文章编号：1673-5196(2021)06-0050-06

基于QM-DBSCAN的风力机数据清洗方法

郑玉巧*，刘玉涵，何正文，董 博，魏剑峰

（兰州理工大学 机电工程学院，甘肃 兰州 730050）

摘要：针对风电场风速-功率异常数据难以清洗的问题，提出一种基于QM-DBSCAN算法的风电场数据清洗方法. 首先选取最能代表风力机运行状况的风速-功率数据作为研究对象，根据异常数据的分布特征进行分类；然后分别利用四分位法、标准DBSCAN算法及基于QM-DBSCAN方法识别和剔除异常数；最后通过spearman系数进一步验证所提方法的有效性. 研究结果表明：QM-DBSCAN方法的剔除效果最好，较四分位法和标准DBSCAN法的spearman系数分别提高0.003 5和0.004 7.

关键词：风力机；异常数据清洗；四分位法；DBSCAN；QM-DBSCAN

中图分类号：TK83 **文献标志码**：A

A novel cleaning method for wind turbine data based on QM-DBSCAN

ZHENG Yu-qiao, LIU Yu-han, HE Zheng-wen, Dong Bo, Wei Jian-feng

(School of Mechanical and Electrical Engineering, Lanzhou Univ. of Tech., Lanzhou 730050, China)

Abstract: For the issue of wind Speed-Power data hard cleaning in wind farms, a novel method based on QM-DBSCAN is proposed. Firstly, the wind condition which can best represent the operating state of the wind turbine is selected as the research object, and the anomalous data are classified according to the distribution characteristics. Then, the quartile method, standard DBSCAN algorithm and QM-DBSCAN method were used to identify and eliminate the abnormal data. The Spearman correlation coefficient was adopted to verify the effectiveness of the proposed method. The results indicated that the QM-DBSCAN method had the best elimination effect, which was 0.003 5 and 0.004 7 higher than the quartile method and the DBSCAN method, respectively.

Key words: wind turbine; abnormal data cleaning; the Quartile Method; DBSCAN; QM-DBSCAN

随着环境污染与能源危机等问题日益成为全球关注的焦点，风力发电作为最具开发与商业化前景的可再生能源发电形式越来越受到各国广泛关注[1-2]. 风电功率的周期性、随机性及间歇性等特征导致大规模风电并网对电力系统的影响越来越明显. 利用数据挖掘与机器学习等算法研究风力发电的风速-功率规律，以解决风电并网对电力系统的影响已日益成为风电领域的一项重要课题. 获得准确的风速-功率记录数据是前述课题的重要基础.

风电场所安装的数据采集与监测控制系统(supervisory control and data acquisition, SCADA)是风电场的风速-功率记录数据来源，该记录数据对风速-功率相关研究尤为重要. 但在实际发电过程中，由于电力系统消纳能力的限制，使得风电场需要弃风，导致SCADA系统记录的风速-功率数据存在大量异常数据簇. 这些数据簇具有集中、横向分布等特点，因此不能有效表征风力机实际运行过程中风速与功率间的关联关系. 由此可见，如何清洗风速-功率数据中大量异常数据簇尤为关键.

诸多学者已对风速-功率数据清洗进行了大量研究. 娄建楼等[3]提出最优组内方差清洗算法对风速-功率数据进行清洗，改变了传统方法对多维度数据的依赖性. 赵永宁等[4]采用四分位法剔除风速-功率数据中的分散型数据，并使用k均值聚类剔除堆积型数据. 朱倩雯等[5]在研究风速-功率数据时序重构时同样采用四分位法清洗了约22.67%的异常数

收稿日期：2021-03-24

基金项目：国家自然科学基金(51965034)，兰州市人才创新创业项目(2018-RC-25)

通讯作者：郑玉巧(1977-)，女，甘肃庄浪人，副研究员，博导. Email: zhengyuqiaolut@163.com

据,四分位法清洗效果显著. 邹同华等[6]研究表明采用 Thompson tau-四分位法清洗风速-功率异常数据簇效果显著,有效解决了风电机组风速-功率异常数据处理方法清洗时间长、模型复杂的问题. 张小奇等[7]在发电功率计算研究的过程中使用 k 均值聚类清洗数据,并取得了良好的效果. 丁明等[8]在研究风力机发电量短期预测问题中使用自组织神经网络对功率数据进行清洗,进一步提升了风力机发电量短期预测精度. 杨茂等[9]在风速-功率数据中考虑风向变化的同时还对上升风与下降风进行了有效区分,据此建立异常数据识别模型,该成果明确了异常数据簇的类型. 张东英等[10]对风速-功率数据中的时序限风区间进行识别,全天整体识别率超过 70%. Shen 等[11]将风速-功率异常数据簇分为 4 类,并采用分组进行的四分位法剔除异常数据. Long 等[12]对风速-功率数据进行二值图像的转化,并建立数据点与图像像素间的映射关系,最后通过数学形态学方法(MMO)提取异常数据,这种方式效果显著. Ying 等[13]通过散点图拟合出风速-功率曲线,将曲线之外的数据识别为异常数据,该方式区分接近正常数据的异常数据簇的准确度有待进一步提高.

基于上述研究,针对在清洗风速-功率数据工作中正常数据与异常数据接近时以及产生密集堆积异常数据簇时出现清洗失效的情况,本文提出了一种用于识别清洗风速-功率数据的 QM-DBSCAN 聚类算法. 该算法根据异常数据簇类别与特点进行区分,采用 QM-DBSCAN 聚类算法对风速-功率数据展开数据清洗工作. 最后将该算法与广泛认同的四分位法、标准 DBSCAN 算法的识别清洗效果进行比较,验证所提方法的有效性.

1　数据清洗方法

1.1　四分位法

四分位数是指将一个排好序的数据样本平均划分成四部分的 3 个数据点,每部分所包含的数据量是整个数据样本数据量的 25%. 一个升序排列的样本集合 $X=[x_1,x_2,\cdots,x_n]$,四分位法(quartile method,QM)求解方法如下[14]:

1) 计算第二个四分位数 Q_2

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

2) 计算第一个和第三个四分位数 Q_1 及 Q_3

当 $n=2k(k=1,2,\cdots)$时,从 Q_2 将样本 X 分为两个子序列,分别求解两子序列的中位数 Q_2'、Q_2'',则有 $Q_1=Q_2'$,$Q_3=Q_2''$.

当 $n=4k+3(k=1,2,\cdots)$时,有

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

当 $n=4k+1(k=1,2,\cdots)$时,有

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

四分位距表示为

$$I_{QR}=Q_3-Q_1\tag{4}$$

根据四分位距可确定样本 X 中异常值的内限为

$$[R_1,R_2]=[Q_1-1.5I_{QR},Q_3+1.5I_{QR}]\tag{5}$$

超出内限$[R_1,R_2]$的值都为异常值.

1.2　基于密度的带噪声空间聚类算法

基于密度的带噪声空间聚类算法(density-based spatial clustering of applications with noise,DBSCAN)是一种典型的密度聚类算法. 该算法可发现任意形状和数量的簇类,通过聚类将样本数据划分成不同的簇类,依此可判别出风电机组 SCADA 数据中的异常数据. DBSCAN 算法中涉及两个重要参数,分别为邻域半径 Eps 和邻域内最少包含点数 Minpts. 给定一组风电机组 SCADA 数据点集合 $H=\{h_1,h_2,\cdots,h_n\}$,相关定义如下:

Eps 邻域:表示集合 H 中任意一点 p 的邻域半径 Eps 范围内点的集合.

核心点:若集合 H 中任意一点 p 的 Eps 邻域内至少包含了 Minpts 个数据点,则将点 p 标记为核心点.

边界点:不属于核心点,但属于在某个核心点的 Eps 邻域内的数据点.

噪声点:既不属于核心点,也不属于边界点的数据点,即异常点.

以 Minpts=4 为例,DBSCAN 算法聚类过程示意图如图 1 所示.

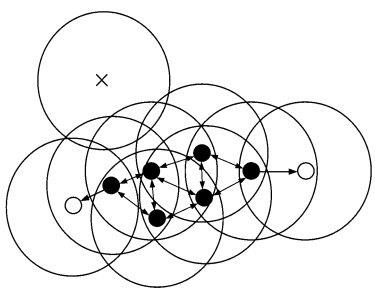

图 1　DBSCAN 聚类示意图

Fig. 1　The schematic diagram of DBSCAN method

图 1 中，黑色圆圈代表核心点，白色圆圈代表边界点，×代表噪声点. 将数据点集合 H 中全部数据点对象标记为未访问状态，通过 DBSCAN 算法随机选取一个未访问的数据点 p 开始聚类；将 p 记为访问点，检查 p 的邻域半径 Eps 内是否至少包含 Minpts 个数据点对象，若包含，则将点 p 标记为核心点，核心点 p 创建一个新簇 C，并将待选集合 S 中不属于其他簇的数据点对象添加至簇 C 中，直到集合 H 中的数据点全部访问完毕；最后将未添加的数据点划分为噪声点，即异常点.

2 风电场异常数据的分类

以国内某风电场 F24 号风电机组为研究对象，选取 2018 年 7 月 1 日零时至 2019 年 7 月 1 日零时的 SCADA 数据为基础数据，提取风速-功率数据总计 54 607 组，散点图如图 2 所示.

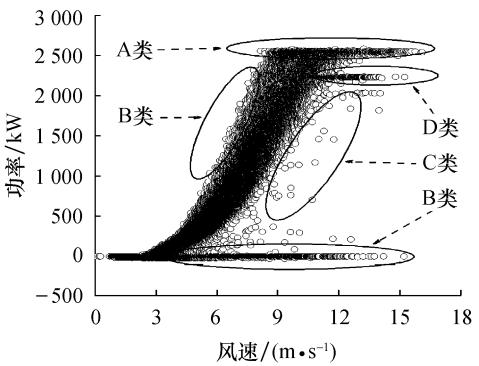

图 2 风力机异常数据示意图

Fig. 2 The display of wind turbine abnormal data

图 2 中根据异常数据的分布特点将异常数据大致分为 4 类，分别表示为 A 类数据、B 类数据、C 类数据及 D 类数据. 其中，A 类数据为超限数据，即记录功率值超过额定功率；B 类数据为错误数据，即高风速下记录功率值为 0 或负值；C 类数据为分散异常数据，即无规律、低密度分散在风功率曲线附近的记录功率值；D 类数据为弃风限电数据，即 1 条及以上的横向密集散点，散点靠近风功率曲线附近时难以与正常数据分离.

3 案例分析

整理国内某风电场 F24 号风电机组由 2018 年 7 月 1 日零时至 2019 年 7 月 1 日零时的风速-功率数据记录. 其中，A 类数据共计 1 775 组，B 类数据共计 9 522 组，总计 11 297 组数据. 由于 A 类数据、B 类数据的分布较集中且易于清洗，所以可直接剔除.

3.1 四分位法的应用

对于含有 C 类数据、D 类数据的风速-功率数据序列，为得到显著清洗效果，可分别从横向和纵向两个维度依次进行清洗.

1) 横向清洗

剔除 A 类数据、B 类数据后的风速-功率数据序列总计 43 309 组，按照风电机组输出功率从小到大的顺序对该数据重新进行排序，排序后的数据组记为 N. 然后以 50 kW 的间隔对数据组 N 等距划分区间，分别记为 N_i（$i=1,2,3,\cdots,n;n=50$）.

在每个子数据组 N_i 内按照风速从小到大的顺序再次重新进行排序. 根据式(1～3)分别计算每个新子数据组内的 Q_1 位置、Q_2 位置及 Q_3 位置，根据式(4)和式(5)确定 R_1 位上限和 R_2 位下限位置，以此确定正常数据范围，将范围外的数据视为异常值进行剔除. 通过计算得到异常数据共计 223 组，异常值剔除后的风速-功率散点图如图 3 所示.

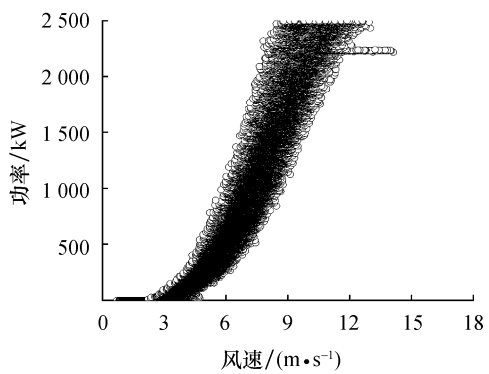

图 3 横向清洗后风速-功率散点图

Fig. 3 The wind speed-power scatter after transverse cleaning by the quartile method

由图 3 可知，经过横向清洗后多数 C 类数据已被剔除，但仍存在部分 C 类数据和 D 类数据未被剔除，因此对横向清洗后的风速-功率数据再次进行纵向清洗.

2) 纵向清洗

横向清洗后的风速-功率数据序列总计 43 086 组，按照风电机组输出风速从小到大的顺序重新排序，排序后的数据组记为 M，然后以 0.5 m/s 的间隔对数据组 M 等距划分区间. 由于筛选后的风速最小值为 0.71 m/s，最大值为 14.1 m/s，故按照[0.5，14.5]的范围划分为 28 个子数据组，分别记为 M_i（$i=1,2,3,\cdots,n;n=28$）.

在每个子数据组 M_i 内按照功率从小到大的顺序再次重新排序. 根据式(1～3)分别计算每个新子数据组内的 Q_1 位置、Q_2 位置及 Q_3 位置，根据

式(4)和式(5)确定 R_1 位上限和 R_2 位下限位置,以此确定正常数据的范围,将范围外的数据视为异常值进行剔除. 统计得到异常数据共计 584 组,异常值剔除后的风速-功率散点图如图 4 所示.

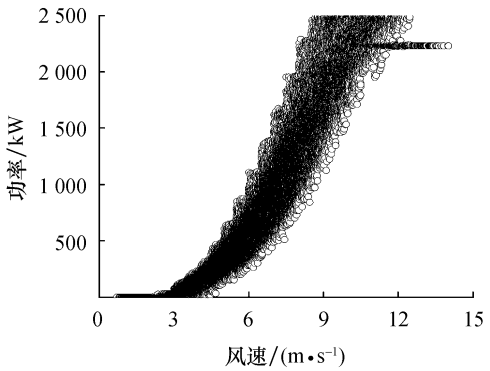

图 4　纵向清洗后风速-功率散点图

Fig. 4　The wind speed-power scatter after longitudinal cleaning by the quartile method

由图 4 可知,纵向清洗后仍存在 1 条中部积聚型异常数据,观察到该异常数据大致呈现为 1 条输出功率为 2 220 kW 左右的横向数据带,为弃风限电数据. 可见四分位法虽然可识别并剔除多数的 C 类数据和 D 类数据,但是部分区间内的 C 类数据和 D 类数据量过多会导致四分位法错误地把异常数据判定为正常值而继续保留,无法达到识别并剔除异常数据的目的.

3.2　DBSCAN 算法的应用

3.2.1　参数的设定

确定邻域半径 Eps 和邻域内最少包含点数 Minpts 是使用 DBSCAN 算法进行聚类的前提,Eps 和 Minpts 的取值将直接影响到聚类效果. 若 Eps 取值过大,则导致所有样本点都被划分到同一个簇;若 Eps 取值过小,则样本数据集合中可能没有核心点,且导致所有点可能都被标记为噪声. DBSCAN 算法提出者 Martin Ester 将 Minpts 的取值设置为 4,矩阵第 k 列称为 k 最近邻距离值,做出 k 距离图,通过观察法确定 k 距离图中曲线从平缓到陡峭的数据点,并以此作为 Eps 取值. 但是主观观察存在一定的误差,导致 Eps 取值不够准确. 因此用如下方法来确定 Eps 取值:

1) 由于 k 距离曲线的平缓性和陡峭性可以通过斜率值体现出来,所以求解 k 距离图中的每个点相对于下一点的斜率值,得到斜率数据组 P;

2) 然后计算斜率数据组 P 中所有非零斜率的平均值和标准差;

3) 最后找出第一个大于平均值和标准差之后的斜率值,此斜率值所对应的 k 距离值即为 Eps 取值[15].

3.2.2　使用 DBSCAN 算法

为达到良好的聚类效果,将风速-功率数据按照 0.5 m/s 的风速区间间隔进行划分,部分风速区间内的 Eps 和 Minpts 取值如表 1 所列.

表 1　QM 方法处理后部分风速区间内的 Eps 和 Minpts 取值

Tab. 1　The value of Eps and Minpts in part wind speed invertal after QM method processing

风速区间/(m·s^{-1})	Eps	Minpts
0～0.5	0.26	4
0.5～1	0.06	4
1～1.5	0.02	4
1.5～2	0.01	4
…	…	…

聚类完成后绘制所有风速区间内的风速-功率散点图,如图 5 所示.

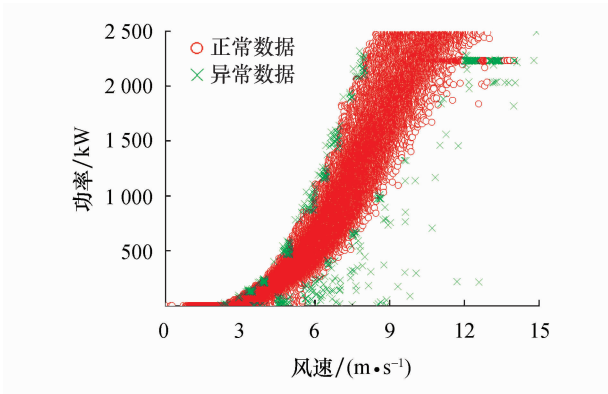

图 5　DBSCAN 聚类后风速-功率散点图

Fig. 5　The clustering of wind speed-power scatter by the DBSCAN

将图 5 中异常数据剔除后的风速-功率散点图如图 6 所示.

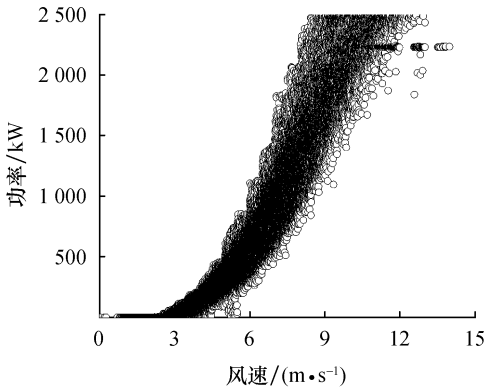

图 6　DBSCAN 处理后风速-功率散点图

Fig. 6　The cleaning effect of the wind speed-power scatter by DBSCAN

由图 6 可知，图中仍存在部分 C 类数据和 D 类数据未被剔除，剔除效果并不理想. 同时，DBSCAN 算法的参数 Eps 受到聚类区间内所有数据点的影响，风速区间内异常数据过多或数据分布范围分散都会影响 Eps 参数取值，导致误将异常数据当做正常数据保留，进而出现图 6 所示的情况.

3.3 QM-DBSCAN 算法的应用

从以上分析可知，四分位法和 DBSCAN 算法均可识别清洗异常数据，但实际应用中在异常数据过多或数据分布范围分散的情况下，单独使用四分位法或 DBSCAN 方法都无法达到理想的清洗效果. 由于两种方法各有优势，四分位法可识别并剔除多数分布范围分散的数据，而 DBSCAN 法在数据分布范围集中时具有优良的聚类效果，故可以考虑应用四分位法剔除分布范围分散的异常数据，然后再应用 DBSCAN 法进行聚类. 基于此，提出一种基于 QM-DBSCAN 算法的数据清洗方法，具体步骤如下：

1）对风速-功率数据应用四分位法进行横向清洗；

2）横向清洗后的风速最小值为 0.71 m/s，最大值为 14.1 m/s；然后将横向清洗后的风速-功率数据按照 0.5 m/s 的风速区间间隔进行划分；

3）由 3.2 节给出的参数确定方法确定每个风速区间内的 Minpts 和 Eps 取值，部分风速区间内的 Eps 和 Minpts 取值如表 2 所列.

表 2 QM-DBSCAN 处理后部分风速区间内的 Eps 和 Minpts 取值

Tab. 2 The value of Eps and Minpts in part wind speed invertal after QM-DBSCAN method processing

风速区间/(m·s^{-1})	Eps	Minpts
0～0.5	0.06	4
0.5～1	0.01	4
1～1.5	0.01	4
1.5～2	0.68	4
…	…	…

4）确定每个风速区间的 Eps 和 Minpts 取值后，对每个风速区间内进行聚类；然后对每个风速区间内的正常数据与异常数据进行标记并汇总，绘制风速-功率散点图，如图 7 所示. 剔除异常数据后得到风速-功率散点图，如图 8 所示.

由图 4、图 6 和图 8 可知，四分位法和 DBSCAN 法清洗结果都未处理掉的异常数据在图 8 中已剔除，因此提出的 QM-DBSCAN 方法效果显著. 为进一步从定量方面描述，选用 spearman 系数作为评价指标进行对比.

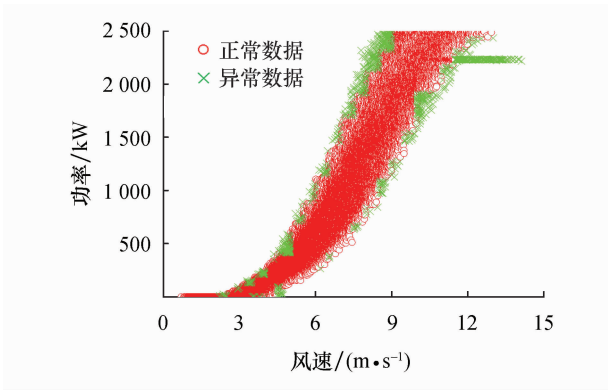

图 7 QM-DBSCAN 聚类后风速-功率散点图

Fig. 7 The clustering of wind speed-power scatter by QM-DBSCAN

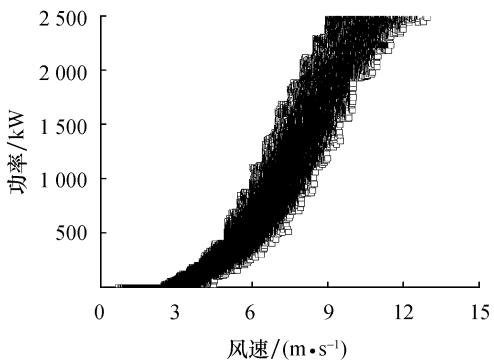

图 8 QM-DBSCAN 处理后风速-功率散点图

Fig. 8 The cleaning of wind speed-power scatter by QM-DBSCAN

3.4 评价指标

相关系数可以从定量角度描述两个变量间的相关程度. 由于风电机组的风速与功率有很强的相关性，故可用相关系数作为评价指标判定风速-功率数据清洗前后的相关性强弱，相关性越强代表清洗效果越好. 由于 spearman 相关系数具有很强的通用性，故选作评价指标对 3 种数据清洗方法（四分位法、DBSCAN 法、QM-DBSCAN 法）进行对比评价. 给定一组风速-功率数据 $S=\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，其中，$i=1,2,\cdots,n$. 则 spearman 相关系数求解公式为

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{6}$$

分别求解 3 种不同方法对风速-功率数据处理后的 spearman 系数，如表 3 所列.

由表 3 可知，QM-DBSCAN 方法清洗后的风速-功率数据的 spearman 系数较四分位法、DBSCAN 法分别提高了 0.003 5 和 0.004 7，进一步说

表 3　3 种方法的 spearman 系数

Tab. 3　The spearman correlation coefficient by three methods

清洗方法	Cov(X,Y)	Var(X)	Var(Y)	Spearman 系数
四分位法	71 328 622	254 692	22 767 886 889	0.936 6
DBSCAN	70 727 899	251 464	22 735 459 190	0.935 4
QM-DBSCAN	66 729 140	235 878	21 357 496 346	0.940 1

明所提出的 QM-DBSCAN 法优于其他 2 种方法.

4　结论

1) 选用风速-功率数据为清洗对象,根据异常数据的分布特征将其分为 4 类,分别表示为 A 类数据、B 类数据、C 类数据和 D 类数据.

2) 识别并剔除风速-功率异常数据时所应用的 QM-DBSCAN 法优于四分位法和 DBSCAN 法,spearman 系数较四分位法和 DBSCAN 法分别提高了 0.003 5 和 0.004 7,QM-DBSCAN 法的剔除效果最好.

3) 本文所提出的方法可适用于其他研究对象的异常数据清洗工作,具有一定的推广意义.

参考文献:

[1] 刘　波,贺志佳,金　昊. 风力发电现状与发展趋势 [J]. 东北电力大学学报,2016,2(36):8-13.

[2] Council GWE. Global wind statistics [EB/OL]. [2020-11-05]. https://gwec.net/gwec-wind-power-industry-to-install-71-3-gw-in-2020-showing-resilience-during-covid-19-crisis/.

[3] 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.

[4] 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(21):39-46.

[5] 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.

[6] 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.

[7] 张小奇,张振宇,孙骁强,等. 基于机群划分方法的风电场理论发电功率计算研究 [J]. 高电压技术,2019,45(1):284-292.

[8] 丁　明,张　超,王　勃,等. 基于功率波动过程的风电功率短期预测及误差修正 [J]. 电力系统自动化,2019,43(3):2-12.

[9] 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.

[10] 张东英,李伟花,刘燕华,等. 风电场有功功率异常运行数据重构方法 [J]. 电力系统自动化,2014,38(5):14-18,24.

[11] 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.

[12] LONG H,SANG L W,WU Z J,*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.

[13] YING Z,ZHENG Q,TIAN S S. Study on the abnormal data rejection and normal condition evaluation applied in wind turbine farm [C]//7th International Symposium on Precision Mechanical Measurements. Xiamen:SPIE,2016.

[14] 王　钦,蒋怀光,文福拴,等. 智能电网中大数据的概念、技术与挑战 [J]. 电力建设,2016,37(12):1-10.

[15] 宋董飞,徐　华. DBSCAN 算法研究及并行化实现 [J]. 计算机工程与应用,2018,54(24):52-56,122.