DOI: 10.19783/j.cnki.pspc.210364

基于改进自适应密度峰值算法的日负荷曲线聚类分析

姚黄金，雷 霞，付鑫权，胡 益

（西华大学电气与电子信息学院，四川 成都 610039）

摘要：电力市场的逐步开放以及大量可再生能源的接入使用户具有更多的用电自由，导致电力用户类型多样化、用户间负荷特性差异逐渐增大、负荷数据的类簇分布情况复杂化。为解决传统聚类算法面对不均衡负荷数据集时聚类效果不佳以及缺乏自适应能力等问题，提出一种改进自适应密度峰值聚类(Improved self-adaptive Density Peak Clustering, ISDPC)算法。首先，基于 K-最近邻(K-Nearest Neighbor, KNN)和相对密度的思想定义了一种新的密度度量方式。然后在决策图中拟合分段函数得到最优类簇数目。最后，通过构造加权 KNN 图改进样本分配策略。试验结果表明，与传统聚类算法相比，所提方法聚类结果更加精确、具备自适应能力、鲁棒性更强。

关键词：负荷曲线聚类；密度峰值聚类；自适应；KNN；鲁棒性

Cluster analysis of daily load curves based on an improved self-adaptive density peak clustering algorithm

YAO Huangjin, LEI Xia, FU Xinquan, HU Yi

(College of Electrical and Electronic Information, Xihua University, Chengdu 610039, China)

Abstract: The opening electricity market and the incremental penetration of renewable energy provide more consumption choices for users. This results in diversification of power user patterns, increasing differences of load characteristics and giving a complex distribution of load clusters. An improved self-adaptive density peak clustering (ISDPC) algorithm is proposed to ameliorate the clustering results and adaptive abilities of traditional clustering methods for unbalanced load data. First, a new density metric is defined based on the K-nearest neighbor (KNN) and relative density. Secondly, the optimal number of clusters is obtained by a fitting partition function obtained from the decision graph. Finally, the allocation of strategy is improved by a weighted KNN graph. The experimental results show that clustering results obtained from the proposed method perform better in accuracy, robustness, and adaptability.

This work is supported by the National Natural Science Foundation of China (No. 51877181).

Key words: load profiles clustering; density peak clustering; self-adaptation; KNN; robustnes

0 引言

随着电力体制改革朝着市场化的方向不断迈进，售电公司越来越重视用户的用电体验。面对竞争愈加激烈的市场环境，为提升客户服务水平、争夺市场份额，售电公司需要从海量、多样的用电数据中挖掘用户的用电特征，充分掌握用户的用电偏好，从而辅助售电公司进行负荷预测[1-3]、参与需求侧响应[4-5]、制定电价等营销策略[6-7]。对用户日负荷曲线进行聚类分析，科学合理地划分用户群体是售电公司分析用户用电特性的重要手段[8-11]。

目前针对负荷曲线聚类研究大多集中在 3 个方面：对传统聚类算法的性能优化、数据降维以及样本相似性度量的改进。针对基于划分的 k-means 和模糊 C 均值(fuzzy C-mean, FCM)算法需要人为设置类簇数目、对初始类簇中心敏感等缺陷，文献[12]通过引入分位数半径令 k-means 算法能够识别类簇数目、产生较为理想的类簇中心。文献[13]提出一种基于灰狼优化的 FCM 聚类算法，能够快速搜索出最优初始类簇中心，提高 FCM 算法的全局寻优能力。文献[14]将半监督学习与改进 AP 聚类算法相结合，完成了对居民用户的负荷分解。为解决高维数据给聚类分析带来的困扰，文献[15-22]分别通过奇异值分解(SVD)、主成分分析法(PCA)、离散小波变换(DWT)、提取负荷特征指标、自动编码器(CAE)、卷积神经网络(CNN)、分段聚合近似等方法

基金项目：国家自然科学基金项目资助(51877181)

降低负荷数据维度。在对样本相似性度量改进的研究中，文献[23]综合考虑了负荷曲线的分布特性和动态特性，将欧氏距离与动态时间弯曲距离相结合，文献[24]基于推土机距离(EMD)衡量了负荷曲线的纵向随机性。然而，随着电力市场的逐步开放以及大量可再生能源的接入，令用户拥有了更大的用电自由，增大了负荷的随机性、波动性以及用户间负荷特性的差异，导致所采集负荷数据的样本分布情况愈加复杂，产生了类簇形状差异大、分布不均衡的负荷数据集，而传统聚类算法在这种数据集上的聚类效果不佳。此外，面对不同的负荷数据集，若每一次聚类分析都需要人为调整某些参数，显然是不利于实际应用的。针对上述问题，为得到准确的聚类结果，聚类算法需同时满足以下两点要求：1) 能够适用于任意类簇分布情况；2) 具备自适应能力。

文献[25]提出快速搜索和寻找密度峰值的聚类(Clustering by fast search and find of density peaks, DPC)，简称密度峰值聚类。由于该算法聚类速度快、能够快速发现任意形状的类簇、鲁棒性强，已广泛用于图像识别、社区发现等领域[26]。目前，已有文献通过 kd 树算法[27]、类间类内优化[28]、变分模态分解[29-30]等方法对 DPC 算法在负荷聚类时运算速度慢、不适用于多种用户类型、聚类精度不佳等缺陷进行了改进，但这些方法均不能同时满足上述两点要求。

鉴于此，本文提出一种 ISDPC 算法。首先，基于 KNN 相对密度的思想提出了一种新的密度度量方式；然后通过在决策图中拟合分段函数确定最优类簇数目；最后构造出加权 KNN 图，兼顾样本间的属性相似性和结构相似性，并基于图的距离衡量样本与类簇中心的相似度，改进样本分配策略。算例分析结果表明，与传统聚类算法相比，本文所提方法聚类结果更加精确，适用于不同分布情况的负荷数据集，鲁棒性更强，且具备自适应能力。

1 密度峰值算法

1.1 算法原理

DPC 算法有两点基本假设：(1) 每个类簇中心的局部密度高于周围相邻点的密度；(2) 类簇中心之间的距离较远。

1) 局部密度

样本点的局部密度 ρ_i 分为截断核和高斯核两种计算方式，分别如式(1)和式(2)所示。

截断核计算方式为

$$\rho_i=\sum_{i\neq j}\chi(d_{i,j}-d_c) \tag{1}$$

高斯核计算方式为

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

式中：若 $x<0$，则 $\chi(x)=1$，否则 $\chi(x)=0$；$d_{i,j}$ 为样本点 x_i 和 x_j 之间的欧氏距离；d_c 为截断距离($d_c>0$)，通常是选取所有样本点之间距离的前 2%处的值。

2) 最小距离

最小距离 δ_i 为样本点和高于其密度且相距最近样本点之间的距离。

$$\delta_i=\begin{cases}\max\limits_{j\neq i}\left(d_{i,j}\right) & \rho_i\text{为最大}\\ \min\limits_{j:\rho_i>\rho_j}\left(d_{i,j}\right) & \rho_i\text{非最大}\end{cases} \tag{3}$$

3) 决策图

DPC 算法认为同时拥有较高的局部密度 ρ_i 和较大最小距离 δ_i 的样本点为密度峰值点，即类簇中心。通过计算决策值 γ_i 可以直接确定密度峰值点，决策值 γ_i 定义为

$$\gamma_i=\rho_i\cdot\delta_i \tag{4}$$

将 γ 值按降序排列，以降序后的 γ 值为纵坐标、排列顺序为横坐标形成决策图，然后根据决策图选出前 c 个 γ 值远大于剩余样本的点作为类簇中心。

4) 样本分配策略

选出类簇中心后再对剩余样本点进行分配，DPC 算法的分配策略为：按局部密度下降的顺序，将剩余样本点依次分配到比其局部密度更高且距离最近的样本点所属类簇之中。

1.2 算法缺陷分析

传统的 DPC 算法应用于类簇分布情况复杂的日负荷曲线聚类存在以下缺陷：

1) DPC 算法定义的局部密度并未考虑数据内部的结构差异，当类簇之间的密度差异过大时，通过固定的截断距离所计算出的局部密度不能真实地反映样本点的疏密情况。如图 1 所示，无论截断距离如何取值，所有绿色样本点的局部密度均全大于蓝色样本点。由于决策值由局部密度和最小距离的乘积决定，当样本间的局部密度差异过大时，会对类簇中心的确定造成影响。

2) 类簇中心需要通过决策图人为确定。

3) 若因数据样本分布不均匀或存在流型结构，导致某些类簇边缘点相距其他类簇较近时，DPC 算法的分配策略会造成样本的错误分配，并且一旦某个样本点分配错误，后续分配还会进一步放大这一

错误。如图 1 所示，点 B 与点 A 的距离最为接近且局部密度小于点 A，根据样本分配策略则点 B 与点 A 属于同一类簇。

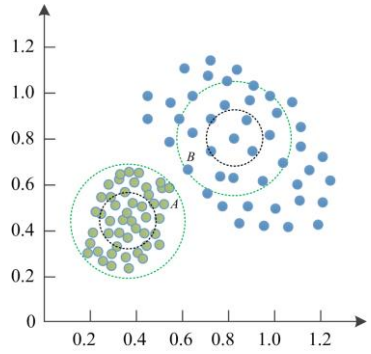


图 1 样本分布图

Fig. 1 Sample distribution diagram

2　改进自适应密度峰值聚类算法

针对 DPC 算法在日负荷曲线聚类时的缺陷，本文提出一种改进自适应密度峰值聚类(ISDPC)算法。

2.1 改进局部密度

为使 DPC 算法不再需要人为设置截断距离，且适用于类簇密度差距较大的数据样本，基于 KNN 和相对密度的思想重新定义了局部密度的计算方法。

DPC 算法设置截断距离判断距离半径内样本点的个数实质上就是计算 ε-最近邻的过程：设置一个扫描参数 ε，然后寻找在扫描半径内的邻居点数量。为了在计算局部密度时考虑样本的内部结构，利用 K-最近邻的思想，通过样本点与其 k 个近邻点之间的平均距离来反映样本点的局部密度，平均距离越短，则该点局部密度越大。

但是 K-最近邻也存在一个与 ε-最近邻相同的问题，即需要手动设置相应参数。为能够根据样本实际情况自适应地获得参数 k，借助自然最近邻(Natural Nearest Neighbor, 3N)算法(该算法可以看作是一种可以自动获得 k 值的 K-最近邻算法)计算得到自然特征值 sup_k 作为 KNN 算法的 k 值。

定义 1(逆近邻)：若样本点 x_j 是样本点 x_i 的其中一个 K-最近邻点，则称样本点 x_i 是样本点 x_j 的逆近邻，记为 $x_i \in \mathrm{RNN}(x_j)$。

定义 2(自然稳定状态)：在自然邻居搜索过程中，若每个数据点都有逆近邻或者当所有逆近邻个数为 0 的数据不变时，自然邻居搜索达到自然稳定状态。

定义 3(自然特征值)：当自然邻居的搜索达到自然稳定状态时，自然邻居的搜索次数即为自然特征值，表示数据样本的平均近邻节点数，记为 sup_k。

自然最近邻算法步骤如下：

1) 初始化搜索指数 $r=1$，逆近邻集合 $\mathrm{RNN}=\varnothing$。

2) 计算每个样本 x_i 的 $\mathrm{KNN}_r(x_i)$、$\mathrm{RNN}(x_i)$。

3) $r=r+1$。

4) 当 $\forall x_j$ 使得 $\mathrm{RNN}(x_j)\neq\varnothing$ 或所有令 $\mathrm{RNN}=\varnothing$ 的 x_j 不再变化时，$sup_k=r-1$，输出 sup_k，否则转跳至 2)。

当数据样本中的各个类簇分布疏密程度差异过大时，最小距离对决策值的影响会减弱，局部密度将直接决定决策值的大小，导致稀疏类簇的类中心无法被发现。通过计算样本点在局部范围内的相对密度来减小这一影响，改进后的局部密度 ρ_i' 定义为

$$\rho_i' = \exp\left(-\frac{\sum\limits_{x_j\in \mathrm{KNN}_k(x_i)} d_{i,j}}{\sum\limits_{x_j\in \mathrm{KNN}_k(x_i)}\ \sum\limits_{x_l\in \mathrm{KNN}_k(x_j)} d_{i,l}}\right) \tag{5}$$

式中，$\mathrm{KNN}_k(x_i)$ 为与 x_i 距离最近的 k 个样本点所构成的集合。改进后的局部密度可以将稀疏类簇样本点的局部密度放大，密集类簇样本点的局部密度缩小，避免了因类簇疏密差异过大而不能准确判断出类簇中心的问题。

2.2 类簇中心选择

类簇中心的决策值 γ 远大于其他样本点，将决策值 γ 按降序排列，则会出现明显的分段现象。基于此，通过两条直线对决策值 γ 进行分段拟合，使函数拟合误差达到最小的分段点即为最优类簇中心数，如图 2 所示。

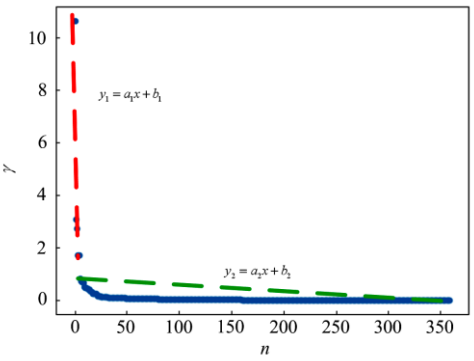


图 2 决策值拟合

Fig. 2 Decision value fitting

类簇中心自动选择方法步骤如下：

1) 将决策值 γ 按降序排列，以 n 个点 (i,ρ_i) 作为数据集 S，记数据集 S 中的第一个点为 S_1，最后

一个点为 S_n。

2) 设置分段点初始值 t=2，分别过点 S_1、S_t 和点 S_1、S_n 作直线，得到 $y_1=a_1x+b_1$ 和 $y_2=a_2x+b_2$。分别计算当 $x=1,2,\cdots,n$ 时对应的拟合值 y^*。

3) 由式(6)计算拟合误差 θ_t。

$$\theta_t=\frac{1}{n}\sum\left(y-y^*\right)^2 \tag{6}$$

4) 令 $t=t+1$，当 $t>n$ 时，转到 5)，否则转到 2)。

5) 找到最小的 θ_t，t 即为最优类簇中心数，前 t 个 γ 值所对应的样本点即为类簇中心 $v=[v_1,v_2,\cdots,v_t]$。

2.3 基于 KNN 图的分配策略

基于 KNN 图的分配策略主要分为 3 步：构建加权 KNN 图、计算最短路径以及样本分配。

首先，构建加权 KNN 图 $G=(V,E,W)$：以所有样本点作为节点集 V，若满足 $x_j\in \mathrm{KNN}_k(x_i)$ 则两节点之间存在边 $e_{i,j}\in E$，边的综合权重 $W_{i,j}$ 定义为

$$W_{i,j}=d_{i,j}+(1-J_{i,j}) \tag{7}$$

$$J_{i,j}=\frac{\left|\Gamma(x_i)\cap\Gamma(x_j)\right|}{\left|\Gamma(x_i)\cup\Gamma(x_j)\right|} \tag{8}$$

式中：$d_{i,j}$ 为样本点 x_i 与 x_j 之间的欧式距离，代表样本之间的属性相似性；$J_{i,j}\in[0,1]$ 为 Jaccard 系数，代表样本之间结构相似性，KNN 图中两节点的公共邻接点越多，Jaccard 系数越大，代表两节点在结构上就越相似；$\Gamma(x_i)$ 为节点 x_i 的邻接点集合。边的综合权重同时兼顾了样本点之间的属性相似性和结构相似性，通过样本在图中的连接结构，扩大了不同类簇节点之间的差异。

然后，通过 Dijkstra 算法计算在加权 KNN 图中各类簇中心到其余节点的最短路径，得到最短路径矩阵 $\boldsymbol{L}$，如式(9)所示。路径越短，节点就与类簇中心越相似。

$$\boldsymbol{L}=\begin{bmatrix} L_{1,1} & L_{1,2} & \cdots & L_{1,c} \\ L_{2,1} & L_{2,2} & \cdots & L_{2,c} \\ \vdots & \vdots & & \vdots \\ L_{n-c,1} & L_{n-c,2} & \cdots & L_{n-c,c} \end{bmatrix} \tag{9}$$

最后，按照式(10)将非类簇中心点分配到所属类簇。

$$C_k=\left\{x_i\left|\arg\min_j\left(L_{i,j}\right)=k,\ j\in[1,t]\right.\right\} \tag{10}$$

2.4 改进自适应密度峰值算法流程

本文所提出的 ISDPC 聚类算法整体流程如图 3 所示。

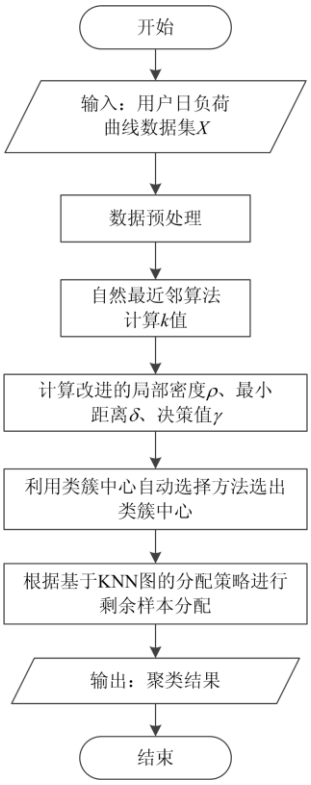


图 3 ISDPC 算法流程图

Fig. 3 Flow chart of ISDPC algorithm

输入为用户日负荷曲线数据集 $X=[x_1,x_2,\cdots,x_N]^{\mathrm{T}}$，其中，$x_i=\left[x_{i1},x_{i2},\cdots,x_{iM}\right]$，$N$ 为样本数量，M 为样本维度。

具体步骤如下。

1) 数据预处理：电力系统在实际运行中，由于测量和信道的误差及系统各种故障及冲击负荷的影响，导致负荷数据中会存在缺失、噪声等问题[31]。因此，有必要对负荷数据进行预处理，包括缺失值填补、曲线平滑滤波、数据归一化处理。

2) 由自然最近邻算法计算得到自然特征值 sup_k 作为 KNN 的 k 值。

3) 分别根据式(5)、式(3)、式(4)计算样本的局部密度 ρ'、最小距离 δ 和决策值 γ。

4) 通过类簇中心自动选择方法，选出类簇中心 v。

5) 根据基于加权 KNN 图的样本分配策略对剩余样本点进行分配。

3 聚类有效性评价指标

聚类评价指标分为外部和内部评价指标[32-33]，二者的主要区别在于外部评价指标需要与依靠聚类

结果相关的外部信息来评价聚类的准确性，而内部评价指标通过计算所划分类簇的类内紧凑度和类间分离度来衡量聚类效果的优劣。由于实际日负荷曲线的聚类分析往往是缺少外部结果信息的，故本文以轮廓系数(Silhouette Coefficient, SC)和戴维森指数(Davies-Bouldin index, DBI) 2 个内部评价指标评价各类聚类算法的聚类效果。

若日负荷曲线数据集 $\boldsymbol{X}$ 被划分为 t 个类簇 C_1, $C_2,\cdots,C_t$，SC 定义如下。

$$SC(x_i)=\frac{b(x_i)-a(x_i)}{\max\{a(x_i),b(x_i)\}} \tag{11}$$

$$SC=\frac{1}{N}\sum_{i=1}^{N}SC(x_i) \tag{12}$$

式中：$b(x_i)$ 为 x_i 与非同一类簇样本点之间的平均距离；$a(x_i)$ 为 x_i 与同一类簇样本点之间的平均距离。类间越分散、类内越紧凑，即 $b(x_i)$ 越大、$a(x_i)$ 越小，则聚类效果越好。

DBI 定义为

$$DBI=\frac{1}{t}\sum_{i=1}^{t}\max_{j\neq i}\left(\frac{\bar{S}_i+\bar{S}_j}{M_{i,j}}\right) \tag{13}$$

式中：$\bar{S}_i$ 为第 i 个类簇的类簇中心到类内各点的平均距离；$M_{i,j}$ 为类簇中心 v_i 和 v_j 的欧氏距离。DBI 的值越小，聚类效果就越好。

4　算例分析

为验证本文所提出的方法在用户负荷曲线聚类中的有效性，本节基于真实负荷数据和模拟负荷数据两个数据集，采用本文方法和传统聚类方法进行聚类分析并比较聚类结果。

实验环境：Intel(R) Core(TM) i5-4460 CPU@ 3.20 GHz，8.00 GB RAM，编程语言为 Python 3.6。

4. 1　算法聚类效果对比

4.1.1 真实负荷数据聚类效果

真实负荷数据集来自美国能源部 OpenEI 公布的工商业用户负荷数据，共 2 260 条工作日负荷曲线，每小时采集一次，每日共计 24 个采样点。

为直观地呈现数据样本的分布情况，通过多维尺度变换(Multiple Dimensional Scalling, MDS)将数据样本从高维映射到二维平面。真实负荷数据集的二维分布如图 4 所示，该数据集中的类簇形状多为球形，类簇形状差异不大。

ISDPC 算法聚类结果如图 5 所示，将用户负荷曲线划分为 7 类，可归为单峰、双峰、三峰、避峰

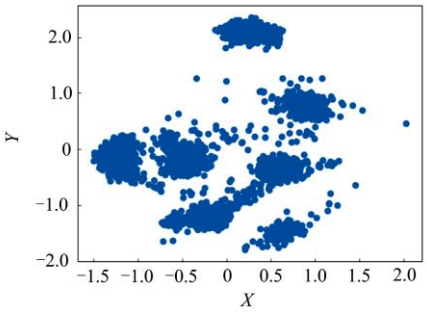


图 4　真实负荷数据集的二维映射

Fig. 4 Two-dimensional mapping of the real load dataset

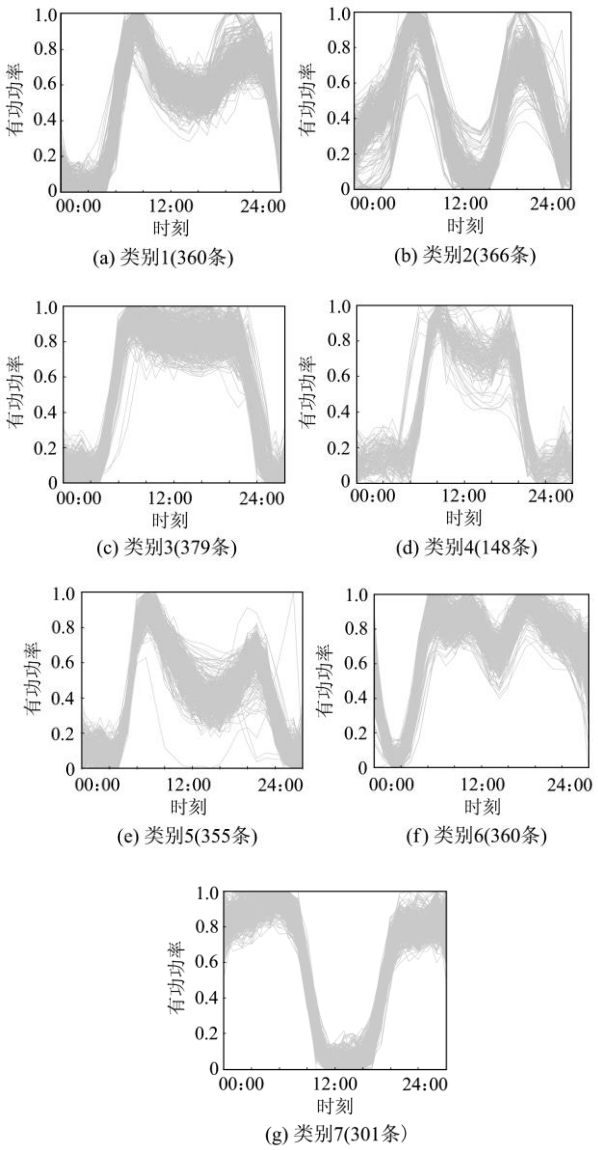


图 5　ISDPC 算法对真实负荷数据集的聚类结果

Fig. 5 Clustering results of ISDPC algorithm for the real load dataset

4 型。类别 3、4 为单峰型负荷，主要包含学校、零售业、医院等行业用户，用电集中在白天。类别 3 的用户呈现日间长时段高峰用电特性，用电时间为 06：00 左右到 20：00；类别 4 的用户在白天用电时长略短于类别 3，且午间休息时段的用电量略有下降。类别 1、2、5 为双峰型负荷，主要包含居民、酒店住宿等行业用户，具有早高峰和晚高峰用电特性。类别 5 用户早高峰时段用电量高于晚高峰用电量；类别 1、2 用户早晚高峰时段用电量接近，但是类别 2 用户午间用电量极低，具有明显的午休现象。类别 6 属于三峰型负荷，可能是餐饮行业用户，在早中晚时段均存在高峰用电现象，在上午工作时段和午间休息时段用电有所下降，凌晨时段用电量大幅下降。类别 7 为避峰型负荷，可能是高能耗企业用户，为了降低用电成本，选择在夜间进行生产任务，呈现夜间用电特性。

DPC 算法会因为设置的截断距离参数不同而导致聚类效果发生变化，如表 1 所示当截断距离 $d_c=0.396$ 时聚类效果最佳，图 6 为该截断距离下的决策图，从图中可以看出最优类簇数为 7；图 7 为最终的聚类结果，对比图 5 和图 7 可以看出，DPC 算法错误地将属于类别 2 中的部分负荷曲线划分到了类别 5 中，即证明了 ISDPC 算法的样本分配策略优于 DPC 算法。

表 1 不同截断距离下 DPC 算法聚类效果

Table 1 Clustering effect of DPC algorithm in different cutoff distance

截断距离	最优类簇数	SC 指标(↑)	DB 指标(↓)
0.245	6	0.470	1.017
0.283	6	0.471	1.017
0.320	7	0.565	0.638
0.358	7	0.562	0.652
0.396	7	0.569	0.638
0.433	7	0.569	0.662
0.471	6	0.421	1.410

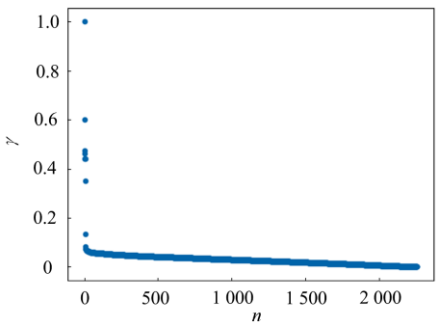


图 6 DPC 算法决策图

Fig. 6 Decision graph of DPC algorithm

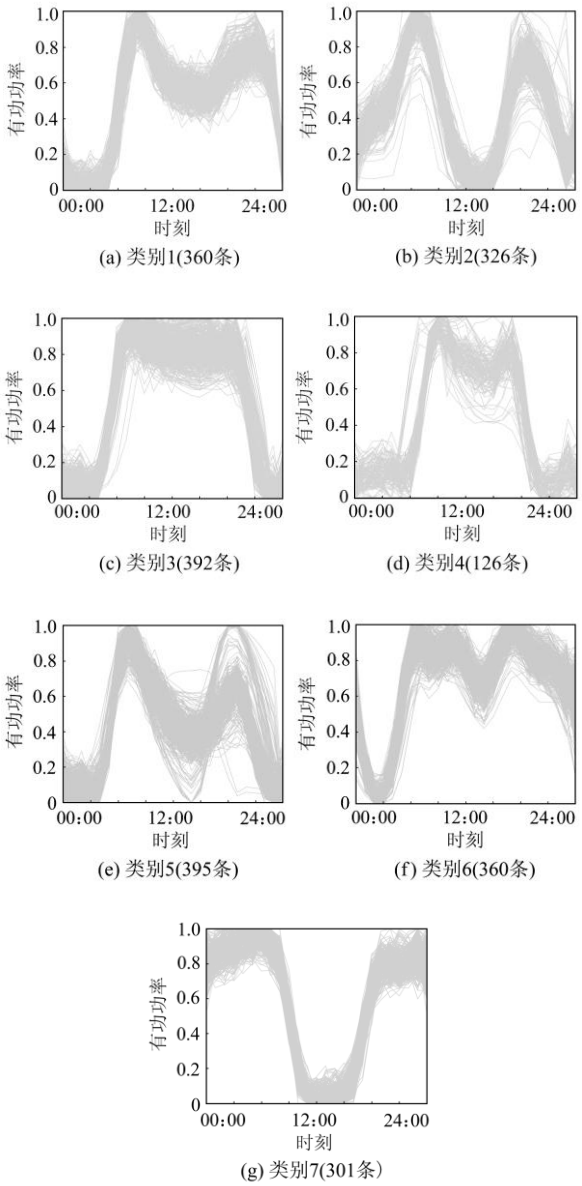


图 7 DPC 算法在真实负荷数据集中的聚类结果

Fig. 7 Clustering results of DPC algorithm in the real load dataset

表 2 为四种算法的聚类效果对比，对于分布均匀、类簇形状相差不大的负荷数据集而言，四种算法均有较好的聚类效果，且 ISDPC 和 k-means 算法的聚类效果略优于另外两种算法。

4.1.2 模拟负荷数据聚类效果

模拟负荷数据集是以 7 类典型日负荷曲线为基础，通过添加的噪声所形成的负荷数据集。对每个类簇设置不同比例的扰动和样本数量得到类簇分布不均衡的模拟负荷数据集，其二维分布如图 8 所示。

表 2 不同算法在真实负荷数据集下的聚类效果对比

Table 2 Clustering comparison of different algorithms for the real load dataset

	最优类簇数	SC 指标(↑)	DB 指标(↓)	运行时间/s
DPC	7	0.569	0.662	5.397
ISDPC	7	0.585	0.606	5.751
k-means	7	0.587	0.608	5.813
FCM	7	0.545	1.025	18.864

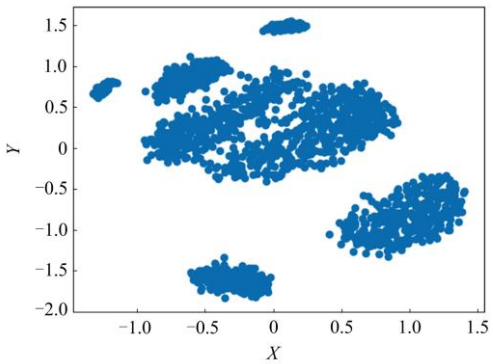


图 8 模拟负荷数据集的二维映射

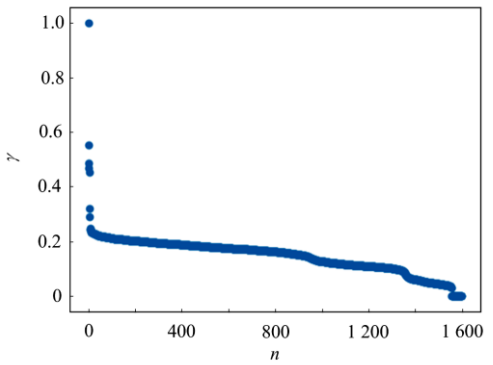
Fig. 8 Two-dimensional mapping of the simulated load dataset

四种算法的聚类效果如表 3 所示，对于不均衡负荷样本的聚类分析，DPC、k-means、FCM 算法的效果均不理想，只有 ISDPC 算法能够准确划分各个类簇，图 9 为在模拟负荷数据集中 ISDPC 算法所得决策图和结果分布图。DPC 算法在不断调整截断距离参数的过程中始终无法得到较好的聚类结果，

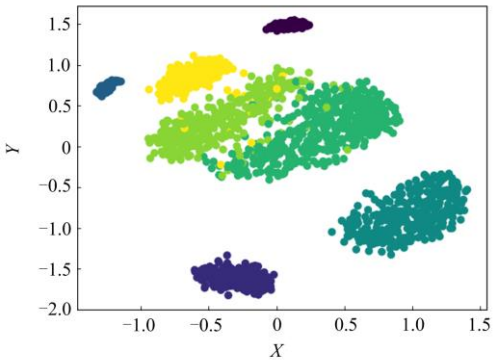
表 3 不同算法在模拟负荷数据集下的聚类效果对比

Table 3 Comparison of the clustering effect of different algorithms in the simulated load dataset

	最优类簇数	SC 指标(↑)	DB 指标(↓)	运行时间/s
DPC	—	—	—	—
ISDPC	7	0.553	0.606	4.914
k-means	3	0.441	0.786	3.887
FCM	2	0.434	0.984	9.644



(a) ISDPC决策图

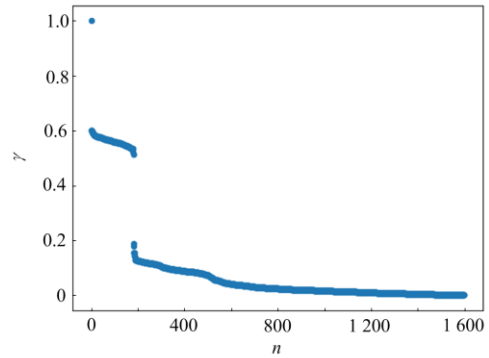


(b) ISDPC聚类结果分布图

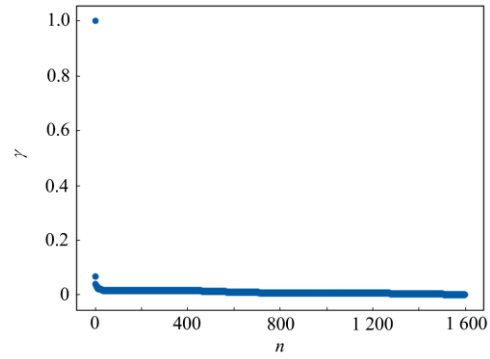
图 9 ISDPC 算法在模拟负荷数据集中的聚类结果

Fig. 9 Clustering results of ISDPC algorithm in the simulated load dataset

出现了以下两种情形：当截断距离较小时，发现的密度峰值点(类簇中心)远超过真实数量，如图 10(a)；当截断距离较大时，仅有 2 个密度峰值点，如图 10(b)。由此可见，当负荷数据集的类簇分布不均衡时，通过固定的截断距离无法准确识别类簇中心；而本文通过改进 DPC 算法的局部密度计算方式，在不设置任何参数的情况下仍然能够准确识别类簇中心。



(a) 较小截断距离下的DPC决策图



(b) 较大截断距离下的DPC决策图

图 10 DPC 算法在不同截断距离下的决策图

Fig. 10 Decision graphs of DPC algorithm with different cutoff distances

4.2 算法性能检验

4.2.1 算法鲁棒性检验

分别使用 ISDPC 和 k-means 算法对不同噪声比例下的模拟负荷数据集进行聚类分析，通过最优类簇数、SC 指标、准确率检验算法的鲁棒性。其中，准确率 T 为分类正确的负荷曲线数与总数之比。表 4 为不同噪声比例 r 下两种方法的结果对比。不难看出，随着噪声比例的增大，SC 指标逐渐下降，最优类簇数目出现偏差，准确率 T 逐渐降低。本文方法在 $r=30\%$ 之前的聚类结果完全正确，鲁棒性较好。而 k-means 算法在 $r=20\%$ 时聚类结果就开始出现偏差，其鲁棒性较差，原因是 k-means 仅以样本之间的欧氏距离作为相似性度量，噪声扰动较大时，会使得相近类别的样本被划分在一起。

表 4 算法鲁棒性对比

Table 4 Comparison of algorithm robustness

噪声比 r/%	ISDPC			k-means		
	最优类簇数	Sihouette 指标	准确率/%	最优类簇数	Sihouette 指标	准确率/%
5	7	0.915	100	7	0.915	100
10	7	0.909	100	7	0.909	100
15	7	0.740	100	7	0.740	100
20	7	0.689	100	6	0.629	87.5
25	7	0.527	100	6	0.478	82.4
30	5	0.479	84.6	4	0.455	78.8
35	6	0.474	83.3	4	0.439	62.5

4.2.2 算法速度比较

在不同规模的模拟负荷数据集下分别执行本文方法、DPC 算法、k-means 算法和 FCM 算法，比较各自的运行时间。如图 11 所示，由于引入了 KNN 和图计算，故本文方法的速度慢于 DPC 算法；其次，本文方法在样本数量小于 40 000 时运行速度略快于 k-means 算法，之后比 k-means 算法慢，但始终远胜于 FCM 算法。

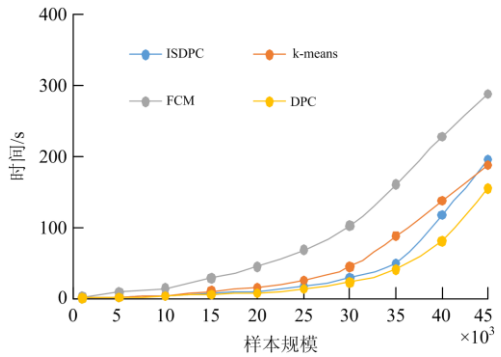


图 11 算法效率比较

Fig. 11 Comparison of algorithm efficiency

5 结论

针对当前的负荷数据集存在类簇形状多样、分布不均衡等问题，本文提出一种改进自适应密度峰值聚类算法。算例结果表明：1) 在不同分布情况的数据集中，本文方法均能够准确划分出各个类簇，且具备自适应能力。2) 相较于传统聚类算法，本文方法在聚类效果、鲁棒性、运算速度等方面均表现出显著的优越性。综上所述，本文方法能够较好地辅助售电公司分析电力用户的用电特性，不需要人为更改任何参数就能够应用于不同的负荷数据样本。

但是本文方法的运算时间随着样本数量的增加呈现指数增长趋势。因此，如何在保证聚类效果不改变的情况下提高算法速度，使其适用于大数据，是接下来的研究内容。

参考文献

[1] DINESH C, MAKONIN S, BAJIĆ I V. Residential power forecasting using load identification and graph spectral clustering[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2019, 66(11): 1900-1904.

[2] SHANG C, GAO J, LIU H, et al. Short-term load forecasting based on PSO-KFCM daily load curve clustering and CNN-LSTM model[J]. IEEE Access, 2021, 9: 50344-50357.

[3] 刘达，雷自强，孙堃. 基于小波包分解和长短期记忆网络的短期电价预测[J]. 智慧电力, 2020, 48(4): 77-83.
LIU Da, LEI Ziqiang, SUN Kun. Short-term electricity price forecasting based on wavelet packet decomposition & long-term and short-term memory networks[J]. Smart Power, 2020, 48(4): 77-83.

[4] LIN S, LI F, TIAN E, et al. Clustering load profiles for demand response applications[J]. IEEE Transactions on Smart Grid, 2019, 10(2): 1599-1607.

[5] 刘鑫，吴红斌，王鲸杰，等. 市场环境下考虑需求响应的虚拟电厂经济调度[J]. 中国电力，2020，53(9): 172-180.
LIU Xin, WU Hongbin, WANG Jingjie, et al. Economic dispatch of a virtual power plant considering demand response in electricity market environment[J]. Electric Power, 2020, 53(9): 172-180.

[6] JIA M, WANG Y, SHEN C, et al. Privacy-preserving distributed clustering for electrical load profiling[J]. IEEE Transactions on Smart Grid, 2021, 12(2): 1429-1444.

[7] 贾雪枫，李存斌. 考虑短期负荷影响的 DeepESN 电力市场实时电价预测研究[J]. 智慧电力，2021，49(1): 64-70.
JIA Xuefeng, LI Cunbin. Real-time electricity price

forecasting of electricity market using deepesn considering short-term load impact[J]. Smart Power, 2021, 49(1): 64-70.

[8] 王继业, 季知祥, 史梦洁, 等. 智能配用电大数据需求分析与应用研究[J]. 中国电机工程学报, 2015, 35(8): 1829-1836.
WANG Jiye, JI Zhixiang, SHI Mengjie, et al. Scenario analysis and application research on big data in smart power distribution and consumption systems[J]. Proceedings of the CSEE, 2015, 35(8): 1829-1836.

[9] 夏成文, 许凯帅, 鲍玉昆, 等. 基于单值聚类分析的区域居民概率负荷预测研究[J]. 电力信息与通信技术, 2021, 19(1): 1-10.
XIA Chengwen, XU Kaishuai, BAO Yukun, et al. Research on probabilistic load forecasting for regional residential users based on single value cluster analysis[J]. Electric Power Information and Communication Technology, 2021, 19(1): 1-10.

[10] 张铁峰, 顾明迪. 电力用户负荷模式提取技术及应用综述[J]. 电网技术, 2016, 40(3): 804-811.
ZHANG Tiefeng, GU Mingdi. Overview of electricity customer load pattern extraction technology and its application[J]. Power System Technology, 2016, 40(3): 804-811.

[11] 靳冰洁, 林勇, 罗澍忻, 等. 基于负荷特性聚类及 Elastic Net 分析的短期负荷预测方法[J]. 中国电力, 2020, 53(9): 221-228.
JIN Bingjie, LIN Yong, LUO Shuxin, et al. A short-term load forecasting method based on load curve clustering and elastic net analysis[J]. Electric Power, 2020, 53(9): 221-228.

[12] 刘季昂, 刘友波, 程明畅, 等. 基于分位数半径动态 K-means 的分布式负荷聚类算法[J]. 电力系统保护与控制, 2019, 47(24): 15-22.
LIU Ji'ang, LIU Youbo, CHENG Mingchang, et al. Distributed load clustering algorithm based on dynamic K-means of quantile radius[J]. Power System Protection and Control, 2019, 47(24): 15-22.

[13] 吴亚雄, 高崇, 曹华珍, 等. 基于灰狼优化聚类算法的日负荷曲线聚类分析[J]. 电力系统保护与控制, 2020, 48(6): 68-76.
WU Yaxiong, GAO Chong, CAO Huazhen, et al. Clustering analysis of daily load curves based on GWO algorithm[J]. Power System Protection and Control, 2020, 48(6): 68-76.

[14] 汪繁荣, 向堃, 刘辉. 基于改进 AP 聚类与优化 GRNN 的非侵入式负荷分解研究[J]. 工程科学与技术, 2020, 52(4): 56-65.
WANG Fanrong, XIANG Kun, LIU Hui. Research on non-intrusive load decomposition based on improved AP clustering and optimized GRNN[J]. Advanced Engineering Sciences, 2020, 52(4): 56-65.

[15] 陈烨, 吴浩, 史俊祎, 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.

[16] 梁京章, 黄星舒, 吴丽娟, 等. 基于 KPCA 和改进 K-means 的电力负荷曲线聚类方法[J]. 华南理工大学学报(自然科学版), 2020, 48(6): 143-150.
LIANG Jingzhang, HUANG Xingshu, WU Lijuan, et al. Clustering method of power load profiles based on KPCA and improved K-means[J]. Journal of South China University of Technology (Natural Science Edition), 2020, 48(6): 143-150.

[17] 黄景林, 彭显刚, 简胜超, 等. 基于深度学习与不平衡样本集的输电线路故障分类[J]. 智慧电力, 2021, 49(2): 114-119.
HUANG Jinglin, PENG Xiangang, JIAN Shengchao, et al. Transmission line fault classification based on deep learning and imbalanced sample set[J]. Smart Power, 2021, 49(2): 114-119.

[18] JIANG Z, LIN R, YANG F, et al. A fused load curve clustering algorithm based on wavelet transform[J]. IEEE Transactions on Industrial Informatics, 2018, 14(5): 1856-1865.

[19] 曹华珍, 吴亚雄, 李浩, 等. 基于海量数据的多维度负荷特性分析系统开发[J]. 电力系统保护与控制, 2021, 49(6): 155-166.
CAO Huazhen, WU Yaxiong, LI Hao, et al. Development of a multi-dimensional load characteristic analysis system based on massive data[J]. Power System Protection and Control, 2021, 49(6): 155-166.

[20] WANG Y, CHEN Q, KANG C, et al. Sparse and redundant representations-based smart meter data compression and pattern extraction[J]. IEEE Transactions on Power Systems, 2016, 32(3): 2142-2151.

[21] RYU S, CHOI H, LEE H, et al. Convolutional autoencoder based feature extraction and clustering for customer load analysis[J]. IEEE Transactions on Power Systems, 2019, 35(2): 1048-1060.

[22] 唐俊熙, 曹华珍, 高崇, 等. 一种基于时间序列数据挖掘的用户负荷曲线分析方法[J]. 电力系统保护与控制, 2021, 49(5): 140-148.

TANG Junxi, CAO Huazhen, GAO Chong, et al. A new user load curve analysis method based on time series data mining[J]. Power System Protection and Control, 2021, 49(5): 140-148.

[23] 宋军英, 崔益伟, 李欣然, 等. 基于欧氏动态时间弯曲距离与熵权法的负荷曲线聚类方法[J]. 电力系统自动化, 2020, 44(15): 87-94.
SONG Junying, CUI Yiwei, LI Xinran, et al. Load curve clustering method based on Euclidean-dynamic time warping distance and entropy weight method[J]. Automation of Electric Power Systems, 2020, 44(15): 87-94.

[24] 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 Wenhu, WU Qinghua, et al. Clustering method of user electricity behavior considering load longitudinal randomness[J]. Electric Power Automation Equipment, 2018, 38(9): 39-44, 53.

[25] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492.

[26] 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.

[27] 陈俊艺, 丁坚勇, 田世明, 等. 基于改进快速密度峰值算法的电力负荷曲线聚类分析[J]. 电力系统保护与控制, 2018, 46(20): 85-93.
CHEN Junyi, DING Jianyong, TIAN Shiming, et al. An improved density peaks clustering algorithm for power load profiles clustering analysis[J]. Power System Protection and Control, 2018, 46(20): 85-93.

[28] 王帅, 杜欣慧, 姚宏民, 等. 面向含多种用户类型的负荷曲线聚类研究[J]. 电网技术, 2018, 42(10): 3401-3412.
WANG Shuai, DU Xinhui, YAO Hongmin, et al. Research on load curve clustering with multiple user types[J]. Power System Technology, 2018, 42(10): 3401-3412.

[29] 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.

[30] 刘君, 余思伍, 陈沛龙, 等. 基于聚类分析的变压器有载分接开关储能弹簧故障识别[J]. 高压电器, 2020, 56(7): 159-165, 172.
LIU Jun, YU Siwu, CHEN Peilong, et al. Fault recognition for on-load tap changer storage spring of power transformer by clustering analysis algorithm[J]. High Voltage Apparatus, 2020, 56(7): 159-165, 172.

[31] HUANG L, YANG Y, ZHAO H, et al. Time series modeling and filtering method of electric power load stochastic noise[J]. Protection and Control of Modern Power Systems, 2017, 2(3): 269-275.

[32] 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.

[33] 徐杰彦, 许雯旸, 褚渊, 等. 区域尺度住宅建筑日用电负荷模型构建方法研究[J]. 中国电力, 2020, 53(8): 29-39.
XU Jieyan, XU Wenyang, CHU Yuan, et al. Residential electricity load model construction in district scale[J]. Electric Power, 2020, 53(8): 29-39.

收稿日期：2021-04-06；　　修回日期：2021-07-01

作者简介：

姚黄金(1996—)，男，硕士研究生，研究方向为负荷画像；E-mail: yaohuangjin@qq.com

雷　霞(1973—)，女，通信作者，博士，教授，研究方向为电力市场、电网规划和调度、电网弹性；E-mail: Snow_lei@mail.xhu.edu.cn

付鑫权(1997—)，男，硕士研究生，研究方向为负荷预测。E-mail: 745034344@qq.com

（编辑　周金梅）