

考虑信息动态表达的异常用电模式识别云边协同方法

刘慧自,汪颖,胡文曦,肖先勇

(四川大学 电气工程学院,四川 成都 610065)

摘要:异常用电识别是用电稽查、计量装置运行状态辨识的重要内容,对维护电网的安全运行和保障正常用户权益有重要的意义。已有方法为了识别用户的多元用电模式,在保证识别准确性的基础上容易造成计算过于复杂的问题,而考虑效率的简单计算方法又难以准确度量不同用电模式的相似性,因此难以兼顾计算效率与准确性;此外,将用电数据上传至云端集中计算会占用大量的网络带宽和计算资源,进一步限制了异常辨识的应用。为此,提出了一种考虑信息动态表达的异常用电模式识别云边协同方法。根据边缘端和云端的计算资源合理分配协作任务,实现了异常用电的云边协同识别。针对边缘服务器算力有限的问题,对用电数据进行动态压缩重表达,在缩减数据量的同时保证数据信息的准确性。云端在收到压缩数据后以分段加权动态时间规整距离作为压缩数据相似性度量的依据,基于自适应参数选择的密度聚类算法识别异常用电。基于实际数据集验证了所提方法的有效性。

关键词:异常用电识别;信息动态表达;云边协同;数据压缩;相似性度量

中图分类号:TM 715

文献标志码:A

DOI:10.16081/j.epae.202204051

0 引言

异常用电识别不仅可以规范用户的用电行为,还能减少电力企业的经济损失以及维护电网的安全运行^[1-2]。但现有的人工排查方法难以应对不同异常用电原因所呈现的多样性、随机性等特点,导致识别结果的准确率不高,且效率较低^[3]。因此,根据电力计量数据实现异常用电的自动识别具有重要的意义。

随着人工智能算法的发展,基于回归模型、分类模型、聚类模型等的识别方法被广泛应用^[4]。基于回归模型的异常检测器通过回归方法预测未来负荷,将与预测结果偏差较大的实际值识别为异常。例如:文献[5]利用用电量指标构建异常用电评判体系,提出了一种基于逻辑回归模型的二次筛选异常用电识别方法,提高了识别准确率;文献[6]针对传统的基于回归模型的异常检测器预测和检测性能不理想的问题,提出了新的异常检测架构,较大地提高了预测和检测精度。然而,基于回归模型的方法需要为每个消费者建立回归模型,这需要大量的样本数据,并消耗大量的计算资源。分类模型则是根据对象的标签和特征训练其与类别的一种映射关系,并基于此进行异常识别。例如:文献[7]结合历史用

电数据和支持向量机的自动特征提取方法进行异常用电识别,并针对训练集进行了专门的参数优化;在此基础上,文献[8]利用轮廓图识别数据集分布,进而检测异常窃电损耗。虽然基于分类模型的方法具有较高的异常识别精度,但要求有大量带标签的数据集进行模型训练,这在实际中难以获取。聚类模型也被广泛地应用于异常识别,聚类可以根据数据间的相似度自动地识别异常,且无需带标签的数据。例如:文献[9]结合K-means、模糊C均值、自组织映射等聚类算法,构建了半自动特征提取方法,并通过优化多个参数实现了基于特征的异常用电识别;文献[10]基于密度聚类算法给出了异常用电波动区间的离群度,并综合分析当前的用电量以获取异常用电得分。相比于分类模型,聚类模型无需带标签的异常数据集,且基于密度的聚类算法能根据数据间的相似度找到高密度区间,在异常识别领域有较大的优势,但其缺点是计算非常费时,且结果的准确性依赖于有效的相似性度量手段。已有研究中的相似性度量方法普遍采用欧氏距离进行度量,但其只考虑了负荷曲线对应时间点的距离特性,难以反映曲线的整体动态特征。动态时间规整DTW(Dynamic Time Warping)距离能更充分地反映时间序列的相似度特性^[11-12],但在效率上存在一定的劣势,采用其处理高维曲线时难以满足在线、实时的应用要求。

此外,将高维用电数据上传至云端集中计算会占用大量的网络带宽和计算资源,而边缘计算直接在用户侧的网络边缘就地对数据进行处理,既提高了系统的响应速度,也降低了对网络传输的需求,因此边缘计算在电力行业中有着广泛的应用^[13-14]。然而,边缘服务器的算力通常较低,无法胜任计算量大

收稿日期:2021-10-17;修回日期:2022-01-28

在线出版日期:2022-04-17

基金项目:国家自然科学基金资助项目(52077145);中央高校基本科研业务费专项资金资助项目

Project supported by the National Natural Science Foundation of China(52077145) and the Fundamental Research Funds for the Central Universities

的数据处理任务,因此边缘端适合与云端协同完成任务。在本文的架构中,为了减轻云边数据传输的压力以及分担云端数据处理的任务,有必要将数据压缩这个前期步骤部署在边缘端。目前常见的压缩方法有分段聚合近似PAA(Piecewise Aggregation Approximation)、符号聚合近似、主成分分析等^[15-16]。上述方法大多使用固定的压缩率对原始数据进行重表达,但是不同的用户甚至是同一用户在不同的时间内用电曲线的整体特征和变化趋势具有多样性,使用固定压缩率容易使原始数据丢失部分细节信息,进而影响后续异常模式识别的准确性。

因此,本文从兼顾准确性与计算效率的角度出发,提出了一种考虑信息动态表达的异常用电模式识别云边协同方法。在边缘端通过信息动态表达的PAA对用户的用电数据进行信息重表达,在缩减数据量的基础上保留特征信息,以减少云边数据传输的带宽占用。云端在接收到压缩数据后以一种新型的分段加权DTW距离作为相似度计算的依据,有效地减轻了云端的计算压力。最后通过自适应参数选择的密度聚类算法进行异常用电识别。算例结果表明,本文所提方法在显著缩减数据量、提高运算效率的同时,保证了异常用电识别的准确性,有较好的实际应用前景。

1 基于云边协同架构的异常用电识别

明确不同的异常用电原因及其呈现的数据特征,是准确识别异常用电的前提,因此本节先对电网中异常用电的主要原因及其数据特点进行分析。

1.1 异常用电原因

设备自身的健康状况、工作环境和运行状态以及人为因素都是造成异常用电的原因,具体可概括如下。

1)窃电。窃电一般通过绕接电表、篡改电能表的接线方式甚至破坏电能计量装置实现,虽然手段错综复杂,但都是从电能计量装置的原理出发。电能计量装置的计量值与计量电压、计量电流、计量时间及电压和电流的相角差有关,改变其中的一个参数就能使装置出现计量错误,从而达到窃电的目的^[17]。

2)装置异常。电能计量装置虽然在出厂前进行了一系列的测试,但是经过长时间的运行后,难免会出现一些故障,从而引起计量数据异常^[18]。电能计量装置的工作环境也会对其计量的准确性产生影响,例如电网中的谐波会加速电能计量装置的老化,从而使计量出现偏差。

3)负荷异常。电网中的冲击负荷或者某些突发事件,如受极端天气的影响所导致的用电负荷明显增大或减小,也容易造成计量结果出现短时间内的

畸变,进而被识别为异常。发生这种异常的根本原因在于用户的用电习惯在短时间内出现了较大的改变,所以正确识别负荷异常有助于电力企业建立正确的用户用电模式,估计用户负荷需求,优化电力资源调度。

针对以上3种原因导致的异常用电,异常用电识别能快速准确地筛选出异常用电数据,辅助运营管理人员对异常用电进行稽查,减少电力企业的损失,维护电网的安全运行,辅助电力企业的运营决策,具有重大的意义和价值。

1.2 异常用电数据特点

异常用电数据的产生具有随机性,且隐藏在正常用电数据中,常见异常用电数据有以下2种类型。

1)点异常。当一个数据点与其周围数据差别较大时,可以认为其是点异常,如附录A图A1所示。点异常是最简单、最普遍的异常,其表示某个瞬间用电负荷的突然增大或减小,在用电曲线上表现为向上或向下的尖峰。装置异常和负荷异常都有可能导致点异常用电数据的出现。

2)连续异常。连续异常是指一段数据出现明显的波动或者毛刺,或某段数据显著增大或减小,如附录A图A2所示。连续异常表示一段时间内用电负荷的显著增大或减小,在用电曲线上表现为突然出现的波峰或波谷。连续异常出现的原因包括窃电、装置异常、负荷异常等。

异常用电数据具有随机性和多样性,且所占比例小,使得异常用电样本难以获取,因此无需标签的无监督方法更适用于识别异常用电。

1.3 异常用电识别云边协同架构

有效的相似度衡量是基于无监督学习进行异常用电模式识别的核心。正常用电模式通常是相似度较高的聚集族群,而异常用电模式是偏离族群的离群点,因此基于相似度衡量的异常点辨识对异常用电的识别准确性具有重要的作用。然而,鉴于目前用电数据的采样率普遍较高,采用集中式的异常识别方法势必会给数据传输带宽和服务器算力造成巨大的压力,为此本文提出了一种云边协同的分布式异常识别方法。

云计算具有不需要维护本地服务器、数据存储和软件的优点,但由于电网结构复杂、节点规模较大,云端和终端往往相距很远,集中式的方法并不适用,其响应通常较慢,且短时间内终端数据的涌入会占用大量的传输带宽,给云端服务器带来巨大的压力。而边缘服务器因为接近终端设备,处理数据的网络延迟可忽略不计,因此采用云边协同的方式可极大地提升系统的性能。但边缘端的算力较弱,因此有必要将数据压缩模块部署在边缘端。综上所述,数据压缩是实现基于云边协同的用户异常用电

模式识别的基础,若数据压缩造成大量的信息缺失,则会直接影响云端异常识别的准确性。此外,边缘端将压缩后的低维数据上传至云端,还能降低对传输带宽的要求。云端在接收到经压缩的数据后,基于本文所提方法进行异常用电识别,相较于直接处理原始数据的方法,在本文的架构下云端的压力和整个协同架构的延迟都明显小得多。云边协同计算的构架如图1所示。

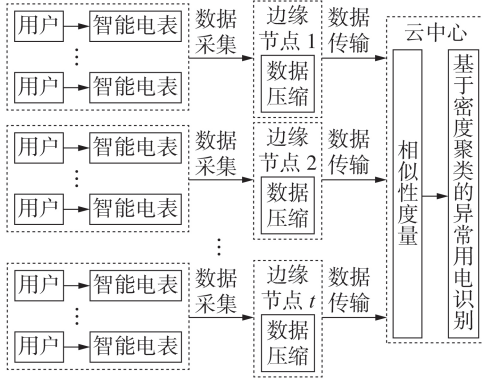


图1 云边协同计算构架

Fig.1 Computing architecture of cloud-edge collaboration

2 信息动态表达的PAA方法

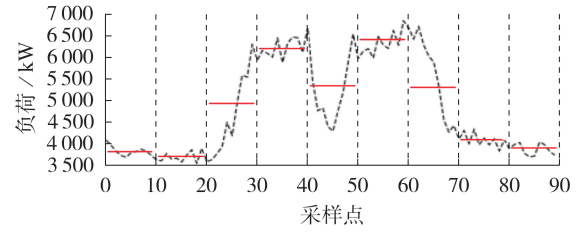
电网中各种仪表的采样率不断提高,使电力企业能掌握到的用户用电数据更为精细,但是这些高精度的数据对算力提出了更高的要求。为了节约计算资源,可以先对数据进行压缩,然而数据压缩往往会带来数据信息的损失,为此本文提出了一种信息动态表达的PAA方法来对用电数据进行压缩处理,在减少数据量的同时保证信息的完备性。

PAA的基本思想为:用等宽度窗口分割时间序列,并用窗口平均值表示每个窗口内的时间序列,从而达到数据压缩处理的目的。对于1条长度为 n 的负荷数据 $X=[x_1, x_2, \dots, x_j, \dots, x_n]$,若 m 能被 n 整除,则将其压缩为1条长度为 m 的数据 $Y=[y_1, y_2, \dots, y_i, \dots, y_m]$, Y 中元素 y_i 的计算式为:

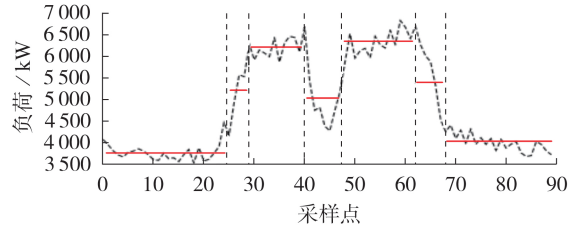
$$y_i = \frac{m}{n} \sum_{j=n(i-1)/m+1}^{ni/m} x_j \quad i=1, 2, \dots, m \quad (1)$$

考虑到不同的用户甚至是同一用户在不同时段内用电曲线的整体特征和变化趋势有所不同,且由异常用电波形特征可知,用电曲线中的大部分波形较为平稳,反而是少部分波动较大的信号段包含了较多的信息,对于其中波动剧烈的数据段而言,若直接利用PAA方法进行压缩处理,则会丢失许多重要的特征信息,无法实现用电信息的动态表达。利用经典PAA方法的压缩效果如图2(a)所示,可见不论用电曲线如何变化,PAA始终以固定的窗口进行数

据压缩,对于一些波动较大的分段而言,均值已明显不能精确表示该段的信息。在此基础上,本文对经典PAA方法的分段策略进行改进,根据数据自身的统计特征选取分段窗口的长度,动态调整压缩率,使用电数据的信息得到动态表达。信息动态表达的PAA方法的压缩效果如图2(b)所示。由图可见:用电曲线平稳处的分段不再受到固定压缩率的限制,而是根据曲线自身的变化进行动态调整,以平衡数据量;对于波动段而言,其压缩窗口明显变窄,使得到的均值具有代表性,从而保留更多的数据信息。



(a) 经典PAA方法



(b) 信息动态表达的PAA方法

--- 原始负荷, — 均值

图2 不同PAA方法的压缩效果

Fig.2 Compression effects of different PAA methods

本文所提信息动态表达的PAA方法的具体步骤如下。

1) 对负荷数据 $X=[x_1, x_2, \dots, x_n]$ 进行差分运算并取绝对值得到 $X'=[x'_1, x'_2, \dots, x'_{n-1}]$,其中 $x'_g = |x_{g+1} - x_g| (g=1, 2, \dots, n-1)$ 。

2) 对于负荷数据 X 中的每个值 $x_j (j=1, 2, \dots, n)$,取其前 k 个和后 k 个值作为一个阈值计算窗口,阈值的计算方法为:

$$\delta_g = \frac{\mu}{\sigma} C_{\text{cnt}} \quad (2)$$

式中: δ_g 为 x_{g+1} 相对 x_g 的最大允许变化量; μ 为计算窗口中数据的平均值; σ 为标准差;由于 μ/σ 没有量纲,因此引入校正参数 C_{cnt} ,用以校正 δ_g 的数量级,使之与 x'_g 相匹配,同时也可以通过控制该参数的大小,得到最匹配边缘服务器算力的数据量。

由式(2)可知,阈值与计算窗口内的用电水平正相关,当出现用电波动大的情况时,阈值会降低以使分段更精细。此外,校正参数的引入使得阈值可根据实际情况下压缩率和压缩精度的要求进行动态调

整,以应对不同情况下对数据量和精度的要求。

3)若 $x'_g > \delta_g$, 则 x'_g 对应的元素 x_g 会作为分段点。找到所有的分段点后,取每个分段数据的平均值可得到改进后的考虑信息动态表达的压缩数据序列。

上述分段方法的关键点在于:利用数据滑动窗口的统计特征计算逐点的阈值作为分段的依据,最大限度地细化分段点的选取,有效地避免基于固定压缩率的经典PAA方法对数据波动段特征信息的丢失。利用该思想可以将数据序列划分为波动段和平稳段,并分别进行近似重表达,这不仅确保了对包含较多信息的波动段的压缩准确度,还增大了对平稳信号段的压缩率,可以达到压缩率和精确度的平衡。整体而言,信息动态表达的PAA方法在保证数据压缩度的情况下尽可能地保留了原序列的特征,充分考虑了用电数据所含信息的动态表达,为后续边缘计算的有效运行创造了可能性。

3 基于分段加权 DTW 距离的异常用电识别算法

为了使云端准确地衡量压缩后的用电数据之间的相似度,本文提出了分段加权 DTW 距离。其通过对分段内的数据取平均值得到综合数据点的方式,衡量用电曲线的局部特征信息,准确度量压缩数据间的相似度,再结合带噪声检测的密度空间聚类 DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 算法得到异常用电识别结果。

3.1 基于分段加权 DTW 距离的相似性度量

用电数据在经过信息动态表达的 PAA 方法压缩后,其长度是不可控的,为此,本文使用分段加权的 DTW 距离来衡量用电数据间的相似度,动态压缩后的数据也能弥补 DTW 距离在运行效率上的劣势。DTW 距离最早被用于解决语音识别领域中的时间对准难题,之后被引入时间序列的相似性研究中。经典 DTW 距离的具体计算步骤如下。

设 2 个时间序列 $A=[a_1, a_2, \dots, a_m]$ 和 $B=[b_1, b_2, \dots, b_n]$, 其长度分别为 m 和 n , 为了求得 A 和 B 之间的 DTW 距离,需先构造一个 $m \times n$ 阶的初始距离矩阵 M , 如式(3)所示。

$$M = \begin{bmatrix} d(a_m, b_1) & d(a_m, b_2) & \dots & d(a_m, b_n) \\ \vdots & \vdots & & \vdots \\ d(a_2, b_1) & d(a_2, b_2) & \dots & d(a_2, b_n) \\ d(a_1, b_1) & d(a_1, b_2) & \dots & d(a_1, b_n) \end{bmatrix} \quad (3)$$

式中: $d(a_i, b_j)$ 为 a_i 和 b_j 这 2 个点之间的距离,通常采用欧氏距离,即 $d(a_i, b_j) = \sqrt{(a_i - b_j)^2}$ 。

计算 DTW 距离的关键是要找到 1 条满足边界条件和单调性的规整路径 $P=\{p_1, p_2, \dots, p_K\}$, 其中 $p_k=(a_i, b_j)_k (k=1, 2, \dots, K)$ 为路径中的第 k 组点,表示

a_i 与 b_j 相匹配,且有 $\max\{m, n\} \leq K \leq m+n-1$ 。满足上述条件的路径 P 有多条,需要找到 1 条最优路径使得根据该路径中各组匹配关系计算得到的距离之和最短。可采用递归法快速求取累加距离,如式(4)所示。

$$D(i, j) = \min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\} + d(a_i, b_j) \quad (4)$$

式中: $D(i, j)$ 为按照规整路径的匹配关系计算得到的终点为 $p_k=(a_i, b_j)_k$ 的累加距离。

则时间序列 A 和 B 之间的 DTW 距离 $d_{\text{DTW}}(A, B)$ 可表示为:

$$d_{\text{DTW}}(A, B) = D(m, n) = \min\left\{\sum_{k=1}^K d(p_k)\right\} \quad (5)$$

式中: $d(p_k)$ 为最优路径中匹配点 p_k 之间的欧氏距离。

由于本文使用的用电数据经过信息动态表达的 PAA 方法压缩处理,当计算时间序列 A 和 B 之间的 DTW 距离时,匹配所得的 2 个点 a_i 和 b_j 实际上是 2 个分段的数据平均值,分别如式(6)和式(7)所示。

$$a_i = \frac{1}{m} \sum_{s=1}^m a_{is} \quad (6)$$

$$b_j = \frac{1}{n} \sum_{y=1}^n b_{jy} \quad (7)$$

式中: a_{is} 为第 i 个数据分段的第 s 个元素; b_{jy} 为第 j 个数据分段的第 y 个元素。 a_i, b_j 包含了分段的信息,相当于分段内的点取相同的权重值得到综合数据点,其包含了曲线的局部特征信息。由于经典 DTW 距离能度量时间序列的整体特征,本文提出的分段加权 DTW 距离在衡量由综合数据点构成的压缩数据的相似度时,考虑了时间序列的整体和局部双重特征。

3.2 基于 DBSCAN 算法的异常用电识别

DBSCAN 算法的核心思想是根据密度可及性关系在数据集中找到最大密度连通样本集,并将该样本集视为同一类。算法的主要步骤如下。

1)输入数据集 D 、DBSCAN 算法的邻域半径 E 和最小聚类点数 M 。

2)从数据集 D 中随机选择一个未被标记的样本 d_i , 并将其标记为已访问,若数据集中的样本 d_j 到 d_i 的 DTW 距离满足式(8)且 d_i 邻域半径 E 内的密度满足式(9),则将 d_i 加入新的一类 C_i 中,其邻域半径 E 内的样本集合 $N_E(d_i)$ 加入候选集 H 。

$$d_{\text{DTW}}(d_j, d_i) \leq E \quad (8)$$

$$\text{card}(N_E(d_i)) \geq M \quad (9)$$

式中: $\text{card}(\cdot)$ 表示括号内集合的元素个数。

3)从候选集 H 中随机选择一个未被访问的样本 d_k , 并将其标记为已访问,若其邻域半径 E 内的密度

满足式(9),且 d_k 没有加入其他类,则将 d_k 加入 C_i 中,实现对类簇的扩展,并将集合 $N_E(d_k)$ 中不属于候选集 H 的元素加入候选集中。

4)重复执行步骤3),直到候选集 H 中所有样本都被访问。

5)如果数据集 D 中的所有样本都已被访问,则输出聚类结果;否则,执行步骤2)~4)。

DBSCAN算法的最大特点在于其能根据数据间的相似度找出数据集中的离群点。在本文的异常用电识别算法中,将用电数据映射为2维平面内的点集,它们之间的距离由数据的分段加权DTW距离决定,由于只是少量的异常数据,因此大部分的点都会分布在一个密集区域内,而少量的用电行为异常数据不会匹配正常用电模式,会明显地零星分布于稀疏区域内。若1条用电数据与其余数据的相似度在2维平面上的映射关系被算法的 E 和 M 这2个参数界定为高密度区,则该数据的相似度为合格;否则,数据的相似度为不合格,该数据就是本文寻找的异常数据。

由上述步骤可以看出,DBSCAN算法识别异常数据完全是自动的,并不需要对聚类结果进行二次处理,这是其优点。但是其缺点也很明显,DBSCAN算法需要预先确定邻域半径 E 和最小聚类点数 M 这2个参数,且对参数的取值很敏感,这会极大地影响算法的性能。如果 E 太小,则很多样本将无法聚集,会出现许多噪声点;如果 E 太大,则很多类簇会合并,导致大部分的样本被分到同一个类簇中,噪声点的数量会比实际结果少。为了确定合适的参数取值,本文采用拐点法^[19]确定邻域半径 E 和最小聚类点数 M 。

计算样本 d_i 到其第 k 个最近邻的DTW距离 $d_{DTW}^k(d_i)$,按照升序绘制曲线,结果如附录A图A3所示。图中,平缓区域表示样本周围的密度基本一致,而陡峭区域意味着密度发生明显变化,拐点为陡峭区域与平缓区域的交界处,显然拐点对应的 d_{DTW}^k 值是一个合适的邻域半径 E 。拐点的精确寻找往往比较困难,因此本文通过 d_{DTW}^k 曲线斜率的变化找出大致的拐点,将拐点对应的 d_{DTW}^k 值作为邻域半径 E 。根据文献[20]中的结论,当 $k>4$ 时, d_{DTW}^k 曲线与 d_{DTW}^4 曲线没有显著差异,且当 $k>4$ 时 d_{DTW}^k 的计算量更多,故本文将 M 设置为4。

本文所提异常用电识别云边协同方法的流程图如图3所示。

4 算例分析

本文选取我国某地区计量中心的长期日负荷曲线作为算例分析对象,每15 min采样一次。为了验证本文所提数据压缩方法、相似性度量方法以及异

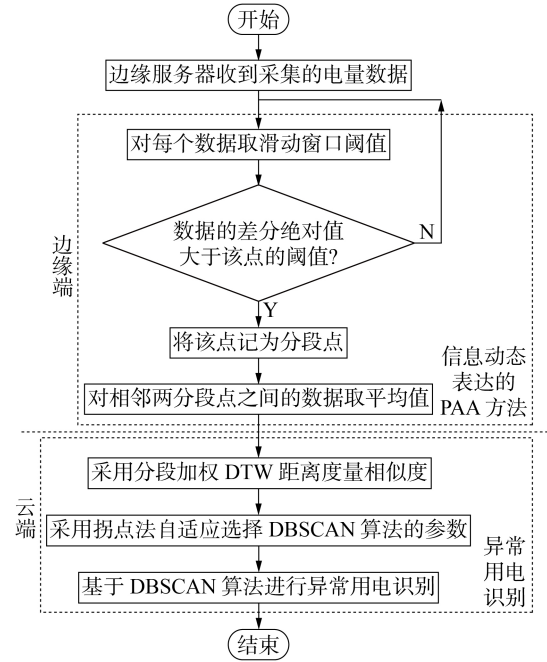


图3 异常用电识别云边协同方法的流程图

Fig.3 Flowchart of cloud-edge collaboration method for abnormal power consumption recognition

常用电识别方法的性能,对基于欧氏距离的自适应参数选择DBSCAN算法(方法1)、基于DTW距离的自适应参数选择DBSCAN算法(方法2)、基于DTW距离的K-means算法(方法3)、基于PAA和分段加权DTW距离的自适应参数选择DBSCAN算法(方法4)、基于信息动态表达的PAA和分段加权DTW距离的经验参数选择DBSCAN算法(方法5)以及基于信息动态表达的PAA和分段加权DTW距离的自适应参数选择DBSCAN算法(本文方法)进行对比分析,并对压缩效果、识别准确度、计算效率进行综合评估。其中:方法1~3为已有常用的云端集中式异常用电识别方法,无数据压缩阶段;而方法4、5为针对云边协同场景下的常用异常用电识别方法,有数据压缩阶段;本文方法由Python程序实现。

4.1 数据压缩效果分析

为了定量地评估本文方法在压缩率和误差上的优势,此处给出压缩率和均方根误差的定义。设原始数据序列 $X=[x_1, x_2, \dots, x_n]$,其经压缩后的序列为:

$$Y = \left[\text{mean}(x_1, x_2, \dots, x_{l_1}), \text{mean}(x_{l_1+1}, x_{l_1+2}, \dots, x_{l_1+l_2}), \dots, \text{mean}\left(x_{\sum_{i=1}^{m-1} l_i + 1}, x_{\sum_{i=1}^{m-1} l_i + 2}, \dots, x_{\sum_{i=1}^m l_i}\right) \right] \quad (10)$$

式中: $\text{mean}(\cdot)$ 表示取括号内元素的均值; l_i 为第 i 个分段的元素数量,且有 $\sum_{i=1}^m l_i = n$ 。则压缩算法的压缩率 γ 为:

$$\gamma = \left(1 - \frac{m}{n}\right) \times 100\% \quad (11)$$

PAA方法的基本思想是用分段的平均值表达整个分段的信息,因此为了计算压缩数据和原始数据之间的均方根误差,将分段内的数据均用该分段内数据的平均值表示^[21],对所有分段进行同样的处理就能得到1条重构序列 $X'' = [x''_1, x''_2, \dots, x''_n]$,以第1个分段为例,即:

$$x''_1 = x''_2 = \dots = x''_{l_1} = \text{mean}(x_1, x_2, \dots, x_{l_1}) \quad (12)$$

该重构序列与原始序列之间的均方根误差 e 为:

$$e = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - x'_i)^2} \quad (13)$$

对图4(a)所示采样点数为96的典型三峰负荷曲线进行压缩处理:若采用经典PAA方法,设压缩窗口 $w=3$,得到的压缩曲线如图4(b)所示,该压缩曲线的剩余采样点数为32,压缩率为66.67%;采用本文考虑信息动态表达的PAA方法对曲线进行压缩,得到的压缩曲线如图4(c)所示,该压缩曲线的剩余采样点数为27,压缩率为71.88%。可见,本文方法对原始曲线的平缓区域采用更大的压缩率,缩减了后续数据分析的数据量。

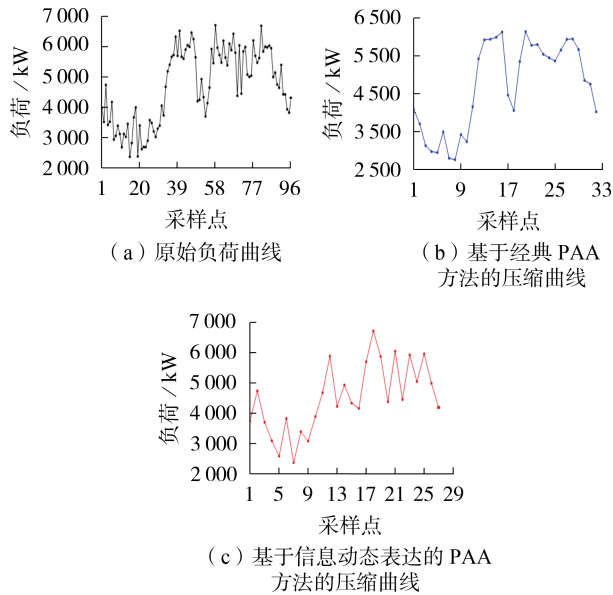
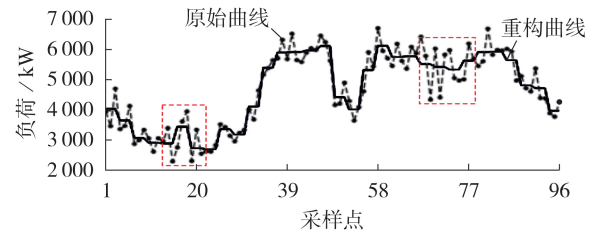


图4 数据压缩效果对比

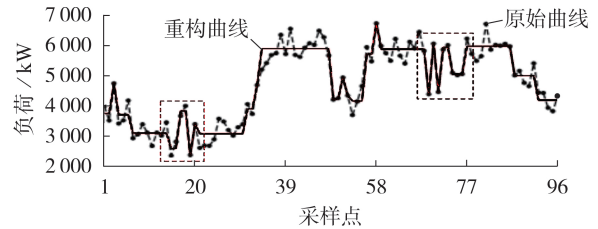
Fig.4 Comparison of data compression effect

虽然经典PAA方法所得压缩曲线与原始负荷曲线形状在直观上基本一致,但数据的压缩效果不能基于压缩曲线与原始曲线的相似度进行直接度量,这是因为不同分段方法会造成原始数据分段结果的不同,且不同分段内的压缩率也不同。因此,需要根据压缩曲线进行数据重构,根据重构数据与原始数据的对比对压缩效果进行有效衡量。不同压缩方法的重构曲线如图5所示。可知,经典PAA方法

的均方根误差为399 kW,而信息动态表达的PAA方法的均方根误差为283 kW。经典PAA方法的压缩窗口固定不变,并不会随着曲线的平稳和波动而变化,在含有信息较多的波动段丢失了大量的信息,导致其均方根误差更大;而本文所提信息动态表达的PAA方法具有压缩率动态特性,在一些突变点所在分段的压缩率明显变小,甚至整个分段直接被保留下来,曲线的特征没有丢失。采用本文所提信息动态表达的PAA方法对数据进行处理,数据中含有的信息能得到动态的表达。



(a) 经典PAA方法



(b) 信息动态表达的PAA方法

图5 不同压缩方法的重构曲线对比

Fig.5 Comparison of reconstruction curves between different compression methods

为了进一步验证本文所提信息动态表达的PAA方法在压缩率和均方根误差方面的优势,对5条负荷曲线分别添加信噪比为15、20、25 dB的噪声,并将经典PAA方法的压缩窗口 w 分别设为2—4,对比信息动态表达的PAA方法和经典PAA方法的性能。为了保证单一变量原则,5条曲线原本的噪声已被平滑,不同压缩方法的压缩率、均方根误差结果见附录A表A1,均值结果如表1所示。

由表A1和表1可知,相较于经典PAA方法,本文所提信息动态表达的PAA方法在压缩率和准确度2个方面都更优,即信息动态表达的PAA方法在压缩率更高的情况下还能更准确地体现原始数据的负荷特征,合理地利用云边计算资源以及缓解带宽的传输压力,使得云边协同架构下的异常用电识别得以实现。进一步分析结果可得:当样本的噪声更大,即曲线的波动程度更大时,本文所提信息动态表达的PAA方法相较于经典PAA方法的压缩效果更明显,这是因为信息动态表达的PAA方法对数据进行动态压缩,较好地保留了波动段数据的特征;当2种压缩方法的压缩率在同一水平时,本文所提信息动态

表 1 不同压缩方法的性能比较
Table 1 Performance comparison of different compression methods

压缩方法	压缩率均值 / %			均方根误差均值 / kW		
	信噪比 为 15 dB	信噪比 为 20 dB	信噪比 为 25 dB	信噪比 为 15 dB	信噪比 为 20 dB	信噪比 为 25 dB
经典 PAA ($w=2$)	50.00	50.00	50.00	579	333	198
经典 PAA ($w=3$)	66.67	66.67	66.67	684	396	241
经典 PAA ($w=4$)	75.00	75.00	75.00	728	432	275
信息动态 表达的 PAA	73.96	76.04	75.21	533	337	217

表达的 PAA 方法明显具有更好的准确性,而若要提高经典 PAA 方法的准确性,则需要牺牲压缩率,增大数据计算量,故难以适用于云边协同的应用场景。

4.2 异常用电识别结果分析

本节对含有异常样本的日负荷曲线进行异常用电识别,根据 3.2 节将由 DBSCAN 算法识别出的噪声点作为异常,并统计正检 TP(True Positive)、误检 FP(False Positive)、漏检 FN(False Negative)的数量,计算查准率 P 和查全率 R ,分别如式(14)和式(15)所示。

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad (14)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad (15)$$

式中: N_{TP} 、 N_{FP} 、 N_{FN} 分别为正检、误检、漏检数量。

本文所提信息动态表达的 PAA 方法的目标之一是通过数据压缩减少算法的计算量,因此对算法的时间复杂度进行分析。虽然传统欧氏距离的时间复杂度小,但其已被证明准确性不佳。而虽然 DTW 距离在准确性上有较好的效果,但其时间复杂度随着曲线数据量的增加而呈现平方倍的增大,2 条长度为 n 的时间序列之间 DTW 距离的时间复杂度为 $O(n^2)$ 。本文所提分段加权 DTW 距离的时间复杂度为 $O((1-\gamma)^2 n^2)$,与压缩率有关:若压缩率为 30%,则其时间复杂度减半;若压缩率达到 70%,则其时间复杂度仅为 DTW 距离的 10%。可见,分段加权 DTW 距离能有效减少计算量,使其适用于云边协同的应用场景。

不同方法的异常用电识别结果如表 2 所示。由表可见:相比于方法 1,方法 2 的效果提升明显,这是得益于 DTW 距离对相似性度量的准确性,但同时也极大地降低了算法的效率;方法 3 的效果明显弱于方法 2,这说明 DBSCAN 算法基于样本密度的自动异常识别显著优于 K -means 算法依赖阈值设置的异常识别;方法 4 使用 PAA 方法对数据进行压缩,可以显著提高算法的运行效率,但是由于 PAA 方法对数

据信息的保留比较粗略,导致方法 4 的效果仍不够理想;方法 5 和本文方法的唯一差别就在于,方法 5 根据经验选择 DBSCAN 算法的参数,而没有使用自适应参数选择方法,由于 DBSCAN 算法对参数敏感,直接导致方法 5 的效果不佳;本文方法改进了数据压缩方法,实现了用电信息的动态表达,极大地提高了数据的重表达准确性,相比于方法 4,其异常用电识别性能明显提高。

表 2 不同方法的异常用电识别结果

Table 2 Abnormal power consumption recognition results of different methods

方法	N_{TP}	N_{FP}	N_{FN}	$P / \%$	$R / \%$
1	23	17	27	57.50	46
2	34	11	16	75.56	68
3	25	24	25	51.02	50
4	32	13	18	71.11	64
5	32	17	18	62.50	50
本文方法	36	12	14	75.00	72

进一步分析不同方法的异常识别结果可发现:方法 4 和方法 2 使用的相似性度量方法不同,但效果基本相同,这说明本文所提分段加权 DTW 距离能准确度量压缩数据的相似性,且比经典 DTW 距离的计算速度更快;本文方法使用了自适应参数选择的 DBSCAN 算法,其效果相较于方法 5 提升较大,验证了本文所提自适应参数选择 DBSCAN 算法的有效性,解决了 DBSCAN 算法对参数选择敏感的问题。

综上可知,相较于其他方法,本文方法的综合性能最优,查准率和查全率均较高,且具有较好的计算效率。因此,本文方法在云边协同架构下有较强的实用性,能在解决计算资源紧张和传输带宽占用大这 2 个问题的同时,对最终的异常用电识别结果影响也不大。本文方法识别所得异常用电识别曲线如图 6 所示,为了便于观察,图中只给出了 38 条正常曲线和 2 条异常曲线。由图可见,2 条异常曲线明显偏离正常曲线。

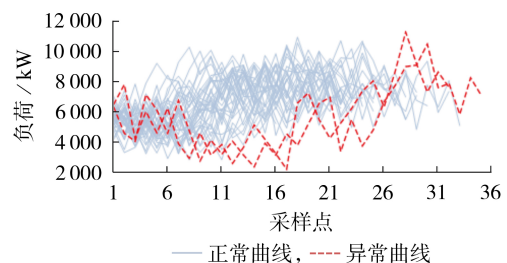


图 6 本文方法的异常用电识别曲线

Fig.6 Abnormal power consumption recognition curves of proposed method

5 结论

为了在云边协同架构下准确识别异常用电,本

文提出了一种考虑信息动态表达的异常用电模式识别云边协同方法,主要包括边缘端的信息动态表达的PAA数据压缩方法、云端的基于分段加权DTW距离的相似性度量及基于DBSCAN算法的异常用电识别3个部分。经过实际数据集进行算例分析,验证了本文方法在数据压缩效果、异常识别准确性、计算效率3个方面均具有较好的性能表现,能实现异常用电识别在云边协同场景的应用,充分考虑了计算资源不足的问题,并减轻了网络传输压力。同时也验证了本文所提信息动态表达的PAA方法对数据压缩的准确性、基于分段加权DTW距离衡量时间序列相似度的合理性以及将其应用在云边协同架构的可行性。

但本文方法尚未考虑异常用电原因的辨识,下一步工作将在已有研究的基础上,针对异常用电原因的识别进行进一步的研究。

附录见本刊网络版(<http://www.epae.cn>)。

参考文献:

- [1] BUZAU M M, TEJEDOR-AGUILERA J, CRUZ-ROMERO P, et al. Detection of non-technical losses using smart meter data and supervised learning[J]. IEEE Transactions on Smart Grid, 2019, 10(3): 2661-2670.
- [2] 赵文清, 沈哲吉, 李刚. 基于深度学习的用户异常用电模式检测[J]. 电力自动化设备, 2018, 38(9): 34-38.
ZHAO Wenqing, SHEN Zheji, LI Gang. Anomaly detection for power consumption pattern based on deep learning[J]. Electric Power Automation Equipment, 2018, 38(9): 34-38.
- [3] 李梓欣, 李川, 李英娜. 用电特征指标降维与极限学习机算法的窃电检测[J]. 计算机应用与软件, 2018, 35(12): 179-186.
LI Zixin, LI Chuan, LI Yingna. Reduced dimension of power consumption characteristics and detection of electricity larceny by extreme learning machine[J]. Computer Applications and Software, 2018, 35(12): 179-186.
- [4] 陈启鑫, 郑可迪, 康重庆, 等. 异常用电的检测方法: 评述与展望[J]. 电力系统自动化, 2018, 42(17): 189-199.
CHEN Qixin, ZHENG Kedi, KANG Chongqing, et al. Detection methods of abnormal electricity consumption behaviors: review and prospect[J]. Automation of Electric Power Systems, 2018, 42(17): 189-199.
- [5] 袁翔宇, 张蓬鹤, 熊素琴, 等. 基于逻辑回归算法的异常用电辨识方法研究[J]. 电测与仪表, 2021, 58(12): 81-87.
YUAN Xiangyu, ZHANG Penghe, XIONG Suqin, et al. Research on identification method of abnormal power consumption based on logistic regression algorithm[J]. Electrical Measurement & Instrumentation, 2021, 58(12): 81-87.
- [6] WANG X L, AHN S H. Real-time prediction and anomaly detection of electrical load in a residential community[J]. Applied Energy, 2020, 259: 114145.
- [7] NAGI J, YAP K S, TIONG S K, et al. Nontechnical loss detection for metered customers in power utility using support vector machines[J]. IEEE Transactions on Power Delivery, 2010, 25(2): 1162-1171.
- [8] JOKAR P, ARIANPOO N, LEUNG V C M. Electricity theft detection in AMI using customers' consumption patterns[J]. IEEE Transactions on Smart Grid, 2016, 7(1): 216-226.
- [9] ZANETTI M, JAMHOUR E, PELLEZZI M, et al. A tunable fraud detection system for advanced metering infrastructure using short-lived patterns[J]. IEEE Transactions on Smart Grid, 2019, 10(1): 830-840.
- [10] 田力, 向敏. 基于密度聚类技术的电力系统用电量异常分析算法[J]. 电力系统自动化, 2017, 41(5): 64-70.
TIAN Li, XIANG Min. Abnormal power consumption analysis based on density-based spatial clustering of applications with noise in power systems[J]. Automation of Electric Power Systems, 2017, 41(5): 64-70.
- [11] 宋军英, 崔益伟, 李欣然, 等. 改进分段线性表示与动态时间弯曲相结合的负荷曲线聚类方法[J]. 电力系统自动化, 2021, 45(2): 89-96.
SONG Junying, CUI Yiwei, LI Xinran, et al. Load curve clustering method combining improved piecewise linear representation and dynamic time warping[J]. Automation of Electric Power Systems, 2021, 45(2): 89-96.
- [12] 石亮缘, 周任军, 李娟, 等. 基于时间序列相似性度量的新能源-负荷特性指标[J]. 电力自动化设备, 2019, 39(5): 75-81.
SHI Liangyuan, ZHOU Renjun, LI Juan, et al. New energy-load characteristic index based on time series similarity measurement[J]. Electric Power Automation Equipment, 2019, 39(5): 75-81.
- [13] 王鹤, 李石强, 于华楠, 等. 基于分布式压缩感知和边缘计算的配电网电能质量数据压缩存储方法[J]. 电工技术学报, 2020, 35(21): 4553-4564.
WANG He, LI Shiqiang, YU Huanan, et al. Compression acquisition method for power quality data of distribution network based on distributed compressed sensing and edge computing[J]. Transactions of China Electrotechnical Society, 2020, 35(21): 4553-4564.
- [14] 孙文文, 何国庆, 刘纯, 等. 物联网背景下应用于光伏发电的边缘计算设备关键技术研究及应用[J]. 电力自动化设备, 2021, 41(7): 38-43.
SUN Wenwen, HE Guoqing, LIU Chun, et al. Research and application of key technologies for edge computing equipment used in photovoltaic power generation under background of IoT[J]. Electric Power Automation Equipment, 2021, 41(7): 38-43.
- [15] NOTARISTEFANO A, CHICCO G, PIGLIONE F. Data size reduction with symbolic aggregate approximation for electrical load pattern grouping[J]. IET Generation, Transmission & Distribution, 2013, 7(2): 108-117.
- [16] 牟婷婷, 陆微, 王兰君, 等. 基于主成分分析的用电模式稳定性分析[J]. 电力系统自动化, 2017, 41(19): 59-65.
MOU Tingting, LU Wei, WANG Lanjun, et al. Stability analysis of consumption mode based on principal component analysis[J]. Automation of Electric Power Systems, 2017, 41(19): 59-65.
- [17] 窦健, 刘宣, 卢继哲, 等. 基于用电信息采集大数据的防窃电方法研究[J]. 电测与仪表, 2018, 55(21): 43-49.
DOU Jian, LIU Xuan, LU Jizhe, et al. Research on electricity anti-stealing method based on power consumption information acquisition and big data[J]. Electrical Measurement & Instrumentation, 2018, 55(21): 43-49.
- [18] 徐严军, 吴蒙, 白佳灵, 等. 多特征提取与深度学习关口计量装置异常事件识别方法[J]. 中国测试, 2021, 47(5): 104-111.
XU Yanjun, WU Meng, BAI Jialing, et al. Identification method of abnormal events of gateway metering device based on multi-feature extraction and deep learning[J]. China Measure-

- ment & Testing Technology, 2021, 47(5):104-111.
- [19] ZHAO Nengwen, ZHU Jing, WANG Yao, et al. Automatic and generic periodicity adaptation for KPI anomaly detection[J]. IEEE Transactions on Network and Service Management, 2019, 16(3):1170-1183.
- [20] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[J]. Knowledge Discovery & Data Mining, 1996, 96(34):226-231.
- [21] 李海林, 梁叶. 分段聚合近似和数值导数的动态时间弯曲方法[J]. 智能系统学报, 2016, 11(2):249-256.
- LI Hailin, LIANG Ye. Dynamic time warping based on piecewise aggregate approximation and data derivatives [J]. CAAI Transactions on Intelligent Systems, 2016, 11(2):249-256.

作者简介:



刘慧自

刘慧自(1999—),男,硕士研究生,主要研究方向为电能质量与优质供电(E-mail:1457018026@qq.com);

汪颖(1981—),女,教授,博士,主要研究方向为电能质量与优质供电(E-mail:769429505@qq.com);

胡文曦(1993—),男,助理研究员,博士,通信作者,主要研究方向为电能质量大数据分析(E-mail:408803186@qq.com);

肖先勇(1968—),男,教授,博士,主要研究方向为电能质量与优质供电(E-mail:xiaoxianyong@163.com)。

(编辑 陆丹)

Cloud-edge collaboration method for abnormal power consumption pattern recognition considering dynamic expression of information

LIU Huizi, WANG Ying, HU Wenxi, XIAO Xianyong

(College of Electrical Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Abnormal power consumption recognition is an important part of the power consumption check and the operation status identification of metering devices, and is of great significance to maintain the safe operation of power grid and protect the rights and interests of normal users. In order to recognize multiple power consumption patterns of users, existing methods tend to cause complicated calculation on the basis of ensuring recognition accuracy. While, simple calculation methods considering efficiency are difficult to accurately measure the similarity of different power consumption patterns, so it is difficult to give consideration to calculation efficiency and accuracy. In addition, uploading power consumption data to the cloud for centralized computing consumes a large amount of network bandwidth and computing resources, which further limits the application of anomaly recognition. Therefore, a cloud-edge collaboration method for abnormal power consumption pattern recognition considering the dynamic expression of information is proposed. According to the computing resources of the edge terminal and the cloud, the cooperative tasks are reasonably allocated to realize the cloud-edge collaboration recognition of abnormal power consumption. Aiming at the problem of limited computing power of edge servers, the dynamic compression and re-expression of power consumption data are carried out to reduce the data amount and ensure the accuracy of data information. After receiving the compressed data, the cloud takes the segmented weighted dynamic time warping distance as the basis for the similarity measurement of compressed data, and identifies abnormal power consumption based on the density clustering algorithm with adaptive parameter selection. The effectiveness of the proposed method is verified based on the actual data set.

Key words: abnormal power consumption recognition; dynamic expression of information; cloud-edge collaboration; data compression; similarity measurement

附录 A

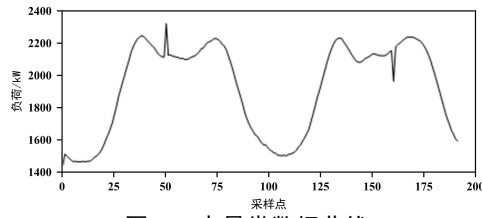


图 A1 点异常数据曲线
Fig.A1 Point abnormal data curve

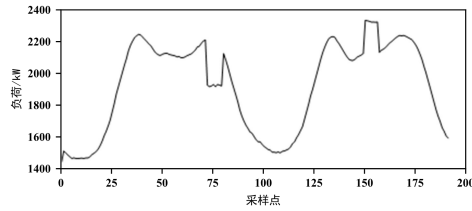


图 A2 连续异常数据曲线
Fig.A2 Continuous abnormal data curve

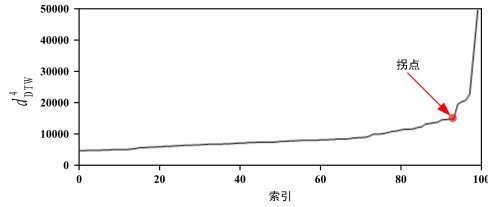


图 A3 100 个样本的 d_{DTW}^4
Fig.A3 d_{DTW}^4 values of 100 samples

表 A1 不同压缩方法的性能比较

Table A1 Performance comparison among different compression methods

信噪比/dB	样本	经典 PAA 方法 $w = 2$		经典 PAA 方法 $w = 3$		经典 PAA 方法 $w = 4$		本文压缩方法	
		压缩率/%	均方根 误差/kW	压缩率/%	均方根 误差/kW	压缩率/%	均方根 误差/kW	压缩率/%	均方根 误差/kW
15	1	50	638	66.67	739	75	765	75	591
	2	50	579	66.67	703	75	775	77.08	553
	3	50	626	66.67	709	75	755	70.83	526
	4	50	509	66.67	634	75	661	73.96	465
	5	50	543	66.67	637	75	683	72.92	532
	平均	50	579	66.67	684	75	728	73.96	533
20	1	50	370	66.67	436	75	441	77.08	369
	2	50	339	66.67	403	75	471	76.04	327
	3	50	349	66.67	406	75	440	73.96	335
	4	50	294	66.67	367	75	398	76.04	312
	5	50	315	66.67	368	75	408	77.08	344
	平均	50	333	66.67	396	75	432	76.04	337
25	1	50	222	66.67	274	75	273	70.83	177
	2	50	206	66.67	241	75	307	75	225
	3	50	198	66.67	243	75	275	73.96	249
	4	50	177	66.67	222	75	259	79.17	211
	5	50	189	66.67	224	75	260	77.08	225
	平均	50	198	66.67	241	75	275	75.21	217