基于节点日负荷曲线的深度嵌入式聚类及其改进方法对比研究

陈 谦,陈嘉雯,王苏颖,史 锐

(河海大学能源与电气学院,江苏 南京 211100)

摘 要:基于日负荷曲线的负荷节点分类是负荷建模的重要环节,详略得当的分类结果保留了负荷节点的内在特性,可提升电力系统仿真计算的效率。当前基于人工智能的节点聚类方法进展迅速,然而总体上针对数据深层特征提取的适应性仍存在不足。采用了基于改进的深度嵌入式算法的日负荷曲线聚类方法,利用神经网络可有效提取数据的深层特征的能力。进而,提出一种先升维后聚类的改进方法,通过算例对比分析,验证了本文所提算法的可行性,以及所提升维—重构聚类方法的正确性。

关 键 词:负荷建模;日负荷曲线聚类;深度嵌入式;升维-重构聚类

DOI:10.19781/j.issn.1673-9140.2023.01.015 **中图分类号**:TM712 **文章编号**:1673-9140(2023)01-0130-08

Comparative study on deep embedded clustering and its improved methods based on node daily load curve

CHEN Qian, CHEN Jiawen, WANG Suying, SHI Rui

(College of Energy and Elctrical Engineering, Hohai University, Nanjing 211100, China)

Abstract: Load node classification based on daily load curve is an important part of load modeling. The detailed and appropriate classification results retain the internal characteristics of load nodes and can improve the efficiency of power system simulation calculation. At present, the node clustering method based on artificial intelligence has made rapid progress. However, the overall adaptability to data deep feature extraction is still insufficient. This paper presents the daily load curve clustering method based on the improved deep embedded algorithm, which uses the ability of neural network to effectively extract the deep features of the data. Then, an improved method of increasing the dimension first and then clustering is proposed. Through the comparative analysis of numerical examples, the feasibility of the proposed algorithm and the correctness of the improved dimension reconstruction clustering method are verified.

Key words: load modeling; daily load curve clustering; deep embedded; dimension increasing-reconstruction clustering

负荷建模是电力系统建模不可或缺一部分,负荷数据中含有丰富信息,能体现客户用电模式,了解用电特点,提升预测用电量精确度,指导电价制定等[1-4],研究基于日负荷曲线的负荷节点分类技术对建立准确的负荷模型具有重要意义。

传统的负荷节点分类方法为统计综合法,该方法需要耗费大量的财力物力,且十分耗时。目前多采用基于人工智能方法的日负荷曲线聚类方法[5],

收稿日期:2022-03-28;**修回日期**:2022-09-20

基金项目:国家自然科学基金(51837004)

通信作者:陈嘉雯(1997—),女,硕士研究生,主要从事人工智能算法在负荷建模中的应用研究;E-mail:506137005@qq.com

这大大提高了工作效率。电力负荷曲线聚类实质上是利用无监督的算法，将没有标签的负荷数据根据曲线的相似性将其划分到不同的类簇中以提取它的群体特性。当前研究大多采用降维算法和聚类算法结合的思路，文献[6]采用主成分分析算法（principle component analysis，PCA）对高维数据进行降维，采用K-shape算法挖掘负荷数据的形状特征进行聚类，但是主成分分析方法主要适用于线性数据，对于负荷数据这类非线性数据的特征难以准确提取；文献[7]采用核主成分分析（kernel principle component analysis，KPCA）对负荷数据进行降维并利用DK-Means算法进行聚类，核主成分分析相较主成分分析更适用于非线性数据。这些降维算法都能有效降低数据的时空复杂度，但对曲线的时序特征不能做很好的保留。

近年来，随着神经网络的发展，采用神经网络对电力系统进行分析研究成为热点问题[8]，因此，许多学者也尝试采用神经网络的方法对日负荷数据进行降维和聚类。将深度学习技术运用于日负荷曲线的聚类研究，能够很好地提取数据中复杂的特征，获得更好的效果[9-11]。文献[12]采用自编码器与模糊C均值聚类实现特征提取和用电模式区分，该方法中特征提取与聚类任务处于分离状态，这会降低聚类的质量。

综合考虑以上情况，本文提出首先采用改进的深度嵌入式聚类算法（improved deep embedded clustering，IDEC）对日负荷曲线进行聚类研究，该方法先使用自动编码器的编码器部分对日负荷曲线进行降维和特征提取，送入K-means层得到初始聚类中心；然后对提取的负荷特征利用KL散度进行软分配；最后联合优化重构损失和聚类损失，得到聚类结果。考虑到之前的研究均采用降维－聚类或者直接聚类的思路，但是目前用于聚类的负荷数据大多不超过288维（每5 min一个采集点），现代计算机的计算能力很强，本文在此基础上提出采用升维—聚类的改进方法并在IDEC算法中进行实践。先将日负荷数据映射到一个更高的维度上，再进行特征提取和聚类，并将其与降维后聚类以及直接聚类思路下的聚类结果进行对比。

1　深度嵌入式聚类算法

1.1　自动编码器

自动编码器（autoencoder）是一种经过训练的神经网络，可以尝试将输入复制到输出。其内部含有一个描述输入代码的隐藏层Z。自动编码器由2个部分组成，分别是编码器$z=f_{W(x)}$和解码器$x'=g_{W(z)}$，编码器用于对数据进行编制、转换为可用以通讯、传输和存储的信号形式，解码器用于数据的重构。欠完备的自动编码器和去噪自动编码器是2种常用的自编码器。

1）欠完备的自动编码器。其控制隐藏层的维数低于输入数据的维度，学习这种欠完备的表示迫使其捕获数据的显著特征。

2）去噪自动编码器。其目标最小化为

$$L=\left\| x-g_{W}\left(f_{W(\tilde{x})} \right) \right\|_{2}^{2} \tag{1}$$

其中，$\tilde{x}$为被某种形式的噪声破坏的x的副本，因此去噪自动编码器必须从这种损坏的副本中恢复x，而不是简单地复制它们的输入[13]。这样，去噪自动编码器能够迫使编码器和解码器捕获数据生成自动分布的结构。

本文采用的改进深度嵌入式聚类算法，去噪自动编码器在其中用于预训练，欠完备的自动编码器将在初始化之后被添加到深度嵌入式聚类算法框架中。

1.2　深度嵌入式聚类算法

深度嵌入式聚类算法（deep embedded clustering，DEC）的网络结构如图1所示[14]。它从预训练自动编码器开始，然后移除解码器部分，通过优化以下目标，对剩余编码器部分进行微调。

优化目标为

$$L=\mathrm{KL}(P\|Q)=\sum_{i}\sum_{j}p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{2}$$

式中，q_{ij}为嵌入点z_i和聚类中心μ_i之间的相似性，由students-t分布测定：

$$q_{ij}=\frac{\left(1+\|z_i-\mu_j\|^2\right)^{-1}}{\sum_j\left(1+\|z_i-\mu_j\|^2\right)^{-1}} \tag{3}$$

目标分布p_{ij}定义为

$$p_{ij}=\frac{q_{ij}^2/\sum_i q_{ij}}{\sum_j\left(q_{ij}^2/\sum_i q_{ij}\right)} \tag{4}$$

如式(4)所示,目标分布P由Q定义,其中P为真实分布,Q为P的拟合分布,所以对L的最小化优化是自我训练的一种形式。

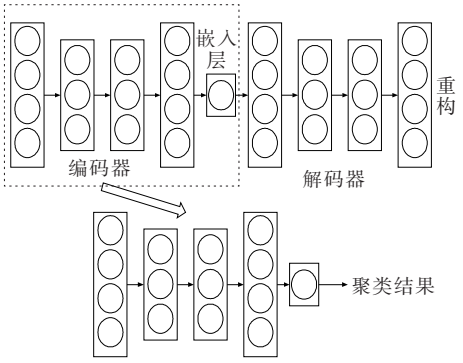


图1 DEC网络结构

Figure 1 Network structure of DEC

f_W为编码器的映射,即$z_i=f_W(x_i)$,其中x_i是数据集X中的元素。预训练后,所有嵌入点z_i可以通过f_W提取,然后对$\{z_i\}$使用K-means聚类得到初始的聚类中心簇$\{\mu_i\}$。根据式(2)计算L,对样本进行标签分配,样本x_i的预测标签为$\arg\max_j q_{ij}$。

在反向传播的过程中,L关于z_i和μ_i的导函数$\partial L/\partial z_i$和$\partial L/\partial\mu_i$可以很容易求得,然后它们分别被传递用来更新f_W和集群中心μ_j。

$$\mu_j=\mu_j-\lambda\frac{\partial L}{\partial\mu_j} \tag{5}$$

总的来说,DEC算法分为2个阶段,第1阶段用深度自编码器对参数进行初始化;第2阶段是参数优化通过迭代计算辅助目标分布和最小化KL散度,其中包括计算嵌入点和簇中心之间软分配和更新深度映射f_W,并通过使用辅助目标分布学习当前的高置信度分配来细化聚类中心。

DEC的最大贡献是聚类损失(具体来说是目标分布P),其原理是使用高度机密的样本作为监督,使每个集群中的样本分布更加密集,但是无法保证将边缘附近的样本拉向正确的聚类。

1.3 升维—改进的深度嵌入式聚类算法

改进的深度嵌入式算法(improved deep embedded clustering,IDEC)的网络结构如图2所示,由自动编码器和聚类层组成。聚类层连接在自动编码器的隐藏层之后[14]。

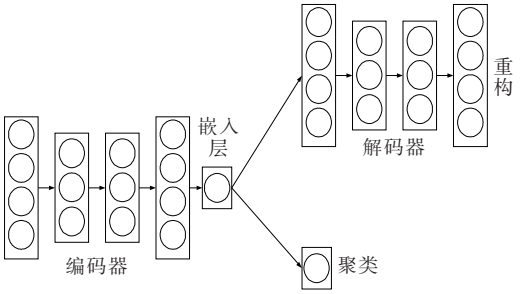


图2 IDEC网络结构

Figure 2 Network structure of IDEC

考虑一个有n个样本组成的数据集X,每一个样本$x_i\in\mathbb{R}^d$,d为数据的维度。聚类数K是先验知识,第j个聚类中心用$\mu_j\in\mathbb{R}^d$来表示。$s_i\in\{1,2,\cdots,K\}$表示分配给聚类样本x_i的聚类指标。定义非线性映射$f_W:x_i\to z_i$和$g_W:z_i\to x_i'$,其中z_i是x_i在低维空间上的嵌入点,x_i'是x_i经过自动编码器重构后的样本。

在本文所采用的升维—聚类思想方法中,z_i为x_i在高维空间上的嵌入点,x_i'为x_i经过自动编码器重构后的样本,即编码过程是升维过程而重构过程是降维过程。

相比DEC算法先对自动编码器进行预训练再将自动编码器的解码器部分移除,IDEC不移除解码器部分,而是对聚类损失和重构损失进行联合优化,其损失函数定义为

$$L=L_r+\gamma L_c \tag{6}$$

其中,L_r为重构损失,L_c为聚类损失,$\gamma>0$为一个控制嵌入空间扭曲程度的系数。当$\gamma=1$和$\gamma\equiv0$时,就相当于DEC算法。

聚类损失计算方法与DEC算法一致,采用KL散度,详细计算过程和公式见文1.2。其中重构损失采用均方误差(mean-square error,MSE)来计算:

$$L_r=\sum_{i=1}^{n}\left\|x_i-g_W(z_i)\right\|_2^2 \tag{7}$$

采用小批量随机梯度下降(stochastic gradient descent,SGD)和反向传播进行优化,IDEC中,有3种参数需要优化或更新:自动编码器的权重,聚类中心和目标分布。

更新自动编码器的权重和聚类中心:固定目标分布P,聚类损失L_c相对于嵌入点z_i和聚类中心μ_j

的梯度可以由以下公式计算：

$$\frac{\partial L_c}{\partial z_i}=2\sum_{j=1}^{K}\left(1+\|z_i-\mu_j\|^2\right)^{-1}\left(p_{ij}-q_{ij}\right)\left(z_i-\mu_j\right) \tag{8}$$

$$\frac{\partial L_c}{\partial \mu_j}=2\sum_{i=1}^{n}\left(1+\|z_i-\mu_j\|^2\right)^{-1}\left(p_{ij}-q_{ij}\right)\left(z_i-\mu_j\right) \tag{9}$$

然后，给定一个 mini batch，样本数为 m，学习率为 λ，聚类中心更新为

$$\mu_j=\mu_j-\frac{\lambda}{m}\sum_{i=1}^{m}\frac{\partial L_c}{\partial \mu_j} \tag{10}$$

编码器和解码器的权重更新为

$$W'=\mathrm{W}'-\frac{\lambda}{m}\sum_{i=1}^{m}\frac{\partial L_r}{\partial W'} \tag{11}$$

$$W=W-\frac{\lambda}{m}\sum_{i=1}^{m}\left(\frac{\partial L_c}{\partial W}+\gamma\frac{\partial L_c}{\partial W}\right) \tag{12}$$

式(11)、(12)中，W'为编码器的权重；W为解码器的权重。为避免过拟合现象的发生，本文对损失函数进行正则化处理，在原始的损失函数后面加上一个 L1 正则化项，即全部权重 w 的绝对值之和，再乘以 λ/n。

更新目标分布 P：目标分布 P 作为“基本事实”软标签，但也取决于预测的软标签。因此，为了避免不稳定性，P 不应在每次迭代时仅使用一批数据进行更新（使用一小批样本更新自动编码器的权重称为迭代）。在实践中，每 T 次迭代使用所有嵌入点更新目标分布。更新目标分布时，分配给 x_i 的标签为

$$s_i=\arg\max_j q_{ij} \tag{13}$$

IDEC 算法运用于图像信息等高维数据时，在编码过程中对数据进行降维，便于特征提取，提高运算效率。而对于日负荷信息，由于其数据自身维度的限制，本文采用升维—聚类的改进方法，即在 IDEC 算法的编码过程中通过全连接层，根据数据本身的维度特点，将数据映射到更高的维度上，使其在更高的维度上进行上述深层特征提取和聚类过程。

1.4　聚类有效性评价指标

本文采用的所有日负荷数据均为无标签数据，数据的类别标签是没有提前给定的，所以在选择聚类评价指标时只能选择内部评价指标来评价聚类效果的好坏。本文选取戴维森堡丁指数（davies-bouldin index，DBI）、卡林斯基—哈拉巴斯指数（calinski-harabasz index，CH）和轮廓系数（silhouette coefficient，SC）这 3 个聚类指标进行定量分析。

1）DBI 指标，又称分类准确性指标，其计算公式为

$$D_{BI}=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\bar{C}_i+\bar{C}_j}{\|w_i-w_j\|_2}\right) \tag{14}$$

其中，$\bar{C}_i$ 为每类内样本到类中心的平均距离，代表了第 i 类中各样本的分散程度，$\bar{C}_j$ 同理。$\|w_i-w_j\|_2$ 为第 i 类与第 j 类的中心之间的距离。DBI 的含义是类内距离之和与类间距离的比值，值越小，聚类效果越好[13]。

2）CH 指标的计算公式为

$$C_H=\frac{S_{SB}}{S_{SW}}\frac{m-k}{k-1} \tag{15}$$

其中，m 为训练样本数，k 为类别数，S_{SB} 为类内平方误差和，用来度量类内的紧密度，S_{SW} 为类间平方误差和，用来度量类间的分离度。类间数据的平方和越小越好，类间的平方和越大越好，所以 CH 的值越大，聚类效果越好。

3）SC 指标的计算公式为

$$S_C=\frac{b(k)-a(k)}{\max\{a(k),b(k)\}} \tag{16}$$

式中，$a(k)$ 为样本 k 到同一类内其他样本的平均距离，$a(k)$ 越小，说明样本 k 越应该被分到该簇；$b(k)$ 为样本 k 到其他类样本的最小平均距离，$b(k)$ 越大表示样本 k 越不属于其他簇。S_C 的值应介于[−1，1]，越接近 1 表示样本 k 聚类越合理，越接近−1，表示样本 k 越应该被分到其他类中；越接近 0，表示样本应该在边界上。所有样本 S 的均值即为轮廓系数，该值越大，聚类效果越好。

2　曲线聚类流程及数据处理

2.1　日负荷曲线聚类的流程

采用改进的深度嵌入式聚类算法的日负荷曲线聚类的总体流程如图 3 所示，主要分为 5 个步骤。

1）数据预处理。本文采用 2021 年华中电网某地 110 kV 变电站的实测数据进行研究，需要对数据进行预处理，包括缺失数据和异常数据的剔除，数据归一化处理。

2）采用自编码器的编码器部分预处理后的数据进行降维或升维，提取潜在聚类特征。通过

K-means得到初始聚类结果。

3）将聚类损失和重建损失组合起来得到损失函数并对其进行优化，利用损失函数迭代来更新聚类结果。比较前两次的标签分布结果，若小于阈值则停止训练。

4）得到聚类结果并输出，对每一类的结果求均值得到各类曲线的聚类中心。

5）计算聚类效果评价指标值。

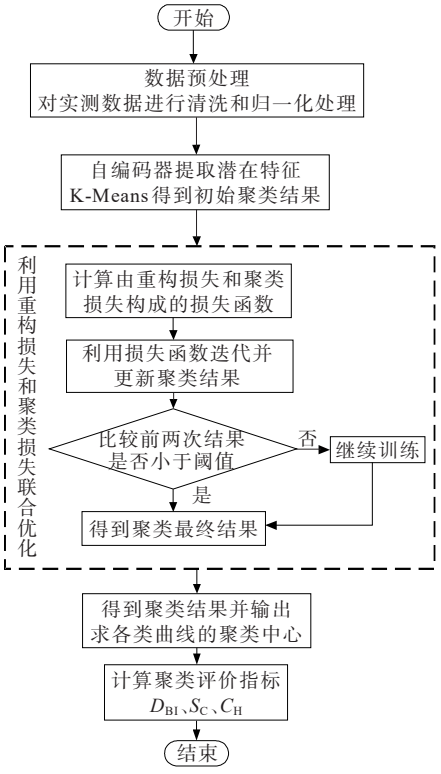


图3 日负荷曲线聚类流程

Figure 3 Clustering flow chart of daily load curve

2.2 数据预处理

数据的完整和准确性是聚类分析的前提，但在实际的测量和提取过程中不可避免地造成一些数据的缺失和错误，所以需要对数据集进行预处理。本文选用的数据是2021年多个地市电网的110 kV变电站实测的日负荷数据，该数据集以5 min为粒度，每天共288个采集点的数据。

首先对缺失和异常数据进行检测，对缺失数据的判断依据为数据记录中出现连续的15个负荷值为0或者缺失数据，数据记录中出现的30个负荷值为0或者缺失数据以及负荷记录中所有负荷值都一样的数据。而对突然升高或降低的数据即与前一数据点相比，升高或降低超过25%的数据，则认定为异常数据，需要对缺失数据和异常数据进行剔除[15]。

2.3 数据集生成

对预处理后的数据集进行进一步处理，原数据集为每5 min采集一次负荷数据，一天共288个采集点。为了对比升维后聚类，直接聚类和降维后聚类在不同维度数据集中的聚类效果，本文对原数据集进行人为精简处理，即在原始数据集中每隔2个数据点提取一个有效数据点，得到每15 min采集一次的数据集，一天共96个数据点；在原始数据集中每隔11个数据点提取一个有效数据点，得到每60 min采集一次的数据集，一天共24个数据点。将原数据集和处理后的2个数据集编号，原数据集每5 min采集的为1号数据集，每15 min采集的为2号数据集，每60 min采集的为3号数据集。

3 算例分析

研究选取2021年华中电网某地市110 kV变电站的实测日负荷数据，经处理得到3个采集频率分别为5、15、60 min的数据集，数据集含有1 092条日负荷数据，数据均选用3、4月工作日。PC配置为Intel(R) Core(TM) i5-8250U CPU/8.00 GHz RAM/GPU NVDIA GTX1080Ti。

为简化研究，聚类个数的确定以对数据集1采用K-means算法直接聚类的聚类评价指标为标准确定，通过对比分析DBI、SC和CH参数在不同聚类个数时的值，如图4所示。由图4可知，选取$K=3$作为本文的聚类个数。

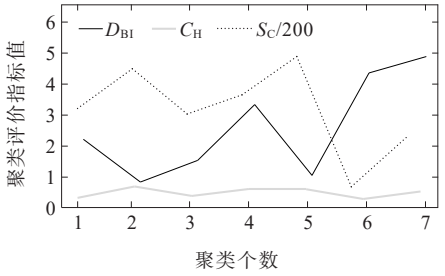


图4 不同聚类个数时D_{BI}、C_H和S_C值

Figure 4 D_{BI}, C_H and S_C values for different cluster numbers

3.1 IDEC算法的网络结构

为避免神经网络隐含层数量对结果造成影响，本文固定隐含层数量为3，其网络结构如表1所示。$x=\{24,96,288\}$。

表 1　网络结构与参数

Table 1　Network structure and parameters

类型	输入尺寸（升维\降维）	输出尺寸（升维\降维）	连接位置
Input	x*1	x*1	—
Encoder1	x*1\x*1	($2x$)*1\($x/2$)*1	Input
Encoder2	($2x$)*1\($x/2$)*1	($4x$)*1\($x/4$)*1	Encoder1
Encoder3	($4x$)*1\($x/4$)*1	($8x$)*1\($x/8$)*1	Encoder2
Embedding	($8x$)*1\($x/8$)*1	($8x$)*1\($x/8$)*1	Encoder3
Decoder3	($8x$)*1\($x/8$)*1	($4x$)*1\($x/4$)*1	Embedding
Decoder2	($4x$)*1\($x/4$)*1	($2x$)*1\($x/2$)*1	Decoder3
Decoder1	($2x$)*1\($x/2$)*1	x*1\x*1	Decoder2
Clustering	($8x$)*1\($x/8$)*1	3*1	Embedding

整个模型分为两部分进行训练，第 1 部分为预训练部分，训练目标是使得嵌入点是输入样本的有效特征表示，采用 Adam 优化器，批尺寸设置为 256，迭代次数为 500 次，通过最小化重构损失 L_r 来对部分参数进行调优；第 2 部分为优化聚类过程，训练目标是对日负荷数据进行更为准确的聚类。学习率设置为 0.001，优化器采用 Adam 优化器，通过损失函数 $L=L_r+\gamma L_c$ 来迭代微调网络。其最大迭代次数不超过 20 000，每间隔 140 次更新一次迭代结果，比较 2 次结果，小于阈值 0.001 则中止迭代，输出最终的聚类结果。

3.2　训练结果

3.2.1　IDEC 算法有效性研究

当前基于人工智能的聚类算法均采用降维—聚类和直接聚类的方法，为了研究 IDEC 的有效性，在本节中也采用先降维再聚类的方法。

使用降维—改进深度嵌入式聚类算法对数据集 1 进行聚类分析，提取典型日负荷曲线，其结果如图 5 所示。从 00:00—23:55，每 5 min 提取一次特征点，共 288 个时段，该算法将 1 092 条日负荷曲线分成三类。

各类负荷曲线的聚类中心如图 5 所示，可以根据聚类中心对各类站点的运行状况进行分析和挖掘。

由图 5 可知，第 1 类站点应该以居民负荷为主，负荷总量较低，且在早晨、中午和晚上有 3 个明显的高峰时段；第 2 类应该以商业和工业站点为主，在白天时段负荷较高且较为稳定；第 3 类与第 2 类类似，也是在白天时段较为稳定，夜间几乎没有负荷，但是负荷远低于第 2 类，应该为学校等公共服务类站点。

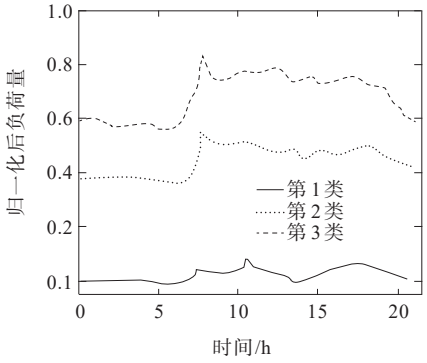


图 5　IDEC 聚类结果

Figure 5　Clustering results of IDEC

为验证本文改进的深度嵌入式聚类算法的有效性，选取聚类分析指标 DBI、CH 和 SC 对其进行定量分析，其中 DBI 值越小聚类效果越好，CH 和 SC 值越大聚类效果越好。通过对 K-means、PCA＋K-means、DEC 和目标重构—深度嵌入式聚类算法 4 种方法对同一数据集的指标进行对比分析，分析结果如表 2 所示。

表 2　不同算法聚类指标对比分析

Table 2　Comparative analysis of clustering indexes for different algorithms

方法	D_{BI}	C_H	S_C
K-Means	1.267 3	890.538 7	0.543 8
PCA＋K-Means	0.958 4	967.859 6	0.653 7
DEC	0.684 7	1 289.673 5	0.721 3
IDEC	0.503 0	1 345.923 4	0.765 8

由表 2 可知，IDEC 相比其他 3 种算法，聚类效果均有显著提升。相较于 K-means 算法，IDEC 的 DBI 降低了约 60.31%，CH 和 SC 分别提高了约 51.14% 和 40.82%。相比于传统降维加聚类算法 PCA＋K-means，IDEC 的各项指标也均有所改进，其中 DBI 下降了 46.27%，CH 和 SC 分别上升了 39.06% 和 17.15%。深度嵌入式聚类算法（deep embedded clustering，DEC）相比 K-means 和 PCA＋K-means 算法，聚类效果已经有了明显提高，但是 IDEC 算法的效果相比 DEC 算法更加优越。相比 DEC 算法，IDEC 算法的 DBI 指标下降了 26.54%，CH 和 SC 分别提高了 4.36% 和 7.66%。由此可以看出，本文所提算法能够有效提高聚类的质量。

3.2.2　升维—聚类方法的有效性研究

为了对比日负荷数据在降维—聚类、直接聚类

和升维—聚类这3种方法下的聚类效果，本文采用改进的深度嵌入式聚类算法进行降维—聚类和升维—聚类操作，其网络结构参数如文3.1所示。选用K-means算法对日负荷数据进行直接聚类。分别对3个数据集采用降维后聚类、升维后聚类和直接聚类，其聚类评价指标和运行时间结果如表3～5所示。数据集3在3种聚类方法下所得结果如图6所示。

表3 数据集1采用不同策略时的聚类评价指标值和运行时间

Table 3 Clustering indexes and running time of dataset 1 with different strategies

策略	D_{BI}	C_H	S_C	运行时间/s
降维—聚类	0.503 0	1 345.923 4	0.765 8	2.835 5
直接聚类	1.267 3	890.538 7	0.543 8	1.926 5
升维—聚类	0.489 4	1 431.655 5	0.755 4	3.167 8

表4 数据集2采用不同策略时的聚类评价指标值和运行时间

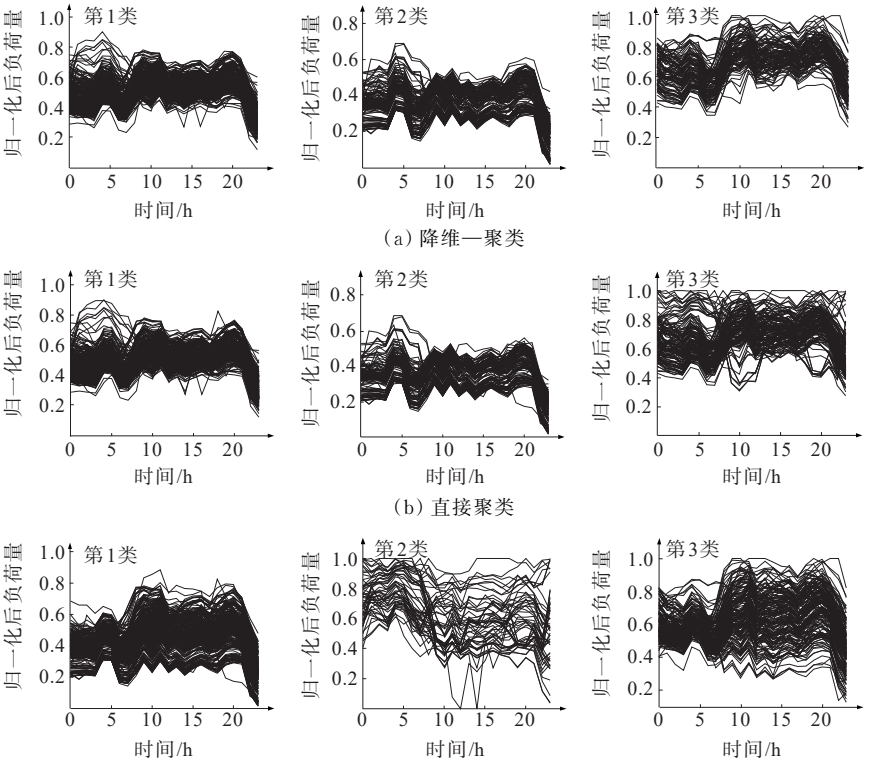
Table 4 Clustering indexes and running time of dataset 2 with different strategies

策略	D_{BI}	C_H	S_C	运行时间/s
降维—聚类	0.618 1	1 238.299 8	0.633 3	2.347 3
直接聚类	1.253 4	970.789 6	0.549 7	1.794 5
升维—聚类	0.607 5	1 419.742 4	0.735 6	2.895 3

表5 数据集3采用不同策略时的聚类评价指标值和运行时间

Table 5 Clustering indexes and running time of dataset 3 with different strategies

策略	D_{BI}	C_H	S_C	运行时间/s
降维—聚类	0.988 7	789.778 6	0.458 7	1.908 9
直接聚类	1.009 8	798.563 8	0.466 6	1.098 3
升维—聚类	0.850 4	1 421.016 7	0.587 4	2.035 4



(a) 降维—聚类

(b) 直接聚类

(c) 升维—聚类

图6 数据集3的聚类结果

Figure 6 Clustering results for dataset 3

由表3～5、图6可知，对于3个数据集，采用升维—聚类的方法均能提高聚类质量，且对于数据集3(采集间隔为60 min)的优化效果最为明显，其DBI值相对于降维—聚类和直接聚类分别下降了13.98%和15.78%，CH值分别提高了79.93%和77.95%，SC值分别提高了28.06%和26.50%。但是相对于直接聚类和降维—聚类方法，升维—聚类会耗费更长的时间。

4 结语

针对当前基于人工智能的聚类算法难以提取数据深层特征的问题，本文提出采用改进深度嵌入式聚类算法，利用神经网络提取负荷曲线的深层特

征，通过与 K-means、KPCA＋K-means 以及 DEC 算法的对比，验证了该算法的有效性。本文提出一种升维—聚类的改进方法，利用神经网络将日负荷数据映射到更高的维度上再进行特征提取和聚类，通过与降维—聚类和直接聚类这 2 种方法的对比，得出该改进方法在数据相对匮乏的情况下，能有效提高聚类质量，但是会耗费更长的时间。在数据条件较差，且计算资源较丰富的时候，采用先升维后聚类的策略效果更佳。

参考文献：

[1] 李笑彤,宋宝同,吕风波,等.基于负荷数据聚类的充电站储能容量规划方法[J].电网与清洁能源,2021,37(1):90-96.
LI Xiaotong, SONG Baotong, LÜ Fengbo, et al. An energy storage capacity planning method of charging station based on load data clustering[J]. Power System and Clean Energy,2021,37(1):90-96.

[2] 楚帅,葛维春,李音璇,等.含海水淡化负荷的可再生能源消纳技术研究综述[J].智慧电力,2021,49(11):14-23.
CHU Shuai, GE Weichun, LI Yinxuan, et al. Review on renewable energy integration technology with seawater desalination load[J].Smart Power,2021,49(11):14-23.

[3] 吴峰,鲍颜红,周华,等.满足大型城市电网供电需求的在线预防控制方法[J].中国电力,2021,54(1):159-166.
WU Feng, BAO Yanhong, ZHOU Hua, et al. An on-line preventive control method for meeting the power supply demand of large-scale urban power grid[J].Electric Power, 2021,54(1):159-166.

[4] 栾乐,马智远,莫文雄,等.综合考虑供用电双方需求的优质电力用户分类方法[J].电力科学与技术学报,2021,36(6):171-181.
LUAN Le, MA Zhiyuan, MO Wewnxiong, et al. A premium user classification method considering the demand of both power company and electricity user[J]. Journal of Electric Power Science and Technology,2021,36(6):171-181.

[5] 陶鹏,张洋瑞,李梦宇,等.基于Graph模型的海量用电数据并行聚类分析[J].电力科学与技术学报,2020,35(6):144-151.
TAO Peng, ZHANG Yangrui, LI Mengyu, et al. Parallel clustering analysis for power consumption data based on graph model[J]. Journal of Electric Power Science and Technology,2020,35(6):144-151.

[6] 张帆,杨翾,商佳宜,等.考虑负荷损失最小的配网孤岛划分策略研究[J].高压电器,2021,57(4):181-188.
ZHANG Fan, YANG Xuan, SHANG Jiayi, et al. Study on island partition strategy of distribution network considering minimum load loss[J]. High Voltage Apparatus,2021,57(4):181-188.

[7] 姚黄金,雷霞,付鑫权,等.基于改进自适应密度峰值算法的日负荷曲线聚类分析[J].电力系统保护与控制,2022,50(3):121-130.
YAO Huangjin, LEI Xia, FU Xinquan, et al. Cluster analysis of daily load curves based on an improved self-adaptive density peak clustering algorithm[J]. Power System Protection and Control,2022,50(3):121-130.

[8] 张慧波,王守相,赵倩宇,等.考虑数据不均衡的居民用户负荷曲线分类方法[J].电力工程技术,2022,41(3):186-193.
ZHANG Huibo, WANG Shouxiang, ZHAO Qianyu, et al. Residential user load curve classification method considering data imbalance[J]. Electric Power Engineering Technology, 2022, 41(3):186-193.

[9] WANG Y L, LI L, YANG Q M. Application of clustering technique to electricity customer classification for load forecasting[C]//Proceedings of the 2015 IEEE International Conference on Information and Automation, Lijiang, China: IEEE,2015.

[10] 张红斌,贺仁睦,刘应梅.基于KOHONEN神经网络的电力系统负荷动特性聚类与综合[J].中国电机工程学报,2003,23(5):2-6+44.
ZHANG Hongbin, HE Renmu, LIU Yingmei. The characteristics clustering and synthesis of electric dynamic loads based on KOHONEN neural network[J]. Proceedings of the CSEE,2003,23(5):2-6+44.

[11] 王剑锋,倪家明,王旭东,等.基于AE-MFCM技术的电力用户响应特性分析方法[J].电力系统及其自动化学报,2021,33(4):47-54.
WANG Jianfeng, NI Jiaming, WANG Xudong, et al. Analysis method for power user response characteristics based on AE-MFCM technology[J]. Proceedings of the CSU-EPSA,2021,33(4):47-54.

[12] 黄冬梅,林孝镶,胡安铎等.基于卷积自编码器的日负荷深度嵌入聚类方法[J].电力建设,2021,42(1):132-138.
HUANG Dongmei, LIN Xiaoxiang, HU Anduo, et al. Deep embedding clustering method for daily load based on convolutional auto-encoder[J]. Electric Power Construction, 2021,42(1):132-138.

[13] GUO X F, GAO L, LIU X W, et al. Improved deep embedded clustering with local structure preservation[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne Australia,2017.

[14] XIE J Y, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//Proceedings of the 33rd International Conference on Machine Learning, New York, USA,2016.

[15] 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):1-5.
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.