

基于快速动态时间弯曲和最小覆盖球的 多日负荷曲线聚类方法

刘晓峰¹, 康进¹, 马翔², 沃建栋², 吕磊炎², 吴浩¹

(1. 浙江大学电气工程学院, 浙江杭州 310027; 2. 国网浙江省电力有限公司金华供电公司, 浙江金华 321016)

摘要:现有负荷曲线聚类研究主要基于单日负荷曲线或多日同时刻的负荷分布开展,忽略了负荷在多日间的波动特性和负荷曲线的时间滞后特性,导致聚类结果的准确度和鲁棒性不足。综合考虑负荷的波动特性和时间滞后特性,提出一种快速动态时间弯曲和最小覆盖球相结合的多日负荷曲线聚类方法。在考虑负荷时间滞后特性的基础上,利用快速动态时间弯曲和多维尺度缩放对负荷曲线进行降维;为每个降维负荷迭代寻找最小覆盖球,并计算不同覆盖球的球间相似度;利用谱聚类算法得到相似度矩阵。算例结果表明,所提方法在准确度和鲁棒性上较传统方法有一定优势。

关键词:多日负荷曲线;快速动态时间弯曲;最小覆盖球;谱聚类;多维尺度缩放

中图分类号: TM 714

文献标志码: A

DOI: 10.16081/j.epae.202205049

0 引言

“十四五”以来,以新能源为主体的新型电力系统得到大力发展,电力负荷作为终端,是新型电力系统建设的重点,也是源网荷储的重要环节之一。为构建清洁、低碳、安全、高效的能源体系,新能源分布式发电、柔性负荷、电动汽车等的比重大幅上升^[1]。该类负荷受天气、调控、用户意愿的影响较大,使得总负荷在多日间的波动较明显,给负荷曲线聚类研究带来了难度。此外,随着用电信息采集装置的部署与数据采集技术的发展,电力系统负荷数据资源得到了快速积累^[2],给多日间负荷波动的研究带来了可能性。因此,有必要且有条件进行多日负荷曲线聚类研究,探索负荷在多日间的波动特性,进而提升负荷调控能力以及深化电力体制改革^[3-4]。

负荷曲线聚类是电力系统负荷特性分析的重要手段之一^[5]。负荷曲线聚类研究通常包括负荷特征提取及降维、负荷间的相似度定义、聚类算法这3个方面。近年来,众多研究者针对上述3个方面,从算法效率、聚类准确率等多个角度对传统负荷曲线聚类研究进行了改进及深化。

在负荷特征提取及降维方面,现有研究主要从2个角度开展。一部分研究针对特定的聚类方法,选取合适的负荷特征向量:文献[6]通过离散小波变

换提取负荷曲线的频域特征,并与模糊K-modes聚类方法结合得到聚类结果;文献[7]提出一种谱聚类特征向量的优化选取方法,提高了谱聚类算法的效率。另一部分研究采用传统降维方法对负荷曲线进行降维:文献[8]利用奇异值分解对日负荷曲线进行降维,从而提取特征;文献[9]比较自组织映射、主成分分析等降维技术在电力系统负荷曲线聚类中的应用效果。

在负荷间的相似度定义方面,现有研究主要从3个角度开展,分别为采用欧氏距离等点对点的距离作为相似度定义^[10],采用动态时间弯曲等考虑负荷曲线时间滞后特性的相似度定义^[11],以及利用前2种方式综合衡量负荷曲线间的相似度。

在聚类算法方面,常见算法有层次聚类算法、划分聚类算法、密度聚类算法、模型聚类算法、谱聚类算法等^[5]。由于谱聚类算法能够划分任意形状的数据集,且对稀疏矩阵的处理效果优于其他算法,因此本文采用该算法。

从聚类对象上而言,上述研究大多集中于单日负荷曲线。实际上多日间的负荷曲线可能差异较大,导致算法在不同日间应用的稳定性较差,因此有必要开展多日负荷曲线的降维与聚类研究。目前多日负荷曲线聚类方法主要是统计多日同一时刻的负荷分布情况以表征负荷的多日波动情况^[12-13]。由于负荷曲线属于时间序列,不同日的相同用电行为未必发生在同一时刻,因此应考虑日负荷曲线的时间滞后特性。

针对现有研究的不足,本文提出一种基于快速动态时间弯曲FDTW(Fast Dynamic Time Warping)与最小覆盖球的多日负荷曲线聚类方法。首先,基于FDTW与多维尺度缩放MDS(Multi-Dimensional

收稿日期:2022-02-28;修回日期:2022-05-04

在线出版日期:2022-05-16

基金项目:国家自然科学基金资助项目(U2066601);国网浙江省电力有限公司科技项目(5211JH1900M3)

Project supported by the National Natural Science Foundation of China(U2066601) and the Science and Technology Project of State Grid Zhejiang Electric Power Co., Ltd. (5211JH1900M3)

Scaling)提取负荷曲线的时间滞后特性,并将负荷曲线降至3维;然后,为每个负荷迭代寻找最小覆盖球,并基于最小覆盖球定义负荷间的相似度计算公式;最后,将相似度矩阵应用于基于 k -最近邻法的谱聚类算法。算例结果表明,本文所提方法在准确度和鲁棒性上较传统方法有一定优势。

1 基于FDTW和最小覆盖球的多日负荷曲线聚类

1.1 多日负荷曲线的特点

为了说明多日负荷曲线的特点,图1给出了某地区一台10 kV专变2周(2019年8月1日至2019年8月14日,其中8月1日为周四)内的日负荷曲线图。由图1(a)所示的归一化功率曲线可见:普通日负荷曲线、曲线3和曲线4均为双峰曲线,其形状大致相似,且这些曲线数量占85%以上,因此可以确定该专变下主要为双峰型负荷;曲线1和曲线2分别具有晚高峰和错峰曲线的特点,与双峰曲线相似度极低,经调查,曲线1和曲线2分别出现在8月9日(周五)和8月10日(周六),因此,异常曲线不完全是由节假日、非工作日产生的。由图1(b)所示的实际功率曲线可见,曲线1和曲线2的实际功率较高,可排除数据采集、日内负荷波动(此处负荷波动表示日内负荷曲线的起伏,以日内负荷波动命名,以与多日间的负荷波动相区别)等异常的影响,因此,推测此类曲线的出现可能是该专变下用户加班、调休等原因所导致的。显然,如果采用单日负荷曲线聚类的方法提取到曲线1或曲线2所对应的负荷,则该负荷的聚类结果就会错误,因此,有必要考虑多日负荷曲线间的波动特性,研究多日负荷曲线的聚类方法。

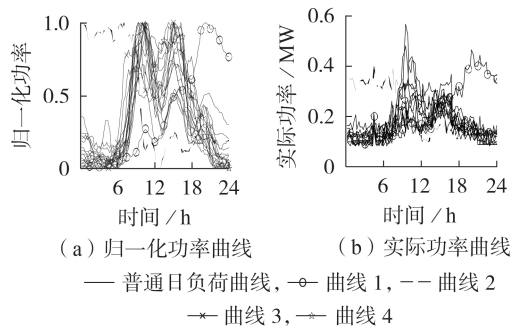


图1 某10 kV专变2周内的日负荷曲线

Fig.1 Daily load curves of a 10 kV specialized transformer within two weeks

曲线3和曲线4分别出现在8月3日(周六)和8月6日(周二),二者均为双峰曲线,但是曲线4相对于曲线3有约2 h的时间滞后,因此,在多日负荷曲线的聚类过程中需考虑负荷曲线的时间滞后特性。

目前,从电网获取有功功率数据的采样间隔为

15 min,相应地,日负荷曲线的数据共有96点,即数据维度为96维。考虑到多日负荷曲线的时间通常为2周及以上,直接对负荷曲线进行聚类时数据维度超过1000维,可能导致聚类的准确度和鲁棒性较低,因此,有必要在多日负荷曲线聚类的过程中对其进行降维。

1.2 算法思路

针对上述多日负荷曲线的特点,本文提出一种基于FDTW和最小覆盖球的多日负荷曲线聚类方法。该方法主要包含负荷曲线降维、基于最小覆盖球的相似度定义、谱聚类这3个步骤。

1)负荷曲线降维。将多日负荷曲线按日分割,考虑负荷曲线的时间滞后特性,采用FDTW形成日负荷曲线距离矩阵,并采用MDS算法将该距离矩阵降至3维。

2)基于最小覆盖球的相似度定义。在3维空间中为每个负荷迭代寻找最小覆盖球,并以3种方式度量球间的相似度,以反映多日间的负荷波动大小,并减少负荷异常日的影响。

3)谱聚类。采用 k -最近邻法根据球间相似度矩阵构造相似度图,并将其应用于谱聚类算法,得到多日负荷曲线的聚类结果。

所提方法的流程图如附录A图A1所示。

2 基于FDTW和MDS的负荷曲线降维

2.1 降维思路

设多日负荷曲线的样本数为 n (即为10 kV专变和公变数),持续时间为 m 日,每日采样点为96点,将负荷数据构成的矩阵记为 $X \in \mathbf{R}^{n \times 96m}$ 。首先,对数据矩阵 X 进行曲线平滑及max-min归一化处理;其次,考虑负荷曲线的时间滞后特性,将 n 个负荷按日分割,得到重构的负荷数据矩阵 $Y \in \mathbf{R}^{nm \times 96}$,并将矩阵 Y 中有数据采集异常的日负荷曲线剔除,得到 $Y_1 \in \mathbf{R}^{s \times 96}$ ($s \leq nm$);然后,利用FDTW计算日负荷曲线之间的距离,得到日负荷曲线距离矩阵 $D \in \mathbf{R}^{s \times s}$;最后,利用MDS对矩阵 D 进行降维,得到3维负荷数据矩阵 $Z \in \mathbf{R}^{s \times 3}$ 。

2.2 FDTW

动态时间弯曲DTW(Dynamic Time Warping)在计算2个时间序列的距离时自动扭曲时间序列(即点与点间的匹配在时间轴上进行局部的提前或滞后,但不打乱时间序列内部各点的先后次序),使得2个序列的形态尽可能一致,得到最短距离^[11]。

给定长度分别为 x, y 的时间序列,首先,计算2个序列各点之间的距离,形成距离矩阵 $d \in \mathbf{R}^{x \times y}$,然后,设定起始条件 $L(1,1)=d(1,1)$,递推规则如式(1)所示,构造路径矩阵 $L \in \mathbf{R}^{x \times y}$,最终得到的 $L(x,y)$ 即为2个序列的DTW距离。

$$L(i, j) = \min(L(i, j-1), L(i-1, j), L(i-1, j-1)) + d(i, j) \quad (1)$$

在利用DTW计算负荷曲线间的距离时算法复杂度较高,且可能会造成时间尺度过度扭曲,因此,本文采用FDTW^[14]限制时间扭曲尺度,即算法搜索范围,使时间复杂度由 $O(t^2)$ 降为 $O(\tau t)$,其中 $t = \max(x, y)$, τ 为限制算法搜索的范围。

由于图1(a)中曲线3和曲线4间的时间滞后约为2h,因此,本文取 $\tau=2$ h,即8个点。FDTW提取负荷曲线时间滞后特性后形成距离矩阵 $D \in \mathbf{R}^{s \times s}$,其第 i 行第 j 列的元素 $D(i, j)$ 为第 i 条和第 j 条日负荷曲线间的FDTW距离。

2.3 MDS

MDS是一种能在降维过程中保持数据样本间距离或者相似性的方法^[15]。本文将MDS应用于FDTW所得的距离矩阵 $D \in \mathbf{R}^{s \times s}$,将其降至3维矩阵 $Z \in \mathbf{R}^{s \times 3}$,使得任意2条日负荷曲线在3维空间中的欧氏距离尽可能等于矩阵 D 所反映的距离,即:

$$\min \left[\sum_{1 \leq i, j \leq s} (\|z_i - z_j\| - D(i, j))^2 \right]^{1/2} \quad (2)$$

式中: z_i 和 z_j 分别为第 i 条和第 j 条日负荷曲线在3维空间中的表示。在 s 个已知条件下求取3维未知坐标,由于 $s \gg 3$,因此式(2)有解,本文采用梯度下降法对该式进行求解。

3 基于最小覆盖球的相似度定义

3.1 最小覆盖球的生成

本文的最小覆盖球是指包含3维空间中给定点的最小球^[16],其生成的具体流程如下。

1)以离均值点最近的3个点构造初始最小覆盖球。

2)计算球外各点与球心的欧氏距离,对球外负荷点按照相应欧氏距离从小到大的顺序重新排序。

3)计算将球外距球心欧氏距离最小的3个点包围进来后的最小覆盖球。

4)重复步骤2)、3),直至60%的点被包含进球内。

5)重复步骤2),并计算将球外距球心欧氏距离最小的1个点包围进来后的最小覆盖球。

6)计算球内负荷点数量和成本指标 I_{cost} , I_{cost} 为球体积增大的比例与球内点数量增加的比值的比值,反映了在最小覆盖球中添加某点所需要付出的成本,其表达式为:

$$I_{\text{cost}} = \frac{(r_+/r_-)^3}{n_+/n_-} \quad (3)$$

式中: r_- 、 n_- 和 r_+ 、 n_+ 分别为添加某点前、后的最小覆

盖球半径和球内负荷点数。

7)当 $I_{\text{cost}} > 2$ 时,认定该点为异常负荷点,输出包围该点前的最小覆盖球,迭代结束,否则,重复步骤5)、6)。

8)得到最小覆盖球矩阵 $B \in \mathbf{R}^{n \times 4}$,其中前3列为球心坐标,第4列为球半径。

最小覆盖球实质上体现了剔除负荷异常日后各负荷点的最优包络,其球心及半径反映了该负荷的日负荷曲线特征及波动特性。

3.2 相似度定义

2个最小覆盖球之间存在相离、相交、相含3种情形,因此,须对3种情形分别定义相似度计算方式。为此,先明确3种情形下的相似度定义要求。

1)两球相离应表示2个负荷的相似度为0。

2)两球相交时的相似度应与两球相交的方向无关,而仅与相交部分的体积占比有关。

3)两球相含时的相似度不仅与被含的体积相关,还与两球半径之差和球心间距离相关。具体地:当两球半径之差一定时,球心间距离越小则相似度越高;当球心间距离一定时,两球半径之差越小则相似度越高。

根据上述3个要求,对给定半径分别为 r_1 和 r_2 、体积分别为 V_1 和 V_2 的2个球,可定义相似度如下:

$$\varepsilon_{\text{si}} = \begin{cases} 0 & \text{相离} \\ \frac{V_{\text{ol}}}{\sqrt{V_1 V_2}} & \text{相交} \\ \frac{r_{\min}^2}{\sqrt{(r_1 r_2)^3}} \frac{r_1 + r_2 - d_{12}}{2} & \text{相含} \end{cases} \quad (4)$$

式中: V_{ol} 为两球相交部分的体积,其计算公式如式(5)所示; $r_{\min} = \min(r_1, r_2)$; d_{12} 为球心间距离。

$$V_{\text{ol}} = \pi h_1^2 \left(r_2 - \frac{1}{3} h_1 \right) + \pi h_2^2 \left(r_1 - \frac{1}{3} h_2 \right) \quad (5)$$

式中: h_1 、 h_2 的含义如图2所示。

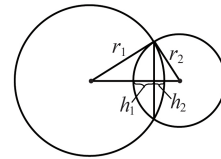


图2 两球相交示意图

Fig.2 Schematic diagram for intersection of two balls

显然,该相似度定义满足上述3个要求。当两球相离时相似度最小,此时 $\varepsilon_{\text{si}}=0$;当两球重合时, $r_1=r_2$ 且 $d_{12}=0$,此时 ε_{si} 达到最大值1;由于式(6)、(7)恒成立,因此, ε_{si} 处于 $[0, 1]$ 之间,其值越大,则2个负荷越相似。

$$V_{\text{ol}} \leq \min(V_1, V_2) \leq \sqrt{V_1 V_2} \quad (6)$$

$$r_{\min}^2(r_1 + r_2 - d_{12}) \leq r_{\min}^2(r_1 + r_2) \leq 2r_{\min}^2 r_{\max} = 2r_1 r_2 r_{\min} \leq 2\sqrt{(r_1 r_2)^3} \quad (7)$$

式中: $r_{\max} = \max(r_1, r_2)$ 。式(6)反映了大小球相交的体积小于等于小球的体积, 小于等于大小球体积的几何平均值; 式(7)反映了大小球半径之和小于等于2倍的大球半径, 小球半径小于等于大小球半径的几何平均值。

下面论证该相似度定义满足两球相离、相交、相含时的边界条件。两球外切为两球相离、相交的边界条件, 此时 $V_{oi} = 0$, 式(4)中相离和相交下的计算结果相同, 因此边界条件成立; 两球内切为两球相交、相含的边界条件, 此时 $V_{oi} = 4\pi r_{\min}^3/3$, $r_1 + r_2 - d_{12} = 2r_{\min}$, 式(4)中相交和相含下的计算结果相同, 因此边界条件成立; $r_{\min} = 0$ 且 $d_{12} = 0$ 为两球相离、相含的边界条件, 此时式(4)中相交和相含下的计算结果均为0, 边界条件成立。综上, 该相似度的定义具有一致性。

考虑到本文的谱聚类算法将距离矩阵作为输入, 距离越小, 则2个负荷越相似, 因此定义 $d_s = 1/\varepsilon_{si}$, 将相似度转化为距离, 以此得到对称距离矩阵 $D_s \in \mathbb{R}^{n \times n}$ 。

4 谱聚类及聚类评价指标

4.1 基于k-最近邻法的谱聚类

谱聚类算法是一种基于谱图理论的聚类算法, 它将数据点映射为无向图, 将点与点之间的相似性表示为带权重的边, 从而将聚类问题转化为图划分问题^[10]。谱聚类的关键是相似度图的构造, 本文采用k-最近邻法来构造相似度图。k-最近邻法构造相似度图的思路为: 对于包含n个点的数据集, 令其中每个点 x_i 对应相似度图上的一个节点 v_i , 若 v_j 属于 v_i 的前k个最近邻节点集, 则称 v_j 是 v_i 的k近邻, 并连接 v_i 和 v_j , 其权重为距离矩阵 D_s 的第i行第j列元素。由于k近邻关系不对称(v_j 是 v_i 的k近邻, 而 v_i 未必是 v_j 的k近邻), 因此, 通过上述方法构造的图是有向的。通常有2种方法将k-最近邻法构造的相似度有向图转化为无向图: 只有2点互为对方的k近邻, 才将2点连接; 只要2点间有k近邻关系, 就将2点连接。本文采用第2种方法构造无向图。

基于上述k-最近邻法构造的相似度无向图, 谱聚类的流程^[7]如下。

1) 由相似度图构建邻接矩阵 W , 其第i行第j列元素 w_{ij} 为节点 v_i, v_j 间的权重, 若两节点间不连接, 则权重为0。

2) 由邻接矩阵 W 构建度矩阵 D_d , 其第i行第i列元素为和节点 v_i 相连的所有边的权重之和, 即:

$$D_d(i, i) = \sum_{j=1}^n w_{ij} \quad (8)$$

3) 计算拉普拉斯矩阵 $L_d = D_d - W$, 进而构建标准化后的拉普拉斯矩阵 $D_d^{-1/2} L_d D_d^{-1/2}$ 。

4) 将 $D_d^{-1/2} L_d D_d^{-1/2}$ 最小的k个特征值对应的特征向量 f 标准化, 形成 $n \times k$ 维的特征矩阵 F 。

5) 对特征矩阵 F 进行k-means聚类。

4.2 聚类评价指标

聚类评价指标是用来评价聚类质量的指标, 常用的聚类评价指标包括轮廓系数、CH 指标 CHI (Calinski-Harabasz Index) 和 DB 指标 DBI (Davies-Bouldin Index)^[17]。

1) 轮廓系数。

轮廓系数结合了聚类的凝聚度和分离度, 用于评估聚类的效果, 其值处于 $[-1, 1]$ 之间, 其值越大表示聚类效果越好。样本i的轮廓值 $s(i)$ 定义为:

$$s(i) = \begin{cases} \frac{b(i) - a(i)}{\max(a(i), b(i))} & |C_i| > 1 \\ 0 & |C_i| = 1 \end{cases} \quad (9)$$

式中: $a(i)$ 为样本i到所在类中其他样本的平均距离; $b(i)$ 为样本i到其他类中样本的平均距离的最小值; $|C_i|$ 为样本i所在类的样本数。整个数据集上轮廓值 $s(i)$ 的均值为轮廓系数 I_{sc} , 即:

$$I_{sc} = \frac{1}{n} \sum_{i=1}^n s(i) \quad (10)$$

2) CHI。

CHI 定义为分离度与紧密度的比值, 其值越大表示类自身越紧密, 类与类之间越分散, 即更优的聚类结果。CHI 的计算公式为:

$$I_{CHI} = \frac{S_B}{S_W} \frac{n - k_c}{k - 1} \quad (11)$$

式中: I_{CHI} 为 CHI; S_B 为类中各点与类中心的距离平方和; S_W 为各类中心点与数据集中心点的距离平方和; k_c 为指定的聚类数。

3) DBI。

DBI 综合考虑了类内紧密性和类间分散性, 其值越小表示聚类效果越好。DBI 的计算公式为:

$$I_{DBI} = \frac{1}{k_c} \sum_{i=1}^{k_c} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (12)$$

式中: I_{DBI} 为 DBI; S_i 为类i中样本点到其所属类中心的距离平方和; M_{ij} 为类i与类j的类中心间距离。

5 算例分析

5.1 算例构造

算例选用某地区2019年8月1日至2019年8月14日共2周的10 kV专变和公变共8 230条负荷曲

线。为验证所提方法的合理性与优越性,本文基于轮廓系数、CHI、DBI这3个指标,将本文方法与5种对比方法的聚类结果进行比较。5种对比方法分别为:基于欧氏距离的单日法(简称单日法),取一天的负荷数据,以欧氏距离衡量相似度并进行聚类^[13];基于欧氏距离的连续日法(简称连续日法),取多日负荷以欧氏距离衡量相似度并进行聚类;基于FDTW的连续日法(简称FDTW法),取多日负荷以FDTW距离衡量相似度并进行聚类;基于欧氏距离的均值方差法(简称均值方差法),提取负荷曲线均值和方差以形成新的2维数组,并采用欧氏距离衡量相似度以及进行聚类;基于欧氏距离的最小覆盖球法(简称覆盖球法),在第2节降维步骤中以欧氏距离代替FDTW计算日负荷曲线的距离矩阵,并将其代入最小覆盖球算法。

5.2 成本指标阈值的确定

图1中曲线1和曲线2与其他曲线间的差异极大,因此,本文所提方法利用式(3)所示的 I_{cost} 指标,在最小覆盖球的生成过程中剔除此类异常日负荷曲线。若 I_{cost} 阈值过小,则可能导致部分正常日负荷曲线被剔除;而若 I_{cost} 阈值过大,则可能导致部分异常日负荷曲线无法被剔除。因此,本文需要确定 I_{cost} 阈值。

为确定最优的 I_{cost} 阈值,本文取采样比例为 n_{ch} (采样比例为所选数据集中数据总数占总数据集中数据总数的比例,本文采用随机采样,此处取 $n_{\text{ch}}=20\%,40\%,\dots,100\%$)的算例数据,在不同 I_{cost} 阈值下,分别计算所提方法聚类结果的轮廓系数指标以评价聚类效果。轮廓系数越大,聚类效果越好,对应的 I_{cost} 阈值也越合理。不同情况下 I_{cost} 的取值如下。

1)理想情况下,假设多日负荷曲线足够多,且这些曲线均可在某一条日负荷曲线上叠加随机高斯扰动得到,此时不存在异常日负荷曲线,则3维空间中的各点将呈球状均匀分布,球体积增大的比例与球内点数量增加的比例一致,因此, I_{cost} 恒等于1。

2)实际情况下,多日负荷曲线中存在异常日负荷曲线,球内异常点变得稀疏,此时球体积增大的比例超过球内点数量增加的比例, I_{cost} 大于1。

综上, I_{cost} 阈值大于等于1才有剔除异常日负荷曲线的意义,因此,本节将 I_{cost} 阈值分别取为1.0、1.5、 $\dots$ 、3.0,并设置不剔除异常日负荷曲线的结果进行对比。表1为不同 I_{cost} 和 n_{ch} 取值下重复采用本文方法聚类10次得到的轮廓系数平均值。可以看出,不剔除异常日负荷曲线时轮廓系数指标最差,当 I_{cost} 阈值为2.0时轮廓系数指标较优,因此,设置 I_{cost} 阈值可有效提高聚类质量,本文取 I_{cost} 阈值为2.0。

图3为在 I_{cost} 阈值为2.0下图1所示负荷曲线对应的最小覆盖球。由图可见,除曲线1和曲线2外,

表1 不同 I_{cost} 和 n_{ch} 取值下的轮廓系数平均值
Table 1 Average value of silhouette coefficient under different values of I_{cost} and n_{ch}

$n_{\text{ch}}/\%$	轮廓系数平均值					
	$I_{\text{cost}}=1.0$	$I_{\text{cost}}=1.5$	$I_{\text{cost}}=2.0$	$I_{\text{cost}}=2.5$	$I_{\text{cost}}=3.0$	不剔除
20	0.884	0.652	0.853	0.741	0.811	0.552
40	1.011	0.995	1.112	0.851	0.774	0.681
60	0.625	0.662	0.758	0.590	0.601	0.556
80	0.854	0.961	0.992	1.001	0.883	0.701
100	0.781	0.742	0.901	0.821	0.772	0.622

其余曲线均在最小覆盖球内或球面上。该球的球心坐标为 $(-1.7759, -0.0625, 0.1279)$,半径为0.9826。曲线1、曲线2对应的点坐标分别为 $(1.4086, 1.4757, 1.5234)$ 、 $(1.3765, -3.7484, -2.8960)$,其与球心的欧氏距离分别为3.8019和5.7155,相离较远,因此2条曲线被剔除。可见, I_{cost} 阈值为2.0时可有效剔除异常日负荷曲线。

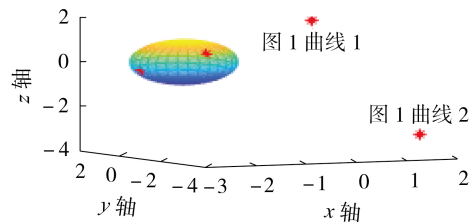


图3 图1所示负荷曲线对应的最小覆盖球
Fig.3 Minimum covering ball corresponding to load curves shown in Fig.1

5.3 负荷曲线-最小覆盖球-相似度间的一致性

为了说明2周负荷曲线、最小覆盖球、相似度间的一致性关系,选取附录A图A2所示的2周负荷曲线中的4条曲线进行分析。由图可看出:曲线1—3为典型的双峰型负荷曲线,它们之间的相似度较高,且曲线1和曲线3的相似度最高;曲线4为晚高峰型负荷曲线,与曲线1—3的相似度均较低。因此,这4条负荷曲线涵盖了极不相似、较相似、十分相似这3种情况。

附录A图A3为附录A图A2中4条负荷曲线对应的最小覆盖球。由图可以看出:曲线1—3对应的最小覆盖球之间均相交,且曲线1和曲线3对应的最小覆盖球相交程度最高;曲线4对应的最小覆盖球与曲线1—3对应的最小覆盖球之间均相离。

附录A表A1为附录A图A2中4条负荷曲线间的相似度 ε_{si} ,其中曲线自身之间的相似度显然为1。由表可以看出:曲线1—3间的 ε_{si} 均大于0,且曲线1和曲线3间的 ε_{si} 最高,为0.842;曲线4与曲线1—3间的 ε_{si} 均为0。

综上可知:2周负荷曲线间相似度高的,其对应的最小覆盖球间的相交程度较高,对应的相似度 ε_{si} 也较高;2周负荷曲线间相似度低的,其对应的最小

覆盖球间的相交程度较低(甚至可能相离),对应的相似度 ε_{si} 也较低。因此,本文方法中2周负荷曲线、最小覆盖球、相似度具有一致性,验证了本文方法的合理性。

5.4 算例结果的比较分析

为确定最佳聚类数,本文采用轮廓系数、DBI、CHI作为评价指标,将3个指标最优时对应的聚类数作为各方法的最佳聚类数。经过比较,各方法的最佳聚类数均为4。表2为不同方法的聚类性能比较。由表可见:均值方差法、覆盖球法及本文方法在3个指标上均有明显优势,这说明无论是单日负荷曲线还是多日负荷曲线聚类,直接聚类的效果均较差,而均值方差法、覆盖球法及本文方法这类通过提取负荷特征进行聚类的方法效果较好;本文方法比其他方法在轮廓系数和DBI上均有所改善,在CHI上仅小于覆盖球法,但是差距较小。这说明本文方法的聚类效果有较明显的优势。此外,由表中各方法的计算时间可见,本文方法耗时较长,具有一定劣势,本文方法的时间损耗主要由FDTW和最小覆盖球算法产生,由于本文研究是基于MATLAB平台展开的,而MATLAB对于FDTW和最小覆盖球算法的优化十分有限,且无法使用多核运算,因而大幅影响了效率。虽然本文方法的计算时间约为1 h,但是聚类结果的应用可以大幅缩短后续负荷预测、负荷建模的耗时,因此仍具有一定的实用性。

表2 6种方法的聚类性能比较

Table 2 Comparison of clustering performance among six methods

方法	I_{SC}	I_{CHI}	I_{DBI}	时间/s
单日法	0.5188	404.79	1.1679	1.743
连续日法	0.5021	243.87	1.2280	158.4
FDTW法	0.4331	193.28	1.3686	2575
均值方差法	0.5578	537.50	1.0465	14.38
覆盖球法	0.5829	679.72	0.9321	1292
本文方法	0.5897	648.10	0.9010	3874

图4为均值方差法、覆盖球法及本文方法的负荷曲线聚类中心。

图4(a)为均值方差法的负荷曲线聚类中心。其中,类1曲线为典型居民负荷曲线,类2曲线为典型商业负荷曲线,类3曲线为典型工业负荷曲线,类4曲线为非典型负荷曲线。由图可见,类4曲线近似为类1和类3曲线的平均曲线,曲线特点不明确。

图4(b)为覆盖球法的负荷曲线聚类中心,4类曲线含义与图4(a)中相同。由于各负荷曲线在聚类前均经过max-min归一化处理,因此,曲线数值分布在 $[0,1]$ 之间,而类4曲线的数值在各时刻均保持在0.5附近,这说明该类负荷曲线在各时刻的数值波动均较大,负荷曲线较杂乱,聚类结果不合理。

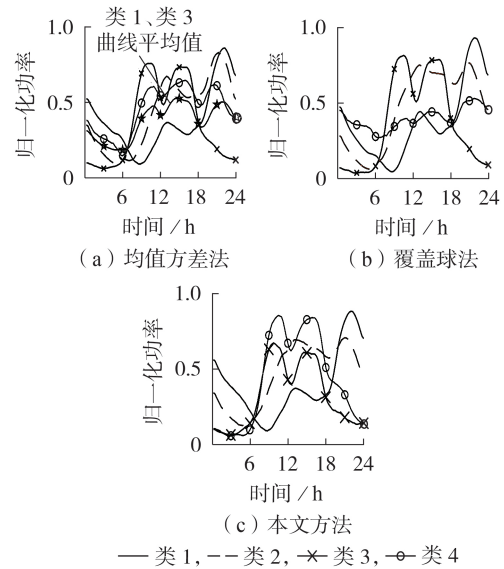


图4 3种方法的负荷曲线聚类中心

Fig.4 Clustering centers of load curves for three methods

表3 3种方法下各类负荷曲线数量占比

Table 3 Percentage of each type of load curve under three methods

方法	负荷曲线数量占比/%			
	类1	类2	类3	类4
均值方差法	24.8	24.2	31.5	19.5
覆盖球法	24.6	20.9	28.2	26.3
本文方法	31.5	31.4	18.4	18.7

表3为均值方差法、覆盖球法及本文方法下各类负荷曲线数量占比。由表可见,均值方差法和覆盖球法的类4曲线数量占比分别为19.5%和26.3%,数量较多,可排除采样异常的影响。因此,可判断这2类曲线的出现均是聚类算法的缺陷所导致的,其实际意义不明确。

图4(c)为本文方法的负荷曲线聚类中心。其中,类1曲线为典型的居民负荷曲线,类2曲线为典型的商业负荷曲线,类3和类4曲线为数值不同、数量大致相同的典型工业负荷曲线。由图可见,本文方法聚类结果中各类曲线的定位和特点明确,具有更好的实用性。

5.5 算例鲁棒性分析

为检验本文方法的鲁棒性,在上述算例的基础上叠加比例为 $\pm r$ ($r=5\%, 10\%, \dots, 25\%$)的均匀分布扰动,以模拟在实际用户负荷曲线采样过程中因随机因素造成的日内负荷波动。

为说明不同程度的扰动对聚类结果的影响,本文设置 I_m 指标描述不同干扰下各方法的聚类结果与无干扰下聚类结果的一致程度,其值越高则表示该方法的鲁棒性越高。 I_m 的计算公式为:

$$I_m = n_m / n_{all} \quad (13)$$

式中: n_m 为某方法在不同干扰下与在无干扰下聚类结果一致的样本数量; n_{all} 为该方法的样本总数。

图5为不同程度扰动下6种方法的 I_m 指标。由图可知:随着负荷曲线扰动程度的增加,各方法的 I_m 指标均呈下降趋势;随着扰动程度的增加,单日法、连续日法、FDTW法的 I_m 指标下降较明显,这说明这3种方法的鲁棒性明显不足;均值方差法、覆盖球法及本文方法由于考虑了多日负荷曲线并对其进行降维,因此,在各干扰下均有较高的 I_m 指标,且本文方法的 I_m 指标最高,这说明本文方法具有较高的鲁棒性。

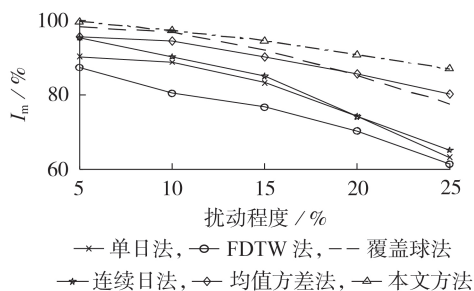


图5 不同干扰下的 I_m 指标

Fig.5 Index of I_m under different interferences

6 结论

本文针对目前负荷曲线聚类准确度和鲁棒性的不足,提出一种基于FDTW和最小覆盖球的多日负荷曲线聚类方法。利用FDTW和MDS将负荷曲线降至3维,并在3维空间中对每个负荷寻找最小覆盖球。在此基础上,定义最小覆盖球间的相似度,进而得到相似度矩阵。最后利用谱聚类算法得到聚类结果。所提方法具有以下特点:能同时考虑日负荷曲线内的时间滞后特性和多日负荷曲线间的波动特性;在最小覆盖球的迭代过程中自动剔除异常负荷。算例结果表明,该方法可有效提高聚类准确度及鲁棒性,具有较好的实用价值。

附录见本刊网络版(<http://www.epae.cn>)。

参考文献:

- [1] 王守相,陈建凯,王洪坤,等. 综合考虑电动汽车充电与储能及可中断负荷调度的配电网两阶段灵活性提升优化方法[J]. 电力自动化设备,2020,40(11):1-10.
WANG Shouxiang, CHEN Jiankai, WANG Hongkun, et al. Two-stage flexibility improvement optimization method of distribution network considering EV charging and scheduling of energy storage and interruptible loads[J]. Electric Power Automation Equipment, 2020, 40(11): 1-10.
- [2] 范志雄,李立新,王松岩,等. 人工智能技术在电网调控中的应用研究[J]. 电网技术,2020,44(2):401-411.
FAN Shixiong, LI Lixin, WANG Songyan, et al. Application analysis and exploration of artificial intelligence technology in power grid dispatch and control[J]. Power System Technology, 2020, 44(2): 401-411.
- [3] LIN S F, LI F X, TIAN E W, et al. Clustering load profiles for demand response applications[J]. IEEE Transactions on Smart Grid, 2019, 10(2): 1599-1607.
- [4] JIANG Z B, WU H, ZHU B Q, et al. A bottom-up method for probabilistic short-term load forecasting based on medium voltage load patterns[J]. IEEE Access, 2021, 9: 76551-76563.
- [5] SI C, XU S L, WAN C, et al. Electric load clustering in smart grid: methodologies, applications, and future trends[J]. Journal of Modern Power Systems and Clean Energy, 2021, 9(2): 237-252.
- [6] 张江林,张亚超,洪居华,等. 基于离散小波变换和模糊K-modes的负荷聚类算法[J]. 电力自动化设备,2019,39(2):100-106,122.
ZHANG Jianglin, ZHANG Yachao, HONG Juhua, et al. A load clustering algorithm based on discrete wavelet transform and fuzzy K-modes[J]. Electric Power Automation Equipment, 2019, 39(2): 100-106, 122.
- [7] 丁明,黄冯,邹佳芯,等. 改进谱聚类与遗传算法相结合的电力时序曲线聚类方法[J]. 电力自动化设备,2019,39(2):93-99,114.
DING Ming, HUANG Feng, ZOU Jiaxin, et al. Power time series curve clustering method combining improved spectral clustering and genetic algorithm[J]. Electric Power Automation Equipment, 2019, 39(2): 93-99, 114.
- [8] 陈焯,吴浩,史俊伟,等. 奇异值分解方法在日负荷曲线降维聚类分析中的应用[J]. 电力系统自动化,2018,42(3):105-111.
CHEN Ye, WU Hao, SHI Junyi, et al. Application of singular value decomposition algorithm to dimension-reduced clustering analysis of daily load profiles[J]. Automation of Electric Power Systems, 2018, 42(3): 105-111.
- [9] 张斌,庄池杰,胡军,等. 结合降维技术的电力负荷曲线集成聚类算法[J]. 中国电机工程学报,2015,35(15):3741-3749.
ZHANG Bin, ZHUANG Chijie, HU Jun, et al. Ensemble clustering algorithm combined with dimension reduction techniques for power load profiles[J]. Proceedings of the CSEE, 2015, 35(15): 3741-3749.
- [10] 徐胜蓝,司曹明哲,万灿,等. 考虑双尺度相似性的负荷曲线集成谱聚类算法[J]. 电力系统自动化,2020,44(22):152-160.
XU Shenglan, SI Caomingzhe, WAN Can, et al. Ensemble spectral clustering algorithm for load profiles considering dual-scale similarities[J]. Automation of Electric Power Systems, 2020, 44(22): 152-160.
- [11] 邵翔,郭谋发,游林旭. 基于改进DTW的接地故障波形互相关度聚类选线方法[J]. 电力自动化设备,2018,38(11):63-71,78.
SHAO Xiang, GUO Moufa, YOU Linxu. Faulty line selection method using mutual correlation cluster of grounding fault waveforms based on improved DTW method[J]. Electric Power Automation Equipment, 2018, 38(11): 63-71, 78.
- [12] 冯志颖,唐文虎,吴青华,等. 考虑负荷纵向随机性的用户用电行为聚类方法[J]. 电力自动化设备,2018,38(9):39-44,53.
FENG Zhiying, TANG Wenhua, WU Qinghua, et al. Users' consumption behavior clustering method considering longitudinal randomness of load[J]. Electric Power Automation Equipment, 2018, 38(9): 39-44, 53.
- [13] LI R, LI F R, SMITH N D. Multi-resolution load profile clustering for smart metering data[J]. IEEE Transactions on Power Systems, 2016, 31(6): 4473-4482.
- [14] WANG J H, ZHAO Y C. Time series K-nearest neighbors classifier based on fast dynamic time warping[C]//2021 IEEE International Conference on Artificial Intelligence and Computer Applications. Dalian, China: IEEE, 2021: 751-754.
- [15] KONG L C, QI C Q, QI H D. Classical multidimensional scaling: a subspace perspective, over-denoising, and outlier detec-

tion[J]. IEEE Transactions on Signal Processing, 2019, 67(14): 3842-3857.

- [16] 来疆亮,王守觉. 最小球覆盖几何算法及其在模式识别中的应用[J]. 模式识别与人工智能, 2006, 19(2): 271-276.

LAI Jiangliang, WANG Shoujue. An geometrical algorithm for minimum covering sphere and application in pattern recognition[J]. Pattern Recognition and Artificial Intelligence, 2006, 19(2): 271-276.

- [17] 孙玉芹,王松雷,黄冬梅,等. 考虑簇间重叠关系的负荷曲线多重聚类集成算法[J]. 电网技术, 2022, 46(5): 1982-1989.

SUN Yuqin, WANG Songlei, HUANG Dongmei, et al. Multiple clustering integration algorithm for load curves considering overlapping relationships between clusters [J]. Power System Technology, 2022, 46(5): 1982-1989.

作者简介:



刘晓峰

(E-mail: zjuwuhao@zju.edu.cn)。

刘晓峰(1998—),男,硕士研究生,主要研究方向为电力系统负荷大数据分析(E-mail: 22010016@zju.edu.cn);

康进(1998—),男,硕士研究生,主要研究方向为电力系统负荷大数据分析(E-mail: 22110115@zju.edu.cn);

吴浩(1973—),男,副教授,博士研究生导师,博士,主要研究方向为电力系统分析与控制以及负荷特性分析、建模及调控

(编辑 王锦秀)

Clustering method for multi-day load curves based on fast dynamic time warping and minimum covering sphere

LIU Xiaofeng¹, KANG Jin¹, MA Xiang², WO Jiandong², LÜ Leiyan², WU Hao¹

(1. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China;

2. Jinhua Power Supply Company of State Grid Zhejiang Electric Power Co., Ltd., Jinhua 321016, China)

Abstract: The current researches of load curve clustering are developed mainly based on single-day load curve or load distribution at the same time in multiple days, which ignore the inter-day load fluctuation characteristic and time lag characteristic of load curves, causing insufficient accuracy and robustness of clustering results. Comprehensively considering the fluctuation and time lag characteristics of loads, a clustering method of multi-day load curves is proposed combining fast dynamic time warping and minimum covering sphere. On the basis of considering the time lag characteristic of loads, fast dynamic time warping and multi-dimensional scaling method are used to reduce the dimension of load curves. The minimum covering sphere is iteratively found for each dimension reduced load, and the similarities between different covering spheres are calculated. The spectral clustering algorithm is used to obtain the similarity matrix. Case results show that the proposed method has certain advantages than the traditional method in terms of accuracy and robustness.

Key words: multi-day load curves; fast dynamic time warping; minimum covering sphere; spectral clustering; multi-dimensional scaling

附录 A

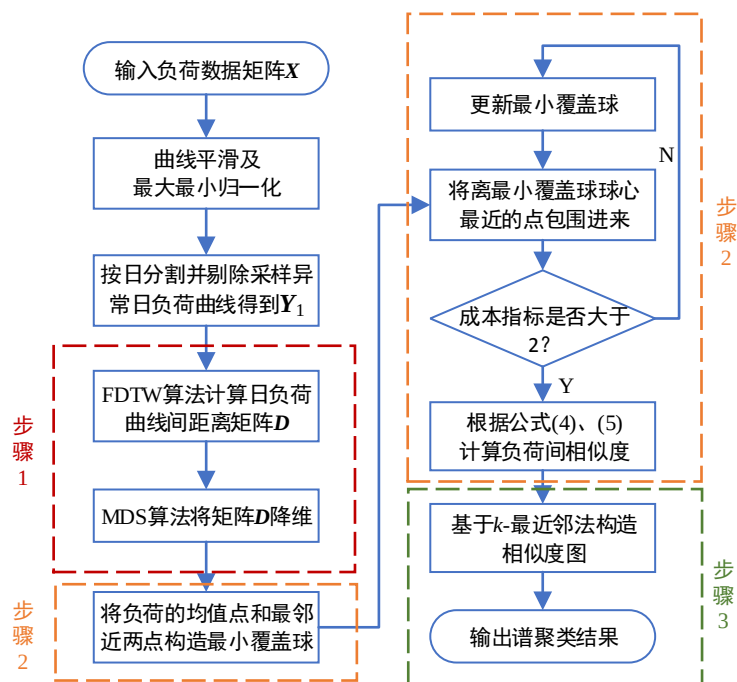


图 A1 所提方法流程图
Fig.A1 Flowchart of proposed method

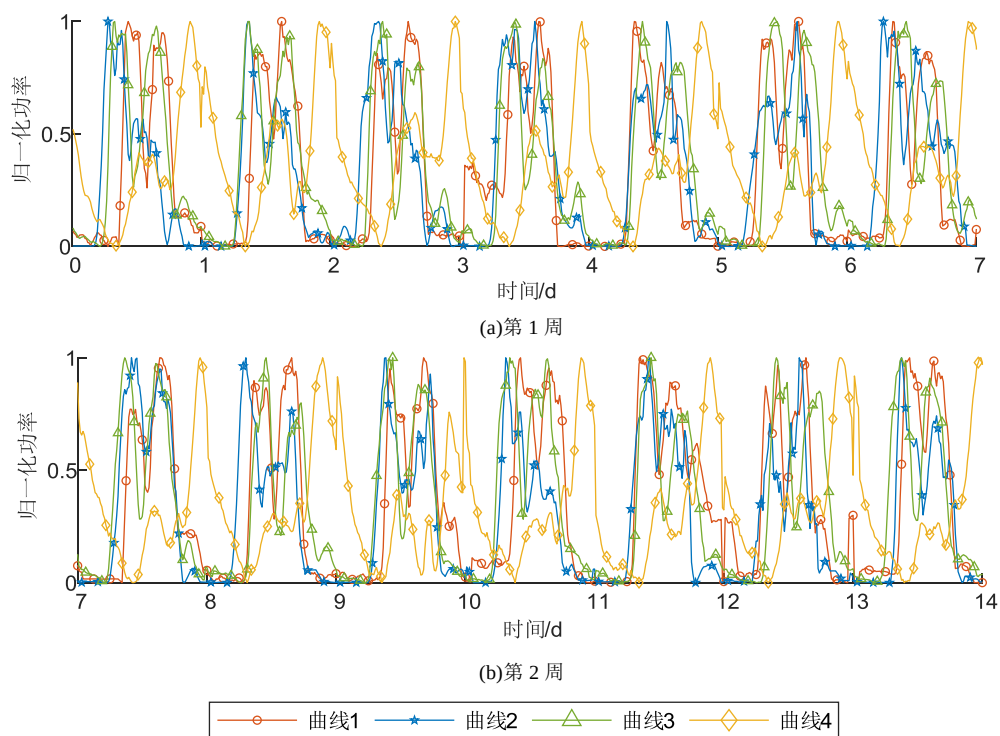


图 A2 4 条示例 2 周负荷曲线
Fig.A2 Four example load curves for two week

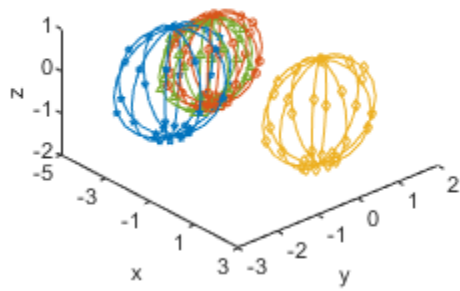


图 A3 4 条负荷曲线对应的最小覆盖球
Fig.A3 Minimum coverage ball of four load curves

表 A1 4 条负荷曲线间的相似度
Table A1 Similarity among four load curves

曲线	$\mathcal{E}_{\text{sl}}$			
	曲线 1	曲线 2	曲线 3	曲线 4
曲线 1	1	0.501	0.842	0
曲线 2	0.501	1	0.517	0
曲线 3	0.842	0.517	1	0
曲线 4	0	0	0	1