基于 MADM-QM 的风电机组风功率异常数据处理方法

莫丰源[1]，王卫华[2,3]，郭 前[3]
(1.湖南科技大学 机电工程学院，湖南 湘潭 411201；2.苏州城市学院 实验室建设与管理中心，江苏 苏州 215104；3.苏州城市学院 智能制造与智慧交通学院，江苏 苏州 215104)

摘 要：针对风电机组非正常运行时导致远程中央监控与数据采集(SCADA)系统所采集的风速-功率数据中存在大量的横向、纵向分布的异常值问题，文章提出了一种基于中值绝对偏差法(MADM)和四分位法(QM)的异常数据清洗方法，即 MADM-QM 算法。首先，基于风速-桨距角关系模型，通过对风速区间的风速-桨距角数据集中绝对中位差(MAD)的求解，清洗掉±4.5MAD 外的风速-桨距角数据；然后，基于风速-功率关系模型，先对功率区间的风速-功率数据集中异常值进行剔除，再对风速区间的风速-功率数据集中异常值进行剔除，完成异常数据的清洗；最后，以某风电场复杂工况下风电机组的实际运行数据为算例进行验证，并与 MADM，QM 和基于密度的空间聚类(DBSCAN)法进行对比分析。结果表明，MADM-QM 算法不仅能够有效识别异常数据，而且能够高效完成异常数据清洗，相比其他 3 种方法，MADM-QM 算法处理异常数据效率良好且清洗质量最优。

关键词：风电机组；风功率；数据清洗；MADM-QM；SCADA 数据
中图分类号：TK81 **文献标志码**：A **文章编号**：1671-5292(2025)03-0339-07

0 引言

风能作为一种可再生能源，具有环保、储备潜力大的优势。风电机组是将风能转化为电能的关键装备[1]，通常安装在环境复杂的偏远山区以及近海或者海上区域，人工监测日常运行状态工作显得非常困难。而风电机组故障发生率较高，因此，风电机组通常安装远程中央监控与数据采集(SCADA)系统来对风电机组运行状况进行远程监控。

实际发电过程中，由于电力系统消纳能力有限，导致风电机组在工作过程中不得不舍弃一部分风能，因此 SCADA 系统所记录的风速、功率数据中存在大量异常数据[2]，这些数据具有横、纵向分散分布的特点，严重影响对风电机组运行规律的分析[3]、故障诊断[4-6]、发电性能评估[7-9]等。除此之外，风速和功率预测也需要高质量的历史数据[10,11]。

已有不少学者对风速-功率数据预处理方法进行了较深入的研究，文献[12]通过数据范围检查和一致性检验来清洗异常数据，但这种方法对特别分散的异常数据清洗效果不好。文献[13]使用四分位法(QM)对风速-功率数据进行清洗，而当存在大量异常数据时，单一的 QM 清洗效果不佳。文献[14]使用 k 最近邻算法来清洗异常数据，该算法原理简单，但需要大量的正常数据进行训练，且存在学习速度慢和泛化能力差的问题。文献[15,16]利用聚类算法对异常数据进行清洗，取得了较好效果，但聚类算法容易受到模型自定义参数的影响。文献[17]利用图像处理技术清洗异常数据，但这种方法实现难度很大，不适用于工程实践。

在现有研究的基础上，本文基于风速-桨距角-功率的关系模型，利用中值绝对偏差法(MADM)的鲁棒性与 QM 较好的数据描述性，并将二者结合，提出了一种新颖的组合方法 MADM-QM 算法，利用其对风功率中的异常数据进行识别和清洗。

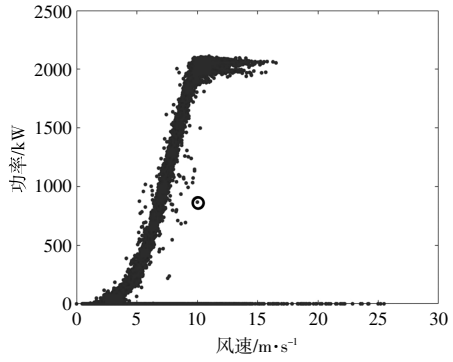
1 问题描述

风是一个随机场，风速和风向会随机变化，并且具有变化频率高、变化幅度大的特点。风电机组容易受到恶劣天气(如暴风雨、强降雪等)或者自身故障的影响，从而导致 SCADA 系统收集的秒级数据波动幅度极大[18]。基于测量的风速和功率

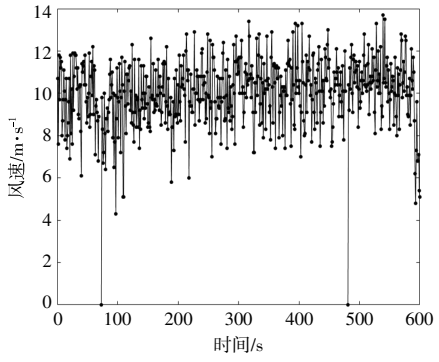
收稿日期：2023-09-09。
基金项目：江苏省高等学校基础科学(自然科学)研究面上项目(23KJB510026)。
通信作者：王卫华(1986-)，男，硕士，讲师，研究方向为电力装备数据分析与处理、机电系统动态行为及其控制。E-mail:bjorker@163.com

数据所形成的风功率曲线通常用于监测风电机组的发电性能以及故障诊断,而在工程应用中,步长为 10 min 等级的 SCADA 数据被认为是风电机组功率曲线评估的最优选择[19]。由于风的不稳定性,导致 10 min 等级的风电功率数据出现大量的异常数据[20],这会造成风电机组出现故障误报、性能评估不准确、风速功率预测精度低等问题。因此,需要一种能够有效识别并过滤掉风电功率数据中各类异常数据的方法,以高效清洗异常的风速-功率数据。

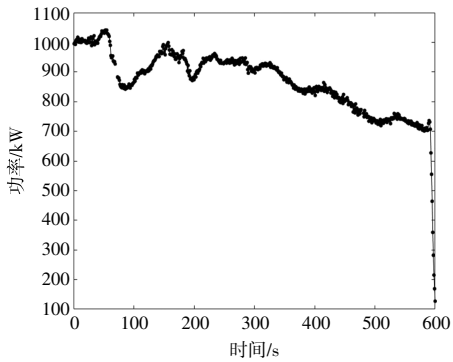
图 1 为风电机组的风功率异常点示意图。



(a)10 min 等级风速-功率散点



(b)(a)图圈中点的 1 s 等级风速数据



(c)(a)图圈中点的 1 s 等级功率数据

图 1 风电机组的风功率异常点示意图

Fig.1 Diagram of wind power anomaly of wind turbine

图 1(b)的风速在 0~14 m/s 内波动,图 1(c)的功率从 1 000 kW 下降到 100 kW,说明图 1(a)圈中点的 1 s 级风速和功率数据在这 10 min 内的变化起伏很大,导致数据经过 10 min 平均压缩处理后偏离整体功率曲线,这种点被认为是须要过滤或清洗掉的异常点。

2 数据清洗方法

2.1 中值绝对偏差法

绝对中位差(MAD)是一种鲁棒统计量[21],而 MADM 常被用于异常值检测。假定存在数据集 $X=[X_1,X_2,X_3,\cdots,X_n]$,该数据集的 MAD 的计算式为

$$S_{\mathrm{MAD}}=\mathrm{median}(|X_i-X_m|) \tag{1}$$

式中:X_i 为数据集中第 i 个数据;X_m 为数据集的中位数。

根据 MAD,可确定数据集 X 异常值的内限为

$$[M_1,M_2]=nS_{\mathrm{MAD}} \quad n\neq 0 \tag{2}$$

超出$[M_1,M_2]$的值均为异常值。

2.2 QM

QM 的关键是找到数据集中的四分位数[22],四分位数可以将一个排好序的数据集平均划分为 4 个部分,每部分所包含的数据量是整个数据集数据量的 1/4。QM 的具体计算步骤如下。

①计算第二个四分位数 Q_2

$$Q_2=\begin{cases} x_{\frac{n+1}{2}} & n=2k+1;k=0,1,2,3,\cdots \\ \dfrac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2} & n=2k;k=0,1,2,3,\cdots \end{cases} \tag{3}$$

②计算第一个四分位数 Q_1 和第三个四分位数 Q_3

当 $n=2k$ 时,将数据集从 Q_2 位置分为两个子数据集,然后分别求解这两个子数据集的中位数 Q_{m1},Q_{m2},其中 $Q_{\mathrm{m1}}<Q_{\mathrm{m2}}$,则:

$$\begin{cases} Q_1=Q_{\mathrm{m1}} \\ Q_3=Q_{\mathrm{m2}} \end{cases} \tag{4}$$

当 $n=4k+1$ 时,则:

$$\begin{cases} Q_1=0.25x_k+0.75x_{k+1} \\ Q_3=0.75x_{3k+1}+0.25x_{3k+2} \end{cases} \tag{5}$$

当 $n=4k+3$ 时,则:

$$\begin{cases} Q_1=0.75x_{k+1}+0.25x_{k+2} \\ Q_3=0.25x_{3k+2}+0.75x_{3k+3} \end{cases} \tag{6}$$

由式(4)~(6)所求出的 Q_1,Q_3 可得四分位距(I_{QR})为 Q_3-Q_1。根据 I_{QR},数据集的异常值内限可以定义为

$$[F_1,F_2]=[Q_1-1.5I_{QR},Q_3+1.5I_{QR}] \quad (7)$$

超出[F_1,F_2]的数据均被判定为异常值。

2.3　风电功率数据清洗方法

基于风功率数据中异常数据的分布特点,本文提出了一种用于异常数据清洗的方法,即 MADM-QM 方法,流程如图 2 所示。

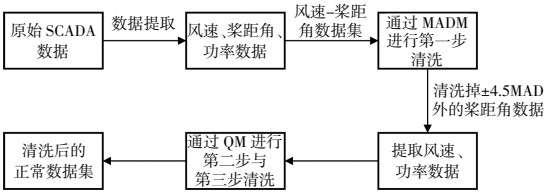


图 2　数据清洗流程

Fig.2 The flow chart of cleaning abnormal data

首先,从原始 SCADA 数据中提取出风速、桨距角、功率数据;然后,采用 MADM 进行第一步数据清洗;最后,采用 QM 进行第二、三步清洗,以获得正常数据集。具体流程如下。

①采用 MADM 进行第一步清洗

基于风速-桨距角组成的关系模型 $Y=f(x)$(Y 为桨距角;f(·) 为风速与桨距角之间的函数关系),形成风速-桨距角数据集。将风速按照升序的方式进行排序,以 ΔV 作为一个风速区间对数据集进行划分,将风速-桨距角数据集划分成 n 个子数据集(图 3),应用式(1),(2)计算每个子数据集中桨距角数据的 MAD,清洗掉±4.5MAD 外的风速-桨距角数据[21]。

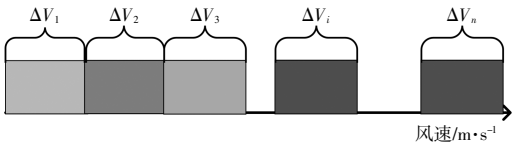


图 3　风速区间划分示意图

Fig.3 Schematic diagram of wind speed interval

②采用 QM 进行第二、三步清洗

基于风速-功率组成的关系模型 $P=q(y)$(P 为功率;q(·)为风速与功率之间的函数关系),形成风速-功率数据集。

第二步清洗:首先按照功率的大小进行递增式排序,并以 ΔP 作为一个功率区间进行划分(图 4);然后,在每个区间按照风速的大小重新进行递增式排序;最后,采用 QM 对异常数据进行清洗。

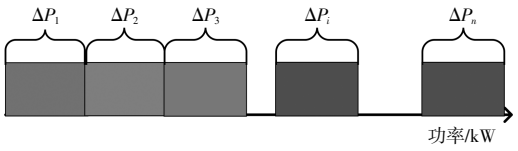


图 4　功率区间划分示意图

Fig.4 Schematic diagram of wind power interval

第三步清洗:首先按照风速的大小进行递增式排序,并以 ΔV 作为一个风速区间进行划分;然后,对每个区间按照功率的大小重新进行递增式排序;最后,采用 QM 对异常数据进行清洗。

3　实例验证

为验证 MADM-QM 方法的有效性,以某风电场两台 2 MW 风电机组为对象,分别选取不同年份两台风电机组的两组 SCADA 数据,风电机组的切入风速、额定风速和切出风速分别为 3,11 m/s 和 25 m/s。算例中,均选用 Intel(R)Xeon(R)Gold 6138 处理器,内存为 256 GB,64 位的服务器,以 Python 为算法语言工具。

3.1　案例 1

案例 1 中,风电机组的 SCADA 数据采集时间为 2018 年 01-12 月,数据间隔为 10 min 等级,提取总数为 51 047 组的风速、桨距角和功率数据。

3.1.1　实例验证

图 5 为风电机组 10 min 级风速-功率散点图。

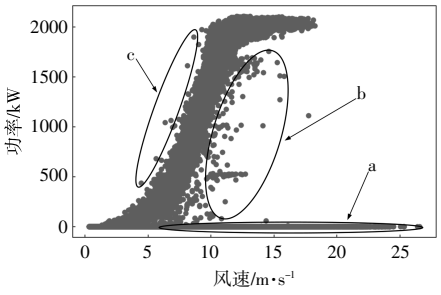


图 5　原始 10 min 级风速-功率散点图

Fig.5 Wind speed-power scatter plot on level of original 10 minutes

由图 5 可知,异常数据主要有 a,b,c 3 类。a 类为风速很大,但功率为零或负值的数据;b 类为风速较大、功率较小或低于额定功率的数据,即弃风数据[23],亦或是故障数据;c 类为风速很小,但功率很大的数据。

根据图 2 数据清洗的步骤,先采用 MADM 进行第一步数据清洗,设定 ΔV 为 0.5 m/s,清洗掉每个 ΔV 步距区间中±4.5MAD 外的风速、桨距角散点,如图 6 中的三角形点。

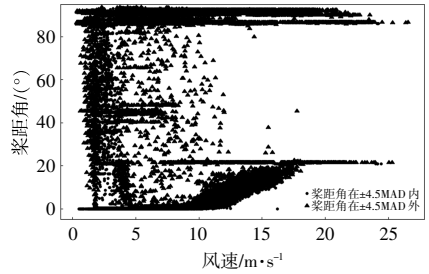


图 6 风速–桨距角散点图

Fig.6 Wind speed–pitch angle scatter plot

图 7 为经过 MADM 清洗后的风速–功率散点图。

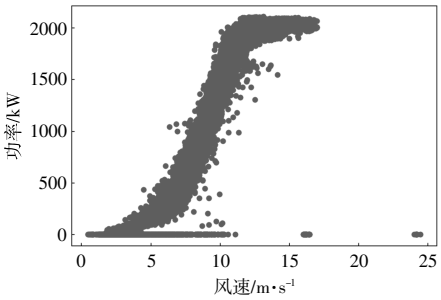


图 7 经过 MADM 清洗后的风速–功率散点图

Fig.7 Wind speed–power scatter plot after cleaning anomaly by MADM

由图 7 可知,经过 MADM 清洗后,已经清洗掉大量的 a 类异常数据,但只过滤掉了小部分 b,c 类异常数据,故须要采用 QM 进行第二步与第三步清洗。

在第二步与第三步清洗前,先设定 ΔV 为 0.5 m/s,ΔP 为 100 kW。图 8 为采用 MADM–QM 方法对原始的风速、功率数据进行清洗后得到的风速–功率散点图。由图 8 可知,a,b,c 3 类异常数据均已经消失。

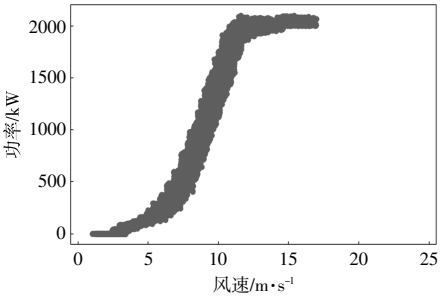


图 8 经过 MADM–QM 清洗后的风速–功率散点图

Fig.8 Wind speed–power scatter plot after cleaning anomaly by MADM–QM

图 9,10 分别为经过 QM 和 DBSCAN 清洗后的风速–功率散点图。

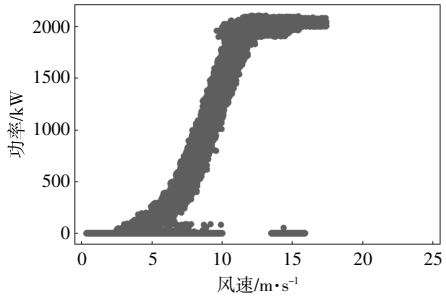


图 9 经过 QM 清洗后的风速–功率散点图

Fig.9 Wind speed–power scatter plot after cleaning anomaly by QM

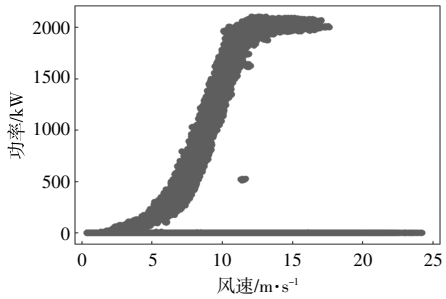


图 10 经过 DBSCAN 清洗后的风速–功率散点图

Fig.10 Wind speed–power scatter plot after cleaning anomaly by DBSCAN

由图 7~10 可知:当风速为 0~3 m/s 时,这 4 种算法均达到良好的清洗效果;当风速为 3~11 m/s 时,只有 MADM–QM 能够将 a,b,c 类异常数据完全清洗干净,MADM,QM,DBSCAN 均未能将大量的 a 类异常数据清洗掉,QM,DBSCAN 能够清洗大量的 b,c 类异常数据,但还有小部分未能准确清洗,MADM 对 b,c 类异常数据的清洗效果是最差的;当风速为 11~25 m/s 时,MADM,QM,DBSCAN 均未能准确清洗掉 a,b 类异常数据点,而 MADM–QM 的清洗效果最好。

3.1.2 算法对比

利用 Spearman 相关系数和算法运行时间对 MADM–QM 方法与 MADM,QM,DBSCAN 3 种常用的数据清洗方法进行对比,结果如表 1 所示。其中 Spearman 相关系数越接近于 1,代表清洗效果越好[24]。

表 1　采用 MADM,QM,DBSCAN 和 MADM-QM 方法的异常数据清洗结果

Table 1 The results of cleaning anomaly by MADM,QM, DBSCAN and MADM-QM

清洗方法	Spearman 相关系数	算法运行时间/s
MADM	0.973 8	1.369 8
QM	0.742 7	1.908 6
DBSCAN	0.571 4	386.136 7
MADM-QM	0.993 3	2.801 8

由表 1 可知:MADM-QM 不仅比 DBSCAN 的 Spearman 相关系数高,并且在算法运行时间上也远优于 DBSCAN;MADM-QM 的运行时间比 QM 慢 1 s 左右，但 MADM-QM 的清洗效果比 QM 提升了约 0.25; MADM-QM 的运行时间比 MADM 慢 1.5 s 左右，但 MADM-QM 的清洗效果比 MADM提升了 0.02。

3.2　案例 2

风电机组的 SCADA 数据采集时间为 2019 年 08-12 月,数据间隔为 10 min 等级,提取总数为 14 523 组的风速、桨距角、功率数据。风电机组 10 min 级风速-功率散点如图 11 所示。

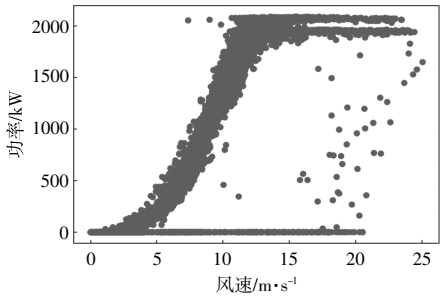


图 11　原始 10 min 级风速-功率散点图

Fig.11 Wind speed-power scatter plot on level of original 10 minutes

3.2.1　实例验证

根据前文所述的数据清洗方法，先采用 MADM 进行第一步清洗后(图 12),仍有稀少的 b, c 类异常数据点存在，采用 QM 进行第二步和第三步清洗后，原始数据中的所有异常数据均消失了(图 13)。

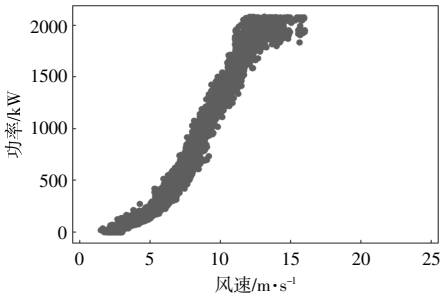


图 12　经过 MADM 清洗后的风速-功率散点图

Fig.12 Wind speed-power scatter plot after cleaning anomaly by MADM

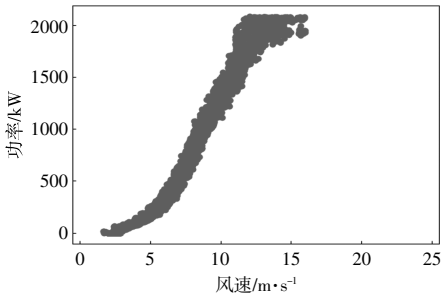


图 13　经过 MADM-QM 清洗后的风速-功率散点图

Fig.13 Wind speed-power scatter plot after cleaning anomaly by MADM-QM

3.2.2　算法对比

利用 Spearman 相关系数和算法运行时间对本文所提出的方法与 MADM,QM,DBSCAN 3 种常用的数据清洗方法对进行比,结果如表 2 所示。

表 2　采用 MADM,QM,DBSCAN 和 MADM-QM 方法的异常数据清洗结果

Table 2 The results of cleaning anomaly by MADM,QM, DBSCAN and MADM-QM

清洗方法	Spearman 相关系数	算法运行时间/s
MADM	0.992 9	0.610 1
QM	0.987 0	0.922 2
DBSCAN	0.934 2	37.852 7
MADM-QM	0.995 6	1.016 5

由表 2 可知：在运行时间方面,MADM-QM 远胜过 DBSCAN,但略慢于 MADM 和 QM;在数据清洗效果方面,MADM-QM 的清洗效果略优于其他 3 种方法。

图 14,15 分别为经过 QM 和 DBSCAN 清洗后的风速-功率散点图。

由图 12~15 可知:当风速为 0~3 m/s 时,4 种

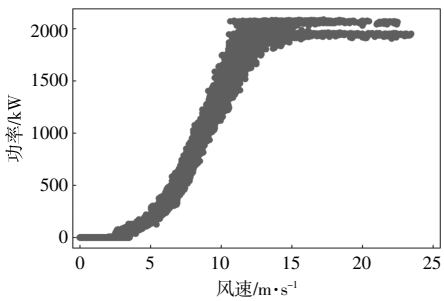


图 14　经过 QM 清洗后的风速-功率散点图

Fig.14 Wind speed-power scatter plot after cleaning anomaly by QM

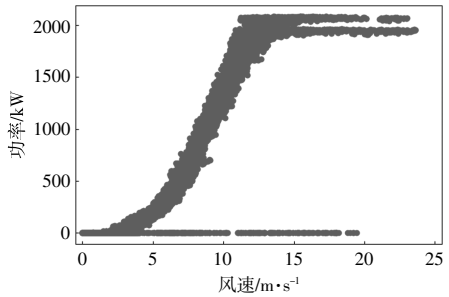


图 15 经过 DBSCAN 清洗后的风速-功率散点图

Fig.15 Wind speed-power scatter plot after cleaning anomaly by DBSCAN

算法的清洗效果均较为良好；当风速为 3~11 m/s 时,MADM,QM,DBSCAN 进行数据清洗后还存在少量的 b,c 类异常数据,且 DBSCAN 算法不能有效清洗 a 类异常数据，而 MADM-QM 对 a,b,c 3 类异常数据的清洗效果是最佳的；当风速为 11~25 m/s 时,DBSCAN 依然不能清洗掉 a 类异常数据，而这 4 种方法对于 b,c 类异常数据的清洗均具有良好效果。

案例 1,2 的实例验证与算法对比表明，相较于其他 3 种方法,采用 MADM-QM 方法可以有效识别并高效且稳定地清洗来自 SCADA 系统的原始风速-功率数据中的异常数据。因此，采用 MADM-QM 方法对风速-功率数据进行清洗具有合理性和可行性。

4 结论

本文提出了一种将 MADM 与 QM 相结合的风功率异常数据清洗方法，并将其应用到实际风场运行的原始 SCADA 数据上,得到以下结论。

①MADM-QM 方法能够对风电机组风速-功率数据集中的异常数据进行有效识别并加以清除，将该方法应用于风电机组异常数据处理是可行的。

②相较于 QM,MADM,DBSCAN 方法，MADM-QM 对风电机组风速-功率数据集中的异常数据的清洗质量最优。

③在清洗效率方面,MADM-QM 方法比 DBSCAN 方法更高，接近于 QM,MADM 两种方法。

参考文献：

[1] Dai J C,Liu D S,Wen L,et al.Research on power coefficient of wind turbines based on SCADA data[J]. Renewable Energy,2016,86:206-215.

[2] Liang G Y,Su Y H,Wu X Y,et al.Abnormal data cleaning for wind turbines by image segmentation based on active shape model and class uncertainty [J]. Renewable Energy,2023,216:118965.

[3] 杨茂,翟冠强,苏欣.基于风特征分析的风电机组异常数据识别算法[J].中国电机工程学报,2017,37(S1):144-151.

[4] 李刚,齐莹,李银强,等.风力发电机组故障诊断与状态预测的研究进展[J].电力系统自动化,2021,45(4):180-191.

[5] 彭安群.基于数据挖掘的风力发电机组齿轮箱故障诊断系统研究[D].兰州:兰州理工大学,2012.

[6] 宿忠娥,祁建宏,效迎春.数据挖掘技术在风力发电机组故障诊断中的应用与研究[J].自动化与仪器仪表,2018(2):13-15.

[7] Xiao Z,Zhao Q C,Yang X B,et al.A power performance online assessment method of a wind turbine based on the probabilistic area metric [J].Applied Sciences,2020,10(9):3268.

[8] Zhan J,Wang R L,Yi L Z,et al.Health assessment methods for wind turbines based on power prediction and mahalanobis distance [J].International Journal of Pattern Recognition & Artificial Intelligence,2019,33(2):17.

[9] Zhang F,Wen Z J,Liu D S,et al.Calculation and analysis of wind turbine health monitoring indicators based on the relationships with SCADA data[J].Applied Sciences,2020,10(1):410.

[10] Wang J Z,An Y N,Li Z W,et al.A novel combined forecasting model based on neural networks,deep learning approaches, and multi-objective optimization for short-term wind speed forecasting [J].Energy,2022,251:123960.

[11] 李东升.风力发电机功率组合预测模型研究[D].秦皇岛:燕山大学,2017.

[12] Schlechtingen M,Santos I F,Achiche S.Using data-mining approaches for wind turbine power curve monitoring:A comparative study [J].IEEE Transactions on Sustainable Energy,2013,4(3):671-679.

[13] 朱倩雯,叶林,赵永宁,等.风电场输出功率异常数据识别与重构方法研究[J].电力系统保护与控制,2015,43(3):38-45.

[14] Kusiak A,Zheng H Y,Song Z.Models for monitoring wind farm power-science direct [J].Renewable Energy,2009,34(3):583-590.

[15] Zhu A F,Xiao Z,Zhao Q C.Power data preprocessing method of mountain wind farm based on POT-DBSCAN [J].Energy Engineering,2021,118(3):549-563.
[16] 孔维胜,朱海鹏,王晓东,等.基于 DBSCAN 的风机功率异常数据清洗[J].计算机科学与应用,2021,11(10):2517-2528.
[17] Liang G Y,Su Y H,Chen F,et al.Wind power curve data cleaning by image thresholding based on class uncertainty and shape dissimilarity[J].IEEE Transactions on Sustainable Energy,2021,12(2):1383-1393.
[18] 郭鹏,陈思.基于运行数据的风电机组本地风向波动特性及偏航控制研究[J].太阳能学报,2020,41(6):77-85.
[19] 吴莎, 康慨, 汪健, 等. 用于风机功率曲线评估的 SCADA 数据时间步长研究[J].电气时代,2019(9):54-56.
[20] 封焯文,朱世平,赵志华,等.风功率异常数据检测方法对比研究[J].电工电能新技术,2021,40(7):55-61.
[21] Martin C M S,Lundquist J K,Clifton A,et al.Wind turbine power production and annual energy production depend on atmospheric stability and turbulence[J].Wind Energy Science,2016,1(2):221-236.
[22] Shen X J,Fu X J,Zhou C C.A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm [J].IEEE Transactions on Sustainable Energy,2019,10(1):46-54.
[23] 王新,王政霞.基于改进 bin 算法的风电机组风速-功率数据清洗[J].智能科学与技术学报,2020,2(1):62-71.
[24] 郑玉巧,刘玉涵,何正文,等.基于 QM-DBSCAN 的风力机数据清洗方法 [J]. 兰州理工大学学报,2021,47(6):50-55.

Wind speed-power abnormal data processing method of wind turbine based on MADM-QM

Mo Fengyuan[1], Wang Weihua[2,3], Guo Qian[3]
(1.School of Mechanical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China; 2.Laboratory Development and Management Centre, Suzhou City University, Suzhou 215104, China; 3.School of Intelligent Manufacturing and Smart Transportation, Suzhou City University, Suzhou 215104, China)

Abstract: Aiming at the problem that there are a large number of horizontal or vertical distribution outliers in the wind speed-power data collected by SCADA system when wind turbine is in abnormal operation, an abnormal data processing method based on median absolute deviation method (MADM) and quartile method (QM) is proposed to solve it, namely MADM-QM algorithm. Firstly, based on the relationship model of wind speed-pitch angle, the wind speed-pitch angle data outside of ±4.5 MAD are discarded by solving the median absolute deviation (MAD) in the wind speed-pitch angle data set of the wind speed interval. Secondly, based on the wind speed-power relationship model, the abnormal values in the wind speed-power data set of the power interval are eliminated, and then the abnormal values in the wind speed-power data set of the wind speed interval are eliminated to complete the abnormal data processing. Finally, the actual operation data of wind turbine under complex working conditions of a wind farm are taken as examples for verification, and comparison with MADM, QM and density-based spatial clustering (DBSCAN) method. The results indicate that the proposed method can not only effectively identify abnormal data but also efficiently and stably clean them. Compared with the other three methods, to a certain extent, it proves that MADM-QM can achieve good efficiency of abnormal data processing and optimal cleaning quality on the abnormal data.

Keywords: wind turbine; wind speed-power; data cleaning; MADM-QM; SCADA data