

考虑负荷纵向随机性的用户用电行为聚类方法

冯志颖¹,唐文虎¹,吴青华¹,陆国俊²,栾乐²

(1. 华南理工大学 电力学院, 广东 广州 510640; 2. 广州供电局有限公司, 广东 广州 510000)

摘要: 针对不考虑负荷纵向随机性所导致的数据损失和用户误分类的问题, 提出了一种考虑负荷纵向随机性的基于推土机距离(EMD)的用户用电行为识别新方法。该方法通过统计电力用户同一时刻多天的负荷分布情况, 从横向和纵向2个角度全面表征用户的用电行为。并结合EMD和欧氏距离度量不同用户用电行为的差异程度。以一组国际通用的居民用电负荷作为算例进行分析, 算例结果表明, 在横向特性较为相似的用户中, 该方法能够很好地提取用户的纵向特性。定性和定量分析均表明, 该方法对用户负荷的聚类效果精细合理。

关键词: 智能电网; 用电行为; 负荷曲线; 纵向随机性; 聚类方法; 推土机距离

中图分类号: TM 714

文献标识码: A

DOI: 10.16081/j.issn.1006-6047.2018.09.007

0 引言

智能电网是能源与电力行业发展的必然趋势^[1]。随着智能电网理论与实践的推进, 电网建设正朝着数字化、互动化和分布化的方向发展^[2]。分布式可再生能源和电动汽车等多元化负荷的广泛接入, 加大了负荷预测和需求侧管理等工作的难度, 对电力用户进行精细化分类成为了负荷预测、需求侧管理、用电定价等方面的重要前提。对用户进行合理精细的分类, 掌握用户的用电行为, 对于各电力企业制定运维和营销策略具有重要意义。用户用电负荷具有较大的不确定性, 日负荷曲线展现了用户1 d中的用电行为, 表现为负荷的横向特性; 而一段时期(如一周或一个月)中不同的日负荷曲线也会存在差异, 表现为负荷的纵向特性。不同用户的纵向差异程度具有明显的区别。

国内外已有不少学者对负荷聚类进行了研究。主要的研究方法大致可分为改进相似度度量方法^[3-6]、改进数据预处理方法^[7-9]和改进聚类算法^[10-13]。在改进相似度度量方法方面, 文献[3]针对传统欧氏距离度量负荷曲线形态的不足, 提出以余弦相似度代替欧氏距离, 能较好地度量负荷形态之间的差异; 文献[4]利用双向夹逼的思想, 结合层次聚类对同一用户的不同负荷日进行聚类, 该方法主要用于寻找相似日进而进行负荷预测, 但不适用于聚类不同的电力用户; 文献[5]提出以余弦相似度作为外层聚类根据、以欧氏距离作为内层聚类根据的双层聚类方法, 其能够识别不同的负荷形态和大、小负荷。在改进数据预处理方面, 文献[6]利用

6个日负荷特性指标取代高维的时间序列进行聚类, 提高了算法的运算速度和鲁棒性, 但在一定程度上会造成负荷数据失真; 文献[7]利用快速小波变换提取负荷时间序列的特性指标; 文献[8]通过多尺度分析解决电力负荷数据的数据量大、波动性和不确定性的问题; 文献[9]比较了多种降维方法的效果, 指出主成分分析和Sammon映射的聚类效果最好。在改进聚类算法方面, 用于负荷聚类的聚类算法主要有Kernel K-means^[10]、自组织映射神经网络^[11]、密度空间聚类^[12]、C-vine Copula混合模型^[13]等算法。文献[10]引入核函数改进传统的K-means算法, 将数据映射到高维空间, 从而提高数据的可分性; 文献[11]利用自组织映射神经网络, 对原始数据进行低维映射, 实现了聚类效果的可视化; 文献[12]结合引力搜索算法和密度空间聚类对居民用电数据进行聚类, 并分析各类用户的需求侧响应潜力; 文献[13]提出一种基于模型的C-vine Copula混合模型聚类算法, 具有较好的聚类效果, 但其计算过程复杂, 计算时间较长。

已有的研究大多针对负荷的横向特性, 对于多日的日负荷曲线的预处理, 只是简单地取平均值, 或者剔除异常值后选取其中的某天作为典型负荷日等, 缺乏对负荷纵向特性的考虑, 而且预处理过程会造成部分有用信息的丢失。在以下的场景中, 对用户进行聚类时应该考虑负荷纵向波动性: 用户每天的负荷曲线变化很大, 难以用一条典型负荷曲线表征其用电行为; 异常用电行为分析; 区间或概率性负荷预测等。针对这种不确定性, 文献[8]提出一种利用高斯混合模型来反映用户的纵向随机性的方法, 取得了一定的效果, 但算法计算过程复杂。

本文提出一种同时考虑负荷横向与纵向特性的负荷聚类算法, 并首次引入推土机距离(EMD)算法, 以衡量在同一时刻不同用户之间的负荷纵向距离; 然后对所有时刻进行综合, 得到用户之间的综合

收稿日期: 2017-11-29; 修回日期: 2018-07-13

基金项目: 国家高技术研究发展计划(863计划)项目(2015AA050201); 国家自然科学基金资助项目(51477054)
Project supported by the National High Technology Research and Development Program of China(863 Program)(2015AA050201) and the National Natural Science Foundation of China(51477054)

距离;最后利用美国能源部提供的公开负荷数据,与传统的只考虑负荷横向特性的层次聚类算法以及其他相似度量方法进行比较,证明本文方法的合理性与准确性。

1 负荷行为识别方法

1.1 负荷的纵向随机性

电力用户的用电行为除了可用日负荷曲线描述之外,还应包括负荷的纵向随机性。负荷的纵向随机性表现为用户在一段时间的负荷波动性。某些用户的用电负荷每天的差异较小,表现为较大的稳定性;某些用户的用电负荷受外界影响较大,表现为较大的波动性。此外,同一用户在不同时刻表现出来的波动性也有所不同。

附录中图 A1 给出了 2 个用户一个月的用电行为,对同一时刻 30 d 的负荷值进行统计,形成一个箱形。图中,箱体表示负荷数据集中范围,即上下四分位间距;上、下 2 条竖线表示数据向上和向下的延伸范围;点 * 表示超出 Whisker 上限或下限(延伸至距框顶部或底部 1.5 倍框高范围内的最大或最小数据点)的负荷值。由图 A1 可看出,2 个用户在横向具有较相似的特性,均为双峰型负荷;但纵向具有较明显的差异,用户 1 在全天各个时刻的用电随机性都较低,用户 2 在负荷早高峰期以及 16:00—19:00 时间段的用电随机性较大,在负荷低谷期的用电随机性较小。由此可见,对多日的日负荷曲线进行简单的取平均值或者取某天作为典型负荷日进行负荷聚类并不能真实地反映用户不同的用电行为。

当用户多日的日负荷曲线形成一个非凸集时,取平均值后的曲线将不包含于此集中,即该平均后的曲线并不能代表用户的用电行为,这种取平均值的处理不仅会导致数据的丢失,还会导致错误的分类^[9]。不同的负荷聚类方法关注的角度不同,适用于不同的领域。当用户每天的负荷曲线差异较大,或者应用于异常用电行为分析、区间或概率性负荷预测等场景时,对用户进行聚类时应该考虑负荷的纵向波动性。

1.2 相似度量方法

相似度量(距离度量)是衡量不同对象之间相互关系的方法^[15]。定义相似度量函数应该满足以下 4 个条件:非负性、自身距离为 0、对称性以及三角不等性,分别如式(1)~(4)所示。

$$d(x, y) \geq 0 \quad (1)$$

$$d(x, x) = 0 \quad (2)$$

$$d(x, y) = d(y, x) \quad (3)$$

$$d(x, y) + d(y, z) \geq d(x, z) \quad (4)$$

其中, $d(x, y)$ 为对象 x 和 y 之间的距离。

除了下文重点介绍的 EMD 之外,衡量 2 个概率

分布之间差异的距离函数还有 Bhattacharyya 距离(巴氏距离)^[15]、KL 距离(相对熵)^[16]和卡方系数^[17],分别用 D_B 、 D_{KL} 和 $D_{\text{chi-sqr}}$ 表示,其离散形式的具体计算公式分别如式(5)~(7)所示。

$$D_B(P, Q) = -\ln \left(\sum_{i=1}^k \sqrt{P(x_i)Q(x_i)} \right) \quad (5)$$

$$D_{KL}(P, Q) = \frac{1}{2} \sum_{i=1}^k P(x_i) \ln \frac{P(x_i)}{Q(x_i)} + \frac{1}{2} \sum_{i=1}^k Q(x_i) \ln \frac{Q(x_i)}{P(x_i)} \quad (6)$$

$$D_{\text{chi-sqr}}(P, Q) = \sum_{i=1}^k \frac{[P(x_i) - Q(x_i)]^2}{P(x_i) + Q(x_i)} \quad (7)$$

其中, P 和 Q 为 2 个概率测度。

本文将在 2.4 节中分别利用包括 EMD 在内的 4 种相似度量方法对用户负荷进行聚类,比较各自的聚类效果。

1.3 EMD 简介

Peleg 等在文献[18]中最早提出 EMD, Rubner 等在文献[19]中提出将 EMD 应用于图像识别中,自此 EMD 被应用于衡量 2 个概率分布之间的差异。文献[20]证明了 EMD 满足相似度量函数的 4 个条件。Earth Mover's Distance 被译为推土机距离或者地球移动距离,顾名思义,其思想来源于一个生活问题:假设有若干数量的土堆,每个土堆的大小不相同且分布的位置不相同;同时存在若干数量的土坑,每个土坑的大小不相同且位置不相同;对于每个土堆-土坑对的运输成本是给定的(以距离表示),任务是把土堆搬动并填到土坑里,通过规划运输方案,使运输成本最低。

在直方图中体现 EMD 的示意图如附录中图 A2 所示,2 个直方图分别表示土堆和土坑。假定 2 个直方图中分别存在 m 和 n 个直方,则 2 个直方图可以被分别描述为 $P = \{(p_1, w_{p1}), (p_2, w_{p2}), \dots, (p_m, w_{pm})\}$ 和 $Q = \{(q_1, w_{q1}), (q_2, w_{q2}), \dots, (q_n, w_{qn})\}$,其中 p_i 和 q_j 分别表示土堆和土坑直方的位置, w_{pi} 和 w_{qj} 表示相对应直方的高度(土堆和土坑的大小)。2 个直方图之间的 EMD 可以转化为如式(8)所示的线性规划问题^[17]。

$$\left\{ \begin{array}{l} \min \sum_{i=1}^m \sum_{j=1}^n f_{ij} d_{ij} \\ \text{s.t.} \quad f_{ij} \geq 0 \quad 1 \leq i \leq m; 1 \leq j \leq n \\ \sum_{j=1}^n f_{ij} \leq w_{pi} \quad 1 \leq i \leq m \\ \sum_{i=1}^m f_{ij} \leq w_{qj} \quad 1 \leq j \leq n \\ \sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min \left(\sum_{i=1}^m w_{pi}, \sum_{j=1}^n w_{qj} \right) \\ d_{ij} = |p_i - q_j| \quad 1 \leq i \leq m, 1 \leq j \leq n \end{array} \right. \quad (8)$$

其中, f_{ij} 为第 i 个直方到第 j 个直方的运输数量。

上述问题即为求取最佳的运输量 $F=[f_{12}, f_{13}, \dots, f_{mn}]$, 满足一系列的约束条件, 使得整体的运输成本

$\sum_{i=1}^m \sum_{j=1}^n f_{ij} d_{ij}$ 最小。EMD 的规范化定义为:

$$f_{\text{EMD}}(P, Q) = \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} d_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (9)$$

1.4 算法流程

对电力用户的横向和纵向特性进行聚类的思路如下: 采集用户负荷数据集, 该数据集包含了一段时间(如一个月)的日负荷曲线数据; 归一化原数据; 统计每个用户每一时刻的负荷值分布特性, 形成该用户关于该时刻的分布直方图; 利用 EMD 衡量 2 个用户在同一时刻之间的负荷差异; 对所有时刻的距离值取均方根, 得到 2 个用户的距离值, 依此类推得到其他所有 2 个用户之间的距离值; 利用所得到的距离信息, 对所有用户进行层次聚类; 根据聚类有效性指标确定聚类数, 得到最终聚类结果。

1.4.1 数据归一化

电力负荷包含基荷和易变负荷两部分, 归一化的目的在于去除日负荷曲线中的基荷部分^[4], 突出曲线的走势与形状, 同时强调纵向上的随机性。假定 $L_t^k = \{l_1^k, l_2^k, \dots, l_T^k\}$ 为某一用户某一天的负荷数据, l_t^k 为归一化后的负荷值, 则具体的归一化处理计算公式如式(10)~(12)所示。

$$l_{\min}^k = \min \{l_1^k, l_2^k, \dots, l_T^k\} \quad (10)$$

$$l_{\max}^k = \max \{l_1^k, l_2^k, \dots, l_T^k\} \quad (11)$$

$$l_t^k = \frac{l_t^k - l_{\min}^k}{l_{\max}^k - l_{\min}^k} \quad t=1, 2, \dots, T \quad (12)$$

其中, $l_{\max}^k, l_{\min}^k$ 分别为第 k 日的最大、最小负荷值。

1.4.2 异常尖峰数据的识别和修正

由式(12)可知, 负荷序列的最大负荷值出现异常, 即出现异常尖峰, 会对归一化后的数据造成影响, 恶化分布特性。因此在进行数据归一化之前, 须先进行异常尖峰数据的识别和修正, 如式(13)~(15)所示。若 $\Delta l^k > \mu + 3\sigma$, 则认为该天最大负荷值异常, 对该天最大负荷值作如式(16)、(17)所示的修正, 否则不作处理。

$$\Delta l^k = [(l_{\max}^k - l_{t-1}^k) + (l_{\max}^k - l_{t+1}^k)]/2 \quad (13)$$

$$\mu = \sum_{k=1}^D \Delta l^k / D \quad (14)$$

$$\sigma^2 = [\sum_{k=1}^D (\Delta l^k - \mu)^2] / D \quad (15)$$

$$\Delta l^{k'} = \mu + 3\sigma \quad (16)$$

$$l_{\max}^{k'} = (2\Delta l^{k'} + l_{t-1}^k + l_{t+1}^k) / 2 \quad (17)$$

其中, l_{t-1}^k 和 l_{t+1}^k 分别为最大负荷值相邻 2 个时刻的负荷值; D 为总天数。

1.4.3 负荷数据的纵向统计

对负荷数据进行归一化后, 每个用户每天的负荷值将归一化到区间 $[0, 1]$ 中, 再将区间 $[0, 1]$ 精细划分为 20 个区间, 每个小区间长度为 0.05。对负荷数据的纵向分布特性进行统计, 将若干天同一时刻的负荷值分进上述 20 个区间中, 得到某一用户关于该时刻的分布直方图。用户 i 在时刻 k 的负荷分布情况如式(18)所示。

$$P_{ik} = \{(p_{ik}^1, w_{ik}^1), \dots, (p_{ik}^j, w_{ik}^j), \dots, (p_{ik}^{20}, w_{ik}^{20})\} \quad (18)$$

其中, p^j 为第 j 个直方(负荷区间)的中心位置; w^j 为第 j 个直方的高度, 即位于该负荷区间的天数。

图 1 为 2 个用户在时刻 08:00 30 d 的负荷分布情况, 可见与用户 2 相比, 用户 1 在该时刻的用电行为随机性更大。从定量角度出发, 2 个用户在该时刻的负荷分布情况可以表示为: $P_{18} = \{(0.025, 0), (0.075, 0), \dots, (0.425, 1), (0.475, 8), \dots, (0.575, 5), (0.625, 12), (0.675, 1), \dots, (0.925, 0), (0.975, 0)\}$; $P_{28} = \{(0.025, 0), (0.075, 0), \dots, (0.575, 1), (0.625, 11), (0.675, 16), (0.725, 1), (0.775, 1), \dots, (0.925, 0), (0.975, 0)\}$ 。

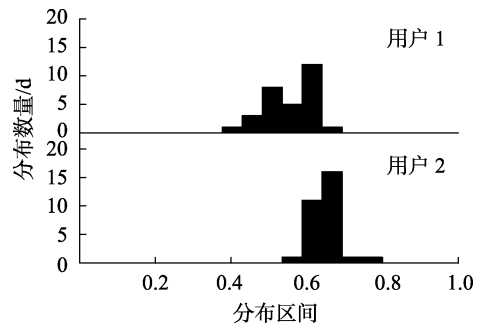


图 1 2 个用户在同一时刻的负荷分布直方图

Fig.1 Distribution histograms of two users' power consumption at the same time

1.4.4 用户用电行为差异性的衡量

利用 1.3 节介绍的 EMD 对 2 个用户在同一时刻的负荷分布直方图进行计算。由式(8)可以看出, 利用 EMD 计算 2 个直方图之间的差异时, 并不要求二者的直方面积一致, 因此对数据的完整性要求较低, 即使 1 d 中某一时刻的数据缺失, 甚至一整天的数据缺失, 也不影响计算结果。对 2 个用户所有时刻的距离值求取均方根, 该结果即为 2 个用户用电行为差异的度量值。具体的计算公式如式(19)、(20)所示。

$$d_{ij}^k = f_{\text{EMD}}(P_{ik}, P_{jk}) \quad (19)$$

$$D_{ij} = \sqrt{(d_{ij}^1)^2 + (d_{ij}^2)^2 + \dots + (d_{ij}^T)^2} \quad (20)$$

1.4.5 聚类效果评价指标

由于目前对双向负荷聚类的研究很少,尚缺乏综合衡量此聚类效果的指标,故本文定义如下 2 个聚类评价指标:负荷横向波动系数 TLFC (Transverse Load Fluctuation Coefficient) 和负荷纵向波动系数 LLFC (Longitudinal Load Fluctuation Coefficient),分别用于衡量用户横向与纵向特性的一致性,以二者的乘积作为聚类效果的综合指标。分别用 f_{TLFC} 和 f_{LLFC} 表示某类用户的 TLFC 值和 LLFC 值,其计算公式分别如式(21)、(22)所示。

$$\begin{cases} f_{\text{TLFC}} = \frac{1}{24} \sum_{t=1}^{24} \sqrt{\frac{1}{N_{\text{users}}} \sum_{j=1}^{N_{\text{users}}} (\bar{l}_j^t - \bar{l}^t)^2} \\ \bar{l}_j^t = \frac{1}{D} \sum_{k=1}^D l_{jk}^t \\ \bar{l}^t = \frac{1}{N_{\text{users}}} \sum_{j=1}^{N_{\text{users}}} \bar{l}_j^t \end{cases} \quad (21)$$

$$\begin{cases} f_{\text{LLFC}} = \frac{1}{24} \sum_{t=1}^{24} \sqrt{\frac{1}{N_{\text{users}}} \sum_{j=1}^{N_{\text{users}}} (v_{\text{arj}}^t - \bar{v}^t)^2} \\ v_{\text{arj}}^t = \sqrt{\frac{1}{D} \sum_{k=1}^D (l_{jk}^t - \bar{l}_j^t)^2} \\ \bar{v}^t = \frac{1}{N_{\text{users}}} \sum_{j=1}^{N_{\text{users}}} v_{\text{arj}}^t \end{cases} \quad (22)$$

其中, l_{jk}^t 为用户 j 在第 k 天时刻 t 的负荷值; N_{users} 为该用户总数。

TLFC 和 LLFC 的数学意义是计算同一类别各个用户之间的“负荷均值的方差”以及“负荷方差的方差”。从物理意义上来分析,式(21)中的 $\bar{l}_j^t$ 和式(22)中的 v_{arj}^t 分别表示用户 j 在时刻 t 的负荷均值和负荷方差,并以此值表示该用户在此时刻的横向特性和纵向特性。 $\bar{l}^t$ 和 $\bar{v}^t$ 分别表示该类用户在时刻 t 的横向聚类中心和纵向聚类中心,通过计算各个用户的横(纵)向特性到横(纵)向聚类中心的距离,并对各个时刻的值取平均,即可得到最终的 f_{TLFC} 和 f_{LLFC} 。其数值越小,表示该类用户的横(纵)向一致性越高。

2 算例分析

本文选取一组由美国能源部公开能源信息网站提供的居民用电负荷数据为研究对象,数据集包含 936 个居民用户 1 a 的用电负荷数据,每 60 min 采集一次,1 d 共计 24 个测量点。

2.1 确定聚类数目

本文所选用的基于层次的聚类方法本质上是无需预先确定聚类数,可以在得到树状图之后根据不同的需要确定聚类数目,并据此对树状图进行分割,但这种处理方法对于聚类数的确定过于主观。为了

使聚类数的确定更具客观性,本文利用 3 个有效性指标确定聚类数,并选取其中 2 个聚类指标: HS (Homogeneity-Separation) 指标和 CH (Calinski-Harabasz) 指标综合确定最佳聚类数。2 个指标值的具体计算公式分别如式(23)、(24)所示。

$$\begin{cases} f_{\text{HS}}(a) = |f_{\text{Hom}}(a) - f_{\text{Sep}}(a)| \\ f_{\text{Hom}}(a) = \frac{2}{\sum_{i=1}^a n_i(n_i - 1)} \sum_{i=1}^a \sum_{t \in C_i, s < t} R(s, t) \\ f_{\text{Sep}}(a) = \frac{1}{\sum_{i,j=1; i < j}^a n_i n_j} \sum_{i,j=1; i < j}^a \sum_{s \in C_i, t \in C_j} R(s, t) \end{cases} \quad (23)$$

$$f_{\text{CH}}(a) = \frac{\text{tr}[\mathbf{S}_B(a)] / (a-1)}{\text{tr}[\mathbf{S}_W(a)] / (a-1)} \quad (24)$$

其中, a 为聚类的数目; $\text{tr}(\cdot)$ 表示矩阵的迹; $\mathbf{S}_B(a)$ 为聚类数 a 下的类间离差矩阵; $\mathbf{S}_W(a)$ 为聚类数 a 下的类内离差矩阵; $R(s, t)$ 为样本 s 和样本 t 之间的距离; n_i 和 n_j 分别为第 i 个聚类 C_i 和第 j 个聚类 C_j 的样本数。

HS 和 CH 指标均以曲线中的最大值点对应的聚类数作为最佳聚类数。由附录中图 A3 可知,利用 HS 和 CH 指标得到的最佳聚类数均为 4,故选取 4 作为本算例的聚类数。

2.2 用户聚类结果

选取聚类数为 4 进行层次聚类后,将每一类中所有用户的负荷箱形图进行综合,垂直方向上矩形盒越长,表明该类用户在这一时刻的纵向波动性较大。用户的负荷横向特性可由矩形箱的中位数点所连曲线表示。通过聚类得到的每一类用户的综合箱形图如图 2 所示,图中负荷为标么值。

由图 2 可以看出,4 类负荷均为双峰型负荷。在横向特性方面,第 1—3 类用户的谷期出现在时间段 12:00—16:00,第 4 类用户谷期出现在时间段 08:00—10:00;第 3 类用户的早高峰负荷明显高于其他 3 类用户;第 1、2 类用户的负荷横向特性较为相似。在纵向特性方面,第 2 类用户的整体波动性小于其他 3 类用户;第 2 类用户的谷期波动性高于峰期,而第 1、3 类用户则与之相反,第 4 类用户谷期和峰期负荷波动性接近,负荷上升期波动较大。由此可见,从定性角度来看,EMD 很好地对用户的负荷横向及纵向特性的差异进行衡量。

根据信息熵的理论可知,某一变量差异度越大,其中包含的信息量就越多,该变量对系统评估就越有价值^[21]。引申至工程技术角度,某类用户的负荷纵向随机性越大,表示该类用户受外部因素影响越大。作为电力调度中心和用户之间的中间机构的

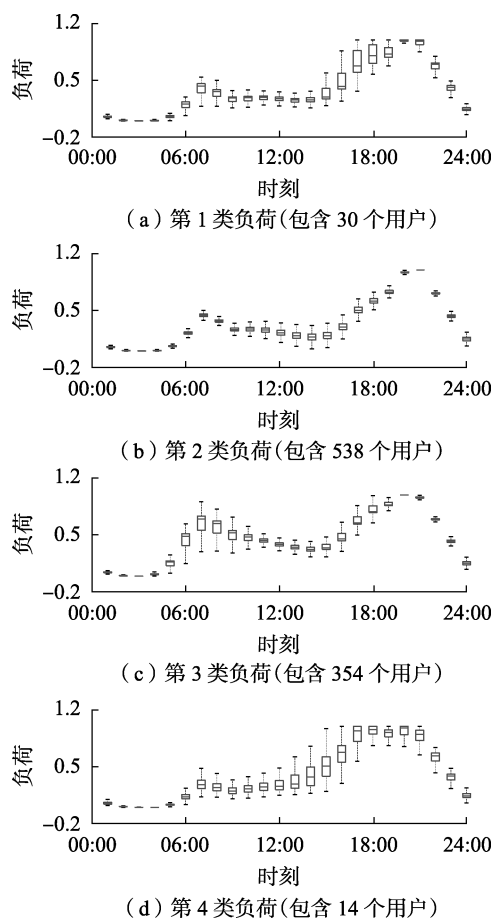


图2 基于EMD聚类算法的用电行为箱形图

Fig.2 Box diagram of power consumption behavior based on EMD clustering algorithm

“负荷聚合商^[22]”,可对该类用户进行更加深入的数据挖掘,从而制定更加精细化的分时调度控制策略。以上4类用户中,第2类用户所包含的信息量最少;第1、3和4类用户分别在时间段16:00—19:00、05:00—09:00以及12:00—16:00有较高的数据挖掘价值。

2.3 聚类结果比较

为了进一步地从定量的角度证明本文所提方法的合理性,采用传统方法,即利用相同的数据集,取相同的聚类数,将一个用户多日的负荷数据取平均值作为该用户典型日负荷曲线,然后利用层次聚类算法聚类典型日负荷曲线。得到2种方法下的各类包含的用户数如附录中表A1所示。经过对比分析可知,类别3和类别4的用户标号一致,即聚类结果相同,类别1和类别2的聚类效果存在差异。

为了比较2种方法的聚类效果,附录中表A2比较了本文方法和传统方法以及未进行聚类时各个类别的综合聚类指标($f_{\text{TLFC}} f_{\text{LLFC}}$),表中数值越小,则表示同一类别中各个用户在一个半月内的用电行为越相近。由此可知,相较于未进行聚类的指标值,2种方法中各个类别的综合聚类指标值都有较大程度的降

低。然后比较本文方法和传统方法,在类别1和类别2中,本文方法的综合聚类指标比传统方法分别低4.1%和1.4%。结果表明本文方法相较于传统方法具有更高的合理性。

2.4 不同度量方法的比较

为证明采用EMD度量用户行为差异的有效性,分别利用EMD、Bhattacharyya距离、KL距离和卡方系数对该数据集进行聚类,得到的聚类结果如附录中表A3所示。可见Bhattacharyya距离和KL距离完全无法对不同的用户进行区分。观察二者所形成的距离矩阵,可以发现以下问题:当2个概率分布差异较大时,

$\sum_{i=1}^a \sqrt{P(x_i)Q(x_i)}$ 将趋近于0,从而导致Bhattacharyya距离出现无穷大值;KL距离中 $P(x_i)$ 和 $Q(x_i)$ 均出现在分母位置和对数函数的真数上,当其中某一个为0时,会导致无法计算距离值。

卡方系数也能粗略地对用户进行区分,但计算卡方系数所形成的4个类别的综合聚类指标分别为0、0.004 1、 3.551×10^{-4} 和 1.355×10^{-5} ,与附录中表A2结果相比,聚类效果明显不如EMD。进一步地观察利用卡方系数所形成的距离矩阵可知,矩阵内存在大量完全相同的距离值,造成该现象的原因在于:当2个概率分布重合部分较小,甚至完全分离时,无论分离得多远,计算得到的卡方系数都将保持不变。因此当样本之间差异较大时,卡方系数并不十分可靠。由此可见,采用EMD度量2个概率分布的差异具有较广的使用场合和更高的准确性。

3 结论

本文同时考虑了电力用户的负荷横向特性和负荷纵向特性,首次提出了基于EMD算法的负荷聚类方法。EMD算法相较于其他度量方法能够更好地度量2个分布之间的差异。比较了考虑和不考虑负荷纵向特性的聚类效果,结果表明本文方法能够同时对负荷的横向和纵向差异进行度量。在用户的负荷横向特性较为相似的情况下,通过对负荷纵向特性的有效反映,使得本文方法下的综合聚类指标比传统方法更低。由于EMD算法涉及较多的线性规划运算,运行效率不高,下一步的研究方向在于如何提高算法的运行速度,以及如何结合气象等因素进一步提高聚类的合理性。

附录见本刊网络版(<http://www.epae.cn>)。

参考文献:

- [1] 鞠平,周孝信,陈维江,等. “智能电网+”研究综述[J]. 电力自动化设备,2018,38(5):2-11.
JU Ping, ZHOU Xiaoxin, CHEN Weijiang, et al. “Smart grid plus” research overview [J]. Electric Power Automation Equipment, 2018,38(5):2-11.

- [2] 郝然, 艾芊, 肖斐. 基于多元大数据平台的用电行为分析构架研究[J]. 电力自动化设备, 2017, 37(8): 20-27.
HAO Ran, AI Qian, XIAO Fei. Architecture based on multivariate big data platform for analyzing electricity consumption behavior[J]. Electric Power Automation Equipment, 2017, 37(8): 20-27.
- [3] 黄宇腾, 侯芳, 周勤, 等. 一种面向需求侧管理的用户负荷形态组合分析方法[J]. 电力系统保护与控制, 2013, 41(13): 20-25.
HUANG Yuteng, HOU Fang, ZHOU Qin, et al. A new combination electrical load analysis method for demand side management[J]. Power System Protection and Control, 2013, 41(13): 20-25.
- [4] 贾慧敏, 何光宇, 方朝雄, 等. 用于负荷预测的层次聚类与双向夹逼结合的多层次聚类法[J]. 电网技术, 2007, 31(23): 33-36.
JIA Huimin, HE Guangyu, FANG Chaoxiong, et al. Load forecasting by multi-hierarchy clustering combining hierarchy clustering with approaching algorithm in two directions[J]. Power System Technology, 2007, 31(23): 33-36.
- [5] 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, 40(5): 1495-1501.
WANG Xinghua, CHEN Zhuoyou, PENG Xiangang. A new combination electrical load analysis method based on bilayer clustering analysis[J]. Power System Technology, 2016, 40(5): 1495-1501.
- [6] 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.
- [7] METS K, DEPUYDT F, DEVELDER C. Two-stage load pattern clustering using fast wavelet transformation[J]. IEEE Transactions on Smart Grid, 2016, 7(5): 2250-2259.
- [8] LI R, LI F, SMITH N D. Multi-resolution load profile clustering for smart metering data[J]. IEEE Transactions on Power Systems, 2016, 31(6): 4473-4482.
- [9] 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, ZUANG 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.
- [10] 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.
- [11] 李智勇, 吴晶莹, 吴为麟, 等. 基于自组织映射神经网络的电力用户负荷曲线聚类[J]. 电力系统自动化, 2008, 32(15): 66-70, 78.
LI Zhiyong, WU Jingying, WU Weilin, et al. Power customers load profile clustering using the SOM neural network[J]. Automation of Electric Power Systems, 2008, 32(15): 66-70, 78.
- [12] 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.
- [13] SUN M, KONSTANTELOS I, STRBAC G. C-vine Copula mixture model for clustering of residential electrical load pattern data[J]. IEEE Transactions on Power Systems, 2017, 32(3): 2382-2393.
- [14] 李海林, 郭崇慧. 时间序列数据挖掘中特征表示与相似性度量研究综述[J]. 计算机应用研究, 2013, 30(5): 1285-1291.
LI Hailin, GUO Chonghui. Survey of feature representations and similarity measurements in time series data mining[J]. Application Research of Computers, 2013, 30(5): 1285-1291.
- [15] COMANICIU D, RAMESH V, MEER P. Real-time tracking of non-rigid objects using mean shift[C]//IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head Island, SC, USA: IEEE, 2000: 142-149.
- [16] 杨坚, 罗四维, 刘蕴辉. 一种基于广义 KL 距离和几何曲率的模型选择准则[J]. 电子学报, 2005, 33(12): 2272-2277.
YANG Jian, LUO Siwei, LIU Yunhui. New model selection criterion based on Kullback-Leibler information divergence and geometric curvature[J]. Acta Electronica Sinica, 2005, 33(12): 2272-2277.
- [17] HUANG L, BATAINEH M A, ATKIN G E, et al. Identification of transcription factor binding sites based on the Chi-Square (χ^2) distance of a probabilistic vector model[C]//2009 International Conference on Future Biomedical Information Engineering (FBIE). Sanya, China: IEEE, 2009: 73-76.
- [18] PELEG S, WERMAN M, ROM H. A unified approach to the change of resolution: space and gray-level[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1989, 11(7): 739-742.
- [19] RUBNER Y, TOMASI C, GUIBAS L J. The earth mover's distance as a metric for image retrieval[J]. International Journal of Computer Vision, 2000, 40(2): 99-121.
- [20] 李展, 彭进业, 温超. 基于 EMD 距离的多示例聚类[J]. 计算机科学, 2011, 38(7): 235-239.
LI Zhan, PENG Jinye, WEN Chao. Multi-instance clustering based on EMD[J]. Computer Science, 2011, 38(7): 235-239.
- [21] LIN Z, WEN F, HUANG J, et al. Evaluation of black-start schemes employing entropy weight-based decision-making theory[J]. Journal of Energy Engineering, 2010, 136(2): 42-49.
- [22] 孙玲玲, 高赐威, 谈健, 等. 负荷聚合技术及其应用[J]. 电力系统自动化, 2017, 41(6): 159-167.
SUN Lingling, GAO Ciwei, TAN Jian, et al. Load aggregation technology and its applications[J]. Automation of Electric Power Systems, 2017, 41(6): 159-167.

作者简介:



冯志颖

冯志颖(1994—),男,广东广州人,硕士研究生,主要研究方向为电力系统数据分析(**E-mail**: epzyfeng@mail.scut.edu.cn);

唐文虎(1974—),男,江苏盐城人,教授,博士研究生导师,博士,国家青年千人计划特聘教授, IET Fellow, 主要研究方向为电力设备智能化、电力系统状态风险评估和

计算机智能及应用等(**E-mail**: wenhutang@scut.edu.cn);

吴青华(1953—),男,湖北武汉人,教授,博士研究生导师,博士,国家千人计划特聘教授,广东省海外引进创新团队带头人, IEEE Fellow, IET Fellow, InstMC Fellow, 主要研究方向为电力系统及其自动化。

(下转第 53 页 continued on page 53)

Error state correlation analysis based on random matrix theory for electronic transformer

HU Chen¹, ZHANG Zhu¹, JIAO Yang¹, LI Hongbing¹, CHEN Gang²

(1. State Key Laboratory of Strong Electromagnetic Engineering and New Technology, Huazhong University of Science and Technology, Wuhan 430074, China; 2. Electric Power Research Institute of State Grid Jiangsu Electric Power Company, Nanjing 211103, China)

Abstract: Since the influencing factors of electronic transformer error involve interactions between each other in complex field conditions, resulting in difficult decoupling implementation. Therefore, it is difficult to confirm the relationship between the error state of electronic transformer and the influencing factors. In order to solve this problem, an error state correlation analysis method for electronic transformer is proposed based on high-dimensional random matrix theory. The evaluation matrix is constructed in real-time using moving time-window and then extended based on Kalman filter in the case of sparse matrix. Two evaluation indexes of correlation level, i.e. d_{MSR} and I_{MSR} are proposed and the steps of correlation analysis are confirmed. The proposed method is adopted to analyze the operating data of the electronic transformer error state monitoring platform, the results show that it can be utilized to confirm the influencing degree of single or multiple influencing factors on the error state of electronic transformer.

Key words: electronic transformers; random matrix; influencing factor; single-ring law; evaluation index; error state

(上接第44页 continued from page 44)

Users' consumption behavior clustering method considering longitudinal randomness of load

FENG Zhiying¹, TANG Wenhui¹, WU Qinghua¹, LU Guojun², LUAN Le²

(1. School of Electric Power, South China University of Technology, Guangzhou 510640, China; 2. Guangzhou Power Supply Co., Ltd., Guangzhou 510000, China)

Abstract: To address the problem of data loss and misclassification caused by not considering the longitudinal randomness of load, a novel user's consumption behavior recognition method based on EMD (Earth Mover's Distance), which considers the longitudinal randomness of load, is proposed. The approach takes both the transverse and longitudinal perspectives into account to fully characterize the power consumption behaviors of electricity consumers by collecting the load distribution of electricity users at the same moment for a few days. The EMD is combined with the Euclidean distance to measure the difference degrees among different user's electricity consumption behaviors. A group of international general household electricity load is employed as a benchmark example. Case study results indicate that the proposed method can better capture the longitudinal characteristics of customers with similar transverse characteristics. Both quantity and quality analysis show that the method is accurate and feasible for clustering the users' load.

Key words: smart grid; consumption behavior; load curve; longitudinal randomness; clustering algorithms; earth mover's distance

附录 A

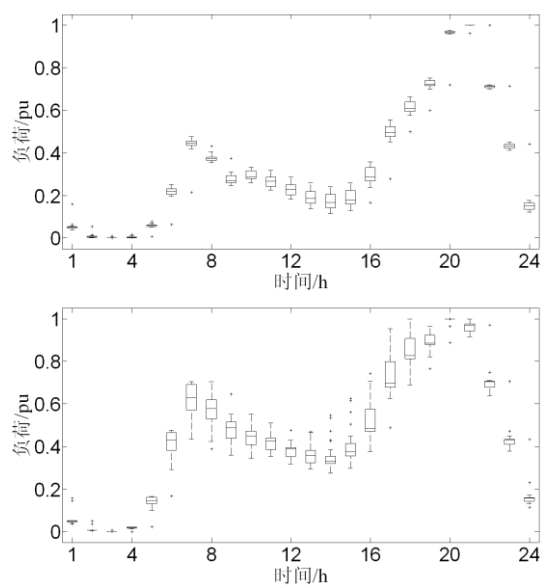


图 A1 2 个用户的用电行为箱形图

Fig. A1 Box diagrams of two users' power consumption behavior

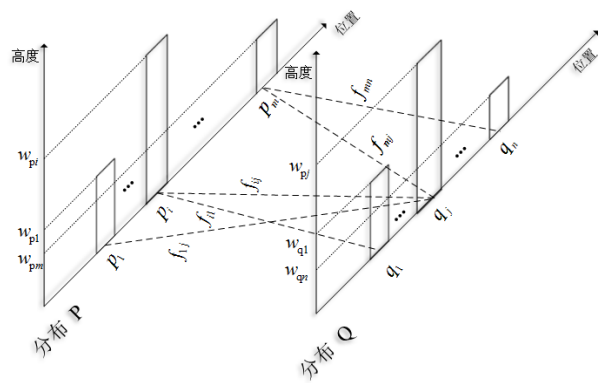
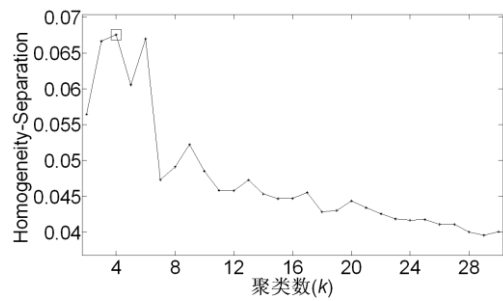


图 A2 EMD 示意图

Fig. A2 Schematic diagram of the EMD problem



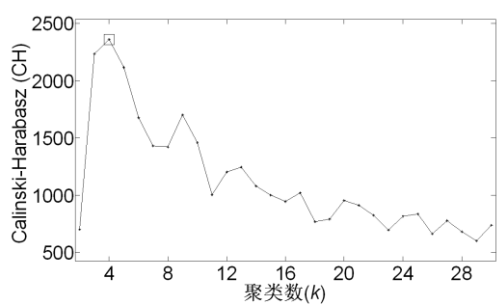


图 A3 聚类数和有效性指标的关系

Fig.A3 Relationship between cluster number and validity index

表 A1 各类用户簇包含的用户数

Table A1 Number of users of each type of user cluster				
方法	类别 1	类别 2	类别 3	类别 4
本文方法	30	538	354	14
传统方法	42	526	354	14

表 A2 类内用户综合聚类指标

Table 2 Comprehensive clustering index within class				
方法	类别 1	类别 2	类别 3	类别 4
本文方法	6.605e-4	1.846e-5	3.359e-4	9.660e-10
传统方法	6.876e-4	1.872e-5	3.359e-4	9.660e-10
未聚类	9.275e-4			

表 A3 不同度量方法的聚类结果

Table A3 Clustering results of different measurement methods	
相似度度量	各类别包含的样本数
EMD	30, 538, 354, 14
Bhattacharyya 距离	933, 1, 1, 1
KL 距离	933, 1, 1, 1
卡方系数	1, 2, 225, 708