DOI: 10.16452/j.cnki.sdkjzk.2023.06.012 文章编号:1672-3767(2023)06-0106-11

基于滑动四分位和可行搜索圆算法的风速-功率异常数据清洗方法

白星振[1],隋舒婷[2],葛磊蛟[3],朱爱莲[4],赵 康[1],顾志成[3]
(1. 山东科技大学 电气与自动化工程学院,山东 青岛 266590;2. 国网山东省电力公司 阳谷县供电公司,山东 聊城 252300;
3. 天津大学 智能电网教育部重点实验室,天津 300072;4. 青岛龙发热电有限公司,山东 青岛 266317)

摘 要:受气候条件、运行环境等因素影响,风电场 SCADA 系统所采集的原始风速和风电功率数据常存在大量奇异点,较难反映风电机组真实性能。本研究提出一种基于滑动四分位和可行搜索圆算法的风速-功率异常数据清洗方法。首先,分析了原始数据的时序特征和异常数据分布特点,将数据分为分散型异常数据和堆积型异常数据两类;然后,运用滑动四分位算法实现了分散型异常数据的识别,提出可行搜索圆(FSC)算法,剔除堆积型异常数据,获得符合风电机组出力特性的数据主带;最后,以我国北方某风电机组实际运行数据为例验证,表明本研究方法能较好地识别异常数据,与传统方法相比,本方法清洗效率高、效果好,具有一定的通用性。

关键词:风速-功率;数据清洗;机组出力;滑动四分位算法;可行搜索圆算法
中图分类号:TP711 **文献标志码**:A

Wind speed-power abnormal data cleaning method based on the algorithm of sliding quartile and feasible search circle

BAI Xingzhen[1], SUI Shuting[2], GE Leijiao[3], ZHU Ailian[4], ZHAO Kang[1], GU Zhicheng[3]
(1. College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China; 2. Shandong Power Supply Company, Yanggu Power Supply Company, Liaocheng 252300, China;
3. Key Laboratory of Smart Grid of Ministry of Education, Tianjin University, Tianjin 300072, China;
4. Qingdao Longfa Thermal Power Company Limited, Qingdao 266317, China)

Abstract: Affected by climatic conditions, operating environments and other factors, the original wind speed and power data collected by the SCADA system for wind power farms can hardly reflect the actual performance of the wind turbine for the existence of many singular points. A wind speed-power abnormal data cleaning method based on the algorithm of sliding quartile and feasible search circle was proposed in this paper. Firstly, the timing characteristics of the raw data and the distribution characteristics of the abnormal data were analyzed. The data was divided into two types: scattered anomaly data and stacked anomaly data. Then, the moving quartile algorithm was used to achieve the identification of the decentralized abnormal data. A feasible search circle (FSC) algorithm was further proposed to eliminate the accumulated abnormal data. And the main data band meeting the wind turbine output characteristics was obtained. Finally, the actual operation data of a wind turbine in northern China was taken as an example to verify the method proposed in this paper. The results show that the proposed method can better identify

收稿日期:2021-10-26
基金项目:国家自然科学基金项目(51807134);省部共建电工装备可靠性与智能化国家重点实验室(河北工业大学)开放基金项目(EERI_KF20200014);山东省自然科学基金项目(ZR2020MF071)
作者简介:白星振(1977—),男,山东临朐人,副教授,博士,主要从事电网状态估计方面的研究工作. E-mail:xzbai@163.com
葛磊蛟(1984—),男,天津人,副教授,博士,主要从事智能配电网态势感知、云计算和大数据方面的研究工作,本文通信作者. E-mail:legendglj99@tju.edu.cn

abnormal data. Compared with traditional methods, the proposed algorithm has certain generality for its high efficiency and excellent results in cleaning data.

Key words: wind speed-power; data cleaning; unit output; sliding quartile algorithm; feasible search circle algorithm

风能作为一种环境友好且经济实用的可再生能源,是国家"碳达峰、碳中和"的主力军[1-3]。风电场建成投运后,根据实际运行条件下风速和功率数据得到的风电机组出力曲线不仅能够表征风电机组的实际运行情况,也能够衡量机组的出力水平[4]。良好的风电机组出力曲线对风速和功率数据的质量有很高要求,但受地理条件、运行环境、弃风限电等多因素影响,风电场数据采集与监视控制(supervisory control and data acquisition,SCADA)系统实时采集的风电场运行数据常存在较多的奇异数据,严重影响了风速和功率本应有的整体分布规律和对应关系,无法直接用于风电机组的性能分析和风电场调度管理,因此对采集的风电场运行数据进行识别和清洗[4-5]非常必要。

目前,针对风速-功率异常数据识别和清洗方面,国内外学者进行了一些研究。已有的数据清洗算法主要存在三类问题。一是奇异数据的漏检,如文献[4]采用最优组内方差算法识别异常数据,但仅能识别位于功率曲线下方的堆积型异常数据,且需要大量的迭代计算;文献[5]采用基于密度的局部离群因子算法,能有效识别分散型异常数据,但不能识别成堆出现的异常数据。二是在识别异常数据的同时删除了大量正常运行的数据,不能保证正常运行数据的完整性,如文献[6]采用四分位法和 k-means 算法剔除异常数据,但没有给出具体的剔除标准,会导致正常运行数据被误删,而且对于不同的弃风数据簇,聚类算法中 k 值的选取目前没有统一的标准;文献[7]基于 Copula 理论建立了风电机组的等效置信边界模型,但需要多台风机运行数据,数据需求量大,且数据删除率高,常出现误删。三是清洗效率低下,如文献[8]建立非线性模型判别异常数据,但清洗效率低且对样本数据的需求量大;文献[9]采用四分位法和基于密度的聚类算法确定出越限风功率数据的识别边界,但算法本身存在较多控制参数且数据处理速度较慢。

综上,现有的数据清洗算法主要存在数据删除率高、异常数据漏检、清洗效率低等问题。本研究提出一种基于滑动四分位和可行搜索圆(feasible search circle,FSC)算法的异常数据清洗方法,该方法首先删除原始风速-功率数据中功率小于等于零的数据,然后利用滑动四分位算法清洗掉位于风功率曲线周围的功率散点,最后利用 FSC 算法完成分布密集的异常数据簇的识别和清洗。实验结果表明,本研究所提的基于滑动四分位和 FSC 算法的清洗方法能够实现多类型异常数据的识别和清洗,具有运算简单、清洗时间短、数据删除率低的优点。

1 风速-功率历史数据分析

标准风功率曲线是在标准工况下根据风电机组设计参数计算得出[10],较难体现风电机组的实际运行状况。实际风电场运行环境下,风电机组 SCADA 系统采集到的实际风功率曲线数据主带周围分布着大量的异常数据点,除了由系统误差和随机误差等原因产生的少量散点之外,大多数异常数据点是由机组非正常运行引起的。这些数据点一般不能用于功率曲线建立和机组性能分析,必须进行针对性剔除。异常数据可根据其空间分布特征分为两大类:分散型异常数据和堆积型异常数据[17]。图 1 为我国北方某个风电机组一年内的实测风速-功率曲线与标准功率曲线对比图,图 1 中(1)、(2)处为分散型异常数据,(3)、

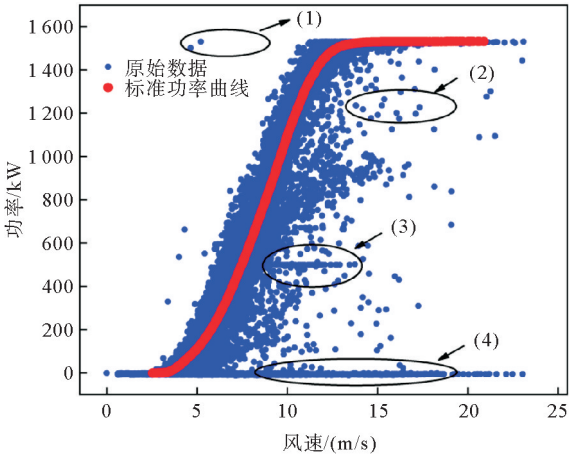


图 1 实测风速-功率散点图

Fig. 1 Scatter diagram of measured wind speed-power

(4)处为堆积型异常数据。剔除上述两大类异常数据点之后,保留的数据主带是机组的常规发电状态数据,能较好地反映风电机组的正常工作范围和出力特性。因此,本研究的目的是实现异常数据的有效清洗。

2 异常数据的识别和分析

异常数据产生的原因及其在风功率曲线中存在的形式多样,目前尚无一个算法实现多类型异常数据的完全识别和清洗,因此本研究根据异常数据的空间分布特点,采用滑动四分位算法处理分散型异常数据,采用 FSC 算法处理堆积型异常数据。

2.1 滑动四分位法

四分位算法[9]通常用于处理分散型异常数据,该算法需要提前进行数据分区且受异常数据量的影响较大,一般不能单独使用。本研究采用的滑动四分位法将滑动窗口[11]与四分位法结合。设原始数据集为:

$$W=\{(v_1,p_1),(v_2,p_2),\cdots,(v_n,p_n)\}。\tag{1}$$

式中:v_i 和 p_i 分别表示第 i 个数据样本的风速和功率值,数据按照功率升序排列,即 $p_i<p_{i+1}$,$i\in(1,n-1)$。设置窗口宽度为 L,窗口每次移动的步长为 m,取第一个数据点作为滑动窗口的起点,窗口框住长度为 L 的数据段,对窗口内的数据用四分位算法进行处理,同时窗口在数据集上顺序滑动。记第 i 个窗口内的风速数据为 $v=\{v_{i,1},v_{i,2},\cdots,v_{i,L}\}$,四分位法的算法步骤如下。

1) 计算样本的第 2 分位数即中位数 Q_2。

$$Q_2=\begin{cases}v_{i,\frac{L+1}{2}}, & L\text{ 为奇数};\\ (v_{i,\frac{L}{2}}+v_{i,\frac{L+2}{2}})/2, & L\text{ 为偶数}。\end{cases}\tag{2}$$

2) 计算第 1、3 个分位数 Q_1 和 Q_3。Q_2 将 v 分为两部分(Q_2 不包含在两部分数据之内),分别计算两部分的中位数 Q_1、Q_3。当 $L=4k+3(k=0,1,2,\cdots)$时,有

$$\begin{cases}Q_1=0.75v_{i,k+1}+0.25v_{i,k+2},\\ Q_3=0.25v_{i,3k+2}+0.75v_{i,3k+3}。\end{cases}\tag{3}$$

当 $L=4k+1(k=0,1,2,\cdots)$时,有

$$\begin{cases}Q_1=0.25v_{i,k}+0.75v_{i,k+1},\\ Q_3=0.75v_{i,3k+1}+0.25v_{i,3k+2}。\end{cases}\tag{4}$$

3) 计算第 1 分位数 Q_1 和第 3 分位数 Q_3 之间的差值 I_{QR},称作四分位距。

$$I_{\mathrm{QR}}=Q_3-Q_1。\tag{5}$$

根据四分位距确定样本中数据的正常值范围为 $[Q_1-1.5I_{\mathrm{QR}},Q_3+1.5I_{\mathrm{QR}}]$,位于此范围之外的数据称为异常值。

滑动四分位算法中不同的 L 和 m 取值对应着不同的数据处理精度,通过窗口宽度和步长的设置,可以保证一次处理的数据量,灵活改变算法的精度,可较好地实现风功率曲线分散型异常数据点的剔除。

2.2 可行搜索圆法

经过滑动四分位算法初步识别之后,大部分位于风功率曲线周围的无规律散点被清除,但位于曲线中部的横向堆积型异常数据仍存在,该类数据多是由弃风限电等原因引起,对于此类异常数据,本研究提出可行搜索圆数据清洗算法。

弃风限电是指因当地电网接纳能力不足、风电不稳定等原因造成的部分风机被迫降低出力或暂停出力的现象,弃风限电产生的原因是多方面的[12-14]。当风电场发生弃风限电时,风电机组的输出功率将不再与风速呈对应关系,而是被限制在一定范围内。此时,在同一风速范围内,正常运行时的功率和发生弃风时的功率相差较大,所以通过弃风数据点偏离正常数据点的程度进行识别,故提出 FSC 算法。考虑到正常运行数据存在一定的波动范围,发展轨迹不是单一的,故选用欧氏距离作为评判标准,以搜索圆的半径阈值作为搜索的边界值,搜索圆的圆心和半径确定后,可在搜索圆的范围内对正常数据进行全方位搜索,极大减少了漏判和误判率。设经过四分位算法初步处理之后的数据样本集合为:

$$W_1=\{(v'_1,p'_1),(v'_2,p'_2),\cdots,(v'_n,p'_n)\}。\tag{6}$$

式中:v'_i 和 p'_i 表示第 i 个数据样本的风速和功率值,数据按照风速升序排列,即 $v'_i<v'_{i+1},i\in(1,n-1)$。

设置一个长度为 2 的可变数据窗口 $[(a_1,b_1),(a_2,b_2)]$,将数据集的第一个数据点(风速数值最小的)看作正常数据点,并将其作为窗口的左端点,即:

$$(a_1,b_1)=(v'_1,p'_1)。\tag{7}$$

将数据集的下一个数据点放在该数据窗口内,以窗口的左端点为圆心,阈值 δ 为半径作圆,判断数据集的下一个数据点是否在此圆包含的范围内。若在,则该数据点属于当前的数据窗口,为正常数据点,此时将窗口的左端点取出,用该点作为窗口的新的左端点判断下一个数据点;若不在,则该点距离正常数据点较远,视作异常数据点舍弃,继续判断接下来的数据点。

FSC 算法示意图如图 2 所示,s_i、s_j、s_l 分别是相邻的 3 个数据点。假设 s_i 为正常点,则 $s_i\in D$,D 是长度为 2 的数据窗口,检验 s_j 时,以 s_i 为圆心,δ 为半径作圆,比较 s_i 和 s_j 之间的欧氏距离 d_1 和 δ 之间的大小关系:

$$\begin{cases}s_j\in D, & d_1\leqslant\delta;\\ s_j\notin D, & d_1>\delta;\end{cases}\tag{8}$$

$$d_1=\sqrt{(v'_j-v'_i)^2+(p'_j-p'_i)^2}。\tag{9}$$

进而判断 s_j 可加入窗口。此时,将 s_i 取出,用 s_j 作为窗口的左端点,判断接下来的数据点 s_l。此时,比较数据点 s_j 和 s_l 之间的欧氏距离 d_2 和 δ 之间的大小关系,

$$d_2=\sqrt{(v'_l-v'_j)^2+(p'_l-p'_j)^2}。\tag{10}$$

由图 2 可知,s_l 不可加入窗口,即 s_j 不是弃风数据点,此时将 s_l 舍去,继续用 s_j 判断下一个数据点。

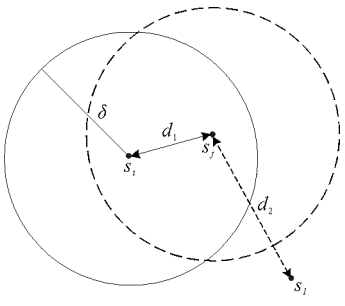


图 2 算法示意图

Fig. 2 Schematic diagram of algorithm

2.3 滑动四分位和 FSC 算法清洗流程

分析风电机组的实际运行情况发现,原始的风功率曲线中位于曲线中部的横向堆积型异常数据十分常见,若直接在竖直方向(沿着风速变化的方向)使用滑动四分位算法进行数据清洗,四分位算法的内限会受到堆积型异常数据的严重影响,不仅无法正确识别异常数据,还会产生一系列新的异常数据点,影响总体的数据识别效果,不利于后续 FSC 算法半径阈值的选取,而在水平方向(沿着功率变化的方向)使用滑动四分位算法可避免此类问题。滑动四分位和 FSC 算法的具体步骤如下:

1) 数据预处理。将原始数据中风速位于切入风速和切出风速之间而功率小于等于零的数据点清除,此类数据在风功率曲线中表现为位于曲线底部的横向密集数据带,如图 1 中的(4)处所示;

2) 在水平方向运用滑动四分位算法消除分散型异常数据。根据曲线周围分散型异常数据的分布情况,定义合适宽度和步长的滑动窗口(选取方法见 2.4 节),对每个窗口中的数据采用四分位算法进行识别清洗,可消除功率曲线中的大部分分散型异常数据,避免了数据散点对后续堆积型异常数据的清洗产生干扰,方便 FSC 算法半径阈值的确定;

3) 对经过滑动四分位算法处理之后的数据用 FSC 算法进行二次处理,将数据按照风速升序顺序排列,将第一个数据点作为初始圆的圆心,依次沿着风速增大的方向进行异常点搜索,直至遍历完所有数据点,即可得出正常数据集。

滑动四分位和 FSC 算法流程如图 3 所示。

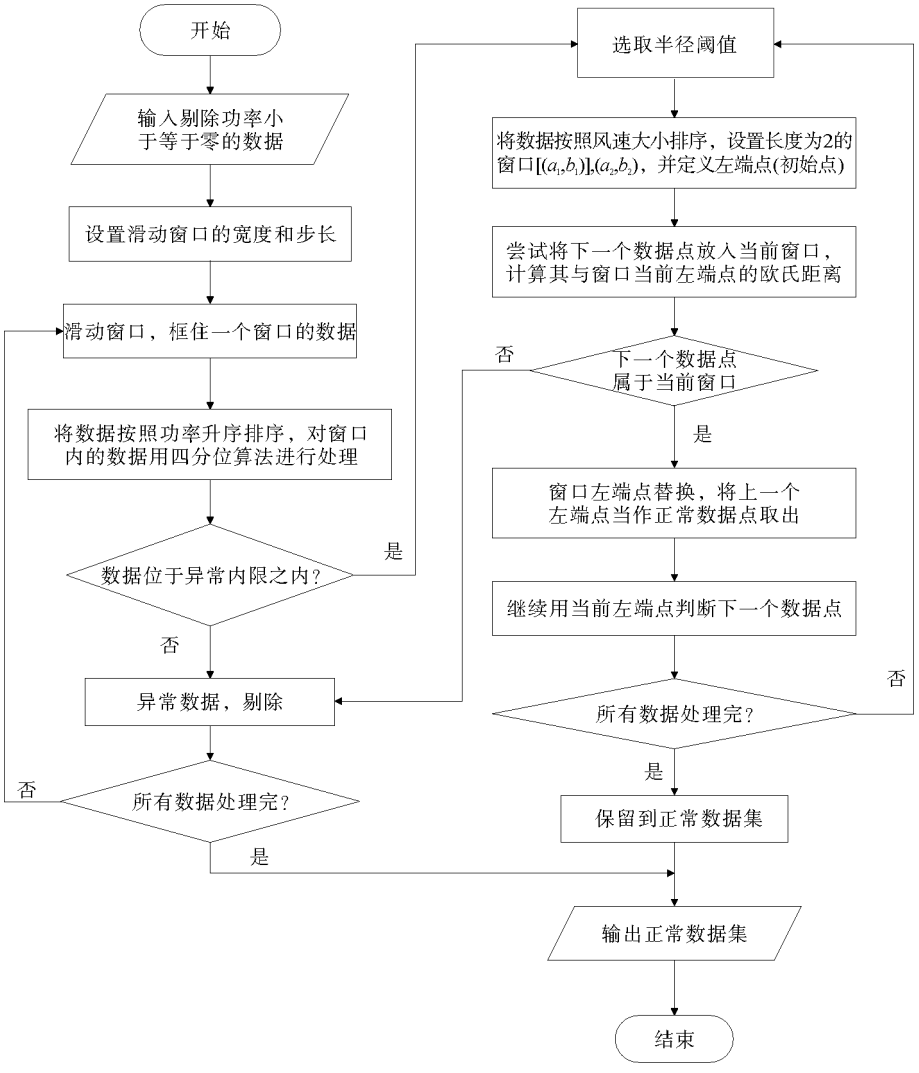


图 3 滑动四分位和 FSC 算法流程图

Fig. 3 Algorithm flow chart of sliding quartile and FSC

2.4 阈值的选取

2.4.1 宽度和步长的选取

滑动四分位算法的滑动窗口区间的宽度和步长的阈值的选取参照式(11)：

$$L = m = \frac{N_{\mathrm{um}}}{N} 。 \tag{11}$$

式中：N_{um} 为功率数据的总数，L 和 m 需为整数，将风功率数据进行 N 等分。

2.4.2 半径阈值的选取

良好的半径阈值应能较完整地识别堆积型异常数据，半径阈值的选取依赖于堆积型异常数据区间集合，堆积型异常数据区间的选取参考文献[6]，FSC 算法半径阈值的选取流程如下。

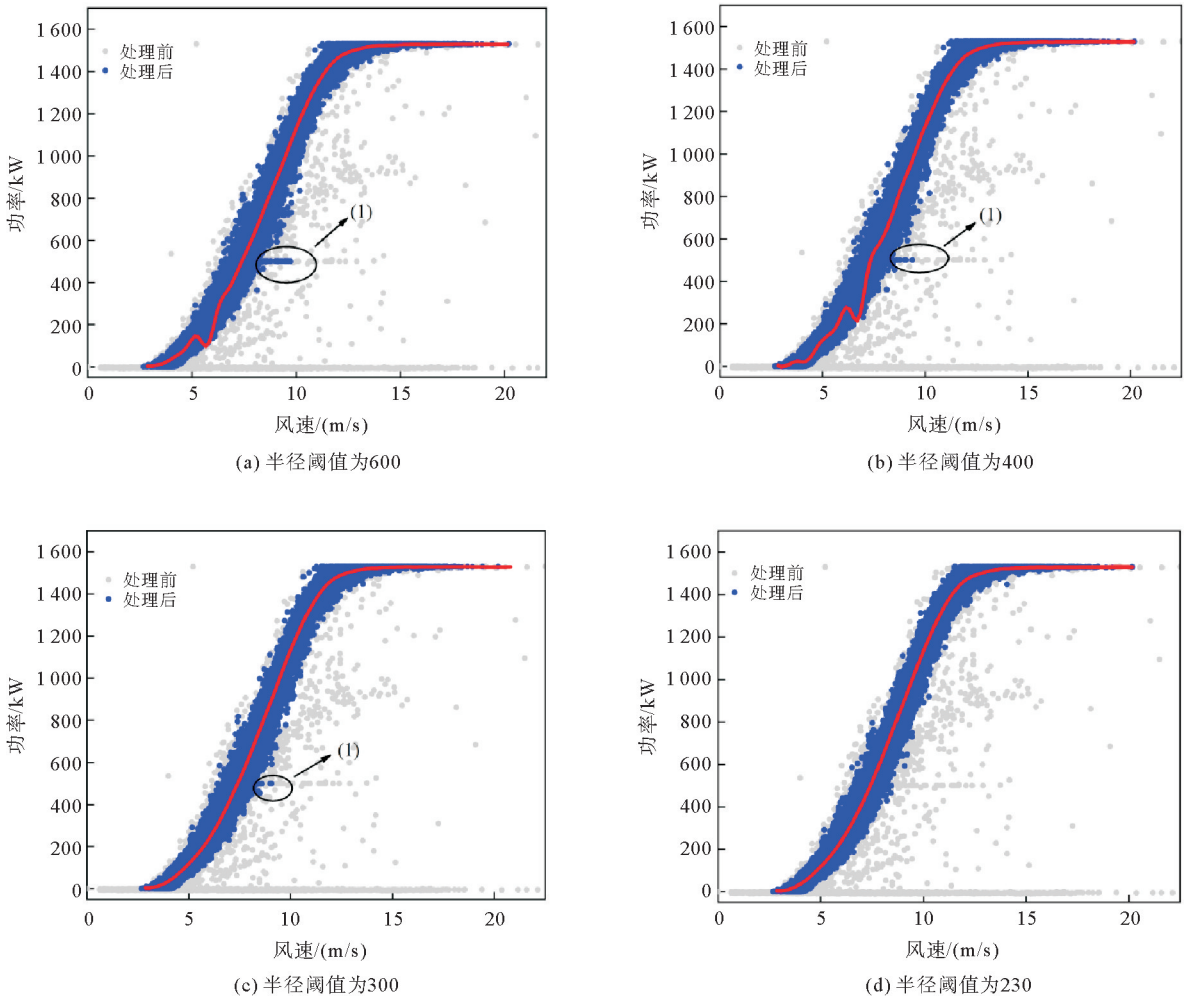
1) 确定阈值所在的区间 $[a,c]$。求得堆积型异常数据区间集合为 $Z=\{U_1,U_2,\cdots,U_m\}$，其中 $U_m=\{(v_1,p_1),(v_2,p_2),\cdots,(v_n,p_n)\}$，且满足 $v_i<v_{i+1}$，$i\in[2,n]$（n 为该堆积型异常数据区间内的数据总数），求该区间内的功率方差值，依次计算每个异常数据区间的方差值，取方差的最小整数值作为阈值区间的右端点，区间的左端点可取为右端点的 1/3，如此取值不仅可以清洗掉堆积型数据点，又可以对分散型异常数据值进行二次识别。

2) 求区间 $[a,c]$ 的中点 b，判断当阈值为 b 时，经 FSC 算法处理之后的数据是否含有堆积型异常数据。

若含有堆积型异常数据,则半径阈值的最佳取值位于区间[a,b],若处理之后的数据不含有堆积型异常数据,则半径阈值的最佳取值位于区间[b,c]。

3) 继续在区间内进行二分搜索,直到找出最佳的半径阈值,该阈值能基本实现堆积型异常数据的清洗。

以北方某风电场风电机组数据为例,对采用滑动四分位清洗过后的风速-功率数据进行堆积型异常数据区间的判断,确定半径阈值过程如图 4 所示。



(a) 半径阈值为600　(b) 半径阈值为400

(c) 半径阈值为300　(d) 半径阈值为230

图 4　半径阈值选取过程示意图

Fig. 4　Schematic diagram of the selection process of the radius threshold

1) 半径阈值的初始范围为[200, 600],当半径阈值为 600 时,FSC 算法的处理效果如图 4(a)所示,图中(1)处仍显示异常数据簇,显然不满足清洗要求。

2) 计算初始区间的中点为 400,当半径阈值为 400 时,FSC 算法的清洗效果如图 4(b)所示,仍有部分堆积型异常数据清洗不彻底,故可知最佳半径阈值应位于区间[200, 400]。

3) 计算区间[200,400]的中点为 300,当半径阈值为 300 时,FSC 算法的清洗效果如图 4(c)所示,仍存在少量异常数据点,故可知最佳半径阈值位于区间[200, 300]。

4) 依次类推,重复上述步骤,最终取整数确定半径阈值为 230 时,处理效果如图 4(d)所示,处理效果最为合适。

为了更好地说明本研究提出的半径阈值确定方法的数据识别效果,对清洗过后的风功率数据采用 bin 算法[15]建模,以均方根误差(root mean square error,RMSE)和平均绝对误差(mean absolute error,MAE)

作为评价指标,对比不同半径阈值的建模误差大小[16]。其中,每个 bin 区间的风速和功率均值分别为:

$$\bar{v}=\frac{1}{N}\sum_{i=1}^{N}v_i\ , \tag{12}$$

$$\bar{p}=\frac{1}{N}\sum_{i=1}^{N}p_i\ 。 \tag{13}$$

均方根误差 R_{MSE} 和平均绝对误差 M_{AE} 的计算式分别为:

$$R_{\mathrm{MSE}}=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{p}_j-p'_j)^2}\ , \tag{14}$$

$$M_{\mathrm{AE}}=\frac{1}{n}\sum_{j=1}^{n}|\hat{p}_j-p'_j|\ 。 \tag{15}$$

式中:n 为数据的个数,$\hat{p}_j$ 为理想功率值,p'_j 为三次样条插值得到的拟合功率值。对不同阈值的功率曲线数据建模的 R_{MSE} 和 M_{AE} 值如表 1 所示,建模曲线为图 4 中红色曲线。

表 1 风电机组等值功率曲线的评价指标值

Table 1 Evaluation index value of equivalent power curve of wind turbine

半径阈值	R_{MSE}	M_{AE}
600	109.22	69.44
400	100.92	64.77
300	96.87	61.95
230	70.87	44.99

结合表 1 和图 4 可知,当半径阈值选择为 230 时,通过 bin 算法建模得到的风功率曲线最接近于标准的风功率曲线,且建模的 R_{MSE} 和 M_{AE} 取得最小值。同时,比较图 4(c)和图 4(d)可知,只有少量异常数据点存在时,对功率曲线的建模影响很小,故为了方便计算,选取半径阈值时可取近似。

3 实验结果与分析

3.1 数据来源

为了说明本研究所提算法和流程的有效性,选取我国北方某风电机组的实际运行数据进行实验验证,取该风电机组 2014 年 9 月—2015 年 8 月、2015 年 9 月—2016 年 8 月、2016 年 9 月—2017 年 8 月共 3 年的原始数据,数据每 10 min 记录一个点。该风电机组的基本参数为:切入风速为 3 m/s,切出风速为 23 m/s,额定功率为 1 500 kW,额定风速为 12 m/s。该风电机组 3 年间每连续 12 个月的原始数据如图 5 所示。

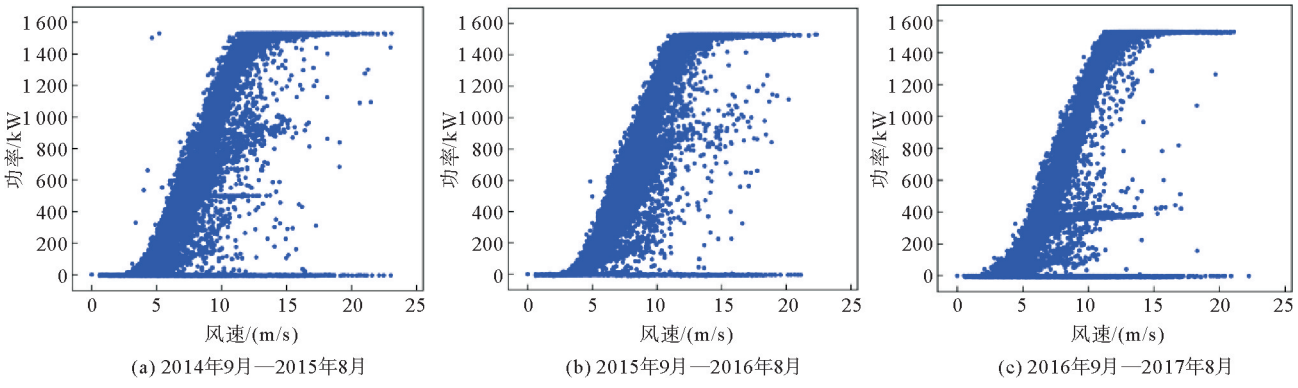


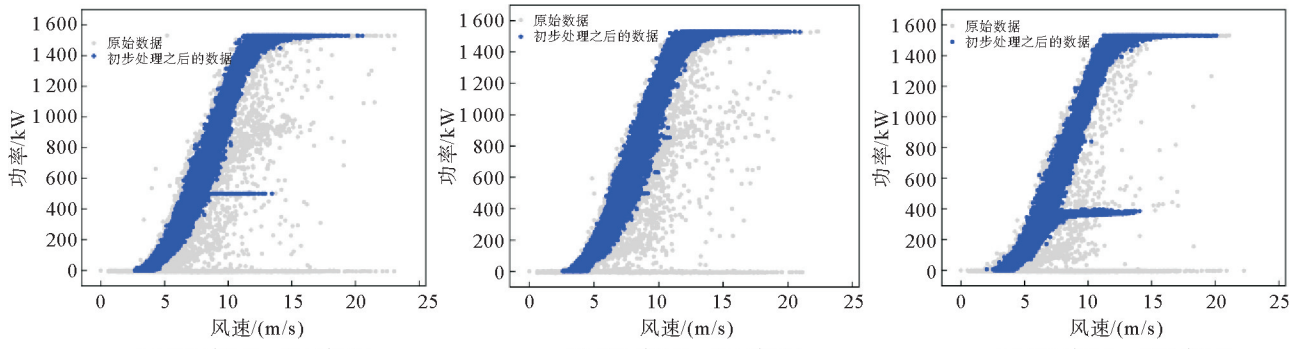
图 5 实测风速-功率散点图

Fig. 5 Scatter diagram of measured wind speed-power

由图 5 可知,该风电场的风电机组 3 年来的风功率实测数据中存在不同类型的异常数据,但具体分布位置和数量有明显差别。其中,图 5(a)中风功率实测数据中两类异常数据均比较典型,图 5(b)中风功率实测数据中主要为分散型异常数据,图 5(c)中风功率实测数据中主要为堆积型异常数据。

3.2 数据清洗测试

对图 5(a)、5(b)、5(c)风电机组的风速-功率实测数据运用滑动四分位算法进行预处理,窗口宽度、步长分别取为 40、40,41、41,20、20,初步处理效果如图 6 所示。

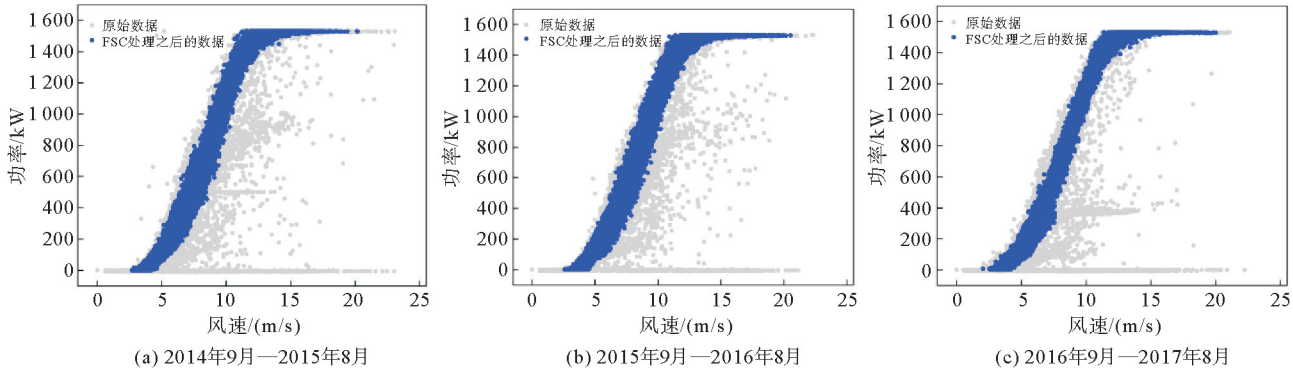


(a) 2014年9月—2015年8月　(b) 2015年9月—2016年8月　(c) 2016年9月—2017年8月

图 6　风功率数据初步处理效果图

Fig. 6　Preliminary processing effect of wind speed-power data

由图 6 可知,原始数据经过滑动四分位算法首次处理之后,分散型异常数据基本被剔除,但位于曲线中部横向分布的堆积型异常数据并未被识别,接下来用 FSC 算法进行下一步处理。按照 2.4 节所述最佳阈值的选取规则,将半径阈值分别设置为 230、200、120,清洗结果与原始数据对比如图 7 所示。由图 7 可知,本研究提出 FSC 算法能够进一步完成堆积型异常数据的清洗。



(a) 2014年9月—2015年8月　(b) 2015年9月—2016年8月　(c) 2016年9月—2017年8月

图 7　FSC 算法数据清洗效果图

Fig. 7　Data cleaning renderings by FSC algorithm

3.3 算法对比分析

采用本研究所提算法与 Thompson tau-四分位算法[17]、LOF 算法[18]分别对该风电机组前述数据进行分析,比较 3 种算法在清洗效果、数据删除率、清洗时间等方面的差距。对比分析时,Thompson tau-四分位算法的舍弃水平设置为 0.01;为便于比较,LOF 算法的邻域点的个数为 20,且 LOF 算法的数据删除率设置为与本研究算法相同。表 2 统计了 3 种方法对该机组 3 年内的数据删除率、清洗效率情况,Thompson tau-四分位算法、LOF 算法的数据清洗效果分别如图 8、图 9 所示。

由表 2 可知,3 种算法的数据删除率、算法效率存在较大差距,具体表现为:

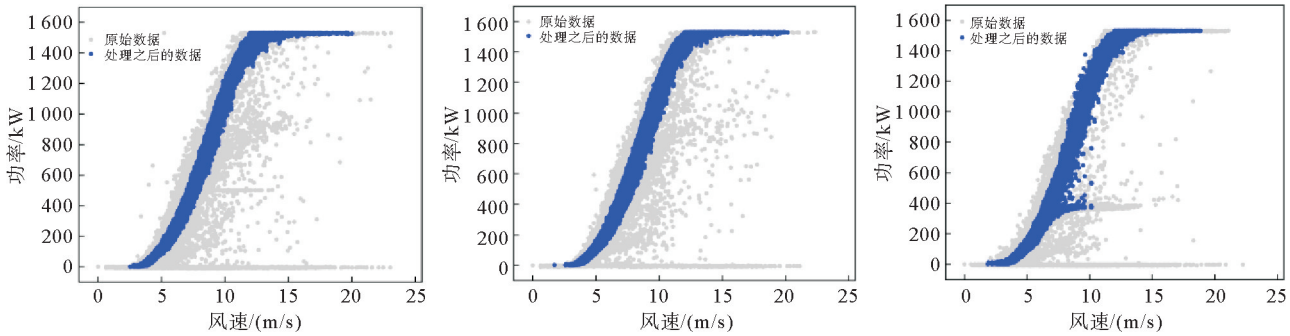
1) 数据删除率。用本研究算法对 2014 年 9 月—2016 年 8 月数据进行处理,数据删除率分别为 7.24%和 9.59%,而采用 Thompson tau-四分位算法处理,数据删除量分别为 21.49%和 21.06%,表明 Thompson tau-四分位算法在处理这两年的数据时存在过识别现象,一部分正常数据被误删。结合图 7(c)和图 8(c)可知,Thompson tau-四分位算法无法准确识别 2016 年 9 月—2017 年 8 月的堆积型异常数据。

表 2 不同方法的数据清洗效果

Table 2 Data cleaning effects of different methods

清洗算法	数据年限	原始数据量	预处理后数据量	剩余数据量	删除率/%	清洗时间/s
Thompson tau-四分位	2014 年 9 月—2015 年 8 月	52 560	37 273	29 263	21.49	183
	2015 年 9 月—2016 年 8 月	52 705	36 483	28 799	21.06	134
	2016 年 9 月—2017 年 8 月	52 560	29 728	23 121	22.22	160
LOF	2014 年 9 月—2015 年 8 月	52 560	37 273	34 576	7.24	32.13
	2015 年 9 月—2016 年 8 月	52 705	36 483	32 984	9.59	30.45
	2016 年 9 月—2017 年 8 月	52 560	29 728	22 161	25.45	28.32
本算法	2014 年 9 月—2015 年 8 月	52 560	37 273	34 576	7.24	5.24
	2015 年 9 月—2016 年 8 月	52 705	36 483	32 984	9.59	5.06
	2016 年 9 月—2017 年 8 月	52 560	29 728	22 161	25.45	4.81

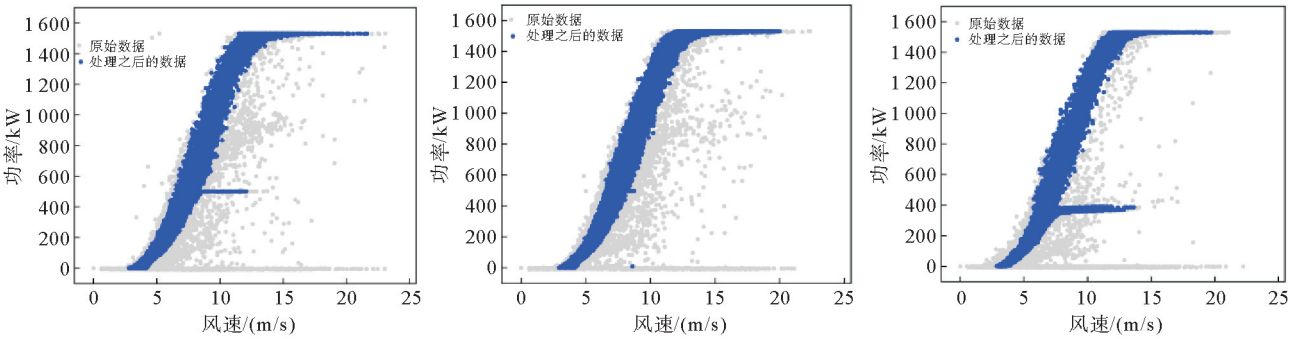
注:预处理是将功率小于等于零的数据剔除;删除率=(1−剩余数据/预处理之后剩余数据)×100%。



(a) 2014年9月—2015年8月 (b) 2015年9月—2016年8月 (c) 2016年9月—2017年8月

图 8 Thompson tau-四分位算法清洗效果

Fig. 8 Data cleaning effect by Thompson tau-quartile algorithm



(a) 2014年9月—2015年8月 (b) 2015年9月—2016年8月 (c) 2016年9月—2017年8月

图 9 LOF 算法清洗效果

Fig. 9 Data cleaning effect by LOF algorithm

2) 算法效率。本研究算法对该风电机组 3 年的数据清洗时间均在 5 s 左右,Thompson tau-四分位算法的清洗时间约为 2～3 min,LOF 的清洗时间约为 30 s,表明本研究算法效率比 Thompson tau-四分位算法和 LOF 算法均有很大提升。

此外,对比图 7、图 8 可知,Thompson tau-四分位算法可以处理分散型异常数据,但当堆积型异常数据过多时,该算法将失去使用价值。

由表 2 和图 7、图 9 可知,LOF 算法和本研究算法在异常数据识别率相同的情况下,LOF 算法对于该风

电机组 2014 年 9 月—2015 年 8 月实测风功率数据中存在的分散型异常数据处理效果尚可,但基本无法识别该风电机组产生的堆积型异常数据。该风电机组 2015 年 9 月—2016 年 8 月产生的异常数据多为分散型异常数据,故 LOF 算法基本能保留完整的数据主带,但对于以堆积型异常数据为主的 2015 年 9 月—2016 年 8 月实测风功率数据处理效果欠佳。

4 结论

本研究分析了风电机组风速-功率异常数据的来源及分布特征,提出基于滑动四分位和 FSC 算法的风速-功率异常数据清洗方法,首先采用滑动四分位算法实现了数据散点的清洗,然后针对滑动四分位算法不能有效清洗的横向堆积型异常数据提出了 FSC 算法。验证结果表明,本研究提出的基于滑动四分位和 FSC 算法的风速-功率异常数据清洗方法,通过设置算法半径阈值,能够很好地识别并清洗风电机组产生的异常数据,清洗效果不受堆积型异常数据的影响。同时,该方法考虑了风功率数据的运行主带,在实现异常数据清洗的同时尽可能地保留了正常运行的数据点,清洗效率较高,具有一定的实际应用价值。

参考文献:

[1] 李昱君,于永进.促进风电消纳的源-荷联合优化调度[J].山东科技大学学报(自然科学版),2021,40(4):118-126.
LI Yujun,YU Yongjin. Source-load joint optimization scheduling for promoting wind power accommodation[J]. Journal of Shandong University of Science and Technology(Natural Science),2021,40(4):118-126.

[2] SHEN G,XU B,JING Y X,et al. Monitoring wind farms occupying grasslands based on remote-sensing data from China's GF-2 HD satellite: A case study of Jiuquan city,Gansu province,China[J]. Resources,Conservation and Recycling,2017,12(1):128-136.

[3] 史坤鹏,赵伟,李婷,等.计及熵权指标及关联度排序的风电历史数据挖掘[J].电测与仪表,2017,54(4):1-5.
SHI Kunpeng,ZHAO Wei,LI Ting,et al. Study on mining in the historical data of wind power based on entropy-weight index and correlation sorting[J]. Electrical Measurement & Instrumentation,2017,54(4):1-5.

[4] 娄建楼,胥佳,陆恒,等.基于功率曲线的风电机组数据清洗算法[J].电力系统自动化,2016,40(10):116-121.
LOU Jianlou,XU Jia,LU Heng,et al. Wind turbine data-cleaning algorithm based on power curve[J]. Automation of Electric Power Systems,2016,40(10):116-121.

[5] ZHENG L,HU W,MIN Y. Raw wind data preprocessing:A data-mining approach[J]. IEEE Transactions on Sustainable Energy,2017,6(1):11-19.

[6] 赵永宁,叶林,朱倩雯.风电场弃风异常数据簇的特征及处理方法[J].电力系统自动化,2014,38(27):39-46.
ZHAO Yongning,YE Lin,ZHU Qianwen. Characteristics and processing method of abnormal data clusters caused by wind curtailments in wind farms[J]. Automation of Electric Power Systems,2014,38(27):39-46.

[7] 胡阳,乔依林.基于置信等效边界模型的风功率数据清洗方法[J].电力系统自动化,2018,42(15):18-23.
HU Yang,QIAO Yilin. Wind power data cleaning method based on confidence equivalent boundary model[J]. Automation of Electric Power Systems,2018,42(15):18-23.

[8] ANDREW K,SONG H,SONG Z. Models for monitoring wind farm power[J]. Renewable Energy,2009,34(3):583-590.

[9] SHEN X,FU X,ZHOU C,et al. 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.

[10] WANG P,ZHANG Y,YUAN S,et al. Analysis and application of the relationship between wind power curve and power generation based on operating data[J]. Journal of Physics:Conference Series,2020,1676(1):12205-12210.

[11] 陈俊生,李剑,陈伟根,等.采用滑动窗口及多重加噪比堆栈降噪自编码的风电机组状态异常检测方法[J].电工技术学报,2020,35(2):346-358.
CHEN Junsheng,LI Jian,CHEN Weigen,et al. A Method for detecting anomaly conditions of wind turbines using stacked denoising autoencoders with sliding window and multiple noise ratios[J]. Transactions of China Electrotechnical Society,2020,35(2):346-358.

[12] AGARWAL A,KHANDEPARKAR K. Distributing power limits:Mitigating blackout through brownout[J/OL]. Sustainable Energy Grids and Networks,2021,26(1). DOI:10.1016/j.segan.2021.100451.

[13] 田书欣,程浩忠,曾平良,等. 基于调频层面的风电弃风分析[J]. 电工技术学报,2015,30(7):18-26.
TIAN Shuxin,CHENG Haozhong,ZENG Pingliang,et al. Analysis on wind power curtailment at frequency adjustment level[J]. Transactions of China Electro-technical Society,2015,30(7):18-26.

[14] 杨玉龙,王子善,杨震,等. 考虑风电不确定性的蓄热式电采暖消纳弃风经济性分析[J]. 电测与仪表,2020,57(13):47-54.
YANG Yulong,WANG Zishan,YANG Zhen,et al. Day-ahead Economic analysis of electric heating with storage system for heating considering the uncertainty of wind power[J]. Electrical Measurement & Instrumentation,2020,57(13):47-54.

[15] 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.
WANG Xin,WANG Zhengxia. Wind speed-power data cleaning of wind turbine based on improved bin algorithm[J]. Chinese Journal of Intelligent Science and Technology,2020,2(1):62-71.

[16] WANG Y,HU Q,SRINIVASAN D,et al. Wind power curve modeling and wind power forecasting with inconsistent data [J]. IEEE Transactions on Sustainable Energy,2019,10(1):16-25.

[17] 邹同华,高云鹏,伊慧娟,等. 基于 Thompson tau-四分位和多点插值的风电功率异常数据处理[J]. 电力系统自动化,2020,44(15):156-162.
ZOU Tonghua,GAO Yunpeng,YI Huijuan,et al. Processing of wind power abnormal data based on Thompson tau-quartile and multi-point interpolation[J]. Automation of Electric Power Systems,2020,44(15):156-162.

[18] 范晓泉,杜大军,费敏锐. 风电异常测量数据智能识别方法研究[J]. 仪表技术,2017(1):10-14.
FAN Xiaoquan,DU Dajun,FEI Minrui. Research on the intelligent identification method for abnormal measurement data of the wind power[J]. Instrumentation Technology,2017(1):10-14.

(责任编辑:齐敏华)