

风功率异常数据检测方法对比研究

封焯文¹, 朱世平¹, 赵志华¹, 孙铭仁², 董 密², 宋冬然²

(1. 中国能源建设集团湖南省电力设计院有限公司, 湖南 长沙 410007;

2. 中南大学自动化学院, 湖南 长沙 410083)

摘要:风电机组风功率预测和功率曲线建模等工作的开展依赖于历史运行数据。然而,历史数据中积累了大量的异常数据,导致上述工作难以有效开展。国内外学者已经提出了多种异常数据检测方法,然而对不同方法的优缺点与适用场合还缺少整体认识。为此,本文对基于密度的聚类算法、局部离群因子算法、Thompson-tau 四分位法和孤立森林四种常用的风功率异常值检测方法进行了对比研究。为评价不同检测方法,提出了基于标准功率曲线的评价指标。实验结果表明,孤立森林算法相比其他三种方法具有更高的精度,能应对不同分布的异常数据,且清洗时间较短。

关键词:风功率数据;异常值特点;数据清洗;孤立森林

DOI: 10.12067/ATEEE2012009

文章编号: 1003-3076(2021)07-0055-07

中图分类号: TM247

1 引言

风能利用是国家发展绿色经济的重要标志,其发展受到世界各国的重视。风电机组容量不断增大,结构愈加柔性化,依靠人力运维的维护成本也越来越高。基于历史数据对风机运行状态进行分析可提高分析效率,节省大量人力物力。例如,基于风功率历史数据可开展风电机组风速预测和功率曲线建模等工作^[1]。然而,受环境变化和机组故障等异常因素影响,风电机组在实际运行过程中经常产生大量的异常功率数据,给后续数据挖掘工作造成干扰。因此,需要对风功率历史数据中的异常数据进行准确识别和处理,进而获得有效数据集。

目前,异常值检测主要有三种基本方法:①基于风机运行机理简单剔除;②基于突变点特征进行异常检测,如密度、距离等;③基于概率模型建立边界,剔除边界外异常值,如设置阈值或建立风功率曲线等效模型等。文献[2]提出将 Thompson-tau 与四分位法相结合建立边界的清洗方法,该方法清洗时间短,在数据样本较多下有较好的效果。文献[3]提出了用于识别异常数据的云分段最优熵算法,通过比较每个数据的熵与阈值来判断异常数据。异常数

据点偏离正常数据集较远时,导致数据集方差较大。文献[4]提出最优组内方差算法,通过对比每个风速区间内数据点与前面数据点的方差来判断异常点。然而,上述方法易受到偏离程度较大的点影响,难以去除正常数据集周围的异常数据。文献[5]提出基于密度的聚类算法(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)进行异常值检测,然而在异常数据聚集时识别效果不好。针对上述缺点,文献[6]提出将 DBSCAN 与四分位法结合,获得了较好的检测结果。文献[7]采用基于密度的局部离群因子(Local Outlier Factor, LOF)算法,把具有足够高密度的区域划分为簇,实现了分散型异常数据的有效识别。文献[8]提出基于 DBSCAN 和 LOF 的 DLOF 算法,该方法相比于上述两种算法精度更高,但是所需时间更长,因此不适用于处理实时数据。文献[9]在检测异常数据时同时考虑风向数据的正确性,将二维异常检测扩展至三维,通过更多的判定条件达到更好的检测效果。文献[10]采用四分位法与变点分组算法结合的方法,但清洗后数据存在明显的阶梯状使功率曲线模型失真。在更改分组宽度后虽然会得到改善,但对于不同的机组清洗效果差别较大,泛用性低。文献[11]

收稿日期: 2020-12-11

基金项目: 湖南省战略性新兴产业-科技攻关与重大科技成果转化项目(2018GK4002)

作者简介: 封焯文(1983-),男,湖南籍,高级工程师,硕士,研究方向为新能源发电与电气技术;

宋冬然(1983-),男,湖南籍,副教授,博士,研究方向为新能源控制与工业智能(通讯作者)。

提出一种新颖的异常检测方法,该方法将风速功率散点图转化为灰度图像,通过判断功率曲线的形状来识别异常数据。文献[12]结合风机运行过程与数据不确定性统计,提出一种基于置信等效边界模型的风功率数据清洗方法,但需要针对不同数据集调整单一或混合模型。文献[13]采用孤立森林(Isolated Forest, IF)算法,通过在二叉树模型中分离单个数据所用的步数来计算该数据的异常评分,并结合等效边界对异常数据进行识别。虽然上述方法已经较为成熟,但是目前对它们的清洗效果与优缺点还缺少整体认知。

因此,本文系统地对比研究了主流数据清洗方法,并得到了一些有益结论。本文其余部分安排如下:第2节详述了异常数据产生原因,第3节介绍了四种数据清洗方法并给出了评价指标,第4节是方法应用与实例分析,第5节给出了结论。

2 异常数据产生原因

在运行过程中,风电机组受到多种不利因素的影响会采集到大量异常数据。根据文献[10],异常数据可分为四类,如图1所示。

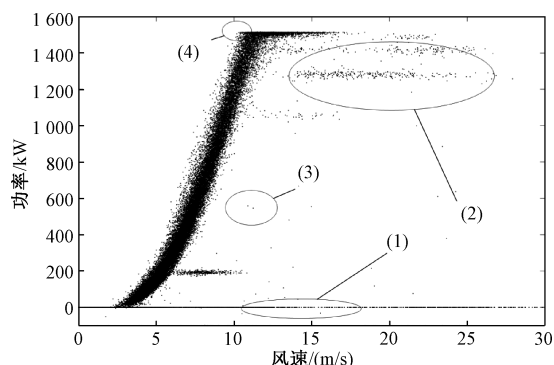


图1 异常数据分布情况

Fig.1 Abnormal data distribution

针对这四类异常数据,各自的分布特征以及产生原因可表述如下。

(1) 聚集分布在曲线底部的数据。曲线底部有大量功率为0或负值的点,这些异常数据的可能原因是机组故障导致机组处于计划或非计划关闭状态。采集到的功率值小于零的情况,这是由风电机组此时处于停机状态,而机组测控系统自耗电所导致。

(2) 零散或聚集分布在曲线中部的数据。此时风电机组可以正常运作,但由于电网消纳能力不足、风力发电不稳定、建设工期不匹配等原因而出现弃

风限电,使得此时的输出功率低于额定功率。

(3) 零散分布在曲线四周的数据。功率曲线数据呈现零散或聚集性的随机分布,源于信号噪声、传感器失灵等因素影响产生的分布随机不固定偏差。

(4) 零散或聚集分布在曲线的顶部的数据。由于机舱风速传感器或通讯故障,导致监控系统采集的风速测量数据异常。

3 异常值检测方法与评价指标

基于风机运行机理简单剔除的方法通常与基于突变点特征进行异常检测或基于概率模型建立边界剔除边界外异常值的方法结合使用。因此,本文选择基于突变点特征异常检测的DBSCAN与LOF方法,以及基于概率模型建立边界,剔除边界外异常值的Thompson-tau四分位法与IF四种方法进行对比研究。考虑正常数据集中于功率曲线周围,而异常数据通常偏离功率曲线,清洗后的数据与参考功率曲线之间的误差,即清洗结果在功率曲线周围的集中程度,能间接反映清洗效果。因此,本文采用基于标准功率曲线的评价指标,用于评价不同异常值检测方法的优劣。

3.1 常见异常值检测方法

3.1.1 DBSCAN

基于密度的聚类算法(DBSCAN)通过判断点 P 周围半径为 eps 内点的个数是否小于某个值 $minpts$,从而判断点 P 是否为核心对象。

若集合中存在一点 O 同时在核心对象 p 和 q 的邻域内,则核心对象 p 和 q 密度相连。DBSCAN的目的便是找到所有密度相连的数据点,以此建立正常数据集^[5,6]。

3.1.2 LOF

局部离群因子算法(LOF)将与点 P 第 K 远的点之间的距离定义为点 P 的第 K 距离记为 $N_k(P)$ 。第 K 距离内的所有点为点 P 的第 K 邻域。定义点 P 的局部可达密度为:

$$lrd_k(P) = 1 / \left(\frac{\sum_{O \in N_k(P)} D(O, P)}{N_k(P)} \right) \quad (1)$$

式中, $D(O, P)$ 为 P 第 K 邻域内点 O 与 P 的距离。由式(1),得到所有点第 K 邻域的局部可达密度,并据此计算点 P 的局部离群因子:

$$LOF_k(P) = \frac{\sum_{O \in N_k(P)} lrd_k(O)}{N_k(P)} / lrd_k(P) \quad (2)$$

式中, $lrd_k(O)$ 为点 P 第 K 邻域内某一点 O 的第 K

邻域局部可达密度。若点 P 局部离群因子接近 1 则说明点 P 与周围点的密度接近,如果大于 1 点 P 周围点的密度小于其他点即异常值。因此若 $LOF_k(P)$ 大于 1,则点 P 为稀疏点即异常点^[7]。

3.1.3 Thompson-tau 四分位法

Thompson-tau 四分位法将风速分区,分别计算每个区间的功率平均值 P_i 与标准差 S_i 。由功率平均值 P_i 得到每个区间功率样本数据偏差的绝对值 $\delta_{i,j} = |P_{i,j} - P_i|$ 。当区间内某个数据点偏差绝对值较大时,表明该点在此区间内过大或过小,由此判断该点是否为异常点。Thompson-tau 法中 τ 值的计算如下:

$$\tau = \frac{t_{\alpha/2}(m-1)}{\sqrt{m}\sqrt{m-2} + t_{\alpha/2}^2} \quad (3)$$

式中, t 为功率样本数据的 t 分布值; α 为显著性水平,其值影响功率数据的充裕度,通常取显著水平 $\alpha = 0.01$ 。

当某一点 $\delta_{i,j} > \tau S_i$ 时,该点为异常数据;反之,该点为正常数据。在第一步使用 Thompson-tau 判断后,再结合四分位法进行二次检测。首先找到每个风速区间内功率数据的上四分位数 $Q_{3,i}$ 与下四分位数 $Q_{1,i}$,然后得到四分位距 $I_i = Q_{3,i} - Q_{1,i}$ 。四分位法中四分位上限 $W_{u,i}$ 与下限 $W_{d,i}$ 计算公式为:

$$\begin{cases} W_{u,i} = Q_{3,i} + 1.5I_i \\ W_{d,i} = Q_{1,i} - 1.5I_i \end{cases} \quad (4)$$

功率在四分位上下限之间的数据为最后的正常数据^[2]。

3.1.4 孤立森林

孤立森林算法 (IF) 的主要思想是:给定 n 个样本数据 $X = \{X_1 \cdots X_n\}$,特征维度为 d ,随机选择特征 q 和其分隔值 p ,递归分割数据集 X 来构建孤立树,直到无法继续分割或达到预设最大高度。

孤立树中样本点 x 的路径长度 $h(x)$ 定义为从 iTree 的根节点到叶子节点所经过的边的数量。由 n 个样本组成的数据集,生成模型树的平均路径长度 $c(n)$ 定义如式 (5) 所示,其中 $H(i)$ 为调和数,通常设置为 $\ln(i) + 0.577\ 215\ 664\ 9$ 。

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (5)$$

树中每个样本 x 的异常得分定义为:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (6)$$

当 $E(h(x))$ 接近 0 时, s 靠近 1,即 x 的异常得分接近 1 时样本 x 被判定为异常。当 $E(h(x))$ 接

近 $n-1$ 时, s 靠近 0,样本 x 被判定为正常^[13]。

3.2 评价指标

由图 1 中异常数据分布可知,正常数据集中在功率曲线周围,异常值明显偏离功率曲线。因此,通过比较不同方法清洗后的结果与厂家提供的标准功率曲线之间的误差能够判断不同清洗方法的性能^[14]。清洗后的数据在标准功率曲线周围集中程度高,表明清洗效果好;反之,清洗效果差。

针对标准功率曲线中 0.5 m/s 的区间(区间按照厂商提供的标准功率曲线上的数据点的风速间隔来选择)划分清洗后的数据集,计算每个区间内数据与标准功率曲线之间的离散程度:

$$AAD_i = \frac{1}{N_i} \sum_{j=1}^{N_i} |P_i - P_{i,j}| \quad (7)$$

$$RMSE_i = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (P_i - P_{i,j})^2} \quad (8)$$

式中, AAD_i 与 $RMSE_i$ 分别为第 i 个风速区间的平均绝对误差和均方根误差; N_i 为第 i 区间内数据量; P_i 为第 i 区间标准功率曲线的值; $P_{i,j}$ 为第 i 区间内第 j 个功率数据。

4 实例分析

本节选择 3 个风场实际风电机组运行数据进行方法应用,并基于所提出的评价指标对不同异常值检测方法进行性能对比。具体为:扶余三井子风场#2 机组,青径云霄风场#11 机组与祥云天峰山风场#5 机组,采样间隔均为 10 min。本次实验软件平台为 Matlab2020a,硬件平台为 CPU:AMD 4800U 主频 1.8 GHz,内存 16 G。

4.1 数据描述

不同清洗方法效果可能受数据样本量的影响。当样本量过少时,数据无法体现风电机组的风速-功率分布特性^[15];反之,清洗时间长且清洗效果不会有明显提升。在参考其他风功率数据清洗的文献后,本文研究选择了半年的数据样本。3 个风场不同风电机组历史数据原始散点如图 2(a)~图 2(c)所示,其中灰色数据点为原始数据散点图,曲线为厂家提供的标准功率曲线。从图 2 可见,三台机组异常数据分布差异明显。

4.2 异常值检测结果

根据经验与现有文献[2,5,6,7,13],将四种方法的参数限定在一定范围内,然后在范围内多次实验得到每种方法的最优参数。四种方法的参数分别

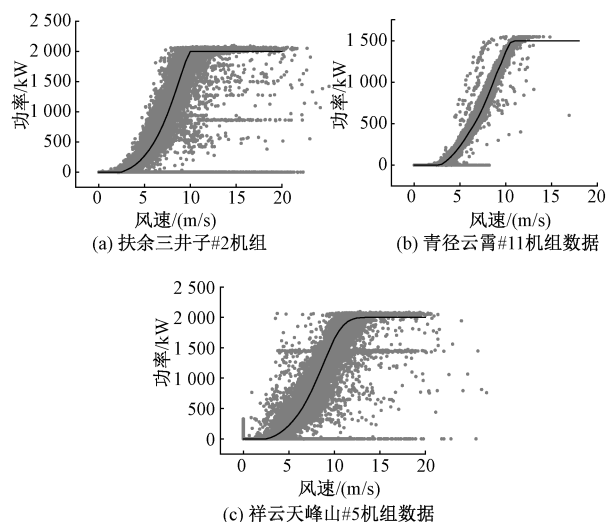


图2 原始数据散点图

Fig.2 Scatter plot of raw data

设置为:Thompson-tau 四分位法风速区间长度取 0.1 m/s ^[2]。DBSCAN 中两次四分位法区间间隔分别取 0.1 m/s 和 1.25% 额定功率, minpts 设置为 5, eps 设置为 2.5% 额定功率^[5]。LOF 中第 K 距离选择 10, LOF 阈值选择 1.5。IF 异常数据量设定为 20% , 树的数量为 100, 树枝为 256。

清洗结果分别如图3~图5所示。其中,浅灰色点为原始数据集,灰色点为剔除异常数据后的正常数据集。不同清洗方法之间的效果差异明显。

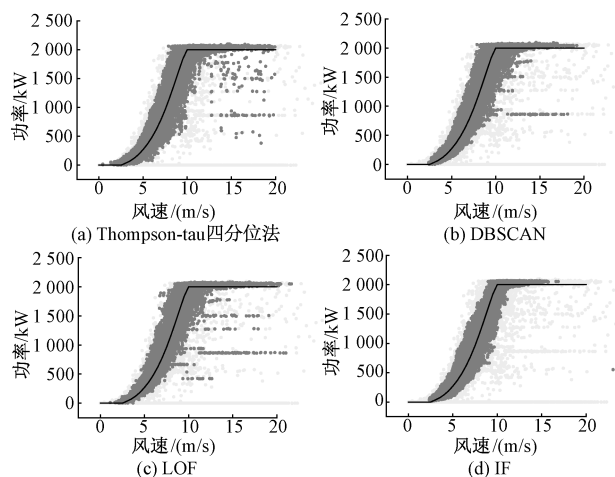


图3 三井子#2 机组清洗结果

Fig.3 Cleaning results of #2 plant in Sanjingzi wind farm

从图3和图5可见,三井子#2 机组和天峰山#5 机组异常数据分布广,同一风速区间下有多个异常数据与正常数据偏差较大,并且有很多聚集的异常数据,Thompson-tau 四分位法(后文简称 T-四分位法)、DBSCAN 和 LOF 无法完全准确识别这些异常

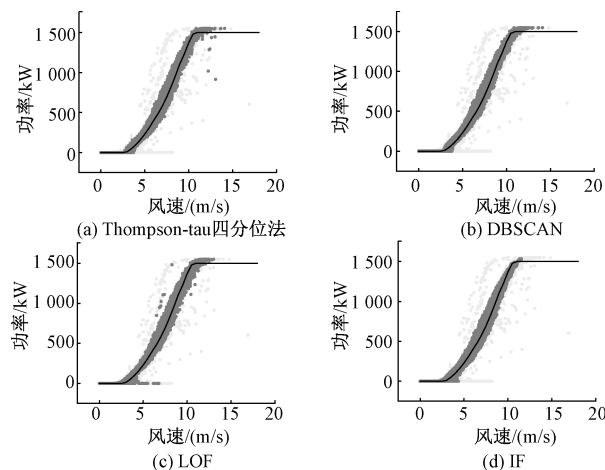


图4 云霄风场#11 机组清洗结果

Fig.4 Cleaning results of #11 plant in Yunxiao wind farm

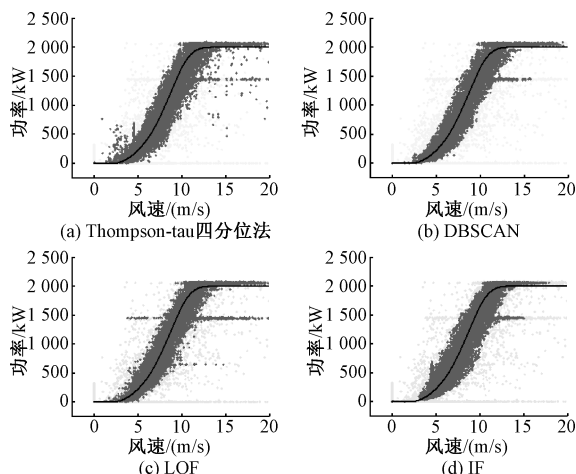


图5 天峰山#5 机组清洗结果

Fig.5 Cleaning results of #5 plant in Tianfengshan wind farm

数据。T-四分位法需要计算一段风速区间内功率的上下边界,当区间内有多个异常数据与正常数据相差较大时,下四分位数变小,上四分位数变大,相同显著性水平下数据区间变大。因此,功率的边界范围被拉大,导致部分异常数据落入边界内而被误识别为正常数据。LOF 与 DBSCAN 都通过判断数据周围的密度来区分异常数据和正常数据,其中 LOF 计算局部离群因子时不但要计算自身的密度,还要计算周围数据点的密度,因此若一个区域内异常数据与正常数据密度的差距不明显就会有大量异常数据误识别为正常数据或大量正常数据被误删除,所以这两种算法更适合处理异常数据分布稀疏的情况(见图4)。而当异常数据在一定区域内分布密集时,这两种方法无法区分异常数据和正常数据的密度差别。IF 在二叉树中对原数据集进行多次

划分将所有数据分离,偏离较远的数据在树中的孤立程度十分明显。但是从图 4 可见,边缘的正常数据也容易被孤立而出现误删除。

不同清洗方法的数据删除率和清洗时间如表 1~表 3 所示。其中数据删除率 $D\%$ 为:

$$D\% = \frac{S_0 - S_1}{S_0} \times 100\% \quad (9)$$

式中, S_0 为原数据集的数据量; S_1 为使用数据清洗方法剔除异常数据后剩余数据集的数据量。

表 1 三井子风场#2 机组数据清洗结果

Tab.1 Cleaning results of #2 plant in Sanjingzi wind farm

	数据删除率(%)	清洗时间/s
T-四分位法	43.04	0.29
DBSCAN	46.21	1.97
LOF	46.01	8.32
IF	37.96	5.57

表 2 云霄风场#11 机组数据清洗结果

Tab.2 Cleaning results of #11 plant in Yunxiao wind farm

	数据删除率(%)	清洗时间/s
T-四分位法	27.54%	0.27
DBSCAN	27.56%	2.89
LOF	48.65%	10.44
IF	26.45%	6.86

表 3 天峰山风场#5 机组数据清洗结果

Tab.3 Cleaning results of #5 plant in Tianfengshan wind farm

	数据删除率(%)	清洗时间/s
T-四分位法	35.23	0.29
DBSCAN	41.09	1.98
LOF	48.58	9.12
IF	33.19	8.05

根据表 1~表 3 可知,IF 的数据删除率最低,对原数据集完整性的破坏最小。T-四分位法的清洗时间最短。虽然 DBSCAN 和 LOF 都是基于密度的方法,但 DBSCAN 的时间复杂度通常小于 $O(N^2)$, LOF 的时间复杂度为 $O(N^3)$ 。因此,LOF 的清洗时间远大于 DBSCAN。

四种方法具有不同的数据删除率。对于三井子风场#2 机组与天峰山风场#5 机组,DBSCAN 与 LOF 的数据删除率较高,分别在 46%与 40%左右。这两种方法的数据删除率明显高于 T-四分位法与 IF,原因在于这两台机组的异常数据分布范围广,且密度较大。由前文分析可知,DBSCAN 与 LOF 对这类异

常数据分布情况的清洗效果较差;为了获得更好的异常数据清洗效果,只能将阈值的范围设置得更小,进而导致了更多正常数据被误删除。在云霄风场#11 机组上,LOF 的数据删除率远高于其他三种方法,主要原因是该机组异常数据分布更稀疏。根据前文分析,LOF 在计算局部离群因子时更易受到数据疏密程度的影响;要获得和 DBSCAN 相近的清洗效果,阈值的设定要比 DBSCAN 更严格,这就导致了 LOF 的数据删除率更大。为进一步分析四种方法的清洗效果,4.3 节对比分析了不同方法下的检测精度。

4.3 不同方法检测精度对比

检测并剔除异常数据后,分别计算不同方法清洗结果与标准功率曲线之间的离散程度,以此评价不同方法清洗性能。基于式(7)和式(8),得到计算结果如表 4~表 6 所示。

表 4 三井子风场#2 机组清洗结果与标准功率曲线的误差

Tab.4 Error between cleaning result and ideal power curve of #2 plant in Sanjingzi wind farm

	AAD	RMSE
T-四分位法	0.692 4	1.889 4
DBSCAN	0.056 9	0.727 4
LOF	0.215 3	0.883 1
IF	0.052 7	0.623 4

表 5 云霄风场#11 机组清洗结果与标准功率曲线的误差

Tab.5 Error between cleaning result and ideal power curve of #11 plant in Yunxiao wind farm

	AAD	RMSE
T-四分位法	0.025 2	0.492 7
DBSCAN	0.023 4	0.471 6
LOF	0.190 1	0.579 6
IF	0.015 7	0.301 6

表 6 天峰山风场#5 机组清洗结果与标准功率曲线的误差

Tab.6 Error between cleaning result and ideal power curve of #5 plant in Tianfengshan wind farm

	AAD	RMSE
T-四分位法	0.786 1	2.485 7
DBSCAN	0.267 6	1.646 7
LOF	0.431 0	1.617 2
IF	0.072 9	0.929 7

由表 4~表 6 能够明显看出,由 IF 得到的三台机组的清洗结果与标准功率曲线的误差也是最小的,即异常数据清洗更彻底。LOF 和 T-四分位法的误差明显高于另外两种方法。

综合来看,IF 虽然误删除了部分边缘的正常数据,但在四种方法中数据删除率最低,对机组数据完整性的破坏最小,清洗后的结果与标准功率曲线之间误差最小,在标准功率曲线附近最集中,并且对不同机组通用性高;T-四分位法更适用于对时间敏感的情况;DBSCAN 与 LOF 在异常数据分布稀疏时有更好的效果。

5 结 论

本文基于实际机组运行数据对现有的四种方法开展了对比研究,并得出各自优缺点:

(1)IF 算法具有最好的清洗效果,其次是 DBSCAN;Thompson-tau 四分位法程序运行时间最短;LOF 检测效果最差。

(2)DBSCAN 与 LOF 等基于突变点特征检测的方法受到异常数据分布情况影响较为严重,这两种适用于异常数据分布稀疏的情况。

(3)IF 与 Thompson-tau 四分位法等基于建立等效边界的方法会受到边界数据的影响。Thompson-tau 四分位法容易受到偏离程度大的异常数据影响,从而使阈值被拉大,导致部分异常数据落入阈值中,适用于对时间敏感或异常数据距功率曲线较近的情况。IF 对密集的异常数据具有良好的分离效果,并且对于不同机组的通用性高,但易将边缘的正常数据误删除,适用于大多数异常数据清洗的场合。

参考文献 (References):

- [1] 柯联锦,潘文霞,朱建红 (Ke Lianjin, Pan Wenxia, Zhu Jianhong). 一种基于风功率预测修正的储能电池容量需求分析 (Analysis of capacity requirements of battery energy storage based on forecasted wind power correction) [J]. 电工电能新技术 (Advanced Technology of Electrical Engineering and Energy), 2013, 32 (3): 57-61.
- [2] 邹同华,高云鹏,伊慧娟,等 (Zou Tonghua, Gao Yunpeng, Yi Huijuan, et al.). 基于 Thompson tau-四分位和多点插值的风电功率异常数据处理 (Wind power abnormal data processing based on Thompson-tau quartile and multi-point interpolation) [J]. 电力系统自动化 (Automation of Electric Power Systems), 2020, 44 (15): 156-162.
- [3] 杨茂,杨琼琼 (Yang Mao, Yang Qiongqiong). 基于云分段最优熵算法的风电机组异常数据识别研究 (Research on abnormal data recognition of wind turbines based on cloud segmented optimal entropy algorithm) [J]. 中国电机工程学报 (Proceedings of the CSEE), 2018, 38 (8): 2294-2301.
- [4] 娄建楼,胥佳,陆恒,等 (Lou Jianlou, Xu Jia, Lu Heng, et al.). 基于功率曲线的风电机组数据清洗算法 (Data cleaning algorithm for wind turbines based on power curve) [J]. 电力系统自动化 (Automation of Electric Power Systems), 2016, 40 (10): 116-121.
- [5] Yan J, Zhang H, Liu Y, et al. Uncertainty estimation for wind energy conversion by probabilistic wind turbine power curve modelling [J]. Applied Energy, 2019, 239: 1356-1370.
- [6] Zhao Y, Ye L, 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.
- [7] Zhang Y, Yang G, Hao X, et al. Research on identification and processing method for abnormal data of residential electric power consumption [A]. 2019 IEEE 3rd International Electrical and Energy Conference [C]. Beijing, China, 2019. 900-904.
- [8] 范晓泉,杜大军,费敏锐 (Fan Xiaoquan, Du Dajun, Fei Minrui). 风电异常测量数据智能识别方法研究 (Research on the intelligent identification method for abnormal measurement data of the wind power) [J]. 仪表技术 (Instrumentation Technology), 2017, (1): 10-14.
- [9] 杨茂,杨春霖,杨琼琼,等 (Yang Mao, Yang Chunlin, Yang Qiongqiong, et al.). 计及风向信息的风电功率异常数据识别研究 (Research on the identification of wind power anomaly data taking into account wind direction information) [J]. 太阳能学报 (Acta Solar Energy), 2019, 40 (11): 3265-3272.
- [10] 沈小军,付雪姣,周冲成,等 (Shen Xiaojun, Fu Xuejiao, Zhou Chongcheng, et al.). 风电机组风速-功率异常运行数据特征及清洗方法 (Data characteristics and cleaning methods of abnormal wind speed-power operation of wind turbines) [J]. 电工技术学报 (Transactions of China Electrotechnical Society), 2018, 33 (14): 3353-3361.
- [11] Long H, Sang L, Wu Z, et al. Image-based abnormal data detection and cleaning algorithm via wind power curve [J]. IEEE Transactions on Sustainable Energy, 2020, 11 (2): 938-946.
- [12] 胡阳,乔依林 (Hu Yang, Qiao Yilin). 基于置信等效边界模型的风功率数据清洗方法 (Wind power data

- cleaning method based on confidence equivalent boundary model) [J]. 电力系统自动化 (Automation of Electric Power Systems), 2018, 42 (15): 18-23, 149.
- [13] 李新鹏, 高欣, 阎博, 等 (Li Xinpeng, Gao Xin, Yan Bo, et al.). 基于孤立森林算法的电力调度流数据异常检测方法 (Power dispatch flow data anomaly detection method based on isolated forest algorithm) [J]. 电网技术 (Power System Technology), 2019, 43 (4): 362-371.
- [14] 韩文卓, 王方雨, 戈志华 (Han Wenzhuo, Wang Fangyu, Ge Zhihua). 基于广义最小二乘的风力机风速-功率曲线拟合方法研究 (Research on wind speed-power curve fitting model of wind turbine based on generalized least squares method) [J]. 电工电能新技术 (Advanced Technology of Electrical Engineering and Energy), 2018, 37 (9): 67-73.
- [15] 钟嘉庆, 李茂林, 江静, 等 (Zhong Jiaqing, Li Maolin, Jiang Jing, et al.). 基于 Copula 理论的风/光出力预测误差分析方法的研究 (Method of wind/solar output forecast error analysis based on Copula theory) [J]. 电工电能新技术 (Advanced Technology of Electrical Engineering and Energy), 2017, 36 (6): 39-46.

Comparative study on detection methods of wind power abnormal data

FENG Zhuo-wen¹, ZHU Shi-ping¹, ZHAO Zhi-hua¹, SUN Ming-ren²,
DONG Mi², SONG Dong-ran²

(1.Hunan Electric Power Design Institute Co., Ltd. CEEC, Changsha 410007, China;

2.School of Automation, Central South University, Changsha 410083, China)

Abstract: Wind power prediction and power curve modeling of wind turbines rely on historical operating data. However, a large amount of abnormal data accumulated in historical data makes it difficult to carry out the above-mentioned work effectively. Scholars at home and abroad have proposed a variety of abnormal data detection methods, but there is still a lack of overall understanding of the advantages and disadvantages of different methods and their applicable occasions. To this end, this paper compares four common wind power outlier detection methods, including density-based clustering algorithm, local outlier factor algorithm, Thompson-tau quartile method and isolated forest. In order to evaluate different detection methods, an evaluation index based on power curve modeling error is proposed. The experimental results show that the isolated forest algorithm has higher accuracy than the other three methods, can deal with differently distributed abnormal data, and has a shorter cleaning time.

Key words: wind power data; outlier characteristics; data cleaning; isolated forest