

DOI: 10.19783/j.cnki.pspc.200409

一种基于时间序列数据挖掘的用户负荷曲线分析方法

唐俊熙¹, 曹华珍¹, 高崇¹, 吴亚雄¹, 石颖²

(1. 广东电网有限责任公司电网规划研究中心, 广东 广州 510080;

2. 北京清软创新科技股份有限公司, 北京 100085)

摘要: 针对目前用户负荷曲线分析方法研究中传统方法在单一用户典型负荷曲线提取以及负荷曲线特征提取的不足, 提出了一种基于时间序列数据挖掘的用户负荷曲线分析方法。该方法首先基于分段聚合近似方法对单一用户的负荷曲线降维, 并基于符号聚合近似方法对该用户一段时期内的负荷曲线重表达, 用符号化序列表示该用户的负荷曲线, 提取该用户的典型负荷曲线。然后结合不同用户典型负荷曲线的负荷特性、指标特征和时间序列特征, 基于 k -means 算法对不同用户的典型负荷曲线聚类分析, 分析不同类型用户的用电特征。以 UCI 一个测试数据集进行算例分析, 结果表明所提方法能够挖掘出用户的典型用电行为特征, 并提升用户负荷曲线分析效率与聚类质量。
关键词: 数据挖掘; 符号聚合近似; 典型负荷曲线; k -means; 聚类分析

A new user load curve analysis method based on time series data mining

TANG Junxi¹, CAO Huazhen¹, GAO Chong¹, WU Yaxiong¹, SHI Ying²

(1. Grid Planning & Research Center, Guangdong Power Grid Co., Ltd., Guangzhou 510080, China;

2. Beijing Tsingsoft Innovation Technology Co., Ltd., Beijing 100085, China)

Abstract: There is a shortage of traditional methods for extracting the typical user load curve and the feature extraction of the load curve in the current user load curve analysis method. This paper proposes a user load curve analysis method based on time series data mining. First, the method reduces the dimension of the load curve of a single user based on piecewise aggregate approximation, and re-expresses the load curve of the user for a period based on the symbolic aggregate approximation method, and represents the user load curve with a symbolized sequence to extract the typical load curve of the user. Then, combined with the load characteristics and time series characteristics of the typical load curve of different users, based on the k -means algorithm, the typical load curves of different users are clustered to analyze the power consumption characteristics of different types of users. A UCI test data set is used for case analysis, and the results show that the proposed method can mine the typical power consumption characteristics of users and improve the user load curve analysis efficiency and clustering quality.

This work is supported by the Science and Technology Project of China Southern Power Grid Co., Ltd. (No. GDKJXM20173251).

Key words: data mining; symbolic aggregate approximation; typical daily load curve; k -means; cluster analysis

0 引言

随着智能电网的发展和高级量测体系的建设, 电力系统积累海量的基础用电数据, 这些数据背后蕴含了丰富的信息, 充分利用海量的电力大数据挖掘出数据背后的有用信息具有重要意义^[1-3]。电力用户是电力系统的重要组成部分, 深入挖掘电力用户

负荷数据的内在规律对电力系统规划、运行等具有重要意义, 因而受到广泛的关注与研究。数据挖掘的核心目的就是就是从海量数据中提取知识、挖掘知识, 而用户的用电数据中隐藏着其用电习惯^[4-5], 研究和挖掘用电数据^[6], 可以帮助电网了解用户的个性化、差异化服务需求^[7], 从而使电网公司进一步拓展服务的深度和广度。

对于单一电力用户的典型负荷曲线提取, 传统方法^[8]主要是在用户一段时期内所有负荷曲线中,

基金项目: 中国南方电网公司科技项目资助 (GDKJXM20173251)

手动选取一条与年平均负荷率相近且无畸变的负荷曲线作为该用户的典型负荷曲线。然而传统方法忽略了不同用户的负荷特征差异性, 不具有通用性且误差较大。文献[8]提出一种基于改进模糊聚类的典型负荷曲线选取方法, 能够辨识负荷数据中的异常负荷曲线形态, 剔除其影响; 文献[9]通过取工作日负荷的平均值方法来提取典型负荷曲线, 从而削弱异常日负荷的影响; 文献[10]提出将典型负荷曲线的选取问题转化为基于统计学习的多元分类问题。电力用户负荷曲线容易受到故障或计量装置误差的影响而产生异常负荷曲线形态, 同时不同用户的负荷曲线差异较大, 其异常负荷曲线形态也各有不同, 上述研究中不能有效地剔除用户的异常负荷曲线, 存在一定缺陷。

此外, 随着智能电表的广泛普及, 电力用户的负荷数据采样频率越来越高, 已经可以较为精细地掌握负荷的波动情况^[11]。在对这些高维的负荷数据进行聚类分析时, 其计算量大且极为耗时, 因此有必要对负荷数据降维处理, 提取用户负荷曲线的关键特征, 保证聚类有效性的同时可以减小数据计算的复杂度。目前对负荷数据降维主要是通过定义相关负荷特性指标^[12-13]来描述负荷曲线关键特征。文献[14]分析了负荷曲线聚类分析的计算复杂性和特征选择的有效性, 提出一种特征优选的策略。但以上研究的特征选择都是基于负荷特性指标, 没有从时间序列数据挖掘的角度提取用户负荷曲线特征, 从而影响负荷曲线的聚类效果。

综上所述, 目前用户负荷曲线分析方法研究主要存在以下两点不足: 1) 缺乏有效的方法提取用户的典型负荷曲线, 对用户异常负荷曲线的剔除效果不佳, 这对于分析用户的典型用电特征将产生一定影响; 2) 常规的负荷特征提取不能反映负荷曲线的内在特征, 对负荷曲线的特征提取不充分, 将导致聚类效果不佳, 误差较大。

本文针对上述问题, 在现有研究的基础上, 提出一种基于时间序列数据挖掘的用户负荷曲线分析方法。该方法首先基于分段聚合近似方法对单一用户的负荷曲线降维, 并基于符号聚合近似方法对该用户一段时期内负荷曲线重表达, 用符号化序列表示该用户的负荷曲线, 同时辨识并剔除该用户的异常负荷曲线, 提取该用户的典型负荷曲线; 然后以此为基础, 提取获得不同用户的典型负荷曲线, 结合不同用户典型负荷曲线的负荷特性指标特征和时间序列特征, 基于 k -means 算法对不同用户的典型负荷曲线聚类分析, 分析不同类型用户的用电特征。最后通过算例分析表明, 所提方法能够挖掘出不同

类型用户的典型用电行为特征, 取得了较好的用户分类效果。

本文主要的创新点包括以下两点: 1) 采用时间序列数据挖掘中的分段聚合近似方法和符号聚合近似方法, 分析和筛选出用户最频繁的负荷曲线形态, 进而提取该用户典型负荷曲线; 2) 提出基于时间序列分析方法提取不同用户的负荷曲线特征, 所提取特征能够有效提高聚类质量。

1 单一用户的典型负荷曲线提取

日负荷曲线的符号表示方法流程主要包括 5 个部分: 原始负荷数据收集、数据预处理、数据降维、低维负荷数据重表达和提取该用户的典型负荷曲线形态。单一用户的典型负荷曲线提取基本流程图如图 1 所示。

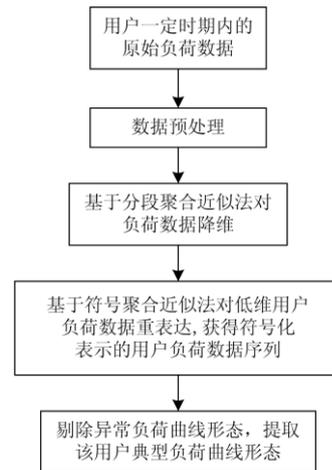


图 1 单一用户的典型负荷曲线提取基本流程图

Fig. 1 Basic flow chart of typical load curve extraction for a single user

1.1 原始负荷数据预处理

数据预处理对于任何数据挖掘方法都是至关重要的。本文定义距离平均值大于 3 倍标准差的数据为异常值, 同时为了使数据更能够体现负荷用电行为的变化特征, 采用 Z 标准化对数据处理。

假设原始日负荷数据为 $X = \{x_1, \dots, x_n\}$, 首先剔除原始日负荷数据中的异常值, 然后再对处理后数据进行 Z 标准化, 即将日负荷数据变换为均值为 0, 标准差为 1, 即 $X' = \{x'_1, \dots, x'_n\}$ 。

$$x'_i = \frac{x_i - \mu}{\sigma}, i = 1, 2, \dots, n \quad (1)$$

式中: x_i 和 x'_i 分别表示第 i 个时刻实际用电负荷和标准化后的用电负荷; n 表示 1 天中负荷采样点数; μ 和 σ 分别表示日负荷数据的均值和标准差。

间最常见的负荷曲线形态, 由式(3)可以提取该用户这段时间的典型负荷曲线。

2 不同用户的典型负荷曲线的聚类分析

时间序列的聚类可以基于原始样本数据进行聚类, 也可以基于时间序列的特征或模型聚类^[17]。特征表示方法可以将高维空间的时间序列映射到低维特征空间, 实现数据降维, 同时还能够有效反映原时间序列的信息。

用户典型负荷曲线反映了该用户的用电行为习惯, 而用户的行为习惯受气象、用户类型和电价政策等的影响较大, 故用户的典型负荷曲线存在多样性。选择有效的用户典型负荷曲线特征, 能够反映该用户的用电行为的内在特征, 同时可以减小聚类计算的复杂度, 提高分析性能。

本文对用户负荷曲线特征的选取将结合常规的负荷特性指标特征和基于时间序列的特征, 然后以 K-means 算法为基础, 对不同用户的典型负荷曲线进行聚类分析, 选择聚类评价指标作为评价聚类有效性和确定最佳聚类数^[18], 比较不同的特征提取方法对聚类结果的影响。

2.1 负荷曲线特征

2.1.1 常规的负荷特征提取

常规的负荷特征是指电力系统常用的负荷特性指标, 如负荷率、峰谷差、平均负荷等 15 个指标, 这些指标能够反映用电负荷的特点和用电性质。文献[12]从全天、峰期、平期、谷期 4 个角度, 选取 6 种常用日负荷特性指标, 较全面地反映了各类用户的用电特性, 如表 2 所示。

表 2 负荷特性指标

Table 2 Load pattern indexes

负荷曲线特征	定义	时段
负荷率	$A_1=P_{av}/P_{max}$	全天
最高利用小时率	$A_2=P_{sum}/(24P_{max})$	全天
日峰谷差率	$A_3=(P_{max}-P_{min})/P_{max}$	全天
峰期负载率	$A_4=P_{av,peak}/P_{av}$	08:00—11:00 18:00—21:00
平期负载率	$A_5=P_{av,sh}/P_{av}$	06:00—08:00 11:00—18:00 21:00—22:00
谷期负载率	$A_6=P_{av,val}/P_{av}$	22:00—24:00 00:00—06:00

注: P 代表负荷; 下标 sum、av、max、min 分别代表总量、均值、最大值、最小值; peak、val、sh 分别代表峰期、谷期、平期。

2.1.2 基于时间序列的负荷曲线特征选取

常规的负荷特征提取无法反映负荷曲线的复杂度和波动性等特征, 本文从时间序列数据挖掘角度引入 4 个负荷曲线特征。

1) 分箱熵^[19](Binned Entropy, BE)

把负荷序列 X 的取值进行分箱操作, 可以把 $[\min(X), \max(X)]$ 这个区间等分为 k 个小区间, 那么负荷序列的取值将分散这 k 个箱中。根据这个等距分箱的情况, 就可以计算出这个概率分布的熵。

$$BE(X) = - \sum_{k=0}^{\min(\max bins, len(X))} p_k \log(p_k) \cdot 1_{(p_k > 0)} \quad (4)$$

式中: p_k 表示负荷序列的取值落在第 k 个桶的比例(概率); $\max bins$ 表示桶的个数; $len(X)$ 表示时间序列的长度。

2) 时间序列复杂度不变距离^[20](Complexity-Invariant Distance, CID)

CID 值可以衡量负荷曲线数据的复杂度, CID 值越大说明该负荷曲线越复杂, 波动性越大, 具有更多的峰谷。

$$CID = \sqrt{\sum_{i=0}^{n-2lag} (x_i - x_{i+1})^2} \quad (5)$$

式中: x_i 表示第 i 个时刻用户负荷数据; n 表示 1 天中负荷采样点数; lag 表示滞后阶数。

3) 时间序列的非线性度量^[21](Nonlinear Metrics, NM)

$$NM = \frac{1}{n-2lag} \sum_{i=0}^{n-2lag} x_{i+2lag}^2 \cdot x_{i+lag} \cdot x_i \quad (6)$$

式中: x_i 表示第 i 个时刻用户负荷数据; n 表示 1 天中负荷采样点数; lag 表示滞后阶数。

4) 平均绝对值变化(Mean Absolute Change, MAC)

$$MAC = \frac{1}{n} \sum_{i=1}^{n-1} |x_{i+1} - x_i| \quad (7)$$

式中: x_i 表示第 i 个时刻用户负荷数据; n 表示 1 天中负荷采样点数。平均绝对值变化是指后一时刻的负荷值和前一时刻的负值绝对值之差的算术平均值, 可以反映负荷曲线的波动性。

2.2 用户典型负荷曲线的聚类分析

本文以 k -means 聚类算法^[22]为基础对用户典型负荷曲线进行聚类分析, 研究不同用户的典型负荷曲线的相似性。 k -means 聚类算法是一种基于划分的硬聚类算法。以误差的平方和(Sum of Squared Errors, SSE)作为衡量聚类效果的目标函数。 k -means 具体算法步骤如下。

步骤1 将 m 个用户的一天 96 点典型负荷曲线 $M = \{X_1, \dots, X_m\}$, 提取负荷曲线特征后 $M = \{T_1, \dots, T_m\}$, 从原来 $m \times 96$ 维下降为 $m \times 10$ 维, 聚类为 k 个曲线簇 $S = \{S_1, S_2, \dots, S_k\}$, 并任意选择 k 个初始簇中心 $\mu_i (i = 1, 2, \dots, k)$ 。

步骤2 计算每条负荷曲线特征 $T_j (j = 1, 2, \dots, m)$ 与簇中心 μ_j 的欧式距离 $d = \|T_j - \mu_j\|$, 并根据最小距离划分负荷曲线, 形成类簇 $S_i (i = 1, 2, \dots, k)$ 。

步骤3 以每个类簇中日负荷曲线的均值作为更新后的新聚类中心, 即

$$\mu_i = \frac{1}{|S_i|} \sum_{T_j \in S_i} T_j \quad (8)$$

步骤4 重复以上步骤, 直到簇中心不再变化。

本文采用“手肘法”^[23]来确定 k -means 的最佳聚类数目 k 。该方法通过计算 SSE 并绘制 k -SSE 曲线图, 观察图中的拐点来确定聚类数目 k 。

2.3 聚类评价指标

采用 Silhouette(SIL)指标和 Calinski-Harabas (CH)指标来进行聚类效果评价^[18]。

1) SIL 指标

SIL 指标是对群集内聚性和群集内分离度的度量。SIL 指标值越高, 聚类效果越佳, 取值-1 到 1。通过式(9)计算系数。

$$SIL = \frac{b - a}{\max(a, b)} \quad (9)$$

式中: a 为样本与同一聚类中所有其他点之间的平均距离; b 为样本与下一个最近聚类中所有其他点之间的平均距离。

2) CH 指标

CH 指标通过计算类内样本与该类聚类中心距离平方和(即类内离差矩阵)来量度紧密度, 通过计算类间离差矩阵量度分离度。CH 指标值为分离度与紧密度比值, CH 指标值越大, 聚类效果更佳。

$$CH = \frac{\text{tr}(\mathbf{B}_k) m - k}{\text{tr}(\mathbf{W}_k) k - 1} \quad (10)$$

式中: m 为训练集样本数; k 为类别数; \mathbf{B}_k 为类别

之间的协方差矩阵; \mathbf{W}_k 为类别内部数据的协方差矩阵; tr 为矩阵的迹。

3 算例分析

本文所提出的基于时间序列数据挖掘的用户负荷曲线分析方法具体流程如图 4 所示。

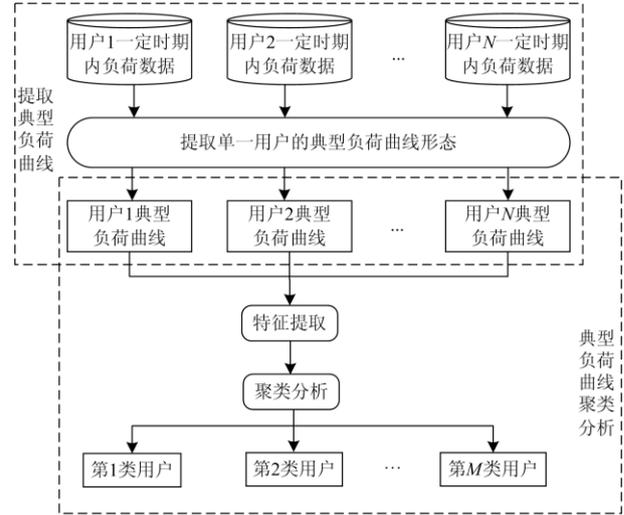


图 4 用户负荷曲线分析方法流程图

Fig. 4 User load curve analysis method flowchart

本文以 UCI Electricity Load Diagrams 2011 年—2014 年数据集^[24]作为算例进行分析。该数据集包括 370 个用户的用电数据, 采集时间从 2011 年至 2014 年, 每隔 15 min 采样一次, 即每天 96 个采样点, 一共 140 256 个样本点。考虑到不同用户的数据开始采集时间不同, 本文取 2012 年数据进行分析, 剔除在 2012 年没有完整记录的用户, 剩余 317 个用户。

3.1 单一用户的典型负荷曲线提取

选取 317 个用户中的一个用户 MT_166 为例, 提取该用户的典型用电行为。图 5 是用户 MT_166

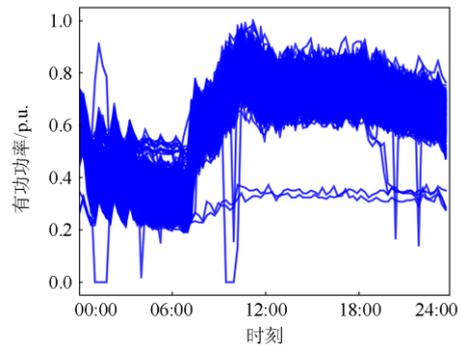


图 5 用户 MT_166 所有日负荷曲线

Fig. 5 All daily load curves of user MT_166

一年所有的日负荷曲线图, 由图 5 可知该用户的负荷曲线存在异常值及异常的负荷曲线形态。

对于用户 MT_166 的所有日负荷曲线, 先通过 PAA 对日负荷曲线降维, 再用 SAX 方法对每条负荷曲线进行转化, 符号化表示序列表示该用户的所有日负荷曲线。这里选择 PAA 分段数 $w=4$, SAX 参数字母集 $\alpha=3$ 。表 3 是用户 MT_166 所有日负荷曲线的符号化表示序列的结果。

由表 3 可知, 用户 MT_166 出现次数最多的负荷曲线符号化序列为 abcc, 其中符号化序列 abcc 所表示的负荷曲线如图 6 所示。对比图 5 和图 6 已剔除了该用户的异常负荷曲线, 再由式(3)得到该用户的典型用电行为, 并作为该用户的典型负荷曲线。

表 3 用户 MT_166 所有负荷曲线的符号化表示序列结果

Table 3 Symbolic representation results of all load curves of user MT_166

SAX 字符	个数	SAX 字符	个数
abcc	181	bbcc	3
abcb	88	aaaa	2
abbb	66	aabb	2
accc	12	abca	2
bccc	9	aacb	1

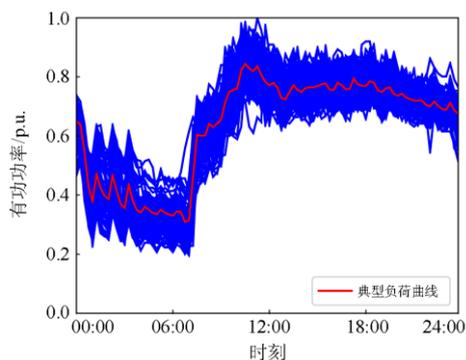


图 6 abcc 对应的负荷曲线

Fig. 6 Daily load curve corresponding to abcc

3.2 不同用户的典型负荷曲线的聚类分析

运用 3.1 节的方法, 可以提取所有 317 个用户的典型负荷曲线, 如图 7 所示。

基于 2.1 节所提出的负荷曲线特征, 提取所有用户典型负荷曲线特征, 再基于 k -means 算法对所有用户进行聚类分析。采用“肘形法”来确定 k -means 的最佳聚类数目, 令 k 值从小到大逐渐增大, 观察 SSE 的变化, 并绘制 k -SSE 折线图, 如图 8 所示。

观察图 8 k -SSE 折线图可知, 选择聚类数 $k=4$ 较为合适。首先选择以下两种方法作为两种方法作

为本文的对比方法: 对比方法 1 为直接基于原始负荷数据聚类, 对比方法 2 为基于常规的负荷特征聚类。不同方法的聚类结果如图 9—图 11 所示。

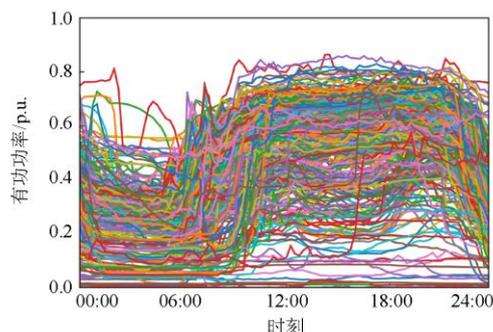


图 7 所有用户的典型负荷曲线

Fig. 7 Typical daily load curve of all users

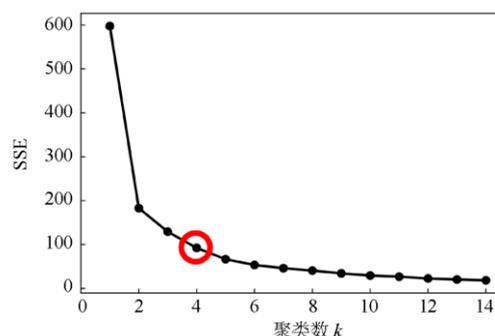


图 8 k -SSE 折线图

Fig. 8 k -SSE line chart

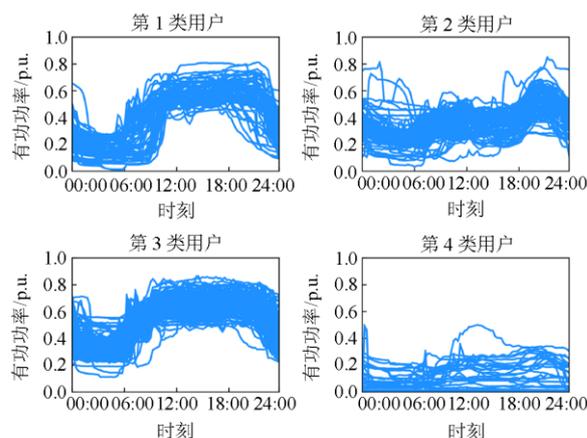


图 9 本文方法的聚类结果

Fig. 9 Clustering results of proposed method

本文方法的聚类结果如图 9 所示, 可以清楚地看出各类用户负荷曲线的特点: 第 1 类和第 3 类用户的用电行为较为类似, 都属于单峰型用户, 不同之处在于, 第 1 类用户用电高峰期较长, 主要集中在 08:00—21:00, 而第 3 类用户的用电高峰期主要

集中在 11:00—20:00, 本文所提方法的聚类结果可较好地地进行区分; 第 2 类用户为双峰型, 用电高峰期主要集中在 09:00—12:00 和 18:00—21:00; 第 4 类用户负荷波动小, 负荷曲线平坦。对比方法 1 的聚类结果如图 10 所示, 其聚类结果与本文方法相似, 但是第 1 类和第 3 类用户的分类效果不佳, 且第 2 类用户聚类结果较差。对比方法 2 的聚类结果如图 11 所示, 其聚类质量最差, 不能明显地区分出 4 类用户, 无法反映各类用户的用电特点。

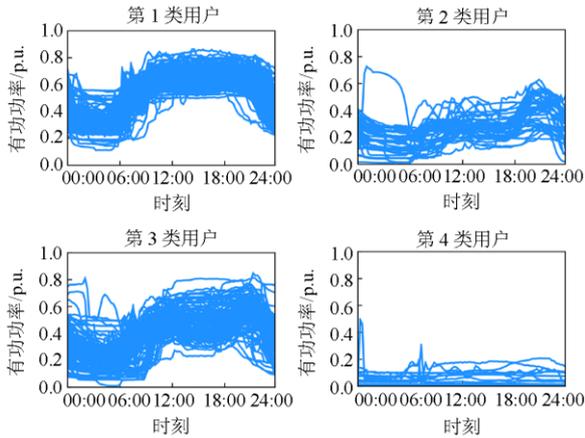


图 10 对比方法 1 的聚类结果
Fig. 10 Clustering results of contrast method 1

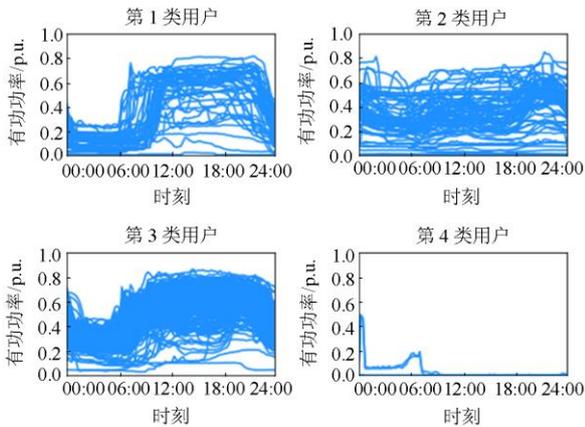


图 11 对比方法 2 的聚类结果
Fig. 11 Clustering results of contrast method 2

因此, 综合观察图 9—图 11 聚类分析结果可知: 直接基于原始负荷数据进行聚类分析, 能够大致区分出不同用户负荷曲线之间的差别, 但是无法精确地对每个用户分类; 基于常规的负荷特征进行聚类分析, 聚类质量不好, 说明常规的负荷特征提取丢失了较多负荷曲线的细节信息, 不能有效反映负荷曲线形态的差异; 本文所提方法的聚类结果和分类效果较优于其他两种方法, 能够反映出负荷曲

线形态差异, 可以准确地地区分出每个用户的用电特点, 聚类质量较好。

下面比较本文所提方法与传统数据驱动特征降维方法的效果。选择两种常用的数据降维方法对原始负荷数据降维, 分别是主成分分析 (Principal Component Analysis, PCA) 和局部线性嵌入 (Locally Linear Embedding, LLE), 再基于约简特征后的数据进行聚类分析, 聚类结果如图 12 和图 13 所示。其中, PCA 算法设置提取主成分方差贡献率达到 98% 的主成分, 最终算法确定主成分个数为 8; LLE 算法的参数设置为: 最邻近个数取 30, 低维特征空间的维数取 8。

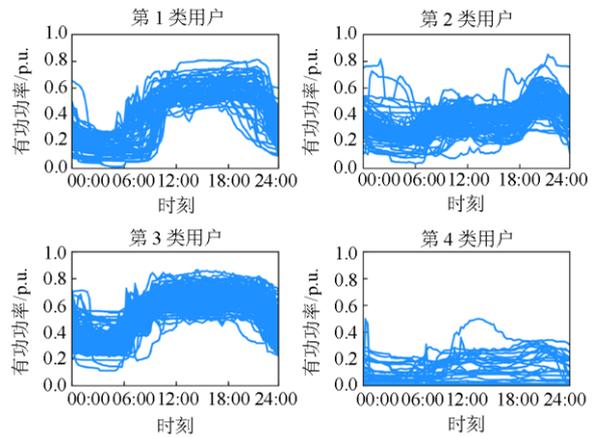


图 12 PCA+k-means 的聚类结果
Fig. 12 Clustering results of PCA+k-means algorithm

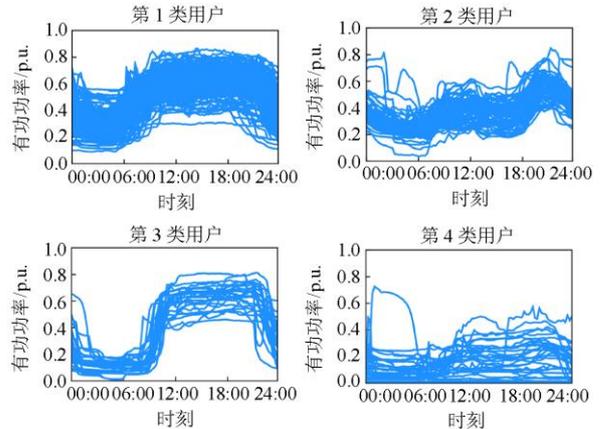


图 13 LLE k-means 的聚类结果
Fig. 13 Clustering results of LLE+k-means algorithm

观察图 12 和图 13 可知, PCA+k-means 与本文方法的聚类结果相似, 聚类质量较好; 而 LLE+k-means 方法的聚类质量稍差, 对于第 4 类用户的分类效果不佳。但是数据驱动特征降维的方法不能反映每个特征维度的物理含义, 本文所提方法

的特征提取结合时间序列分析方法和基于负荷数据曲线形态特征, 每个特征都有明确的含义, 更能反映负荷数据的形态特征。

对上述不同方法的聚类效果进行数值上的比较, 表 4 对比了不同方法聚类结果中各类用户的数目, 表 5 基于聚类评价指标对比了不同方法的聚类效果。

表 4 不同方法的聚类用户数目比较

Table 4 Comparing the number of clustering curves of different algorithms

方法	第 1 类用户	第 2 类用户	第 3 类用户	第 4 类用户
本文方法	27	80	132	78
对比方法 1	27	86	132	72
对比方法 2	73	199	42	3
PCA+k-means	23	82	137	75
LLE+k-means	33	74	184	26

表 5 实验方法的性能比较

Table 5 Performance comparison of experiment methods

方法	SIL	CH
本文方法	0.416	247.2
对比方法 1	0.303	187.5
对比方法 2	0.326	221.9
PCA+k-means	0.407	244.5
LLE+k-means	0.344	190.3

由表 4 可知, 本文方法与对比方法 1 和 PCA+k-means 方法的不同类型用户的聚类用户数目相近, 其他两种方法则相差较大。由表 5 可知本文方法的 SIL 和 CH 指标均高于其他 4 种方法, 说明本文方法的分类效果较好, 提取的负荷曲线特征能够提高聚类质量, 本文方法可行。

4 结论

本文通过引入时间序列数据挖掘中的符号聚合近似分析方法, 提出了一种用户典型负荷曲线提取方法, 该方法能够较好地提取用户的用电行为, 有效地反映用户负荷曲线的整体趋势和消除噪声。然后结合不同用户典型负荷曲线的负荷特性指标特征和时间序列特征, 基于 k -means 聚类算法对不同用户的典型负荷曲线进行聚类分析。算例表明本文所提方法应用于负荷曲线提取与聚类分析可大大提升分析效率与聚类质量, 验证了方法的可行性与有效性。

研究不同用电用户的负荷曲线特征, 挖掘用户的用电规律, 可以帮助电网公司更深入地了解不同用户的用电特性差异, 为电网公司制定供电策略、优化供电设施规划建设提供指引。随着电力大数据

技术广泛使用和需求侧响应的推广, 下一步将重点研究如何考虑需求侧响应对用户负荷曲线的影响以及如何将本文方法运用于实际系统中。

参考文献

- [1] 赵俊华, 董朝阳, 文福拴, 等. 面向能源系统的数据科学: 理论、技术与展望[J]. 电力系统自动化, 2017, 41(4): 1-11, 19.
ZHAO Junhua, DONG Zhaoyang, WEN Fushuan, et al. Data science for energy systems: theory, techniques and prospect[J]. Automation of Electric Power Systems, 2017, 41(4): 1-11, 19.
- [2] WANG Y, CHEN Q, HONG T, et al. Review of smart meter data analytics: applications, methodologies, and challenges[J]. IEEE Transactions on Smart Grid, 2018, 10(3): 3125-3148.
- [3] 王毅, 张宁, 康重庆, 等. 电力用户行为模型: 基本概念与研究框架[J]. 电工技术学报, 2019, 34(10): 76-88.
WANG Yi, ZHANG Ning, KANG Chongqing, et al. Electrical consumer behavior model: basic concept and research framework[J]. Transactions of China Electrotechnical Society, 2019, 34(10): 76-88.
- [4] 刘季昂, 刘友波, 程明畅, 等. 基于分位数半径动态 K-means 的分布式负荷聚类算法[J]. 电力系统保护与控制, 2019, 47(24): 15-22.
LIU Ji'ang, LIU Youbo, CHENG Mingchang, et al. A distributed load clustering algorithm based on quantile radius dynamic K-means[J]. Power System Protection and Control, 2019, 47(24): 15-22.
- [5] 彭显刚, 赖家文, 陈奕. 基于聚类分析的客户用电模式智能识别方法[J]. 电力系统保护与控制, 2014, 42(19): 68-73.
PENG Xiangang, LAI Jiawen, CHEN Yi. Application of clustering analysis in typical power consumption profile analysis[J]. Power System Protection and Control, 2014, 42(19): 68-73.
- [6] 席雅雯, 吴俊勇, 石琛, 等. 融合历史数据和实时影响因素的精细化负荷预测[J]. 电力系统保护与控制, 2019, 47(1): 80-87.
XI Yawen, WU Junyong, SHI Chen, et al. A refined load forecasting based on historical data and real-time influencing factors[J]. Power System Protection and Control, 2019, 47(1): 80-87.
- [7] 吴亚雄, 高崇, 曹华珍, 等. 基于灰狼优化聚类算法的日负荷曲线聚类分析[J]. 电力系统保护与控制, 2020, 48(6): 68-76.
WU Yaxiong, GAO Chong, CAO Huazhen, et al. Clustering analysis of daily load curves based on GWO

- algorithm[J]. Power System Protection and Control, 2020, 48(6): 68-76.
- [8] 徐邦恩, 蔺红. 基于改进模糊聚类的典型日负荷曲线选取方法[J]. 电测与仪表, 2019, 56(4): 21-26.
XU Bang'en, LIN Hong. Selection method of typical daily load curve based on improved fuzzy clustering[J]. Electrical Measurement & Instrumentation, 2019, 56(4): 21-26.
- [9] 赵岩, 李磊, 刘俊勇, 等. 上海电网需求侧负荷模式的组合识别模型[J]. 电网技术, 2010, 34(1): 145-151.
ZHAO Yan, LI Lei, LIU Junyong, et al. Combinational recognition model for demand side load profile in Shanghai Power Grid[J]. Power System Technology, 2010, 34(1): 145-151.
- [10] 孟令奎, 段红伟, 黄长青, 等. 一种基于语义聚类的典型日负荷曲线选取方法[J]. 华北电力大学学报: 自然科学版, 2013, 40(1): 43-48.
MENG Lingkui, DUAN Hongwei, HUANG Changqing, et al. A semantic clustering method for selecting the typical day load curve[J]. Journal of North China Electric Power University: Natural Science Edition, 2013, 40(1): 43-48.
- [11] LI Huang, YANG Yongbiao, ZHAO Honglei, et al. Time series modeling and filtering method of electric power load stochastic noise[J]. Protection and Control of Modern Power Systems, 2017, 2(3): 269-275. DOI: 10.1186/s41601-017-0059-8.
- [12] 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.
- [13] 罗滇生, 杜乾, 别少勇, 等. 基于负荷分解的居民差异化用电行为特性分析[J]. 电力系统保护与控制, 2016, 44(21): 29-33.
LUO Diansheng, DU Qian, BIE Shaoyong, et al. Analysis of differentiation residential electricity consumption characteristic based on power load decomposition[J]. Power System Protection and Control, 2016, 44(21): 29-33.
- [14] 陆俊, 朱炎平, 彭文昊, 等. 智能用电用户行为分析特征优选策略[J]. 电力系统自动化, 2017, 41(5): 58-63, 83.
LU Jun, ZHU Yanping, PENG Wenhao, et al. Feature selection strategy for electricity consumption behavior analysis in smart grid[J]. Automation of Electric Power Systems, 2017, 41(5): 58-63, 83.
- [15] 孙友强. 时间序列数据挖掘中的维数约简与预测方法研究[D]. 合肥: 中国科学技术大学, 2014.
- SUN Youqiang. Research on dimensionality reduction and prediction methods in time series data mining[D]. Hefei: University of Science and Technology of China, 2014.
- [16] 宋伟, 张帆, 叶阳东, 等. 基于 SAX 方法的时间序列分类问题的多阶段改进研究[J]. 计算机工程与科学, 2016, 38(5): 988-996.
SONG Wei, ZHANG Fan, YE Yangdong, et al. A multi-phased improvement for time series classification based on symbolic aggregation approximation representation[J]. Computer Engineering & Science, 2016, 38(5): 988-996.
- [17] 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.
- [18] ROUSSEEUW P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational Application Math, 1987, 20: 53-65.
- [19] CHRIST M, KEMPA-LIEHR A W, FEINDT M. Distributed and parallel time series feature extraction for industrial big data applications[EB/OL]. [2017-05-19]. <https://arxiv.org/abs/1610.07717v3>.
- [20] GUSTAVO E, BATISTA A, KEOGH E J, et al. CID: an efficient complexity-invariant distance for time series[J]. Data Mining and Knowledge Discovery, 2014, 28(3): 634-669.
- [21] SCHREIBER T, SCHMITZ A. Discrimination power of measures for nonlinearity in a time series[J]. Physical Review E, 1997, 55(5): 5443.
- [22] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
- [23] 张宪超. 数据聚类[M]. 北京: 科学出版社, 2017.
Electricity load diagrams 2011-2014 data set[EB/OL]. [2011-02-14]. https://archive.ics.uci.edu/ml/datasets/Electricity_LoadDiagrams20112014#.

收稿日期: 2020-04-17; 修回日期: 2020-06-23

作者简介:

唐俊熙(1989—), 男, 工学硕士, 工程师, 从事电力系统规划与运行研究; E-mail: tangjunxi89@163.com

曹华珍(1974—), 女, 工学硕士, 高级工程师, 从事电力系统规划与运行研究; E-mail: 729875469@qq.com

高崇(1983—), 男, 工学硕士, 高级工程师, 从事电力系统规划与运行研究。E-mail: 25903257@qq.com

(编辑 周金梅)