# 一种基于 Pearson 相关系数的电力用户负荷曲线聚类算法

### 王星华,许炫壕,周亚武

(广东工业大学 自动化学院 广州 510006)

摘 要: 提出一种基于 Pearson 相关系数作为聚类判据的负荷曲线聚类算法——Pearson 相关聚类( Pearson Correlation Clustering , PCC)。首先对负荷数据进行数据清理以及去噪处理 再选择合适的降维算法以降低数据处理的复杂性。提出利用 Pearson 相关系数阈值作为聚类中心选择依据的方法解决初始聚类中心选择的随机性; 利用电力负荷曲线数据与聚类中心之间的 Pearson 相关系数进行聚类 以 DBI 指标作为聚类效果的评价标准 ,分析了不同系数对聚类效果的影响。算例结果表明 ,该算法相比传统算法运行时间短 , 鲁棒性强 聚类效果更好。

关键词: Pearson 相关系数; 负荷曲线分类; 降维; 中值滤波; 聚类有效性

中图分类号: TM714 文献标志码: A 文章编号: 2095 - 6843(2017) 05 - 0397 - 05

DOI:10.13625/j.cnki.hljep.2017.05.004

## A clustering algorithm of power userload curves based on Pearsoncorrelation coefficient

WANG Xinghua XU Xuanhao ZHOU Yawu

(School of Automation Guangdong University of Technology Guangzhou 510006 China)

Abstract: A Pearson Correlation Clustering ( PCC) algorithm based on Pearson correlation coefficient as a clustering criterion is proposed. Firstly, the load data is cleaned and noiseeliminationis made, then the appropriate dimensionality reduction algorithm is selected to reduce the complexity of data processing. The Pearson correlation coefficient threshold is proposed to be used as the method of clustering center selection to solve the stochasticity of the initial clustering center selection. Clusteringis madebyusing Pearson correlation coefficient between the power load curve data and the clustering center and the influence of different coefficients on the clustering effect is analyzed by taking the DBI indexas the clustering effect. The results show that compared with the traditional algorithm, the algorithm has a short run time, strong robustness and better clustering effect.

**Key words**: Pearson correlation coefficient; load curve classification; dimensionality reduction; median filter; clustering validity

随着智能电网的普及,电网运行、维护以及管理过程产生了大量电力数据。大数据产生于智能电网系统的各个环节中,其中用户侧大数据占了很大比重,因此有必要对用户侧大数据进行研究,为智能电网规划和运行提供有益参考[1]。用户侧大数据中,包含了大量电力用户的日/月负荷曲线,通过对这些电力消费信息进行专门的分析研究并获取相应的负荷模式,可以为电网公司的决策提供依据<sup>[2]</sup>。

当前国内对电力负荷聚类分析研究已经起步,

收稿日期: 2017 - 06 - 27。

作为辅助电力负荷聚类分析的工具,各种聚类算法及其改进算法不断地被提出并应用于具体工程。 文献 [3]以函数型数据分析和 k - means 算法结合对电力负荷进行分类分析。文献 [4]将余弦相似度和欧氏距离结合组成双层聚类算法应用于负荷形态组合识别。文献 [5]针对 FCM 聚类算法易受初始聚类中心选择影响及极易陷入局部最小值的缺陷 提出了对初始聚类中心进行改进的 FCM 聚类算法。

以上研究均以欧氏距离作为聚类的相似性判据。但以欧式距离作为相似性判据易受噪声和尖峰值影响,且欧式距离反映的是曲线间距离,而负荷曲线聚类更侧重于曲线轮廓的聚类。因此,为了更加充分反映曲线相似性,综合考虑以上文献中聚

类算法的优缺点,本文提出一种基于 Pearson 相关系数的负荷曲线聚类算法——Pearson 相关聚类(Pearson Correlation Clustering, PCC)。聚类过程中,首先对负荷数据进行数据清理以及去噪处理,选择合适的降维算法进行降维,再利用 Pearson 相关系数作为相似性判据选取初始聚类中心,并对负荷曲线进行聚类分析。与传统聚类算法进行对比的结果表明,该算法运行效率高、鲁棒性强,能够得到较好的聚类结果。

## 1 数据预处理

#### 1.1 数据来源

MS

576

本文研究算例数据采用两个数据集: 一是来源于某市 5 类用户类别的 2 340 个电力客户同一日期的 24 点日负荷曲线; 二是某市 4 类用户类别的 1 023个电力客户某年的 12 点月负荷曲线。两数据源的用户数分别如表 1、2 所示。

表 1 数据 1 用户类别 Table 1 User classification of data 1

MT	AM	MI	DM
2.42	605	4.42	20.4

表 2 数据 2 用户类别 Table 2 User classification of data 2

A	В	С	D
435	8	413	167

#### 1.2 数据清理及去噪

为了避免数据采集过程中出现的偏差对聚类结果的影响、采用以下措施进行数据清理及去噪:

- 1) 错误数据清理。采集设备停止工作或者传输过程中断 导致部分电力用户缺失数据甚至无数据 因此当某一负荷曲线数据全为 0 时 ,则剔除该数据。
- 2) 异常数据修复。考虑到负荷功率的暴增或者骤降情况,当负荷曲线功率相邻点的差值相差较大时 利用文献 [6] 中基于 Lagrange 插值的 Neville 算法对曲线  $X = \{x_1, x_2, \dots, x_n\}$  进行插值修复。
- 3)数据去噪。针对数据中"噪声"所导致的曲线毛刺问题,采取中值滤波方法对数据源负荷曲线进行去噪处理。中值滤波的原理为:给定一个输入数据 重新计算每一个x的输出值y,得到新的输出值相当于y = new(x)。其中 new 函数的操作是,从在以x为中心,长度为2l的原信号中(区间为[x-l+1,x+l]),提取出这段区间内中间的那个值,作为y = new(x)的结果。经实验结果表明 l 取

2 时效果最佳。

#### 1.3 降维算法

随着智能电网时代的到来,电力数据的数量呈海量增长,采集设备的精度提高使数据维数也越来越高。为了提高算法的效率和可行性,有必要对数据进行降维操作。假设每条负荷曲线的维数为W, 经降维后将适当缩小维数为W, 即降维的目的是用尽可能低维数的负荷曲线代替原来的负荷曲线,并保证原负荷曲线不会因降维而失真过多。

文献 [7]对 Sammon 映射、SOM 映射、PCA 以及 FE 降维算法进行对比分析 ,分析得出 PCA 降维算 法所需时间最短 ,且可以最大程度地保护数据原始 信息 ,因此本文数据降维采用 PCA 降维算法。

分别对两数据源使用 PCC 算法进行降维聚类 以及未降维聚类 所用时间对比如图 1 所示。

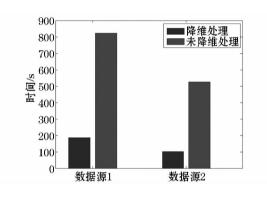


图 1 降维前后聚类时间对比图

Fig. 1 Time comparison between before and after data dimensionality reduction

对比降维前后聚类用时可知,降维处理可很大程度减少聚类算法耗时。

## 2 PCC 聚类算法原理

#### 2.1 Pearson 相关系数

在统计学中,考察两个事物(数据变量)之间相关程度的系数称为相关系数。如果有两个变量  $X = \{x_1, x_2, \dots, x_n\}$  和  $Y = \{y_1, y_2, \dots, y_n\}$  根据最终计算出的相关系数的大小可以有以下三层含义:

- 1) 相关系数为 0 X、Y 两变量间无关系;
- 2) X 的值增大(减小)  $_{,Y}$  值增大(减小) ,两个 变量为正相关 相关系数在 0.00 与 1.00 之间;
- 3) X 的值增大(减小), Y 值减小(增大),两个变量为负相关相关系数在 -1.00 与 0.00 之间。

由上可知,相关系数的绝对值越大,相关性越强相关系数越接近于1或-1,相关性越强,相关系数越接近于0相关性越弱。

Pearson 相关也称为积差相关,它以两变量间的 离均差积和与离均差平方和为算术基础进行计算。

Pearson 相关系数 p 的计算公式为

$$p = \frac{L_{XY}}{\sqrt{L_{XX}L_{YY}}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

X 的离均差平方和  $L_{xx}$ 为

$$L_{XX} = \sum_{i=1}^{n} x_i - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Y的离均差平方和 $L_{yy}$ 为

$$L_{YY} = \sum_{i=1}^{n} y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

 $X \setminus Y$  间的离均差积和  $L_{xy}$  为

$$L_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x) (\sum_{i=1}^{n} y)}{n}$$

Pearson 相关系数 p 的判断标准如表 3 所示。

表 3 Pearson 相关系数 p 的判断标准

Table 3 Judgement criterion of Pearson correlation

coefficient

<i>p</i>	相关程度
0.0 ~ 0.19	极低相关
0.2 ~ 0.39	低度相关
0.4 ~ 0.69	中度相关
0.7 ~ 0.89	高度相关
0.9 ~ 1.00	极高相关

#### 2.2 基于 Pearson 相关系数的初始聚类中心选取

为了提高算法的运行效率和聚类精确性,需要对数据源中大量数据选取合适的初始聚类中心。本文初始聚类中心的选取原理是:事先设定一个初始相关系数阈值,使得到的初始聚类中心之间的相关系数尽量小于该阈值,即各初始聚类中心之间的相关性尽可能小,可以使原本随机生成的初始聚类中心变成有目的性的选取,提高算法精确性。

设  $X = \{x_1, x_2, \dots, x_n\}$  为待聚类数据集 ,设类间最大相关系数阈值为  $\partial$  ,聚类数目为 K ,初始聚类中心选取步骤如下:

Step 1 计算任意两组数据之间的 Pearson 相关系数 并生成系数矩阵 Q 将系数最大的两个数据归为一类 并取两个数据的中点作为第一个聚类中心。

Step 2 根据事先设定的用于聚类中心选择的系数阈值  $\partial$  利用系数矩阵 Q 检测与第一个聚类中

心的两组数据的系数都小于 ∂ 的所有数据 ,并在其中选择系数最高的两个数据定为一类 ,且取两个数据的中点为第二类聚类中心。

Step 3 同理,在剩下的数据中整理与已经确定的数据的系数都小于∂的数据,并在这些数据中选择系数最高的两个数据定为一类,且取两个数据的中点为新一类聚类中心。

Step 4 重复 Step 3 直至找到 K 类为止。

此流程利用系数矩阵 *Q* 及系数阈值对初始聚类中心进行判断,以数据点和已确定数据点的相关系数进行比较,避免了传统方法中将聚类中心与全部数据点进行比较,大幅度减少了所需运算量,且精度比传统随机性选择方法要高。

#### 2.3 PCC 算法流程

根据已得到的系数矩阵 Q 和已选定的初始聚类中心,PCC 算法的原理为: 假设初始聚类中心为 K 类 根据系数矩阵 Q ,分别比较某一数据点到 K 类 聚类中心的相关系数,与哪一类的相关系数最大即最相似则归为哪一类。 归类后重新计算聚类中心,重复归类,直至满足聚类结束迭代条件。 PCC 算法具体流程如下:

Step 1 由初始聚类中心选取规则得出初始聚 类中心  $m_i = \{m_1, m_2, \dots, m_K\}$  K 为聚类个数。

Step 2 设  $y_j$  为第 j 个样本 找出  $y_j$  与  $m_i$  所有的最大相关系数  $Q_{\max}$  将其归入对应的聚类中心所属的类中。

Step 3 设  $n_i$  是第 i 类已聚类样本的个数  $g_{ij}$ 为第 i 类中第 j 个样本,由下式重新计算聚类中心  $m_i$ :

$$\overline{m}_i = \frac{1}{n} \sum_{i=1}^{n_i} y_{ij}$$

Step 4 设 t 为迭代次数 按下式计算 Y 中所有样本的系数误差偏移量 J(t) 并与前一次系数误差偏移量 J(t-1) 比较:

$$J(t) = \frac{N_{Q \ge r}}{N}$$

式中: N 为样本数据总个数; r 为系数误差偏移系数;  $N_{0 \ge r}$  为样本数据各类别中  $Q \ge r$  的样本总数。

Step 5 若  $J(t) - J(t-1) < \varepsilon$  ,即 J(t) 趋于稳定 则算法结束 否则转 Step 2 重复迭代。

由以上 PCC 算法流程 ,结合聚类有效性指标检验 设最小聚类数为  $L_{\min}$ ,最大聚类数为  $L_{\max}$ ,有效性指标为  $I_{\mathrm{DBI}}$  聚类算法总流程如图 2 所示。

#### 2.4 聚类有效性指标

为确定最佳聚类数和最佳聚类结果 引入聚类 有效性指标进行聚类效果评价。文献 [7]对 SSE、

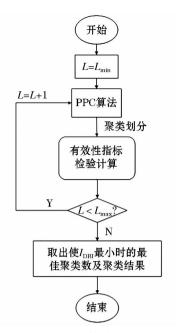


图 2 PCC 算法总流程

Fig. 2 Flowchart of PCC

CHI、DBI 三种指标进行评价,验证得 DBI 指标曲线的拐点更为直观,计算公式简单且变化范围小,更适合作为电力用户负荷曲线的聚类有效性指标。因此本文采用的聚类有效性指标为 DBI 指标,DBI 计算公式如下:

$$I_{\text{DBI}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{R}_{k}$$

$$R_{k} = \max_{k \neq j} \frac{d(X_{k}) + d(X_{j})}{d(c^{k} \varphi_{j})}$$

式中: K 为聚类数目;  $d(X_k)$  和  $d(X_j)$  为矩阵内部距离;  $d(c_k, c_j)$  为向量间的距离。 $I_{\mathrm{DBI}}$  越小则表明聚类结果越好。

## 3 算例分析

#### 3.1 电网实际负荷曲线聚类结果

如上所述 本文采用两个数据集 ,经数据清理、去噪后 ,数据源数据分类如表 4、5 所示。

表 4 数据源 1 清理后数据分类

Table 4 Data classification of data source 1 after cleaning

MS	MT	AM	MI	DM	总计
558	334	675	434	272	2 273

表 5 数据源 2 清理后数据分类

Table 5 Data classification of data source 2 after cleaning

A	В	С	D	总计	
428	6	408	142	984	

设定该算例初始聚类中心选取阈值  $\partial = 0.7$  ,系数误差偏移系数 r = 0.7  $\varepsilon = 0.0005$  。

以聚类有效性指标  $I_{DBI}$  判定聚类最佳结果 ,如图 3 所示。结果表明: 对于数据源 1 ,当聚类数为 6 时  $I_{DBI}$  最小 最佳聚类数为 6; 对数据源 2 ,当聚类数为 4 时  $I_{DBI}$  最小 最佳聚类数为 4。

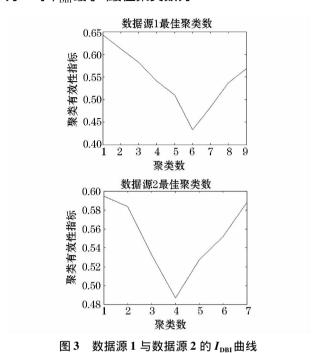


Fig. 3 The best clustering number of dataset 1 and dataset 2 with  $I_{DRI}$ 

经 PCC 算法聚类后 两数据源的聚类结果如图 4、5 所示。

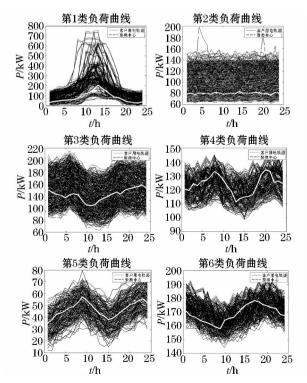


图 4 数据源 1 聚类结果

Fig. 4 Clustering result of data source 1

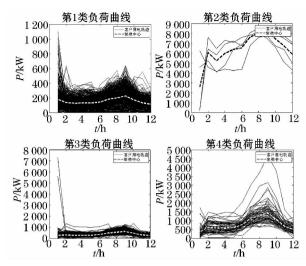


图 5 数据源 2 聚类结果

Fig. 5 Clustering result of data source 2

保持  $\partial$  = 0.7 不变 分别设定 r = 0.6 和 r = 0.8 , 对数据源 1 进行聚类 ,由聚类有效性指标  $I_{DBI}$  判定 聚类最佳结果 ,如图 6 所示。

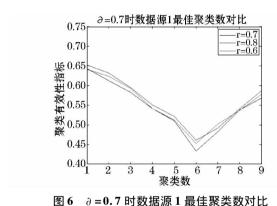


Fig. 6 The Comporison of the best clustering number of dataset 1 for  $\partial = 0.7$ 

保持 r = 0.7 不变 ,分别设定  $\partial = 0.6$  和  $\partial = 0.8$  , 对数据源 1 进行聚类 ,由聚类有效性指标  $I_{DBI}$  判定 聚类最佳结果 ,如图 7 所示。

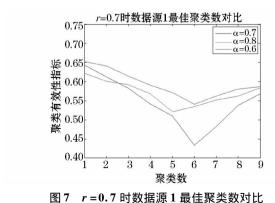


Fig. 7 The comparison of the brst clustering mumber of dataset 1 for r = 0.7

#### 3.2 结果分析及对比

由图 4 可知,经 PCC 算法聚类之后,数据源 1 分为 6 类,负荷类型主要为单峰型、双峰型、平稳型以及避峰型 具体分类数如表 6 所示。表 6 结果表明 数据源 1 分类结果与实际结果偏差不大,能较好地将数据进行真实分类。

表 6 数据源 1 分类数目

Table 6 Number of data source 1 classifications after clustering

_	1	2	3	4	5	6
	550	332	682	232	216	264

图 6 中 ,当  $\theta$  = 0. 7 保持不变时 ,可知三条曲线的极值点都出现在聚类数为 6 时 ,且 r = 0. 7 时  $I_{DBI}$  值最小 ,即 r 值的变化不影响最佳聚类数的变化 ,只影响  $I_{DBI}$ 。图 7 中 ,当保持 r = 0. 7 不变时 ,三条曲线的极值点位置存在差异  $I_{DBI}$ 最小的曲线为  $\theta$  = 0. 7 ,此时的最佳聚类数为 6 ,说明  $\theta$  取不同值时最佳聚类数有区别 ,且  $I_{DBI}$  也不尽相同。综上所述 ,适合于PCC 算法的最佳取值应为  $\theta$  = 0. 7 和 r = 0. 7。

将传统 FCM 聚类算法和 k – means 算法应用于本文两数据源中,并与 PCC 算法的聚类有效性指标进行比较 如图 8 所示。由图 8 结果可得出 PCC 算法在两数据源分类中的聚类有效性指标都比传统算法理想 因此可知 PCC 算法聚类效果较好,分类结果合理。

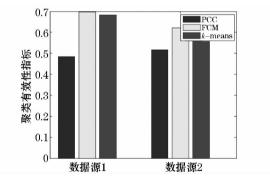


图 8 PCC 算法与传统 FCM 算法和 *k* – means 算法聚类结果对比

Fig. 8 Clustering result comparison between PCC algorithm and FCM or k – means algorithm

## 4 结 语

1)提出一种基于 Pearson 相关系数作为聚类判据的负荷曲线聚类算法——PCC,通过对负荷数据进行数据清理以及去噪处理,选择合适的降维算法进行降维之后,以改进方法选取初始聚类中心,利用 Pearson 相关系数作为聚类相似性判据,对电力(下转第415页)

品电缆进行测试 结果符合耐寒电缆的使用要求。

### 参考文献:

- [1] 王春江. 电线电缆手册. 第1册[M]. 北京: 机械工业出版 社 2002.
  - WANG Chunjiang. Wire and cable manual. Volume 1 [M]. Beijing: China Machine Press, 2002.
- [2] GB/T8815-2008. 电线电缆用软聚氯乙烯塑料[S]. 2009.
- [3] 王贵恒. 高分子材料成型加工原理[M]. 北京: 化学工业出版社 ,1995.
  - WANG Guiheng. Polymeric material molding and processing principle [M]. Beijing: Chemical Industry Press, 1995.
- [4] 韩吉彬,田洪池 徐欣 等. 增塑剂 DOP 对 NBR/PP 共混物性能的影响[J]. 橡胶工业 2005 52(8):472-475.

- HAN Jibin , TIAN Hongchi , XU Xin , et al. Effect of DOP on the properties of NBR/PP TPV [J]. China Rubber Industry , 2005 , 52(8): 472 475.
- [5] 罗绍武, 习有建 毛继红. 聚氯乙烯电缆料电气性能的分析[J]. 聚氯乙烯 2006(8):9-12.
  - LUO Shaowu , XI Youjian , MAO Jihong. Analysis on the electrical behaviors of PVC cable materials [J]. Polyvinyl Chloride , 2006(8):9-12.
- [6] 王娟娟 冯晓燕 ,王颖 ,等. 聚氯乙烯共混改性研究进展[J]. 绝缘材料 2003(3):40-45.
  - WANG Juanjuan , MA Xiaoyan , WANG Ying , et al. The advance of PVC blending modification [ J ]. Insulation Materials , 2003(3):40-45.

(编辑 陈银娥)

#### (上接第386页)

WANG Jiuhe ,LI Huade ,WANG Liming. Direct power control system of three phase boost type PWMrectifiers [J]. Proceeding of the CSEE. 2006  $\,$  26(18):54  $\,$  -60( in Chinese) .

- [7] Malinowski M ,Kazmierkowski M ,Blaabjerg P K ,et al. Virtual flux based direct power control of three phase PWM rectifiers [J]. IEEE Transaction on industrial Applications. 2001 ,37 (4): 1019 1027.
- [8] Bouafia A ,Gaubert J P ,Krim F. Predictive direct power control of three – phase pulsewidth modulation (PWM) rectifier using space – vector modulation (SVM) [J]. IEEE Transactions on Power E– lectronics 2010 25(1):228 –236.
- [9] Yin B ,Oruganti R ,Panda S K ,et al. A simple single input sin-

gle – output (SISO) model for a three – phase PWM rectifiers [J]. IEEE Transaction on Power Electronics. 2009 24(3):620 –631.

- [10] Briz F ,Degner M W ,Lorenz R D. Analysis and design of current regulation using complex vectors [J]. IEEE Transaction on industrial Applications. 2000 36(3):815-825.
- [11] 陈耀军,钟炎平. 基于合成矢量的电压型 PWM 整流器电流控制研究[J]. 中国电机工程学报 2006 26(2):143-148.

  CHEN Yaojun ZHONG Yanping. Study on the current control for voltage source PWM rectifier using complex vector [J]. Proceedings of the CSEE 2006 26(2):52-56(in Chinese).

(编辑 侯世春)

#### (上接第401页)

用户负荷曲线进行聚类分析。

- 2) 利用 Pearson 相关系数作为聚类判据 ,比传统基于欧氏距离的判据方法更能合理地反映负荷曲线相似性 ,同时在聚类速度上也有较大改进。
- 3) 本文研究角度为横向角度聚类 即每一用户都选定同一天负荷曲线聚类。如何从纵向角度 ,即从同一用户多天的负荷曲线中 ,结合温度等因素找寻典型曲线 ,再对不同用户的典型曲线进行聚类是接下来需要研究的方向。

## 参考文献:

- [1] 宋亚奇 周国亮 ,朱永利. 智能电网大数据处理技术现状与挑战[J]. 电网技术 2013 37(4):927-935.

  SONG Yaqi ,ZHOU Guoliang , ZHU Yongli, Present status and
  - SONG Yaqi ,ZHOU Guoliang , ZHU Yongli. Present status and challenges of big data processing in smart grid [J]. Power System Technology ,2013 ,37(4): 927 -935.
- [2] 赵腾,张焰,张东霞. 智能配电网大数据应用技术与前景分析[J]. 电网技术 2014 38(12):3305-3312.

  ZHAO Teng, ZHANG Yan ZHANG Dongxia. Application technology of big data in smart distribution grid and its prospect analysis [J]. Power System Technology, 2014 38(12):3305-3312.
- [3] 张欣 高卫国 苏运. 基于函数型数据分析和 k means 算法的电力用户分类[J]. 电网技术 2015 39(11):3153 3162.
  ZHANG Xin, GAO Weiguo SU Yun. Electricity consumer arche-

- types study based on functional data analysis and k means algorithm [J]. Power System Technology ,2015 , 39 ( 11 ) : 3153 –3162.
- [4] 王星华 陈卓优 彭显刚. 一种基于双层聚类分析的负荷形态组合识别方法[J]. 电网技术,2016,38(1):1-7.
  - WANG Xinghua CHEN Zhuoyou , PENG Xiangang. A new combinational electrical load analysis method based on bilayer clustering analysis [J]. Power System Technology 2016 , 38(1):1-7.
- [5] 张慧哲,王坚. 基于初始聚类中心选取的改进 FCM 聚类算法[J]. 计算机科学 2009 36(6):206-209.

  ZHANG Huizhe, WANG Jian. Improved fuzzy C means clustering algorithm based on selecting initial clustering centers [J]. Computer Science, 2009 36(6): 206-209.
- [6] 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):
  - 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.
- [7] 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.

(编辑 陈银娥)