DOI: 10.7667/PSPC171386

基于改进快速密度峰值算法的电力负荷曲线聚类分析

陈俊艺[1]，丁坚勇[1]，田世明[2]，卜凡鹏[2]，朱炳翔[1]，黄事成[1]，周 凯[1]

(1. 武汉大学电气工程学院，湖北 武汉 430072；2. 中国电力科学研究院，北京 100192)

摘要：为解决传统聚类算法对大数据背景下高维海量、类簇形状差异巨大的电力负荷曲线进行聚类分析时存在的聚类结果不稳定、聚类效果较差、聚类速度慢和内存消耗过大等问题，提出一种改进的快速密度峰值聚类算法。首先应用主成分分析法对归一化后的负荷曲线集进行降维处理，以减少样本向量间欧式距离的计算量和加快后续操作。然后利用 kd 树算法对降维后的数据进行快速 K 近邻搜索生成 KNN 矩阵。最后以 KNN 矩阵代替原算法的距离矩阵作为输入数据。在基于 KNN 改进的样本局部密度和距离计算准则的基础上，运用快速密度峰值算法对负荷曲线进行聚类分析。通过实验和算例分析验证了所提改进算法的实用性和有效性。

关键词：电力大数据；负荷曲线聚类；快速密度峰值算法；主成分分析；kd 树；KNN 算法

An improved density peaks clustering algorithm for power load profiles clustering analysis

CHEN Junyi[1], DING Jianyong[1], TIAN Shiming[2], BU Fanpeng[2], ZHU Bingxiang[1], HUANG Shicheng[1], ZHOU Kai[1]

(1. School of Electrical Engineering, Wuhan University, Wuhan 430072, China;
2. China Electric Power Research Institute, Beijing 100192, China)

Abstract: Aiming at the problems of poor stability of clustering results, poor effectiveness in clustering, slow speed and high memory consumption when making traditional clustering analysis for a large dimensionality huge number of load profiles with huge difference between the clusters under the background of the big data, an improved density peaks clustering algorithm is proposed. Firstly, principle components analysis method is used to reduce dimensions of load curves after normalization in order to reduce the calculation of the Euclidean distance between the sample vectors and to speed up the subsequent operations. Then, the kd tree algorithm is used to carry out the fast k-nearest neighbor search to generate KNN matrix. Finally, the KNN matrix is used to replace the original distance matrix as the input data. Based on the KNN improved local density and distance calculation criterion, the density peaks clustering algorithm is used to cluster the load profiles. Experiments and case analysis show that the proposed method is practicable and effective.

This work is supported by National High-tech R & D Program of China (863 Program) (No. 2015AA050203).

Key words: power big data; load profiles clustering; density peaks clustering algorithm; PCA; kd tree; KNN algorithm

0 引言

随着智能配电网信息化、自动化、互动化程度的不断提高[1]，各种先进的配用电自动化和管理系统得到广泛的应用，由此产生并积累了海量多源异构数据[2-3]。对这些数据进行有效挖掘和科学合理地利用，可以有效提升智能配电网的运行管理水平，同时也是大数据背景下电力企业发展的必然要求。

负荷曲线聚类分析是配用电领域的基础，在配用电大数据挖掘中占据非常重要的地位。在负荷预测研究中，通常先对历史负荷数据或气象数据等进行聚类分析，根据聚类分析结果提取与待预测日相似的训练样本集来构建预测模型[4]。与传统的按电价、行业等用户分类方法不同，通过对用户的负荷曲线聚类分析，可以更好地将不同用电模式的用户区分开，为开展电价制定与需求侧响应提供基础条件[5]。在异常用电检测方面，主要通过负荷曲线聚类技术提取典型负荷模式来检测用户是否存在异常用电行为[6]。因此，研究适用于当今大数据环境的负荷曲线聚类技术具有重要意义。

基金项目：国家高技术研究发展计划(863 计划)(2015AA050203)；国家电网公司科技项目“智能配用电大数据应用关键技术深化研究”

传统的聚类算法包括基于划分、基于层次、基于密度、基于模型和基于网格的聚类算法[7]等。基于划分的聚类算法，如 k-means[8]，虽然运行速度非常快，但存在以下缺陷：确定初始聚类中心和聚类数目的理论仍不完善，只能发现球状的类簇，在迭代过程中易陷入局部最优，易受离群点和噪声点影响。基于层次的聚类算法[9]的时空复杂度过高，无法应用于大数据环境，而且难以选择子簇的合并或分裂点，导致其可伸缩性差。基于密度的聚类算法 DBSCAN[10]虽然可以发现任意形状的类簇，但它的两个关键参数 Eps(邻域半径)和 MinPts(任意簇中数据对象的最小数目)难以确定，设置稍有不当就会导致聚类效果不尽人意。基于模型的聚类算法，如期望最大化算法 EM[11]，是基于概率分布规律的，但大多数情况下我们无法得知数据真实的概率分布，并且由于其需要进行大量迭代求解，对大数据的处理非常慢。基于网格的聚类算法[12]一般用于处理空间数据，不适用于负荷曲线这种时间序列数据。

为了弥补传统聚类方法的不足，提高负荷曲线聚类的准确性和实用性，许多学者提出了大量的改进算法。文献[13-14]分别引入模拟退火遗传算法和量子粒子群算法，克服模糊 C 均值聚类算法(Fuzzy C-means Clustering, FCM)易陷入局部最优及易受初始聚类中心影响的不足，较好地改善了 FCM 的聚类效果。文献[15]提出一种结合主成分分析降维的集成聚类算法对负荷曲线进行聚类，效果优于传统单一聚类算法。文献[16]先利用 KPCA 对数据降维，再利用 Kernel k-means 对降维后的数据聚类，有效提高了负荷曲线聚类的准确性。然而，运用这些算法对大数据背景下高维海量、类簇形状差异巨大的负荷曲线进行聚类分析时，聚类结果稳定性、聚类效果、聚类速度和内存消耗等方面仍不甚理想。

2014 年，一种称为快速搜索与发现密度峰值聚类算法(Clustering by Fast Search and Find of Density Peaks, CFSFDP)[17]发表于 Science 期刊。这种新型聚类算法能发现任意形状的类簇，同时能检测出离群点，可以自动确定聚类数目，在对样本数据进行类簇分配时无需进行迭代，只需一步即可完成分配，能快速对高维海量数据进行聚类分析，聚类结果稳定且效果好。因此，该算法一经提出便被应用于众多领域，取得了很好的效果。与此同时，一些学者针对该算法存在的问题进行了相应的改进[18-21]，使得该算法适应性更强。

鉴于此，结合大数据背景下电力负荷曲线具有高维海量、类簇形状差异巨大等特点，本文提出一种改进的快速密度峰值聚类算法(I-CFSFDP)。该改进算法利用 K 近邻(K-Nearest Neighbors, KNN)的思想改进了原算法的样本局部密度和距离计算准则。首先利用主成分分析(Principal Component Analysis, PCA)对归一化的负荷数据降维处理，然后用 kd 树对降维后的数据进行快速 K 近邻搜索生成 KNN 矩阵，最后以 KNN 矩阵代替原算法的距离矩阵作为输入，在新的局部密度和距离计算准则基础上运用快速密度峰值算法进行负荷曲线聚类分析。实验和算例分析表明，相比于原算法和传统聚类算法，本文算法聚类结果稳定且效果较好，还能有效减少原算法的内存消耗和执行时间，更适用于大数据背景下高维海量电力负荷曲线的聚类分析。

1 快速搜索与发现密度峰值聚类算法

1.1 CFSFDP 算法

快速搜索与发现密度峰值聚类算法(CFSFDP)的基本思想主要基于两点：1) 每个聚类中心的局部密度高于其领域样本点的密度；2) 聚类中心与比它密度更高的点的距离相对较大。

每个样本点的局部密度 ρ_i 定义如式(1)。

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right) \tag{1}$$

式中：若 $x<0$，则 $\chi(x)=1$，否则，$\chi(x)=0$；d_{ij} 是样本点 i 和 j 间的距离(一般用欧式距离度量)；d_c 是截断距离，它的设定对聚类效果影响很大，如果 d_c 过小，可能导致某一类簇被强制分裂成几部分，如果 d_c 过大，可能会将几个类簇合并成一类。文献[17]指出，合适的 d_c 应该使得平均每个样本点的 d_c 范围内的样本数量占整个数据集的比例为 τ(τ=1%~2%)。

为了降低截断距离 d_c 对聚类结果的影响，文献[18]提出利用高斯核函数改进局部密度：

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

每个样本点与密度比它更高的点的最小距离 δ_i 定义如下：

$$\delta_i = \min_{j:\rho_j>\rho_i}\left(d_{ij}\right) \tag{3}$$

式中，对于密度是全局最高的样本点 i，令 $\delta_i=\max_j(d_{ij})$。

应用 CFSFDP 算法的基本步骤如下所示。

输入：数据集 X，参数 τ

输出：聚类结果

1) 计算数据集的距离矩阵；

2) 计算截断距离 d_c；

3) 利用式(2)和式(3)计算每个样本的局部密度 ρ 和距离 δ；

4) 以局部密度 ρ 为横坐标，距离 δ 为纵坐标建立二维决策图，找出 ρ 和 δ 都较大的点作为聚类中心。某些情况下，这种决策图可能难以选出聚类中心[17]，此时可以降序排列后的 γ 值($\gamma_i=\rho_i\delta_i$, i=1, 2,…, n)为纵坐标、新序列号 i 为横坐标建立决策图，选出 γ 较大的点作为聚类中心。

5) 对剩余样本进行分配，每个样本均与密度比它大且距离最近的样本同属一类。

1.2 CFSFDP 算法的不足

CFSFDP 算法计算局部密度 ρ 时没有考虑到数据集的局部结构，当数据集各类簇间密度差异较大时会导致错误聚类(几个类簇合并或某个类簇分裂)。

设数据集规模为 n，CFSFDP 算法的内存消耗主要来自存储规模为 $n\times n$ 的距离矩阵。执行时间主要来自计算距离矩阵、截断距离 d_c、样本局部密度 ρ 和距离 δ，其中计算距离矩阵所占比重最大，这在数据规模和维数很高时更明显。CFSFDP 算法的内存消耗和执行时间会随着数据规模的增大而迅速增长，因此算法的应用会受到数据规模的限制，难以更好地应用于高维海量负荷曲线聚类分析。

2　改进的快速密度峰值聚类算法

2.1 相关理论

2.1.1 K 近邻算法及其 kd 树实现

K 近邻算法(KNN)是一种简单实用的监督学习分类方法，广泛应用于文本分类、聚类分析、预测分析、模式识别等领域。它的基本思路是通过计算待分类样本与所有训练样本间的距离，取出 K 个距离最小的训练样本(称为待分类样本的 K 近邻)，则待分类样本的类别由其 K 近邻中出现最多的类别决定[22]。

KNN 算法的实现主要有两种方法[23]：一是线性扫描，即计算每一个待分类样本与所有训练样本间的距离，这对高维海量的训练集来说耗时巨大，几乎无法实现；二是通过特殊的结构存储训练数据，以减少计算距离的次数，其中经典有效的方法之一是 kd 树。

kd 树是一种对 k 维空间中的实例点进行存储以便对其进行快速检索的树形数据结构[23-24]。应用 kd 树对 k 维搜索空间进行划分，在待分类样本的邻域空间内进行相关搜索即可得到其 K 近邻样本，从而避免了穷举搜索计算，极大地减少了搜索的计算量。

2.1.2 主成分分析法降维

降维技术主要分为特征选择和特征提取。

特征选择是从原始冗余的特征集中选出能够包含数据集大部分信息的特征子集。基于特征选择的负荷曲线降维一般是以峰、平、谷时耗电率等能够反映负荷曲线时间特性的特征构建原始特征集，再从中选出最优的特征子集以达到降维的目的[25-26]。利用特征选择降维的优点是降维后的特征子集物理意义明确，但现有的原始特征集还无法完全反映负荷曲线的时间特性，从中选出的特征子集显然会丢失更多的负荷曲线信息，而且目前还没有如何构造完善的原始特征集的理论。

特征提取是将原始高维特征空间经过某种形式的变换，形成新的低维特征空间。虽然利用特征提取得到的新特征的物理意义不明确，但新特征的确包含了数据集的大部分信息，可以在保证较高聚类准确率的同时提高聚类效率。文献[15]指出相比其他降维方法(如 Sammon 映射、自组织映射等)，主成分分析法(PCA)降维的负荷曲线的聚类准确率和效率是最高的。

主成分分析法是一种重要的线性降维方法。它的基本思想是将 p 个原始变量经过线性变换后转换为 m 个互相正交的新变量($m<<p$)，这 m 个新变量尽可能多地包含原始变量的信息，称为原始变量的主成分[27]。

2.2 改进的快速密度峰值聚类算法

2.2.1 CFSFDP 算法改进思路

本文在文献[18-21]的基础上，主要对原算法的内存消耗和执行时间两方面进行改进，得到改进的快速密度峰值聚类算法(I-CFSFDP)。

1) CFSFDP 算法内存消耗的改进

为了避免 CFSFDP 算法在计算样本局部密度时没有考虑到数据集的局部结构，引入 KNN 的思想对局部密度计算准则进行改进，改进的样本局部密度 ρ_i 定义如下：

$$\rho_i = \sum_{j\in \mathrm{KNN}(i)} \exp\left(-d_{ij}^{\ 2}\right) \tag{4}$$

式中，KNN(i)表示样本 i 的 K 近邻样本集合。式(4)表明样本 i 密度随着样本 i 到其 K 近邻距离的增大而减小，新的 ρ_i 仅考虑了其 K 近邻样本，可以更多地反映样本 i 的局部结构。

为了降低 CFSFDP 算法的内存消耗，还需要利用 KNN 思想改进样本距离 δ 的计算准则。由于除了少数样本点(如聚类中心)外，其余每个样本点 i 至少存在一个密度高于该样本点 i 密度的近邻样本，显然该样本点 i 与密度比它更高的点的最小距离 δ_i 只需从其 K 个近邻样本中寻找。因此，基于

KNN 改进的样本距离 δ_i 定义如下：

$$\delta_i=\begin{cases}\min\limits_{\{j\in \mathrm{KNN}(i)\}\wedge\{\rho_j>\rho_i\}}\left(d_{ij}\right), \\ \text{if } \exists j \text{ s.t.}\left\{j\in \mathrm{KNN}\left(i\right)\right\}\wedge\left\{\rho_j>\rho_i\right\} \\ \min\limits_{j:\rho_j>\rho_i}\left(d_{ij}\right), \\ \text{otherwise}\end{cases} \tag{5}$$

其中，对于局部密度是全局最高的样本点 i，令

$$\delta_i=\max_{j\neq i}\left(\delta_j\right) \tag{6}$$

上述基于 KNN 算法计算样本局部密度 ρ 和距离 δ 时 K 的取值由式(7)得到。

$$K=p\times n \tag{7}$$

式(7)表明每个样本的 K 近邻数占整个数据规模的比例为 p，p 一般很小，则 $K<<n$。

改进的样本距离 δ 计算流程如下：

(1) 将数据集的样本局部密度降序排列，设为 $\rho'=(\rho_1',\rho_2',\cdots,\rho_n')$，令 $i=2$；

(2) 根据式(5)计算密度 ρ_i'对应的距离 δ_i'，$i=i+1$；

(3) 如果 $i\leq n$，返回步骤(2)；否则，令 $i=1$，根据式(6)计算密度 ρ_i'对应的距离 δ_i'。

2) CFSFDP 算法执行时间的改进

在大规模数据集上应用线性扫描方法建立 KNN 矩阵是非常耗时的，而应用经典的 kd 树算法可以实现快速 K 近邻搜索建立 KNN 矩阵，具体过程可参考文献[23-24]。由于在训练集规模远大于空间维数时($n\geq 2^k$)应用 kd 树进行 K 近邻搜索效率很高，当二者差别不大时，效率会迅速下降，几乎接近线性扫描。并且负荷曲线随着维数增高，其等距性越明显[28]，距离测度越不准确[29]，导致相似性度量偏差很大，极大地降低了聚类效果。因此，为了提高 K 近邻搜索效率，进一步加快 KNN 矩阵的建立，同时提高聚类效果，有必要在应用 kd 树前先对数据集进行降维处理。除此以外，对数据集降维还可以减少各样本向量间欧式距离的计算量和计算时间。由上节分析可知，宜采用 PCA 降维。

因此，为了有效降低 CFSFDP 算法执行时间，应先对数据集进行 PCA 降维处理，然后采用 kd 树算法建立 KNN 矩阵，再利用式(4)—式(7)计算样本的 ρ 和 δ。

2.2.2 I-CFSFDP 算法流程

I-CFSFDP 算法整体流程如图 1 所示。

输入：负荷曲线数据集 X，K 近邻比例参数 p

输出：聚类结果

(1) 为了全面反映各负荷曲线的形态特征，避免量纲和幅值差异对负荷曲线聚类的影响，对数据集 X 极值归一化处理：

$$x_t'=\frac{x_t-x_{\min}}{x_{\max}-x_{\min}} \tag{8}$$

式中，x_t、x_t'、$x_{\min}$、$x_{\max}$ 分别是第 t 时刻的负荷、负荷曲线归一化后第 t 时刻的值、日最小负荷和日最大负荷。

(2) 对归一化后的数据集进行 PCA 降维得 $\boldsymbol{X}^*$。

(3) 采用 kd 树算法建立 $\boldsymbol{X}^*$的 KNN 矩阵。

(4) 根据式(4)—式(7)计算样本局部密度 ρ 和距离 δ。

(5) 为了避免ρ和δ数量级的差异导致两者的决策权重不一样，分别对 ρ 和 δ 归一化处理再形成 γ 曲线构成决策图，选出 γ 较大的点作为聚类中心。

(6) 对剩余样本进行分配，每个样本均与密度比它大且距离最近的样本同属一类。

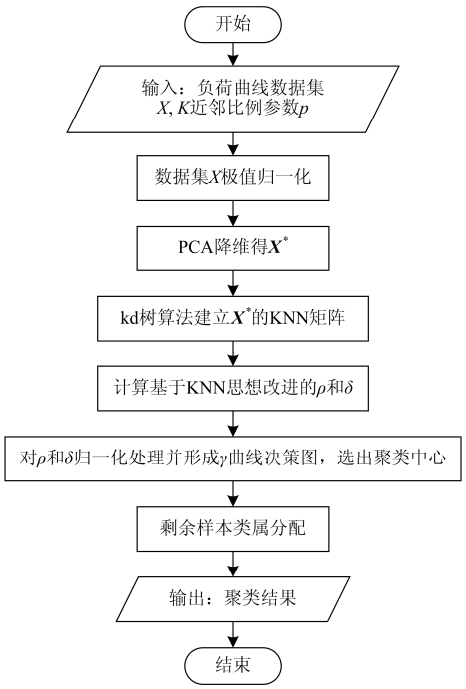


图 1 I-CFSFDP 算法整体流程

Fig. 1 Overall flow chart of I-CFSFDP algorithm

2.2.3 I-CFSFDP 算法性能分析

设数据集规模为 n，I-CFSFDP 算法的内存消耗主要来自存储规模为 $n\times K$ 的 KNN 矩阵，与 CFSFDP 算法的内存消耗相比明显降低了。由于 $K<<n$，因此当数据规模很大时内存消耗会有更明显的降低。

I-CFSFDP 算法的执行时间主要来自 4 部分：PCA 降维、计算 KNN 矩阵、计算样本局部密度 ρ 和距离 δ，其中计算 KNN 矩阵所占比重最大。与 CFSFDP 算法相比，I-CFSFDP 算法在计算 KNN 矩阵时利用 PCA 降维和 kd 树算法进行优化，极大地

减少了执行时间。由式(4)—式(6)可知，除了少数样本点外，I-CFSFDP 算法计算每个样本的 ρ 和 δ 时基本只涉及其 K 个近邻样本的运算。而由式(1)—式(3)可知，CFSFDP 算法基本需要涉及$(n-1)$个样本的运算，这也在一定程度上减少了算法执行时间。除此以外，由于 CFSFDP 算法计算截断距离时需要对规模巨大的距离矩阵上三角元素进行升序排列，其所需时间远高于 I-CFSFDP 算法的 PCA 降维时间。

综上所述，I-CFSFDP 算法能考虑到数据集的局部结构，避免类簇密度差异巨大时出现错误聚类，而且能有效减少原算法的内存消耗和执行时间。

3　实验与算例分析

3.1　实验环境

本文采用的数据集来源于美国能源部 OpenEI (open energy information)公布的商业用户负荷数据[30]，涵盖 16 个行业类别共 14 976 个用户，采样间隔为 1 h，合计 24 个量测点，随机选择夏季 6~10 月中 10 个工作日总共近 15 万条负荷曲线为研究对象。

实验环境：单台计算机，配置为 Intel(R) Core(TM) i5-3210M 2-core CPU@2.50 GHz，操作系统为 Windows7，内存为 4 GB(2.91 GB 可用)，编程语言为 MATLAB R2013a。

3.2　聚类算法性能对比实验

取研究对象不同规模的子集进行如下测试：1) 分别采用 I-CFSFDP 算法和 CFSFDP 算法进行聚类分析，对比二者各部分的内存消耗和执行时间，实验结果如表 1 所示；2) 分别采用几种传统聚类算法和改进前后的 CFSFDP 算法进行聚类分析，对比各算法的总体内存消耗和执行时间，实验结果如图 2 所示。为了保证算法执行时间的客观性，每种算法在不同规模的数据集上各运行 20 次，取运行时间的平均值作为算法在该数据规模下的执行时间。

表 1　I-CFSFDP 算法和 CFSFDP 算法性能对比

Table 1 Performance comparisons between I-CFSFDP and CFSFDP

数据集规模	I-CFSFDP 运行时间/s						CFSFDP 运行时间/s					
	数据降维	计算 KNN 矩阵	主程序			合计	计算截断距离 d_c	计算距离矩阵	主程序			合计
			计算局部密度 ρ	计算距离 δ	其他				计算局部密度 ρ	计算距离 δ	其他	
2 000	0.070 1	0.074 2	0.000 445	0.071	0.030 7	0.246 4	0.211 6	0.158 3	0.080 1	0.086 2	0.034 8	0.571
4 000	0.077 1	0.107 4	0.000 656	0.115 2	0.032 5	0.332 9	0.857 3	0.463 3	0.273 3	0.281 5	0.045 3	1.920 7
6 000	0.080 2	0.153 9	0.001 1	0.151 4	0.033 7	0.420 3	1.88	1.065	0.610 2	0.648 2	0.064 9	4.268 3
8 000	0.084 3	0.211 8	0.001 6	0.181 1	0.035 3	0.514 1	2.997 1	1.738 5	1.069 9	1.137 1	0.087 1	7.029 7
10 000	0.090 6	0.281 1	0.002 2	0.210 6	0.037 3	0.621 8	—	—	—	—	—	—
20 000	0.111 6	0.792 5	0.007 8	0.408 4	0.043 5	1.363 8	—	—	—	—	—	—
60 000	0.198 5	6.488 6	0.070 7	1.516 1	0.074	8.347 9	—	—	—	—	—	—
100 000	0.288 4	24.925	0.183 4	3.271 5	0.116 1	28.784 4	—	—	—	—	—	—
140 000	0.392 6	40.883 4	0.385	5.701 7	0.163	47.525 7	—	—	—	—	—	—

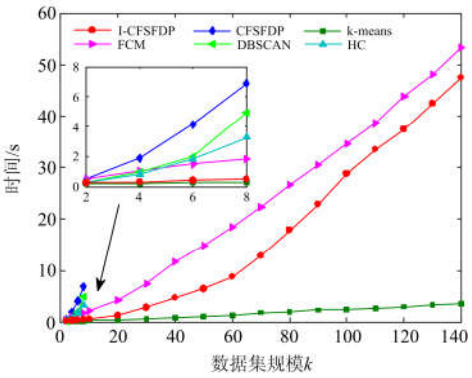


图 2　各种聚类算法性能对比

Fig. 2 Performance comparisons between each clustering algorithm

从表 1 可以看出，CFSFDP 算法大致只能处理 8 000 条负荷曲线，当数据集规模超过 8 000 时会因为算法所需内存消耗过大导致计算机内存空间不足而无法继续执行(既内存溢出)，I-CFSFDP 算法在数据集规模超过 140 000 时才出现内存溢出，这验证了 I-CFSFDP 算法能有效减少原算法的内存消耗。

I-CFSFDP 算法各部分及总的执行时间均小于 CFSFDP 算法，而且随着数据集规模的增大，I-CFSFDP 算法执行时间的减少量更大，这说明了 I-CFSFDP 算法能有效减少原算法的执行时间。

从图 2 可以看出，在总体执行时间方面，I-CFSFDP 算法仅劣于 k-means 算法，较优于 FCM 算法，远胜于其他聚类算法；在总体内存消耗方面，

k-means 和 FCM 算法能处理规模超过 140 000 的数据集，I-CFSFDP 算法在数据集规模超过 140 000 时出现内存溢出，其余算法大致在数据集规模超过 8 000 时出现内存溢出，表明 I-CFSFDP 算法内存消耗在比较合理的范围内，远小于大部分聚类算法。

综上所述，I-CFSFDP 算法所需的内存消耗和执行时间均不高，远胜于大部分传统聚类算法，验证了 I-CFSFDP 算法在高维海量负荷曲线聚类分析中的适用性。注意到虽然 k-means 算法的内存消耗和执行时间远远优于 I-CFSFDP 算法，但其聚类效果较差且结果不稳定，下节算例分析将作出证明。

3.3 算例分析

3.3.1 I-CFSFDP 算法聚类分析

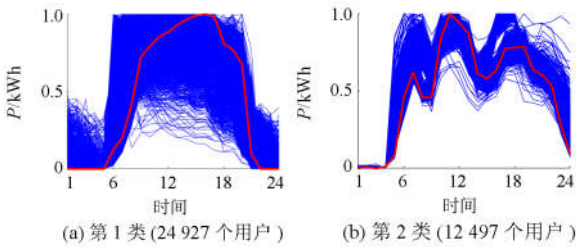
为了验证 I-CFSFDP 算法对高维海量电力负荷曲线的聚类效果，取研究对象中的 100 000 条负荷曲线作为输入数据，运用 I-CFSFDP 算法(p=0.2%)进行聚类分析，聚类簇如图 3 所示。

负荷曲线集被分成 8 类，可以归为单峰、双峰、三峰及多峰 4 种类型。每类负荷涵盖的用户信息如表 2 所示。

单峰型负荷包括第 1、4、5、8 类负荷，主要涵盖卫生、教育、商业零售业等行业的用户，总体特征为白天用电水平较高且变化不大，中午时段部分用户因午休、人流量少等导致用电量略有下降。而同属单峰型的这 4 类负荷除了曲线形态存在差异外，峰期时段也不同，分别为 9:00—18:00、8:00—17:00、9:00—16:00、7:00—17:00。

双峰型负荷包括第 6 类负荷，主要涵盖住宅公寓的用户，峰期时段为 6:00—9:00、17:00—21:00，而且由于住宅家用电器晚上的用电量远高于早上的用电量，导致第二个峰值显著高于第一个峰值。

三峰型负荷包括第 2、7 类负荷，主要涵盖餐饮业、大型住宿业等行业的用户。这 2 类负荷除曲线形态不同外，同样存在峰期时段的差异。第 2 类负荷峰期时段为 6:00—8:00、10:00—13:00、17:00—19:00。第 7 类负荷峰期时段为 6:00—8:00、17:00—19:00、19:00—22:00。



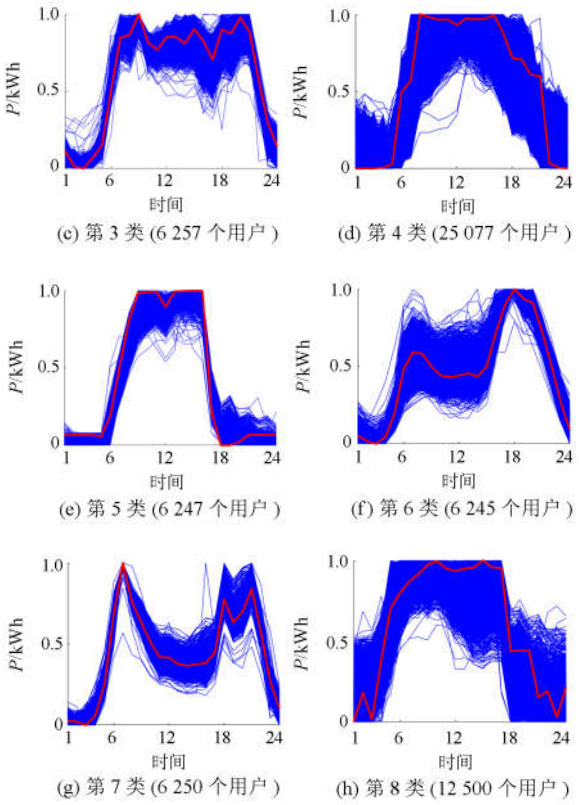


图 3 I-CFSFDP 算法聚类簇

Fig. 3 Clusters of I-CFSFDP

多峰型负荷包括第 3 类负荷，主要涵盖小型住宿业等行业的用户，该类型负荷在 7:00—21:00 时间段内存在多个负荷峰值，且峰值大小相差不大。

一般来说，大部分行业类别相同的用户应该被分到相同的聚类簇中，但也可能存在少数相同行业类别的用户分到不同聚类簇。因此，由上述分析可看出，聚类结果合理，表明 I-CFSFDP 算法对高维海量电力负荷曲线的聚类效果较好，结果稳定，具有较高的工程实用价值。

3.3.2 传统聚类算法聚类分析

限于篇幅，本文仅采用传统聚类算法中的 k-means 算法对上述数据集进行聚类分析，以对比验证 I-CFSFDP 算法的聚类效果和稳定性。由于 k-means 算法受初始聚类中心的影响，因此为了得到较好的聚类效果，采用 k-means 算法对数据集运行 20 次，取最优的聚类结果，聚类簇如图 4 所示，聚类簇用户信息如表 2 所示。

从聚类结果可看出，k-means 算法不稳定，聚类效果较差，例如第 2 类负荷无法区分多峰型的小型酒店和单峰型的办公楼、中学等用户，第 5 类负

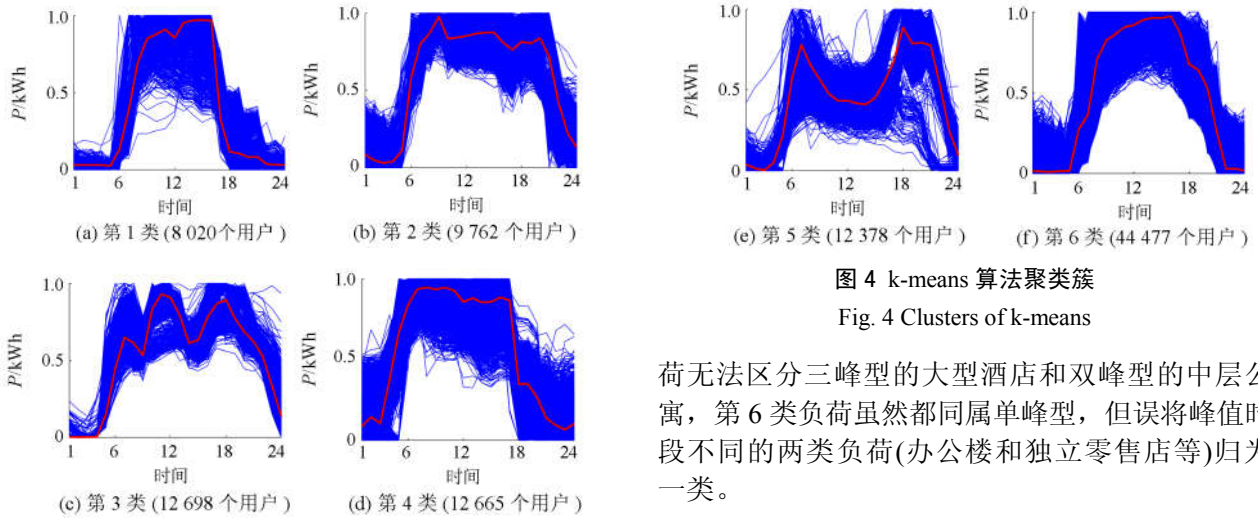


图 4 k-means 算法聚类簇

Fig. 4 Clusters of k-means

荷无法区分三峰型的大型酒店和双峰型的中层公寓，第 6 类负荷虽然都同属单峰型，但误将峰值时段不同的两类负荷(办公楼和独立零售店等)归为一类。

表 2 I-CFSFDP 算法和 CFSFDP 算法聚类簇用户信息

Table 2 Electricity customers information of I-CFSFDP and CFSFDP clusters

用户类别	I-CFSFDP 算法聚类簇用户信息								k-means 算法聚类簇用户信息					
	第 1 类	第 2 类	第 3 类	第 4 类	第 5 类	第 6 类	第 7 类	第 8 类	第 1 类	第 2 类	第 3 类	第 4 类	第 5 类	第 6 类
全服务餐厅	0	6 248	0	0	0	1	0	1	0	2	6 243	0	5	0
医院	0	0	1	4	0	0	0	6 245	1	34	5	6 201	1	8
大型酒店	0	0	0	0	0	1	6 249	0	0	0	0	0	6 250	0
大型办公楼	3	0	0	6 247	0	0	0	0	0	778	0	1	1	5 470
中型办公楼	1	0	0	6 249	0	0	0	0	4	1	0	1	0	6 244
中层公寓	0	1	5	0	0	6 243	1	0	0	14	202	0	6 034	0
门诊部	0	0	0	0	0	0	0	6 250	0	0	0	6 250	0	0
小学	177	0	0	6 073	0	0	0	0	433	0	0	10	11	5 796
快速服务餐厅	0	6 248	0	1	0	0	0	1	0	1	6 248	0	1	0
中学	6 020	0	0	230	0	0	0	0	73	1 679	0	14	11	4 473
小型酒店	0	0	6 250	0	0	0	0	0	0	6 246	0	0	4	0
小型办公楼	1	0	1	6 248	0	0	0	0	1 262	2	0	0	0	4 986
独立零售店	6 245	0	0	5	0	0	0	0	0	12	0	59	12	6 167
沿公路商业区	6 240	0	0	10	0	0	0	0	0	16	0	122	45	6 067
超级市场	6 240	0	0	10	0	0	0	0	0	977	0	4	3	5 266
仓库	0	0	0	0	6 247	0	0	3	6 247	0	0	3	0	0

4　结论

针对传统聚类算法难以很好地对大数据背景下高维海量、类簇形状差异巨大的电力负荷曲线进行聚类分析的情况，本文提出一种改进的快速密度峰值聚类算法。通过实验与算例分析，表明该算法聚类结果稳定且效果较好，还能有效减少原算法的内存消耗和执行时间，可以较好地实现高维海量电力负荷曲线聚类分析，辅助大数据环境下的负荷预测、电价制定与需求侧响应、用电稽查等。

K 近邻比例参数 p 选取对聚类效果影响较大，本文利用经验参数 p 进行聚类，取得较好的效果，基于数据集结构选取精确的 K 近邻比例参数 p 以进一步改善负荷曲线聚类效果是下阶段的研究重点。

参考文献

[1] 惠晓林，孙振权. 智能配电网与物联网的融合[J]. 物联网技术, 2011(8): 31-35.
HUI Xiaolin, SUN Zhenquan. Integration of intelligent distribution network and internet of things[J]. Internet of Things Technologies, 2011(8): 31-35.

[2] 刘科研，盛万兴，张东霞，等. 智能配电网大数据应用需求和场景分析研究[J]. 中国电机工程学报，2015, 35(2): 287-293.
LIU Keyan, SHENG Wanxing, ZHANG Dongxia, et al. Big data application requirements and scenario analysis in smart distribution network[J]. Proceedings of the CSEE, 2015, 35(2): 287-293.

[3] 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.

[4] 林顺富, 郝朝, 汤晓栋, 等. 基于数据挖掘的楼宇短期负荷预测方法研究[J]. 电力系统保护与控制, 2016, 44(7): 83-89.
LIN Shunfu, HAO Chao, TANG Xiaodong, et al. Study of short-term load forecasting method based on data mining for buildings[J]. Power System Protection and Control, 2016, 44(7): 83-89.

[5] 陈明照, 毛坚, 杜宗林, 等. 基于聚类法的工业用户需求侧管理(DSM)方案分析与研究[J]. 电力系统保护与控制, 2017, 45(7): 84-89.
CHEN Mingzhao, MAO Jian, DU Zonglin, et al. Analysis on demand side management scheme of industrial enterprise based on clustering method[J]. Power System Protection and Control, 2017, 45(7): 84-89.

[6] 田力, 向敏. 基于密度聚类技术的电力系统用电量异常分析算法[J]. 电力系统自动化, 2017, 41(5): 64-70.
TIAN Li, XIANG Min. Abnormal power consumption analysis based on density-based spatial clustering of applications with noise in power systems[J]. Automation of Electric Power Systems, 2017, 41(5): 64-70.

[7] CHICCO G, NAPOLI R, PIGLIONE F. Comparisons among clustering techniques for electricity customer classification[J]. IEEE Transactions on Power Systems, 2006, 21(2): 933-940.

[8] 李朝晖, 尹晓博, 杨海晶, 等. 基于改进的 k-means 聚类算法的季节性负荷特性分析[J]. 电网与清洁能源, 2018, 34(2): 53-59.
LI Zhaohui, YIN Xiaobo, YANG Haijing, et al. Seasonal load characteristics analysis based on improved k-means clustering algorithm[J]. Power System and Clean Energy, 2018, 34(2): 53-59.

[9] 王红斌, 陈扬, 高雅, 等. 基于数据挖掘的预警技术在一体化输电设备监测中的应用研究[J]. 电网与清洁能源, 2014, 30(1): 55-58.
WANG Hongbin, CHEN Yang, GAO Ya, et al. Application of early warning technology in power transmission equipment condition monitoring based on data mining[J]. Power System and Clean Energy, 2014, 30(1): 55-58.

[10] 王桂兰, 周国亮, 赵洪山, 等. 大规模用电数据流的快速聚类和异常检测技术[J]. 电力系统自动化, 2016, 40(24): 27-33.
WANG Guilan, ZHOU Guoliang, ZHAO Hongshan, et al. Fast clustering and anomaly detection technique for large-scale power data stream[J]. Automation of Electric Power Systems, 2016, 40(24): 27-33.

[11] 李芬, 李春阳, 闫全全, 等. 基于变分贝叶斯学习的光伏功率波动特性研究[J]. 电力自动化设备, 2017, 37(8): 99-104.
LI Fen, LI Chunyang, YAN Quanquan, et al. Photovoltaic output fluctuation characteristics research based on variational Bayesian learning[J]. Electric Power Automation Equipment, 2017, 37(8): 99-104.

[12] 张西芝. 网格聚类算法的研究[D]. 郑州: 郑州大学, 2006.

[13] 周开乐, 杨善林. 基于改进模糊 C 均值算法的电力负荷特性分类[J]. 电力系统保护与控制, 2012, 40(22): 58-63.
ZHOU Kaile, YANG Shanlin. An improved fuzzy C-means algorithm for power load characteristics classification[J]. Power System Protection and Control, 2012, 40(22): 58-63.

[14] 张少敏, 赵硕, 王保义. 基于云计算和量子粒子群算法的电力负荷曲线聚类算法研究[J]. 电力系统保护与控制, 2014, 42(21): 93-98.
ZHANG Shaomin, ZHAO Shuo, WANG Baoyi. Research of power load curve clustering algorithm based on cloud computing and quantum particle swarm optimization[J]. Power System Protection and Control, 2014, 42(21): 93-98.

[15] 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.

[16] 赵文清, 龚亚强. 基于 Kernel K-means 的负荷曲线聚类[J]. 电力自动化设备, 2016, 36(6): 203-207.
ZHAO Wenqing, GONG Yaqiang. Load curve clustering based on Kernel K-means[J]. Electric Power Automation Equipment, 2016, 36(6): 203-207.

[17] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014(334): 1492-1496.

[18] WANG S, WANG D, LI C, et al. Comment on "Clustering by fast search and find of density peaks"[J]. arXiv preprint arXiv: 1501.04267, 2015.

[19] JI C, LEI Y. Parallel clustering by fast search and find of density peaks[C] // 2016 International Conference on Audio, Language and Image Processing, July 11-12,

2016, Shanghai, China: 563-567.

[20] DU M, DING S, JIA H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis[J]. Knowledge-Based Systems, 2016(99): 135-145.

[21] 谢娟英, 高红超, 谢维信. K 近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学: 信息科学, 2016, 46(2): 258-280.
XIE Juanying, GAO Hongchao, XIE Weixin. K-nearest neighbors optimized clustering algorithm by fast search and finding the density peaks of a dataset[J]. Science China: Information Sciences, 2016, 46(2): 258-280.

[22] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[23] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[24] BENTLEY J L. Multidimensional binary search trees used for associative searching[J]. Communications of the ACM, 1975, 18(9): 509-517.

[25] 傅军栋, 杨姚, 罗善江. 智能小区居民用电负荷特征权重分析[J]. 电力系统保护与控制, 2016, 44(18): 41-45.
FU Jundong, YANG Yao, LUO Shanjiang. Residential electricity load features weighting analysis in smart community[J]. Power System Protection and Control, 2016, 44(18): 41-45.

[26] 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.
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.

[27] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1): 37-52.

[28] PIAO M, SHON H S, LEE J Y, et al. Subspace projection method based clustering analysis in load profiling[J]. IEEE Transactions on Power Systems, 2014, 29(6): 2628-2635.

[29] PARSONS L, HAQUE E, LIU H. Subspace clustering for high dimensional data: a review[J]. SIGKDD Explore, 2004, 6(1): 90-105.

[30] ERIC W. Commercial and residential hourly load profiles for all TMY3 locations in the United States[EB/OL]. [2017-09-10]. http: // en.openei.org/datasets /dataset/ commercial and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states.

收稿日期：2017-09-18；　修回日期：2017-11-28

作者简介：

陈俊艺(1994—)，男，硕士研究生，研究方向为电力系统大数据分析；E-mail: junyichen@whu.edu.cn

丁坚勇(1957—)，男，博士，教授，博士生导师，研究方向为电力系统运行与控制、电力系统规划及可靠性；

田世明(1965—)，男，教授级高工，研究方向为能源互联网、大数据分析。

(编辑　魏小丽)