**基于孤立森林算法的风电场异常数据识别方法研究**

周寰育[[1]](#footnote-0)，邱颖宁1，冯延晖1

南京理工大学能源与动力工程学院，江苏 南京 210094

摘要

随着我国风电领域的快速发展，许多已经投产的风电场产生大量具有研究价值的历史数据。通过对风电机组运行数据的分析可以对风机运行状态进行判断和预测，及时发现故障，对机组运行的安全性和经济性有重要的意义。风电机组的数据分析建立在SCADA系统采集到的数据基础上，但采集的原始数据中往往包含大量的无效异常数据。由异常数据和正常数据混合的原始数据进行分析会导致分析结果产生很大偏差，因此本文基于孤立森林算法实现对风电场SCADA数据的以无监督方式进行异常数据识别和清洗。

关键词

孤立森林算法 风电场 异常数据识别

**1 引言**

在2020-2024年这五年中，全球计划新增风电装机的一半以上将来自中国和美国这两大市场，而全球风电累计装机将在2024年之前突破1TW[1]。我国无论是海上新装机容量还是陆上新装机容量均居世界第一的位置[2]。目前我国存在部分风电场发电效率未能达到预期目标，其中原因包括尾流影响、天气影响、风电场选址规划影响等多方面因素。为了找到影响风电场发电效率的原因，需要对风电场的历史数据进行分析和研究。

随着风电发展技术的日益成熟,越来越多的国家将风电作为从传统能源向可再生能源转变以及应对全球气候变化的重要战略。我国存在大量已经投入运行的风电场。通过这些已经运行的风电场产生的历史数据可以指导后续风电场的优化选址，风机布局，尾流控制等多方面研究工作。风电机组的数据分析建立在SCADA系统采集到的数据基础上，然而由于停机限电或极端恶劣天气以及数据采集设备异常等原因，风电场历史数据中包含较多的异常数据。这类数据通常没有分析价值，如果使用未经处理的原始数据进行分析会导致分析结果产生很大偏差，对运行状态的判断造成很大影响,因此,需要对原始SCADA 数据进行异常数据的识别和清洗。

蒋勋等人以大数据为背景，提出了一种数据清洗方法。针对大数据的多维度性，给出了面向知识服务数据清洗的基本框架，并对其局限性进行了分析[3]。文献颜永龙等人基于反向传播神经网络和最小二乘支持向量机，引入信息熵建立组合预测模型，对风电机组状态参数进行异常检测，在不重复建立参数模型情况下，可在不同风电机组中进行状态参数异常检测，而参数模型的建立需要多次训练测试[4]。包从剑利用切比雪夫原理设定一个阚值区间，通过这个阈值区间来检测“噪音数据”，并采用遗传神经网络的预测模型来填补空缺值，以实现用户用电量的数据清洗[5]。Madnick等人提出了一种基于基于功率曲线的风电机组数据清洗算法，组内最优方差清洗算法。但该方法需要对较大的采样数据进行分析才准确，并且该方法只能剔除部分偏差较大的异常点，而且不能进行异常值修正和缺失值插补，对于一些要求连续性的数据不适用[6]。因此，对于风电机组SCADA数据清洗方法需要做更深入的研究。

**2 风电场数据集描述及算法原理**

**2.1风电场数据集描述**

本文采用位于内蒙古某平原地区的风电场数据集进行研究分析，风电场由33台相同型号的1500MW风机构成。其风电场内各风机位置关系如图1所示，图中圆圈代表以风机自身为原点三倍直径范围画圆。

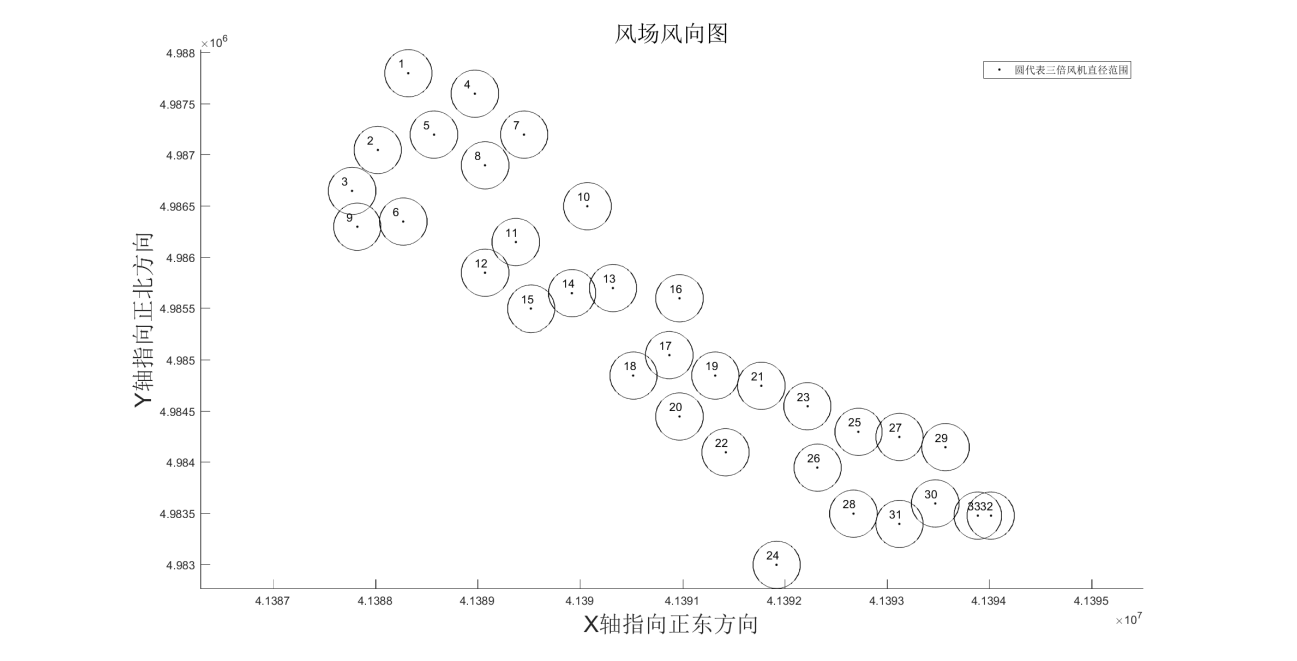


图1 风电场风机位置示意图

每台机组的SCADA数据集中包含由厂家监测和收集以10min为记录间隔的15维包含风电机组的工作状况、外部环境和运行状态等数据。共记录2021年3月至9月半年的SCADA数据。其中SCADA包含的数据类型如表1所示

表1 原始SCADA数据特征类型

|  |  |  |  |
| --- | --- | --- | --- |
|  | 数据特征 |  | 数据特征 |
| 1 | 有功功率 | 9 | 电网电压 |
| 2 | 无功功率 | 10 | 电网频率 |
| 3 | 桨叶角度 | 11 | 实际功率 |
| 4 | 实际风速 | 12 | 机舱位置 |
| 5 | 绝对风向 | 13 | 齿轮箱转速 |
| 6 | 发电机转速 | 14 | 变频器转速 |
| 7 | 环境温度 | 15 | 给定力矩 |
| 8 | 实际力矩 |  |  |

**2.2孤立森林算法原理**

孤立森林算法是基于 Ensemble 的异常检测方法，因此具有线性的时间复杂度。且精准度较高，在处理大数据时速度快，所以目前在工业界的应用范围比较广。针对于不同类型的异常，要用不同的算法来进行检测，而孤立森林算法主要针对的是连续型结构化数据中的异常点。

Maurizio Collu等人利用孤立森林算法和深度学习神经网络，通过高频SCADA数据进行功率预测。在预测模型中风速、机舱方向、偏航误差、叶片俯仰角和环境温度被视为输入特征，而风力被视为输出特征[7]。Conor McKinnon等人通过支持向量机、孤立森林等算法对兆瓦级风力发电机进行异常状态检测，结果表明支持向量机与孤立森林的识别异常准确率比较近似[8]。James Carroll等人通过孤立森林算法比较了三种不同时间序列窗口长度下检测风电机组变桨系统的异常，发现该算法可以在所有不健康风机的液压和电动变桨系统发生故障前大约12至18个月检测到异常活动，并且在故障前的一段时间内，可以发现异常活动呈上升趋势[9]。

使用孤立森林的前提是，将异常点定义为那些 “容易被孤立的离群点” 可以理解为分布稀疏，且距离高密度群体较远的点。从统计学来看，在数据空间里，若一个区域内只有分布稀疏的点，表示数据点落在此区域的概率很低，因此可以认为这些区域的点是异常的。孤立森林算法的理论基础有两点：1.异常数据占总样本量的比例较小；2.异常点的特征值与正常点的差异较大。

孤立森林中的 “孤立”(isolation)指的是 “把异常点从所有样本中孤立出来”。大多数基于模型的异常检测算法会先 ”规定“ 正常点的范围或模式，如果某个点不符合这个模式，或者说不在正常范围内，那么模型会将其判定为异常点[10]。

孤立森林包括以下四个特点：1.在训练过程中，每棵孤立树都是随机选取部分样本；2.不同于 KMeans、DBSCAN 等算法，孤立森林不需要计算有关距离、密度的指标，可大幅度提升速度，减小系统开销；3.因为基于 Ensemble，所以有线性时间复杂度。通常树的数量越多，算法越稳定；4.由于每棵树都是独立生成的，因此可部署在大规模分布式系统上来加速运算[11]。

该算法的思想即用一个随机超平面对一个数据空间进行切割，生成子空间。不断循环直到每子空间里面只包含一个数据点为止。如图2所示，蓝色X2表示待切割数据空间中的密集区域数据点，而红色X1表示切割完成的数据空间中稀疏区域的数据点。该算法利用将数据点不断切割为独立数据点空间的次数作为评价指标。因此异常数据由于分布稀疏，其被切割的次数会明显小于正常数据。因此孤立森林算法可以基于无监督学习的方法实现对于异常数据的识别和清洗[12]。

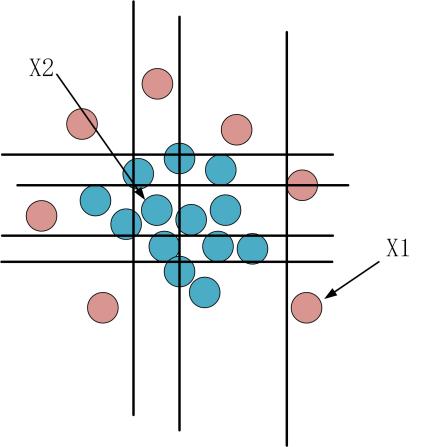


图2 孤立森林分割数据空间操作

由于该算法效率高、不需要进行预训练。孤立森林算法对异常数据的定义与风电场中对异常机组和异常数据点的定义相符。因此，利用该算法通过风电场历史SCADA数据对存在异常的风机以及存在异常状态的时刻点的数据进行无监督识别。

**3 基于孤立森林的风电场异常数据识别方法**

异常数据识别与具体应用场景紧密相关。风电场异常数据识别在特征选择时，要尽量过滤不相关的特征。因此选择与风电机组运行状态密切相关以及与风况采集设备相关的数据特征进行分析。因此从原始SCADA数据中4个主要特征保留，如表2所示。

表2 保留主要特征类型

|  |  |
| --- | --- |
| 1 | 有功功率 |
| 2 | 实际风速 |
| 3 | 绝对风向 |
| 4 | 机舱位置 |

基于孤立森林算法对风电场异常数据的识别分为两个部分，分别对应不同异常原因进行识别，最终实现保留正常数据的目的，其流程图如图3所示。

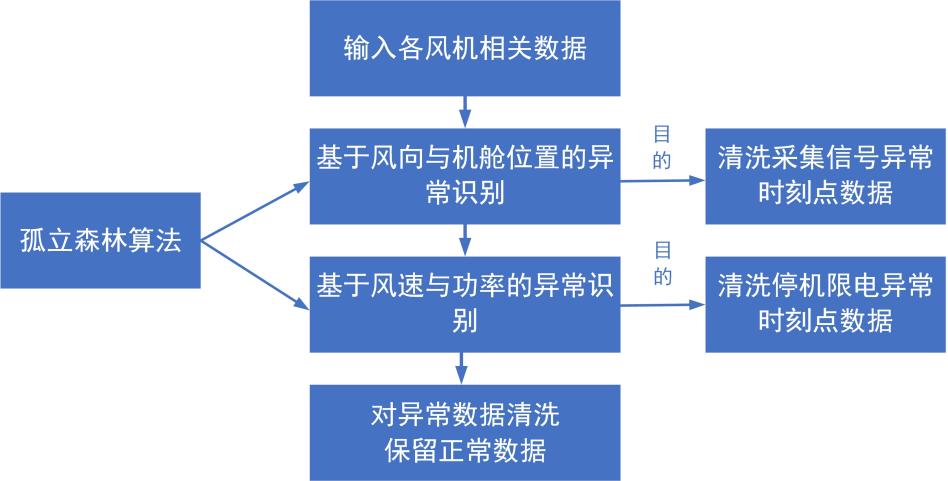


图3 基于孤立森林的风电场异常数据识别流程

**3.1基于风向与机舱位置的异常数据识别**

首先对风电场内风电机组的风向分布进行分析。分别绘制各风机风向-功率玫瑰散点图，如图4所示(仅展示三台风机,其他风机分布与这三台相似)

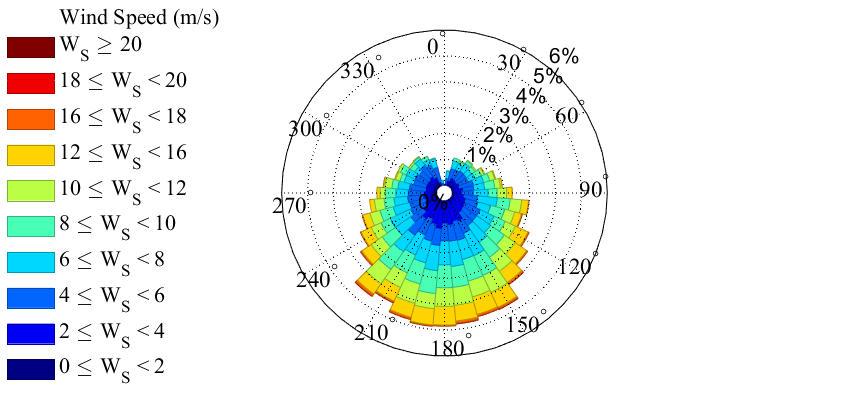


图4(a) 1号风机玫瑰图

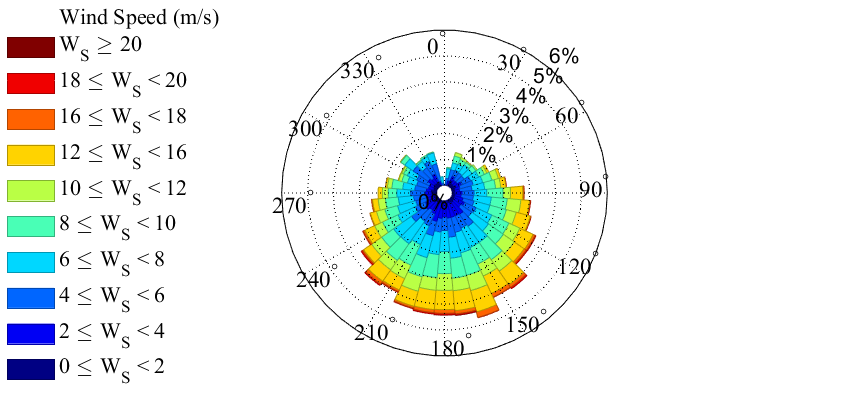


图4(b) 12号风机玫瑰图

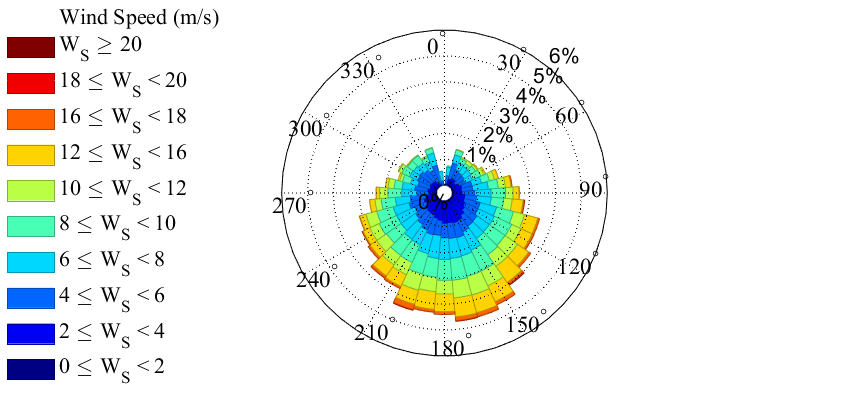


图4(c) 32号风机玫瑰图

通过各风机玫瑰图可以观察到全部风机均出现在0°附近数据缺失情况。通过对比各风机SCADA数据发现由于风机SCADA采集设备的错误记录，将风机10分钟内的平均风向和机舱位置进行了简单的平均值处理，而忽略了实际风向以及机舱位置角度值所代表的空间方位。例如表3所示，某些时刻10分钟内风向和机舱位置的平均值不处于最大值最小值范围内。

表3 某时刻风向与机舱位置最大最小平均值

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 风向最大值 | 风向最小值 | 风向平均值 | 机舱位置最大值 | 机舱位置最小值 | 机舱位置平均值 |
| 时刻1 | 355.7° | 8.2° | 179.8° | 349.8° | 7.3° | 175.2° |
| 时刻2 | 339.8° | 10.8° | 251.2° | 342.2° | 5.5° | 241.2° |

对于因采集设备计算错误导致产生异常数据，并于正常数据相互掺杂的现象，通过人工方法无法应对大量数据的异常数据识别和清洗。因此通过孤立森林算法可以实现快速高效的识别异常数据。孤立森林算法流程如图5所示

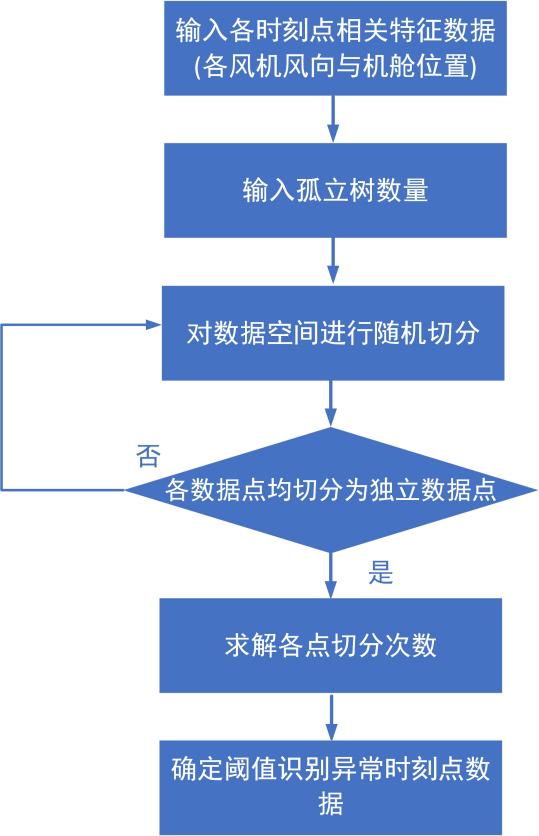


图5 基于风向与机舱位置的异常数据识别

最终通过孤立森林算法识别结果通过风向-机舱位置关系图可视化，如图6所示，从图中可以发现对于明显原理密集区域的离群点，孤立森林可以很好的将其识别出来，并将其进行清洗，保留正常数据。并统计出异常数据点较多的五台风机，如表4所示。

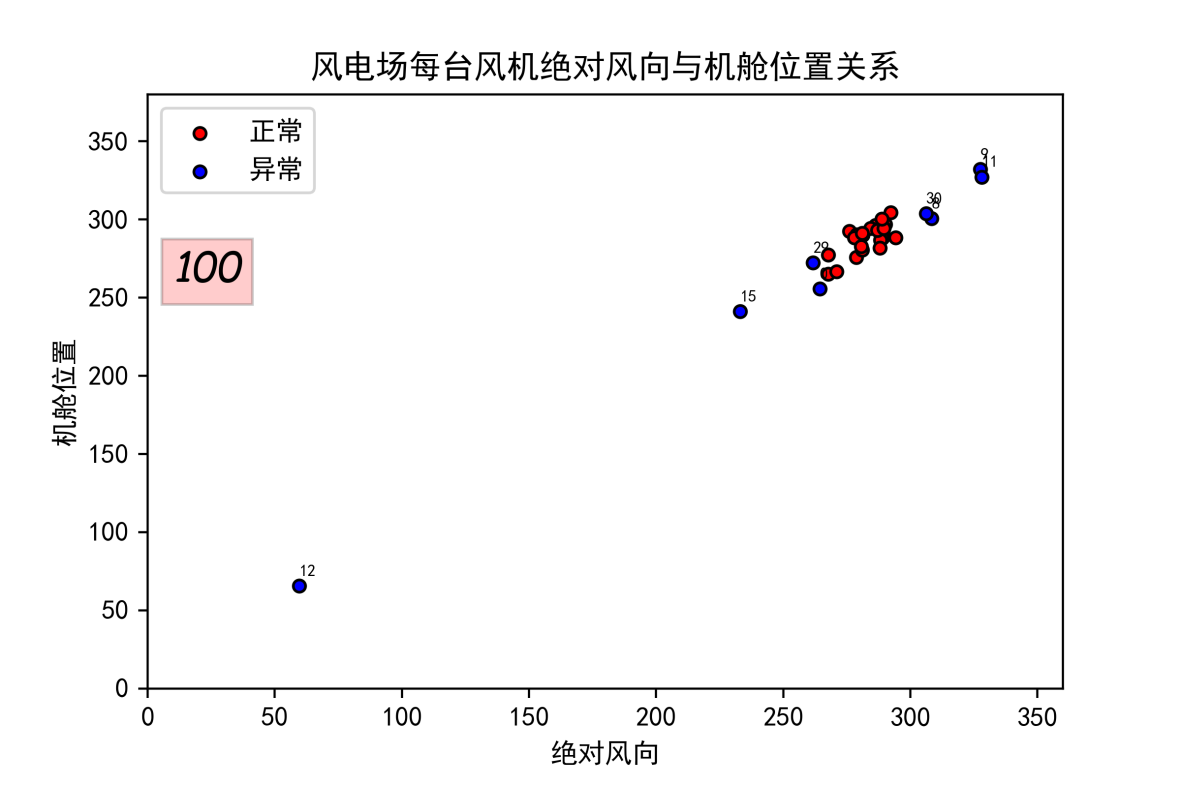


图6 孤立森林算法结果可视化

表4 异常风机时刻出现次数与频率

|  |  |  |
| --- | --- | --- |
| 风机编号 | 出现异常次数 | 出现异常频率 |
| 13 | 19619 | 91% |
| 16 | 18281 | 85% |
| 10 | 8564 | 40% |
| 12 | 7828 | 36% |
| 31 | 1628 | 7.6% |

**3.2基于风速与功率的异常数据识别**

当基于风向与机舱位置的异常识别完成后，需要针对停机限电原因的异常数据进行清洗。未清洗前的风速-功率关系散点图如图7所示，蓝色框包含的为停机限电引起的不同风速条件下功率均不发生改变，且始终低于该风速对应的理论功率。这部分数据会对风电场数据分析产生干扰。因此需要进行异常识别及清洗。

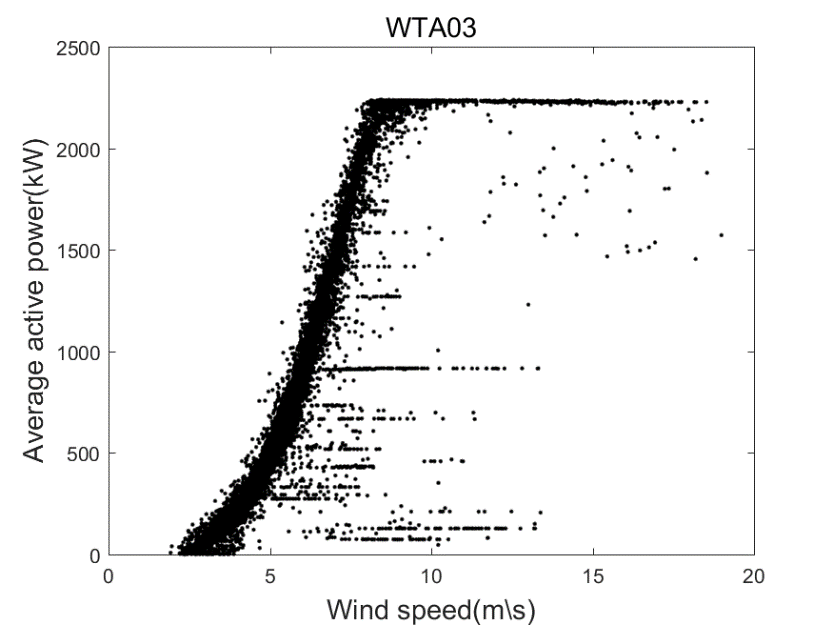


图7 清洗前功率散点图

以往普遍的异常数据识别方法需要使用风机半径、风机额定功率、风机转速等信息基于经验公式进行异常识别。但该方法对于大批量数据而言其识别成本较高，识别时间较慢。因此通过孤立森林算法实现无监督的异常识别。其流程如图8所示。

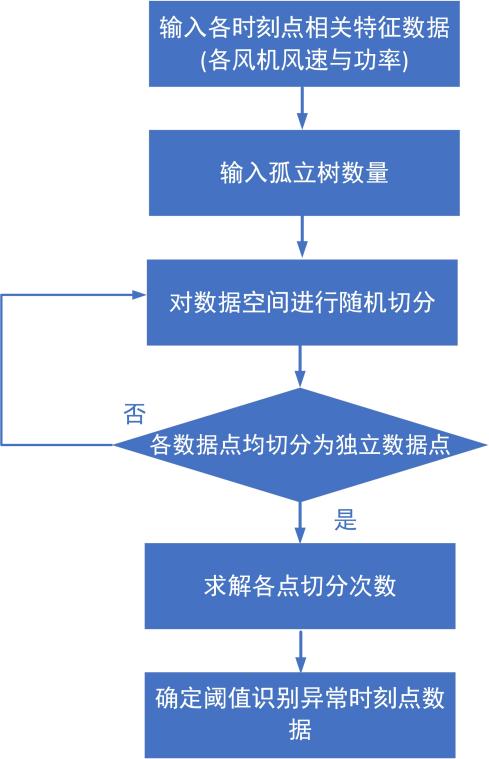


图8 基于风速与功率的异常数据识别

基于风向与功率的异常数据识别结果如图9所示，图中可以发现其停机限电数据被完全识别且清洗完成。

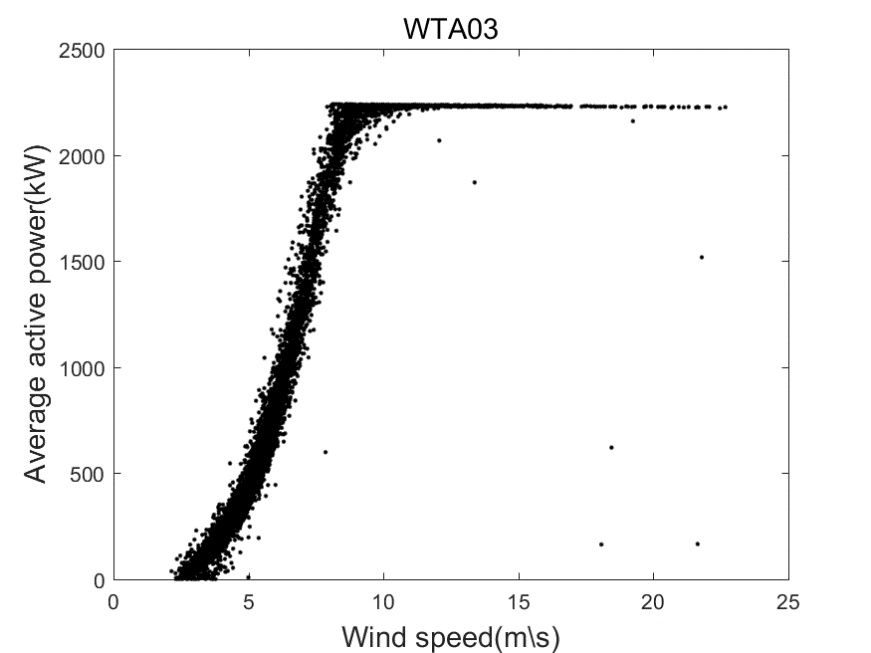


图9 清洗后功率散点图

**4 结论与展望**

通过风向数据与机舱位置数据采用孤立森林算法可以有效识别因SCADA采集计算错误导致的异常数据。并统计出现异常较多的风机组，对于后续分析风电场发电效率降低原因分析具有指导意义。风速数据和功率数据采用孤立森林算法可以在不获取额外的风机信息的情况下，有效识别因停机限电导致的异常数据。通过孤立森林算法将不利于风电场数据分析的异常数据清洗，可以明显提高数据的质量。通过高质量数据可以实现风电场影响因素分析，找出影响发电效率的主要原因。后续的研究可以利用风轮转速等高维数据实现风电机组不同运行状态的无监督识别，以及引入评价指标对算法的无监督识别有效性进行合理监控。

参考文献

1. Global Wind Report 2021.Global Wind Energy Council (GWEC).2021.
2. 乐威.新能源背景下我国风力发电现状和未来发展方向探索[J].绿色环保建材,2020(11):165-166.
3. 蒋勋,刘喜文.大数据环境下面向知识服务的数据清洗研究[J].图书与情报, 2013(5):6.
4. 颜永龙，李剑，李辉，等．采用信息熵和组合模型的风电机组异常检测方法[J]．电网技术，2015，39(3)：737-743．
5. 包从剑.数据清洗的若干关键技术研究[D].江苏大学,2007.
6. Wang,R.Y.;Kon,H.B.;Madnick,S.Data quality requirements analysis and modeling[J].In: Proceedings of the 9th International Conference on Data Engineering. Vienna: IEEE Computer Society,1993,670-677
7. LIN Zi,LIU Xiaolei,COLLU M，Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks[J].International Journal of Electrical Power & Energy Systems，2020(118):105835．
8. McKinnon C, Carroll J, McDonald A, et al. Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data[J]. Energies, 2020, 13(19): 5152.
9. McKinnon C, Carroll J, McDonald A, et al. Investigation of isolation forest for wind turbine pitch system condition monitoring using SCADA data[J]. Energies, 2021, 14(20): 6601.
10. Liu F T,Kai M T,Zhou Z H.Isolation forest[C]//Eighth IEEE International Conference on Data Mining．IEEE Computer Society,2008：413-422．
11. 季一木，张永潘，郎贤波，等．面向流数据的决策树分类算法并行化[J]．计算机研究与发展，2017，54(9)：1945-1957．

[12] Liu F T,Kai M T,Zhou Z H.Isolation-based anomaly detection [J].ACM Transactions on Knowledge Discovery from Data，2012,6(1)：1-39．

1. 基金项目：国家重点研发计划，项目编号；2019YFE0104800，项目名称：海上风电场智能运行控制技术研究 [↑](#footnote-ref-0)