DOI:10.12158/j.2096-3203.2022.03.022

# 考虑数据不均衡的居民用户负荷曲线分类方法

张慧波[1]，王守相[1]，赵倩宇[1]，任杰[2]，王海[2]
（1. 智能电网教育部重点实验室（天津大学），天津 300072；
2. 国网冀北张家口风光储输新能源有限公司，河北 张家口 075061）

**摘 要**：由于用户用电行为的多样性和随机性，负荷数据存在分布不均衡的问题，传统负荷曲线分类方法在处理不均衡数据时性能较差。为此，提出一种改进 *K*-means 与长短期记忆（LSTM）神经网络-卷积神经网络（CNN）分类模型结合的负荷曲线分类方法。首先，为提升 *K*-means 算法对不均衡数据的聚类效果，基于密度峰值聚类（DPC）算法思想，提出一种相对 *k* 近邻密度峰值（RKDP）初始聚类中心选取方法，将其作为 *K*-means 算法的初始中心进行聚类；然后，为提高 RKDP-*K*-means 处理高维负荷数据的性能，采用 LSTM 自编码器进行特征降维后再聚类获得精准类别标签；最后，基于 LSTM 神经网络和 CNN 分别提取负荷特征构建负荷曲线分类模型，实现对大规模负荷曲线的分类。算例选取了爱尔兰智能电表数据集和伦敦负荷数据集进行实验，验证了所提算法在大规模负荷曲线分类时的有效性和实用性。

**关键词**：负荷曲线分类；不均衡数据；改进 *K*-means；自编码器；长短期记忆（LSTM）神经网络；卷积神经网络（CNN）
**中图分类号**：TM73 **文献标志码**：A **文章编号**：2096-3203（2022）03-0186-08

## 0 引言

用户用电行为分析是电网分析规划的重要环节。随着智能采集装置的广泛应用，用户用电活动可通过智能电表采样并以负荷曲线等形式表现，其数据具有体量大、速度快、价值密度低等特征。针对用户负荷数据特点，研究高效的负荷曲线分类方法有助于电力公司从海量用电侧数据中挖掘用户潜在用电规律，对开展负荷预测、需求响应、电价决策等工作有着重要意义[1—2]。

目前，负荷曲线分类方法主要有无监督聚类、有监督分类以及无监督与有监督相结合等。近来年关于负荷曲线无监督聚类所做的研究主要集中于改进聚类算法[3—4]和改进聚类特征 2 个方面。在算法方面，文献[5]提出一种分段聚类方法对建筑负荷曲线分类，能够更高效地获取建筑的日典型用电模式。在聚类特征改进方面，主要聚焦在特征提取方法[6—8]和相似度度量计算方法[9—10]，文献[11]提出一种基于负荷曲线斜率分段的形状聚类方法，能够更好地捕捉曲线的形状特征；文献[12]采用样本皮尔逊相关系数距离作为相似度度量，算例表明优于欧几里得距离。在负荷有监督分类方面，应用最广泛的是反向传播神经网络（back propagation neural network，BPNN）[13—14]，但 BPNN 存在梯度爆炸、梯度消失等问题。在无监督与有监督结合方面，负荷数据作为无标签数据，利用无监督聚类获得类别标签，训练有监督学习模型进行分类，可将无监督与有监督的优势相结合，实现海量负荷数据的高效分类，其首先应获得训练集的精准类别标签[15—17]。

不均衡数据是指数据集中归属于某一类别的样本数量和密度与其他类别有较大差异。由于用户用电行为的随机性与多样性，负荷数据同样存在不均衡的现象，某些类别的负荷数量远少于其他类别的负荷数量。传统的 *K*-means 算法处理此类数据时容易出现“均匀效应”[18—19]，小类会吞噬大类中的部分样本，而传统分类方法同样在小样本类别上分类效果欠佳。目前在负荷曲线分类时考虑不均衡数据问题的研究较少，文献[20]改进密度峰值聚类（density peak clustering，DPC）算法实现了对多类别分布不均衡的负荷曲线聚类，但该算法计算复杂度较高，难以处理海量负荷数据；文献[21—23]利用过采样技术处理类别不平衡问题后训练分类模型，但其前提是训练集需要精准的类别标签，而负荷数据是无标签数据，难以获得准确的类别信息。

为了解决上述问题，提出一种无监督与有监督相结合的负荷曲线分类方法。首先，采用长短期记忆（long short-term memory，LSTM）神经网络自编码器对负荷曲线进行特征降维；然后，基于相对 *k* 近邻密度峰值（related *k*-nearest neighbor density peaks，RKDP）初始聚类中心选取方法改进 *K*-means 获得训练集精准类别标签；最后训练搭建的 LSTM-卷积

收稿日期：2021-12-19；修回日期：2022-02-25
基金项目：河北省重点研发计划资助项目（20312102D）

神经网络(convolutional neural network,CNN)分类模型,实现大规模负荷数据分类。

## 1 RKDP 初始聚类中心选取方法

DPC 算法的核心思想为:聚类中心本身的局部密度大,即其被小于其密度的邻居所包围;聚类中心与其他具有更大密度的数据点之间有相对大的距离[24]。在 DPC 算法中,每个数据点 $i$ 有 2 个重要参数:局部密度 $\rho_i$ 与相对距离 $\delta_i$。

基于高斯核计算数据点 $i$ 的局部密度 $\rho_i$ 为:

$$\rho_i = \sum_{j \neq i} \mathrm{e}^{-(d_{i,j}/d_c)^2} \tag{1}$$

式中:$d_{i,j}$ 为数据点 $i$,$j$ 之间的距离;$d_c$ 为截断距离,即距离阈值。与数据点 $i$ 距离小于 $d_c$ 的点越多,该点的局部密度 $\rho_i$ 就越大。

相对距离 $\delta_i$ 为数据点 $i$ 与其他密度比它大的数据点的所有距离中的最小值,计算公式为:

$$\delta_i = \begin{cases} \min\ d_{i,j} & \exists \rho_j > \rho_i \\ \max\ d_{i,j} & \forall \rho_j \leqslant \rho_i \end{cases} \tag{2}$$

根据 DPC 算法的核心思想,将相对距离大且局部密度值大的点选定为聚类中心,然后将剩余数据分配到密度比它高的最近数据点所在类别,快速完成聚类。然而,DPC 算法在数据集密集程度不均时效果较差,这是由于该算法定义的局部密度是由全局数据进行计算,未考虑数据内部局部结构差异。当数据集不同类别间密集程度差异较大时,全局范围内密度较高的点可能全分布在密集类别中,容易忽略密度稀疏的类别,难以找到正确的初始聚类中心[25]。因此,通过计算数据点与其近邻点间相对密度可能更能反映该点是否为潜在的聚类中心。

文中基于 DPC 算法思想,提出 RKDP 初始聚类中心选取方法,该方法须提前设定 2 个参数:聚类中心数 $K$ 和 $k$ 近邻的参数 $n$,其具体流程如下。

(1) 首先,通过数据点 $i$ 与其近邻点的距离来计算其局部密度,新的局部密度 $\rho_i$ 计算公式如式(3)所示,$N_i$ 为 $i$ 的 $n$ 个近邻点集合。

$$\rho_i = \sum_{j \in N_i} \mathrm{e}^{-d_{i,j}^2} \tag{3}$$

(2) 计算数据点 $i$ 与其近邻点间的相对密度 $\tilde{\rho}_i$ 作为依据,判断其是否为潜在的聚类中心。数据点 $i$ 相对密度 $\tilde{\rho}_i$ 计算公式为:

$$\tilde{\rho}_i = n\rho_i \Big/ \sum_{j \in N_i} \rho_j \tag{4}$$

式中:$\rho_j$ 为数据点 $i$ 其近邻点 $j$ 的局部密度。通过计算相对密度,可使稀疏类别的数据点局部密度放大,密集类别的数据点局部密度缩小,从而降低密集程度不一对聚类中心选取的影响。$\tilde{\rho}_i > 1$,说明数据点 $i$ 在其局部范围内密度相对其他点较大,也就有可能为聚类中心。选择 $\tilde{\rho} > 1$ 的数据点构成潜在聚类中心集合 $C$。

(3) 基于 DPC 算法思想(聚类中心有着较大的局部密度与相对距离)引入聚类中心权值 $\gamma_i$ 来选择初始聚类中心,计算公式如下:

$$\gamma_i = \tilde{\rho}_i^{\,*} \delta_i^{\,*} \quad i \in C \tag{5}$$

式中:$\tilde{\rho}_i^{\,*}$,$\delta_i^{\,*}$ 分别为集合 $C$ 中数据点的 $\tilde{\rho}_i$ 和 $\delta_i$ 归一化后的结果。将数据点按 $\gamma_i$ 值从大到小进行排列,选取前 $K$ 个点作为初始聚类中心。

## 2 LSTM 自编码器

自编码器(auto-encoder,AE)是一种常用于特征提取与降维的神经网络,包括编码与解码 2 个过程,其基本结构如图 1 所示,包括输入层、隐藏层和输出层 3 个部分[8]。AE 的思想就是在输出层最大程度重构输入数据,同时通过隐藏层提取输入数据的隐藏特征,通过设置隐藏层神经元数量小于输入数据维度即可实现特征降维。

LSTM 神经网络是一种改进的时间循环网络,依靠其独特的门控结构和记忆单元可有效处理长时间序列,目前在时序预测、分类等领域有广泛的应用。LSTM 神经网络基本单元主要包括遗忘门、输入门和输出门 3 个门控单元[23—24]。

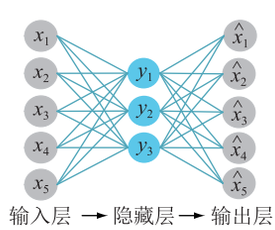


图 1 AE 结构

Fig.1 Structure of AE

文中将传统 AE 与 LSTM 神经网络相结合,提取负荷数据的时序特征,所提出的 LSTM-AE 网络结构如表 1 所示。

表 1 LSTM-AE 网络结构

Table 1 Network structure of LSTM-AE

| 网络 | 名称 | 神经元数量 |
|---|---|---|
| 编码器 | LSTM 网络层 | 64 |
| | LSTM 网络层 | 32 |
| | Dense 层 | 8 |
| | Reshape 层 | |
| 解码器 | LSTM 网络层 | 32 |
| | LSTM 网络层 | 64 |

## 3 基于 LSTM-CNN 的负荷曲线分类模型

CNN 近年来在深度学习领域被广泛应用，其内部基于局部连接和共享权值的方式可有效提取数据的潜在特征[26—27]。文中使用 CNN 提取负荷数据的深层次特征，同时与 LSTM 神经网络提取的时序特征拼接作为特征向量，实现特征增强，从而提升分类模型对不平衡数据的处理能力。所提出的 LSTM-CNN 分类模型如图 2 所示，主要包括 CNN 子模块、LSTM 子模块以及分类模块。CNN 子模块主要由 2 层的一维卷积层与池化层组成。Reshape 层转换输入数据维度，2 层卷积层提取数据特征，激活函数为 Relu；池化层对卷积层提取特征进行下采样，实现特征约简。LSTM 子模块由 2 层 LSTM 网络层构成，神经元数量均为 64，激活函数为 Relu，用于提取负荷的内在时序特征。分类模块中，特征拼接层对 LSTM 及 CNN 子模块提取的特征进行拼接，输出为一维特征向量；第一层全连接层实现特征降维，激活函数为 Relu，数量为 32；第二层全连接层激活函数设置为 Softmax，其神经元数量取决于负荷类别数，输出最后的分类结果。

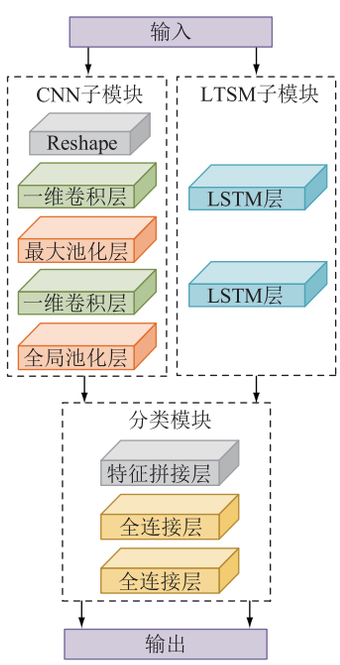


图 2 LSTM-CNN 分类模型结构

Fig.2 Structure of LSTM-CNN classification model

## 4 算例分析

### 4.1 实验数据及评价指标

#### 4.1.1 实验数据介绍

由于负荷数据缺乏类别标签，无法直接测试所提方法对不均衡数据的分类能力，文中基于 UCI 数据集中的 $D_{\text{Iris}}$，$D_{\text{Wine}}$，$D_{\text{Seed}}$ 数据集来验证 RKDP 初始聚类中心选取方法的有效性，同时选取 Synthetic Control 时序数据集对所提出的 LSTM-CNN 分类模型进行测试。最后，选取伦敦智能电表数据集 $D_{\text{L}}$ 及爱尔兰负荷数据 $D_{\text{I}}$ 作为实际负荷数据进行负荷聚类及分类实验（采样时间间隔均为 30 min，即每天有 48 个采样点），验证所提方法的有效性。文中所使用的实验平台处理器型号为 AMD Ryzen Threadripper 3970X，操作系统为 Windows 10，所用编程语言为 Python 3.7，所提出的神经网络模型采用 keras 深度学习框架搭建。

#### 4.1.2 评价指标介绍

在聚类性能评估指标方面，对于有类别标签的数据集，选取调整互信息（adjusted mutual information，AMI）$i_{\text{AMI}}$、调整兰德系数（adjusted rand index，ARI）$i_{\text{ARI}}$ 和 Fowlkes-Mallows 指数（fowlkes-mallows index，FMI）$i_{\text{FMI}}$ 3 项指标来描述聚类结果与实际标签的吻合程度，上限均为 1，其值越接近 1 表示聚类效果越好。对于无标签负荷数据，选取常用的轮廓系数（silhouette coefficient，SC）$i_{\text{SC}}$ 和戴维森堡丁指数（Davies-Bouldin index，DBI）$i_{\text{DBI}}$，$i_{\text{SC}}$ 值越大、$i_{\text{DBI}}$ 越小意味着类内距离越小，类间距离越大，聚类效果越好[28—30]。在分类模型评估指标方面，直接选择分类准确率作为分类模型的评价指标。

### 4.2 RKDP-*K*-means 性能测试

#### 4.2.1 RKDP 有效性验证

首先将 RKDP-*K*-means 算法直接与 *K*-means 算法进行对比，验证 RKDP 初始聚类中心选取方法能够提升 *K*-means 方法对不均衡数据的聚类精度。基于 $D_{\text{Iris}}$，$D_{\text{Wine}}$，$D_{\text{Seed}}$ 3 个真实数据集，采用随机抽样法分别构建不平衡比例为 3∶1，5∶1，10∶1的数据集，聚类数均为各数据集的类别数，$k$ 近邻参数在 3～20 之间选取，每种不平衡比例下重复 5 次，即每个数据集进行 15 次实验，2 种方法的 $i_{\text{ARI}}$，$i_{\text{AMI}}$，$i_{\text{FMI}}$ 及其平均值分别见表 2 和表 3，$i_{\text{Iter}}$ 为迭代次数均值。

由表 2 和表 3 可知，*K*-means 算法聚类精度随着不平衡比例加重逐渐下降，以 $D_{\text{Wine}}$ 数据集为例，数据不平衡比例由 3∶1变为 10∶1时，$i_{\text{ARI}}$ 指标由 0.858变为 0.670，而 RKDP-*K*-means 算法由 0.876 变为 0.804，仍保持较高水平；在各指标平均值方面，相对于 *K*-means 算法，RKDP-*K*-means 算法的 $i_{\text{ARI}}$，$i_{\text{AMI}}$，$i_{\text{FMI}}$ 均有提升，且迭代次数减少。综上，文中所提出的 RKDP 初始聚类中心选取方法能够有效提升 *K*-means 算法对不平衡数据的处理能力。

#### 4.2.2 聚类效果对比分析

为了更加客观地验证所提算法处理不均衡数

表 2 *K*-means 实验结果

Table 2 Experimental results of *K*-means

| 不平衡比例 | 数据集 | $i_{ARI}$ | $i_{AMI}$ | $i_{FMI}$ | $i_{Iter}$ |
|---|---|---|---|---|---|
| 3:1 | $D_{Wine}$ | 0.858 | 0.837 | 0.911 | 7.067 |
| | $D_{Iris}$ | 0.673 | 0.701 | 0.797 | 5.067 |
| | $D_{Seed}$ | 0.690 | 0.661 | 0.809 | 7.800 |
| 5:1 | $D_{Wine}$ | 0.785 | 0.774 | 0.871 | 5.667 |
| | $D_{Iris}$ | 0.698 | 0.716 | 0.819 | 5.400 |
| | $D_{Seed}$ | 0.675 | 0.649 | 0.805 | 8.067 |
| 10:1 | $D_{Wine}$ | 0.670 | 0.689 | 0.808 | 9.133 |
| | $D_{Iris}$ | 0.625 | 0.649 | 0.789 | 4.933 |
| | $D_{Seed}$ | 0.657 | 0.630 | 0.802 | 6.308 |
| 平均值 | | 0.703 | 0.701 | 0.823 | 6.605 |

表 3 RKDP-*K*-means 实验结果

Table 3 Experimental results of the RKDP-*K*-means

| 不平衡比例 | 数据集 | $i_{ARI}$ | $i_{AMI}$ | $i_{FMI}$ | $i_{Iter}$ |
|---|---|---|---|---|---|
| 3:1 | $D_{Wine}$ | 0.876 | 0.852 | 0.922 | 5.400 |
| | $D_{Iris}$ | 0.703 | 0.721 | 0.817 | 7.000 |
| | $D_{Seed}$ | 0.721 | 0.681 | 0.828 | 6.533 |
| 5:1 | $D_{Wine}$ | 0.850 | 0.824 | 0.911 | 4.600 |
| | $D_{Iris}$ | 0.749 | 0.752 | 0.850 | 4.733 |
| | $D_{Seed}$ | 0.727 | 0.673 | 0.837 | 4.733 |
| 10:1 | $D_{Wine}$ | 0.804 | 0.773 | 0.890 | 5.000 |
| | $D_{Iris}$ | 0.649 | 0.668 | 0.804 | 5.000 |
| | $D_{Seed}$ | 0.697 | 0.647 | 0.827 | 6.308 |
| 平均值 | | 0.753 | 0.732 | 0.854 | 5.479 |

据的有效性,将 RKDP-*K*-means 算法与基于划分的 *K*-means、基于空间密度的聚类(density-based spatial clustering of applications with noise, DBSCAN)[31]、基于层次的凝聚聚类(agglomerative clustering, AG)及基于图论的谱聚类(spectral clustering, SP)4 种方法进行对比。其中,*K*-means、AG 及 SP 聚类数设置为 3,DBSCAN 邻域半径以 0.02 为步长,在 0.1~0.5 之间选取,邻域内最少样本数在 5~25 之间选取,RKDP-*K*-means 的 $k$ 近邻参数在 3~20 之间选取。所有结果均为最佳参数下测得,每组不平衡数据同样重复 5 次,表 4 为 5 种方法的准确率。

由表 4 可知,RKDP-*K*-means 算法在各数据集下均优于 *K*-means 算法,以 $D_{Wine}$ 数据集为例,随着不平衡比例加大,*K*-means 的准确率从 0.957 变为 0.829,RKDP-*K*-means 从 0.964 变为 0.915,仍有较高准确度。整体上看,RKDP-*K*-means 算法的准确率均值均优于其他 4 种方法。因此,RKDP-*K*-means 算法在处理不平衡数据时具有优势。

### 4.3 实际负荷数据聚类分析

采用实际负荷数据来对 LSTM-AE 的性能进行评价分析。从 $D_L$ 数据集中随机选取 500 条负荷曲线为实验对象,基于 *K*-means 算法计算不同聚类数下的 $i_{SC}$ , $i_{DBI}$ 指标,结果如图 3 所示,当聚类数目为 4 时,2 项指标所反映的聚类效果较好,因此设置聚类数为 4。分别采用 LSTM-AE、主成分分析(principal component analysis, PCA)、核主成分分析(kernel PCA, KPCA)、AE 4 种降维方法(维度均设置为 8)降维后采用 RKDP-*K*-means 聚类以及 *K*-means, RKDP-*K*-means 不降维直接聚类进行对比,重复 10 次试验。同时基于 $D_I$ 重复上述实验进行验证,聚类中心数设置为 3,结果如表 5 所示。

表 4 5 种聚类算法准确率

Table 4 Accuracy of five clustering algorithms

| 不平衡比例 | 数据集 | *K*-means | DBSCAN | AG | SP | RKDP-*K*-means |
|---|---|---|---|---|---|---|
| 3:1 | $D_{Wine}$ | 0.957 | 0.732 | 0.966 | 0.974 | 0.964 |
| | $D_{Iris}$ | 0.854 | 0.856 | 0.837 | 0.823 | 0.870 |
| | $D_{Seed}$ | 0.881 | 0.850 | 0.872 | 0.875 | 0.892 |
| 5:1 | $D_{Wine}$ | 0.923 | 0.778 | 0.944 | 0.940 | 0.948 |
| | $D_{Iris}$ | 0.838 | 0.865 | 0.844 | 0.799 | 0.853 |
| | $D_{Seed}$ | 0.869 | 0.853 | 0.858 | 0.861 | 0.890 |
| 10:1 | $D_{Wine}$ | 0.829 | 0.806 | 0.881 | 0.888 | 0.915 |
| | $D_{Iris}$ | 0.808 | 0.898 | 0.799 | 0.778 | 0.820 |
| | $D_{Seed}$ | 0.847 | 0.863 | 0.813 | 0.843 | 0.877 |
| 平均值 | | 0.867 | 0.833 | 0.868 | 0.864 | 0.892 |

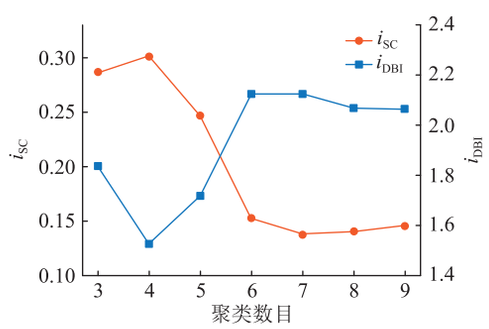


图 3 SC 和 DBI 与聚类数目关系

Fig.3 Relationship between SC, DBI and cluster number

表 5 LSTM-AE 有效性验证实验结果

Table 5 Results of validity verification of LSTM-AE

| 数据集 | 指标 | *K*-means | RKDP-*K*-means | LARK | PCA | KPCA | AE |
|---|---|---|---|---|---|---|---|
| $D_L$ | $i_{SC}$ | 0.317 | 0.342 | 0.463 | 0.360 | 0.357 | 0.346 |
| | $i_{DBI}$ | 1.552 | 1.540 | 1.408 | 1.515 | 1.512 | 1.635 |
| $D_I$ | $i_{SC}$ | 0.189 | 0.219 | 0.269 | 0.250 | 0.249 | 0.232 |
| | $i_{DBI}$ | 2.438 | 2.436 | 2.220 | 2.377 | 2.382 | 2.586 |

为表述方便,将经 LSTM-AE 降维后 RKDP-*K*-means 聚类命名为 LARK 聚类算法。由表 5 可知,RKDP-*K*-means 在 $i_{SC}$ 指标上优于 *K*-means 算法, $i_{DBI}$ 指标基本持平。与 RKDP-*K*-means 聚类相比,LARK

算法的各项指标均有较大幅度的改善，在 $D_L$ 和 $D_I$ 数据集上，$i_{SC}$ 指标分别提升0.121 和 0.05，这表明 LSTM-AE 能够提升 RKDP-*K*-means 的聚类精度。通过对比 4 种降维方法可以发现，LSTM-AE 的特征提取能力优于其他 3 种方法。

### 4.4 LSTM-CNN 分类模型测试

#### 4.4.1 LSTM-CNN 网络测试

首先，使用 Synthetic Control 时序数据集测试所提出的 LSTM-CNN 分类模型，与相同结构的 LSTM 网络模型以及传统支持向量机（support vector machine，SVM）模型进行对比，训练集与测试集比例为 1∶1，神经网络迭代次数设置为 100，优化器为 adam，损失函数为 MSE；SVM 算法中核参数为径向基，分类准确率如表 6 所示。由表 6 可知，3 种方法在训练集上均有 100%的准确率，在测试集上，LSTM 模型最弱，仅有 95.3%，而所提出的 LSTM-CNN 分类模型与 SVM 均达到了 97.7%的准确率。

表 6 3 种方法分类准确率对比

Table 6 Classification accuracy comparison of three methods %

| 数据集 | SVM | LSTM | LSTM-CNN |
|---|---|---|---|
| 训练集 | 100 | 100 | 100 |
| 测试集 | 97.7 | 95.3 | 97.7 |

为了验证 LSTM-CNN 模型对不平衡时序数据的分类性能，基于 Synthetic Control 时序数据集构建了 15 种类别不平衡数据集，与 LSTM 模型和 SVM 模型进行对比，传统 SVM 模型处理不平衡时序数据性能较弱，准确率均值仅有 80.7%，LSTM 模型准确率均值为 87.9%，而所提 LSTM-CNN 模型相对其他 2 种方法有着更好的分类性能，准确率高达92.2%。由此可见，提出的基于 LSTM-CNN 模型能够有效处理时序不平衡数据分类问题。

#### 4.4.2 实际负荷数据分类测试

（1）算法分类性能测试。基于 $D_L$ 和 $D_I$ 负荷数据，分别随机选取 10 万条负荷曲线，按照 3∶7构造训练集与测试集，基于 LARK 获得训练集标签数据，训练 LSTM-CNN 模型实现对测试集的分类，与 *K*-means 和 LARK 直接聚类进行对比，$D_L$，$D_I$ 的聚类中心数分别设为 6 和 8，$i_{SC}$ 和 $i_{DBI}$ 指标如表 7 所示。由表 7 可知，文中方法聚类精度优于 LARK 算法，在 2 个数据集上，$i_{SC}$ 指标分别提升 0.043 和 0.044。*K*-means 算法虽然在 $D_I$ 上 $i_{DBI}$ 指标最小，但其 $i_{SC}$ 指标仅有 0.074，文中方法 $i_{SC}$ 指标相较于 *K*-means 提升 0.118，$i_{DBI}$ 指标提升 0.172，整体上看，所提出的分类方法分类性能优于其他 2 种方法。

表 7 3 种方法 SC、DBI 对比

Table 7 Comparison of SC，DBI of three methods

| 数据集 | 指标 | *K*-means | LARK | 文中方法 |
|---|---|---|---|---|
| $D_L$ | $i_{SC}$ | 0.332 | 0.430 | 0.475 |
| | $i_{DBI}$ | 1.535 | 1.455 | 1.414 |
| $D_I$ | $i_{SC}$ | 0.074 | 0.148 | 0.192 |
| | $i_{DBI}$ | 2.715 | 2.932 | 2.887 |

（2）负荷分类结果。图 4 为基于 $D_I$ 归一化后的负荷分类结果，可以看出用户的用电模式多种多样，8 种典型负荷曲线可大致分为平稳型用电和尖峰型用电。类别 1 一整天始终保持较高的负荷水平，在凌晨用电量较大。类别 5 也是平稳型用电类型，但其负荷水平一直很低。其余 6 种皆为尖峰型用电，但用电高峰时段不同，类别 7 是典型的午间负荷，类别 4 和类别 6 用电高峰分别出现在下午和傍晚，类别 2、类别 3 和类别 8 是典型的晚间负荷，其中类别 3 的用电高峰时间持续较长。通过挖掘用户的典型用电模式，有助于电力公司制定更好的售电方案，提高服务水平。

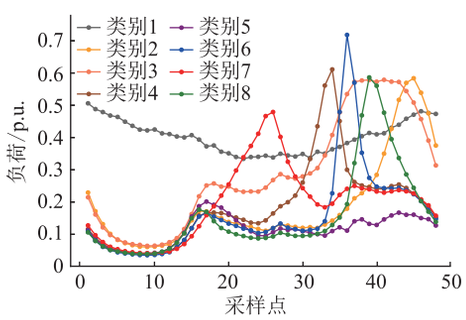


图 4 典型负荷曲线

Fig.4 Typical daily load profiles

（3）算法效率测试。文中所提方法包括 LARK 聚类获取样本标签、LSTM-CNN 模型训练及分类 3 个环节，实验对比了 *K*-means、LARK 及文中方法（训练集∶测试集＝3∶7）在不同规模负荷数据集下的计算速度，执行时间如图 5 所示。

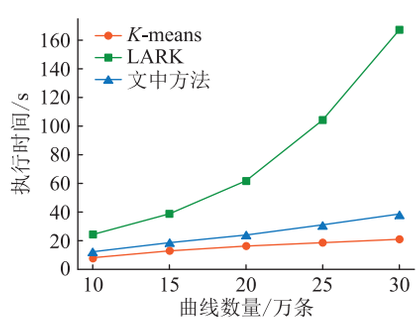


图 5 算法效率对比

Fig.5 Comparison of algorithmic efficiency

从图 5 可以看出，LARK 聚类算法随着数据规

模增加运行时间急剧增大,在对 30 万条负荷曲线分类时,LARK 算法运行时间达到 167 s,而文中方法仅用时 37.4 s,相比于 LARK 算法效率提升 3.46 倍;传统 *K*-means 算法用时 20.5 s,文中方法虽相较于 *K*-means 算法较差,但在分类性能上表现更好,同时文中方法主要耗时在于标签获取与训练分类模型环节(共耗时 34.2 s),分类阶段用时仅 3.22 s,分类模型一旦训练完成后可重复使用。因此,文中所提方法在面对大规模负荷分类时具有效率优势。

## 5 结论

文中提出了一种考虑数据分布不均衡的负荷曲线分类方法,主要包括基于 LSTM-AE 实现负荷数据降维、基于 RKDP-*K*-means 聚类算法获得负荷类别标签及训练 LSTM-CNN 分类模型实现大规模负荷分类三部分。通过算例分析验证了文中方法的有效性,得到以下结论:

(1) 基于 UCI 公共数据集验证了所提出的 RKDP 初始聚类中心选取方法可有效提升 *K*-means 算法对不均衡数据的聚类性能,其中 $i_{\mathrm{ARI}}$ 指标提升 6.6%,迭代次数减少 17.1%;

(2) 在 RKDP-*K*-means 算法对负荷进行聚类分析时,所提出的 LSTM-AE 特征提取方法可有效提升 RKDP-*K*-means 的聚类精度,在伦敦负荷测试集,$i_{\mathrm{SC}}$ 指标提升 35.4%;

(3) 在大规模负荷分类上,基于 LARK 聚类与 LSTM-CNN 分类模型相结合的负荷分类方法相比于 LARK 算法有着更好的负荷分类性能,其中 $i_{\mathrm{SC}}$ 指标提升 29.7%,效率提升 3.46 倍。

**参考文献:**


[1] 朱天怡,艾芊,贺兴,等. 基于数据驱动的用电行为分析方法及应用综述[J]. 电网技术,2020,44(9):3497-3507.
ZHU Tianyi, AI Qian, HE Xing, et al. An overview of data-driven electricity consumption behavior analysis method and application[J]. Power System Technology, 2020, 44(9): 3497-3507.

[2] 夏水斌,张芹,谢玮,等. 电力用户用电信息采集系统建设研究[J]. 自动化与仪器仪表,2018(10):48-50.
XIA Shuibin, ZHANG Qin, XIE Wei, et al. Research on the construction of electrical information acquisition system for power users[J]. Automation & Instrumentation, 2018(10): 48-50.

[3] 丁明,黄冯,邹佳芯,等. 改进谱聚类与遗传算法相结合的电力时序曲线聚类方法[J]. 电力自动化设备,2019,39(2):93-99,114.
DING Ming, HUANG Feng, ZOU Jiaxin, et al. Power time series curve clustering method combining improved spectral clustering and genetic algorithm[J]. Electric Power Automation Equipment, 2019, 39(2): 93-99, 114.

[4] 严强,李扬,樊友杰,等. 基于加权表决集成聚类的居民用电行为回归分析[J]. 电网技术,2021,45(11):4435-4446.
YAN Qiang, LI Yang, FAN Youjie, et al. Regression analysis of residential electricity consumption behavior based on weighted voting ensemble clustering[J]. Power System Technology, 2021, 45(11): 4435-4446.

[5] LI K H, MA Z J, ROBINSON D, et al. Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering[J]. Applied Energy, 2018, 231: 331-342.

[6] 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.
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.

[7] 梁京章,黄星舒,吴丽娟,等. 基于 KPCA 和改进 *K*-means 的电力负荷曲线聚类方法[J]. 华南理工大学学报(自然科学版),2020,48(6):143-150.
LIANG Jingzhang, HUANG Xingshu, WU Lijuan, et al. Clustering method of power load profiles based on KPCA and improved *K*-means[J]. Journal of South China University of Technology (Natural Science Edition), 2020, 48(6): 143-150.

[8] 庞传军,余建明,冯长有,等. 基于 LSTM 自动编码器的电力负荷聚类建模及特性分析[J]. 电力系统自动化,2020,44(23):57-63.
PANG Chuanjun, YU Jianming, FENG Changyou, et al. Clustering modeling and characteristic analysis of power load based on long-short-term-memory[J]. Automation of Electric Power Systems, 2020, 44(23): 57-63.

[9] 宋军英,崔益伟,李欣然,等. 基于欧氏动态时间弯曲距离与熵权法的负荷曲线聚类方法[J]. 电力系统自动化,2020,44(15):87-94.
SONG Junying, CUI Yiwei, LI Xinran, et al. Load curve clustering method based on euclidean dynamic time warping distance and entropy weight[J]. Automation of Electric Power Systems, 2020, 44(15): 87-94.

[10] 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.
XU Shenglan, SI Caomingzhe, WAN Can, et al. Ensemble spectral clustering algorithm for load profiles considering dual-scale similarities[J]. Automation of Electric Power Systems, 2020, 44(22): 152-160.

[11] 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.

[12] KIM N, PARK S, LEE J, et al. Load profile extraction by mean-shift clustering with sample Pearson correlation coefficient distance[J]. Energies, 2018, 11(9): 2397.

[13] 黄奇峰,杨世海,邓欣宇,等. 基于欠完备自编码器的用户

用电行为分类分析方法[J]. 电力工程技术,2019,38(6):24-30.
HUANG Qifeng,YANG Shihai,DENG Xinyu,et al. Classification analysis method for electricity consumption behavior based on undercomplete autoencoder[J]. Electric Power Engineering Technology,2019,38(6):24-30.

[14] 刘洋,刘洋,许立雄. 适用于海量负荷数据分类的高性能反向传播神经网络算法[J]. 电力系统自动化,2018,42(21):96-103.
LIU Yang,LIU Yang,XU Lixiong. High-performance back propagation neural network algorithm for classification of mass load data[J]. Automation of Electric Power Systems,2018,42(21):96-103.

[15] 林顺富,顾乡,汤继开,等. 基于稀疏自动编码器神经网络的负荷曲线分类方法[J]. 电网技术,2020,44(9):3508-3515.
LIN Shunfu,GU Xiang,TANG Jikai,et al. Power load profile classification method based on neural network of sparse automatic encoder[J]. Power System Technology,2020,44(9):3508-3515.

[16] 白明亮,张冬雪,刘金福,等. 基于深度自编码器和支持向量数据描述的燃气轮机高温部件异常检测[J]. 发电技术,2021,42(4):422-430.
BAI Mingliang,ZHANG Dongxue,LIU Jinfu,et al. Anomaly detection of gas turbine hot components based on deep autoencoder and support vectordata description[J]. Power Generation Technology,2021,42(4):422-430.

[17] 于小青,齐林海. 基于流数据聚类算法的电力大数据异常检测[J]. 电力信息与通信技术,2020,18(3):8-14.
YU Xiaoqing,QI Linhai. Power big data anomaly detection based on stream data clustering algorithm[J]. Electric Power Information and Communication Technology,2020,18(3):8-14.

[18] 武森,汪玉枝,高晓楠. 基于近邻的不均衡数据聚类算法[J]. 工程科学学报,2020,42(9):1209-1219.
WU Sen,WANG Yuzhi,GAO Xiaonan. Clustering algorithm for imbalanced data based on nearest neighbor[J]. Chinese Journal of Engineering,2020,42(9):1209-1219.

[19] 代杰杰,滕莹冰,龚越明. 基于区间集聚类分析的电力设备状态异常检测方法[J]. 电力信息与通信技术,2019,17(11):1-6.
DAI Jiejie,TENG Yingbing,GONG Yueming. Power equipment state anomaly detection method based on interval set theory and clustering analysis[J]. Electric Power Information and Communication Technology,2019,17(11):1-6.

[20] 王帅,杜欣慧,姚宏民,等. 面向含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.

[21] 李想,王鹏,刘洋,等. 考虑类别不平衡的海量负荷用电模式辨识方法[J]. 中国电机工程学报,2020,40(1):128-137,380.
LI Xiang,WANG Peng,LIU Yang,et al. Massive load pattern identification method considering class imbalance[J]. Proceedings of the CSEE,2020,40(1):128-137,380.

[22] 唐冬来,郝建维,刘荣刚,等. 基于动态规划的配电台区三相负荷不平衡治理方法[J]. 电力系统保护与控制,2020,48(21):58-66.
TANG Donglai,HAO Jianwei,LIU Ronggang,et al. Control method of three phase load imbalance in a distribution station area based on dynamic programming[J]. Power System Protection and Control,2020,48(21):58-66.

[23] 刘洋,刘洋,许立雄,等. 计及数据类别不平衡的海量用户负荷典型特征高性能提取方法[J]. 中国电机工程学报,2019,39(14):4093-4104.
LIU Yang,LIU Yang,XU Lixiong,et al. A high performance extraction method for massive user load typical characteristics considering data class imbalance[J]. Proceedings of the CSEE,2019,39(14):4093-4104.

[24] RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496.

[25] 赵嘉,姚占峰,吕莉,等. 基于相互邻近度的密度峰值聚类算法[J]. 控制与决策,2021,36(3):543-552.
ZHAO Jia,YAO Zhanfeng,LYU Li,et al. Density peaks clustering based on mutual neighbor degree[J]. Control and Decision,2021,36(3):543-552.

[26] 刘亚珲,赵倩. 基于聚类经验模态分解的 CNN-LSTM 超短期电力负荷预测[J]. 电网技术,2021,45(11):4444-4451.
LIU Yahui,ZHAO Qian. Ultra-short-term power load forecasting based on cluster empirical mode decomposition of CNN-LSTM[J]. Power System Technology,2021,45(11):4444-4451.

[27] 陆继翔,张琪培,杨志宏,等. 基于 CNN-LSTM 混合神经网络模型的短期负荷预测方法[J]. 电力系统自动化,2019,43(8):131-137.
LU Jixiang,ZHANG Qipei,YANG Zhihong,et al. Short-term load forecasting method based on CNN-LSTM hybrid neural network model[J]. Automation of Electric Power Systems,2019,43(8):131-137.

[28] 金伟超,张旭,刘晟源,等. 基于剪枝策略和密度峰值聚类的行业典型负荷曲线辨识[J]. 电力系统自动化,2021,45(4):20-28.
JIN Weichao,ZHANG Xu,LIU Shengyuan,et al. Identification of typical industrial power load curves based on pruning strategy and density peak clustering[J]. Automation of Electric Power Systems,2021,45(4):20-28.

[29] 王德文,周昉昉. 基于无监督极限学习机的用电负荷模式提取[J]. 电网技术,2018,42(10):3393-3400.
WANG Dewen,ZHOU Fangfang. Extraction of electricity consumption load pattern based on unsupervised extreme learning machine[J]. Power System Technology,2018,42(10):3393-3400.

[30] 张彼德,洪锡文,刘俊,等. 基于无监督学习的 MMC 子模块

开路故障诊断方法[J]. 电力系统保护与控制,2021,49(12):98-105.
ZHANG Bide,HONG Xiwen,LIU Jun,et al. Diagnosis method for sub-module open-circuit fault in modular multilevel converter based on unsupervised learning[J]. Power System Protection and Control,2021,49(12):98-105.
[31] 袁昊,金铭,邱昱,等. 基于电力日志特征的 DBSCAN 聚类[J]. 电力信息与通信技术,2019,17(5):68-72.
YUAN Hao, JIN Ming, QIU Yu, et al. DBSCAN clustering based on power log characteristics[J]. Electric Power Information and Communication Technology,2019,17(5):68-72.

作者简介:


张慧波

张慧波(1997),男,硕士在读,研究方向为人工智能方法在配用电系统的应用(E-mail:zhbo@tju.edu.cn);

王守相(1973),男,博士,教授,研究方向为分布式发电、微电网、智能配电网;

赵倩宇(1990),女,博士,讲师,研究方向为电力系统稳定性。

## Residential user load curve classification method considering data imbalance

ZHANG Huibo[1], WANG Shouxiang[1], ZHAO Qianyu[1], REN Jie[2], WANG Hai[2]

(1. Key Laboratory of Smart Grid of Ministry of Education (Tianjin University), Tianjin 300072, China;
2. State Grid Hebei Zhangjiakou Scenic Storage and Transportation New Energy Co., Ltd., Zhangjiakou 075061, China)

**Abstract**: Due to the diversity and randomness of users' electricity consumption behaviors, the imbalance of load data classes is increasingly obvious. Traditional load curve classification technologies have become ineffective to deal with the im-balanced class problem of data. Therefore, an algorithm combing improved *K*-means with long short term memory (LSTM) neural network and convolutional neural network (CNN) classification model is proposed. Firstly, to improve the classifica-tion accuracy of the *K*-means on imbalanced data, a method of relative k-nearest neighbor density peaks (RKDP) based on the density peak clustering algorithm (DPC) is proposed to select the initial clustering centre of *K*-means. Secondly, in order to improve the performance of RKDP-*K*-means in processing high-dimensional load data, an au-to-encoder based on LSTM is used to extract load characteristics from high dimensional data, and com-bined with RKDP-*K*-means to obtain accurate load profiles labels. Finally, based on LSTM neural network and CNN, load characteristics were extracted to construct load curve classification model to realize the classification of large-scale load curve. Different algorithms are employed to classify Ireland smart meter data set and London load data set. The results show the proposed algorithm is more effective and practicable in large-scale load curve classification.

**Keywords**: load curve classification; imbalanced data; improved *K*-means; auto-encoder; long short term memory (LSTM) neural network; convolutional neural network (CNN)

(编辑 钱悦)