文章编号：1674-0629(2025)09-0150-12 中图分类号：TM714 文献标志码：A
DOI：10. 13648/j. cnki. issn1674-0629. 2025. 09. 014

基于改进密度峰值聚类算法的典型负荷曲线提取

彭晓璐[1]，王涛[1]，卢泽钰[1]，廉杰[1]，赵斌[2]，张谦[2]
（1. 国网冀北电力有限公司唐山供电公司，河北 唐山 063000；2. 输变电装备技术全国重点实验室(重庆大学)，重庆 400044）

摘要：针对现有聚类算法在提取典型负荷曲线时存在的非凸簇识别能力不足和参数敏感性等问题，提出基于改进密度峰值聚类(density peak clustering，DPC)算法的典型负荷曲线提取方法。首先，提出基于局部密度和相对距离的自适应聚类中心选取方法，解决传统DPC算法人为选择聚类中心的主观不确定性问题；其次，定义聚类交叉密度和聚类边界密度两个新参数，提出初始聚类校正策略，有效解决非聚类中心点的分配连带错误问题。通过6个二维数据集、4个多维数据集和1个实际REFIT电气负载测量数据集的对比实验表明，所提改进DPC算法在准确率(ACC)、调整兰德指数(ARI)和Fowlkes-Mallows指数(FMI)3个评价指标上均优于传统DPC、K-means和DBSCAN算法，其中ACC、ARI和FMI平均提升25. 40 %、46. 92 %和21. 83 %。算例结果表明，所提改进DPC算法提取的典型负荷曲线更具代表性，可为电力系统灵活性资源优化调控提供更精准的数据支撑。

关键词：负荷聚类；改进DPC算法；聚类交叉密度；聚类边界密度

Typical Load Profile Extraction Based on Improved DPC Algorithm

PENG Xiaolu[1], WANG Tao[1], LU Zeyu[1], LIAN Jie[1], ZHAO Bin[2], ZHANG Qian[2]
（1. Tangshan Power Supply Company of State Grid Jibei Electric Power Co, Hebei Tangshan 063000, China；2. State Key Laboratory of Power Transmission Equipment Technology Chongqing University, Chongqing 400044, China）

Abstract：Aiming at the existing clustering algorithms′ lack of non-convex cluster identification ability and parameter sensitivity in extracting typical load curves, a typical load curve extraction method based on the improved density peak clustering (DPC) algorithm is proposed. Firstly, an adaptive cluster centre selection method based on local density and relative distance is proposed to solve the subjective uncertainty problem of artificially selecting cluster centres in the traditional DPC algorithm. Secondly, two new parameters of cluster cross-density and cluster boundary density are defined, and an initial cluster correction strategy is proposed to effectively solve the problem of assigning cascading errors to non-cluster centre points. Comparison experiments with six 2D datasets, four multidimensional datasets and one actual REFIT electrical load measurement dataset show that the proposed improved DPC algorithm outperforms the traditional DPC, K-means and DBSCAN algorithms in three evaluation indexes, namely, accuracy (ACC), adjustment of rand index (ARI) and Fowlkes-Mallows Index (FMI), where they are better than the traditional DPC, K-means and DBSCAN algorithms. Among them, ACC, ARI and FMI are improved by 25. 40 %, 46. 92% and 21. 83 % on average. The results show that the typical load curve extracted by the proposed improved DPC algorithm is more representative, which can provide more accurate data support for the optimal regulation of power system flexibility resources.

Key words：load clustering；improved DPC algorithm；cluster crossover density；cluster boundary density

基金项目：国家自然科学基金资助项目(52277081)。
Foundation item：Supported by National Natural Science Foundation of China (52277081).

0　引言

负荷曲线是电力系统运行分析的重要基础数据，其数量庞大且形态复杂，准确描述负荷波动规律对于电力系统的稳定运行和优化调度至关重要[1]。电力系统中的典型负荷曲线反映了负荷在不同时间段内的变化趋势和特征，是电力系统灵活性资源优化调控的关键依据。为准确描述负荷波动规律并支撑灵活性资源优化调控，需要使用聚类算法从海量负荷曲线中提取最具代表性的典型负荷曲线[2]。目前，负荷聚类相关研究主要集中在聚类算法改进、特征提取降维、相似性度量三个方面。

在聚类算法改进方面，文献[3]使用深度差分法求解数据集的最佳聚类数，将K-means算法分别和DBSCAN算法、AHC算法结合，通过对比发现K-means-AHC算法表现最佳。文献[4−5]提出自适应k-means++负荷特性聚类算法，通过迭代图切分自动确定最佳聚类数。文献[6]提出K-medios聚类算法，以聚类半径决定聚类数，在荷载模型参数聚类中效果表现良好。文献[7−8]结合云平台特点和考虑智能计量数据特性改进K-means算法，从而提高其对海量用电数据的聚类效率。文献[9]提出基于动态特性聚类方法，应用于频率安全的在线快速评估。文献[10−11]分别利用高斯混合模型和相似日聚类以实现用户分类与风光功率预测。文献[12]提出基于改进高斯混合模型的变电站负荷聚类算法，增强传统高斯混合聚类算法的计算速度和稳定性。文献[13−14]通过符号聚合近似降维和集成聚类提高预测精度。改进算法虽然优化聚类效果，但仅存在于特定场景，负荷聚类准确性有待提高。

在特征提取降维方面，文献[15]提出SVD-KICIC聚类算法，识别有效的负荷曲线特征，降低数据维度。文献[16]通过多级离散小波变换(discrete wavelet transform，DWT)转换降低负荷曲线维度，优化聚类结果。文献[17]从用户负荷曲线提取多个日负荷特性指标，使用德尔菲法配置权重，实现降维目的。文献[18]提出基于卷积神经网络和自注意力机制的典型运行方式提取方法。文献[19]对多种降维方法研究后发现，集成主成分分析降维的算法性能最佳。文献[20−22]分别利用信息熵分段聚合、参数优化变分模态分解(variational mode decomposition，VMD)和自组织映射神经网络实现降维与聚类。尽管特征提取降维可以提高聚类效率，减轻了维度灾难，但仍未解决基础聚类算法的局限性，特别是负荷曲线的非凸簇识别能力不足的问题。

在相似性度量方面，文献[23]提出基于时空联合聚类方法，通过时间窗移动实现时空联合数据的聚类分析。文献[24]提出基于用户负荷特征的特征优选策略，提高负荷聚类准确率。文献[25]基于形态特征开发自适应分段聚合方法，提高算法速度。文献[26−27]分别结合余弦距离、高斯核函数和欧氏距离度量相似性，实现稳定有效的负荷聚类。文献[28−30]使用动态时间归整法对特征点序列进行度量，提高算法效率并准确反映时间序列特征。文献[31]提出基于时间序列变密度处理的聚类方法，改善负荷聚类效果，真实反映居民用电特性。改进相似性度量虽然增强了聚类算法对特定样本的区分能力，但对于算法的参数敏感性和聚类中心选择的主观性问题仍亟待解决。

综上所述，现有负荷聚类研究主要存在两个问题：一是负荷曲线多样性和用电随机性导致非凸簇形成，多数算法识别能力不足，降低典型负荷曲线代表性；二是虽然DBSCAN算法可识别非凸簇，但参数敏感性高，而密度峰值聚类(density peak clustering，DPC)算法参数简单但需主观选取聚类中心，影响聚类准确性。

针对上述聚类算法在提取典型负荷曲线时存在的非凸簇识别能力不足、参数依赖性强导致聚类结果不稳定的问题，本文提出一种改进的DPC算法。首先，在保留传统DPC算法对非凸簇识别优势的基础上，通过初始聚类中心的自适应选取，有效消除传统DPC算法人为选取聚类中心的主观偏差；其次，设计初始聚类校正机制，有效抑制传统算法非聚类中心分配时的分配连带错误；最后，采用不同算法对多种维度的数据集和实际数据集进行对比验证，结合聚类评价指标对聚类结果进行对比分析，验证了所提改进DPC算法相较于传统DPC算法、K-means和DBSCAN算法在提取实际数据集的典型负荷曲线时具有更高的聚类准确率和稳定性。

1　基于DPC算法的负荷曲线聚类算法研究

DPC算法是一种基于密度的聚类算法，可以识别非凸簇，且对样本密度敏感，能够准确地聚类密

度较高的簇，适用于负荷曲线聚类。

1.1 DPC算法原理分析

DPC算法应用于负荷曲线聚类时，一条n维的日负荷曲线即为一个含n个数据的数据点集合，日负荷曲线的维度指日负荷曲线的采样时段数量，如采样间隔为1 h的日负荷曲线，其采样时段数为24，即其维度为24。由此引出局部密度和相对距离2个变量。

其中，第x条负荷曲线的局部密度（截断核）ρ_x定义式如下。

$$\rho_x = \sum_{x \neq y} \chi\left(d_{x,y} - d_c\right) \tag{1}$$

式中：$d_{x,y}$为第x个数据点和第y个数据点的欧式距离；d_c为截止距离；$\chi(\cdot)$为判断函数。对于较为庞大的数据集，可以用高斯核来计算局部密度，其定义式如下。

$$\rho_x = \sum_{x \neq y} \exp\left(-\frac{d_{x,y}^2}{d_c^2}\right) \tag{2}$$

欧式距离$d_{x,y}$的计算公式为：

$$d_{x,y} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

式中：n为数据维度；x_i和y_i分别为x个数据点和第y个数据点的第i维数据。应用于负荷曲线聚类时，x_i和y_i分别为x条负荷曲线和第y条负荷曲线的第i个时段的负荷值。截止距离d_c的定义式如下。

$$d_c = \mathrm{sort}(d)_{\mathrm{round}(Np\%)} \tag{4}$$

式中：N为数据集的数据点总数；p为百分比参数，通常取1~10；$\mathrm{sort}(d)$表示对数据集所有数据点之间的欧氏距离做升序排序；$\mathrm{round}(Np\%)$表示对$Np\%$四舍五入取整，即对所有欧氏距离升序排序后，取第$\mathrm{round}(Np\%)$的距离值为截止距离。$\chi(m)$的表达式如下：

$$\chi(m) = \begin{cases} 1, m < 0 \\ 0, m \geqslant 0 \end{cases} \tag{5}$$

第x个数据点的相对距离δ_x定义如下。

$$\delta_x = \begin{cases} \max\limits_{y}\left(d_{x,y}\right), \rho_x = \max(\rho) \\ \min\limits_{y:\rho_y > \rho_x}\left(d_{x,y}\right), \rho_x \neq \max(\rho) \end{cases} \tag{6}$$

式中：$\max\left(d_{x,y}\right)$为其他数据点到第x个数据点的最大距离；$\min\left(d_{x,y}\right)$为其他数据点到第x个数据点的最小距离；$\max(\rho)$为第x个数据点的最大局部密度；ρ为局部密度。当第x个数据点的局部密度最大时，则其可能为聚类中心，其相对距离为其他数据点到第x个数据点的最大距离；当第x个数据点的局部密度不是最大时，其相对距离为所有局部密度大于数据点到它的最短距离。

计算所有数据点的局部密度ρ和相对距离δ后，以ρ为横轴、δ为纵轴构建决策图，决策图示例如图1所示。

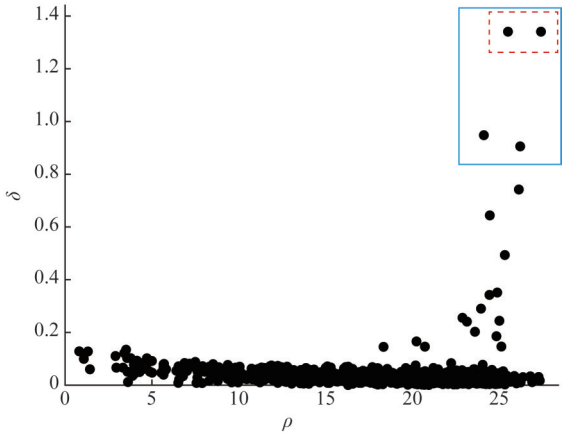


图1 DPC算法决策图示例

Fig. 1 Example of the decision graph for density peaks clustering algorithm

DPC算法需要人为选择局部密度和相对距离较大的数据点作为聚类中心，将剩余的数据点分配给最近的局部密度大于自身局部密度的聚类中心，得到聚类结果。如图1所示，红色实线和蓝色虚线是其中两种选择方式，不同的选择方式将产生不同的聚类数量和聚类结果。因此，DPC算法聚类结果的准确性受使用者的主观影响严重，存在不确定性。同时，在DPC算法中，一旦某个数据点被错误的分配，那么在它截止距离范围内局部密度较低的点将被连带分配到错误聚类中。

1.2 聚类评价指标选取

常用的聚类评价指标有准确率（accuracy，ACC）、精度、兰德指数、调整兰德指数（adjusted rand index，ARI）、召回率和FMI（Fowlkes-Mallows index）等，为了比较不同聚类算法的性能，本文引入3个聚类评价指标：ACC、ARI和FMI，其中ACC反映了样本聚类的正确程度，ARI能够有效排除聚类随机性带来的影响，量化了聚类结果和

随机分配的关系，FMI综合考量了聚类精度和召回率，以避免出现典型负荷曲线识别失误。使用这3个聚类评价指标可以较为全面的评价聚类算法的聚类效果。

ACC的值*Acc*计算公式为：

$$A_{CC}=\frac{\sum_{j=1}^{m} r_j}{n} \tag{7}$$

式中：m为聚类数量；r_j为正确分配到聚类C_j的数据点数量；n为数据点总数；A_{cc}为正确识别的数据点数量与数据点总数的比值。A_{cc}用于衡量聚类算法将数据点分配到聚类中的准确程度，A_{cc}仅考虑样本聚类的正确率，未考虑聚类的精度和召回率。

ARI和FMI的值A_{RI}和F_{MI}计算公式分别为：

$$A_{RI}=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{8}$$

$$F_{MI}=\frac{a}{\sqrt{(a+b)(a+c)}} \tag{9}$$

式中：a为实际和实验情况下均属于同一个聚类的数据点对数目；b为实际情况下属于同一个聚类而实验情况下不属于同一个聚类的数据点对数目；c为实际情况下不属于同一个聚类而实验情况下属于同一个聚类的数据点对数目；d为实际和实验情况下均不属于同一个聚类的数据点对数目。A_{RI}考虑了随机分配的影响，F_{MI}为聚类精度和召回率的几何平均值，二者均用于衡量聚类结果与真实类别之间的相似度。

以上3个聚类评价指标的最大值均为1，3个指标的值越大，表示算法的聚类准确性、聚类精度越好，聚类性能越强。

2　基于改进DPC算法的典型负荷曲线提取研究

通过DPC算法原理分析可知，传统DPC算法存在主观选取聚类中心的不确定性和存在分配连带错误的问题，为解决上述问题，提出初始聚类中心的自适应选取方法，引入了聚类交叉密度和聚类边界密度参数，并基于此提出DPC算法的改进策略。

2.1　改进DPC算法的初始聚类中心自适应选取

作为聚类中心的数据点一般同时具有较大的局部密度和较大的相对距离，因此，为了避免局部密度小而相对距离大的噪声点和局部密度大而相对距离小的边界点被选为聚类中心，本文定义了两个阈值，以提高初始聚类中心选取的准确性。

设$D_{ATA}=\{x_1, x_2, \cdots, x_i\}$为某$n$天的负荷曲线数据，由式(1)和式(3)计算出各个负荷曲线的局部密度和相对距离，并根据式(10)、式(11)和式(12)。确定初始聚类中心。

$$\Omega_\rho=\{x_i|\rho_i\geqslant\overline{\rho}\} \tag{10}$$

$$\Omega_\delta=\{x_i|\delta_i\geqslant\overline{\delta}\} \tag{11}$$

$$\Omega_0=\Omega_\rho\cup\Omega_\delta \tag{12}$$

式中：Ω_ρ为局部密度大于平均局部密度的负荷曲线数据点的集合；Ω_δ为相对距离大于平均相对距离的负荷曲线数据点的集合；x_i为D_{ATA}中的数据点；ρ_i为其对应的局部密度；δ_i为其对应的相对距离；$\overline{\rho}$为所有数据点局部密度的平均值；$\overline{\delta}$为所有数据点相对距离的平均值；Ω_0为初始聚类中心集。

2.2　改进DPC算法初始聚类校正策略

由确定的初始聚类中心可聚类得到核心点集合Ω和初始聚类集合C。实际情况下，一些非核心点的局部密度和相对距离也大于平均值，因此初始聚类中心集合不仅包含实际的核心点还包含一些非核心点，初始聚类并不准确，需要对初始聚类进行校正。

定义两个聚类C_i和C_j的聚类交叉密度为$J_{i,j}$，其定义式如下。

$$J_{i,j}=\max\{\rho_x|x\in Y_i\cap Y_j\} \tag{13}$$

式中：Y_i为和聚类C_i中数据点的距离小于截止距离d_c的数据集；Y_j为和聚类C_j中数据点的距离小于截止距离d_c的数据点的集合。

定义聚类C_i的边界点集合为B_i，如下所示。

$$B_i=\{x_j|x_j=\mathrm{sort}(\rho_j)[:m]\} \tag{14}$$

$$m=\mathrm{ceil}[\mathrm{length}(C_i)\times 20\%] \tag{15}$$

式中：$[:m]$表示取集合前m个元素。m的值由式(15)得到，ceil为向上取整，length(C_i)表示C_i中数据点数量，即聚类C_i的边界点为C_i中局部密度处于较低的20%范围的数据点。

定义两个聚类C_i和C_j的聚类边界密度为$B_{i,j}$，其定义式如下。

$$B_{i,j}=\frac{\overline{\rho(B_i)}+\overline{\rho(B_j)}}{2} \tag{16}$$

式中$\overline{\rho(B_i)}$为聚类C_i的边界点集合B_i中各数据点的局部密度平均值。

对于初始聚类中的 C_i 和 C_j，如果满足 $J_{i,j} > B_{i,j}$，则将聚类 C_i 和 C_j 合并为一个聚类。初始聚类改进策略的原理如下。

如果本应属于同一个聚类中的数据点被错误的分在了两个聚类 C_i 和 C_j 中，如图 2（a）左图所示，其中红色的圆是以某数据点为圆心、截止距离 d_c 为半径的圆，该圆内包含的数据点数量就是该点的局部密度。C_i 和 C_j 的聚类交叉数据点集 $Y_i \cap Y_j$ 即为图 2（a）右图中红色圈数据点的集合，聚类交叉密度 $J_{i,j}$ 为其中局部密度最大数据点的局部密度值。显而易见的，两个聚类中局部密度处于较低的数据点集合即边界点集合 B_i 和 B_j 为右图中蓝色圈数据点集合，B_i 和 B_j 的平均局部密度即聚类边界密度 $B_{i,j}$ 显然小于聚类交叉密度 $J_{i,j}$，故可以判断 C_i 和 C_j 实际上应同属一个聚类。

如果两个聚类 C_i 和 C_j 被正确的划分，如图 2（b）所示，那么这两个聚类的聚类交叉数据点集 $Y_i \cap Y_j$ 应仅包含两个聚类边缘的极少数数据点，或为空集，那么最边缘数据点的局部密度即 $J_{i,j}$ 应小于 $B_{i,j}$，即可判断 C_i 和 C_j 不属于同一个聚类。

改进策略的具体步骤如下。

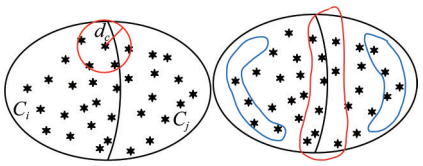
步骤一：根据聚类集合 C 中各聚类间的聚类交叉密度和聚类边界密度，并形成聚类交叉密度矩阵 $\boldsymbol{J}_{p\times p}$ 和聚类边界密度矩阵 $\boldsymbol{B}_{p\times p}$，其中 p 是 C 包含的聚类个数，并定义初始比较次数 $n = 0$。

步骤二：按照聚类中心局部密度降序取 $\boldsymbol{J}_{p\times p}$ 和 $\boldsymbol{B}_{p\times p}$ 中的元素进行比较，例如：假设 Ω 中第 4 个和第 7 个聚类中心是 Ω 中局部密度最高的数据点，则先对 $J_{4,7}$ 和 $B_{4,7}$ 进行比较，若不满足 $J_{4,7} > B_{4,7}$，则取局部密度第一和第三的参数进行比较，依此类推，直到满足 $J_{i,j} > B_{i,j}$ 或所有聚类两两比较完成。

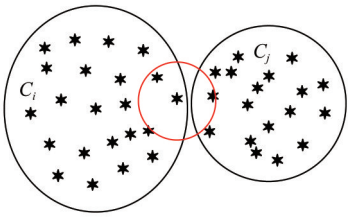
步骤三：若满足 $J_{i,j} > B_{i,j}$，从 Ω 中删除原本的两个聚类中心，重新选择两个聚类中局部密度最大的数据点作为聚类中心添加到 Ω 中，C 中两个聚类合并。返回步骤一，计算新的聚类交叉密度矩阵和聚类边界密度矩阵，并继续循环。若最终仍不满足 $J_{i,j} > B_{i,j}$，即现有 C 中的聚类已经无法合并，停止循环，输出聚类结果。

2.3 改进的DPC算法流程

对DPC算法改进后，实现了初始聚类中心自适应选取和初始聚类校正，其流程如图3所示。



(a) 两个聚类可合并



(b) 两个聚类不可合并

图2 DPC算法初始聚类改进策略示意图

Fig. 2 Schematic diagram of the initial clustering improvement strategy of the DPC algorithm

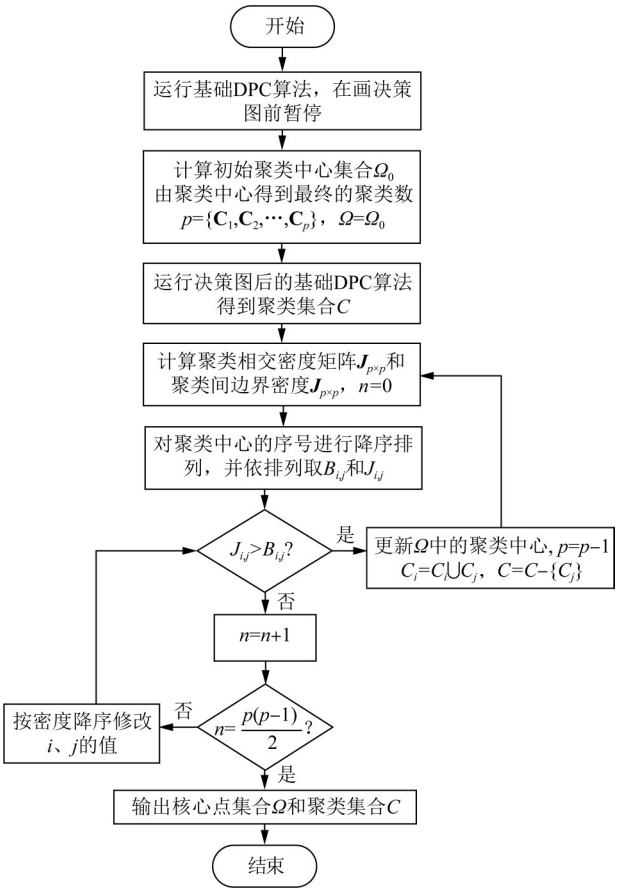


图3 改进的DPC算法流程图

Fig. 3 Flowchart of the improved DPC

2.4 改进DPC算法复杂度分析

本文提出的改进DPC算法的算法复杂度主要由以下6部分组成。

1)计算各数据点的欧式距离，复杂度$O(n^2)$；

2)计算各数据点的局部密度，复杂度$O(n^2)$；

3)计算各数据点的相对距离，复杂度$O(n^2)$；

4)计算选择出初始聚类中心，复杂度$O(n)$；

5)分配各数据点到不同聚类，复杂度$O(n)$；

6)计算聚类交叉密度、聚类边界密度以及对聚类进行合并，复杂度为$O(n^2)$；

综上，本文改进DPC算法的算法复杂度为$O(n^2)$，与传统DPC算法相同，即本文对传统DPC算法的改进没有增加算法的复杂度。同时，选取其他3种较为常见的聚类算法与改进DPC算法对比，其中：K-means算法的复杂度为$O(n)$，DPC算法的复杂度为$O(n^2)$，DBSCAN算法的复杂度为$O(n^2)$。与常见聚类算法相比，改进的DPC算法复杂度位于平均水平。

3　算例分析

采用6个二维数据集、4个多维数据集和1个实际数据集对改进DPC算法和3种常见的先进算法检验对比，3种算法分别为：DPC算法、K-means算法和DBSCAN算法；并采用改进DPC算法对实际数据集进行聚类，有效提取其典型负荷曲线。

3.1　基于多种维度数据集的改进DPC算法性能检验

数据集参数如表1所示。对于不同数据集，各算法的参数设置如表2所示。

表1　数据集参数

Tab. 1　Parameters of data set

数据集	维数	对象数	聚类数
Moons	2	1 000	2
Spiral	2	312	2
Flame	2	240	2
Jain	2	373	2
Aggregation	2	788	7
R15	2	600	15
Iris	4	150	3
Ecoli	7	336	8
Wdbc	30	569	2
Segment	19	2 310	7

其中，6个二维数据集的聚类效果图如图4—9所示。图中每个点代表一个二维数据点，横纵坐标值代表该点在二维坐标系中的位置坐标，无物理含义。

表2　算法参数设置

Tab. 2　algorithmic parameter

数据集	改进DPC	DPC	DBSCAN		K-means
	百分比参数p	百分比参数p	领域半径	领域样本数阈值	聚类数
Moons	2	2	0. 12	2	2
Spiral	2	2	0. 12	1	2
Flame	6	6	0. 9	5	2
Jain	1	0. 5	2. 6	4	2
Aggregation	0. 5	0. 2	1	2	7
R15	2	2	0. 3	4	15
Iris	2. 4	4	0. 7	2	3
Ecoli	0. 3	17	0. 11	4	8
Wdbc	1. 44	4	9. 2	4	2
Segment	0. 3	3	2	3	7

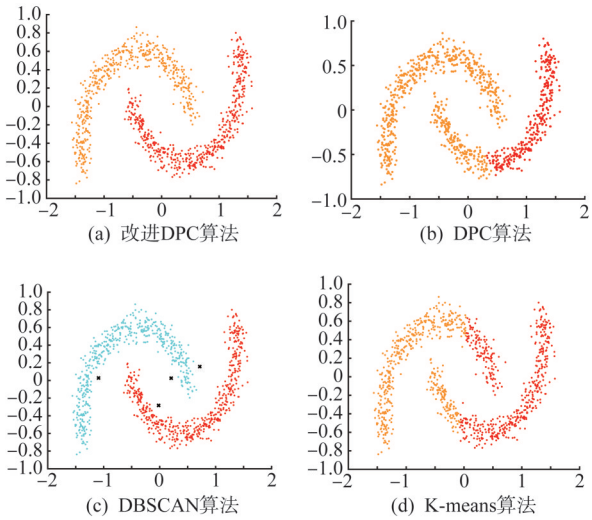


图4　Moons数据集聚类效果图

Fig. 4　Moons dataset clustering effect diagram

由图4可知，改进DPC算法和DBSCAN算法对Moons数据集的聚类较为准确；DPC算法的聚类出现部分错误，其原因是该算法在分配非核心数据点时出现了连带分配错误；K-means算法出现错误的原因是该算法无法有效识别非凸簇。

由图5可知，前3种基于密度的算法均可以准确地对Spiral数据集进行聚类，而K-Means算法无法识别出非凸簇，导致聚类效果不佳。

由图6可知，改进DPC和传统DPC算法对Flame数据集的聚类效果较好，DBSCAN算法错误地将部分数据点识别为噪声，K-means算法聚类效果最差。

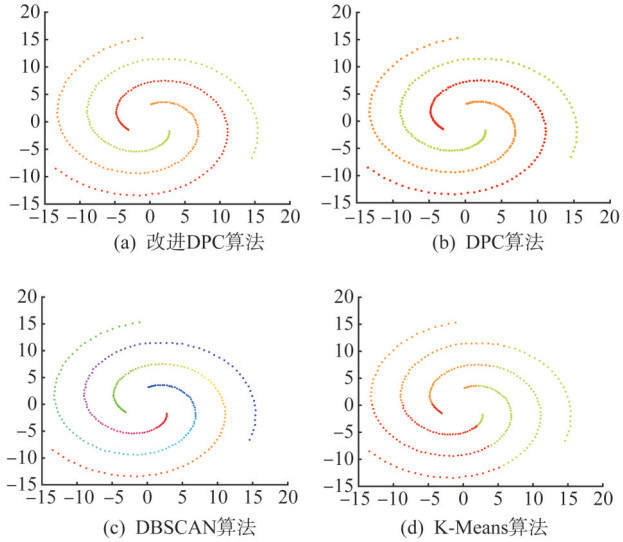


图 5 Spiral 数据集聚类效果图

Fig. 5 Spiral dataset clustering effect diagram

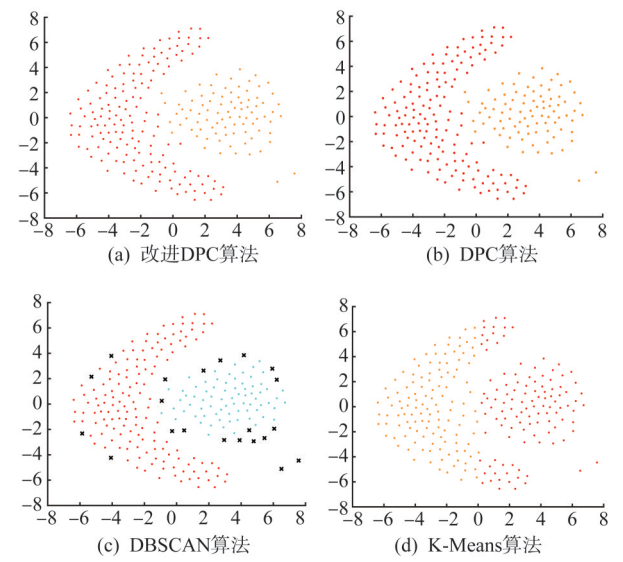


图 6 Flame 数据集聚类效果图

Fig. 6 Flame dataset clustering effect diagram

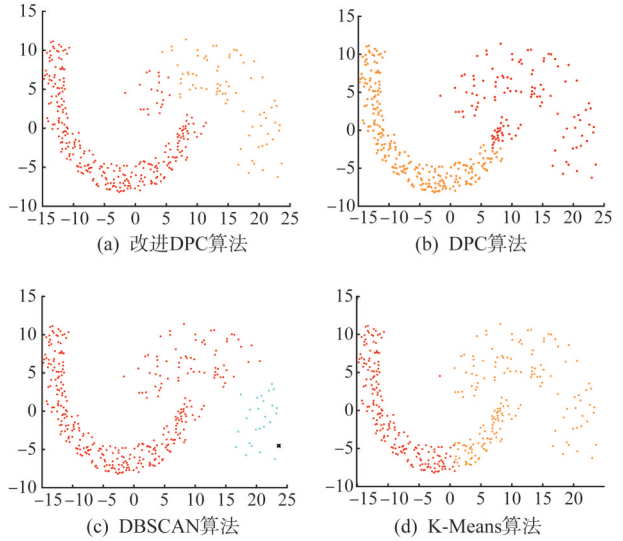


图 7 Jain 数据集聚类效果图

Fig. 7 Jain dataset clustering effect diagram

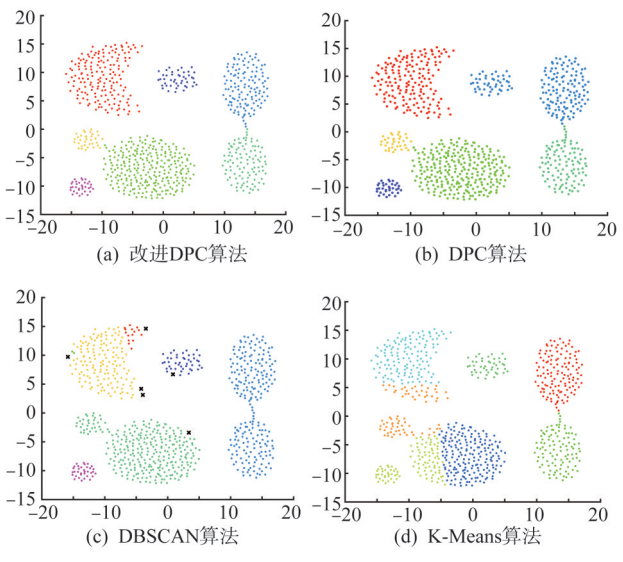


图 8 Aggregation 数据集聚类效果图

Fig. 8 Aggregation dataset clustering effect diagram

由图 7 可知，改进 DPC 算法在分配非核心点时，错误地将部分相对距离更近的另一个聚类中的数据点分配到本聚类中，但聚类效果依然优于其他 3 种算法。

由图 8 可知，改进 DPC 算法对 Aggregation 数据集的聚类效果较好；传统 DPC 算法仅聚类出 6 个簇，原因是在决策图中仅选择了 6 个局部密度和相对距离较大的聚类中心，由于主观不确定性导致聚类效果较差；DBSCAN 算法的聚类效果相对更差，其原因是 DBSCAN 算法的参数设置困难，难以找到适用于该数据集的参数；K-means 算法聚类效果差的原因是无法识别非凸簇。

由图 9 可知，改进 DPC 算法和传统 DPC 算法的聚类效果较好。DBSCAN 算法错误地将部分数据点识别为噪声。4 种方法对不同数据集的聚类效果比较如表 3 所示。

对比图 9 和表 3 可知，改进 DPC 算法对本节中 10 个数据集的聚类效果均好于其他 3 种聚类算法，ACC、ARI 和 FMI 分别比 DPC、DBSCAN 和 K-means 算法平均高出 25.40%、46.92% 和 21.83%。对于 2 维数据集，改进 DPC 算法能保持较高的准确率，对于 Jain 数据集，改进 DPC 算法的 ARI 值未达

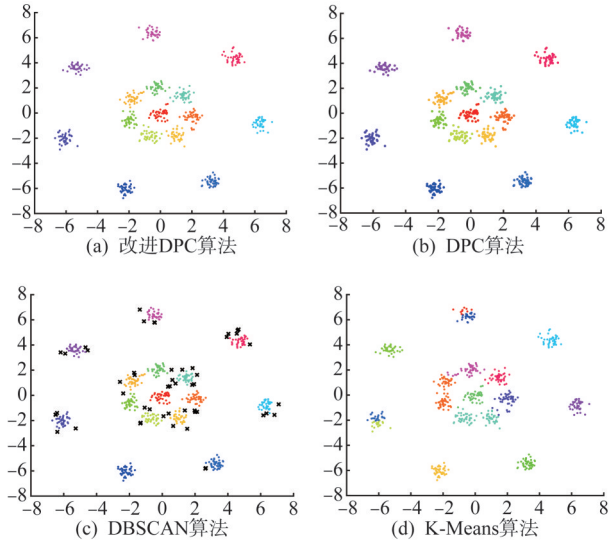


图 9　R15 数据集聚类效果图

Fig. 9　R15 dataset clustering effect diagram

到 0. 9 以上的原因是 Jain 数据集中两类元素在某些区域的密度过于接近，算法难以对不同类的元素进行有效分类；对于高维数据集，改进 DPC 算法显著优于其他 3 种算法。

3. 2　基于实际数据集的改进 DPC 算法性能校验

采用 REFIT 电气负载测量数据集测试 4 种聚类算法的聚类性能，该数据集包含 20 个家庭用户近两年的负荷数据，取样间隔为 8 s。经过数据处理，得到 20 个家庭用户 340 天的负荷数据，并将其取样间隔扩大为 30 min，得到 340×48 条负荷记录，将 20 个家庭用户同时间的负荷相加，作为负荷数据。

REFIT 数据集作为实际数据集，没有聚类标签，为了定量分析改进 DPC 算法对实际数据集的聚类效果，首先选择数据集中负荷较为相似的日负荷曲线 200 条作为测试用的负荷数据集，然后从该组负荷数据集中随机挑选一定比例的负荷数据，人为修改其数据，使其与原负荷数据产生较大差异，并为得到的数据打上聚类标签。

选取 20% 的负荷数据增加其某些时刻的负荷并为其打上“聚类 2”标签，选取 10% 的负荷数据减少其某些时刻的负荷并为其打上“聚类 3”标签，给剩余 70% 未做修改的负荷数据打上“聚类 1”标签，从而模拟得到带聚类标签的实际负荷数据集。修改后的负荷曲线数据如图 10 所示，其中蓝色曲线为“聚类 1”负荷，红色曲线为“聚类 2”负荷，绿色曲线为“聚类 3”负荷。

表 3　聚类效果比较

Tab. 3　Comparison of clustering effects

数据集	评价指标	改进 DPC	DPC	DBSCAN	K-means
Moons	ACC	**1**	0. 752	0. 996	0. 589
	ARI	**1**	0. 556	0. 992	0. 556
	FMI	**1**	0. 632	0. 996	0. 608
Spiral	ACC	**1**	**1**	**1**	0. 343
	ARI	**1**	**1**	**1**	−0. 007
	FMI	**1**	**1**	**1**	0. 330
Flame	ACC	**1**	**1**	0. 968	0. 855
	ARI	**1**	**1**	0. 843	0. 500
	FMI	**1**	**1**	0. 925	0. 758
Jain	ACC	**0. 943**	0. 923	0. 806	0. 788
	ARI	**0. 767**	0. 711	0. 256	0. 325
	FMI	**0. 916**	0. 869	0. 801	0. 713
Aggregation	ACC	**0. 999**	0. 997	0. 798	0. 863
	ARI	**0. 998**	0. 995	0. 776	0. 733
	FMI	**0. 998**	0. 995	0. 835	0. 792
R15	ACC	**0. 997**	**0. 997**	0. 925	0. 806
	ARI	**0. 992**	**0. 992**	0. 855	0. 815
	FMI	**0. 993**	**0. 993**	0. 866	0. 832
Iris	ACC	**0. 961**	0. 907	0. 688	0. 819
	ARI	**0. 887**	0. 758	0. 562	0. 602
	FMI	**0. 922**	0. 840	0. 760	0. 732
Ecoli	ACC	**0. 712**	0. 503	0. 522	0. 515
	ARI	**0. 671**	0. 362	0. 338	0. 253
	FMI	**0. 783**	0. 517	0. 524	0. 434
Wdbc	ACC	**0. 896**	0. 790	0. 618	0. 833
	ARI	**0. 629**	0. 317	−0. 015	0. 440
	FMI	**0. 823**	0. 756	0. 711	0. 736
Segment	ACC	**0. 572**	**0. 572**	0. 148	0. 402
	ARI	**0. 409**	0. 323	0. 000	0. 218
	FMI	**0. 549**	0. 465	0. 373	0. 330

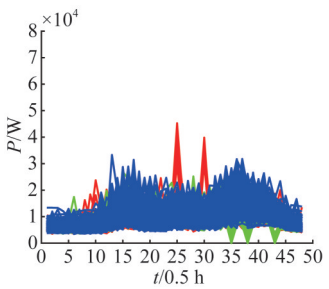


图 10　修改后的负荷数据曲线

Fig. 10　Modified load data curves

采用4种算法对修改后的负荷曲线进行聚类，4种算法的参数设置如表4所示，参数设置方法参照文献[32]。

表4 算法参数设置

Tab. 4 Algorithm parameters settings

	改进DPC	DPC	DBSCAN	K-means
参数设置	p=5	p=2	领域半径=5 500 邻域样本数阈值=8	K=3

得到的结果如图11和表5所示。其中，红色、蓝色和绿色曲线分别是被聚为同一类的曲线，黑色曲线是被DBSCAN算法识别为噪声的曲线，在图(c)中用黑色方框标出。

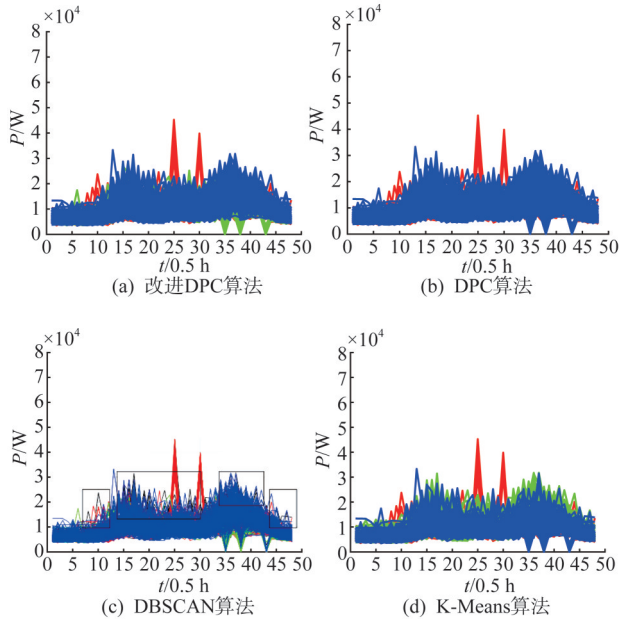


图11 4种算法的聚类结果

Fig. 11 Clustering results of 4 algorithms

由图11可知，改进DPC算法的聚类效果最优异，其聚类结果与标签差异很小；DBSCAN算法可以较为准确地聚类出“聚类2”曲线，但是将部分“聚类3”曲线错误地分配到了“聚类1”中，且将部分曲线错误地识别为噪声；同样，K-means算法也能较好地聚类出“聚类2”曲线，但是将“聚类1”曲线和“聚类3”曲线混杂；DPC算法将所有“聚类3”曲线错误地分配到“聚类1”中，主要原因是在决策图中，主观选择聚类中心的局限性较大。

普通DPC算法和改进DPC算法在该次聚类中在决策图中选择的聚类中心如图12所示，实际情况下，改进DPC算法不会产生决策图，此处为作进一步说明而绘制了决策图。

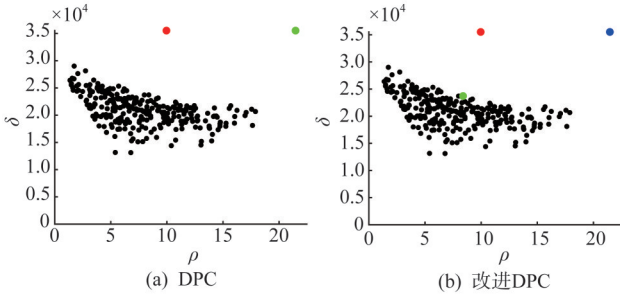


图12 2种DPC算法选择的聚类中心在决策图中的对比

Fig. 12 Comparison of clustering centers selected by 2 DPC algorithms in decision graph

由于决策图中存在局部密度ρ和相对距离δ都较大的2个点，在数据集实际的聚类数未知时，DPC算法通常仅会选择该2点作为聚类中心，使含3种类别的数据集被错误的聚为2类。改进DPC算法不需要在决策图上主观选择聚类中心，避免了主观选择的不确定性。

表5 4种算法的聚类结果

Tab. 5 Clustering results of 4 algorithms

算法	参数	ACC值	ARI值	FMI值
改进DPC	0.05	**0.993**	**0.977**	**0.989**
DPC	0.05	0.899	0.710	0.890
DBSCAN	5 500/8	0.849	0.755	0.878
K-means	3	0.694	0.391	0.664

由表5可知，对于REFIT电气负载测量数据集，改进DPC算法的ACC、ARI和FMI聚类评价指标值均最大，尤其是ARI的值远远高于其他算法，即改进DPC算法的聚类结果与标签值的相似程度最高，其聚类效果优于其他3种算法。由于K-means算法无法识别非凸簇，且在高维数据集上表现较差，仅将欧氏距离相对较近的数据点简单地识别为同一类，而不考虑各数据点在高维空间的具体布局，因此聚类效果最差，其ACC、ARI和FMI在4种算法中最低；DBSCAN算法可以识别非凸簇，但其参数设置困难，且将部分正常曲线错误地识别为异常曲线，导致聚类效果较差；传统DPC算法存在主观不确定性从而仅将数据集分为2类，导致聚类效果较差，传统DPC算法与DBSCAN算法相比，ACC和FMI较高，但ARI较低，其原因是

DBSCAN算法将数据集分为3类，聚类结果与标签的相似程度高于传统DPC；改进DPC算法自适应选择了正确的聚类中心，其聚类效果最优。

综上可知，本文提出的改进DPC算法在高维数据的聚类上优于其他算法，且其对于传统DPC算法的最大优势在于其能够自动选取聚类中心，有效避免了主观选取聚类中心的不确定性。

3.3　典型负荷曲线提取有效性验证

使用4种算法对REFIT电气负载测量数据集聚类，聚类结果如表6所示。

表6　算法参数和聚类结果

Tab. 6　Algorithm parameters and clustering results

	改进DPC	DPC	DBSCAN	K-means
参数设置	p=5	p=2	领域半径=5 500 邻域样本数阈值=8	K=3
典型负荷曲线集合数量	278	335	274	137

由表6可知，改进DPC算法和DBSCAN算法的典型负荷曲线集合数量较为合理，约占总曲线数量的80%左右，该数量的曲线可代表负荷在大部分时间的用电情况。而DPC算法将98%以上的曲线都划分进典型负荷曲线集合，比例过高，K-means算法只将40%曲线划分为典型负荷曲线，比例过低，均不能有效代表的典型用电情况。

各算法得到的最大聚类集合即为节点的典型负荷曲线集合。对于改进DPC算法和传统DPC算法，取最大聚类的聚类中心作为典型负荷曲线；对于DBSCAN算法，取最大聚类中密度最大的数据点作为典型负荷曲线；对于K-means算法，取和最大聚类质心最近的数据点作为典型负荷曲线。4种算法得到的典型负荷曲线及其集合如图13所示。

对比4种算法聚类出的典型负荷曲线集合可知，改进DPC算法聚类出的典型负荷曲线集中少有含有极端负荷的特殊负荷曲线，而传统DPC算法和DBSCAN算法聚类出的典型负荷曲线集中，包含大量突出的极端负荷曲线，聚类效果不佳；K-means算法聚类出的典型负荷曲线集合中曲线数量仅有137条，将很多本该属于典型负荷曲线集合的曲线分配到了其他集合中，导致典型负荷曲线集合的特征缺失，因此提取出的典型负荷曲线不准确。

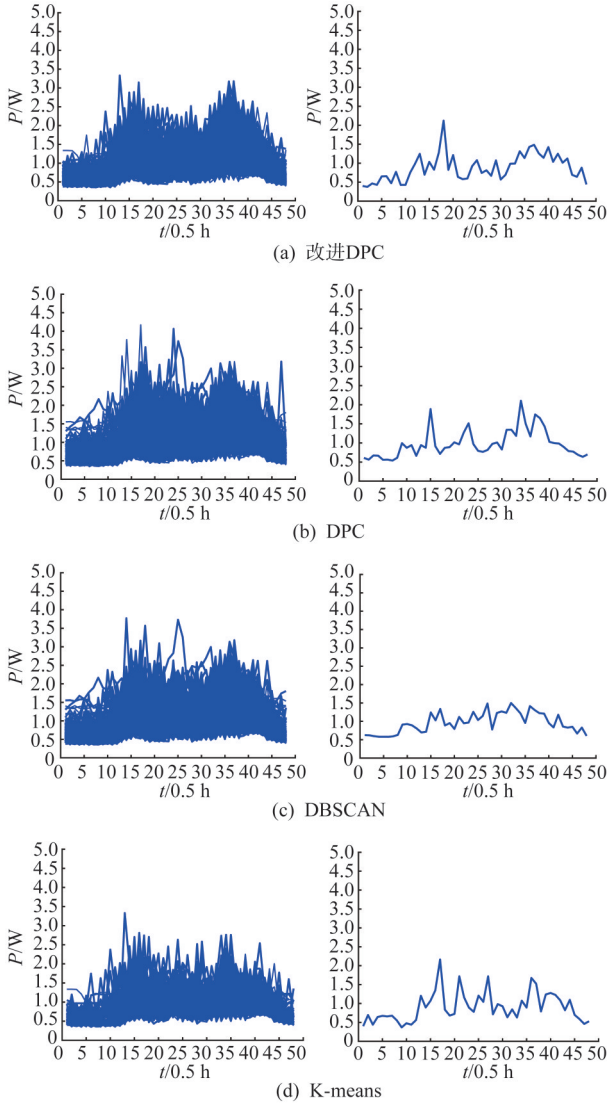


图13　典型负荷曲线及其集合

Fig. 13　Typical load curves and their aggregation

改进DPC算法提取出的典型负荷曲线通过对大量负荷曲线聚类分析得出，能够代表大部分负荷曲线的特征。在实际应用中，电力系统运行通常具有一定的规律性和稳定性，大部分时间的负荷曲线较为相似，因此提取出一条具有代表性的典型负荷曲线对于电力系统灵活性资源优化调控等应用具有重要意义。综上所述，本文所提改进DPC算法相较于其他几种聚类算法，提取出的典型负荷曲线准确性更高，更具代表性。

4　结论

针对DPC算法选取聚类中心的主观不确定性和数据点的分配连带错误等问题，本文提出了基于改

进DPC算法的典型负荷曲线提取方法。首先，提出自适应确定初始聚类中心的方法，进而使用DPC算法聚类，得到初始聚类的集合。其次，提出初始聚类中心改进策略，通过比较两个参数的大小校正初始聚类。最后，使用改进DPC算法和其他常见聚类算法对多种维度的数据集和实际数据集进行聚类，并对结果进行对比验证，得到的提升效果如下。

1)所提改进DPC算法可以自适应确定初始聚类中心，并通过初始聚类改进策略确定最终的聚类中心，克服了选取聚类中心的主观不确定性问题。

2)所提改进DPC算法在多种维度的数据集验证中，聚类评价指标的值均优于其他算法，其中ACC、ARI和FMI分别平均比DPC、DBSCAN和K-means算法高出25.40%、46.92%和21.83%，验证了改进DPC算法可提升典型负荷曲线的非凸簇识别能力。

3)在实际REFIT电气负载测量数据集的应用中，所提改进DPC算法对实际数据集进行聚类，可有效提取出更具代表性的典型负荷曲线，为电力系统灵活性资源优化调控提供了更精准和可靠的数据支撑。

参考文献

[1] 孟琰，肖居承，洪居华，等. 计及需求响应不确定性的节点负荷准线：概念与模型[J]. 电力系统自动化，2023，47(13)：28－39.
MENG Yan，XIAO Jucheng，HONG Juhua，et al. Nodal customer directrix load considering demand response uncertainty：concept and model[J]. Automation of Electric Power Systems，2023，47(13)：28－39.

[2] 孙毅，毛烨华，李泽坤，等. 面向电力大数据的用户负荷特性和可调节潜力综合聚类方法[J]. 中国电机工程学报，2021，41(18)：6259－6271.
SUN Yi，MAO Yehua，LI Zekun，et al. A Comprehensive clustering method of user load characteristics and adjustable potential based on power big data[J]. Proceedings of the CSEE，2021，41(18)：6259－6271.

[3] 张文韬. 基于改进K-means聚类的配电网负荷预测方法研究[D]. 成都：电子科技大学，2023.

[4] 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.

[5] 闵永智，郝大宇，王果，等. 基于深度自适应K-means++算法的电抗器声纹聚类方法[J]. 电力系统保护与控制，2025，53(8)：1－13.
MIN Yongzhi，HAO Dayu，WANG Guo，et al. Reactor voiceprint clustering method based on deep adaptive K-means++algorithm[J]. Power System Protection and Control，2025，53(8)：1－13

[6] ZHANG X R，HILL D. Clustering of uncertain load model parameters with k-medoids algorithm[C]//IEEE Power and Energy Society General Meeting PESGM，August 05－10，2018，Portland，OR，USA. New York：IEEE，2018：108－123.

[7] 张素香，刘建明，赵丙镇. 等. 基于云计算的居民用电行为分析模型研究[J]. 电网技术，2013，37(6)：1542－1546.
ZHANG Suxiang，LIU Jianming，ZHAO Bingzhen，et al. Cloud computing-based analysis on residential electricity consumption behavior[J]. Power System Technology，2013，37(6)：1542－1546.

[8] LI R，SMITH N，LI F R. Multi-resolution load profile clustering for smart metering data[J]. IEEE Transactions on Power Systems，2016，31(6)：4473－4482.

[9] 鲍颜红，吴峰，任先成，等. 基于动态特性节点聚类的低阶仿真频率安全分析方法[J]. 南方电网技术，2023，17(3)：47－55，64.
BAO Yanhong，WU Feng，REN Xiancheng，et al. Low-order simulation frequency security analysis method based on dynamic characteristic node clustering[J]. Southern Power System Technology，2023，17(3)：47－55，64.

[10] HINO H，SHEN H Y，MURATA N，et al. A versatile clustering method for electricity consumption pattern analysis in households[J]. IEEE Transactions on Smart Grid，2013，4(2)：1048－1057.

[11] 王凌梓，沈海波，邓力源，等. 基于相似日聚类与WOA-BiLSTM-Copula算法的短期风光功率相关性概率区间预测方法[J/OL]. 南方电网技术：1－8[2024－12－12]. https://nfdwjs.csg.cn/gateway-web/en/debutDetail.html？serialNum=20240201001.
WANG Lingzi，SHEN Haibo，DENG Liyuan，et al. Short-term wind and photovoltaic power correlation probability interval prediction method based on similar day clustering and WOA-BiLSTM-Copula algorithm[J/OL]. Southern Power System Technology：1－8[2024－12－12]. https://nfdwjs.csg.cn/gateway-web/en/debutDetail.html？serialNum=20240201001.

[12] 余浩，高镱滴，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.

[13] WANG Y，CHEN Q X，KANG C Q，et al. Clustering of electricity consumption behavior dynamics toward big data applications[J]. IEEE Transactions on Smart Grid，2016，7(5)：2437－2447.

[14] 李一，杨茂，苏欣. 基于集成聚类及改进马尔科夫链模型的光伏功率短期预测[J]. 南方电网技术，2023，17(10)：113－122.
LI Yi，YANG Mao，SU Xin. Short-term prediction of photovoltaic power based on integrated clustering and improved Markov chain model[J]. Southern Power System Technology，2023，17(10)：113－122.

[15] ZHANG Y K，ZHANG J，YAO G，et al. Method for clustering daily load curve based on SVD-KICIC[J]. Energies，2020，

13(17)：24－42.

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

[17] 刘思，李林芝，吴浩，等. 基于特性指标降维的日负荷曲线聚类分析[J]. 电网技术，2016，40(3)：797－803.
LIU Si，LI Linzhi，WU Hao，et al. Cluster analysis of daily load curves using load pattern indexes to reduce dimensions[J]. Power System Technology，2016，40(3)：797－803.

[18] 付小标，姜旭，李欣蒙，等. 基于卷积自注意力聚类算法的高比例新能源电力系统典型运行方式提取[J/OL]. 南方电网技术：1－10[2025－03－03]. https://nfdwjs.csg.cn/gateway-web/en/debutDetail.html? serialNum=20240924001.
FU Xiaobiao，JIANG Xu，LI Xinmeng，et al. Typical operation mode extraction of high proportion new energy power system based on convolutional self-attention clustering algorithm[J/OL]. Southern Power System Technology：1－10[2025－03－03]. https://nfdwjs. csg. cn/gateway-web/en/debutDetail.html? serialNum=20240924001.

[19] 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，(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，(15)：3741－3749.

[20] 林顺富，田二伟，符杨，等. 基于信息熵分段聚合近似和谱聚类的负荷分类方法[J]. 中国电机工程学报，2017，37(8)：2242－2252.
LIN Shunfu，TIAN Erwei，FU Yang，et al. Power load classification method based on information entropy piecewise aggregate approximation and spectral clustering[J]. Proceedings of the CSEE，2017，37(8)：2242－2252.

[21] 王建元，张宇辉，刘铖. 基于参数优化VMD和改进K聚类判据融合的配电网故障选线方法[J]. 南方电网技术，2023，17(7)：135－145.
WANG Jianyuan，ZHANG Yuhui，LIU Cheng. Fault line selection method of distribution network based on the fusion of parameter optimized variational modal decomposition and improved k clustering criterion [J]. Southern Power System Technology，2023，17(7)：135－145.

[22] 李智勇，吴晶莹，吴为麟，等. 基于自组织映射神经网络的电力用户负荷曲线聚类[J]. 电力系统自动化，2008，32(15)：66－70，78.
LI Zhiyong，WU Jingying，WU Weilin，et al. Power customers load profile clustering using the SOM neural network[J]. Automation of Electric Power Systems，2008，32(15)：66－70，78.

[23] 陈佳俊，陈玉峰，严英杰，等. 基于时空联合聚类方法的输变电设备状态异常检测[J]. 南方电网技术，2015，9(11)：65－72.
CHEN Jiajun，CHEN Yufeng，YAN Yingjie，et al. Anomaly detection of state information of power equipment based on spatiotemporal clustering[J]. Southern Power System Technology，2015，9(11)：65－72.

[24] 陆俊，朱炎平，彭文昊，等. 智能用电用户行为分析特征优选策略[J]. 电力系统自动化，2017，41(5)：58－63，83.
LU Jun，ZHU Yanping，PENG Wenhao，et al. Feature selection strategy for electricity consumption behavior analysis in smart grid[J]. Automation of Electric Power Systems，2017，41(5)：58－63，83.

[25] 王潇笛，刘俊勇，刘友波，等. 采用自适应分段聚合近似的典型负荷曲线形态聚类算法[J]. 电力系统自动化，2019，43(1)：110－118.
WANG Xiaodi，LIU Junyong，LIU Youbo，et al. Shape clustering algorithm of typical load curves based on adaptive piecewise aggregate approximation[J]. Automation of Electric Power Systems，2019，43(1)：110－118.

[26] 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.

[27] 蒋正邦，吴浩，程祥，等. 基于多元聚类模型与两阶段聚类修正算法的变电站特性分析[J]. 电力系统自动化，2018，42(15)：157－163.
JIANG Zhengbang，WU Hao，CHENG Xiang，et al. Analysis of substation characteristics based on multivariate clustering model and two-stage clustering-correction algorithm [J]. Automation of Electric Power Systems，2018，42(15)：157－163.

[28] WANG X Y，MUEEN A，DING H，et al. Experimental comparison of representation methods and distance measures for time series data [J]. Data Mining and Knowledge Discovery，2013，26：275－309.

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

[30] 肖辉，胡运发. 基于分段时间弯曲距离的时间序列挖掘[J]. 计算机研究与发展，2005，42(1)：72－78.
XIAO Hui，HU Yunfa. Data mining based on segmented time warping distance in time series database [J]. Journal of Computer Research and Development，2005，42(1)：72－78.

[31] 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.

[32] 谭鸿伟，吕莉，郝筱萱，等. 基于近邻与代表点的密度峰值聚类算法[J/OL]. 控制工程：1－11[2024－05－10]. https://doi.org/10.14107/j.cnki.kzgc.20240081.
TAN Hongwei，LÜ Li，HAO Xiaoxuan，et al. Density peaks clustering algorithm based on nearest neighbors and representative points[J/OL]. Control Engineering of China：1－11[2024－05－10]. https://doi.org/10.14107/j.cnki.kzgc.20240081.

收稿日期：2025-05-07；**网络首发日期**：2025-08-20

作者简介：

彭晓璐(1989)，女，高级工程师，硕士，研究方向为电力系统及其自动化；

王涛(1985)，男，高级工程师，硕士，研究方向为电网经营；

张谦(1980)，女，通信作者，教授，博士，研究方向为综合能源系统优化调度，zhangqian@cqu. edu. cn。