改进谱聚类与遗传算法相结合的电力时序曲线聚类方法

丁 明[1],黄 冯[1],邹佳芯[2],刘金山[2],宋晓皖[1]
(1. 合肥工业大学 安徽省新能源利用与节能重点实验室,安徽 合肥 230009;
2. 国网青海省电力公司电力科学研究院 青海省光伏发电并网技术重点实验室,青海 西宁 810008)

摘要:为改善传统聚类算法在电力时序数据上的聚类效果,提出一种基于优化特征向量选取的遗传谱聚类算法。针对应用数据结构特点,合理优化谱聚类算法中特征向量的提取过程,避免传统方法可能造成的数据信息缺失问题;采用遗传聚类优化算法对优选后的特征向量进行聚类划分,并将最终划分结果映射回原始数据。以 UCI 标准合成时间序列数据与美国区域电网运营商 PJM 提供的日负荷数据为例,对比分析现有常用聚类算法与所提算法测试结果的聚类有效性指标与形态特征。研究结果表明,所提算法分类效果显著,有较高的聚类质量和算法稳健性,具有工程应用前景。

关键词:时序数据;谱聚类;遗传算法;特征向量提取;负荷聚类
中图分类号:TM 714 **文献标识码**:A DOI:10.16081/j.issn.1006-6047.2019.02.014

0 引言

近年来,随着智能电网逐步成为电力工业的发展方向与趋势,电力系统信息化、数字化的时代已经到来[1]。电力系统中的数据通常种类繁杂、来源广泛,在电力系统的发电、输电、变电、配电、用电等各个环节均有大量、不同类型的电力数据不断产生,如何有效地从各类电力数据中挖掘具有价值的信息是电力系统亟需解决的问题。针对电力数据具有海量、异构且含大量噪音的特点,采用数据挖掘技术分析与处理电力数据,并为电力决策提供参考,已被广泛认为是一种有效的方法[2]。

聚类分析属于数据挖掘技术中的无监督学习方法,它旨在发现未知数据中具有相似模式的对象,并分别归类[3]。目前,聚类分析技术已被广泛应用于电力系统领域的各个方向中,如电力负荷曲线聚类[4]、供电块划分[5]、电力工业异常数据识别[6]、电能质量扰动识别[7]、可再生能源发电预测[8]等。现有的聚类方法一般包括划分聚类算法、层次聚类算法、基于密度聚类算法、基于网格聚类算法与基于模型聚类算法[9]等。在电力系统中应用较为成熟的聚类方法包括 *K*-means 算法、模糊 C 均值 FCM(Fuzzy C-Means)算法、层次聚类以及基于神经网络的聚类算法等。文献[10]采用改进的 *K*-means 算法对全年风电、光伏、负荷曲线进行聚类划分,得到一组可以反映周期内历史数据特征的典型场景,并将其代入以风电接纳能力最大为目标的机组组合模型中验证所提方法的有效性。文献[11]将重采样技术、层次聚类与划分聚类相结合,提出一种用户用电曲线的集成聚类算法。文献[12]提出一种基于云模型确定聚类数目与初始聚类中心的 FCM 算法,从负荷曲线中提取相似的用户用电模式。文献[13]将主成分分析法与层次聚类相结合,从而对全年风电出力序列进行聚类划分。文献[14]利用 *K*-means 算法对气象信息与风电历史数据进行聚类,并结合神经网络方法提高了风力发电预测的精度。文献[15]提出一种基于密度空间聚类与引力搜索算法的居民负荷用电模式分类模型。

由于无论是负荷曲线,还是新能源出力曲线,本质上它们均属于时间序列数据,因此研究适用于时序数据的聚类方法与相关技术,对电力数据分类而言具有重要的参考价值。目前,时序数据聚类的方法主要分为基于原始测度的方法与基于时间序列形态特征的方法这两大类[16]。文献[17]用动态弯曲距离(DTW)替换普通欧氏距离度量,并结合全局时间序列平均法提高了 *K*-means 算法聚类时间序列的能力,但最终的效果并不显著。文献[18]采用基于 MapReduce 框架改进的谱聚类算法,并应用于实际股市数据中,取得了较好的分类效果。

本文提出一种谱聚类特征向量优化选取与遗传聚类算法相结合的聚类优化算法。算法主要分为 2 个步骤:根据算法应用数据的结构特征,合理优化 NJW 算法中的特征向量选取过程;在形成的特征向量聚类空间中,用遗传聚类优化算法取代传统 NJW 算法中的最后一步 *K*-means 聚类,以提高传统 NJW 算法的全局寻优能力,实现对数据的合理聚类分析。将本文方法分别应用于 UCI 标准合成时间序列分析与实际负荷数据聚类中,验证了所提方法的合理性。

1 基于优化特征向量提取的遗传谱聚类

1.1 特征聚类空间的提取

1.1.1 谱聚类理论

谱聚类是从图论中演化出的算法,它将数据聚

收稿日期:2018-04-01;**修回日期**:2018-10-30
基金项目:国家电网公司科技项目(5228001600DX)
Project supported by the Science and Technology Program of SGCC(5228001600DX)

类转换成图的最优划分问题。谱聚类可以识别任意形状的簇并且其聚类性能不受数据维数的约束。

不同的谱聚类算法采用的划分准则不同,但总结起来均包括以下3个步骤:

a. 构建可以表示数据样本集关系属性的矩阵 $\boldsymbol{R}$;

b. 计算 $\boldsymbol{R}$ 的前 k 个特征值对应的特征向量集,以构建数据的特征向量空间 S;

c. 利用聚类方法对特征向量空间 S 中的数据点进行聚类,并将聚类结果映射回原数据空间。

NJW 算法[19]是一种优秀的经典谱聚类子算法,对于给定的聚类数目 k,它通过计算标准 Laplacian 矩阵的前 k 个最小特征值对应的特征向量,采用 K-means 算法对特征向量空间 S 中的数据点聚类,并映射得到原数据的划分结果。

1.1.2 特征空间选取方法

实际上,由于不同类型的数据结构特性一般不同,NJW 算法统一按前 k 个特征向量构成的特征子空间聚类,往往会导致不太理想的聚类划分结果。文献[20]证明了只有当标准 Laplacian 矩阵的第 k 个与第 $k+1$ 个特征值间的差值足够大时,NJW 算法才可能具有良好的聚类划分结果,并且文献[20]给出一个假设:即应按实际需求设定参数 $M(M>k)$,用前 M 个特征向量替代前 k 个特征向量构成特征向量子空间。参考文献[20]的思路,并结合 NJW 算法,本文提出一种参数 M 范围的简化选取方法。

给定聚类数目 k 与数据集 $X=\{X_1,X_2,\cdots,X_i,\cdots,X_n\}$,其中 $X_i=\{X_{i1},X_{i2},\cdots,X_{im}\}$,即每个样本序列,共有 m 维属性。则特征向量空间 S 的产生方法如下。

a. 构建原始数据间的相似度矩阵 $\boldsymbol{W}$,矩阵元素 $W_{ij}=\exp[-d^2(x_i,x_j)/(2\sigma^2)]$,其中 $d^2(x_i,x_j)$ 为两样本点之间的欧氏距离,σ 为样本点之间衰减速度的尺度参数,本文选取其为所有样本数据的标准差。

b. 构建标准 Laplacian 矩阵 $\boldsymbol{L}_{\text{sym}}$。将矩阵 $\boldsymbol{W}$ 每行元素之和作为度矩阵 $\boldsymbol{E}$ 的主对角线元素,其余元素设为0;构建 Laplacian 矩阵 $\boldsymbol{L}=\boldsymbol{E}^{-0.5}\boldsymbol{W}\boldsymbol{E}^{-0.5}$,则标准 Laplacian 矩阵 $\boldsymbol{L}_{\text{sym}}=\boldsymbol{E}^{-0.5}\boldsymbol{W}\boldsymbol{E}^{-0.5}$。

c. 选定参数 M 的范围:首先计算 $\boldsymbol{L}_{\text{sym}}$ 的所有 n 个特征值,并按从小到大顺序排列,即 $\Lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_n\}$,且 $\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$;依次计算 Λ 中所有特征值间的差值序列,即 $G=\{g_1,\cdots,g_{n-1}\mid g_i=\lambda_{i+1}-\lambda_i\}$ $(1\leqslant i\leqslant n-1)$;判定 g_k 是否满足 $\{g_{k-1}<g_k\ \&\ g_k>g_{k+1}\}$,如果满足,则直接令 $M=k$,否则转至下一步;令 i 从 k 开始依次递增,直至 $i=l_1$ 时,满足条件 $\{g_{l-1}<g_l\ \&\ g_l>g_{l+1}\}$,即找到当 $i>k$ 时,序列 G 的第1个极大值,继续令 i 递增,并找到序列 G 的第2个极大值,此时 $i=l_2$。则参数 M 的取值范围为 $[l_1+1,l_2+1]$,其中 $k<l_1<l_2<n$。

d. 按上述步骤并结合实际数据选定参数 M 后,求取前 M 个特征值对应的特征向量矩阵,单位化后形成矩阵 $\boldsymbol{Y}$,用于聚类的特征向量空间 S。

1.2 遗传聚类优化算法

由1.1.2节可得到前 M 个特征向量构成的特征矩阵 $\boldsymbol{Y}$,传统 NJW 算法采用 K-means 算法对矩阵 $\boldsymbol{Y}$ 按行进行聚类,即当且仅当 $\boldsymbol{Y}$ 的第 i 行被划分到类 j 中时,将数据点 X_i 划分到聚类 j 中。

由于 K-means 算法对初始中心的选取十分敏感,其爬山式的寻优算法往往不能得到全局最优解,因而导致 NJW 算法的性能不稳定。将遗传算法与聚类算法相结合可以提高聚类算法全局寻优能力[21],因而本文将遗传聚类算法引入 NJW 算法中,替代算法最后一步的 K-means 聚类,完成对特征向量的聚类,具体操作如下文所述。

1.2.1 个体编码与种群初始化

本文算法采用对簇心序列进行实数编码的形式,并结合随机分配的方法生成初始种群,步骤如下:对于给定的聚类数目 k,为初始样本集 X 的每个样本随机指定其所属类号,再将每个样本依次按类号划入其所属簇中;定义所有类的簇心序列的集合为个体,每个簇心序列对应为该算子中的基因;对于指定的种群数目 N,重复上述步骤 N 次,可得到 N 个不同的个体的集合并完成种群初始化。

1.2.2 适应度函数

遗传聚类算法的目标函数 T 与 K-means 算法保持一致,即最小化所有样本数据的总类内方差,则定义算法的适应度函数如下:

$$f=\left[\frac{1-(T_i-T_{\min})}{T_{\max}-T_{\min}+0.000\ 1}\right]^{\alpha} \tag{1}$$

其中,$1\leqslant i\leqslant N$;T_i 为种群中第 i 个个体的目标函数值;$T_{\min}$ 为种群中所有个体的最小目标函数值;$T_{\max}$ 为所有个体的最大目标函数值;α 为适应值淘汰加速指数,本文取 $\alpha=2$。

1.2.3 等位基因匹配

本文算法的选择、交叉、变异过程与传统遗传算法保持一致。但考虑到基因(即类心序列)在个体中排列的无序性,为防止个体间的错误交叉,本文在算法交叉过程前对需要交叉的两个体(如个体 A、B)先进行等位基因排序,即将个体间距离最相近的各基因逐位匹配,具体步骤如下。

首先生成两个体间的对应基因位的距离矩阵 $\boldsymbol{D}$:

$$\boldsymbol{D}=\begin{bmatrix} D_{11} & \cdots & D_{1j} & \cdots & D_{1k} \\ \vdots & & \vdots & & \vdots \\ D_{i1} & \cdots & D_{ij} & \cdots & D_{ik} \\ \vdots & & \vdots & & \vdots \\ D_{k1} & \cdots & D_{kj} & \cdots & D_{kk} \end{bmatrix} \tag{2}$$

其中,D_{ij}为个体 A 第 i 位基因 A^i 与个体 B 第 j 位基因 B^j 间的欧氏距离。

其次找出矩阵 $\boldsymbol{D}$ 中最小的元素,若为 D_{ij},则分别视基因 A^i 与基因 B^j 为配对基因,同时将矩阵中的第 i 行与第 j 列元素都置 0。

然后找出矩阵 $\boldsymbol{D}$ 剩余元素中的最小非零元素,若为 $D_{i^*j^*}$,其中 $1\leqslant i^*\neq i\leqslant k$ 且 $1\leqslant j^*\neq j\leqslant k$,同理配对个体 A 的第 i^* 位与个体 B 的第 j^* 位基因,并将该元素所属行与列元素置 0。

重复上述步骤,直至完成两个体的所有基因配对。

1.2.4　类心的重新划分

本文每一代种群中的个体在经过选择、交叉、变异操作后,重新计算原始样本集 V 中各序列与各类心序列间的距离,按距离最近原则重新指定各样本的类号并确定类心,将新个体作为本次迭代的最终结果代入到下一次迭代中[21]。

1.2.5　遗传聚类算法步骤

本文遗传聚类算法的相关参数设置如下:每一代的种群规模为 100,算法的终止迭代次数为 200,交叉概率为 0.5,变异概率为 0.001。算法步骤如下:

a. 将 1.1.2 节生成的特征向量矩阵按行划分为聚类空间中特征样本集,并采用 1.2.1 节的方法为个体进行编码并生成初始种群;

b. 按照 1.2.2 节的方法计算初始种群的适应度函数,并设置迭代次数 $i=1$;

c. 参照 1.2.3、1.2.4 节方法,结合传统遗传算法对第 i 代的种群进行选择、交叉、变异操作并重新划分类心,得到新一代的种群;

d. 判定迭代次数是否达到终止迭代次数,如果达到则令 $i=i+1$,转至步骤 **c**,否则停止迭代过程,将种群中适应度函数最大的个体作为遗传聚类算法的最终结果;

e. 将初始特征样本集按离最终结果(类心序列)距离最近原则重新指定各样本序列的类号,至此得到特征向量空间的聚类划分结果,并将该结果映射回原始数据,至此本文算法结束。

2　聚类有效性评价

2.1　聚类质量评价指标

聚类质量评价指标一般分为外部评价指标与内部评价指标。外部评价指标是指先将聚类算法应用于有明确类别的标准测试数据集,再用有关指标去统计算法在该数据集上划分的准确率;而内部评价指标是指根据预先定义的评价标准,通常是描述聚类划分后簇的一些固有特征及量值,来评价聚类结果的质量。

典型的外部聚类评价指标有 FM(Fowlkes-Mallows)指标[22]、AR(Adjusted-Rand)指标[17]等,它们均是用于测量聚类结果与真实类属信息一致性的经典指标。内部评价指标本文选取误差平方和 SSE(Sum of Squared Error)指标。

2.1.1　外部评价指标

现假设标准测试集共有 k 个类别,包含 n 个样本数据,聚类后亦划分为 k 个簇,现对相关参数进行说明,如表 1 所示。表中,$V_i(1\leqslant i\leqslant k)$ 为标准数据中第 i 类簇中的样本数量;$U_j(1\leqslant j\leqslant k)$ 为聚类划分后第 j 类簇中的样本数量;$n=\sum_{i=1}^{k}V_i=\sum_{j=1}^{k}U_j$,为样本总数量;$n_{ij}$为 V_i 与 U_j 中相同样本的数量,且 $n_{i\cdot}=\sum_{j=1}^{k}n_{ij}$,$n_{\cdot j}=\sum_{i=1}^{k}n_{ij}$。

表 1　聚类划分前、后的簇

Table 1　Groups before and after clustering

标准	相同样本数量				每行和
	U_1	U_2	…	U_k	
V_1	n_{11}	n_{12}	…	n_{1k}	$n_{1\cdot}$
V_2	n_{21}	n_{22}	…	n_{2k}	$n_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
V_k	n_{k1}	n_{k2}	…	n_{kk}	$n_{k\cdot}$
每列和	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot k}$	

a. FM 指标。

令 $Z=\sum_{i=1}^{k}\sum_{j=1}^{k}n_{ij}^2$,FM 指标的计算公式如下:

$$I_{\mathrm{FM}}=\frac{Z-n}{2\left[\sum_{i=1}^{k}\binom{n_{i\cdot}}{2}\sum_{j=1}^{k}\binom{n_{\cdot j}}{2}\right]^{0.5}} \tag{3}$$

FM 指标的值介于 0~1,且其值越大,表示聚类划分后的簇与标准簇越接近,当且仅当矩阵 $\boldsymbol{N}$(由表 1 中的元素 n_{ij} 构成)的每行每列中仅有一个非零元素并且该元素值等于标准簇 $V_i(1\leqslant i\leqslant k)$,即聚类结果与标准簇完全一致时,$I_{\mathrm{FM}}=1$。

b. AR 指标。

同样地,基于表 1 中的内容,并另设相关参数如下:

$$\begin{cases} a=\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2},\ b=\sum_{i=1}^{k}\binom{n_{i\cdot}}{2}-a \\ c=\sum_{j=1}^{k}\binom{n_{\cdot j}}{2}-a,\ d=\binom{n}{2}-a-b-c \end{cases} \tag{4}$$

当 $0\leqslant n\leqslant 1$ 时,令排列组合公式$\binom{n}{2}=0$。AR 指标的计算公式如下:

$$I_{\mathrm{AR}}=\frac{\binom{n}{2}(a+d)-[(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2-[(a+b)(a+c)+(c+d)(b+d)]} \tag{5}$$

与FM指标一样,AR指标值也是在0~1范围内,且其值越靠近1,划分结果越好,并且在标准划分的情况下,$I_{AR}=1$。

2.1.2 内部评价指标

SSE指标为聚类后所有子类中各数据点到对应的聚类中心的距离平方和,具体计算公式如下:

$$I_{SSE}=\sum_{i=1}^{k}\sum_{x\in G_i}\|\boldsymbol{x}-\boldsymbol{o}_i\|^2 \tag{6}$$

其中,$\boldsymbol{o}_i$为类簇G_i的聚类中心;$\boldsymbol{x}$为类簇G_i中的数据点。传统意义上认为,I_{SSE}越小,聚类质量越好。

2.2 聚类稳健性与全局寻优能力指标

聚类算法的稳健性可以定义为同一个算法在相同或相似数据上多次聚类后,其结果相似程度的大小;而全局寻优能力则定义为这些结果是否基本稳定在一个较高的水平的能力。因此算法稳健性与全局寻优能力越高,在实际应用中的可靠性也越好。

选定聚类质量评价指标$I^{(+)}$($I^{(+)}$越大表示聚类结果越好),本文算法拟在同一个数据上重复测试c_I次,则定义聚类稳健性与全局寻优能力指标分别如下:

$$I_{mean}^{(+)}=\frac{1}{c_I}\sum_{i=1}^{c_I}I_i \tag{7}$$

$$I_{SI}^{(-)}=\sum_{i=1}^{c_I}\|I_i-I_{mean}\|^2 \tag{8}$$

其中,I_i为聚类算法在第i次测试的评价结果;I_{mean}和I_{SI}分别为c_I次聚类结果的均值和方差。因此I_{mean}越大,同时I_{SI}越小,则表示该算法稳健性与全局寻优能力越好。

3 算例分析

本文将所提方法分别用于测试UCI标准合成时间序列数据[23]与美国最大区域电网运营商PJM官网提供的实际负荷数据[24]中,并对聚类结果进行综合评价。

3.1 标准时间序列测试

本节选取UCI标准测试数据库中的一套通用合成时间序列数据集作为本文所提算法的验证数据。

该标准数据集由600条时间序列组成,共分为标准、循环、上升趋势、下降趋势、陡升与陡降6类曲线,每类曲线数量均为100条,且长度都为60。附录中图A1给出了每类时间序列曲线的包络线与各类典型曲线的形状。

3.1.1 参数选取

将标准时间序列数据代入本文算法中,并将Laplacian矩阵特征值按从大到小顺序排序,结果如图1所示。

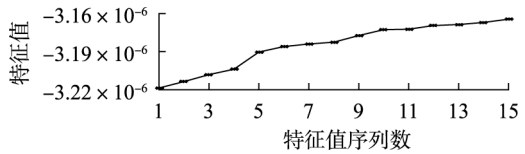


图1 Laplacian矩阵的前15个特征值

Fig.1 First 15 eigenvalues of Laplacian matrix

从图1中可以明显看出,第$k(k=6)$个与第$k+1$个特征值之间的差值并不大。得到特征值的差值序列如图2所示。

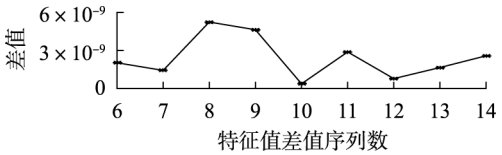


图2 Laplacian矩阵特征值的差值序列

Fig.2 Difference series of Laplacian matrix eigenvalues

从图2中可以明显看到,$g_6<g_5$,而g_8为第1个极大值,g_{11}为第2个极大值,根据本文方法选择M的取值范围为9~12,结合实际情况,本文取$M=11$。

3.1.2 聚类质量评价

从附录中图A1可以看出,这6类标准时间序列曲线具有很明显的区分度,下文将采用本文所提算法与*K*-means算法、FCM算法、层次聚类算法及基于自组织映射SOM(Self-Organizing Maps)神经网络的聚类算法(下文简称SOM算法)对该标准数据集分别进行聚类测试,并结合FM指标、AR指标及SSE指标对各自聚类结果分别进行评估。此外,本文也针对标准时间序列集计算上述指标值,以作为衡量标准。为了避免偶然性,对上述算法均重复做多次测试,并选取其中最好的结果,如表2所示。

表2 标准数据测试结果

Table 2 Test results of standard data

算法	FM指标	AR指标	SSE指标
K-means算法	0.694 7	0.620 1	$9.672\ 0\times10^5$
FCM算法	0.682 4	0.619 3	$1.010\ 7\times10^6$
层次聚类算法	0.677 2	0.563 8	$1.024\ 5\times10^6$
SOM算法	0.610 2	0.518 1	$1.035\ 6\times10^6$
本文算法	0.940 4	0.928 5	$1.097\ 7\times10^6$
标准时间序列集	1.000 0	1.000 0	$1.085\ 2\times10^6$

从表2中可以看出,本文所提算法的FM与AR指标评价结果均要优于*K*-means算法、FCM算法、层次聚类算法与SOM算法,并且指标值均几乎接近于1;本文算法下的SSE指标值均大于其他4种对比算法,但同时注意到,本文算法下的类内方差实际上更接近于标准数据的结果,而且标准数据的类内方差也均大于其他4种对比算法。由此可见,虽然采用*K*-means、FCM、层次聚类及SOM算法取得了更小的类内方差,但实际聚类效果却不如本文算法。为了直观地体现本文所提算法在该标准数据集上的分类效果,本文给出了算法聚类前、后簇的相同元素数目表,如表3所示。

表 3　本文算法下聚类前、后的簇

Table 3　Groups before and after clustering under proposed algorithm

标准	相同样本数量					
	U_1	U_2	U_3	U_4	U_5	U_6
V_1	0	**100**	0	0	0	0
V_2	0	0	**100**	0	0	0
V_3	0	0	0	**99**	1	0
V_4	1	2	1	0	0	**96**
V_5	0	5	0	0	**95**	0
V_6	**92**	8	0	0	0	0

从表 3 中可看出,本文所提算法较好地实现了标准数据的分类,其中标准数据的第 1、2、3、4、5、6 类簇分别对应着聚类后的第 2、3、4、6、5、1 类簇,聚类后的簇中只有极少数数据被错误分类。表 4—7 也分别给出了 *K*-means、FCM、层次聚类与 SOM 算法的分类情况。

从表 4—7 中不难看出,*K*-means 算法几乎无法区分标准数据的第 3 类与第 5 类簇;而 FCM 算法虽然基本可以区分标准数据的每个类,但却有较多的数据被错误分类;层次聚类算法无法区分标准数据

表 4　*K*-means 算法下聚类前、后的簇

Table 4　Groups before and after clustering under *K*-means algorithm

标准	相同样本数量					
	U_1	U_2	U_3	U_4	U_5	U_6
V_1	**100**	0	0	0	0	0
V_2	0	0	**58**	42	0	0
V_3	0	**100**	0	0	0	0
V_4	0	0	0	0	**61**	39
V_5	4	**96**	0	0	0	0
V_6	0	0	0	0	16	**84**

表 5　FCM 算法下聚类前、后的簇

Table 5　Groups before and after clustering under FCM algorithm

标准	相同样本数量					
	U_1	U_2	U_3	U_4	U_5	U_6
V_1	**96**	0	0	0	4	0
V_2	4	0	0	0	**96**	0
V_3	0	0	**64**	36	0	0
V_4	0	34	0	0	0	**66**
V_5	1	0	29	**70**	0	0
V_6	0	**74**	0	0	1	25

表 6　层次聚类算法下聚类前、后的簇

Table 6　Groups before and after clustering under hierarchical clustering algorithm

标准	相同样本数量					
	U_1	U_2	U_3	U_4	U_5	U_6
V_1	0	0	0	**100**	0	0
V_2	12	22	25	**41**	0	0
V_3	0	0	0	0	0	**100**
V_4	0	0	0	0	**100**	0
V_5	0	0	0	0	0	**100**
V_6	0	0	0	0	**100**	0

表 7　SOM 算法下聚类前、后的簇

Table 7　Groups before and after clustering under SOM algorithm

标准	相同样本数量					
	U_1	U_2	U_3	U_4	U_5	U_6
V_1	0	0	**100**	0	0	0
V_2	0	0	**100**	0	0	0
V_3	41	**59**	0	0	0	0
V_4	0	0	0	0	**61**	39
V_5	**42**	17	0	41	0	0
V_6	0	0	0	0	20	**80**

的第 3、5 类与第 4、6 类簇;SOM 算法无法区分标准数据的第 1 类与第 2 类簇,同时其他类的数据也有较多被错误划分。综上所述,本文算法在标准数据上的聚类质量要高于 *K*-means、FCM、层次聚类与 SOM 算法。

3.1.3　*算法全局寻优能力与稳健性分析*

将 NJW 算法($M=11$ 或 6)、*K*-means 算法、FCM 算法与本文算法均重复对该标准数据集做 20 次测试,并选用 FM 指标来评价聚类结果,结果如图 3 所示。

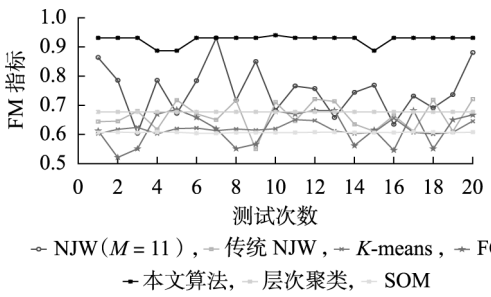


图 3　不同算法下多次测试指标对比

Fig.3　Comparison of multiple test indicators under different algorithms

从图 3 中可以直观地看出,采用传统 NJW 算法时无法收敛到全局最优解,当选取前 $M=11$ 个特征向量时,NJW 算法的聚类质量有了明显提高,但 20 次聚类结果相差较大,收敛到全局最优解的概率也较低,而本文所提算法以很大概率收敛到全局最优解,这验证了本文用遗传聚类算法取代 NJW 算法的最后一步 *K*-means 聚类后,算法的全局寻优能力与稳健性均得到了较大的提高。此外不难发现,*K*-means、FCM、层次聚类与 SOM 算法 20 次的聚类质量均明显低于本文算法。

为定量分析上述算法 20 次的聚类结果,选取 2.2 节中的指标对上述结果进行综合评价,结果见表 8。

从表 8 中可见,本文算法下的 I_{mean} 指标均高于其他算法,指标值 I_{SI} 仅高于层次聚类算法与 SOM 算法,这验证了本文算法具有较高的稳健性与全局寻优能力。此外,虽然层次聚类算法与 SOM 算法的指标值 I_{SI} 较小,但其 I_{mean} 指标却较低,这说明了层次聚类算法与 SOM 算法虽然稳健性较高,但全局寻优能力却不如本文算法。

表8 算法稳健性与全局寻优能力

Table 8 Stability and global optimization ability of algorithms

算法	$I_{\text{mean}}^{(+)}$	$I_{\text{SI}}^{(-)}$
本文算法	0.924 3	0.005 21
传统 NJW 算法	0.662 8	0.045 74
NJW 算法(选取 $M=11$)	0.752 2	0.135 92
K-means 算法	0.625 0	0.009 86
FCM 算法	0.604 2	0.056 07
层次聚类算法	0.677 2	0
SOM 算法	0.606 9	0.000 03

3.2 实际负荷数据集测试

实际负荷数据来源于 PJM 公司官网 Midatl 区域 2017 年 3 月 17 日至 2018 年 3 月 18 日全年的日负荷数据,如附录中图 A2 所示。

参照 3.1 节标准数据集分类数目,首先设置聚类数目为 6 类,采用本文算法聚类后,各类簇曲线轮廓如图 4 所示。

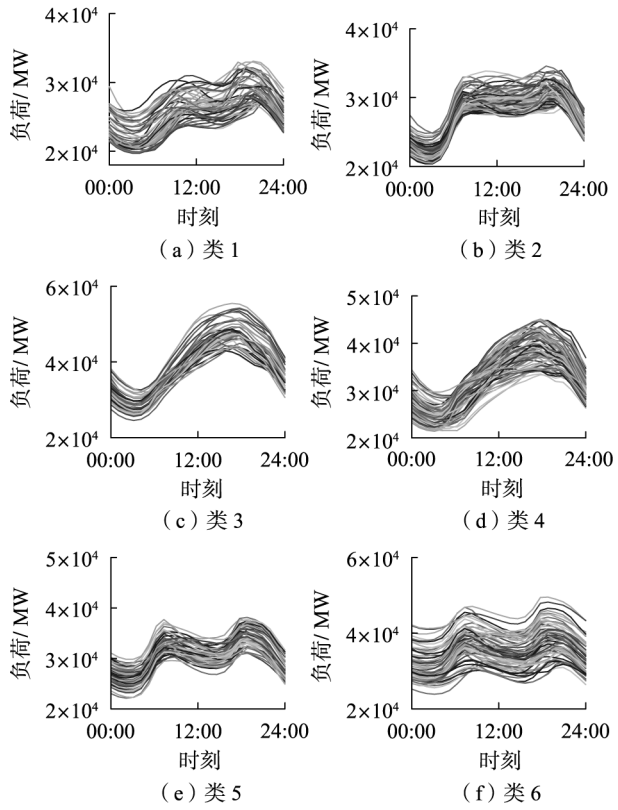


图4 本文算法下的聚类结果(6类)

Fig.4 Clustering results under proposed algorithm (six categories)

由图 4 可见,本文所提算法较好地实现了全年日负荷曲线的聚类划分,每类簇中的曲线形态特征基本一致,但不难发现,其中类 3 与类 4 以及类 5 与类 6 的曲线轮廓较为相似,故进一步设置聚类数目为 4 类,并继续采用本文算法聚类,结果如图 5 所示。

由图 5 可见,本文算法将数据聚为 4 类后,各类间曲线形态差异明显,各类中曲线形态特征较为相似。此外可以看出,之前的类 3 与类 4 合并为现在的类 3,类 5 与类 6 也合并为此时的类 4,另外类 1 与类 2 基本不变,这一方面说明了将该日负荷数据集划分为 4 类较为合理,另一方面也验证了本文算法在实际负荷数据中也具有良好的聚类划分性能。本文也同时采用 *K*-means 算法、FCM 算法、层次聚类算法与 SOM 算法对该负荷数据进行测试对比,结果如附录中图 A3—A6 所示。

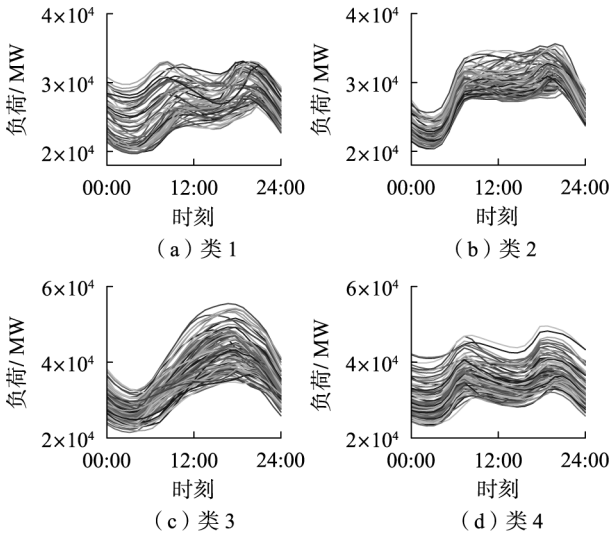


图5 本文算法聚类结果(4类)

Fig.5 Clustering results under proposed algorithm (four categories)

由以上结果可见,无论是采用 *K*-means、FCM 算法,还是 SOM 算法,其聚类后各簇内的曲线形态差异均较为明显,而各簇间的曲线轮廓却较为相似;层次聚类算法下各簇曲线数目相差较大,簇内曲线形态差异亦为明显。这进一步表明了 *K*-means、FCM、层次聚类与 SOM 算法在实际负荷数据中的聚类效果亦不如本文算法。

4 结论

本文提出了一种结合谱聚类特征向量选取与遗传算法的全局性收敛聚类算法,对传统 NJW 谱聚类算法做了以下改进:提出了一种选取谱聚类特征向量个数方法,用于指导实际工程应用中特征向量个数的合理选取范围;采用遗传聚类优化算法取代传统 NJW 谱聚类算法的 *K*-means 聚类过程,增强了算法的全局寻优能力与稳健性。

以聚类质量评价指标 FM、AR、SSE 和聚类稳定性、全局寻优能力指标为标准,通过标准分类数据集与实际负荷数据集验证了本文算法比现有的 *K*-means、FCM、层次聚类与 SOM 算法具有更好的分类效果、聚类质量和全局寻优性能,并同时拥有良好的稳健性。

附录见本刊网络版(http://www.epae.cn)。

参考文献:

[1] 张东霞,苗新,刘丽平,等. 智能电网大数据技术发展研究[J]. 中国电机工程学报,2015,35(1):2-12.
ZHANG Dongxia,MIAO Xin,LIU Liping,et al. Research on development strategy for smart grid big data[J]. Proceedings of the CSEE,2015,35(1):2-12.

[2] HONG T,CHEN C,HUANG J,et al. Guest editorial big data analytics for grid modernization[J]. IEEE Transactions on Smart Grid,2016,7(5):2395-2396.

[3] 周开乐,杨善林,丁帅,等. 聚类有效性研究综述[J]. 系统工程理论与实践,2014,34(9):2417-2431.
ZHOU Kaile,YANG Shanlin,DING Shuai,et al. On cluster validation[J]. Systems Engineering-Theory & Practice,2014,34(9):2417-2431

[4] 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.

[5] 韩俊,谈健,黄河,等. 基于改进 *K*-means 聚类算法的供电块划分方法[J]. 电力自动化设备,2015,35(6):123-129.
HAN Jun,TAN Jian,HUANG He,et al. Power-supplying block partition based on improved *K*-means clustering algorithm[J]. Electric Power Automation Equipment,2015,35(6):123-129.

[6] 吴军基,杨伟,葛成,等. 基于 GSA 的肘形判据用于电力系统不良数据辨识[J]. 中国电机工程学报,2006,26(22):23-28.
WU Junji,YANG Wei,GE Cheng,et al. Application of GSA-based elbow judgment on bad-data detection of power system[J]. Proceedings of the CSEE,2006,26(22):23-28.

[7] 徐志超,杨玲君,李晓明. 基于聚类改进 S 变换与直接支持向量机的电能质量扰动识别[J]. 电力自动化设备,2015,35(7):50-58.
XU Zhichao,YANG Lingjun,LI Xiaoming. Power quality disturbance identification based on clustering-modified S-transform and direct support vector machine[J]. Electric Power Automation Equipment,2015,35(7):50-58.

[8] 丁志勇,杨苹,杨曦,等. 基于连续时间段聚类的支持向量机风电功率预测方法[J]. 电力系统自动化,2012,36(14):131-135.
DING Zhiyong,YANG Ping,YANG Xi,et al. Wind power prediction method based on sequential time clustering support vector machine[J]. Automation of Electric Power Systems,2012,36(14):131-135.

[9] 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.

[10] 丁明,解蛟龙,刘新宇,等. 面向风电接纳能力评价的风资源/负荷典型场景集生成方法与应用[J]. 中国电机工程学报,2016,36(15):4064-4071.
DING Ming,XIE Jiaolong,LIU Xinyu,et al. The generation method and application of wind resources/load typical scenario set for evaluation of wind power grid integration[J]. Proceedings of the CSEE,2016,36(15):4064-4071.

[11] 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.

[12] 宋易阳,李存斌,祁之强. 基于云模型和模糊聚类的电力负荷模式提取方法[J]. 电网技术,2014,38(12):3378-3383.
SONG Yiyang,LI Cunbin,QI Zhiqiang. Extraction of power load patterns based on cloud model and fuzzy clustering[J]. Power System Technology,2014,38(12):3378-3383.

[13] 王洪涛,刘旭,陈之栩,等. 低碳背景下基于改进通用生成函数法的随机生产模拟[J]. 电网技术,2013,37(3):597-603.
WANG Hongtao,LIU Xu,CHEN Zhixu,et al. Power system probabilistic production simulation based on improved universal generating function methods in low-carbon context[J]. Power System Technology,2013,37(3):597-603.

[14] WU W,PENG M. A data mining approach combining *K*-means clustering with bagging neural network for short-term wind power forecasting[J]. IEEE Internet of Things Journal,2017,4(4):979-986.

[15] 苏适,李康平,严玉廷,等. 基于密度空间聚类和引力搜索算法的居民负荷用电模式分类模型[J]. 电力自动化设备,2018,38(1):129-136.
SU Shi,LI Kangping,YAN Yuting,et al. Classification model of residential power consumption mode based on DBSCAN and gravitational search algorithm[J]. Electric Power Automation Equipment,2018,38(1):129-136.

[16] 宋辞,裴韬. 基于特征的时间序列聚类方法研究进展[J]. 地理科学进展,2012,31(10):1307-1317.
SONG Ci,PEI Tao. Research progress in time series clustering methods based on characteristics[J]. Progress in Geography,2012,31(10):1307-1317.

[17] YAO X,WEI H L. Improving *K*-means clustering performance using a new global time-series averaging method[C]// Proceedings of the 9th International Conference on Electronics,Computers and Artificial Intelligence(ECAI). Targoviste,Romania:IEEE,2017:1-6.

[18] ZHENXIN Q,MENGZHU W,KEJIA T. Research of spectral clustering on trend of big time series[C]// Proceedings of the 4th International Conference on Information Science and Control Engineering(ICISCE). Changsha,China:IEEE,2017:562-568.

[19] LUXBURG U V. A tutorial on spectral clustering[J]. Statistics and Computing,2007,17(4):395-416.

[20] REBAGLIATI N,VERRI A. Spectral clustering with more than K eigenvectors[J]. Neurocomputing,2011,74(9):1391-1401.

[21] 金微. 基于遗传算法的 *K*-means 聚类方法的研究[D]. 南京:河海大学,2007.
JIN Wei. Research on *K*-means clustering method based on genetic algorithm[D]. Nanjing:Hohai University,2007.

[22] DUDOIT S,FRIDLYAND J. A prediction-based resampling method for estimating the number of clusters in a dataset[J]. Genome Biology,2002,3(7):1-21.

[23] HETTICH S,BAY S D. The UCI KDD archive data[DB/OL]. [2018-03-18]. http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.data/.

[24] PJM. Energy market hourly load data[EB/OL]. [2018-03-18]. http://www.pjm.com/markets-and-operations/energy/real-time/hourly-prelim-loads.aspx/.

作者简介:


丁　明

丁　明(1956—),男,安徽合肥人,教授,博士研究生导师,博士,研究方向为电力系统可靠性与安全防御、可再生能源与分布式发电系统、电力电子技术在电力系统中的应用等(**E-mail**:mingding56@126.com);

黄　冯(1993—),男,安徽六安人,硕士研究生,研究方向为新能源与分布式发电技术(**E-mail**:huangfeng93@foxmail.com);

宋晓皖(1994—),女,安徽六安人,硕士研究生,研究方向为新能源与分布式发电技术(**E-mail**:xiaowansongst@163.com)。

(下转第 114 页 continued on page 114)

[J]. IEEE Transactions on Power Electronics,2016,32(4):2622-2630.
[10] KOMURCUGI H,ALTIN N,OZDEMIR S,et al. Lyapunov-function and proportional-resonant-based control strategy for single-phase grid-connected VSI with LCL filter[J]. IEEE Transactions on Power Electronics,2016,63(5):2838-2849.
[11] HU Sijia,ZHANG Zhiwen,LI Yong,et al. A new half-bridge winding compensation-based power conditioning system for electric railway with LQRI[J]. IEEE Transactions on Power Electronics,2014,29(10):5242-5256.
[12] XIE Bin,ZHANG Zhiwen,HU Sijia,et al. YN/VD connected balance transformer-based electrical railway negative sequence current compensation system with passive control scheme[J]. IET Power Electronics,2016,9(10):2044-2051.
[13] LI Yong,LIU Qianyi,HU Sijia,et al. A virtual impedance comprehensive control strategy for the controllably inductive power filtering system[J]. IEEE Transactions on Power Electronics,2017,32(2):920-926.
[14] 宁志毫,罗隆福,张志文,等. 节能滤波型变压器及其整流系统关键问题研究[J]. 电力自动化设备,2012,32(4):20-25.
NING Zhihao,LUO Longfu,ZHANG Zhiwen,et al. Key techniques of rectifier system based on energy-saving and filtering transformer[J]. Electric Power Automation Equipment,2012,32(4):20-25.

作者简介:


许加柱

许加柱(1980—),男,安徽来安人,教授,博士,主要研究方向为交直流电能变换新技术(**E-mail**:xujiazhu@126.com);

王 涛(1994—),男,湖南娄底人,硕士研究生,主要研究方向为交直流电能变换新技术以及储能装置在配电网中的应用(**E-mail**:473260354@qq.com);

崔贵平(1986—),男,湖南衡阳人,博士研究生,主要研究方向为储能装置的应用(**E-mail**:328736407@qq.com);

刘裕兴(1991—),男,湖南邵阳人,博士研究生,主要研究方向为双流制电力机车的牵引传动技术(**E-mail**:503821992@qq.com)。

A novel active power filtering method based on harmonic magnetic potential balance of transformer

XU Jiazhu,WANG Tao,CUI Guiping,LIU Yuxing
(College of Electrical and Information Engineering,Hunan University,Changsha 410082,China)

Abstract:To reduce the voltage withstand capability of APF(Active Power Filter) switch,exploit the capacity potential of both low-voltage large-current switches and devices,save the cost of boosting transformer which connected with APF and grid-side lines,and reduce the nominal voltage of DC stabilized voltage capacitor effectively,a novel active power filtering approach based on the harmonic magnetic potential balance of transformer is proposed. The active power filtering device is connected to the middle of the secondary side winding tap of transformer in this approach,the harmonic current magnetic potential produced by the device and load cancels out each other at the secondary side of transformer consequently,hence the grid-side harmonic current is significantly decreased. The detailed filtering mathematical principle associated with the harmonic magnetic potential balance in the secondary side winding of transformer is analyzed. Filtering simulation on widely used 10 kV/380 V Y/△ and △/Y transformers at the distribution network side is performed. Finally,the effectiveness of the proposed filtering approach is verified by experiments.

Key words:active power filter;power transformers;harmonic suppression;magnetic potential balance

(上接第99页 continued from page 99)

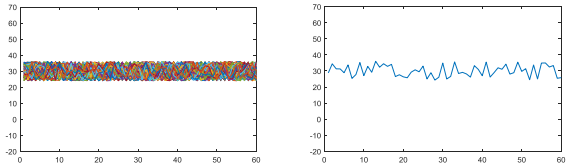
Power time series curve clustering method combining improved spectral clustering and genetic algorithm

DING Ming[1],HUANG Feng[1],ZOU Jiaxin[2],LIU Jinshan[2],SONG Xiaowan[1]
(1. Anhui Key Laboratory of New Energy Utilization and Energy Conservation,Hefei University of Technology,Hefei 230009,China;
2. Key Laboratory of Photovoltaic Power Generation and Grid Integration,State Grid Qinghai Electric Power Research Institute,Xining 810008,China)

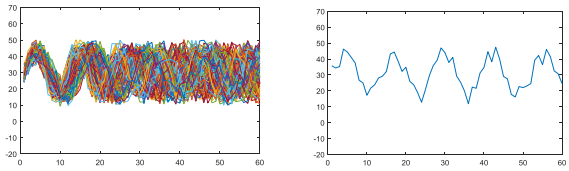
Abstract:To improve the clustering effect of traditional clustering algorithms on power time series data,a genetic spectrum clustering algorithm using selected optimal feature vectors is proposed. According to the characteristics of the application data structure,the extraction process of the feature vectors in the spectral clustering algorithm is optimized to avoid possible lack of data information suffered from traditional approaches. Then,the genetic clustering optimization algorithm is used to cluster the optimized feature vectors,and map the final division results back to the original data. The UCI standard synthetic time series data and the daily load data provided by the USA regional power grid operator PJM are employed as an example. Test results obtained from the traditional clustering algorithm and the proposed algorithm are compared and analyzed in terms of cluster validity indicators and morphological features. These results indicate that the proposed algorithm has remarkable classification effect,high clustering quality and robustness,which exhibits promising engineering applications.

Key words:time series data;spectral clustering;genetic algorithm;feature vector extraction;load clustering

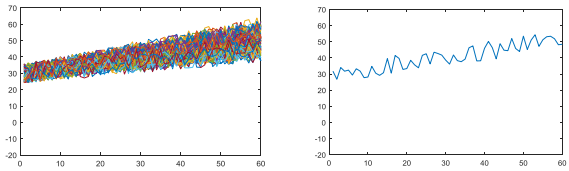
附 录



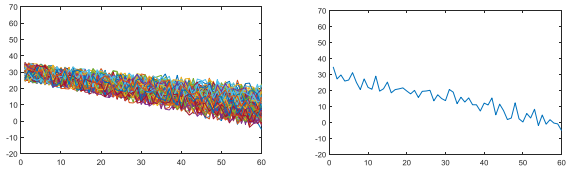
(a)　标准趋势



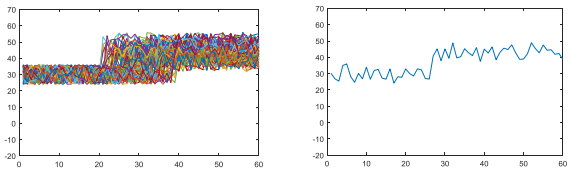
(b)　循环趋势



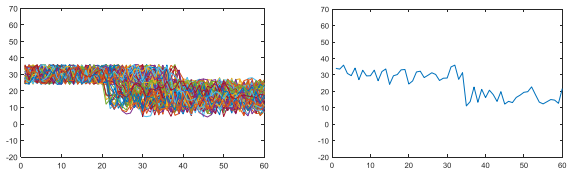
(c)　上升趋势



(d)　下降趋势



(e)　陡升趋势



(f)　陡降趋势

图 A1　标准合成时间序列集

Fig. A1　Standard synthetic control chart time series

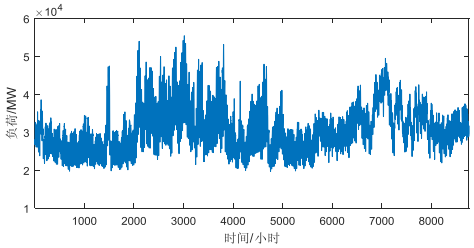


图 A2　全年日负荷数据

Fig. A2　Annual load data curve

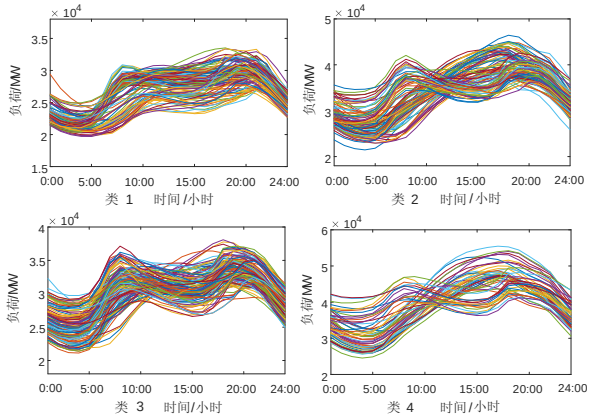


图 A3　*K*-means 算法聚类结果

Fig. A3 Clustering results of the *K*-means algorithm

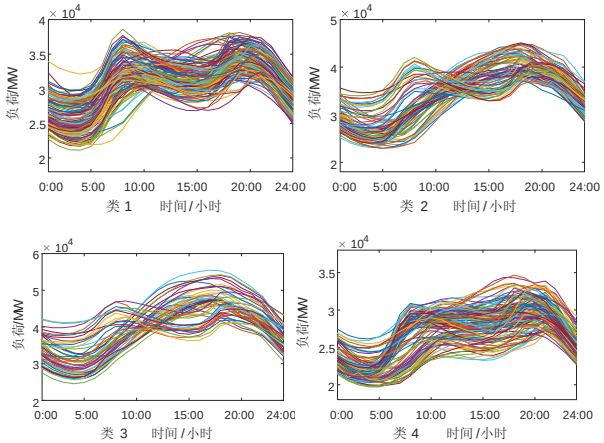


图 A4　FCM 算法聚类结果

Fig. A4　Clustering results of the FCM algorithm

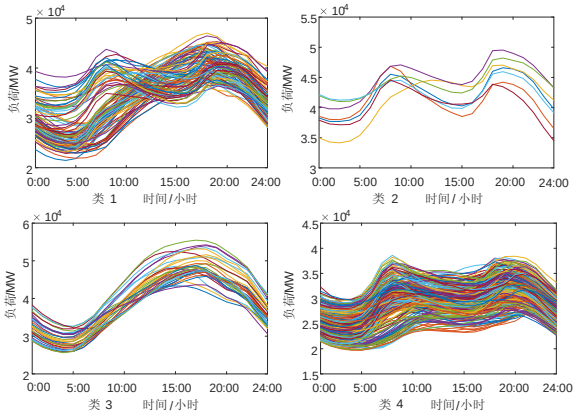
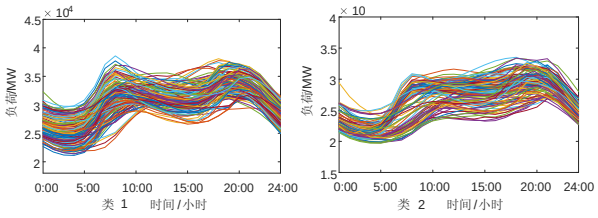


图 A5　层次聚类算法聚类结果

Fig.A5 Clustering results of the Hierarchical algorithm



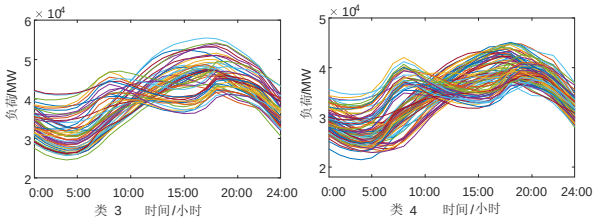


图 A6　SOM 算法聚类结果

Fig.A6　Clustering results of the SOM algorithm