

基于滑动标准差计算的光伏阵列异常数据清洗办法

时珉¹, 尹瑞¹, 胡傲宇², 吴骥³

(1. 国网河北省电力有限公司, 河北 石家庄 050000; 2. 华北电力大学, 北京 102206;
3. 中国电力科学研究院有限公司, 江苏 南京 210000)

摘要: 光伏阵列运行数据中存在大量异常数据, 这些异常数据会对光伏阵列性能分析、建模、故障诊断的实现带来困难。为了有效剔除光伏阵列运行数据中的异常数据, 提出了一种基于滑动标准差的光伏阵列异常数据清洗方法。分析了阵列异常数据的来源及分布特性, 给出了光伏阵列滑动标准差的计算方法。该方法以滑动标准差的曲线上翘作为异常数据的判断依据。最后通过实例分析以及其他方法对比, 说明该算法可以有效降低由于异常数据集中分布带来的异常数据清洗困难。

关键词: 光伏阵列; 功率曲线; 异常数据; 数据清洗; 清洗算法

A novel photovoltaic array outlier cleaning algorithm based on moving standard deviation

SHI Min¹, YIN Rui¹, HU Aoyu², WU Ji³

(1. State Grid Hebei Electric Power Supply Co., Ltd., Shijiazhuang 050000, China; 2. North China Electric Power University, Beijing 102206, China; 3. China Electric Power Research Institute, Nanjing 210000, China)

Abstract: There are a large number of outliers in the PV array operation data. The outlier will bring difficulties to the functions such as PV array performance analysis, modeling, and fault diagnosis. In order to effectively clean the outlier in the PV array operation data, this paper proposes a cleaning method for PV array outlier based on moving standard deviation. It analyzes the source and distribution characteristics of the array outlier data and proposes the algorithm based on moving standard deviation. The curve's rising of the sliding standard deviation set is used as the basis for judging the outlier data. Finally, through the case analysis and comparison of quartile method, the results show that the algorithm can effectively reduce the cleaning error caused by the concentration distribution of the outlier.

This work is supported by National Natural Science Foundation of China (No. U1765104) and Science and Technology Project of State Grid Hebei Electric Power Co., Ltd. (No. 5204BB170007) Research and Application of Regional Landscape Resources and Mid-long Term Forecast Technique for Power Generation Ability.

Key words: PV array; power curve; outlier; data cleaning; cleaning algorithm

0 引言

近年来中国光伏发电发展速度惊人, 2018 年全国光伏发电新增装机 44.06 GW, 全国光伏发电累计装机达到 174.63 GW^[1]。随着光伏装机容量的快速增长, 光伏发电系统智能化运行功能^[2]。上述功能的实现依赖于数据的质量和可靠性。然而在光伏发电系统实际运行过程中存在大量异常数据, 产生这些

异常数据的原因包括数据传输, 维护就显得愈发重要(其中系统性能分析、状态评价、故障诊断、预测性维护是智能运行维护的核心传感器故障, 通信、测量设备故障, 最大功率跟踪异常^[3]以及阵列停机限电等。因此异常数据的清洗在实际工程应用中具有重要意义。

目前研究者们已经在新能源发电系统异常数据识别和数据清洗这一领域做了大量工作, 并取得了很大的成就。常见的方法包括两类: 一类是全局概率统计方法, 另一类是智能聚类方法。文献[4]基于 B 样条平滑和基于内核平滑的技术, 提出了一种自动清理损坏和丢失的负载曲线数据的方法。文献

基金项目: 国家自然科学基金项目 (U1765104); 国网河北省电力有限公司科技项目 (5204BB170007) “区域风光资源及发电能力中长期预测技术研究与应用”

[5]提出了一种基于Copula函数的联合概率模型, 用来排除风功率曲线中的异常数据。文献[6]提出了一种基于变点分组算法和四分位算法清除异常数据的方法, 该方法能够识别风力曲线中的四种异常数据特征。全局概率统计方法的优点在于方法本身成熟的概率统计理论^[7-9], 但异常数据的分布会影响样本数据的分布特征, 使分布参数发生畸变, 从而影响到全局概率统计方法数据筛选结果的准确性^[10]。

对于智能聚类方法方面, 文献[11]提出了一种经验聚类方法, 通过局部异常因子(LOF)算法计算风力发电运行数据库中各个对象之间的离群因子, 然后以这个离群因子为基准, 分析归类运行数据类型。文献[12]提出了一种基于分层聚类算法的地区风电出力典型场景选取方法, 利用分层聚类算法分析风电出力样本, 从而得到样本亲疏关系的聚类树状图。智能聚类方法的优点^[13-14]在于算法可以有效根据数据特征进行分类。但是智能方法泛化能力有限, 其泛化能力取决于算法自身对于海量新鲜样本的学习, 然而在实际过程中样本数据的种类和数据容量都无法得到保证^[15]。另一方面, 智能聚类算法的划分结果难以从物理意义的角度进行解释^[16]。

随着光伏装机规模的不断扩大, 越来越多的学者开始研究光伏发电中的异常数据问题。文献[17]采用短路电流 I_{sc} 作为参考值, 但是短路电流在实际光伏阵列运行过程中难以获取, 无法应用于实际。文献[18]基于最小二乘拟合方法, 提出了单点异常筛查、短期连续异常、长期连续筛查三种情况下的异常数据筛选方法, 该方法对历史数据的质量和精度要求较高, 而且无法排除异常数据造成的分布参数畸变对算法的负面影响。光伏发电系统出力特性和电气参数分布和外部环境关系明确^[19], 考虑这种确定的对应关系, 从而对光伏阵列数据清洗方法进行优化设计, 是有效提高数据清洗准确率的一种途径。

考虑到光伏阵列异常数据的分布特性, 本文提出了一种用于光伏功率曲线分析的基于滑动标准差的异常数据清洗算法, 论文主要工作包括: 首先分析了光伏阵列异常数据的来源与分布特征, 其次详细说明了算法求解步骤, 最后实例说明了算法的性能。

1 光伏阵列运行过程中的异常数据

光伏阵列输出电流、功率和辐照量之间为线性关系^[20-22], 此分布特性可以作为异常数据的判断依据。本文以辐照度-输出功率这组对应关系来说明光伏阵列异常数据的来源和分布特征。图1说明了光

伏阵列实际运行数据中异常数据的分布特征, 由图1可知异常数据可以分为两类。

A类异常数据为曲线底部堆叠的异常数据, 堆叠的数据通常是由瞬时且无法恢复的故障或异常引起的。这类异常数据的主要特征为在辐照度远大于0时, 光伏功率保持为0或接近于0。这种异常数据产生的原因是组件或逆变器设备故障, 通信、传感器终端故障, 发电单元或机组停机。

B类异常数据为曲线周围的散乱数据点, 是功率曲线附近的一些不规则散乱点。分散的异常数据点通常是由短时间内可以恢复的故障或异常引起的, 其来源分别是通信或传感器异常、噪声, 外部输入的随机波动性带来的测量误差, 最大功率跟踪不准确。

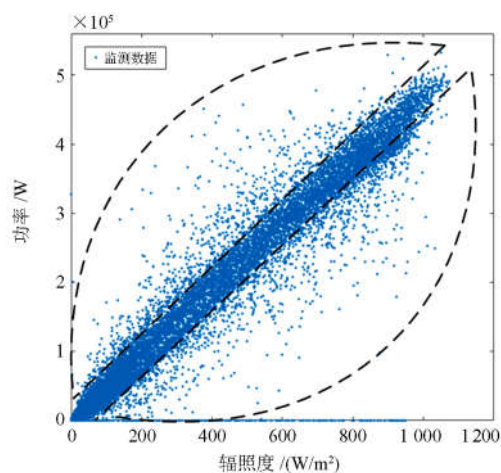


图1 光伏阵列异常数据分布

Fig. 1 PV array outlier data distribution

2 滑动标准差算法

本文以光伏阵列运行数据异常数据分布特征为判定异常的依据。当发现运行数据中存在异常值时, 数据的变化率、平均值、方差、标准差和方差变化率等数据特征将发生变化。选择合适的变化点指标能够准确地识别异常值^[23]。

在本文所提的算法中滑动标准差被作为判断指标。标准差的计算公式为

$$\sigma(r) = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - r)^2} \quad (1)$$

为了让标准差能够更加精确地反映出一组数据的变化程度, 采用多组数滑动计算的方法。现将数据样本集合 U 分成 m 个互相独立的子集合 Y , 即 $U = \{Y_1, Y_2, \dots, Y_m\}$ 。同时设定滑动集合为 $Z_j = \{(x_1, y_1), (x_2, y_2), \dots, (x_a, y_a)\}$, Z_j 表示某子集合 Y_m

中的滑动集合滑动到第 j 个位置, $j=1,2,\dots,n-a+1$, $i=1,2,\dots,n$ 。 a 为滑动集合内数据点总数, 且 $1 < a < n$ 。 n 为子集合数据点总数。

在滑动分组完成后, 逐一计算每个滑动集合的标准差 $\sigma_{m,j}$, 其中 $\sigma_{m,j}$ 表示子集合 Y_m 中的第 j 个标准差值。计算完毕后, 对每个子集合 Y_m 中的标准差值进行变点分析, 识别分类每个子集合中的正常数据与异常数据。最后将分类结果进行整合, 完成数据清洗。

现以一个光伏阵列的实测辐照度与功率数据为例来详细说明滑动标准差算法的步骤、求解特点以及需要注意的细节, 图 2 为滑动标准差算法的流程图。

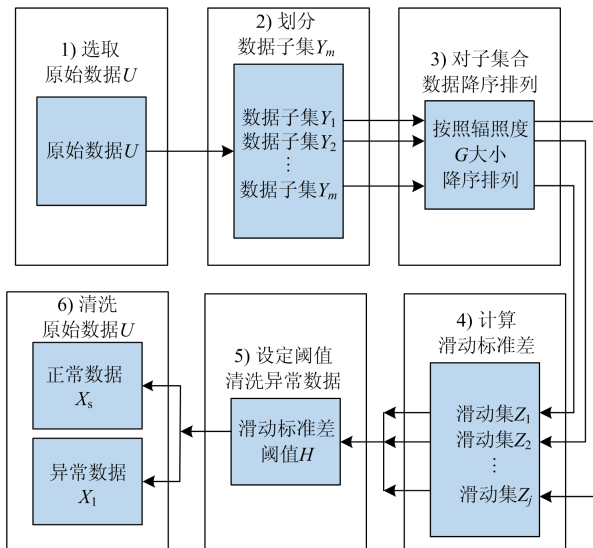


图 2 滑动标准差算法流程图

Fig. 2 Flow chart of sliding standard deviation mutation algorithm

具体算法求解步骤如下, 其中所有辐照度值和功率值已做归一化处理。

1) 选取原始数据 U 。本实例选取了某电站光伏阵列全年的实测辐照度与功率数据作为样本数据。

2) 划分数据子集 Y_m 。将原始数据按照辐照度区间 $T=10 \text{ W/m}^2$ 划分为 120 个子集合。

3) 对子集合数据降序排列。由于每个子集合计算过程相似, 现以第 77 个子集合为例进行步骤详解。第 77 个子集合中共有 130 个功率点, 对数据点按照功率大小进行降序排列, 满足 $y_i < y_{i-1}$, $i \in (2, 130)$, 从而得到排序后的 $Y_{77} = \{(x_1, y_1), (x_2, y_2), \dots, (x_{130}, y_{130})\}$, 其中 x 为辐照度值, y 为功率值。

4) 计算滑动标准差。设滑动集合的数据容量 $a=30$ 。将子集合数据代入滑动集合 Z_j , 其中 $j=1,2,\dots,101$, 代入规律如图 3 所示。然后逐一计算 101 个滑动集合的标准差, 这 101 个滑差值以及其他具体数据如表 1 所示。

5) 设定阈值清洗异常数据。根据计算的滑差值的统计分布规律, 设定该样本数据阈值为 0.02。

设定阈值后, 该子集合的滑差值曲线如图 4(a) 所示。发现在第 3 个滑差值的左侧和第 94 个滑差值右侧的滑差曲线都显著上翘, 且没有趋于平稳的倾向。曲线中间走势平稳说明这段曲线滑差值相近, 数据波动小, 符合正常值范畴。同时结合滑差阈值以及公式(2), 确定该子集合第 4 个数据点以及第 124 个数据点为变化点, 从而确定前 3 个数据点(A 区域)和后 6 个数据点(B 区域)为异常数据点, 其处理结果如图 4(b)所示。

6) 清洗原始数据 U 。最后对剩余的其他组按照同样的步骤进行处理。

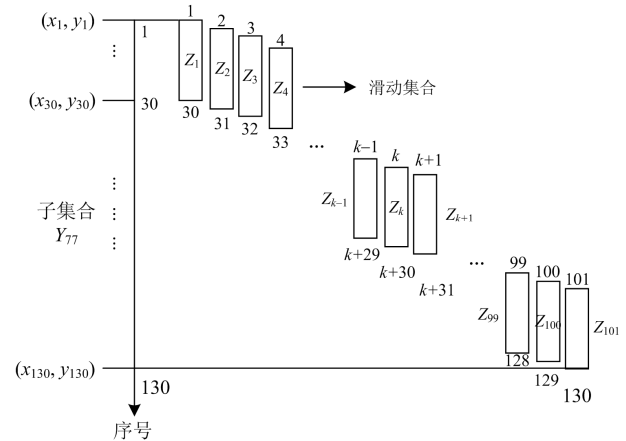


图 3 滑动集合示意图

Fig. 3 Sliding set diagram

表 1 第 77 个子集合滑差值及其他数据
Table 1 77th subset's sliding standard deviations and other types' data

滑动集合	功率/W	环境温度/℃	倾斜面辐照度/(W/m ²)	组件温度/℃	功率归一化	滑差值
Z ₁	441 520.7	34.00	779.90	52.79	0.724	0.031 1
Z ₂	382 031.2	32.50	779.89	55.83	0.626	0.026 8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Z ₁₀₀	416 376	29.49	770.11	51.09	0.683	0.037 7
Z ₁₀₁	407 237.6	32.46	770.01	52.80	0.668	0.121

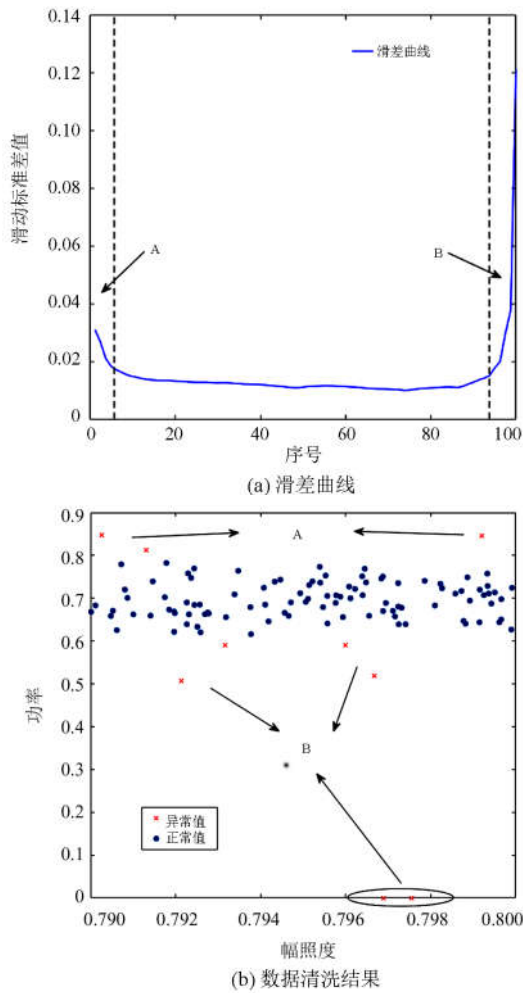


图 4 第 77 个子集结果图

Fig. 4 Result of 77th subset

3 方法验证

3.1 实例分析

以某大型并网光伏电站为分析对象, 选取了三个典型阵列汇流箱一年的实际运行数据作为清洗样本, 编号分别为 5B、19A、37B, 数据的分辨率为 10 min 一个记录点。

以 37B 阵列汇流箱的运行数据为例, 计算不同阈值设定下滑动标准差算法的清洗效果。由于功率和辐照度之间是一个线性关系, 故引入线性相关系数作为衡量标准。计算结果如表 2。图 5 为不同阈值下清洗结果对比图。

由表 2 可以看出, 阈值的设定会影响数据删除率和清洗结果的线性相关系数。在设计阈值范围内, 线性相关系数稳定在 99% 以上, 且随着阈值的减小而小幅提升。同时数据删除率会随着阈值的减

小而大幅增加。

表 2 不同阈值设定下 37B 阵列数据滑动标准差算法清洗结果

Table 2 Data cleaning results of 37B under different thresholds

滑差阈值	原始数据量	剩余数据量	数据删除率	线性相关系数
0.005	33 264	23 075	30.63%	99.75%
0.01	33 264	25 826	22.36%	99.69%
0.02	33 264	28 936	13.01%	99.63%
0.03	33 264	30 054	9.65%	99.43%
0.04	33 264	31 105	6.49%	99.25%

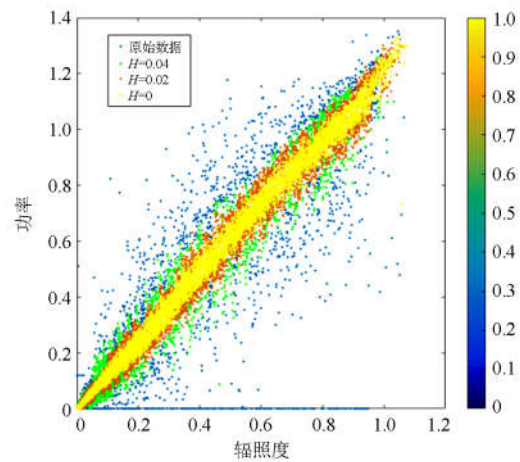


图 5 不同阈值设定下的 37A 数据清洗结果

Fig. 5 Data cleaning results of 37A under different thresholds

根据实际应用场景的不同, 阈值的设定也会随改变, 例如建模时对数据要求较高, 可以将阈值设定为 0~0.01。结合表 2 与图 5 可知阈值设定越小, 线性相关系数越大, 但是相应的数据删除量就会大量增加; 阈值设定越大, 数据删除量越小, 但线性相关系数会随之降低。

3.2 性能对比

本文设计了几种不同异常数据分布场景(底部异常数据堆积、顶部异常数据堆积和正常异常数据分布), 采用四分位法和滑动标准差算法进行对比, 其中滑差方法的阈值设置为 0.02。同时对清洗效果、数据删除率以及线性相关系数等参数指标进行分析, 清洗结果如图 6 所示。

由图 6 可知, 在底部堆积和顶部堆积情况下, 四分位法算法识别结果会因为异常数据的分布而向异常数据堆积一侧偏移, 这会导致一部分正确数据被识别成异常数据, 一部分异常数据被识别成正常数据。而本文方法并不会受到不同异常数据分布的影响, 仍能准确识别正常数据和异常数据。

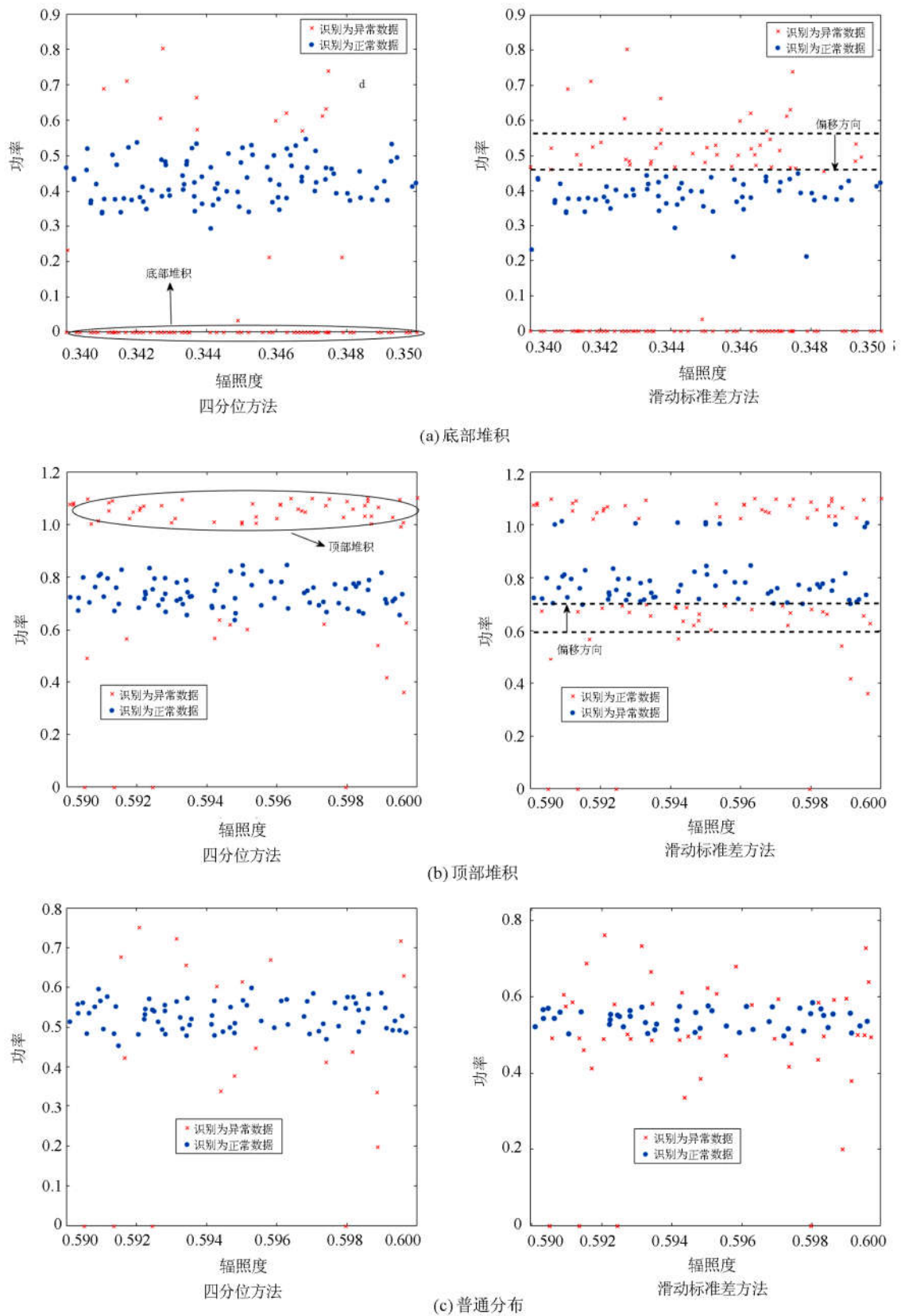


图 6 不同数据情况下的清洗结果

Fig. 6 Cleaning result under different cases

最后分别利用四分位法和滑动标准差方法对上述三个支路一年的实际历史数据进行分析。由图 7 可知, 两种方法都能识别出底部堆叠的异常数据。对于四分位法而言, 其识别的正常数据整体向下偏移, 经分析原始数据发现 37B 数据中在底部堆叠的异常数据占比更大, 从而其影响了四分位算法清洗结果, 导致结果总体向下偏移。另一方面也证明了本文方法并未受到异常数据堆叠分布的影响。

表 3 对两种方法的清洗结果进行了量化。由表 3 可知, 两种方法在清洗后其结果的数据线性相关系

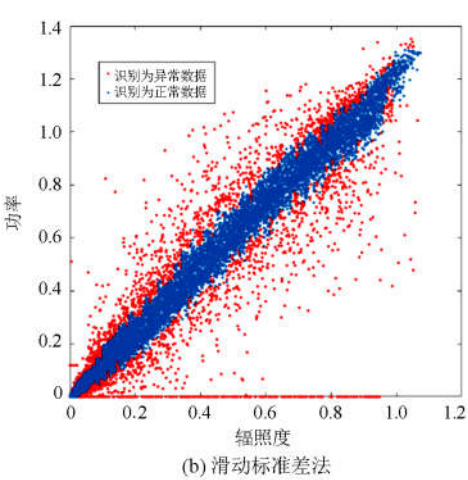
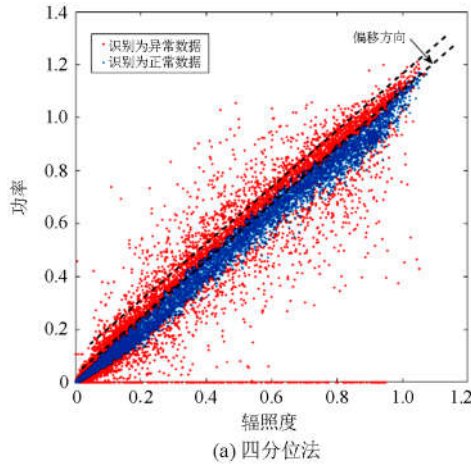


图 7 两种算法下 37B 阵列汇流箱清洗结果

Fig. 7 Cleaning results of two algorithms for 37B data

数都上升到了 99%以上, 而四分位法的数据删除率分别为 44.53%、38.01%、33.56%, 相比本文方法的 15.64%、18.99%以及 13.84%的数据删除率明显要大很多, 同时四分位法的线性相关系数相比本文方法也未见明显提高。这说明四分位法会将一些正常数据识别为异常数据。

表 3 不同清洗方法的清洗结果对比

Table 3 Data cleaning effects of two different algorithms

清洗数据	清洗前线性相关系数	清洗方法	数据删除率	清洗后线性相关系数	线性相关系数变化
5A	94.96%	四分位法	44.53%	99.31%	+4.35%
		滑差法	15.64%	99.00%	+4.04%
19A	94.67%	四分位法	38.01%	99.94%	+5.27%
		滑差法	18.99%	99.75%	+5.08%
37B	96.31%	四分位法	33.56%	99.42%	+3.11%
		滑差法	13.84%	99.33%	+3.02%

4 结论

本文提出了基于滑动标准差的光伏阵列数据清洗方法, 其结果和分析可归纳如下: 分析了光伏阵列异常数据分布特征及来源; 提出了基于滑动标准差算法并对光伏阵列异常数据进行清洗, 清洗结果符合光伏阵列出力数据的分布特征; 本文方法解决了经典四分位法清洗结果易受到异常数据分布的影响, 同时识别准确率高。本文方法可以用于光伏系统性能分析、光伏系统建模、故障诊断等智能光伏发电核心功能的数据预处理中, 对于提高相应算法的精度具有重要意义。

参考文献

[1] 国家能源局. 2018 年全国电力工业统计数据[EB/OL].

http://www.nea.gov.cn/2019-01/18/c_137754977.htm,2019,01,18/2019,01,18.

National Energy Administration. National power industry statistics for 2018[EB/OL]. http://www.nea.gov.cn/2019-01/18/c_137754977.htm,2019,01,18/2019,01,18.

[2] KAMENOPOULOS S N, TSOUTSOS T. Assessment of the safe operation and maintenance of photovoltaic systems[J]. Energy, 2015, 93: 1633-1638.

[3] 葛双冶, 杨凌帆, 刘倩, 等. 基于改进 CPSO 的动态阴影环境下光伏 MPPT 仿真研究[J]. 电力系统保护与控制, 2019, 47(6): 151-157.

GE Shuangye, YANG Lingfan, LIU Qian, et al. Research on photovoltaic MPPT simulation under dynamic shadow environment based on modified CPSO[J]. Power System Protection and Control, 2019, 47(6): 151-157.

- [4] CHEN J, LI W, LAU A, et al. Automated load curve data cleansing in power systems[J]. IEEE Transactions on Smart Grid, 2010, 1(2): 213-221.
- [5] WANG Y, INFELD D G, STEPHEN B, et al. Copula-based model for wind turbine power curve outlier rejection[J]. Wind Energy, 2015, 17(11): 1677-1688.
- [6] SHEN X, FU X, ZHOU 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.
- [7] ZHAO Y, LIN Y E, WANG W, et al. Data-driven correction approach to refine power curve of wind farm under wind curtailment[J]. IEEE Transactions on Sustainable Energy, 2018, 9(1): 95-105.
- [8] 龚莺飞, 鲁宗相, 乔颖, 等. 基于 Copula 理论的光伏功率高比例异常数据机器识别算法[J]. 电力系统自动化, 2016, 40(9): 16-22.
- GONG Yingfei, LU Zongxiang, QIAO Ying, et al. Copula theory based machine identification algorithm of high proportion of outliers in photovoltaic power data[J]. Automation of Electric Power Systems, 2016, 40(9): 16-22.
- [9] 段偲默, 苗世洪, 霍雪松, 等. 基于动态 Copula 的风光联合出力建模及动态相关性分析[J]. 电力系统保护与控制, 2019, 47(5): 35-42.
- DUAN Simo, MIAO Shihong, HUO Xuesong, et al. Modeling and dynamic correlation analysis of wind/solar power joint output based on dynamic Copula[J]. Power System Protection and Control, 2019, 47(5): 35-42.
- [10] YE X, LU Z, QIAO Y, et al. Identification and correction of outliers in wind farm time series power data[J]. IEEE Transactions on Power Systems, 2016, 31(6): 4197-4205.
- [11] ZHENG L, HU W, MIN Y. Raw wind data preprocessing: a data-mining approach[J]. IEEE Transactions on Sustainable Energy, 2015, 6(1): 11-19.
- [12] 林俐, 费宏运, 刘汝琛, 等. 基于分层聚类算法的地区风电出力典型场景选取方法[J]. 电力系统保护与控制, 2018, 46(7): 1-6.
- LIN Li, FEI Hongyun, LIU Ruchen, et al. A regional wind power typical scenarios' selection method based on hierarchical clustering algorithm[J]. Power System Protection and Control, 2018, 46(7): 1-6.
- [13] 伍育红. 聚类算法综述[J]. 计算机科学, 2015, 42(增刊 1): 491-499, 524.
- WU Yuhong. General overview on clustering algorithms[J]. Computer Science, 2015, 42(S1): 491-499, 524.
- [14] 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.
- [15] BINKHONAIN M, ZHAO L. A review of machine learning algorithms for identification and classification of non-functional requirements[J]. Expert Systems with Applications: X, 2019, 1: 100001.
- [16] OKTAR Y, TURKAN M. A review of sparsity-based clustering methods[J]. Signal Processing, 2018, 148: 20-30.
- [17] ZHANG J, ZHANG S, LIANG J, et al. Photovoltaic generation data cleaning method based on approximately periodic time series[C] // 2017 International Conference on Environmental and Energy Engineering (IC3E 2017), March 22-24, 2017, Suzhou, China: 2008-2014.
- [18] YU L, WANG H, CHE J, et al. Outliers screening for photovoltaic electric power based on the least square method[C] // IEEE Chinese Control and Decision Conference, May 28-30, 2016, Yinchuan, China: 2799-2804.
- [19] 郑可轲, 牛玉广. 大规模新能源发电基地出力特性研究[J]. 太阳能学报, 2018, 39(9): 2591-2598.
- ZHENG Keke, NIU Yuguang. Research on renewable power basement output characteristics[J]. Acta Energaie Solaris Sinica, 2018, 39(9): 2591-2598.
- [20] 朱红路, 刘珠慧. 环境因素影响下的光伏系统出力特性分析[J]. 华北电力技术, 2014(8): 50-55.
- ZHU Honglu, LIU Zhuhui. PV system output analysis of environmental factors affect[J]. North China Electric Power, 2014(8): 50-55.
- [21] 孙航, 杜海江, 季迎旭, 等. 光伏分布式 MPPT 机理分析与仿真研究[J]. 电力系统保护与控制, 2015, 43(2): 48-54.
- SUN Hang, DU Haijiang, JI Yingxu, et al. Photovoltaic distributed MPPT mechanism analysis and simulation study[J]. Power System Protection and Control, 2015, 43(2): 48-54.
- [22] SKOPLAKI E, PALYVOS J A. On the temperature dependence of photovoltaic module electrical performance: a review of efficiency/power correlations[J]. Solar Energy, 2009, 83(5): 614-624.
- [23] KHEZRIMOTLAGH D, ZHU J, COOK W D, et al. Data envelopment analysis and big data[J]. European Journal of Operational Research, 2019, 274(3): 1047-1054.

收稿日期: 2019-03-31; 修回日期: 2019-08-25

作者简介:

时 珉(1976—), 男, 本科, 高级工程师, 研究方向为电网调度管理、新能源并网及运行消纳等; E-mail: shimin9999@126.com

尹 瑞(1990—), 男, 博士, 工程师, 研究方向为柔性直流输电技术、灵活交流输电技术、新能源发电; E-mail: 1021207298@zju.edu.cn

胡傲宇(1994—), 男, 通信作者, 硕士研究生, 研究方向为光伏电站故障诊断与智能运维。E-mail: aoyualex@126.com

(编辑 姜新丽)