的详细方法已在 1.2 节描述。经数据分类重建,所得目标机组的完整 SCADA 数据作为后续部分的输入数据,此处将其分为两部分,即训练数据和测试数据。训练数据是用于预警模型训练部分的输入数据;剩余部分为测试数据,用于故障预警部分。数据重构部分框架结构如图 3 所示。

2.2 GMM 预警模型训练

图 4 所示为模型训练部分结构示意图,经所提出的训练流程后,可创建一个稳健的 GMM 预警模型。在训练预警模型之前,需要对训练数据进行数据处理,因为训练数据的选择对于最终模型是否能够代表机组的根本特征有直接关系。

首先,提取目标机组常态数据。基于密度的含噪声空间聚类(density-based spatial clustering of applications

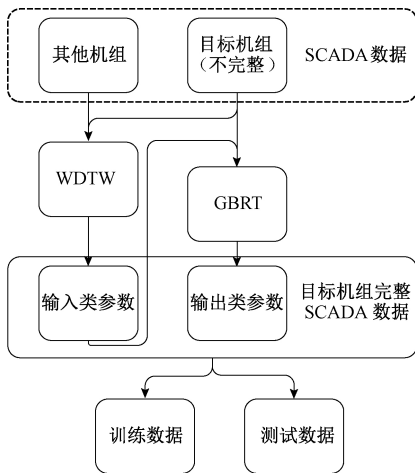


图3 数据分类重建部分框架结构

Fig.3 The frame structure of the classified data reconstruction section

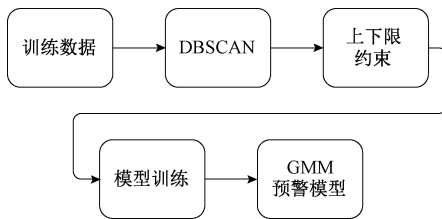


图4 模型训练部分结构

Fig.4 The structure of the model training section

with noise, DBSCAN) 算法是一种数据聚类算法,用于获取潜在的数据类。作为预警模型训练部分的输入数据,异常数据会降低训练 GMM 模型的准确性,最终导致所训练模型精度无法满足预警故障要求。

其次,使用上下限约束去除异常数据,特别是极端风力资源条件下记录的 SCADA 数据项。由于风资源分布具有不均衡特点,应尽可能多地保留与动态特性相关的数据,以提高 GMM 预警模型的准确性。

最后,使用已预处理的输入类数据来构建 GMM 预警模型。GMM 是一种多元分布函数,通过任意多个子函数的加权,可模拟任意分布。GMM 可直接应用,也能和其他模型配合使用,GMM 能够对捕获的数据集进行分类^[22];Ntalampiras 等^[23]使用 GMM 来区分网络-物理关键基础设施的正常状态或异常状态。另一方面,GMM 与其他模型搭合作以实现不同应用目的;张龙等^[24]提出了一种基于 TESPAS 与 GMM 的滚动轴承性能退化评估方法,轴承疲劳试验表明该方法能及时发现轴承早期故障^[24];Chen 将 Gibbs 随机场应用于图像分割,旨在提高基于 GMM 的手势识别准确率^[25]。

一个含有 k 个高斯分布函数的 GMM 模型,其中某高

斯分布函数的概率密度函数表示为:

$$N(x | \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}} \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2}(x - \mu)^T \cdot \Sigma^{-1} \cdot (x - \mu)\right] \quad (5)$$

式中: x 是 D 维输入向量; μ 是第 k 个高斯分布平均值; Σ 是第 k 个高斯分布的协方差矩阵。

GMM 模型的概率密度函数如下:

$$p(x) = \sum_{k=1}^K \pi_k N(x | \mu_k, \Sigma_k) \quad (6)$$

式中: K 是 GMM 模型中高斯函数的数量; π_k 是每个高斯分布函数的权重。 π_k 满足如下约束:

$$\sum_{k=1}^K \pi_k = 1 \text{ 且 } 0 \leq \pi_k \leq 1$$

2.3 故障预警

故障预警部分中,将测试数据及在线 SCADA 数据作为故障预警部分的输入,经过多个步骤处理数据后,使用冗余的、改进的衰退指标在线警告潜在故障。图 5 所示为机组故障预警框架结构示意图。

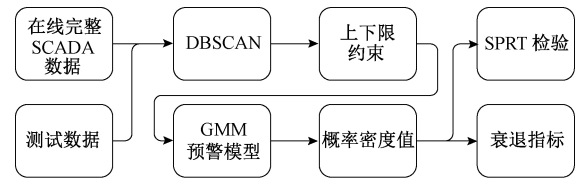


图5 故障预警部分框架结构

Fig.5 The frame structure of the fault warning section

GMM 预警模型的训练参数经过多重数据筛选,在故障预警部分中则应以相同的方式对测试数据进行预处理,由此得出的监测部件故障结论亦更有说服力。DBSCAN 算法和上下限约束可以细化 GMM 模型输入数据,经 GMM 预警模型产生的概率密度值用来指示监测部件参数分布是否发生本质变化。最后,使用衰退指标量化故障演变过程,并辅以序贯概率比检验 (sequential probability ratio test, SPRT)^[26]完成故障预警。

衰退指标 (deterioration degree, DD) 计算公式如下:

$$\text{LOG}(i) = -\log\left(\sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k)\right) \quad (7)$$

$$\text{DD}(i+T) = \frac{\alpha}{T} \sum_{t=\text{end}-T+1}^{\text{end}} \text{DD}_t + (1-\alpha) \text{LOG}^2(i) \quad (8)$$

式中: I 是测试集序列长度且 $1 \leq i \leq I$; α 为记忆权重因子,可控制前时刻指标对下一时刻的影响; T 是一个平滑因子,用于求取范围内的平均值, end 是衰退指标序列的末端序号。

SPRT 的本质是序贯抽样方案,即不确定总的抽样个数,而是先抽取少量样本,根据其结果决定是否继续抽样

或停止抽样。假设某随机变量服从分布 p , 其累积似然比为:

$$S_i = S_{i-1} + \log \Lambda_i$$

(9)

其中, $\log \Lambda_i$ 为其增量, 且 $S_0 = 0$ 。

存在一对假设, 零假设 H_0 及备择假设 H_1 :

$$\begin{cases} H_0: p = p_0 \\ H_1: p = p_1 \end{cases}$$

(10)

抽样开始后, 逐次计算 S_i , 停止抽样的条件由两个常数确定:

$S_i \geq b$, 则接受备择假设 H_1 ;

$S_i \leq a$, 则接受零假设 H_0 ;

$S_i \in (a, b)$, 则继续抽样。

而常数 a 、 b 的确定与犯第 1 类错误的概率 α 和犯第 2 类错误的概率 β 有关。近似地满足如下关系:

$$a \approx \log \frac{\beta}{1 - \alpha}, b \approx \log \frac{1 - \beta}{\alpha}$$

3 案例研究

使用现场风电机组 SCADA 数据检验所提故障预警方法, 以验证其有效性、预警结果可读性。

3.1 故障预警方法参数设置

所采集 SCADA 数据基本信息已在第 2 节介绍, 其中第 14 号风力发电机是本文的研究对象。14 号风机于 2016 年 6 月 23 日由变速箱故障导致离线, 停机时间长达 834 h。

在研究数据重构功能部分之前, 首先需要确定参数属性的选择问题。本文所使用 SCADA 数据中与齿轮箱轴承温度相关的属性有齿轮箱轴承温度, 此处通过研究与齿轮箱轴承温度相关属性的皮尔逊相关系数(见表 2), 确定训练 GMM 模型的参数属性。两个变量 $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ 及 $\mathbf{Y} = \{y_1, y_2, \dots, y_n\}$ 的皮尔逊相关系数 r 的计算公式如下:

$$r(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

(11)

其中, $\bar{x}$ 和 $\bar{y}$ 为两变量的均值。

表 2 相关系数表

Table 2 The correlation coefficient table

相关系数	发电机转速	风速	有功功率	环境温度
齿轮箱轴承温度	0.950 2	0.871 6	0.774 5	-0.279 8

将齿轮箱轴承温度、环境温度、风速、有功功率和发电机转速归一化后, 分别得到齿轮箱轴承温度与环境温

度、风速、有功功率、发电机转速的数据点对应图, 如图 6 所示(归一化后横轴和纵轴无单位)。从图 6 可知: 齿轮箱轴承温度与环境温度之间的相关关系较弱, 而其他几个参数属性相关关系较强。

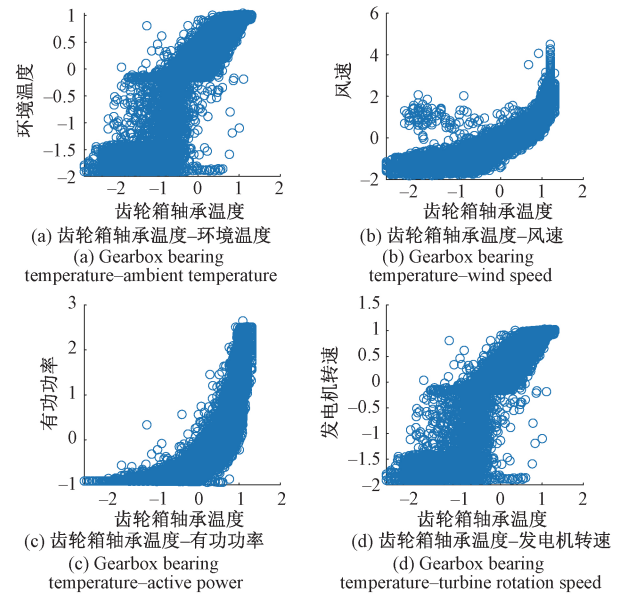


图 6 数据点对应图

Fig.6 Data point mapping diagrams

根据表 2 和图 6 选择发电机转速, 风速, 有功功率和齿轮箱轴承温度作为 GMM 模型训练输入参数(相关系数数值的正负代表两变量之间的相关性的正负)。而由于齿轮箱轴承温度与环境温度之间的关系较弱, 环境温度不作为 GMM 模型训练的输入参数。

1) 数据分类重建部分: 在 WDTW 算法中, 时间序列的长度设定为 200, 并使用其他 15 个风机重建输入类参数。输出类参数由 GBRT 算法重建, 最后, 将输入类和输出类参数组合成完整的 SCADA 数据, 留待第 2 部分训练 GMM 模型。

关于使用 GBRT 算法重建输出类的参数选择, 此处使用 3 个标准评价算法性能: 均方根误差 (root mean square error, RMSE)、平均绝对误差 (mean absolute error, MAE) 和决定系数 (coefficient of determination, R^2), 即式 (12)~(14)。RMSE 表示算法回归结果的准确性, 结果越准确 RMSE 越小。MAE 代表回归结果的一致性, 如果一些回归结果远离真实数据, MAE 将会非常大。 R^2 亦称之为拟合优度, 当 R^2 越接近 1 时, 表示相关变量拟合度越高; 相反, 越接近 0 时, 表示相关变量拟合度越低。

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (observed_i - predicted_i)^2}$$

(12)

$$MAE = \frac{1}{N} \sum_{i=1}^N |observed_i - predicted_i|$$

(13)

$$R^2 = 1 - \frac{\sum_{i=1}^N |\text{observed}_i - \text{predicted}_i|}{\sum_{i=1}^N |\text{observed}_i - \overline{\text{observed}}|} \quad (14)$$

图7所示为以有功功率为例,分别固定回归树数量和叶子节点数量的GBRT算法参数选择图。从图7(a)可以看出, RMSE与MAE在回归树数量[20, 50]范围内到达最低点,同时 R^2 的值在此区间也趋于稳定,图7(b)中RMSE与MAE在叶子节点[10, 20]范围中到达最低点,同时 R^2 的值在此区间也取得极值。需要注意的是, GBRT算法中回归树或叶子节点的数值越大,算法耗时越多。最终,以平衡耗时和准确性为目的,选择30和15来设置GBRT算法中回归树或叶子节点的数量。

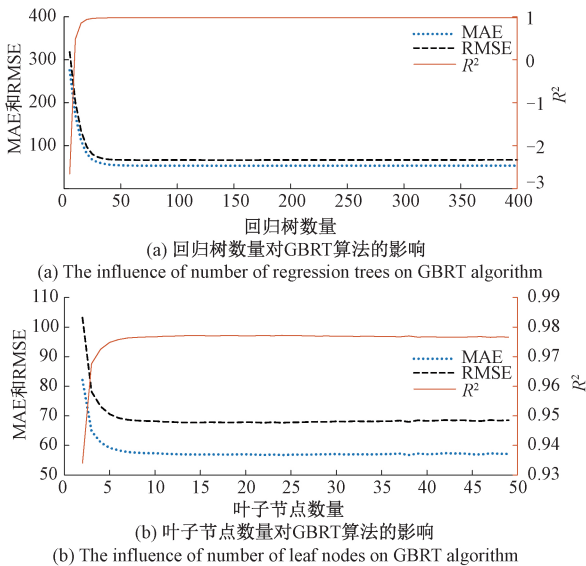


图7 GBRT算法参数设定图

Fig.7 The parameter setting diagram of GBRT algorithm

使用10折验证计算训练样本大小对有功功率、发电机转速和齿轮箱轴承温度重建结果的影响。由于三者之间有相似性,以齿轮箱轴承温度为例,分析训练样本大小对重建结果的影响,如图8所示。在10折验证中样本被分成2类,分别是训练集和验证集,而测试集是与训练样本无关的一段数据,3种集合的RMSE趋势为图8(a)。图8(b)是计算测试集与训练集、验证集与训练集的RMSE偏差趋势,可以代表所训练模型的泛化能力和精度,能说明增加的训练样本对提高模型精度的作用大小。图8(c)是基于重建模型生成的风速—功率曲线,能够代表模型的预测结果。

图8(a)训练集的RMSE处于上升态势,验证集在1000项处达到最低点,测试集在500项内的RMSE值迅速缩小,3个数据集在2000项训练样本后的RMSE值较为平稳。图8(b)中RMSE偏差在2000项后变化

较小,说明再增加训练样本对GBRT模型的泛化能力与精度提高较小。图8(c)使用1000项训练样本时,风速—齿轮箱轴承温度曲线不平滑,波动较大,而使用2000项和2500项训练样本时,风速—齿轮箱轴承温度曲线较平滑,且两者区别不大。综合上述结果,为保证数据分类重建的有效性,同时平衡算法耗时,使用缺失SCADA数据的前2000项(约2周数据)完整数据作为训练样本。

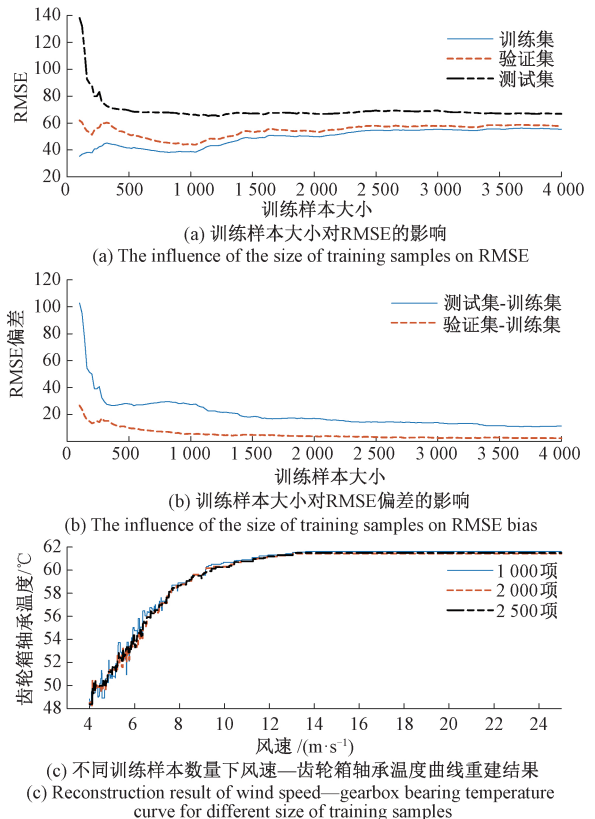
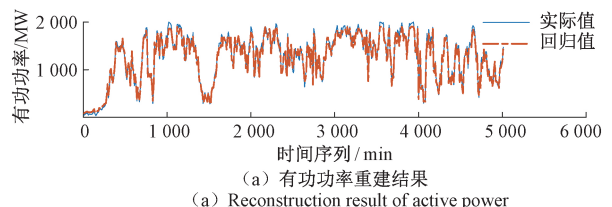


图8 训练样本大小对重建结果的影响

Fig.8 The influence of training sample size on reconstruction result

数据分类重建部分算法参数设置验证:使用上述GBRT算法的参数设置及完整SCADA数据进行训练,连续预测500项“假设丢失”的有功功率、齿轮箱轴承温度、发电机转速数值,所得结果与实际值相比可得GBRT算法验证结果,如图9所示。



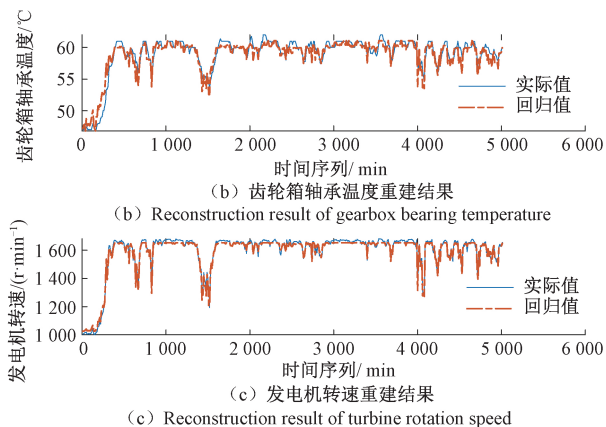


图9 GBRT 算法验证结果

Fig.9 Verification result of GBRT algorithm

2) 模型训练部分: 训练数据集是 5 000 条完整的 SCADA 数据项。在 DBSCAN 算法中, 最小包含点数和邻域半径分别为 50 和 0.02。区间 [6, 13] 包含更多的风机动态特性, 选择此区间精炼有效数据。

3) 故障预警部分: 数据处理相关设置与模型训练部分相同, 使用式(7)和(8)计算衰退指标, 其中 λ 和 T 分别为 0.89 和 3; SPRT 检验中犯第 1 类错误的概率 $\alpha = 0.05$ 和犯第 2 类错误的概率 $\beta = 0.01$ 。

3.2 结果和讨论

1) 准确性

经数据重构后, 输入类参数(如风速)可去除参数中的缺失值, 如图 10 所示, 图中横线为零风速线。由图 1 可知, 获取的风电机组原始时间序列有很多 902 点, 因此有许多线穿过 x 轴; 图 10 为输出类参数数据分类重建的结果。然后, 使用 GBRT 算法重建了相关的输出类参数, 有功功率、齿轮箱轴承温度和发电机转速。从图 11~13 所示可以知道, 输入类数据丢失的信息已经重构, 将其作为输入类数据, 可将 359 个缺失数据分类重建。图 11 与 13 的横线为零值线, 图 12 的横线为 25℃ 线。

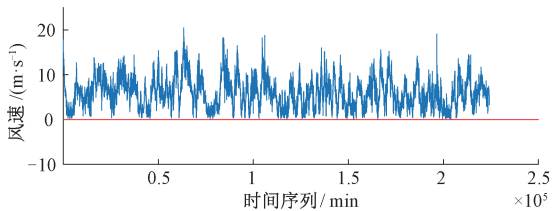


图10 输入类风速重建结果

Fig.10 The reconstruction result of input class wind speed parameter

将上下限约束应用到 GMM 预警模型输入数据上, 如图 14 所示; 训练好的 GMM 预警模型的概率密度散点图如图 15 所示, 颜色条显示三维坐标下的概率密度值。颜色越热, 概率密度值越大。

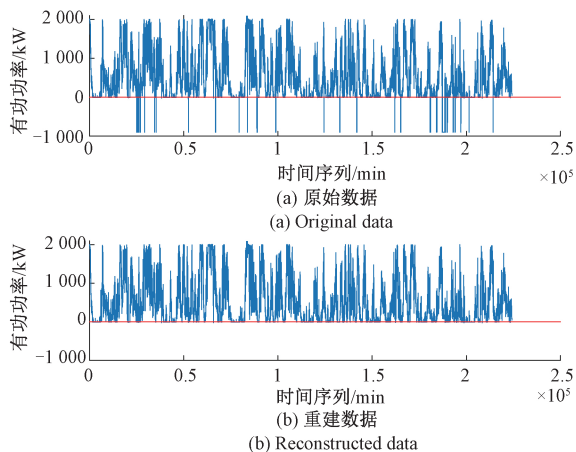


图11 有功功率数据分类重建结果对比

Fig.11 Comparison of the classified data reconstruction results of active power

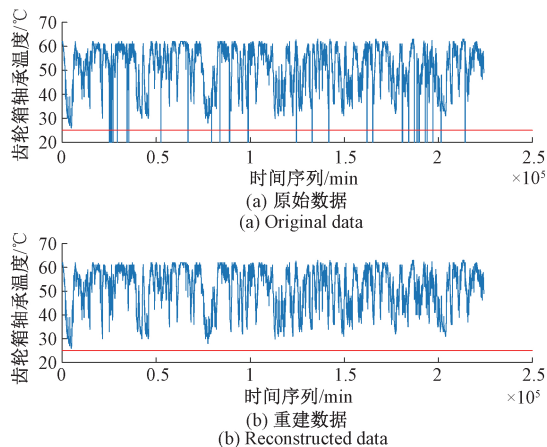


图12 齿轮箱轴承温度数据分类重建结果对比

Fig.12 Comparison of the classified data reconstruction results of gearbox bearing temperature

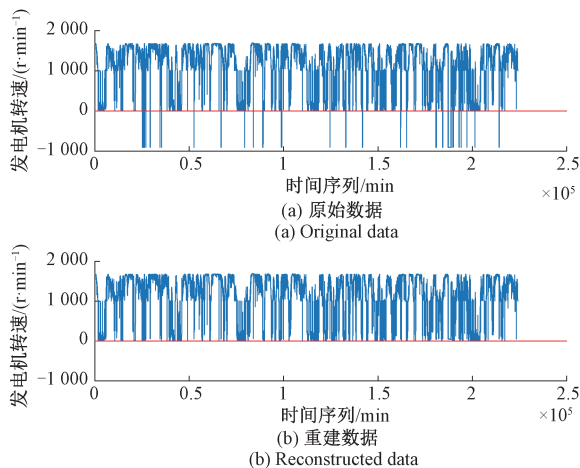


图13 发电机转速数据分类重建结果对比

Fig.13 Comparison of the classified data reconstruction results of turbine rotation speed

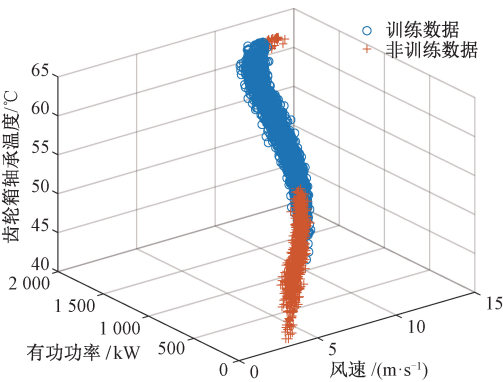


图 14 GMM 预警模型的输入数据

Fi.14 The input data of the GMM warning model

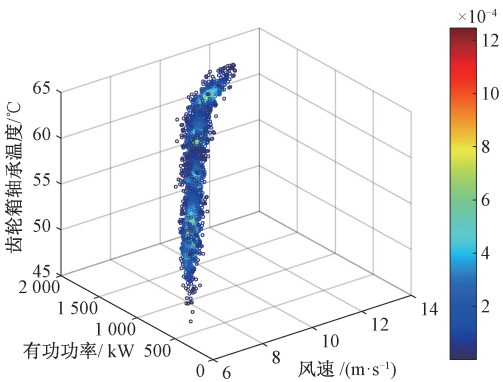


图 15 GMM 预警模型的概率密度

Fig.15 Probability density diagram of the GMM warning model

在线 SCADA 数据输入 GMM 预警模型得到概率密度值,将其转化为衰退指标如图 16 所示。监测开始时,衰退指标略有波动,在 2016 年 2 月 18 日之后波动的趋势增长明显,很容易注意到衰退程度的上限越来越大,且趋势的不稳定性正在增加。表 3 所示为从轻微到严重恶化的 5 个衰退阶段的统计信息。

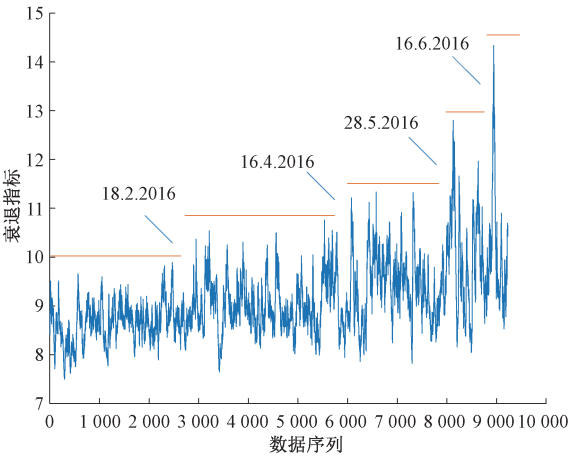


图 16 衰退指标

Fig.16 Deterioration degree index

表 3 衰退阶段

Table 3 The deterioration degree stages

统计值	均值	方差	中数	区间
第 1 阶段 (2016. 1. 1~2016. 2. 18)	8. 643 5	0. 162 6	8. 668 3	2. 393 4
第 2 阶段 (2016. 2. 18~2016. 4. 16)	8. 965 1	0. 248 6	8. 888 5	3. 110 5
第 3 阶段 (2016. 4. 16~2016. 5. 27)	9. 481 4	0. 468 2	9. 506 6	3. 512 5
第 4 阶段 (2016. 5. 28~2016. 6. 16)	9. 477 5	0. 805 8	9. 263 0	4. 756 4
第 5 阶段 (2016. 6. 16~2016. 6. 23)	10. 070 0	1. 073 6	9. 797 5	6. 119 9

以极值来看,在 4 月 16 日之后,极值高于 2 月 18 日之前的第 1 阶段极值 12%,距故障发生(6 月 23 日)还有 2 个月时间,此时故障已经显出征兆;使用 SPRT 检验方法用于辅助预警,也显示已经出现故障预警信号(见图 17)。至 5 月 28 日,极值已经是第 1 阶段的 28%,此时故障征兆明显;6 月 16 日,极值达到第 1 阶段的 44%,此时故障已经开始表现,出现大量异常 SCADA 数据,数据库中-902 错误代码出现频率猛增。6 月 23 日,14 号风机停机检修,更换故障齿轮箱。

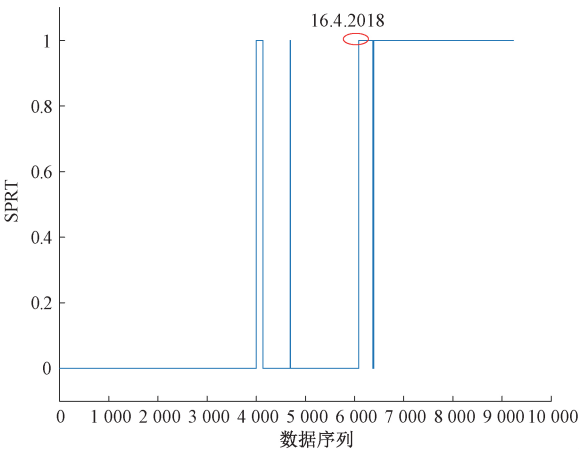


图 17 SPRT 检验辅助预警

Fig.17 Auxiliary warning with SPRT test

对照表 3,以衰退阶段来看,衰退第 1 阶段和第 2 阶段,4 个统计值都有所上升,但第 3 阶段的均值、方差变化较大,说明此阶段故障征兆显现,需要对机组格外关注;而后第 3、4 阶段虽然均值并无明显变化,但是其方差、区间变化较为明显,说明第 4 阶段数据震荡剧烈,侧

面反映了设备衰退加剧;第 5 阶段因故障产生较多异常值,此阶段应安排非计划停机检修,至 6 月 23 日停机时,由于故障严重,只能更换齿轮箱消除故障。

从以上分析可知,共 2 个时间节点预警潜在故障,分别是 4 月 16 日和 5 月 28 日;有 1 个明显时间节点(6 月 16 日)安排故障停机检修齿轮箱,提前排除故障,避免长时间停机。

2) 必要性

倘若缺失 5 月 28 日的 27 个点,衰退指标中对应的峰值将高出第 1 阶段极值 20%、仅高出第 3 阶段极值 5%;同时,对应表 3 第 4 阶段数据为均值 9.399 5,方差 0.603 5,中数 9.2472,区间 3.918 9。由于数据变化较缓,可理解为机组常态衰退,从而错失此处预警时间节点。而 6 月 16 日中 40 个点构成了表 3 第 5 衰退阶段的峰值,如果此时出现数据缺失的现象,将彻底错失预警机会。

假设目标机组在 5 月 28 日及 6 月 16 日的 SCADA 数据“恰好”丢失了上述关键数据点,使用所提方法分别以“丢失关键数据点”及“分类重建丢失数据点”两种思路处理 SCADA 数据,所得结果如图 18 和 19 所示。图 18 为丢失关键数据点后所得的衰退指标图,对比图 16 可明显看出,5 月 28 日至故障发生日(6 月 23 日)之间并无显著变化,无法判断是否存在潜在故障。图 19 为分类重建丢失数据点后所得的衰退指标图,对比图 18 可明显观察到衰退指标的剧烈波动,暗示了目标机组的潜在故障,可根据所得衰退指标准确预警故障,妥善安排检修工作,验证了所提方法的准确性。最后,对比图 19 与 16,可得出两图并无显著差别的结论,可提前预警风机潜在故障,验证了所提方法的必要性。

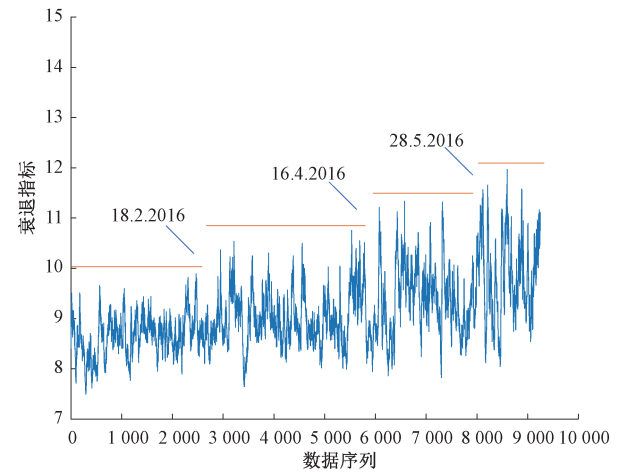


图 18 衰退指标(丢失关键数据点)

Fig.18 Deterioration degree index (missing key data points)

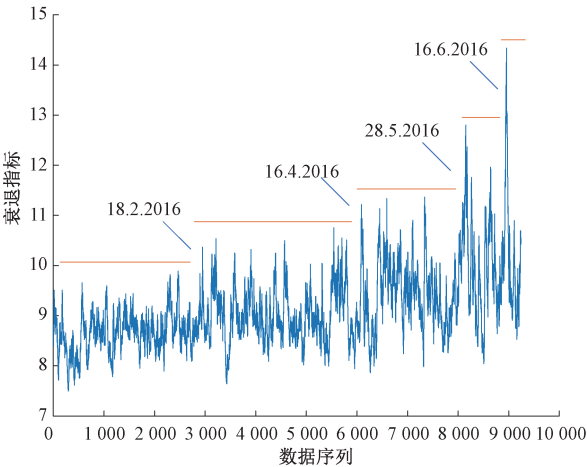


图 19 衰退指标(分类重建丢失数据点)

Fig.19 Deterioration degree index (the missing data points are reconstructed with classified data reconstruction)

4 结 论

本文提出了一种由数据分类重建、模型训练和故障预警 3 部分构成的风电机组故障预警方法。在数据分类重建部分,分别对输入类和输出类数据分类重建,解决了因传输、存储不当带来的部分数据条目缺失问题;在模型训练部分,所训练 GMM 预警模型充分考虑了机组的动态特性及参数相关性,表征了目标机组的本质特征;在故障预警部分,所提衰退指标使得原始 SCADA 数据更加具有可读性,机组故障演变各阶段区分较为清晰、预警及时可靠。案例研究采用福建某风场齿轮箱轴承故障 SCADA 数据,研究结果显示,可提前至少 3 周预警潜在故障,验证了所提方法的有效性、必要性及现场可操作性。

参考文献

[1] LIU W Y, TANG B P, HAN J G, et al. The structure healthy condition monitoring and fault diagnosis methods in wind turbines: A review [J]. Renewable and Sustainable Energy Reviews, 2015(44) : 466-472.

[2] 任玉亭. 振动监测技术在风力发电机组的应用[J]. 内蒙古电力技术, 2010, 28(2) : 8-12.

REN Y T. Application research of vibration supervision technology on wind power generating units [J]. Inner Mongolia Electric Power, 2010, 28(2) : 8-12.

[3] KIM K, PARTHASARATHY G, ULUYOL O, et al. Use of SCADA data for failure detection in wind turbines[C]. ASME, 2011: 2071-2079.

[4] QIU Y, FENG Y, SUN J, et al. Applying thermophysics

- for wind turbine drivetrain fault diagnosis using SCADA data [J]. *IET Renewable Power Generation*, 2016, 10(5): 661-668.
- [5] DAO P B, STASZEWSKI W J, BARSZCZ T, et al. Condition monitoring and fault detection in wind turbines based on cointegration analysis of SCADA data [J]. *Renewable Energy*, 2018(116): 107-122.
- [6] BORCHERSEN A B, KINNAERT M. Model-based fault detection for generator cooling system in wind turbines using SCADA data [J]. *Wind Energy*, 2016, 19(4): 593-606.
- [7] CAO M, QIU Y, FENG Y, et al. Study of wind turbine fault diagnosis based on unscented Kalman filter and SCADA data [J]. *Energies*, 2016, 9(10): 847.
- [8] MITRA S, KOLEY C. An automated SCADA based system for identification of induction motor bearing fault used in process control operation [C]. 2nd International Conference on Control, Instrumentation, Energy & Communication (CIEC), IEEE, 2016: 294-298.
- [9] ZHANG M, WANG T, TANG T, et al. An imbalance fault detection method based on data normalization and EMD for marine current turbines [J]. *ISA transactions*, 2017(68): 302-312.
- [10] LI Z, CHEN J, ZI Y, et al. Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of high speed locomotive [J]. *Mechanical Systems and Signal Processing*, 2017(85): 512-529.
- [11] HANG J, ZHANG J, CHENG M. Application of multi-class fuzzy support vector machine classifier for fault diagnosis of wind turbine [J]. *Fuzzy Sets and Systems*, 2016, 297: 128-140.
- [12] BANGALORE P, PATRIKSSON M. Analysis of SCADA data for early fault detection, with application to the maintenance management of wind turbines [J]. *Renewable Energy*, 2018(115): 521-532.
- [13] XIN W, LIANCHENG S. Wind turbine modeling research based on the combination of SCADA and vibration signals [C]. 2017 4th International Conference on Information Science and Control Engineering (ICISCE), IEEE, 2017: 1331-1335.
- [14] ITAKURA F. Minimum prediction residual principle applied to speech recognition [J]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1975, 23(1): 67-72.
- [15] HEJUN Z, LIEHUANG Z. Encrypted network behaviors identification based on dynamic time warping and k-nearest neighbor [J]. *Cluster Computing*, 2017, doi: 10.1007/s10586-017-1329-y.
- [16] 白敬, 宋爱国, 李会军, 等. 基于工作空间测量的居家脑卒中患者上肢康复训练评估系统 [J]. *仪器仪表学报*, 2018, 39(4): 74-81.
- BAI J, SONG A G, LI H J, et al. Upper extremity rehabilitation training and assessment system for home stroke patients based on work [J]. *Chinese Journal of Scientific Instrument*, 2018, 39(4): 74-81.
- [17] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. *Annals of Statistics*, 2001, 29(5): 1189-1232.
- [18] REZVANBEHBAHANI S, STEARNS L A, KADIVAR A, et al. Predicting the geothermal heat flux in Greenland: A machine learning approach [J]. *Geophysical Research Letters*, 2017, doi: 10.1002/2017gl075661.
- [19] JOHNSON N E, IANIUK O, CAZAP D, et al. Patterns of waste generation: a gradient boosting model for short-term waste prediction in New York city [J]. *Waste Management*, 2017(62): 3-11.
- [20] SUN X, HUANG Z, PENG X, et al. Building a model-based personalised recommendation approach for tourist attractions from geotagged social media data [J]. *International Journal of Digital Earth*, 2019, 12(6): 661-678.
- [21] DIVINA F, GILSON A, GOMÉZ-VELA F, et al. Stacking ensemble learning for short-term electricity consumption forecasting [J]. *Energies*, 2018, 11(4): 949.
- [22] MELZI F N, SAME A, ZAYANI M H, et al. A dedicated mixture model for clustering smart meter data: identification and analysis of electricity consumption behaviors [J]. *Energies*, 2017, 10(10): 1446.
- [23] NTALAMPIRAS S, SOUPIONIS Y. Gaussian mixture modeling for detecting integrity attacks in smart grids [J]. *Electronics*, 2016, 5(4): 82.
- [24] 张龙, 黄文艺, 熊国良, 等. 基于 TESPAR 与 GMM 的滚动轴承性能退化评估 [J]. *仪器仪表学报*, 2014, 35(8): 1772-1779.
- ZHANG L, HUANG W Y, XIONG G L, et al. Bearing performance degradation assessment based on TESPAR

and GMM [J]. Chinese Journal of Scientific Instrument, 2014, 35(8): 1772-1779.

[25] CHEN D, LI G, SUN Y, et al. An interactive image segmentation method in hand gesture recognition [J]. Sensors, 2017, 17(2): 253.

[26] GROSS K C, HUMENIK K E. Sequential probability ratio test for nuclear plant component surveillance [J]. Nuclear Technology, 1991, 93(2): 131-137.

作者简介



刘帅,2013 年于华北电力大学获得学士学位,现为华北电力大学博士研究生,主要研究方向为风电机组故障预警。
E-mail:liushuai@ncepu.edu.cn

Liu Shuai received his B.Sc. degree from North China Electric Power University in 2013. Now he is a Ph. D. candidate in North China Electric Power University. His main research interest is wind turbine fault warning.



刘长良, 分别在 1985 年、1990 年和 2002 年于华北电力学院获得学士学位、硕士学位和博士学位,现为华北电力大学教授,主要研究方向为风电机组故障诊断、火电机组节能、热力系统建模与仿真等。

Liu Changliang received his B.Sc. , M.Sc. and Ph.D. degrees all from North China Electric Power University in 1985, 1990 and 2002, respectively. Now he is a professor in North China Electric Power University. His main research interests include wind turbine fault diagnosis, energy saving of thermal power units, modeling and simulation of thermal system.