

基于 K-Medoids 聚类的分布式光伏台区线损异常感知算法

梁嘉文, 严贝峰, 景楷楠, 李婷婷, 屈志原, 王伟宁
(国网甘肃省电力公司甘南供电公司, 甘肃 甘南 747000)

摘要: 为保证分布式光伏台区稳定运行, 精准有效地划分台区线损数据, 提出基于 K-Medoids 聚类的分布式光伏线损异常感知算法, 精准判断分布式台区线损异常程度。采用局部异常因子(LOF)算法判断分布式光伏台区数据局部异常程度, 并筛选和去除受孤立点影响产生的异常线损数据。采取 K-Medoids 聚类方法聚类分析筛选后的分布式光伏台区数据, 将异常线损率区间结合异常线损数据的聚类中心和欧式距离, 完成台区线损异常感知。并创新性地引入粒度计算优化 K-Medoids 聚类算法聚类中心, 提升异常数据感知效果。试验结果表明, 所提算法可有效避免孤立点对异常感知效果的影响, 精准有效地感知分布式光伏台区线损异常, 并清晰划分台区线损数据类别。

关键词: K-Medoids 聚类; 局部异常因子; 粒度计算; 分布式光伏; 台区线损; 数据异常感知

中图分类号: TM731 文献标志码: A 文章编号: 1673-6540(2022)12-0047-06

doi: 10.12177/emca.2022.139

Distributed Photovoltaic Station Area Line Loss Anomaly Sensing Algorithm Based on K-Medoids Clustering

LIANG Jiawen, YAN Beifeng, JING Kainan, LI Tingting, QU Zhiyuan, WANG Weining
(Gannan Power Supply Company, State Grid Gansu Electric Power Company, Gannan 747000, China)

Abstract: In order to ensure the stable operation of the distributed photovoltaic station area and accurately and effectively divide the line loss data of the station area, a distributed photovoltaic line loss anomaly sensing algorithm based on K-Medoids clustering is proposed to accurately judge the degree of line loss anomaly in the distributed station area. The local anomaly factor (LOF) algorithm is used to judge the degree of local anomaly in the distributed photovoltaic station area data, and the anomalous line loss data generated by the influence of isolated points is filtered and removed. The distributed photovoltaic station area data after the filtering is clustered and analyzed by the K-Medoids clustering method. The anomalous line loss rate interval is combined with the clustering center and Euclidean distance of the anomalous line loss data, and the line loss anomaly sensing of the station area is completed, and the granularity calculation is innovatively introduced to optimize the K-Medoids clustering algorithm clustering center and to improve the sensing of anomalous data. The test results show that the proposed algorithm can effectively avoid the influence of isolated points on the anomalous sensing effect, accurately and effectively perceive the line loss anomalies in the distributed photovoltaic station area, and clearly divide the line loss data categories in the station area.

Key words: K-Medoids clustering; local anomaly factor (LOF); granularity calculation; distributed photovoltaic; line loss in station area; data anomaly sensing

收稿日期: 2022-07-15; 收到修改稿日期: 2022-10-21

作者简介: 梁嘉文(1979—), 男, 本科, 研究方向为信息通信和配网管理。

严贝峰(1990—), 男, 本科, 研究方向为配网运行、10 千伏分线线损。

景楷楠(1995—), 男, 本科, 研究方向为运营监测。

0 引言

分布式光伏为分散建设,可减少对公共电网的影响和危害^[1-4]。但随着分布式光伏数量的增加,大量的线损情况也同时出现,对分布式光伏的使用产生很大影响。若无法及时发现故障则会导致更多的问题,影响电网的稳定性、安全性。因此分布式台区线损异常程度感知对分布式光伏稳定运行有重要作用。

众多相关学者已对此展开大量研究,如陈洪涛等^[5]根据 K-Means 聚类算法判断线损异常,但该方法无法剔除影响线损异常的孤立点,且存在应用过程复杂的问题。张裕等^[6]结合多元线性回归和 K-Means 聚类算法进行了台区线损数据异常的检测,此方法运算简单,容易实现,可收缩性高,但运算开销比较大。

K-Medoids 聚类是 K-Means 的改进算法,可更精准地进行数据检测。因此本文提出基于 K-Medoids 聚类的分布式光伏台区线损异常感知算法,完成对分布式光伏线损异常数据的检测。

1 分布式光伏台区线损异常感知算法

1.1 线损数据聚类算法

采用 K-Means 聚类算法聚类分布式光伏台区线损数据时,会出现聚类线损结果差异小的情况,且对噪声和孤立点比较敏感^[7]。因此提出用中心点对象替换 K-Means 聚类算法中重心点的 K-Medoids 聚类算法。

K-Medoids 聚类算法的核心是已知分布式光伏台区线损数据集包括 n 个分布式光伏台区线损数据对象,将这些数据集分为 k 个聚类,将 O_i 作为原始簇中心,随机获取差异数据对象时,应在 k 聚类内完成。为了给与各数据对象距离最近的簇中分配合理的对象^[8],需要基于这些数据对象到 O_i 的距离进行分配。顺序采集非簇中心点的数据对象 O_{tmp} 的工作,需要在各簇中完成。 E_{tmp} 为消耗代价,用于描述 O_i 被 O_{tmp} 代替过程所产生的分布式光伏台区线损数据聚类成本。 E_i 是线损数据聚类采用 O_i 过程中形成的聚类成本。对比 E_{tmp} 和 E_i 两种聚类成本,获取聚类准则函数为若 $E_{\text{tmp}} < E_i$,则说明聚类还在收敛,用 O_{tmp} 替换 O_i ,完成簇中心的优化,不断迭代聚类,直至存在 $E_{\text{tmp}} \geq$

E_i ,保持簇中心稳定,此时簇结构也保持稳定,该种情况下,聚类准则函数收敛,完成线损数据聚类。

为了使簇内以及簇间的数据对象尽可能地相似以及不相似,需要对聚类准则函数可否收敛进行判断。

K-Medoids 聚类算法利用欧氏距离权衡类内相似性,用下式描述该距离 $d_{(x,y)}$:

$$d_{(x,y)} = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + \cdots + |x_p - y_p|^2} \quad (1)$$

式中: $x = (x_1, x_2, \cdots, x_p)$ 与 $y = (y_1, y_2, \cdots, y_p)$ 分别为数据集中两个 p 维的数据对象。

通过目标函数(误差平方和准则函数) E 评价 K-Medoids 的划分聚类效果^[9],则有:

$$E = \sum_{i=1}^k \sum_{x_j \in C_i} |x_j - o_i|^2 \quad (2)$$

式中: k 为聚类数量; $|x_j - o_i|^2$ 与 o_i 分别为欧氏距离与簇类 C_i 的中心点; E 为数据对象与其所在簇类的中心点误差平方和。

假定待聚类数据集是 n 个台区线损数据对象构建的数据集合 $X = (x_1, x_2, \cdots, x_n)$, p 维数据对象用 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 描述。为了得到数值最小化的目标函数 E ,需要得到 X 的一个划分数据对象 $P_k = \{C_1, C_2, \cdots, C_k\}$, 确保式(2)值最小化。

K-Medoids 聚类算法模型为,输入:分布式光伏台区线损数据集 X ,其由 n 个线损数据对象组成,待划分的聚类数为 k ;输出: k 个聚类簇 C_i ,确保聚类准则函数 E 值最小化。具体步骤如下:

(1) 设置初始簇中心 O_i 时,需要有 n 个数据对象组成数据集,将这些数据集分为 k 个聚类,在 k 聚类中随意选择不同的数据对象;

(2) 重复迭代;

(3) 为了向各数据对象距离最近的簇中分配合理的对象,需要根据式(1)计算获取的数据对象到 O_i 的距离进行分配;

(4) 根据式(2)计算得出聚类成本 E_i ;

(5) 为了计算数据对象 O_{tmp} 的消耗代价 E_{tmp} ,需要顺序采集非簇中心点的数据对象 O_{tmp} ,并在各簇中完成;

(6) 若 $E_{\text{tmp}} < E_i$,则说明聚类还在收敛,用

O_{tmp} 替换 O_i , 不断迭代聚类, 产生 k 个新簇;

(7) 直到 E 值保持稳定。

本文采用 K-Medoids 聚类算法逐个聚类光伏台区线损数据, 过程中设置聚类簇数为 1 个。

1.2 异常数据挑选和去除

聚类中心点和欧氏距离对分布式光伏台区线损异常感知起着重要的作用。首先孤立点对分布式光伏台区线损聚类中心点的干扰, 可以通过 K-Medoids 聚类算法解决, 但孤立点对欧式距离的干扰不能用 K-Medoids 描述。因此要在挑选和去除分布式光伏台区线损异常数据之后, 再进行分布式光伏台区数据聚类, 异常数据即为上文的孤立点。为了正确筛选出分布式光伏台区异常数据^[10], 本文采用了局部异常因子(LOF)算法。

1.2.1 定义局部光伏台区异常

取分布式光伏台区线损数据随意自然数 K , 在数据对象 x_a 具备以下条件时, 可以用数据对象 x_b 与数据对象 x_a 之间的距离描述 x_b 的 K -距离, 即 $K\text{-distance}(x_b)$ 。

(1) 最少有 K 个对象 $x'_a \in D \setminus \{x_b\}$, 满足 $d(x_b, x'_a) \leq d(x_b, x_a)$, d 代表距离;

(2) 最多有 $K-1$ 个对象 $x'_a \in D \setminus \{x_b\}$, 满足 $d(x_b, x'_a) < d(x_b, x_a)$ 。

数据对象 x_b 的 K -距离邻域可描述为

$$N_{k\text{-distance}}(x_b) = \{q \mid d(x_b, q) \leq K\text{-distance}(x_b)\} \quad (3)$$

即所有数据对象与 x_b 的距离对比, 为小于等于 $K\text{-distance}(x_b)$ 。

x_b 对于对象 x_a 的可达距离表达为

$$\text{reach-dist}_k(x_b, x_a) = \max\{K\text{-distance}(x_b), d(x_b, x_a)\} \quad (4)$$

对象 x_b 与其 MinPts -领域的平均可达距离的倒数为对象 x_b 的局部可达密度, 即:

$$\text{lrd}_{\text{MinPts}}(x_b) = \frac{1}{\left[\frac{\sum_{o \in N_{\text{MinPts}}(x_b)} \text{reach-dist}_{\text{MinPts}}(x_b, x_o)}{|N_{\text{MinPts}}(x_b)|} \right]} \quad (5)$$

根据式(3)描述出分布式光伏线损 LOF 为

$$\text{LOF}(x_b) = \sum_{o \in N_{\text{MinPts}}(x_b)} \frac{\text{lrd}_{\text{MinPts}}(x_b, x_o)}{\text{lrd}_{\text{MinPts}}(x_b)} \cdot \frac{1}{|N_{\text{MinPts}}(x_b)|} \quad (6)$$

以上算法说明分布式光伏台区线损 LOF 与数据异常的可能性成正比, 分布式光伏台区线损数据对象 x_b 异常的程度可以用数据对象 x_b 的异常因子描述。

1.2.2 LOF 计算

n 个分布式光伏台区线损数据对象组成分布式光伏台区线损数据库 X , 用邻域中数据对象的最小个数 MinPts 来计算所有分布式光伏台区线损数据对象的 LOF 值^[11-12], 此计算有两步。

(1) 为了得到各个分布式光伏台区线损数据对象的 MinPts -邻域到各个分布式光伏台区线损数据对象的距离, 需要使分布式光伏台区线损数据对象的 MinPts -邻域在分布式光伏台区线损数据库中构成, 在物化数据库导入计算结果, 且使用的方法随差异维数而变化。

(2) 对不同数据对象的局部可达密度以及 LOF 进行运算, 将各个分布式光伏台区线损数据对象的 LOF 值根据物化数据库 M 来描述。

用下式描述对象 x_b 的局部可达密度:

$$\text{LD}(x_b) = \frac{|N(x_b)|}{\sum_{o \in N(x_b)} \text{dist}(x_b, x_o)} \quad (7)$$

式中: $\text{dist}(x_b, x_a)$ 与 $|N(x_b)|$ 分别为 x_b 和 x_a 的 n 维欧式距离与 x_b 的邻域中邻居的数目。

为了得到 x_b 的局部偏离指数, 需要满足 $|N(x_b)| < \text{MinPts}$, 用下式来描述:

$$\text{LDI}(x_b) = \sum_{o \in N(x_b)} [\text{LD}(x_o) / \text{LD}(x_b)] \cdot \frac{1}{|N(x_b)|} \quad (8)$$

LOF 算法分析数据异常的标准是数据对象全部邻居的局部可达密度均值。

1.3 分布式光伏台区线损异常感知流程

本文所提方法先由 LOF 算法对分布式光伏每日线损率的数据进行选择, 再由 K-Medoids 聚类算法聚类处理后的线损数据, 完成分布式光伏台区线损异常感知。分布式光伏台区线损率分为正常、高损和负损。正常的线损率值在 0%~9% 之间, 高损的线损率值大于 9%, 负损的线损率值小于 0%。分布式光伏线损的电量和数据的采集必须在智能表 100% 覆盖的台区进行, 才能精准判断日线损率, 在非智能表下的光伏台区采集数据是无效的。根据本文所提方法的思想得到了如图 1 所示的分布式

光伏台区线损异常感知的流程图。

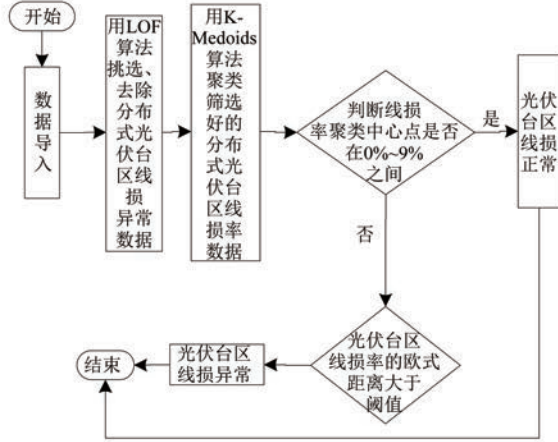


图1 K-Medoids 算法分布式光伏台区线损异常感知流程图

导入分布式光伏台区线损数据,用 LOF 算法挑选和去除分布式光伏线损异常数据后,利用 K-Medoids 算法对 LOF 算法处理后的分布式光伏台区线损率数据进行聚类。在聚类时,影响分布式光伏台区线损异常的因素有两个^[13],分别是聚类中心点和欧氏距离,其分别代表分布式台区线损异常的程度以及每日线损率的上下幅度。为了准确判定分布式光伏台区线损,采用光伏台区日线损率总个数与欧式距离结果的倒数描述测验中的欧氏距离数据差异。如果分布式光伏台区线损出现异常^[14-15],说明日线损率上下幅度大,且欧氏距离比阈值大;反之分布式光伏台区线损正常,日线损率上下幅度较平稳,且数值在聚类中心点区间内。

1.4 引入粒度计算的 K-Medoids 算法优化

利用 K-Medoids 算法进行光伏台区线损数据聚类分析,某种程度上可以理解为在全部可能的粒度集内得出最佳粒度值。因此为了达到 K-Medoids 算法的最佳数据聚类效果,引入粒度计算。

论域 U 以及属性集合 A 构成聚类空间,聚类空间 $T=(U,A)$ 的相似度 $F(x_i, x_j)$ 用下式描述:

$$F(x_i, x_j) = \frac{1}{1 + \sum_{i=1}^{|A|} w_i |x_{il} - x_{jl}|} \quad (9)$$

式中: $|A|$ 以及 x_{il}, x_{jl} 分别为聚类属性集合个数以及 x_i, x_j 的第 l 属性值; w_i 为聚类属性分辨能力,如式(10)所示:

$$w_i = \frac{\sum_{i=1}^e |x_i|^2}{|U|^2} \quad (10)$$

式中: $|U|$ 与 e 分别为数据对象的总数与划分块数。

区间块 $\{x_1, x_2, \dots, x_e\}$ 由属性 l 值分割获取,利用 x_i 进行区间分割获取的数据对象个数为 $|x_i|$, h 为数据对象数,对象平均相似度如下:

$$\bar{F} = \frac{\sum_{i,j=1}^h F(x_i, x_j)}{\sum_{i,j=1}^h h^2} \quad (11)$$

在论域里分割 h 个粒子得到 $\{X_1, X_2, \dots, X_h\}$, 第 i 个区间中的数据对象数目为 $|X_i|$, X_i 粒子的密度如下:

$$g(X_i) = \frac{|X_i|}{|U|} \quad (12)$$

该粒子 h 的平均密度为

$$\bar{G} = \frac{\sum_{i=1}^n g(X_i)}{h} \quad (13)$$

设置有 H 个数据对象在第 i 个粒子中,用 $x_{i1}, x_{i2}, \dots, x_{in}$ 表示,粒子中心点 D_i 用下式描述:

$$D_i = \left\{ x_{ij} \mid \min_{j=1}^H \left| x_{ij} - \frac{1}{H} \sum_{k=1}^H x_{ik} \right| \right\} \quad (14)$$

粒计算的运算过程如下。

(1) 线损数据对象间相似度及平均相似度,由式(9)和式(11)运算。

(2) 设置阈值 d 的取值,且其上限 $\bar{d}$ 值变化幅度为 $\pm 10\%$, 数据对象间的模糊相似矩阵是 $M(i, j)$, 在 $M(i, j) = 1$ 时, 则 $F(x_i, x_j) \geq d$, 反之 $M(i, j) = 0$, 则 $F(x_i, x_j) < d$ 。

(3) 获取各对象类似对象序号的过程,即为依据 $M(i, j)$ 处理各对象获取其粗粒集 G 的过程,粒子即为对象及其所有类似对象构成的集合。

(4) 各个粒子的密度以及 G 的平均密度由式(12)和式(13)描述,有效粒子集 I 包含 $gd(x_i) \geq \bar{G}$ 的有效粒子。

(5) I 内各个粒子的中心根据式(14)运算,在 X 中录入两个随意粒子间的欧氏距离。

(6) 首个聚类中心点 o_1 是 I 内最大密度粒子 Q_1 的中心点,第二个聚类中心点 o_2 是与 Q_1 最远

的密度最大粒子 Q_2 的中心点,由 X 对其余粒子的中心点到 $o_1, o_2, \dots, o_n$ 的距离进行运算,得 $d_{i1}, d_{i2}, \dots, d_{in}$, 取其 d_i 最小值,求 d_i 最大值,粒子 Q_i 的中心是 o_i ,同理粒子 Q_k 的中心是 o_k 。

以簇的数量 K 以及 n 个对象的数据集为输入,得到 K 个簇类中心和聚类准确度。粒计算算法模型如下。

(1) 记录中心点对应的粒子是 $Q_1, Q_2, \dots, Q_k$, 且 K 个簇类中心 $o_1, o_2, \dots, o_k$ 通过粒计算改进的 K-Medoids 聚类算法被初始化。

(2) 为距离数据对象最近的中心点的簇分配其余数据对象,并将该簇记作 w , 对 $S(w)$ 进行运算。

(3) 分别选取在 $Q_1, Q_2, \dots, Q_k$ 里的非代表对象 o_{random} , 将 o_i 用 o_{random} 替换,再更新各个簇内中心点并重新聚类,将新的簇记作 w' , 运算 $S(w')$ 。

(4) 运算 $\Delta S = S(w') - S(w)$, 如果 $\Delta S > 0$, 则 o_i, w 被 o_{random}, w' 替换, 反之保持不变。

(5) 反复运算步骤(3)和步骤(4), 当簇集合中对象保持稳定时, 运算结束。

2 试验分析

以某地区 689 个分布式光伏台区的线损数据为试验对象,验证本文所提算法的有效性。本文采用 LOF 算法检测和去除分布式线损异常数据,统计 LOF 算法不同取值下的分布式台区线损数据分布图,结果如图 2 和图 3 所示。

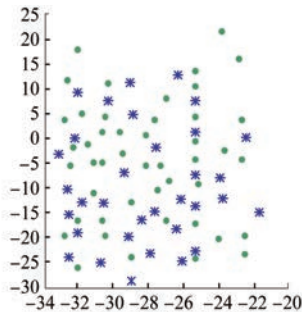


图 2 LOF=0.4 时分布式台区线损数据分布图

图 2 和图 3 中,圆圈代表分布式光伏台区线损正常数据,星号代表分布式光伏异常数据。当 LOF 值为 0.4 时,分布式光伏线损的异常数据数量很多,当 LOF 值为 1 时,分布式光伏线损的

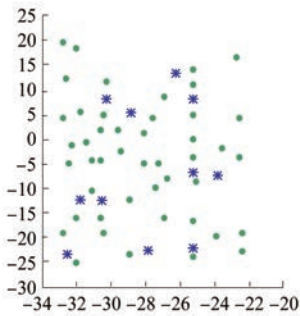


图 3 LOF=1 时分布式台区线损数据分布图

异常数据大幅减少。由此说明, LOF 值越大对由孤立点导致的分布式光伏台区暂态异常数据的挑选和去除效果越好,能够更精准地衡量数据的异常程度,为应用 K-Medoids 算法进行分布式光伏台区线损异常数据感知提供更好数据基础。

利用 K-Medoids 算法对分布式光伏台区线损率进行聚类,设置欧氏距离的阈值为 9。本文从 689 个分布式光伏台区中,选取 4 个光伏台区 15 天内的每日线损率,如表 1 所示。

表 1 4 个分布式光伏台区每日线损率 %

日期	B01	B02	B03	B04
9月1日	-56.96	19.09	-8.10	21.80
9月2日	66.02	16.48	-10.09	19.70
9月3日	-10.50	23.82	-11.68	19.32
9月4日	-60.06	26.30	-13.00	23.89
9月5日	0.32	18.80	-3.96	17.10
9月6日	1.00	8.49	-21.56	9.25
9月7日	0.31	9.89	-19.97	8.93
9月8日	0.81	19.49	-17.84	17.46
9月9日	0.89	-4.47	15.82	16.14
9月10日	0.34	3.69	-0.59	21.87
9月11日	0.77	1.89	-19.82	29.99
9月12日	0.78	-3.08	-8.62	19.66
9月13日	0.56	-5.56	-14.99	15.76
9月14日	0.49	-9.75	-14.84	14.20
9月15日	0.77	-5.63	-16.42	20.88

将以上分布式台区线损率,分别通过 K-Medoids 聚类算法、本文算法进行线损异常感知,得到如表 2 所示的统计结果。

表 2 线损数据感知结果对比

台区名称		B01	B02	B03	B04
K-Medoids 聚类算法	聚类中心点	0.29	3.74	-11.73	17.05
	欧氏距离	371.61	87.23	65.72	26.77
	线损状态	线损异常	线损异常	负损	高损
本文算法	聚类中心点	0.29	2.46	-11.32	16.14
	欧氏距离	3.79	51.02	49	13.97
	线损状态	正常	线损异常	负损	高损

由表 2 可知,K-Medoids 算法将正常光伏台区数据判定为异常光伏台区,存在判定错误的缺点。本文所提算法有效结合了 LOF 算法和 K-Medoids 聚类算法,合理地避免了由孤立点导致的分布式光伏台区出现暂态异常数据对异常感知结果的影响,可以准确感知分布式光伏台区线损异常。

通过本文所提算法进行分布式光伏台区的异常线损感知。先由 LOF 算法对分布式光伏每日线损率的数据进行选择,如图 4(a)所示;再由 K-Medoids 聚类算法聚类处理后的线损数据,如图 4(b)所示,最终完成分布式光伏台区线损异常感知。

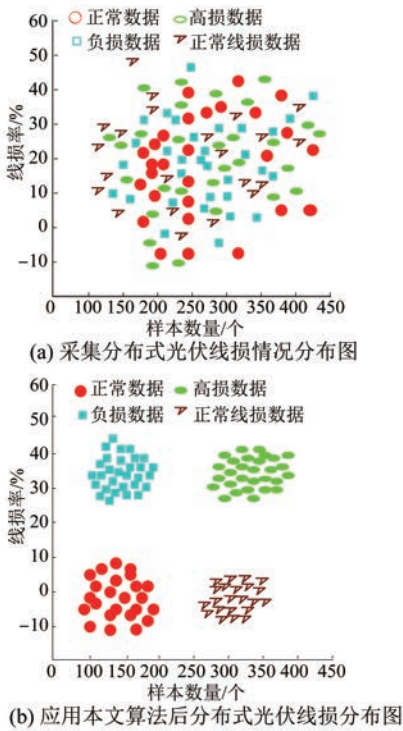


图 4 异常线损感知效果图

图 4(a)中分布式光伏台区线损的正常数据、

高损数据、负损数据、正常线损数据每个类别比较分散,无法准确对数据进行感知,而在图 4(b)中能清楚看到,每个数据都被聚类在各个数据的类别里,说明本文所提算法可有效、准确地解决分布式光伏台区线损异常感知的问题。

3 结 语

本文所提算法将 LOF 与 K-Medoids 算法结合,根据分布式光伏台区线损异常的情况,对光伏台区数据进行采样、筛选、识别,将异常数据进行聚类,保证算法能够准确地感知分布式光伏台区异常数据,并将分布式光伏的异常情况进行聚类。创新性地引入粒度计算优化 K-Medoids 算法,在全部可能粒度集内得出最佳粒度值,以此达到 K-Medoids 算法的最佳数据聚类效果,确保分布式光伏台区异常数据感知的精准性、可靠性。

【参考文献】

[1] 李东东,郭天洋,刘庆飞,等. 计及光伏发电的新能源电力系统惯量评估[J]. 太阳能学报, 2021, 42(5): 174.

[2] 吴虹. 光伏台区同期线损异常原因分析及治理思路[J]. 江苏科技信息, 2019, 36(29): 44.

[3] 林潮雄. 分布式光伏发电影响台区线损处理[J]. 农村电工, 2020, 28(4): 36.

[4] 刘金亮,宋文乐,黄庆,等. 电桩谐波的分布式光伏发电并网台区线损精确计算方法[J]. 电气自动化, 2021, 43(4): 14.

[5] 陈洪涛,蔡慧,李熊,等. 基于 K-Means 聚类算法的低压台区线损异常辨别方法[J]. 南方电网技术, 2019, 13(2): 2.

[6] 张裕,徐依明,张彦,等. 一种新 K-Means 聚类算法的多元线性回归台区线损率预测模型[J]. 电力科学与技术学报, 2021, 36(5): 179.

[7] 伍栋文,于艾清,俞林刚,等. 基于 ICS-K-Means 聚类算法和 WNN 的有源低压台区线损估算方法[J]. 智慧电力, 2022, 50(4): 8.

[8] 盛静文,于艳丽,江开忠. 基于 K-Medoids 聚类的贝叶斯集成算法[J]. 智能计算机与应用, 2021, 11(2): 84.

[9] 王华勇,韩松,肖孝天,等. 改进的电力负荷曲线集成 K-Medoids 聚类算法[J]. 电力科学与工程, 2019, 35(2): 38.

(下转第 80 页)