第 43 卷 第 20 期 2019 年 10 月 25 日
电力系统自动化
Automation of Electric Power Systems
Vol.43 No.20 Oct.25,2019
DOI：10.7500/AEPS20181115008

基于特征指标降维及熵权法的日负荷曲线聚类方法

宋军英[1]，何　聪[2]，李欣然[2]，刘志刚[1]，汤　杰[2]，钟　伟[1]
（1. 国网湖南省电力有限公司，湖南省长沙市 410077；2. 湖南大学电气与信息工程学院，湖南省长沙市 410082）

摘要：日负荷曲线聚类是负荷建模背景下分析负荷特性的基础。针对现有聚类方法在聚类质量、聚类效率等方面的不足，综合运用模糊 C 均值及熵权法原理提出一种基于特征指标降维及熵权法的日负荷曲线聚类方法。首先提取日负荷率、日峰谷差率、日最大利用时间等 7 类降维特征指标替代各采样点负荷数据作为聚类输入；其次，引入熵权法自适应配置各特征指标的权重系数；最后，采用特征加权的模糊 C 均值聚类算法对用电日负荷曲线进行聚类。采用所提方法对某地区日负荷曲线进行聚类分析，算例结果表明该方法在运行效率、鲁棒性、聚类质量等方面具有一定的优越性，聚类结果能真实有效地反映负荷的实际用电特性。

关键词：特征指标降维；熵权法；加权模糊 C 均值算法；负荷曲线聚类

0　引言

电力系统仿真计算中，负荷模型的准确性很大程度上影响仿真结果的可信度，负荷时变性是建立精确负荷模型的瓶颈[1-2]。用户、变电站的负荷特性可由对应日负荷曲线充分体现，且同类用户、变电站具有相似的负荷特性。在某种程度上，变电站日负荷曲线可看作是其供电用户日负荷曲线的叠加[3]，若能实时对变电站所供用户日负荷曲线进行聚类分析，即可实时掌握其用电特性。因此，日负荷曲线聚类是分析变电站用电特性的关键，是解决负荷模型随机时变性的有效途径。

现有研究中，日负荷曲线聚类算法大致可分为直接法、间接法两种[4]。直接法首先对多维采样点功率值作极大值归一化，然后采用模糊 C 均值（FCM）或其改进方法并基于欧氏距离对负荷曲线进行聚类。文献[5]以预处理后的 48 维功率向量作为聚类输入，并基于改进 FCM 算法进行日负荷曲线聚类，文献[6]提出以云模型确定 FCM 算法的初始聚类中心及最佳聚类数，但以欧氏距离作为负荷的相似性判据，易忽略负荷曲线的形态特征对聚类的影响，聚类质量难以保证。此外，随着智能电网技术的快速发展，负荷数据量及维数大幅增加，直接聚类法在计算效率方面面临着巨大的挑战。

间接法结合离散小波变换（DWT）、离散傅里叶变换（DFT）、核主成分分析（KPCA）、奇异值分解（SVD）等数据降维方法对高维日负荷曲线进行降维处理，提取反映负荷曲线特征的降维特征指标进行聚类。文献[7-8]分别采用多级离散小波变换、离散傅立叶变换对日负荷数据进行低维映射从而提取负荷特征，但在降维过程中不可避免地会对原始曲线信息造成一定程度的损失，从而影响聚类质量；文献[9]采用奇异值分解法提取反映负荷特征的相关信息作为降维指标，但较难确定最佳降维指标数目，且权重系数配置方法很难全方位反映降维指标的变化规律及重要程度；文献[10]以核主成分分析对日负荷曲线进行降维处理，但要求数据结构具有全局线性分布，否则很难取得理想的降维效果；文献[11-12]分别利用分位数和差分算法、多维标度法提取原始日负荷曲线特征，提取步骤繁琐且同样存在日负荷曲线原始信息失真问题；文献[13]以日负荷率、日峰谷差率、峰期负载率、谷期负载率等负荷指标作为聚类输入向量，但所选指标尚无法最大限度地保证日负荷曲线的形态特征。文献[14-15]结合周、季度日负荷曲线数据提取特征指标进行聚类，时间跨度较大，聚类结果较难反映实时负荷特性。

针对现有文献在特征指标降维及权重配置方面存在的不足，同时兼顾聚类质量、效率、稳定性等要求，本文提出一种基于特征指标降维及熵权法的日负荷曲线聚类方法。本方法以物理意义明确的 7 个日负荷特征指标对日负荷曲线进行降维处理，并引入熵权法对各指标权重进行配置，以特征加权的模糊 C 均值（FW-FCM）聚类算法对用户日负荷曲线

收稿日期：2018-11-15；修回日期：2019-03-12。
上网日期：2019-05-15
国家自然科学基金资助项目（51577056）。

进行聚类。算例结果表明，本文选取的聚类特征指标合理，能充分反映曲线形态特征，聚类质量、效率较高且算法鲁棒性较好，聚类结果可为统计法综合负荷建模提供指导依据。

1　聚类特征指标选取及权重配置

1.1　负荷曲线降维必要性分析

负荷曲线是负荷用电特性的直观反映，是一类高维且可直观描绘的规则曲线，在进行欧氏距离计算时，不同曲线之间可能会出现不理想的等距性。

以图 1 为例，以曲线 A 和 B 模拟用户 A 和 B、曲线 C 模拟聚类中心曲线，不难发现，曲线 $A\rightarrow C$，$B\rightarrow C$ 的欧氏距离相等，若此时 A，B 到 C 的距离均小于其他曲线到 C 的距离，以距离作为相似性判据，A 和 B 将会被分入同一类中。然而负荷 A 和 B 的用电特性并不一致，即产生误分。上述是一种理想化情况，但在对大量日负荷曲线聚类时，类似误分的情况无法避免。此外，若一个数据集合对象的维数越高，距离测度的意义就越小[16]。因此，需对负荷曲线进行降维处理，提取能充分描述日负荷曲线特征的新指标用于负荷曲线聚类。

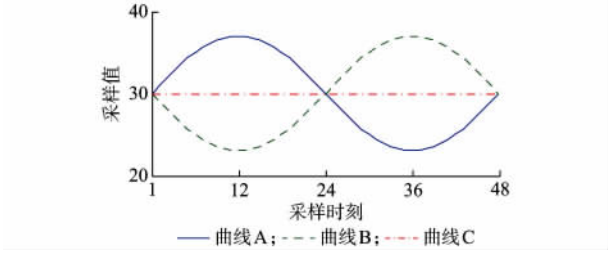


图 1　3 类负荷曲线
Fig.1　Three types of load curves

1.2　聚类特征指标选取

负荷用电特性一方面可用负荷曲线描述，另一方面可用负荷率、峰谷差率、峰期负载率等 10 余种特征指标[16]来体现。若以文献[13,17]所述日负荷率、日峰谷差率、日最大负荷利用时间、峰期、谷期、平期负载率等指标为依据对图 1 中曲线 A 和 B 进行划分，因各指标计算结果相同导致不能有效划分。但若在此基础上加入最大、最小负荷出现时间 2 个指标即可实现其有效划分。鉴于此，本文选取日负荷率、日峰谷差率、日最大负荷利用时间、峰期、谷期负载率、最大负荷出现时间、最小负荷出现时间等 7 个特征指标作为负荷曲线的降维指标。各指标定义及其物理意义见表 1。其中，$P(t)$ 为 t 时刻的功率值，P_{ave}，P_{max}，P_{min}，$P_{ave,peak}$，$P_{ave,low}$ 分别为日负荷功率平均值、最大值、最小值、峰期负荷平均值、谷期负荷平均值；根据负荷用电规律，选择峰期时段为：09:00—12:00，14:30—17:30，19:00—22:00；谷期时段为：22:00—24:00 和 00:00—06:00。

表 1　负荷特征指标
Table 1　Load characteristic index

特征指标	定义	反映物理意义
日负荷率	$k_1=P_{ave}/P_{max}$	全天负荷变化
日峰谷差率	$k_2=(P_{max}-P_{min})/P_{max}$	电网调峰能力
日最大负荷利用时间	$k_3=\left(\int_0^{24}P(t)\mathrm{d}t\right)/P_{max}$	时间利用效率
峰期负载率	$k_4=P_{ave,peak}/P_{ave}$	峰期负荷变化
谷期负载率	$k_5=P_{ave,low}/P_{ave}$	谷期负荷变化
最大负荷出现时间	$k_6=T_{max}$	
最小负荷出现时间	$k_7=T_{min}$	

1.3　特征指标权重配置

记 7 维日负荷特征指标的权重向量为 $\boldsymbol{W}=[w_1,\cdots,w_i,\cdots,w_7]$，$0<w_i<1$，且满足式(1)约束。

$$\sum_{i=1}^{7}w_i=1 \tag{1}$$

特征指标的权重系数能较为客观地反映各指标描述实际日负荷曲线特征的重要程度，其配置差异将在很大程度上影响负荷聚类质量。同时，针对不同的聚类曲线簇对象，同一特征指标的重要程度亦差异较大。为此，以特征指标作为聚类向量进行聚类需对其权重系数进行合理配置。本文采用熵权法原理自适应确定不同聚类曲线簇的特征指标权重。

熵权法广泛用于电能质量评估[18]、竞争力评价[19]等领域，其实质是以熵值的大小衡量指标间的差异进而确定各指标的权重。各评价对象中某项指标差异越大，对应指标的熵值越小，所涵盖的有效信息量越多，在评价对象中的位置越重要，其权值也将越大。本文以熵权法确定权重的具体步骤如下。

1)计算熵值 h_i。在有 m 个指标、n 个被评价对象的问题中，第 i 个指标的熵值可按下式确定。

$$h_i=-k\sum_{j=1}^{n}f_{ij}\ln f_{ij} \tag{2}$$

$$f_{ij}=\frac{r_{ij}}{\sum_{j=1}^{n}r_{ij}} \tag{3}$$

$$k=\frac{1}{\ln n} \tag{4}$$

式中：$i=1,2,\cdots,m$；r_{ij} 为实测数据；f_{ij} 为第 i 个指标下第 j 个被评对象的贡献度。

2)计算熵权 w_i。当同类指标的差异度较小时，h_i 计算结果趋近于 1，若采用式(5)所示的传统计算方法计算各指标的权重，误差较大，无实际指导意义。故本文以式(6)对熵权计算公式进行修正，公式定义详见文献[20]。

$$w_i=\frac{1-h_i}{m-\sum_{i=1}^{m}h_i} \tag{5}$$

$$w_i=\frac{\exp\left(\sum_{t=1}^{m}h_t+1-h_i\right)-\exp(h_i)}{\sum_{l=1}^{m}\left(\exp\left(\sum_{t=1}^{m}h_t+1-h_l\right)-\exp(h_l)\right)} \tag{6}$$

3）由式(2)至式(6)即可求解出7个特征指标对应的权重向量$\boldsymbol{W}$。

2　基于特征指标及权重的FW-FCM算法

2.1　数据降维

设$\boldsymbol{P}=[\boldsymbol{P}_1,\cdots,\boldsymbol{P}_i,\cdots,\boldsymbol{P}_N]^{\mathrm{T}}\in\mathbf{R}^{N\times n}$，$\boldsymbol{P}_i=[P_{i1},\cdots,P_{ij},\cdots,P_{in}]\in\mathbf{R}^{1\times n}$为$N$条日负荷曲线每天$n$个采样数据构成的初始负荷曲线矩阵(每条负荷曲线经处理后均不含缺失和异常数据)。利用表1中7个特征指标对矩阵$\boldsymbol{P}$做特征降维处理，得到$N\times 7$阶特征降维矩阵，记为$\boldsymbol{Y}$矩阵。

2.2　算法实现过程

FW-FCM算法以特征降维矩阵$\boldsymbol{Y}=[\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_i,\cdots,\boldsymbol{Y}_N]^{\mathrm{T}}\in\mathbf{R}^{N\times 7}$，$\boldsymbol{Y}_i=[Y_{i1},\cdots,Y_{ij},\cdots,Y_{i7}]\in\mathbf{R}^{1\times 7}$为输入进行聚类，因需计及权重向量$\boldsymbol{W}$的影响，FW-FCM算法与传统的FCM算法略有差异。具体实现步骤如下。

1）确定隶属度矩阵$\boldsymbol{U}$。设定聚类数$L(2\leqslant L\leqslant\sqrt{0.5N})$[21]，在$\boldsymbol{Y}$矩阵中随机选取$L$行数据作为初始聚类中心矩阵$\boldsymbol{V}=[\boldsymbol{V}_1,\cdots,\boldsymbol{V}_j,\cdots,\boldsymbol{V}_L]^{\mathrm{T}}\in\mathbf{R}^{L\times 7}$，$\boldsymbol{V}_j=[V_{j1},\cdots,V_{jk},\cdots,V_{j7}]^{\mathrm{T}}\in\mathbf{R}^{1\times 7}$。则$\boldsymbol{Y}_i$属于第$j$个聚类中心的隶属度$U_{ij}$如式(7)所示，隶属度矩阵$\boldsymbol{U}=[\boldsymbol{U}_1,\cdots,\boldsymbol{U}_i,\cdots,\boldsymbol{U}_N]^{\mathrm{T}}\in\mathbf{R}^{N\times L}$，$\boldsymbol{U}_i=[U_{i1},\cdots,U_{ij},\cdots,U_{iL}]^{\mathrm{T}}\in\mathbf{R}^{1\times L}$。

$$U_{ij}=\frac{1}{\sum_{l=1}^{L}\left[\frac{\sum_{k=1}^{m}(w_k d_k^2(Y_{ik},V_{jk}))}{\sum_{k=1}^{m}(w_k d_k^2(Y_{ik},V_{lk}))}\right]^{\frac{1}{q-1}}} \tag{7}$$

式中：$0\leqslant U_{ij}\leqslant 1$，$\sum_{j=1}^{L}U_{ij}=1$；$q\in[0,2]$为加权指数；$d_k^2(Y_{ik},V_{jk})$为$Y_{ik}$和$V_{jk}$的欧氏距离。

2）根据下式构建目标函数$F(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})$。

$$F(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})=\sum_{i=1}^{N}\sum_{j=1}^{L}\sum_{k=1}^{m}(U_{ij}w_k d_k^2(Y_{ik},V_{jk})) \tag{8}$$

3）更新聚类中心矩阵$\boldsymbol{V}$。若目标函数$F(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})$未达到最小则按式(9)更新聚类中心矩阵$\boldsymbol{V}$，并返回式(7)重新计算隶属度矩阵$\boldsymbol{U}$，直至目标函数$F(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})$达到最小。

$$V_{jk}=\frac{\sum_{i=1}^{n}((U_{ij})^q y_{ik})}{\sum_{i=1}^{n}(U_{ij})^q} \tag{9}$$

式中：$j=1,2,\cdots,L$；$k=1,2,\cdots,m$。

4）当目标函数$F(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})$达到最小时，即可由矩阵$\boldsymbol{U}$和$\boldsymbol{V}$获得聚类结果。

由上述公式不难发现，计及特征指标及权重的FW-FCM算法对隶属度矩阵、聚类中心矩阵、目标函数均做了精细化调整，无疑会在一定程度上提高聚类的质量。

2.3　聚类有效性检验

聚类有效性是指通过建立聚类有效性指标(PC指标、SC指标、XB指标、SSE指标、CHI指标、DBI指标等[9-10,12,22])对聚类质量进行评价并确定最佳聚类数。XB指标是模糊聚类中应用较为广泛的有效性指标，其定义详见文献[22]：当计及权重向量$\boldsymbol{W}$时，采用式(10)对XB指标进行修正。

$$\lambda_{\mathrm{XB}}=\frac{\sum_{i=1}^{L}\sum_{j=1}^{N}\sum_{k=1}^{m}U_{ij}^q w_k d_k^2(V_{ik},Y_{jk})}{N\min\limits_{i\neq l}\sum_{k=1}^{m}w_k d_k^2(V_{ik},V_{lk})}\quad l=1,2,\cdots,L \tag{10}$$

式(10)中，分子反映类内的紧凑程度，其值越小越紧凑，分母反映类间的分离程度，其值越大越分离越好。因此，指标XB的值越小，聚类效果越好。XB指标最小时对应的L即为最佳聚类数。

基于XB指标确定FW-FCM算法最佳聚类数的流程如图2所示。图中，$L_{\min}=2$，$L_{\max}=\sqrt{0.5N}$。

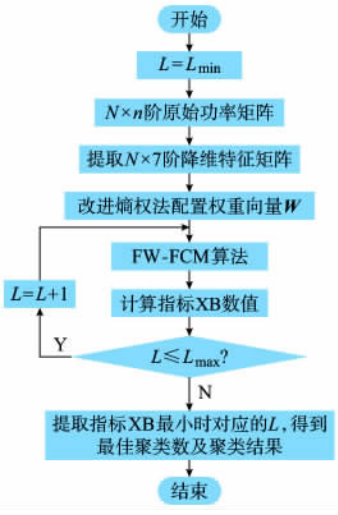


图2　基于XB指标确定FW-FCM算法最佳聚类数流程
Fig.2　Flow chart of determining optimal number of FW-FCM algorithms based on XB index

3　算例分析

为验证本文针对日负荷曲线聚类所提方法的有效性，本节算例的主要工作为：①分别以实际日负荷曲线数据和特殊日负荷曲线数据为基础，采用本文方法和传统聚类方法（本文将传统聚类方法定义为以标幺化的48维功率值作为特征向量，以FCM聚类算法进行聚类）进行聚类，并进行对比分析，同时检验本文选取XB指标作为聚类有效性判定的合理性；②构造模拟数据并加入随机扰动，验证本文算法的鲁棒性；③研究特征指标选取及权重配置差异对本文算法鲁棒性的影响。

3.1　实际日负荷曲线聚类

3.1.1　数据来源

以某市2016年8月22日实测476个典型用户（主要为重工业、轻工业、市政三产用户）的日负荷曲线为研究对象，采样频率为30 min/点，每条曲线共计48个功率量测点。剔除无效数据后（某负荷曲线中缺失或异常量达到采样量的20%，则认定该曲线无效），以文献[5]方法对局部缺失数据进行补全、以文献[23]方法对异常数据进行辨识和调整，获得394条典型用户（重工业89条、轻工业181条、市政三产124条）有效日负荷曲线，即394×48阶初始功率矩阵。

3.1.2　聚类结果及对比分析

提取每条日负荷曲线的7个特征指标，将394×48阶初始功率矩阵转化为394×7阶降维特征矩阵$\boldsymbol{Y}$。结合熵权法原理获得权重向量$\boldsymbol{W}$=[0.059，0.062，0.176，0.137，0.147，0.202，0.217]。以$\boldsymbol{Y}$矩阵及$\boldsymbol{W}$向量作为本文FW-FCM算法的输入，对该394条日负荷曲线进行分类（以下称为本文方法），并与传统聚类方法在聚类结果、聚类质量、聚类效率方面进行对比分析。同时为检验本文选取XB指标作为聚类有效性判定的合理性，计算不同聚类数下的Silhouette，CHI，DBI以及XB指标，并进行比较分析。Silhouette指标定义方法详见文献[9]，CHI指标、DBI指标定义方法详见文献[10]。

本文方法与传统方法在不同聚类数下的XB指标数值分布曲线见图3。附录A图A1为不同聚类数下上述4种有效性指标数值分布曲线；图A2、图A3为2种方法在最佳聚类数下的日负荷曲线聚类结果；图A4为2种方法下3类聚类中心曲线。

由图3不难发现，当聚类数为2～7时，本文方法所得XB指标均小于传统方法，且2种方法确定的最佳聚类数均为3；由附录A图A1可知，4种聚类有效性指标均显示最佳聚类数为3，从而验证了本文选取XB指标作为聚类有效性判定的合理性；由图A2、图A3可知，当聚类数为3时，两种方法下对应聚类类别的曲态分布基本一致。

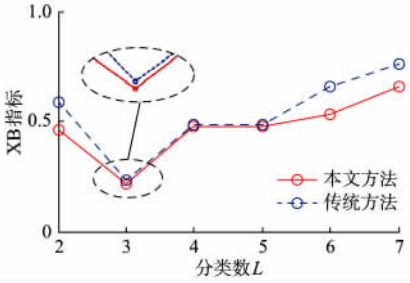


图3　采用2种方法在不同聚类数下的XB指标
Fig.3　XB indices with different classification numbers in two methods

本文方法中隶属于类别1，2，3的日负荷曲线条数分别为86，178，130，对应于传统方法分别为80，191，123。此外，按式(11)计算2种方法的分类准确率c（若某条曲线在聚类前后均属于同一类别则称为分类准确），本文方法结果为95.82%，传统方法为90.54%。

$$c=\frac{\text{分类准确负荷曲线总条数}}{\text{负荷曲线总条数}}\times 100\% \quad (11)$$

可见，本例中本文方法与传统方法的聚类结果具有很高的相似性，且由上述聚类结果与实际仿真数据及分类准确率不难发现，本文方法在聚类质量方面相较于传统方法更具优势。

由附录A图A4可知，2种方法中同一类别的聚类中心曲线形态相似度很高。分析各类曲线的特征，第1类曲线为平峰型，其反映的特性符合重工业用户的用电行为：工作性质为三班制，负荷波动趋于平稳，负荷水平较高；第2类曲线为单峰型，其反映的特性符合轻工业用户以及部分政府、企业办事机构等用户的用电行为：工作时间固定，白天工作量大为用电高峰期，中午休息出现小低谷；第3类曲线为双峰型，其反映的特性符合市政三产用户的用电行为：白天为用电高峰，晚上出现一段晚高峰。上述分析表明，本文的聚类结果较传统方法能够更加准确地反映所属用户的用电性质，具有更好的工程应用价值，可为变电站特性分析提供有力依据。

图4为两种方法在最佳聚类数下目标函数随迭代次数的变化曲线，表2为2种方法在最佳聚类数下运行时间、迭代次数、类类平均距离等方面的性能对比（表中类类平均距离是指每类中所有曲线与聚类中心曲线的平均距离）。由图4及表2可看出，本文方法在最佳聚类数下的迭代次数、运行时间、XB指标以及各类的类类平均距离均远小于传统方法。

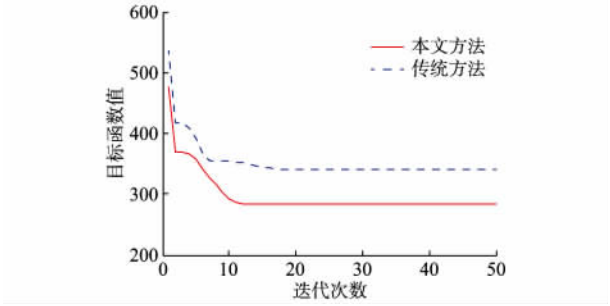


图 4　目标函数随迭代次数变化曲线
Fig.4　Curves of objective function versus number of iterations

综上，本文所提方法不仅可应用于工程实际，且其运行效率、聚类质量等在同等条件下均优于传统聚类方法。

3.2　特殊日负荷曲线聚类

3.2.1　数据来源

为验证本文方法能否有效避免因 1.1 节所述等距性问题造成分类时用户日负荷曲线的误分，在 3.1.2 节的类别 2 中提取同聚类中心曲线形态十分相近的日负荷曲线 100 条（记为第 1 类曲线），将该 100 条曲线在时间轴上往右平移 8 个时刻，获得另 100 条日负荷曲线（记为第 2 类曲线），共计 200 条日负荷曲线。

表 2　最佳聚类数下 2 种方法的聚类性能对比
Table 2　Comparison of clustering performance of two methods with optimal clustering number

聚类方法	运行时间/s			迭代次数	第 1 类类类平均距离	第 2 类类类平均距离	第 3 类类类平均距离	XB 指标
	数据处理	聚类程序	合计					
本文方法	0.43	2.74	3.17	11	0.907	1.658	1.168	0.209
传统方法	0.34	6.38	6.72	17	1.082	1.876	1.295	0.245

3.2.2　聚类结果及对比分析

以上述数据为基础，分别采用本文方法和传统聚类方法进行聚类（聚类数为 2），并对聚类结果进行分析。本例中熵权法获得的权重向量 $\boldsymbol{W}=[0.069,0.104,0.157,0.114,0.167,0.176,0.213]$。2 种方法下的 2 类聚类中心曲线形态与原始 2 类聚类中心曲线均十分相似，如图 5 所示，2 种方法中 200 条日负荷曲线的归属情况、类类平均距离、XB 指标大小如表 3 所示。由表 3 可知，本文方法获得的类别 1 中曲线条数占原始第 1 类曲线的 100%，类别 2 中占原始第 2 类曲线的 100%，不存在误分；传统方法类别 1 占原始第 1 类曲线的 95%，类别 2 占原始第 2 类曲线的 92%，存在一定的误分。

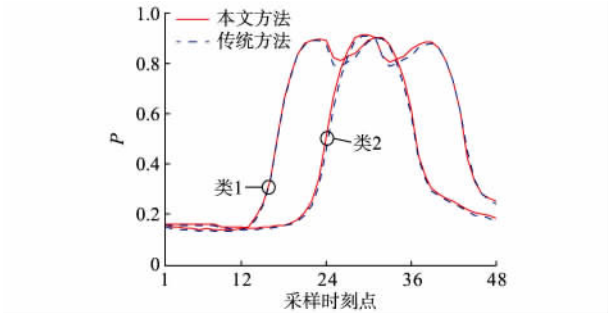


图 5　两种聚类方法下两类聚类中心曲线的对比
Fig.5　Comparison of two cluster center curves in two clustering methods

分析原因在于：传统方法聚类特征指标维数较大，聚类时存在原始曲线中某些曲线到设定聚类中心的欧氏距离相近且均较小，从而被误分到同一类中；而本文方法聚类特征指标维数较少且本例中 2 类原始曲线的聚类特征指标区别较大，同时本方法引入权重系数，很大程度上可避免曲线的误分。比较本例中两类方法的类类平均距离及 XB 指标，本文方法均小于传统聚类方法，由此可得本文方法的聚类效果优于传统聚类方法。

表 3　2 种方法聚类结果对比
Table 3　Clustering result comparison of two methods

方法	类 1 占原始曲线比例/%	类 2 占原始曲线比例/%	第 1 类类类平均距离	第 2 类类类平均距离	XB 指标
本文方法	100	100	2.03	2.12	0.234
传统方法	95	92	2.43	2.38	0.287

综上，在聚类日负荷曲线样本具有一定的形态差异，且差异不是特别明显的情况下，本文方法可以在一定程度上避免因维度较大带来的等距性问题而造成分类时用户日负荷曲线的误分。

3.3　算法鲁棒性检验

为检验本文方法的鲁棒性，结合上文模拟 6 类典型日负荷曲线，并分别在每类典型日负荷曲线的每个采样点加入 r 的扰动，获得每类 150 条，共计 900 条模拟日负荷曲线（$r=30\%$时的 6 类模拟曲线如附录 A 图 A5 所示）。

改变扰动 $r=5\%,10\%,15\%,20\%,25\%,30\%,35\%$，分别采用本文方法及传统聚类方法对相关模拟数据进行聚类分析，同时以最佳聚类数、分类准确率 c、XB 指标大小等指标检验算法的鲁棒性。不同扰动下 2 种算法 3 个指标的对比结果见表 4。

表 4　2 种聚类方法鲁棒性比较
Table 4　Robustness comparison of two clustering algorithms

r/%	本文聚类方法			传统聚类方法		
	最佳聚类数	XB 指标	c/%	最佳聚类数	XB 指标	c/%
5	6	0.01	100.0	6	0.01	100.0
10	6	0.06	100.0	6	0.07	99.8
15	6	0.12	100.0	6	0.15	99.5
20	6	0.19	99.6	6	0.19	98.8
25	6	0.29	99.6	5	0.28	86.2
30	5	0.28	88.5	5	0.31	84.3
35	5	0.32	85.8	5	0.41	82.5
40	5	0.38	83.2	4	0.45	78.9

由表 4 不难得出以下结论。

1)随着扰动的增加，XB 指标数值增大，分类准确度 c 降低，最佳分类数出现一定偏差。表明可用该 3 个指标检验算法的鲁棒性。

2)当扰动比例不大于 25%时，本文方法的最佳聚类数一直为 6，且分类准确率等于或十分接近于 100%，传统聚类方法在扰动为 25%时，最佳聚类数变为 5，聚类准确率大幅降低；当扰动比例为 25%～40%时，本文方法的聚类数均为 5，XB 指标及分类准确率波动较小，鲁棒性较好，而传统聚类方法的最佳聚类数在扰动为 40%时变为 4，XB 指标及分类准确率下降幅度进一步变大，鲁棒性较差。

上述情况出现的原因为：当负荷曲线各采样点的数值在一定范围内变动时，本文方法提取的 7 类特征指标数值与原始曲线提取的对应指标数值差异不大，也即该 7 类指标在一定扰动情况下仍然能较准确地反映原始负荷曲线的特征。此外，在聚类时引入各指标的权重系数，可进一步削弱扰动对聚类的影响。而传统聚类方法以欧氏距离作为相似判据，当扰动比例变大时，曲线与曲线之间的欧氏距离变化较大，从而使得相近类别(如类 3、类 6)中曲线误分的概率变大。

当 r=25%时，将基于本文聚类方法提取的6 类聚类中心曲线与原始 6 类聚类中心曲线进行对比，定义式(12)所示误差指标 E_{err} 来衡量同类别聚类中心曲线的相似程度。

$$E_{\mathrm{err}}=\sum_{i=1}^{n}\left(\frac{|x_i^*-x_i|}{x_i^*}\right)^2\times 100\% \qquad (12)$$

式中：x_i^* 和 x_i 分别为原始聚类中心曲线与本文提取聚类中心曲线上第 i 点的数据。以式(12)计算6 类曲线的误差指标 E_{err}，其结果均小于 8%，表明2 类负荷曲线的形态高度相似。

综上，本文方法耐噪能力强、鲁棒性好。在一定噪声强度下能真实还原原始典型日负荷曲线特征。

3.4　特征指标及权重对聚类质量的影响

3.4.1　特征指标选取对聚类质量的影响

为探究特征指标选取对算法鲁棒性及聚类准确性的影响，分别采用文献[2]所述 4 个指标、文献[13]所述 5 个指标、文献[17]所述 3 个指标、文献[21]所述 6 个指标作为聚类输入向量(本文所选特征指标均已涵盖这些指标)，以 3.3 节模拟数据为基础，采用本文方法配置不同扰动下的各指标的权重系数并计算不同扰动下的最佳聚类数、XB 指标、分类准确率 c，结果如附录 A 表 A1 所示。

由表 A1 可知，随着指标数目的减少，聚类准确性及鲁棒性大幅降低，当聚类指标数为 3 时，无论在小扰动还是大扰动情况下，聚类结果均无法还原真实曲线特性。原因在于，指标数目太少较难真实反映日负荷曲线的形态特征，易导致相近形态日负荷曲线的误分，当扰动量变大时，上述误分情况会更加严重。

需指出，聚类指标数目的增多必然导致聚类时间的增加，但换来了优良的聚类质量和稳定性，且运行时间也在可接受的范围内(本例中，7 类指标聚类的总程序运行时间均保持在 9.5～10.5 s)。因此，选择本文提出的 7 类指标进行日负荷曲线聚类有很大的优越性，同时可满足精细化聚类的要求。

3.4.2　特征指标权重对聚类质量的影响

以 3.3 节数据为基础，以本文 7 类指标为聚类输入向量，在 3 种方式(方式 1：不配置特征指标权重；方式 2：以本文方法配置指标权重；方式 3：修改本文方法配置的权重系数)下，分别计算不同扰动下的最佳聚类数、XB 指标、分类准确率 c，结果如附录 A 表 A2 所示。限于篇幅，只列出 r=10%时方式 2、方式 3 下的权重向量，$\boldsymbol{W}_2$=[0.056，0.264，0.056，0.027，0.042，0.25，0.305]，$\boldsymbol{W}_3$=[0.234，0.086，0.232，0.062，0.096，0.127，0.163]。

由表 A2 方式 1 和方式 2 结果可知，权重系数配置与否对聚类质量及算法稳定性有一定影响，原因在于，曲线簇中各指标反映实际曲线特征的重要程度不同，通过配置权重可进一步弱化扰动对负荷曲线形态特征的影响；而各指标权重大小会因聚类曲线簇的不同而存在较大差异，权重系数的合理配置有利于提高聚类质量。本文通过分析曲线簇中同类指标的离散情况来配置各指标权重大小无疑有一定合理性，方式 2 和 3 的对比结果也证实了本文方法的合理性。

4　结语

本文提出了一种基于特征指标降维及熵权法的

日负荷曲线聚类方法。提取物理意义明确的7类负荷曲线特征作为聚类输入向量，并引入熵权法自适应配置各指标权重，进而采用FW-FCM聚类算法对日负荷曲线进行聚类。算例结果表明：①综合划分能力、聚类质量、聚类效率和算法鲁棒性4个方面来看，本文方法相较于传统方法具有显著的优越性；②本文提取的7类特征指标能较好地反映用户的用电特性，也能较好地识别不同日负荷曲线的形态差异，以该7类指标进行聚类，聚类稳定性更理想，抗干扰能力更强；③本文采用熵权法原理自适应确定不同聚类曲线簇特征指标的权重系数，可进一步提高算法的聚类准确度。

本文以特征指标的加权欧氏距离衡量负荷曲线之间的相似程度，是否可用其他方法进行衡量是本文后续研究的内容之一。此外，本文方法的聚类结果服务于统计法综合负荷建模，曲线形态是本文关注的重点，而不同应用场景所关注的重点有一定差异，探究本文方法在其他场景下的适应性也将是本文后续的研究重点。

本文得到国网湖南省电力公司科技项目(5216A0140090)资助，特此感谢！

附录见本刊网络版(http://www.aeps-info.com/aeps/ch/index.aspx)。

参考文献

[1] 屈星，李欣然，宋军英，等.考虑配电网调压的综合负荷模型[J].电工技术学报，2018，33(4)：759-770.
QU Xing，LI Xinran，SONG Junying，et al. Composite load model considering voltage regulation of distribution network[J]. Transactions of China Electrotechnical Society，2018，33(4)：759-770.

[2] 李欣然，林舜江，刘杨华，等.基于实测响应空间的负荷动特性分类原理与方法[J].中国电机工程学报，2006，26(8)：39-44.
LI Xinran，LIN Shunjiang，LIU Yanghua，et al. A new classification method for aggregate load dynamic characteristics based on field measured response[J]. Proceedings of the CSEE，2006，26(8)：39-44.

[3] 徐振华，李欣然，钱军，等.变电站用电行业负荷构成比例的在线修正方法[J].电网技术，2010，34(7)：52-57.
XU Zhenhua，LI Xinran，QIAN Jun，et al. An online method to modify substation's structural proportion of synthetic load for power consuming-industries[J]. Power System Technology，2010，34(7)：52-57.

[4] 朱文俊，王毅，罗敏，等.面向海量用户用电特性感知的分布式聚类算法[J].电力系统自动化，2016，40(12)：21-27.
ZHU Wenjun，WANG Yi，LUO Min，et al. Distributed clustering algorithm for awareness of electricity consumption characteristics of massive consumers[J]. Automation of Electric Power Systems，2016，40(12)：21-27.

[5] 蒋雯倩，李欣然，钱军.改进FCM算法及其在电力负荷坏数据处理的应用[J].电力系统及其自动化学报，2011，23(5)：1-5.
JIANG Wenqian，LI Xinran，QIAN Jun. Application of improved FCM algorithm in outlier processing of power load[J]. Proceedings of the CSU-EPSA，2011，23(5)：1-5.

[6] 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.

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

[8] ZHONG Shiyin，TAM K S. Hierarchical classification of load profiles based on their characteristic attributes in frequency domain[J]. IEEE Transactions on Power Systems，2015，30(5)：2434-2441.

[9] 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.DOI：10.7500/AEPS20170309008.
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. DOI：10.7500/AEPS20170309008.

[10] 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.

[11] 李阳，刘友波，刘俊勇，等.基于形态距离的日负荷数据自适应稳健聚类算法[J].中国电机工程学报，2019，39(12)：3409-3420.
LI Yang，LIU Youbo，LIU Junyong，et al. Self-adaptive and robust clustering algorithm for daily load profiles based on morphological distance[J]. Proceedings of the CSEE，2019，39(12)：3409-3420.

[12] 王帅，杜欣慧，姚宏民，等.面向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.

[13] 高亚静，孙永健，杨文海，等.基于非参数核密度估计及改进谱多流行聚类的负荷曲线分类研究[J].电网技术，2018，42(5)：1605-1612.
GAO Yajing，SUN Yongjian，YANG Wenhai，et al. Study on load curve's classification based on nonparametric kernel density estimation and improved spectral multi-manifold clustering[J]. Power System Technology，2018，42(5)：1605-1612.

[14] 龚钢军，陈志敏，陆俊，等.智能用电用户行为分析的聚类优选策略[J].电力系统自动化，2018，42(2)：58-63.DOI：10.7500/AEPS20170726009.

GONG Gangjun，CHEN Zhimin，LU Jun，et al. Clustering optimization strategy for electricity consumption behavior analysis in smart grid[J]. Automation of Electric Power Systems，2018，42(2)：58-63. DOI：10.7500/AEPS20170726009.

[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] 国家发展和改革委员会，国家电网公司.电力需求侧管理工作指南[M].北京：中国电力出版社，2007：237-245.
National Development and Reform Commission，State Grid Corporation of China. Guide to power demand side management[M]. Beijing：China Electric Power Press，2007：237-245.

[17] 李欣然，姜学皎，钱军，等.基于用户日负荷曲线的用电行业分类与综合方法[J].电力系统自动化，2010，34(10)：56-61.
LI Xinran，JIANG Xuejiao，QIAN Jun，et al. A classifying and synthesizing method of power consumer industry based on the daily load profile[J]. Automation of Electric Power Systems，2010，34(10)：56-61.

[18] 欧阳森，石怡理.改进熵权法及其在电能质量评估中的应用[J].电力系统自动化，2013，37(21)：156-159.
OUYANG Sen，SHI Yili. A new improved entropy method and its application in power quality evaluation[J]. Automation of Electric Power Systems，2013，37(21)：156-159.

[19] 安晓华，欧阳森，杨家豪.分布式供能系统场景竞争力的分类评价模型及应用[J].电力系统自动化，2016，40(10)：69-75.
AN Xiaohua，OUYANG Sen，YANG Jiahao. Classified evaluation model for scenario competitiveness of distributed energy supply system and its application[J]. Automation of Electric Power Systems，2016，40(10)：69-75.

[20] 王瑞鹏，高永涛，吴顺川，等.基于改进熵-云模型的隧道采空区稳定性评价[J].现代矿业，2017(10)：208-211.
WANG Ruipeng，GAO Yongtao，WU Shunchuan，et al. Stability evaluation of tunnel goaf based on improved entropy-cloud model[J]. Modern Mining，2017(10)：208-211.

[21] 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.

[22] 耿嘉艺，钱雪忠，周世兵.新模糊聚类有效性指标[J].计算机应用研究，2019，36(4)：1001-1005.
GENG Jiayi，QIAN Xuezhong，ZHOU Shibing. New fuzzy clustering validity index[J]. Application Research of Computers，2019，36(4)：1001-1005.

[23] 孔祥玉，胡启安，董旭柱，等.引入改进模糊 C 均值聚类的负荷数据辨识及修复方法[J].电力系统自动化，2017，41(9)：90-95. DOI：10.7500/AEPS20160920002.
KONG Xiangyu，HU Qi'an，DONG Xuzhu，et al. Data identification and correction method with improved fuzzy C-means clustering algorithm[J]. Automation of Electric Power Systems，2017，41(9)：90-95. DOI：10.7500/AEPS20160920002.

宋军英(1969—)，女，博士，教授级高级工程师，主要研究方向：电力系统运行与控制。E-mail：276435879@qq.com

何　聪(1993—)，男，通信作者，硕士研究生，主要研究方向：电力系统仿真与建模。E-mail：749127148@qq.com

李欣然(1957—)，男，博士，教授，主要研究方向：电力系统分析控制、负荷建模。E-mail：lixinran1013@qq.com

（编辑　章黎）

Daily Load Curve Clustering Method Based on Feature Index Dimension Reduction and Entropy Weight Method

SONG Junying[1]，*HE Cong*[2]，*LI Xinran*[2]，*LIU Zhigang*[1]，*TANG Jie*[2]，*ZHONG Wei*[1]

(1. State Grid Hunan Electric Power Corporation，Changsha 410077，China；
2. College of Electrical and Information Engineering，Hunan University，Changsha 410082，China)

Abstract：Daily load curve clustering is the basis of analyzing load characteristics under the background of load modeling. In view of the shortcomings of existing clustering methods in terms of clustering quality and clustering efficiency and based on the principle of fuzzy C-means and entropy weight method，a method of daily load curve clustering based on dimensionality reduction of feature index and entropy weight method is proposed. Firstly，seven kinds of feature indices of dimensionality reduction such as daily load rate，daily peak-to-valley difference rate and daily maximum utilization time are extracted and taken as the clustering input to replace the load data of each sampling point. Secondly，the entropy weight method is introduced to configure the weight coefficient of each feature index adaptively. Finally，the power load curves are clustered with the feature-weighted fuzzy C-means (FW-FCM) clustering algorithm. The daily load curves of a region are clustered using the proposed method. Results show that the method has certain advantages in operating efficiency，robustness，clustering quality and so on. Moreover，the clustering results can reflect actual power consumption characteristics of the load truly and effectively.

This work is supported by National Natural Science Foundation of China (No. 51577056).

Key words：dimensionality reduction of feature index；entropy weight method；feature-weighted fuzzy C-means (FW-FCM) algorithm；load curve clustering

附录 A

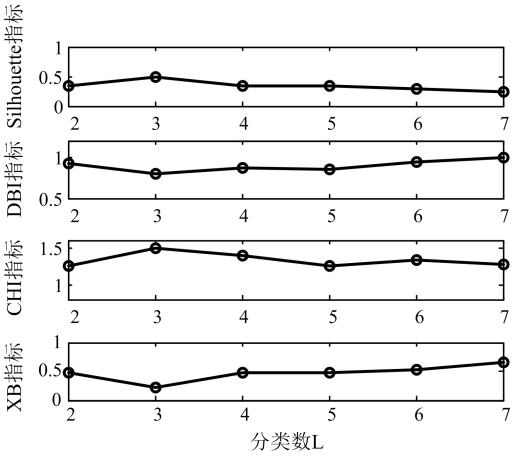
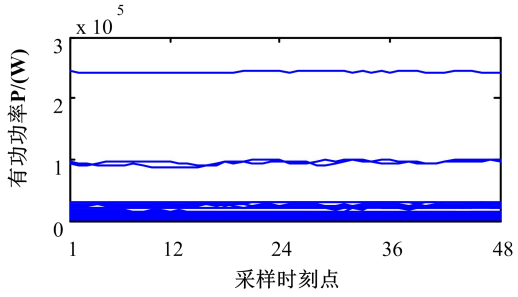
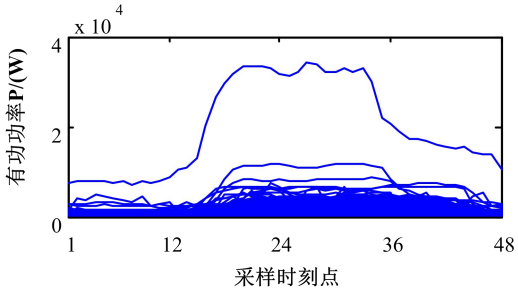


图 A1 有效性指标与聚类数的关系

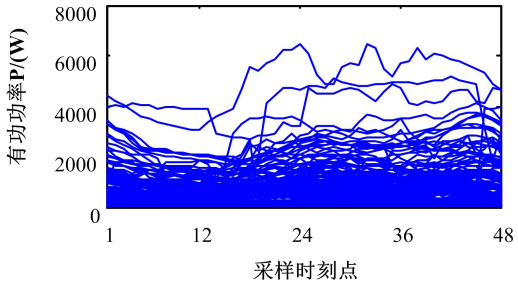
Fig.A1 Relationship between clustering validity indices and number of clusters



(a)类别 1（86 条）



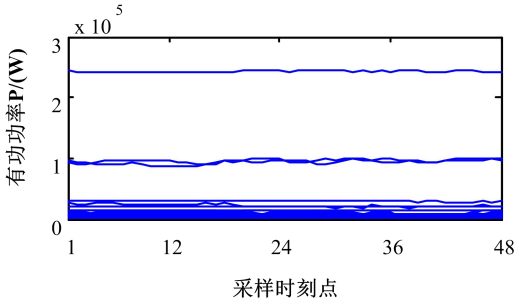
(b)类别 2（178 条）



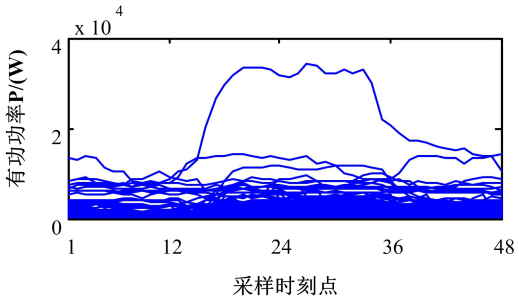
(c)类别 3（130 条）

图 A2 基于本文方法的日负荷曲线聚类结果

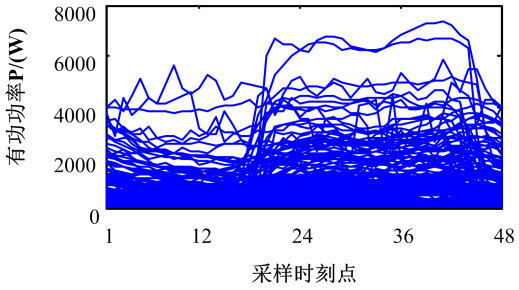
Fig.A2 Daily load curve clustering results based on method proposed



(a)类别 1（80 条）



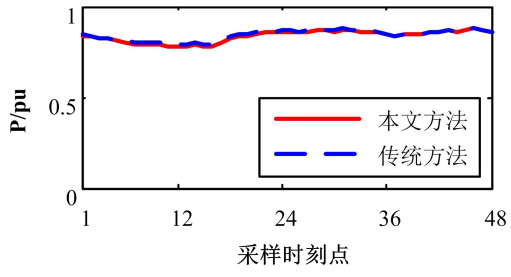
(b)类别 2（191 条）



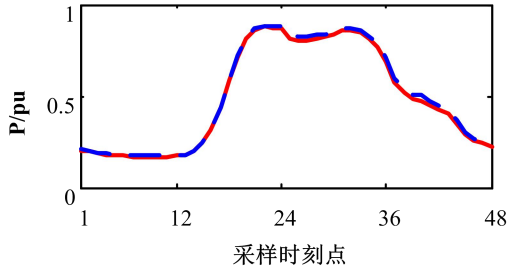
(c)类别 3（123 条）

图 A3 基于传统方法的日负荷曲线聚类结果

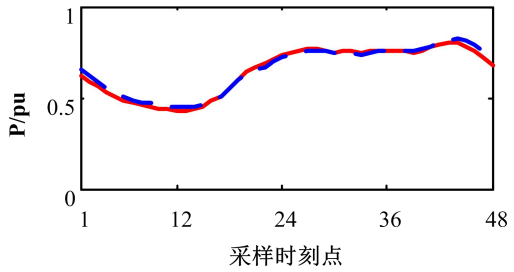
Fig.A3 Daily load curve clustering results based on traditional method



(a)类别 1 聚类中心曲线



(b)类别 2 聚类中心曲线



(c) 类别 3 聚类中心曲线

图 A4 两种方法聚类中心曲线对比结果

Fig.A4 Comparison results of clustering center curves of two methods

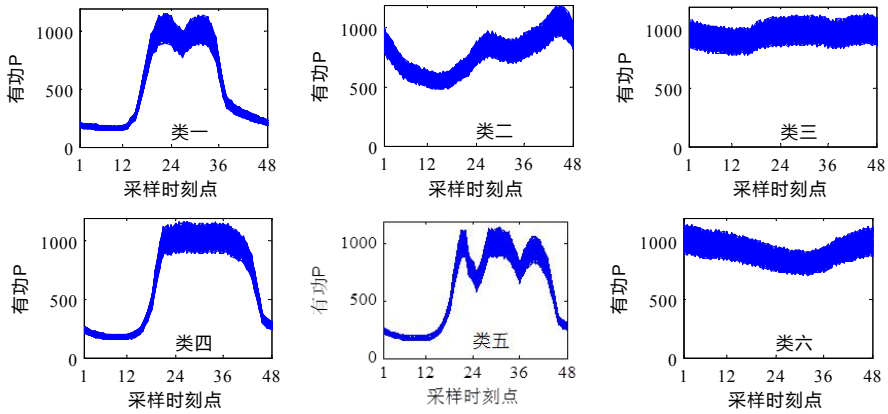


图 A5 *r*=30%时 6 类模拟曲线

Fig.A5 Six types of simulation curves at *r*=30%

表 A1 不同指标数下算法鲁棒性比较

Table A1 Comparison of algorithm robustness under different index numbers

r%	3 指标			4 指标			5 指标			6 指标			本文 7 指标		
	最佳聚类数	XB 指标	c%	最佳聚类数	XB 指标	c%	最佳聚类数	XB 指标	c%	最佳聚类数	XB 指标	c%	最佳聚类数	XB 指标	c%
5	5	0.12	85.8	6	0.06	100	6	0.08	100	6	0.02	100	6	0.01	100
10	5	0.20	83.4	6	0.12	99.2	6	0.15	99.5	6	0.12	99.5	6	0.06	100
15	5	0.32	81.5	5	0.23	89.8	6	0.21	99.4	6	0.18	99.5	6	0.12	100
20	5	0.42	80.5	5	0.29	88.4	5	0.29	90.4	6	0.20	91.7	6	0.19	99.6
25	4	0.54	75.8	5	0.35	85.4	5	0.31	88.8	5	0.27	87.8	6	0.29	99.6
30	4	0.59	72.6	5	0.42	84.5	5	0.35	85.8	5	0.36	84.7	5	0.28	88.5
35	4	0.62	71.5	5	0.47	82.5	5	0.40	84.1	5	0.42	85.8	5	0.32	85.8
40	4	0.69	68.9	4	0.52	75.1	4	0.48	78.4	4	0.48	80.4	5	0.38	83.2

表 A2 不同权重下算法鲁棒性比较

Table A2 Comparison of algorithm robustness under different weights

r%	方式 1			方式 2			方式 3		
	最佳聚类数	XB 指标	c%	最佳聚类数	XB 指标	c%	最佳聚类数	XB 指标	c%
5	6	0.05	100	6	0.01	100	6	0.10	96.8
10	6	0.14	99.8	6	0.06	100	6	0.18	94.6
15	6	0.20	96.5	6	0.12	100	5	0.25	88.5
20	6	0.28	93.7	6	0.19	99.6	5	0.31	86.7
25	5	0.31	87.1	6	0.29	99.6	5	0.38	83.5
30	5	0.39	83.4	5	0.28	88.5	5	0.42	80.6
35	5	0.42	80.4	5	0.32	85.8	4	0.55	72.5
40	4	0.59	76.7	5	0.38	83.2	4	0.62	64.4