基于 K-means++算法的日负荷曲线聚类分析

李 勇[1],王 华[1],雷雯婷[1],汪泉霖[1],刘渝凯[1],吴含欣[2],李婵娓[1]
(1. 国网四川省电力公司天府新区供电公司, 四川 成都 610218;
2. 浙江大学电气工程学院,浙江 杭州 310027)

摘 要:K-means 聚类算法因计算过程简单、收敛速度快,被广泛应用于负荷特性分析。然而,K-means 聚类算法的聚类数难以选择,且随机的初始化质心选择易导致收敛速度慢和陷入局部最优的问题。因此,提出一种基于 K-means++算法的日负荷曲线聚类分析方法,利用启发式随机播种方式选取初始质心,基于肘部法则,利用聚类评价指标度量聚类的密集度和分离度,综合评定确定最佳聚类数,避免初始质心的随机性影响聚类质量。算例表明,所提方法在聚类质量方面表现出较高水平,能够提供较为合理的负荷分类,有助于掌握用户的负荷特性。

关键词:K-means++; 聚类分析; 日负荷曲线
中图分类号:TM 714 **文献标志码**:A **文章编号**:1003-6954(2024)06-0056-05
DOI:10.16527/j.issn.1003-6954.20240608

Cluster Analysis of Daily Load Curves Based on K-means++ Algorithm

LI Yong[1], WANG Hua[1], LEI Wenting[1], WANG Quanlin[1], LIU Yukai[1], WU Hanxin[2], LI Chanxiao[1]
(1. State Grid Tianfu New Area Electric Power Supply Company, Chengdu 610218, Sichuan, China;
2. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, China)

Abstract: The K-means cluster algorithm is widely used for the analysis of load characteristics due to its simple calculation process and fast convergence rate. However, selecting the number of clusters in K-means algorithm is challenging, and the random initialization of cluster centers can lead to slow convergence rate and local optima issues. A cluster analysis method for daily load curves based on K-means++ algorithm is proposed. It selects the initial cluster centers using heuristic random seeding method, combines the elbow method and clustering evaluation metrics to measure the density and separation of clusters and comprehensively determines the optimal number of clusters, which mitigates the impact of randomness of initial cluster centers on clustering quality. The case studies show that the proposed algorithm can achieve high clustering quality, provide a reasonable user classification and facilitate a comprehensive understanding of load characteristics.

Key words: K-means++; cluster analysis; daily load curve

0 引 言

近年来,电力需求快速增长,电动汽车等多元化负荷大量投入,局部区域出现电力负荷紧张的问题[1]。深入了解全网不同时空条件下的负荷特性,挖掘负来趋势,分析数据中潜藏着的用户用电行为习惯,对荷数据之间的内在联系,把握负荷变化的规律和未于合理利用电力数据十分必要[2]。探索有效的数据挖掘算法,对用户进行细分,针对不同用电类别的用户快速准确地挖掘出用电行为、电量消费等大量有价值的信息,有助于电力公司实现对电网整体态势的宏观、实时、精确把握,提高对电网的整体认知,为电网的下一步优化方案提供依据,在电网规划中具有重要意义[3]。

通过数据挖掘中的聚类技术对电力用户负荷曲线进行聚类,是负荷特性分析的重要手段之一[4-5],常用的有基于划分的聚类算法[6-7]、基于层次的聚类算法[8]、基于密度的聚类算法[9-10]、基于模型的

基金项目:国网四川省电力公司科技项目"配电网负荷虚拟聚合及预测关键技术研究"(512920230001)

聚类算法[11]和基于谱图的聚类算法[12-13]等。其中基于划分的 K-means 聚类算法计算过程简单,收敛速度快[14],因而应用最为广泛。

然而,K-means 聚类算法的聚类数量难以选择,且随机的初始化质心选择易导致收敛速度慢和陷入局部最优的问题。针对上述问题,文献[15]采用 canopy 算法的输出结果作为 K-means 算法的输入,不需要人为规定聚类数,但是确定 canopy 算法中 T1、T2 这两个参数成为新问题。文献[16]提出基于 MapReduce 模型的改进 K-means 算法,根据簇内分散程度小、簇间距离大的原则确定聚类数,选择密度最大的 k 个对象作为初始质心,避免了初始质心的随机性使结果陷入局部最优,但这种方法计算复杂度高,对异常值敏感,异常值会对密度计算产生显著影响。文献[6]提出使用 MaxMin 原则选取初始质心,简单易操作,但是这种方法基于样本的实际分布,对于分布不规则的数据集的计算效率较低。

下面提出一种基于 K-means++算法的日负荷曲线聚类分析方法,利用启发式随机播种方式选取初始质心,基于肘部法则,利用聚类评价指标度量聚类的密集度和分离度,综合评定确定最佳聚类数,避免初始质心的随机性影响聚类质量。

1　基于启发式随机播种的 K-means++算法

1.1　K-means 算法

K-means 算法[13]是一种基于划分的迭代式聚类算法,旨在将 n 个数据点划分为 k 个簇,按照相似度将每个数据点分配到离其最近的质心所对应的簇中,使得簇内数据点之间的平方误差准则函数稳定在最小值。通常采用欧氏距离作为样本相似程度的评价指标对样本进行划分,通过迭代计算更新每个簇的质心为该簇所有数据点的平均值,直至质心位置不再发生变化或达到预定的迭代次数。

平方误差准则函数即误差平方和函数(sum of squared errors,SSE),公式为:

$$\min J = \sum_{i=1}^{k}\sum_{m=1}^{N} d_{mi}\|\boldsymbol{x}_m - \boldsymbol{c}_i\|^2 \tag{1}$$

$$d_{mi} = \begin{cases} 1, \boldsymbol{x}_m \in \boldsymbol{R}_i \\ 0, \boldsymbol{x}_m \notin \boldsymbol{R}_i \end{cases} \tag{2}$$

式中:J 为误差准则函数;k 为聚类簇数;$\boldsymbol{R}_i$ 为第 i 类;c_i 为 $\boldsymbol{R}_i$ 的质心;m 为样本编号;N 为样本数;x_m 为第 m 个样本,即待聚类日的相关因素构成的向量;d_{mi} 表示第 m 个样本是否属于 $\boldsymbol{R}_i$。

1.2　K-means++ 算法

一般而言,K-means 聚类的初始质心是随机选择的,且算法的结果对初始值敏感,初始化质心的选择对最后的聚类结果和运行时间有较大的影响,完全随机的选择有可能导致算法收敛很慢或是陷入局部最优。而 K-means++算法可以改善初始质心的随机性。

K-means++ 算法使用启发式随机播种方法找到 K-means 聚类的质心种子。K-means++ 算法的基本原则和操作方式是为每个样本分配不同的概率,使得距离现有质心较远的点更有可能被选为初始质心,以确保初始质心之间的距离尽可能远。

假设簇数为 k, K-means++ 算法按如下方式选择质心:

1)初始化第一个质心:随机均匀选择一个数据样本作为第一个质心。

2)选择其他质心:对于数据集中的第 m 个样本 $\boldsymbol{x}_m$,计算其与已选定的 i 类质心 c_i 的最短欧式距离 $d(\boldsymbol{x}_m, \boldsymbol{c}_i)$。

$$d(\boldsymbol{x}_m, \boldsymbol{c}_i) = \sqrt{\|\boldsymbol{x}_m - \boldsymbol{c}_i\|^2} \tag{3}$$

将每个样本按距离分配给其最近的质心。样本 x_m 被选择为下一个质心的概率 P_m 为

$$P_m = \frac{d^2(\boldsymbol{x}_m, \boldsymbol{c}_i)}{\sum_{\{h;\boldsymbol{x}_h \in \boldsymbol{R}_i\}} d^2(\boldsymbol{x}_h, \boldsymbol{c}_i)} \tag{4}$$

式中,$\boldsymbol{x}_h$ 为 $\boldsymbol{R}_i$ 中第 h 个样本。

也就是说,选择每个后续中心时,每个样本被选中的概率与它到已选最近中心的距离成比例,也即距离较远的数据点更有可能成为下一个质心,以确保簇间距离尽可能远。

3) 重复步骤 2,直到选择出预先设定的 k 个质心为止。

这种初始化方法可以使得初始的质心分布更加广泛,有助于避免 K-means 算法陷入局部最优解的问题,从而提高了聚类结果的稳定性和准确性。文献[17] 通过对几个簇方向的模拟研究证明,在计算簇内点到质心距离平方和时,K-means++的表现一

直优于 K-means,K-means++ 相比 K-means 算法能更快地收敛至更低的总和,运行时间也更快。在几个真实的数据集上进行了初步实验,观察到 K-means++ 在速度和准确性方面都大大优于标准 K-means。

2 基于 K-means++的日负荷曲线聚类

2.1 数据预处理

2.1.1 数据收集

采集数据获得由 n 条日负荷曲线构成的初始负荷曲线集合 $X_0=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,其维度为 $n\times q$,其中 n 为曲线的数量,q 为每条曲线的数据点数量。常用实际采样频率为 15 min,对应 q 为 96。

2.1.2 数据清洗

首先核查 X 中每条负荷曲线的负荷数据,查找缺失数据、负荷突变或负值等异常数据。若某条负荷曲线的数据缺失和异常达到一定百分比,该曲线将被视为无效曲线并被剔除。剔除无效样本后的有效样本的总数记录为 N。

设第 m 条负荷曲线 t 时刻采集的负荷数据 $x_{m,t}$ 为异常数据,则采用平滑修正公式修正异常数据,修正后的数据值 $x'_{m,t}$ 为

$$x'_{m,t}=\left(\sum_{f=1}^{f_1}x_{m,t-f}+\sum_{b=1}^{b_1}x_{m,t+b}\right)/(f_1+b_1) \tag{5}$$

式中:f、b 分别为向前和向后采集数据点;f_1、b_1 分别为向前、向后采集的步长,根据具体要求确定,一般取 5~10。

2.1.3 数据归一化

采用 max-min 归一化方法对负荷数据进行处理,具有同一特征的不同数量级的负荷分到一类,消除量纲的影响,增加可信度,公式为

$$x''_{m,t}=\frac{x_{m,t}-\min\limits_{1\leqslant j\leqslant q}x_{m,j}}{\max\limits_{1\leqslant j\leqslant q}x_{m,j}-\min\limits_{1\leqslant j\leqslant q}x_{m,j}} \tag{6}$$

式中,$x''_{m,t}$ 为归一化后第 m 个样本在第 t 时刻的负荷值。可见,归一化后各样本负荷值分布在[0,1]之间。

经过数据预处理,得到日负荷曲线数据集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$。

2.2 最佳聚类数的确定

设置不同的聚类数得到的聚类结果会有所差异。基于肘部法则,利用聚类评价指标度量聚类的密集度和分离度,综合评定确定最佳聚类数。

肘部法则是一种在聚类分析中用于确定最佳聚类数的常用启发式方法,通过计算不同聚类数下的误差平方和 SSE 来帮助确定最佳的聚类数。绘制聚类数与对应的 SSE 的图表,最佳的聚类数定义为图表中出现“肘部”处的点,即聚类数对性能指标的影响逐渐减弱的拐点处。肘点意味着增加类别无法带来更多回报,即认为此时是最佳的聚类数。

肘部法则简单直观,被广泛应用于聚类分析中,但是某些情况下肘点不清晰,不易直接观察得到而存在主观性。因此,选用聚类评价指标戴维森堡丁指数(Davies-Bouldin index,DBI)度量聚类的簇内密集度和簇间分离度,综合评定聚类数目。

DBI 通过计算簇内距离和簇间距离的比值来度量聚类的紧密度和分离度,计算公式为

$$I_{\mathrm{DBI}}=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(\boldsymbol{c}_i,\boldsymbol{c}_j)}\right) \tag{7}$$

式中:σ_i、σ_j 分别为第 i 类、第 j 类中的元素到相应类的质心的平均距离;$d(\boldsymbol{c}_i,\boldsymbol{c}_j)$ 为第 i 类到第 j 类质心的欧氏距离。DBI 越小意味着类内距离越小,同时类间距离越大,表示聚类结果同类别内部紧密,不同类别分离较远,聚类效果越好。

2.3 聚类流程

基于 K-means++的日负荷曲线聚类分析,算法整体流程如图 1 所示。

3 算例分析

3.1 聚类分析

算例数据来源于某电网某日 10 kV 变压器共 14 826 条负荷曲线,采集频率为 15 min ,共计 96 个量测点。经数据清洗后,本算例共含 14 321 条有效日负荷曲线。

构造算例数据集之后,按照所提方法进行聚类分析。基于肘点法则,选用 DBI 指标度量聚类效果,结果如表 1 和图 2 所示。从图 2 能看出,误差平方和 SSE 随聚类数变化较为平滑,肘点不显著。同时,DBI 在 $k=6$ 时取到极小值,此时 SSE 变化也趋于平缓,因此认为 $k=6$ 是最佳聚类数。

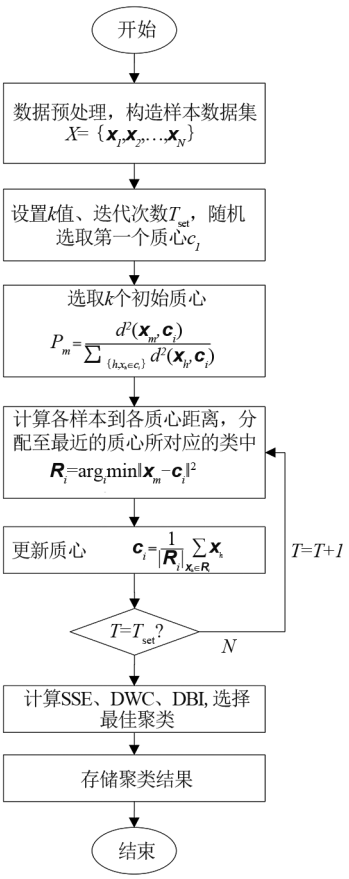

图 1　聚类算法流程

表 1　不同聚类数的聚类效果评价

k	SSE/10^4	DBI	计算时间/s	k	SSE/10^4	DBI	计算时间/s
1	9.27	—	0.02	6	4.15	1.54	0.30
2	7.42	1.83	0.19	7	4.03	1.82	0.40
3	6.56	1.72	0.27	8	3.91	1.94	0.41
4	5.42	1.64	0.24	9	3.80	2.05	1.24
5	4.64	1.56	0.29	10	3.72	2.18	0.53

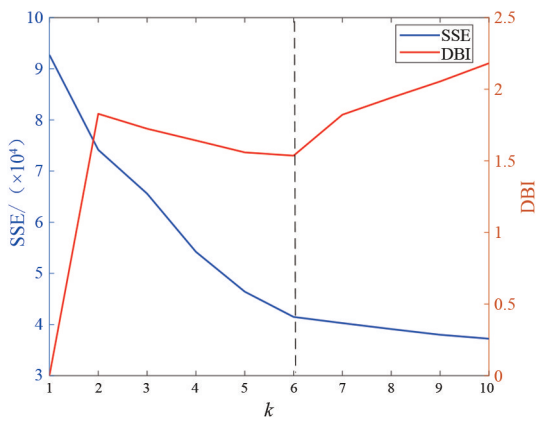

图 2　不同聚类数下 SSE、DBI 值

取 $k=6$，对数据集样本进行聚类，得到结果如图 3 所示。

在图 3 中，观察到 6 种不同形态的负荷曲线展现出明显的负荷特征，表现出良好的聚类效果。第 1 类负荷曲线呈现双峰型，在 10:00 到 22:00 期间维持较高负荷，午间略有回落，主要代表商业负荷；第 2 类负荷表现出稳定的趋势和较高的负荷，主要为大型工业负荷，夜间负荷略有回升，可能存在利用低谷价夜间生产的行为；第 3 类负荷主要表现为晚高峰负荷，8:00 到 12:00 有小高峰，主要代表居民负荷；第 4 类负荷表现出稳定的趋势和较低的负荷，主要为小型工业负荷；第 5 类负荷在 9:00 到 17:00 期间维持较高负荷，午间略有回落，主要为非工业负荷；第 6 类负荷主要为夜间用电，代表着城市行政负

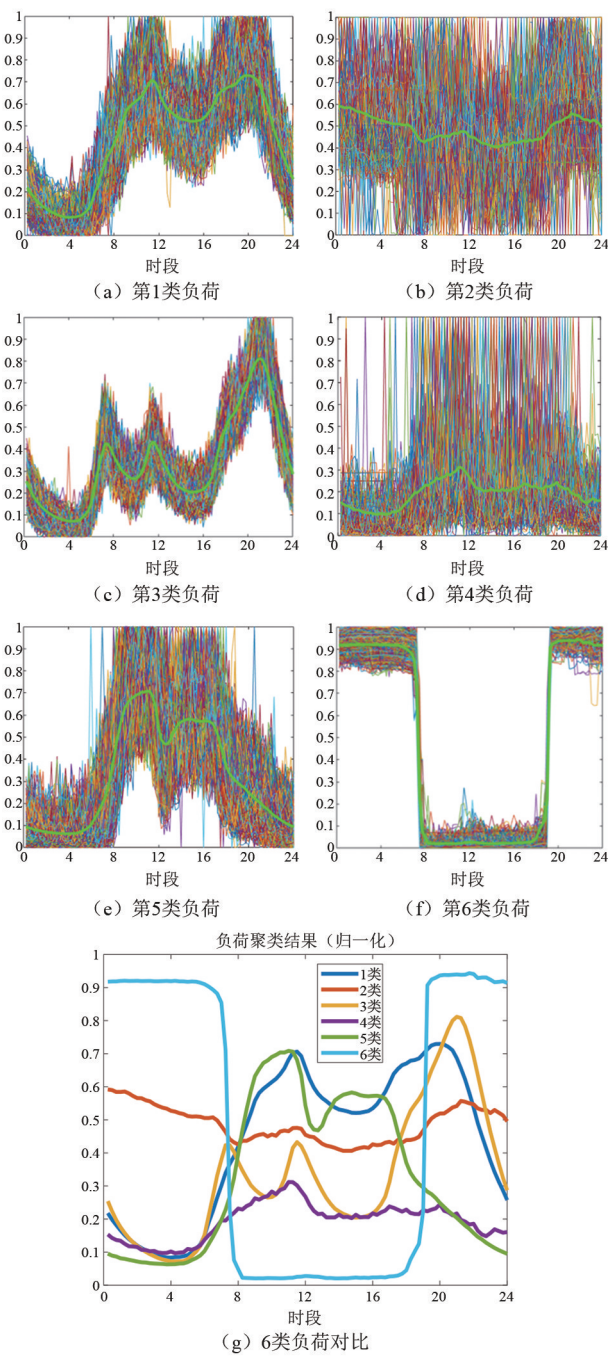

图 3　日负荷曲线聚类结果

荷,例如路灯照明等。

为验证所提算法的合理性和有效性,选取传统 K-means 聚类算法与所提算法进行对比,结果如表 2 所示,通过聚类评价指标和计算时间综合评估两种算法的性能表现。

表 2 不同聚类算法的性能对比

算法	SSE	DBI	计算时间/s
K-means	41 482.90	1.538 3	4.373 7
K-means++	41 482.85	1.536 6	0.577 1

由表 2 可见,在同时设置聚类数为 6 的情况下,使用所提聚类算法进行聚类的运行时间仅为传统 K-means 聚类算法聚类结果的 1/7 左右,同时所提算法 SSE、DBI 指数均小于传统算法,聚类质量占优势。这主要是因为 K-means++算法基于启发式随机播种的初始化方法使得初始的质心分布更加广泛,有助于避免算法陷入局部最优解的问题,算法能更快地收敛至更低的总和,运行时间也更快。综上,所提降维聚类算法在速度和聚类质量方面都优于传统 K-means,能够提供较为合理的负荷分类,有助于充分掌握用户的负荷特性,具有实际应用价值。

3.2 算法鲁棒性分析

为检验所提方法的鲁棒性,在上述算例的基础上,引入均匀分布扰动,其比例为 $\pm r$($r=5\%,10\%,\cdots,40\%$)。这一步骤是模拟在实际采样过程中用户负荷曲线因随机因素造成的日内负荷波动,从而评估所提方法对于在真实环境中可能存在的日内负荷波动的适应能力。

为说明不同程度的扰动对聚类结果的影响,利用调整兰德系数(adjusted Rand index,ARI)量化不同干扰下的聚类结果与无干扰下聚类结果的一致性。调整兰德系数的取值范围为[-1,1],其中 -1 表示两个结果完全不同,0 表示两种结果与随机的相似性相当,1 表示两种结果完全相同。ARI 的计算公式为

$$I_{\mathrm{ARI}}=\frac{I_{\mathrm{RI}}-E(I_{\mathrm{RI}})}{\max I_{\mathrm{RI}}-E(I_{\mathrm{RI}})} \tag{8}$$

式中:I_{RI}为兰德系数,其值等于受干扰下和无干扰下聚类结果一致的样本数量和样本总数的比值,易知,I_{RI}取值范围是[0,1];$E(\cdot)$为变量的数学期望。

ARI 越大,代表受干扰下和无干扰下的聚类结果一致性越高,则说明鲁棒性越好。在不同干扰下进行聚类,对聚类结果进行一致性检验,得到表 3。

表 3 不同扰动下的聚类一致性检验

r/%	SSE/10^4	DBI	ARI	r/%	SSE/10^4	DBI	ARI
5	4.15	1.536 1	0.989 6	25	4.52	1.536 8	0.943 7
10	4.17	1.537 9	0.980 6	30	4.74	1.537 4	0.935 6
15	4.24	1.534 8	0.969 0	35	5.00	1.538 6	0.925 3
20	4.36	1.535 5	0.953 8	40	5.51	1.656 6	0.763 9

从表 3 可以看出,当叠加干扰比例小于 10%时,聚类结果与未受干扰时聚类结果一致性非常高,即所提算法在小干扰情况下保持非常好的稳定性和准确性。当信噪比达到 40%时,所提算法的精度出现大波动,聚类准确率会明显下降。整体来看,所提算法具备一定的抗干扰能力,算法鲁棒性较好。

4 结 论

上面提出了一种基于 K-means++算法的日负荷曲线聚类分析方法,利用启发式随机播种方式选取初始质心,基于肘部法则,利用聚类评价指标度量聚类的密集度和分离度,综合评定确定最佳聚类数,避免初始质心的随机性影响聚类质量。以某地区某日负荷数据为例,选取聚类数目为 6,对日负荷曲线进行聚类分析。聚类结果显示,所得 6 类负荷曲线形态差异显著,展现出明显的负荷特征。该算法聚类质量高,所提算法在聚类质量方面表现出较高水平,能够提供较为合理的负荷分类,有助于充分掌握用户的负荷特性,具有实际应用价值。

参考文献

[1] 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):2056-2068.

[2] 朱栋.典型负荷用电行为模式分析方法及其应用研究[D].南京:东南大学,2017.

[3] 彭大健,裴玮,肖浩,等.数据驱动的用户需求响应行为建模与应用[J].电网技术,2021,45(7):2577-2586.

[4] SI C M Z,XU S L,WAN C,et al. Electric load clustering in smart grid: Methodologies, applications, and future trends[J]. Journal of Modern Power Systems and Clean Energy,2021,9(2):237-252.

[5] 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.

(下转第 107 页)

[8] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision(ICCV), October 22-29, 2017, Venice, Italy. IEEE, 2017: 2961-2969.

[9] ZHU X K, LYU S C, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]// 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), October 11 - 17, 2021, Montreal, BC, Canada. IEEE, 2021:2778-2788.

[10] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[EB/OL]. [2023-07-15]. http://doi.org/10.48550/arXiv.2004.10934.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30:5998-6008.

[12] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition(CUPR), June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016:779-788.

[13] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):12993-13000.

[14] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 09630.

[15] 王润民,桑农,丁丁,等.自然场景图像中的文本检测综述[J].自动化学报,2018,44(12):2113-2141.

作者简介:

陈　亮(1983),男,博士,高级工程师,研究方向为电力系统自动化;

高　杰(1989),男,本科,工程师,研究方向为电气工程及安全管理;

李　诚(1977),男,本科,高级工程师,研究方向电气工程及电力营销。

(收稿日期:2024-01-15)

(上接第 60 页)

[6] 董雷,陈振平,韩富佳,等.基于图卷积神经网络与 K-means 聚类的居民用户集群短期负荷预测[J].电网技术,2023,47(10):4291-4301.

[7] 何哲楠,吴浩,程祥,等. 基于变电站-用户双层结构的变电站负荷聚类研究[J].电网技术,2019,43(8):2983-2991.

[8] 郑乐,徐青山,冯小峰.基于层次聚类算法与 ISA-LSSVM 的短期负荷预测研究[J].电力需求侧管理,2022,24(5):51-57.

[9] 魏勇,李学军,李万伟,等.基于空间密度聚类和 K-shape 算法的城市综合体负荷模式聚类方法[J].电力系统保护与控制,2021,49(14):37-44.

[10] XIANG Y, HONG J H, YANG Z Y, et al. Slope-based shape cluster method for smart metering load profiles[J].IEEE Transactions on Smart Grid, 2020, 11(2):1809-1811.

[11] 许雅婧,黄小庆,曹一家,等.基于 SOM 神经网络聚类的空调负荷聚合方法[J].电力系统及其自动化学报,2015,27(11):26-33.

[12] 徐胜蓝,司曹明哲,万灿,等.考虑双尺度相似性的负荷曲线集成谱聚类算法[J].电力系统自动化,2020,44(22):152-160.

[13] 刘晓峰,康进,马翔,等.基于快速动态时间弯曲和最小覆盖球的多日负荷曲线聚类方法[J].电力自动化设备,2022,42(7):51-58.

[14] LLOYD S. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982,28(2): 129-137.

[15] 程艳柳.基于云计算的智能电网数据挖掘的研究[D].北京:华北电力大学,2013.

[16] 赵莉,候兴哲,胡君,等.基于改进 k-means 算法的海量智能用电数据分析[J].电网技术,2014,38(10):2715-2720.

[17] ARTHUR David, VASSILVITSKII Sergi. K-means++: The advantages of careful seeding [C]//Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'07), January 7-9, 2007, New Orleans, USA.2007:1027-1035.

作者简介:

李　勇(1986),男,博士,高级工程师,从事电力系统安全稳定运行工作;

王　华(1978),男,硕士,高级工程师,从事电力系统生产运行管理工作;

雷雯婷(1988),女,硕士,高级工程师,从事电力系统调度运行工作。

(收稿日期:2024-01-25)