

基于奇异谱分析和改进密度峰值聚类算法的电力负荷曲线聚类方法

赵俊¹, 李鹏¹, 李文超¹, 苏适², 梁俊宇²

(1. 云南大学信息学院, 昆明 650091; 2. 云南电网有限责任公司电力科学研究院, 昆明 650217)

摘要: 针对现有的密度峰值聚类算法在数据局部波动较大或各类负荷数量分布不均匀导致的错分问题, 提出了一种基于奇异谱分析和改进密度峰值聚类算法的电力负荷曲线聚类方法。首先, 采用奇异谱分析将原始电力负荷数据分解为包含主要轮廓信息的低频分量和噪声成分的高频分量。然后, 融合K近邻和自然最近邻思想重新定义了密度峰值聚类算法的局部密度, 并对低频分量进行聚类。最后将其应用于真实的电力负荷数据集, 与其他聚类算法进行对比, 算例结果验证了所提方法在真实数据下的聚类有效性。

关键词: 电力负荷曲线聚类; 奇异谱分析; 密度峰值聚类; 智能电网; 数据驱动

Power Load Curve Clustering Method Based on Singular Spectrum Analysis and Improved Density Peak Clustering Algorithm

ZHAO Jun¹, LI PENG¹, LI Wenchao¹, SU Shi², LIANG Junyu²

(1. School of Information, Yunnan University, Kunming 650091, China; 2. Electric Power Research Institute, Yunnan Power Grid Co., Ltd., Kunming 650217, China)

Abstract: To address the misclassification issues in existing density peak clustering algorithms caused by significant local fluctuations in data or uneven distribution of load categories, a power load curve clustering method based on singular spectrum analysis and improved density peak clustering algorithm is proposed. Firstly, singular spectrum analysis is employed to decompose the original power load data into a low-frequency component containing the main contour information and a high-frequency component representing noise. Secondly, the local density of the density peak clustering algorithm is redefined by integrating the ideas of K-nearest neighbors and natural nearest neighbors, followed by clustering the low-frequency component. Finally, applied to real-world power load datasets, the proposed method is compared with other clustering algorithms, and case study results validate its effectiveness in real data.

Key words: power load curve clustering; singular spectrum analysis; density peak clustering; smart grid; data driven

0 引言

随着主动配电网技术信息化和智能化的迅猛发

展, 电力行业正迎来前所未有的变革。基于智能电表采集的海量的电力大数据作为这一变革的重要推动力量正在深刻地改变电力系统的运行和管理模式^[1-2]。针对用户侧的电力负荷特性分析对智能电网系统规划运行和配电网管理具有重要意义, 能够帮助零售商和配电网运营商提高对消费者用电行为的理解^[3], 由此进行电力用户的社会属性辨识, 更好地实施能源政策和基础设施规划策略^[4-5]。如何高效地以数据驱动的方式对电力负荷曲线进行聚类分析是分析电力负荷特性的重要问题^[6]。

目前, 各种聚类方法已被应用到电力负荷曲线

基金项目: 国家自然科学基金资助项目(62163036); 云南省科技计划重大专项(202302AF080006); 云南省教育厅科学研究基金资助项目(2025Y0152); 云南大学第四届专业学位研究生实践创新项目(ZC-24248916)。

Foundation item: Supported by the National Natural Science Foundation of China (62163036); the Yunnan Science and Technology Major Program (202302AF080006); the Scientific Research Fund of Yunnan Education Department (2025Y0152); the Fourth Practical Innovation Project of Postgraduate Students in the Professional Degree of Yunnan University (ZC-24248916).

聚类中^[7],其中以无监督学习算法为主,主要包括高斯混合模型(Gaussian mixture model, GMM)、K-medoids、和基于密度的带噪声应用空间聚类(density-based spatial clustering of applications with noise, DBSCAN)算法等。然而, GMM存在聚类数目选取敏感、收敛速度慢等缺陷,基于密度的DBSCAN存在参数敏感的问题^[8]。2014年,文献[9]提出了一种快速搜索密度峰值聚类算法(cluster-ing by fast search and find of density peaks, DPC),该算法超参数量较少、对参数取值较不敏感,可实现任意形状类簇的聚类,相较于其他聚类算法具有一定优势,已被广泛应用于交通流量预测、图像识别等领域^[10-11]。近年来,多名学者针对上述聚类方法应用于电力负荷曲线聚类所存在的缺陷进行了改进。文献[12]引入K-means确定初始聚类中心以提高GMM算法收敛的计算速度及稳定性,实现了负荷曲线的高效聚类。文献[13]采用基于正态云模型的相似性度量和改进的密度峰值聚类算法实现了海量典型负荷曲线的聚类。文献[14]提出了一种基于粒计算结合欧式距离与皮尔森相关系数作为相似性度量的聚类算法,克服了传统聚类算法数值和形态单一的局限性。文献[15]将负荷数据看作非平稳信号应用变分模态分解与领域保留嵌入算法实现了负荷数据的多尺度精细化聚类。文献[16]采用线性插值放大多个时间段密度、利用自适应分段聚合近似(adaptive piecewise aggregate approximation, APAA)减小冗余密度,改善了K-medoids的负荷曲线聚类效果。

随着新型配电网的建设深入及用户侧的精细化管理,用电负荷数据在多个时段内呈现出了更强波动性,不同负荷类型的用户数量存在较大的差异,而现有的一些聚类算法并未充分考虑上述问题,由此造成样本的错误分配。鉴于此,本文提出了一种基于奇异谱分析和改进密度峰值算法的负荷曲线聚类方法。针对负荷数据的波动性,利用奇异谱分析将日负荷曲线分为包含主要轮廓特征的低频分量和包含细节与噪声的高频分量,波动信息主要包含在高频分量中;针对数量差异较大的各类负荷曲线形态,重新定义了密度峰值聚类算法中表征各样本疏密关系的局部密度估计,避免较小数量类型样本的错误分配。算例结果表明本文所提出的聚类方法在真实数据下具有更好的聚类效果,鲁棒性更强。

1 基本原理及问题分析

1.1 密度峰值聚类算法概述

DPC算法基于两个基本假设:1)每个聚类中心的局部密度高于非聚类中心的密度,且相互之间距离较大;2)聚类中心被低于自身密度的局部领域数据点包围。实现过程如下。

步骤1:计算每个样本点 x_i 的局部密度 ρ_i ,通常有截断核和高斯核两种计算方式,截断核的计算方式为:

$$\left\{ \begin{array}{l} \rho_i = \sum_{j=1}^n \chi(d_{ij} - d_c) \\ \chi(w) = \begin{cases} 1, & w \leq 0 \\ 0, & w > 0 \end{cases} \end{array} \right. \quad (1)$$

高斯核的计算方式为:

$$\rho_i = \sum_{j \neq i} \exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \quad (2)$$

式中: n 为样本点数量; d_{ij} 为样本点 x_i 与样本点 x_j 之间的距离; d_c 为截断距离; $\chi(w)$ 为指示函数; w 为两个样本点间的距离,即 $d_{ij} - d_c$,若两点距离小于截断距离则计入局部密度,否则不参与密度累加。 d_c 值过大将导致不同簇的合并,其值过小将导致同一簇的分裂,通常选取为总体样本数目的1%~2%。高斯核的计算方式适用于较小规模的数据集,以降低 d_c 取值对聚类结果的影响^[17]。

步骤2:计算样本间相对距离 δ_i ,若 x_i 为其领域内密度最高的点,则

$$\delta_i = \max_j (d_{ij}) \quad (3)$$

否则其相对距离为:

$$\delta_i = \max_{\rho_i > \rho_j} (d_{ij}) \quad (4)$$

式(4)表示局部密度高于 x_i 的点中与 x_i 相距最近点的距离。

步骤3:选取同时拥有较高的 ρ_i 和 δ_i 的样本点作为密度峰值点,即聚类的类簇中心点,以 γ_i 值确定:

$$\gamma_i = \rho_i \delta_i \quad (5)$$

类簇中心点的 γ_i 值应显著大于其他样点。

步骤4:对剩余样本点进行分配,按照局部密度下降的顺序,将剩余样本依次分配到密度更高且距离最小的样本点所属类簇。

1.2 问题分析

应用于日负荷曲线聚类的 DPC 算法存在下列缺陷。

1) 负荷数据的总体变化趋势取决于用户的生产和生活规律,同时也受到外界环境的影响,会在多个时段内产生波动,造成负荷曲线的局部畸变。当数据点存在较大波动时,各负荷曲线与正确类簇中心曲线的距离较大,导致负荷曲线被分配到错误的类簇中,从而影响聚类结果。图 1 为两类不同用电模式用户的电力负荷日曲线集合,红色曲线为某小型办公楼用户的负荷曲线集合,曲线形态为单峰型;蓝色曲线为某餐厅用户的负荷曲线集合,曲线形态为三峰型。黑色曲线为小型办公楼用户某一日的日负荷曲线,受天气因素影响,其负荷数据在其峰值期间存在一定程度的波动,使用 DPC 聚类时该曲线会被错误地归为餐厅用户所对应的三峰型曲线形态的用电模式。

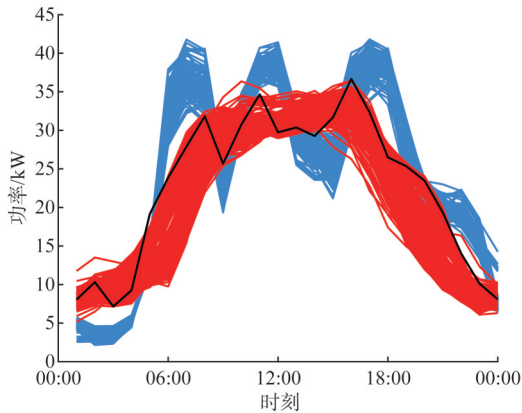


图 1 两类用户多日负荷曲线

Fig. 1 Multi-day load profiles for two types of customers

2) 不同类型用电模式的用户数量差异较大,其总体分布不均衡,对日负荷曲线聚类时类簇之间密度差异过大,使用固定的截断距离所计算出的局部密度难以表征样本点之间的疏密情况,导致样本点分配错误,影响后续一连串样本点的分配。

t 分布随机近邻嵌入 (t-distributed stochastic neighbor embedding, t-SNE) 算法是一种基于概率分布的非线性降维方法,算法通过保持高维数据点与低维空间中模拟点的相似性概率分布,实现数据在低维空间的非线性映射^[18]。图 2 为采用 DPC 算法的日负荷曲线通过 t-SNE 算法转化为二维数据的可视化结果。图中,相同颜色的点被归为同一类,坐

标轴的标度值不具备物理量纲意义,二维空间中的距离表示点与点间的相对距离,距离越小,负荷曲线越相似。蓝色、绿色与黄色样本点类簇分别为工厂、办公楼与超市用户的日负荷曲线集合,其数量均为 800 余条。红色样本点类簇为 200 余条中小学用户的日负荷曲线集合。由于数量上存在差异,且类簇距离较近,黄色、绿色类簇均与红色类簇产生了一定程度的混叠。在更大规模的数据集中,该问题会被进一步放大,导致更多样本点的分配错误。

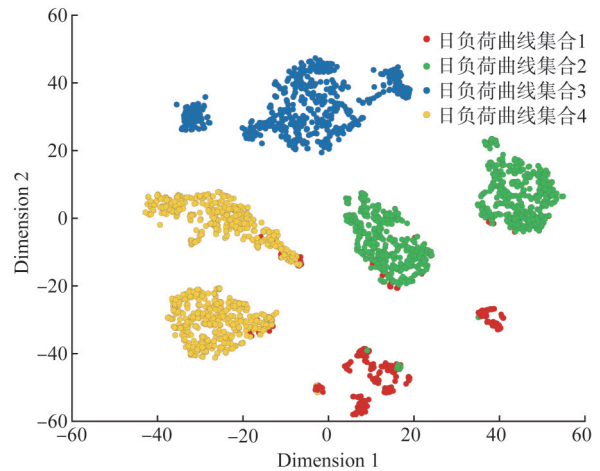


图 2 DPC 算法聚类结果可视化

Fig. 2 Visualization of the clustering results of the DPC algorithm

2 基于奇异谱分析和改进密度峰值算法的负荷聚类方法

针对 1.2 节 DPC 算法应用于负荷曲线聚类存在的缺陷,本文提出了一种基于奇异谱分析和改进密度峰值算法 (singular spectrum analysis-improved clustering density peaks, SSA-iDPC) 的负荷曲线聚类方法。

2.1 基于奇异谱分析的负荷分解

负荷数据主要取决于用户的生产和生活规律,同时还受到天气等不确定因素影响,既具有规律性又具有随机性,有效地提取其中规律性的周期趋势是提升聚类效果的一个重要步骤。小波分析对白噪声敏感性较强、经典模态分解容易受到中断影响,该缺陷会对负荷数据的周期趋势提取造成较大影响^[19]。奇异谱分析 (singular spectrum analysis) 是一种非参数的信号分解技术,能够有效地将时间序列分解为少数可解释的、独立的成分,适用于非线

性、非平稳的信号处理分析^[20],主要包括分解和重构两个部分,具体过程如下。

步骤1:对于给定长度为 N 的负荷数据 $X_N = [x_1, x_2, \dots, x_N]$,构造 $L \times H$ 阶轨迹矩阵 Y :

$$Y = \{y_{ij}\}_{i,j=1}^{L,H} = \begin{bmatrix} x_1 & x_2 & \cdots & x_H \\ x_2 & x_3 & \cdots & x_{H+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{bmatrix} \quad (6)$$

式中: H 为嵌入维度; $L (L \leq N/2)$ 为窗口长度,满足 $L = N - H + 1$ 。

步骤2:对轨迹矩阵 Y 进行奇异值分解(singular value decomposition, SVD)。求解协方差矩阵 YY^T 的 L 个特征值 $\lambda_1, \lambda_2, \dots, \lambda_L$ 以及对应的特征向量 $U_1, U_2, \dots, U_L$,令 n 为轨迹矩阵 Y 的秩,定义 $V_i = YU_i/\sqrt{\lambda_i}$,得到:

$$Y = Y_1 + Y_2 + \cdots + Y_i + \cdots + Y_n \quad (7)$$

式中 $Y_i = \sqrt{\lambda_i} U_i V_i^T$ 。

步骤3:计算各个特征值在所有特征值之和的占比,计为每个特征值的贡献率 G_i ,表示每个分量与原始负荷数据的相关性。计算公式如下:

$$G_i = \frac{\lambda_i}{\sum_{i=1}^L \lambda_i}, i = 1, 2, \dots, L \quad (8)$$

步骤4:对每一个子矩阵 Y_i 进行对角平均化,将每一个分量还原成长度为 N 的序列,对于轨迹矩阵 Y , $L^* = \min(L, H)$, $K^* = \max(L, H)$,当 $L < H$ 时, $y_{ij}^* = y_{ij}$;当 $L \geq H$ 时, $y_{ij}^* = y_{ji}$ 。通过式(9)将各个子矩阵 Y_i 转换为其对应的子序列 $[y_{r1}, y_{r2}, \dots, y_{rk}, \dots, y_{rN}]$ 。

$$y_{rk} = \begin{cases} \frac{1}{k} \sum_{a=1}^k y_{a,k-a+1}^*, & 1 \leq k < K^* \\ \frac{1}{K^*} \sum_{a=1}^{K^*} y_{a,k-a+1}^*, & K^* \leq k \leq L^* \\ \frac{1}{N-k+1} \sum_{a=k-L^*+1}^{N-L^*+1} y_{a,k-a+1}^*, & L^* < k \leq N \end{cases} \quad (9)$$

按照各子矩阵 Y_i 对应的特征值贡献率 G_i 的高低将子序列叠加为多个频率的分量。其中,低频分量包含原始负荷数据的主要信息,高频分量包含原始负荷数据的细节与噪声信息^[21]。

图3是某用户连续5日的日负荷曲线的分解结果,其负荷曲线形态大致为单峰型。第3日和第4日的负荷数据均存在一定程度的波动,经过SSA分

解后的低频分量能够有效地剔除波动,保留负荷曲线的主要轮廓特征及周期趋势;高频分量则包含了原始负荷曲线的细节内容与噪声分量,其在第3和第4日分量的振幅和频率显著大于其他几天。

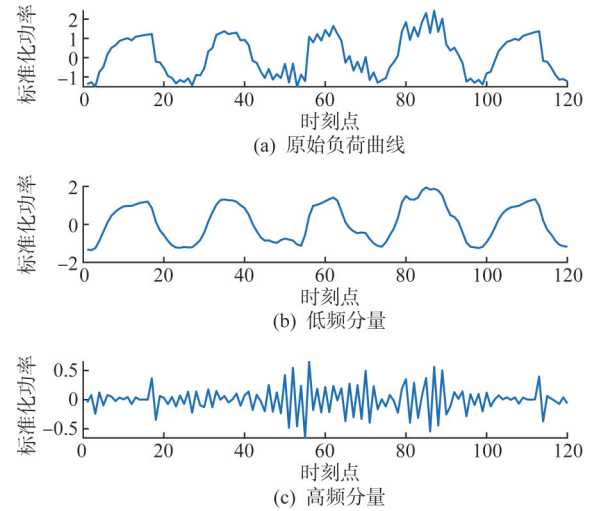


图3 负荷曲线SSA分解结果

Fig. 3 SSA decomposition results of load curves

2.2 改进的密度峰值聚类算法

局部密度是DPC聚类算法的重要部分,传统的DPC算法通过固定的截断距离计算。该算法应用于负荷数据聚类时未充分考虑数据内部的结构差异,在数据分布不均衡时导致类簇之间密度差异过大,无法较好地反映样本点间的疏密情况和非线性特征,从而导致类簇中心距离过近。基于K近邻和自然最近邻的思想,将每个样本点 x_i 的局部密度 ρ_i 重新定义为:

$$\rho'_i = \sum_{x_j \in KNN(x_i)} NNN(i, j) \times \exp\left(-\frac{d_{ij}^2}{d_{ik}^2}\right) \quad (10)$$

式中: $KNN(x_i)$ 为与样本点 x_i 的欧氏距离最小的 K 个样本点所构成的集合; K 近邻数量参考DPC截断距离为总样本点数的1%~2%^[9, 22]; d_{ij} 为样本点 x_i 与样本点 x_j 之间的欧氏距离; d_{ik} 为样本点 x_i 的与其 K 近邻中最远的样本点间的距离; $NNN(i, j)$ 为样本点的自然最近邻(natural nearest neighbors, NNN),其计算方式如式(11)所示。

$$NNN(i, j) = \begin{cases} 1, & i \in KNN(j) \cup j \in KNN(i) \\ 0, & \text{其他} \end{cases} \quad (11)$$

当样本点 x_i 与 x_j 互为 K 近邻领域内的样本点时,两样本点为一对自然最近邻点。改进后的局部密度

估计通过对样本点的 K 个近邻样本进行加权求和,并考虑到样本点间的欧式聚类与样本所处类簇的疏密程度,能够根据样本点的自然最近邻样本点数量对局部密度估计进行调整,避免了样本点因稀疏程度较大时局部密度估计无法准确表征其相互关系而造成的错误分配问题。图 4 给出了与图 2 相同的日负荷曲线采用改进后的局部密度估计的 iDPC (improved density peak clustering) 算法的聚类结果通过 t-SNE 算法在二维空间的分布。对比图 2, iDPC 算法有效解决了红色类簇与黄色、绿色类簇间样本点混叠的问题。

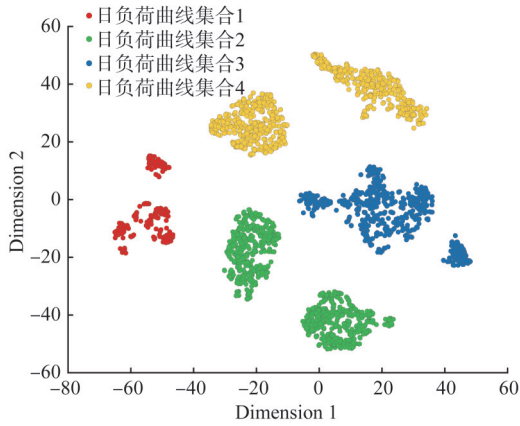


图 4 iDPC 算法聚类结果可视化

Fig. 4 Visualization of the clustering results of the iDPC algorithm

2.3 基于奇异谱分析和改进密度峰值算法的电力负荷聚类方法流程

本文提出的 SSA-iDPC 算法整体流程如图 5 所示,主要分为负荷分解过程和负荷聚类过程。

负荷分解过程如下。

步骤 1: 对输入负荷数据进行数据预处理。这一步骤主要包括数据清洗和数据标准化处理。数据清洗将缺失或异常数据达到总量的 10% 及以上负荷曲线进行剔除,避免异常值对聚类结果造成影响。为避免幅度差异对相似性度量的影响,提高数据的可比性,根据式 (12) 对负荷数据进行 z-score 标准化,将其转换为均值为 0、标准差为 1 的分布。

$$x' = \frac{x - \mu}{\sigma} \quad (12)$$

式中: x' 为 z-score 标准化后的数据; μ 为序列 x 的均值; σ 为序列 x 的标准差。

步骤 2: 对预处理后的数据进行 SSA 信号分

解。将数据集中连续 n 天采样频率为 d 的日负荷数据 $X^{n \times d}$ 转换为一维的时间序列 $X^{1 \times n \times d}$ 进行 SSA 分解。

对于 SSA 分解,需要确定其嵌入维度 H ,较大的 H 值对于长周期信号分解效果较好^[23],而电力负荷数据是以 24 h 进行分隔的较短周期的信号,标准化的负荷曲线平滑掉过多的细节信息导致曲线间轮廓差异过小,故本文将 H 设置为 2,负荷曲线通过 SSA 算法分解为高频分量 X_H 和低频分量 X_L 。

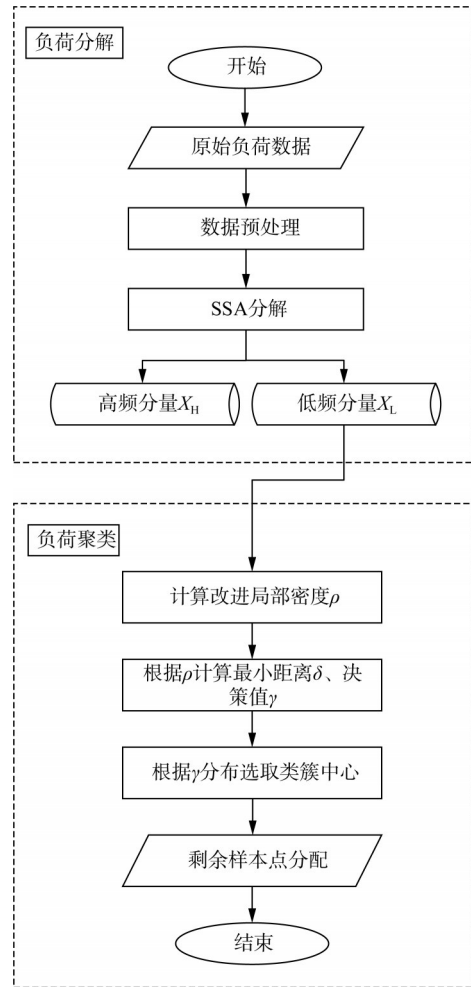


图 5 SSA-iDPC 负荷聚类方法流程图

Fig. 5 Flowchart of SSA-iDPC load clustering method

负荷聚类过程如下。

步骤 1: 计算各样本的改进局部密度 ρ'_i 。选取分解后包含负荷曲线主要轮廓信息的低频分量,根据式 (10) — (11) 计算每个样本点的改进局部密度 ρ'_i 。

步骤 2: 根据每个样本点的 ρ'_i , 计算其对应的最小距离 δ_i 和决策值 γ_i 。

步骤3: 对决策值 γ_i 按照大小排序, 选取类簇中心点。类簇中心点的 γ_i 值应该显著大于其他非类簇中心点。

步骤4: 对剩余样本点进行类簇的分配, 将其分配到与其距离最小且局部密度 ρ'_i 比它大的类簇中。

3 算例分析

本文选取美国能源部开放式能源信息网站(open energy information, OpenEI)公布的商业用户负荷数据和我国云南省某市的工商业负荷数据作为案例。OpenEI数据集中每个用户采样间隔为1 h, 每天采样24点数据, 包含美国多个地区16个行业共计7 200条日负荷曲线; 工商业负荷数据包含超市、商铺和酒店等商业用户和钢铁、冶炼等工业用户在2018年11月至2019年1月的4 000条日负荷曲线, 采样间隔为30 min。

3.1 聚类有效性评价指标

聚类效果的评价指标分为外部和内部的评价指标^[24-26], 外部评价指标需要聚类数据的外部信息来衡量聚类的有效性, 内部评价指标通过考察簇和样本点间的关系来衡量聚类的有效性。对负荷曲线聚类时往往会缺少数据的外部信息, 故本文选取内部评价指标对负荷聚类的有效性进行衡量。本文选取常用的内部评价指标^[27-29]误差平方和(sum of squares of errors, SSE, 其值用 S_{SE} 表示)、DBI指数(Davies Bouldin Index, DBI, 其值用 D_{BI} 表示)、轮廓系数(silhouette coefficient, SC, 其值用 S_c 表示)进行聚类结果的有效性衡量。若负荷曲线集被划分为 K 个类簇, 3个聚类指标计算公式为:

$$S_{SE} = \sum_{i=1}^K \sum_{x_i \in C_i} (x_i^* - x_{ic})^2 \quad (13)$$

式中: x_i^* 为 C_i 簇的聚类中心; x_{ic} 为 C_i 簇内的非中心样本点。SSE指标反应了各个簇的凝聚程度, 其值越小各类簇越紧凑, 聚类效果越好。

$$D_{BI} = \frac{1}{K} \sum_{i=1}^K \max_{i \neq j} \left(\frac{\bar{S}_i + \bar{S}_j}{\|x_i^* - x_j^*\|_2} \right) \quad (14)$$

式中: $\bar{S}_i$ 为第 i 簇中各样本点到聚类中心的平均距离; $\|x_i^* - x_j^*\|_2$ 为第 i 个聚类中心到第 j 个聚类中心的距离。DBI综合考虑了类内样本相似度和类间样本差异度, 其值越小表明聚类效果越好。

$$S_c = \frac{1}{N} \sum_{i=1}^N \frac{b(x_i) - a(x_i)}{\max \{a(x_i), b(x_i)\}} \quad (15)$$

式中: $b(x_i)$ 为样本点 x_i 到其他簇内样本点的最小平均距离; $a(x_i)$ 为 x_i 到所属簇内其他样本点的平均距离。SC的取值范围在-1~1之间, 值越大表明簇与簇之间越疏远, 簇内差异越小, 聚类效果越好。

3.2 聚类效果对比分析

为验证本文提出的SSA-iDPC算法的聚类有效性, 本文设置SSA-DPC、VMD-iDPC、iDPC、DPC、GMM和K-medoids 6种算法作为对比实验。其中SSA分解的嵌入维度 H 取2, VMD-iDPC同样取VMD分解得到的低频分量进行实验。

3.2.1 OpenEI数据集聚类分析

将各算法应用于OpenEI的公开数据集后, 基于决策图分析, SSA-iDPC(图6)选定了6个聚类中心, 而DPC(图7)算法则选定5个聚类中心。VMD-iDPC和iDPC采用改进后的密度估计同样划分出6个聚类中心, SSA-DPC划分出5个聚类中心。为确保实验的有效性, K-medoids和GMM均选取评价指标最佳聚类中心数6。

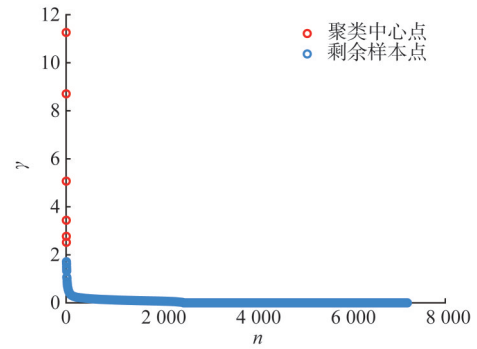


图6 SSA-iDPC算法对OpenEI数据集的聚类决策图

Fig. 6 Decision graph of SSA-iDPC algorithm in OpenEI dataset

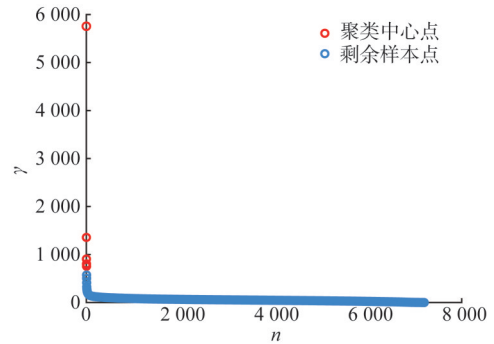


图7 DPC算法对OpenEI数据集的聚类决策图

Fig. 7 Decision graph of DPC algorithm in OpenEI dataset

SSA-iDPC算法的聚类结果如图8所示,从聚类结果可以看出,聚类结果基本准确,不同簇间形态差异较大,负荷曲线大致被分为4类,包括单峰、双峰、三峰和避峰4种曲线形态。第3类和第4类为单峰形态,第3类峰值期略长于第4类,大致在09时—17时,该类别包括医院、中学、办公楼等主要工作集中在白天且持续时间较长的场所。第4类仅包含小学,峰值期大致在09时—16时。第1类和第6类为双峰形态,其中第1类为酒店类,早晚高峰峰值接近。第6类为公寓类,晚高峰峰值高于早高峰。第5类为仅包含餐厅类的三峰形态,主要表现为在07时、11时30分和17时30分附近存在3个峰值幅度接近的3个用电高峰。第2类为避峰形态,主要包含仓库类,其用电峰值期均在夜间。

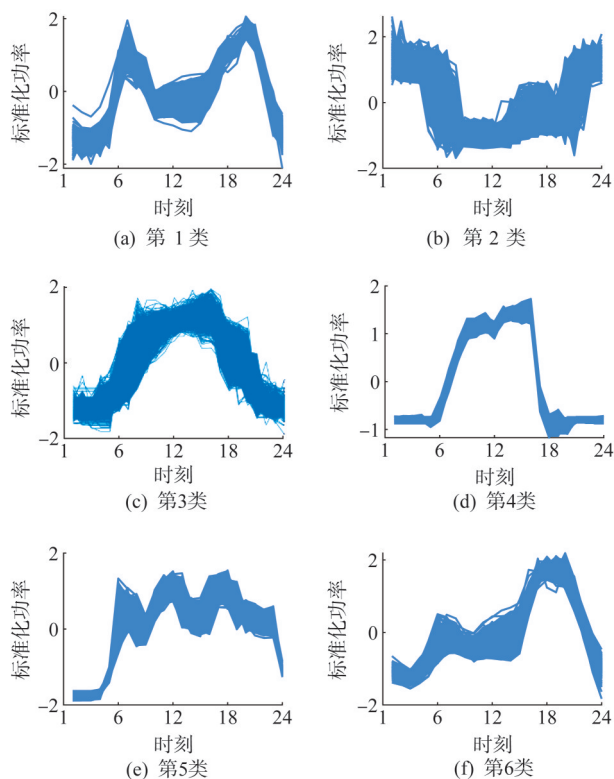


图8 SSA-iDPC算法在OpenEI数据集的聚类结果

Fig. 8 Clustering results of SSA-iDPC algorithm in OpenEI dataset

DPC算法的聚类结果如图9所示,在DPC的聚类结果中,两类双峰型的负荷曲线在划分时被混叠到了同一类中,第3类和第4类两类单峰形态负荷曲线也存在划分的混叠,且二者总体轮廓差异较小,整体聚类效果与图8相比较差,证明了本文算

法的样本分配策略优于DPC算法。K-medoids和GMM算法的聚类结果如图10—11所示。在K-medoids的聚类结果中,同一类避峰形态的负荷曲线被划分为了两类,其整体轮廓差异较小,单峰曲线被统一归为一类。在GMM的聚类结果中,第3类和第4类单峰形态曲线差异较小,第6类曲线混叠了第5类三峰形态曲线。综合上述聚类结果分析,SSA-iDPC的聚类效果最佳,能够将不同形态的曲线区分开。

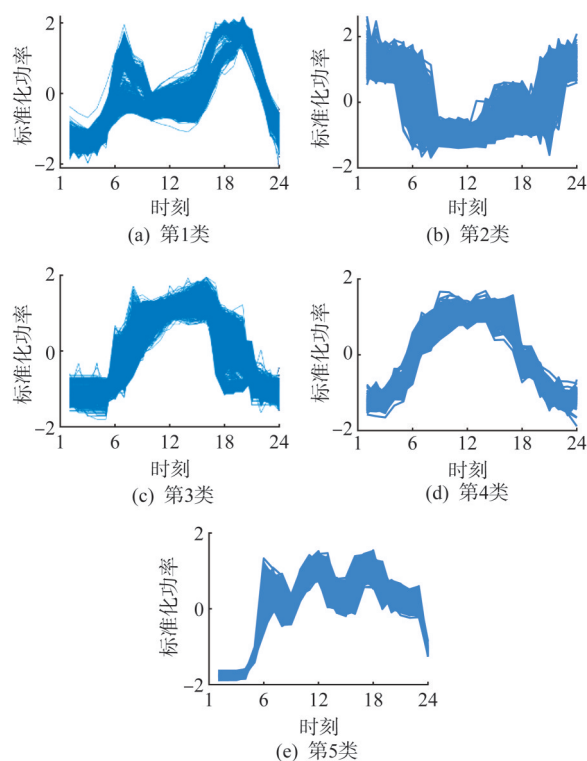


图9 DPC算法在OpenEI数据集的聚类结果

Fig. 9 Clustering results of DPC algorithm in OpenEI dataset

为客观评价7种方法对负荷曲线的聚类效果,表1给出了3.1节中的3个内部聚类评价指标对比。从聚类效果的指标来看,SSA-iDPC的SSE和DBI指标最小,表明其聚类结果的类簇的凝聚程度最高,簇内最为紧密。SSA-iDPC的SC最大,表明其各个类簇间的稀疏程度最大。因此,SSA-iDPC在7种聚类方法中能够较好地划分出正确的类簇,簇间距离最大且簇内距离最小,聚类效果最好。SSA-DPC和iDPC算法相较于原始DPC算法都有着一定程度的提升,通过聚类指标验证了奇异谱分析与改进的局部密度估计对提升聚类效果的有效性。

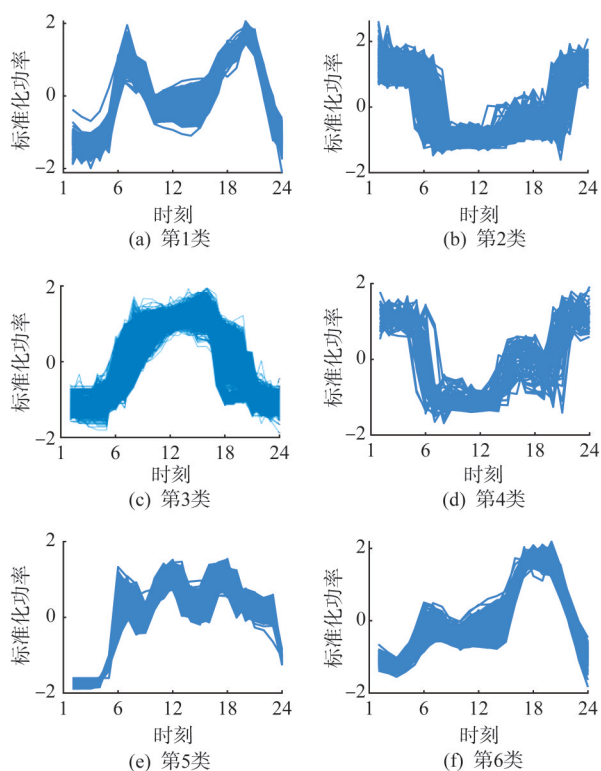


图10 K-medoids算法在OpenEI数据集的聚类结果

Fig. 10 Clustering results of K-medoids algorithm in OpenEI dataset

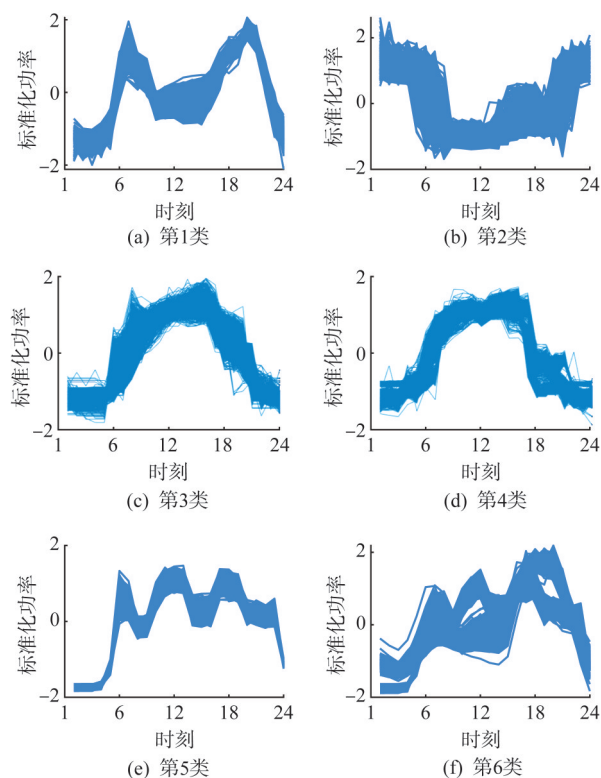


图11 GMM算法在OpenEI数据集的聚类结果

Fig. 11 Clustering results of GMM algorithm in OpenEI dataset

表1 不同算法在OpenEI数据集的聚类效果评价指标对比

Tab. 1 Comparison of clustering effect evaluation indexes of different algorithms in OpenEI dataset

聚类评价指标	最佳类簇数	SSE	DBI	SC
SSA-iDPC	6	8 404.8	0.595 3	0.504 2
VMD-iDPC	6	9 904.0	1.134 9	0.450 0
iDPC	6	8 440.9	0.746 3	0.449 5
DPC	5	9 240.9	0.832 7	0.394 1
SSA-DPC	5	9 240.5	0.826 6	0.409 0
K-medoids	6	10 028.6	1.104 8	0.470 7
GMM	6	9 802.7	1.052 7	0.388 7

3.2.2 工商业数据集聚类分析

将各算法应用于工商业数据集后,各方法均在6个聚类中心数下取得最佳效果。SSA-iDPC算法的聚类结果如图12所示,各类负荷曲线形态差异明显。第1类负荷曲线主要代表夜间工作的工厂,其用电高峰通常出现在21时左右,并持续到次日11时。第2类负荷曲线包含酒店类负荷,其用电峰值分别出现在12时—21时。第3类负荷曲线为主要用电时段为12时—23时的服务业商铺。第4类负荷曲线为超市类,通常在08时—19时期间存在较高的电力负荷。第5类负荷曲线包括小型便利店和小型工厂,主要表现为全天都存在较低的电力负荷。第6类负荷曲线为早、中、晚3个时段均有明显的用电高峰的工厂,在12时和18时左右,由于工人就餐,电力需求有所下降。

DPC算法的聚类结果如图13所示。其中第2、第3类负荷曲线整体形态差异较小,第1类和第4类应属于同一类的负荷曲线被错分为两类,与图12的聚类结果相比较差。K-medoids和GMM的聚类结果如图14—15所示。观察图14,第2类和第6类同属一种类型的负荷曲线被错分为了2类,第4类和第5类同属一种类型的负荷曲线同样被错分为了2类,该算法聚类效果较差,存在较大程度多类负荷曲线的混叠与同类负荷曲线的错分。观察图15,第3类和第5类负荷曲线整体形态差异较小,也出现了同类负荷曲线的错分。综合上述聚类效果,本文算法在复杂的实际应用场景下具有更强的适应性与聚类有效性。

表2给出了各聚类算法在该工商业数据集下的3个内部聚类效果指标对比。根据表2,本文算法的SSE、DBI和SC分别达到9 648.1、1.087 5和

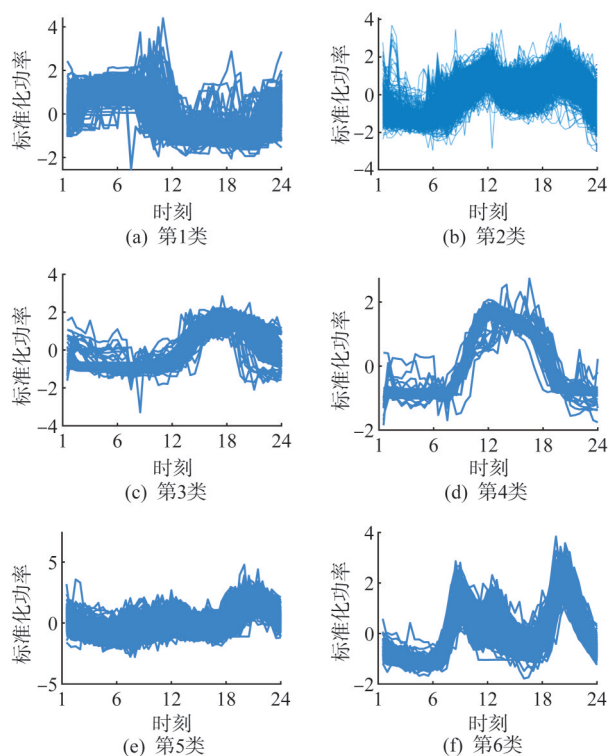


图 12 SSA-iDPC 算法在工商业数据集的聚类结果

Fig. 12 Clustering results of SSA-iDPC algorithm in commerce and industry dataset

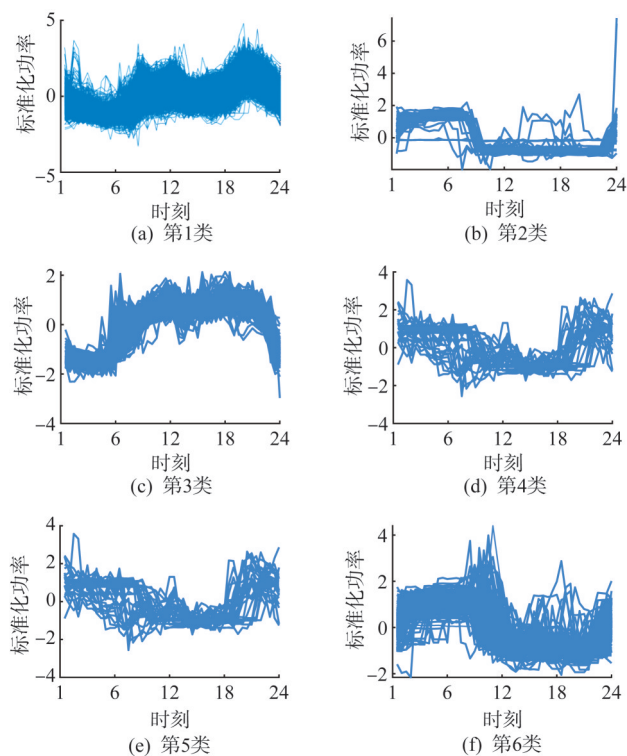


图 14 K-medoids 算法在商业数据集的聚类结果

Fig. 14 Clustering results of K-medoids algorithm in commerce and industry dataset

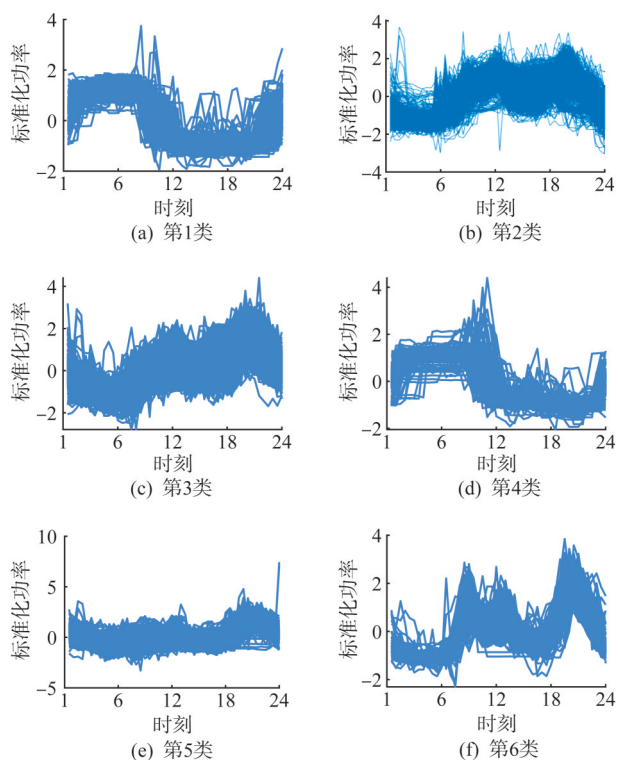


图 13 DPC 算法在工商业数据集的聚类结果

Fig. 13 Clustering results of DPC algorithm in commerce and industry dataset

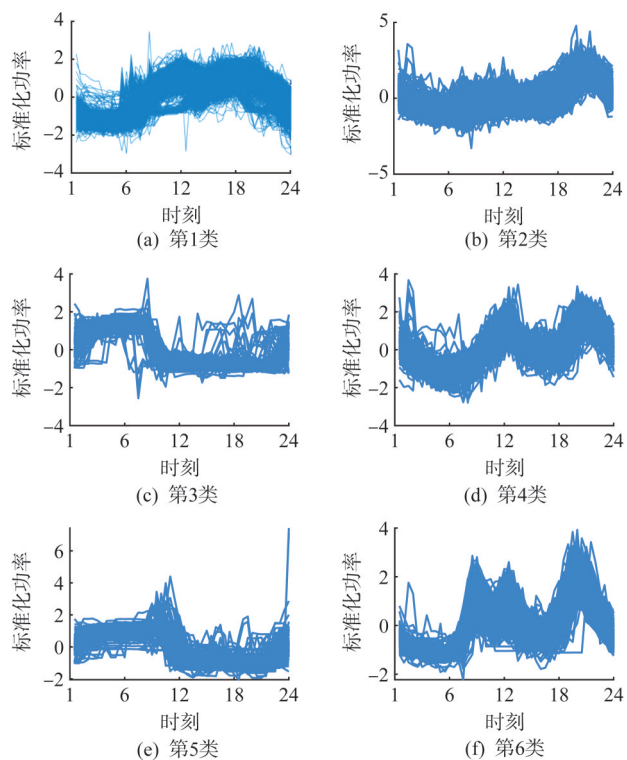


图 15 GMM 算法在商业数据集的聚类结果

Fig. 15 Clustering results of GMM algorithm in commerce and industry dataset

0.318 3, 相较于其他算法均为最佳。由于数据集中存在较多相似度高的负荷曲线, 且负荷曲线存在较大程度波动, 各类算法的整体聚类效果较 OpenEI 数据集有所下降, 本文算法对 DPC 算法的提升较小。

表 2 不同算法在工商业数据集的聚类效果评价指标对比
Tab. 2 Comparison of clustering effect evaluation indexes of different algorithms in commerce and industry dataset

聚类评价指标	最佳类簇数	SSE 值	DBI 值	SC 值
SSA-iDPC	6	9 648. 1	1. 087 5	0. 318 3
VMD-iDPC	6	10 814. 4	2. 835 4	0. 291 3
iDPC	6	9 626. 6	1. 421 1	0. 281 6
DPC	6	10 269. 9	1. 442 8	0. 231 0
SSA-DPC	6	10 243. 8	1. 419 5	0. 253 7
K-medoids	6	10 028. 6	1. 849 9	0. 224 6
GMM	6	10 195. 5	1. 704 2	0. 220 7

3.3 算法性能分析

3.3.1 噪声鲁棒性分析

为模拟聚类算法在不同场景下实际负荷因不同原因产生的波动的聚类有效性, 在原始负荷曲线上添加不同比例的高斯白噪声进行实验。选取 3.2.1 节中 4 类算法均能够划分出的图 8 中的第 1、2、3 和 5 类的双峰、避峰、单峰和三峰 4 类形态负荷曲线样本加以不同比例的噪声进行实验, 通过准确率 r 对聚类效果进行评价, r 为分类正确的负荷曲线数与总数之比。实验结果如表 3 所示。从表 3 中可以看出, 在噪声比例较低时, 各类算法均能准确识别出 4 类负荷曲线。随着噪声比例的增加, 准确率 r 逐渐下降。iDPC 和 DPC 算法在 $d=15\%$ 及之后未能保持 100% 的准确率, 而 SSA-DPC 在 $d=20\%$ 及之后也未能保持 100% 的准确率。相比之下, 本文提出的 SSA-iDPC 在 $d=20\%$ 之前的聚类结果完全正确, 表现出较好的噪声鲁棒性。实验表明, 本文所提出的算法能够在保留原始负荷曲线轮廓形态的同时有效消除部分噪声成分, 减弱噪声及其他相关因素对聚类结果的影响, 即使在中小程度负荷局部波动的情况下, 依然能够保持较高质量的聚类效果。

3.3.2 参数敏感性分析

考虑到改进密度峰值聚类算法中 K 近邻参数的选择直接影响局部密度估计的精度, 进而可能构成算法性能的影响因素。为系统评估该参数的敏感性, 在 3.2.1 节所述数据集上构建了对照实验组:

表 3 不同算法的噪声鲁棒性对比

Tab. 3 Comparison of noise robustness of different algorithms

噪声比 d	分类准确率 r			
	SSA-iDPC	iDPC	DPC	SSA-DPC
0	100	100	100	100
5	100	100	100	100
10	100	100	100	100
15	100	95. 80	87. 97	100
20	100	87. 68	83. 15	98. 92
25	93. 87	85. 73	76. 95	90. 63
30	85. 50	81. 20	63. 40	80. 98

将 K 近邻数设定为总样本量 1%~2% 的区间范围内 (取值间隔为 0.1%) 测试不同参数配置下的聚类效果, 获得实验结果如表 4 所示。根据表 4, 当 K 近邻数量占比在 [1%, 2%] 区间内变化时, 各指标都能够维持在偏差相对较小的范围内, 且整体聚类质量始终稳定, 相较于其他聚类算法仍具有一定优势。

表 4 K 近邻数量对聚类结果的影响

Tab. 4 Influence of the number of K -nearest neighbors on clustering results

K 近邻数量占比/%	聚类评价指标		
	SSE 值	DBI 值	SC 值
1. 0	8 305. 7	0. 785 3	0. 497 5
1. 1	8 404. 8	0. 595 3	0. 504 2
1. 2	8 404. 8	0. 595 3	0. 504 2
1. 3	8 552. 5	0. 796 9	0. 475 4
1. 4	8 404. 8	0. 595 3	0. 504 2
1. 5	9 107. 4	0. 795 5	0. 339 9
1. 6	8 404. 8	0. 595 3	0. 504 2
1. 7	8 404. 8	0. 595 3	0. 504 2
1. 8	8 404. 8	0. 595 3	0. 504 2
1. 9	8 240. 1	0. 784 2	0. 440 7
2. 0	8 404. 8	0. 595 3	0. 504 2

3.3.3 算法复杂度分析

设 n 为负荷曲线数量, 经典 DPC 算法由于需要构建全连接距离矩阵并进行全局排序, 其时间复杂度为 $O(n^2)$ 。本文提出的 SSA-iDPC 算法中, 奇异谱分析 (SSA) 中分解与重构的整体时间复杂度为 $O(n)$; 改进的局部密度估计需要对每个样本点 K 近邻领域内的密度加权 and 进行计算, 其时间复杂度为 $O(n^2 + n^2)$, 故本文算法的时间复杂度为 $O(n^2 +$

$n^2) + O(n) \sim O(n^2)$ 。相较于原始 DPC 算法, 本文算法通过引入可控的计算复杂度增量, 显著提升了聚类效果, 在更大规模数据下不会造成较大计算资源浪费。

4 结语

针对日负荷曲线聚类受数据局部波动和各类样本数量分布不均匀影响导致样本点分配错误的问题, 本文提出了一种基于奇异谱分析和改进密度峰值算法的电力负荷曲线聚类方法。通过奇异谱分析获取包含负荷数据主要轮廓特征的低频分量, 并针对密度峰值聚类算法的局部密度进行改进。算例结果表明, 本文方法能够更准确地划分出不同类型的负荷形态, 具有更强的鲁棒性, 有助于供电部门深刻把握各类电力用户的用电行为与能耗模式, 从而为电力用户提供更加高质量的服务。如何在保证聚类效果的情况下降低算法在更大规模数据下的计算复杂度是未来研究的重点方向。

参考文献

- [1] 汪际峰, 吴小辰, 林火华, 等. 数字电网的概念、特征与架构[J]. 南方电网技术, 2023, 17(12): 36–41, 89.
WANG Jifeng, WU Xiaochen, LIN Huohua, et al. Concepts, features and architectures of digital grids[J]. Southern Power System Technology, 2023, 17(12): 36–41, 89.
- [2] WANG Y, CHEN Q, HONG T, et al. Review of smart meter data analytics: applications, methodologies, and challenges[J]. IEEE Transactions on Smart Grid, 2018, 10(3): 3125–3148.
- [3] CUI X Y, LIU S Y, LIN Z Z, et al. Two-step electricity theft detection strategy considering economic return based on convolutional autoencoder and improved regression algorithm[J]. IEEE Transactions on Power Systems, 2022, 37(3): 2346–2359.
- [4] 卓映君, 周保荣, 姚文峰, 等. “双碳”目标下南方区域能源电力消费及负荷特性预测[J]. 南方电网技术, 2023, 17(9): 132–140.
ZHUO Yingjun, ZHOU Baorong, YAO Wenfeng, et al. Prediction of energy and electricity consumption and load characteristics in the southern region under the carbon peak and carbon neutrality target[J]. Southern Power System Technology, 2023, 17(9): 132–140.
- [5] 李璐, 张泽端, 毕贵红, 等. “双碳”目标下基于系统动力学的发电行业碳减排政策研究[J]. 电力系统保护与控制, 2024, 52(12): 69–81.
LI Lu, ZHANG Zeduan, BI Guihong, et al. Carbon emission reduction policy in the power generation sector based on system dynamics with “dual carbon” targets[J]. Power System Protection and Control, 2024, 52(12): 69–81.
- [6] 李鹏, 黄文琦, 梁凌宇, 等. 人机混合增强决策智能在新型电力系统调控中的应用与展望[J]. 中国电机工程学报, 2024, 44(16): 6347–6367.
- [7] LI Peng, HUANG Wenqi, LIANG Lingyu, et al. Application and prospect of hybrid human-machine decision intelligence in the dispatch and control of next-era power systems[J]. Proceedings of the CSEE, 2024, 44(16): 6347–6367.
- [8] 李婧, 徐胜蓝, 万灿, 等. 基于自适应 k-means++ 算法的电力负荷特性分析[J]. 南方电网技术, 2019, 13(2): 13–19.
LI Jing, XU Shenglan, WAN Can, et al. Electricity load characteristics analysis based on adaptive k-means++ algorithm[J]. Southern Power System Technology, 2019, 13(2): 13–19.
- [9] 谷紫文, 李鹏, 郎恂, 等. 基于变分模态分解和密度峰值快速搜索的电力负荷曲线可控聚类模型[J]. 电力系统保护与控制, 2021, 49(8): 118–127.
GU Ziwen, LI Peng, LANG Xun, et al. A controllable clustering model of the electrical load curve based on variational mode decomposition and fast search of the density peak[J]. Power System Protection and Control, 2021, 49(8): 118–127.
- [10] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492–1496.
- [11] 张清华, 周靖鹏, 代永杨, 等. 基于代表点与 K 近邻的密度峰值聚类算法[J]. 软件学报, 2023, 34(12): 5629–5648.
ZHANG Qinghua, ZHOU Jingpeng, DAI Yongyang, et al. Density peaks clustering algorithm based on representative points and K-nearest neighbors[J]. Journal of Software, 2023, 34(12): 5629–5648.
- [12] 陈叶旺, 申莲莲, 钟才明, 等. 密度峰值聚类算法综述[J]. 计算机研究与发展, 2020, 57(2): 378–394.
CHEN Yewang, SHEN Lianlian, ZHONG Caiming, et al. Survey on density peak clustering algorithm[J]. Journal of Computer Research and Development, 2020, 57(2): 378–394.
- [13] 余浩, 高懿涛, 潘险险, 等. 基于改进高斯混合模型的变电站负荷聚类算法[J]. 全球能源互联网, 2024, 7(5): 591–601.
YU Hao, GAO Yihao, PAN Xianxian, et al. Substation load clustering algorithm based on improved gaussian mixture model[J]. Journal of Global Energy Interconnection, 2024, 7(5): 591–601.
- [14] LU F, CUI X Y, XING J X, et al. Electricity load profile characterisation for industrial users based on normal cloud model and iCFSFDP algorithm[J]. IEEE Transactions on Power Systems, 2023, 38(4): 3799–3813.
- [15] 孙园, 李秋雨, 黄冬梅, 等. 基于粒计算和双尺度相似性的负荷曲线集成聚类算法[J]. 电力建设, 2022, 43(5): 117–126.
SUN Yuan, LI Qiuyu, HUANG Dongmei, et al. Clustering ensemble model based on granular computing and dual-scale similarity[J]. Electric Power Construction, 2022, 43(5): 117–126.
- [16] GU Z, LI P, XUN L, et al. A multi-granularity density peak clustering algorithm based on variational mode decomposition[J]. Chinese Journal of Electronics, 2021, 30(4): 658–668.
- [17] 郭文熙, 李知艺, 尹建兵, 等. 基于时间序列变密度处理的负荷曲线聚类分析[J]. 电力工程技术, 2024, 43(6): 21–32.
GUO Wenxi, LI Zhiyi, YIN Jianbing, et al. Clustering analysis of load curve based on time series density changing processing[J]. Electric Power Engineering Technology, 2024, 43(6): 21–32.
- [18] WANG S L, WANG D K, LI C Y, et al. Comment on

- "Clustering by fast search and find of density peaks"[DB/OL]. (2015-01-18)[2015-01-20]. <https://arxiv.org/abs/1501.04267>.
- [18] 葛阳, 郭兰中, 牛曙光, 等. 基于t-SNE和LSTM的旋转机械剩余寿命预测[J]. 振动与冲击, 2020, 39(7): 223-231, 273.
GE Yang, GUO Lanzhong, NIU Shuguang, et al. Prediction of remaining useful life based on t-SNE and LSTM for rotating machinery[J]. Journal of Vibration and Shock, 2020, 39(7): 223-231, 273.
- [19] 黄雨薇, 彭道刚, 姚峻, 等. 基于SSA和K均值的TD-BP神经网络超短期光伏功率预测[J]. 太阳能学报, 2021, 42(4): 229-238.
HUANG Yuwei, PENG Daogang, YAO Jun, et al. Ultra-short-term photovoltaic power forecast of TD-BP neural network based on SSA and K-means[J]. Acta Energetica Sinica, 2021, 42(4): 229-238.
- [20] AFSHAR K, BIGDELI N. Data analysis and short term load forecasting in Iran electricity market using singular spectral analysis (SSA)[J]. Energy, 2011, 36(5): 2620-2627.
- [21] HASSANI H. Singular spectrum analysis: methodology and comparison[J]. RePEc, 2007, 5(2): 239-257.
- [22] PETERSON L E. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2): 1883.
- [23] ANDREW L R. Analysis of time series structure: SSA and related techniques[J]. Technometrics, 2002, 44(3): 290.
- [24] GOLYANDINA N E, NEKRUTKIN V V, ZHIGLJAVSKY A A. Analysis of time series structure: SSA and related techniques[J]. Technometrics, 2002, 44(3): 290.
- [25] 彭晓璐, 王涛, 卢泽钰, 等. 基于改进密度峰值聚类算法的典型负荷曲线提取[J]. 南方电网技术, 2025, 19(9): 150-161.
PENG Xiaolu, WANG Tao, LU Zeyu, et al. Typical load profile extraction based on improved DPC algorithm [J]. Southern Power System Technology, 2025, 19(9): 150-161.
- [26] 陈仕龙, 吴涛, 郭成, 等. 基于DBSCAN聚类 and 区间回归的多谐波责任划分[J]. 中国电力, 2024, 57(2): 138-148.
CHEN Shilong, WU Tao, GUO Cheng, et al. Division of multi-harmonic responsibilities based on DBSCAN clustering and interval regression[J]. Electric Power, 2024, 57(2): 138-148.
- [27] GUO G, CHEN L, YE Y, et al. Cluster validation method for determining the number of clusters in categorical sequences[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(12): 2936-2948.
- [28] 谢娟英, 周颖, 王明钊, 等. 聚类有效性评价新指标[J]. 智能系统学报, 2017, 12(6): 873-882.
XIE Juanying, ZHOU Ying, WANG Mingzhao, et al. New criteria for evaluating the validity of clustering [J]. CAAI Transactions on Intelligent Systems, 2017, 12(6): 873-882.
- [29] 李丹, 梁云嫣, 缪书唯, 等. 基于高斯混合聚类和改进条件变分自编码的多风电场功率日场景生成方法[J]. 中国电力, 2024, 57(12): 17-29.
LI Dan, LIANG Yunyan, MIAO Shuwei, et al. Daily power scenario generation method for multiple wind farms based on Gaussian mixture clustering and improved conditional variational autoencoder[J]. Electric Power, 2024, 57(12): 17-29.

收稿日期: 2024-09-25; 网络首发日期: 2025-06-01

作者简介:

赵俊(2001), 男, 硕士研究生, 研究方向为电力负荷特性分析, 2396196679@qq.com;

李鹏(1976), 男, 通信作者, 教授, 博士, 研究方向为电力物联网与能源互联网、电力大数据分析建模, lipeng@ynu.edu.cn;

李文超(2002), 男, 硕士研究生, 研究方向为电力负荷预测, 2926250119@qq.com。