DOI: 10.13382/j.jemi.B2003008

基于 BIC 准则和加权皮尔逊距离的居民负荷模式精细识别及预测*

夏 飞[1] 张 洁[1] 张 浩[2] 陆剑峰[2]
(1. 上海电力大学 自动化工程学院 上海 200090; 2. 同济大学 电子与信息工程学院 上海 201804)

摘 要:针对居民日用电负荷的聚类分析和预测问题提出了一种基于居民用电负荷模式精细分类的预测框架。为了提高用于聚类分析的特征质量,首先基于贝叶斯信息准则(BIC)实现特征筛选。然后,采用基于加权皮尔逊距离的密度峰值法实现居民用电负荷曲线形态的准确识别。接下来,通过融合激活函数的方法对长短期记忆(LSTM)预测网络进行改进。最后,利用改进后的 LSTM 网络对精细分类的居民用电负荷模式进行预测。实验结果表明,根据所提出的方法得到的预测误差指标为平均绝对百分误差(MAPE),*MAPE*=6.6792%,提高了负荷预测质量,在居民用电负荷预测中具有较好的效果。

关键词: BIC 特征提取;加权皮尔逊距离;密度峰值法;改进的 LSTM 网络;精细分类;居民负荷预测
中图分类号: TM714;TN0 **文献标识码**: A **国家标准学科分类代码**: 470.40

Fine recognition and prediction of resident load pattern based on BIC criterion and weighted Pearson distance

Xia Fei[1] Zhang Jie[1] Zhang Hao[2] Lu Jianfeng[2]
(1. College of Automation Engineering, Shanghai University of Electric Power, Shanghai 200090, China;
2. School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: Aiming at the problem of clustering analysis and prediction of residential daily electricity load, a prediction framework based on the fine classification of residential power load patterns was proposed. In order to improve the quality of features used for cluster analysis, feature selection was first implemented based on BIC criteria. Then, the CFSFDP algorithm based on weighted pearson distance is used to realize the accurate identification of the shape of the residential electricity load curve. Next, the LSTM prediction network is improved by a fusion activation function method. Finally, the improved LSTM network is used to predict the finely classified residential power load patterns. The experimental results show that the forecast error index obtained by the method proposed is *MAPE*=6.6792%, which improves the quality of load forecasting and has a good effect in the forecast of residential electricity load.

Keywords: BIC feature extraction; weighted Pearson distance; CFSFDP algorithm; improved LSTM network; fine classification; resident load forecast

0 引 言

随着智能电网建设规模的不断扩大、电力系统中智能电表的广泛应用,电力公司相关部门积累了大量的用电数据,通过相应的数据挖掘技术可以挖掘出信息中隐藏的宝贵信息[1]。目前,大量研究通过聚类分析的方法实现用电负荷曲线的模式识别。

对居民用户的用电数据进行聚类,可以分析得到不同居民的用电习惯,从而总结出用电规律以及用电的特征等情况,为进一步的预测做打算。因此,对用户的用电负荷曲线进行聚类分析,挖掘其用电行为,已经成为智能用电大数据挖掘的关键[2]。

负荷曲线聚类对负荷预测[3]、电网规划[4]、需求侧响应[5]等有重要意义,有助于挖掘出用电数据中隐藏的重要信息。但是海量的用电负荷数据为数据处理时间以及

收稿日期: 2020-03-17 Received Date: 2020-03-17
*基金项目:自然科学基金重大项目(71690234)、政府间国际科技创新合作重点专项(2017YFE0100900)、上海市科委创新项目(19DZ1206800)资助

计算复杂度带来了困难,因此聚类前需要先对数据进行降维处理,既可以降低复杂度,又提取了相应的特征。

文献[6]利用 SAX 算法对负荷曲线进行降维并提取特征,然后采用 AP 聚类算法对负荷曲线进行聚类,最后基于聚类结果,对各类用户的用电行为以及需求响应潜力进行分析。文献[7]针对电力负荷曲线的特征,对数据集进行降维处理,然后利用降维后的数据集进行聚类分析。但是上述文献进行特征提取或降维处理时,无法同时考虑数据的实用性和冗余性,为了同时满足实用性和冗余性,本文提出一种基于贝叶斯信息准则(Bayesian information criterions,BIC)的特征选择方法。一方面可以在降低特征维度的同时保证了模型的准确率(即实用性)。另一方面为了避免出现过拟合,将特征维度作为惩罚因子,从而也实现了冗余性的要求[8]。

根据 BIC 准则实现特征选取后,将对居民用电负荷数据行聚类分析。聚类分析就是把数据集中的样本按照相似度进行归类,相似度高的样本归为一类,相似度低的样本则属于不同的类别[9]。因此,相似性度量函数的选取是聚类算法的核心之一,只有选取了合适的相似性度量函数,样本的聚类才能更加准确。目前,对于负荷曲线的聚类算法研究中,大多采用欧氏距离为相似性度量[10-14],欧氏距离是根据几何平均距离来衡量样本间的相似性,它的缺点是不能反映曲线形态及趋势的相似性。皮尔逊距离不同于几何平均距离,它强调两变量之间的变化情况,因此本文提出了一种加权的皮尔逊距离度量方法,在一定程度上可以反映负荷曲线之间变化趋势的相似性。

在聚类方面,经典 K-means[9,15]算法对用电高峰的识别较为准确,但是对于相同用电水平下的不同用电模式无法精准识别,因此需要更加精细的聚类。密度峰值算法[16]于 2014 年提出,很好的解决了传统聚类算法的不足,适用于大规模数据集,可以快速发现任意形状的类簇,同时也可以检测出噪声点和离群点,并且进行簇类分配时不需要进行迭代,聚类效果稳定。因此,本文采用密度峰值算法[17-20]对居民用电负荷进行精细聚类。同时也为用户负荷预测提供了准确的用电模式。

最后,本文采用长短期记忆(long short-term memory,LSTM)深度神经网络对居民负荷进行预测。未改进的 LSTM 长短期记忆神经网络模型学习速度太慢(需要进行复杂的矩阵运算)、有可能陷入局部极值、梯度随着记忆网络的深度循环而逐渐消失等。基于上述问题,本文提出了一种改进 LSTM 的居民负荷预测模型,通过融合激活函数来适应负荷预测中的长时间梯度消失问题。最终,建立了适用于居民负荷预测的高精度 LSTM 网络。

综上所述,本文提出了一种针对居民用电负荷模式的精细分类及预测框架。首先,利用 BIC 准则进行特征筛选。然后采用基于加权皮尔逊距离的密度峰值算法对居民用电负荷数据进行分类,得到较为精准的用电模式。最后,采用改进的 LSTM 网络对居民用电负荷进行预测。该方法的框架如图 1 所示。

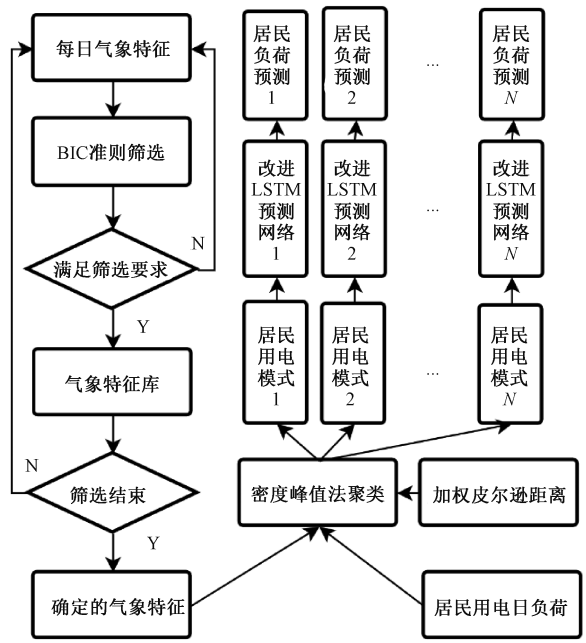

图 1 居民用电负荷精细识别及预测框架

Fig. 1 Framework of fine-grained identification and prediction for residential electricity load

1 基于 BIC 准则的特征选择

在进行居民用电负荷分析时,除了用电日负荷数据外,如果能结合每日气象特征将会提高分析的精确性。但如果不加筛选,将所有气象特征用于聚类分析,不仅增加了聚类时间,而且聚类效果不一定好。因此,本文采用了 BIC 信息量准则对气象特征进行筛选,选择满足一定条件的气象特征进入特征库。通过特征选择可以有效实现降低维度的目的,可以减少数据处理时间,既降低了复杂度,又提取了相应的特征,从而达到更好的聚类效果。

1.1 信息量准则

最早提出的基于信息熵的信息准则为赤池信息量准则(akaike information criterion,AIC)[21],BIC 是在 AIC 准则的基础上增加了惩罚力度[22-23]。

$$BIC = k\ln n - 2\ln\varphi \tag{1}$$

式中:k 为模型参数的个数;φ 为似然函数,n 为样本的数量,$k\ln n$ 为惩罚量。由式(1)可以看到,BIC 准则的惩罚项随着样本容量变化而改变,有效改善了 AIC 准则惩罚因子与样本容量无关的缺陷。BIC 值越小,说明模型的质量越好。

1.2　聚类评价指标

本文通过 BIC 准则进行特征选择时,利用聚类的效果确定各个特征的重要性。评价效果一般通过计算同一簇内的聚合度与不同簇间的分离度来评估聚类效果的好坏[24],即"簇内越紧密、簇间越分离,聚类效果越好"。本文选用轮廓系数(silhouette coefficient, SC)指标来进行聚类评价。

假设数据集 X 被分为 J 个类簇: $C=\{C_1,C_2,\cdots,C_J\}$,则数据集 X 中的某一样本 i 的 SC 指标定义为:

$$I_{SC}=\frac{b(x_i)-a(x_i)}{\max\{a(x_i),b(x_i)\}} \tag{2}$$

式中: $a(x_i)$ 表示 x_i 到同一簇内其他对象之间的平均距离,按照式(3)进行计算; $b(x_i)$ 表示 x_i 到其余类簇的最小平均距离,按照式(4)进行计算。

$$a(x_i)=\frac{\sum_{x_i,x_i'\in C_s,x_i\neq x_i'} dist(x_i,x_i')}{|C_s-1|} \tag{3}$$

$$b(x_i)=\min_{C_t:1\leqslant t\leqslant s,t\neq s}\left\{\frac{\sum_{x_i\notin C_t,x_i'\in C_t} dist(x_i,x_i')}{|C_t|}\right\} \tag{4}$$

式中: $a(x_i)$ 为所属类簇的内聚度,其值越小,说明簇越紧凑; $b(x_i)$ 为所属类簇与其他类簇的分离度,其值越大,说明簇与簇间越分离。由式(2)计算出所有样本 I_{SC} 的均值,即可得到数据集 X 的 I_{SC},其取值范围为[-1,1],该值越大说明聚类效果越好。

1.3　基于 BIC 准则的特征聚类

式(1)中的 k 为聚类模型中聚类簇的个数,似然函数 φ 可以表示为:

$$\varphi=\frac{SSE}{n} \tag{5}$$

式(5)中的剩余平方和(sum of squares for error, SSE)在本模型中表示为:

$$SSE=\sum_{i=1}^{n}(SC-SC^*)^2 \tag{6}$$

式中: SC 与 SC^* 分别表示聚类评价指标的最优值及实际输出的评价指标值。

基于聚类评价指标的最优值及实际输出的评价指标值的剩余平方和为性能指标。本文筛选要求为 SSE 的值不大于 8 的特征,从而实现特征选择。

2　基于加权皮尔逊距离的密度峰值法聚类

为了弥补基于距离的 K-means 特征聚类无法分辨相同用电量下不同用电模式的缺点,提高聚类的细粒度,本文提出采用加权皮尔逊距离的密度峰值法对居民用电负荷进行更精细化的分类。

2.1　加权皮尔逊距离

假设负荷曲线 $L_1=\{x_i\}$, $i=1,2,\cdots,n$ 和 $L_2=\{y_i\}$, $i=1,2,\cdots,n$。则它们的皮尔逊系数和皮尔逊距离为:

$$\rho(L_1,L_2)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{7}$$

$$D(L_1,L_2)=1-\rho(L_1,L_2) \tag{8}$$

与皮尔逊距离相似,加权皮尔逊距离通过调整各维分量 $(x_i-\bar{x})(y_i-\bar{y})$ 的权值大小来控制各维的贡献力度。

假设权值矩阵为 $\boldsymbol{W}=\{w_i\}$, $i=1,2,\cdots,n$。则它们的加权皮尔逊系数及加权皮尔逊距离的计算过程为:

$$m(L_1)=\frac{\sum_{i=1}^{n}w_ix_i}{\sum_{i=1}^{n}w_i},m(L_2)=\frac{\sum_{i=1}^{n}w_iy_i}{\sum_{i=1}^{n}w_i} \tag{9}$$

$$\operatorname{cov}(L_1,L_2,W)=\frac{\sum_{i=1}^{n}w_i(x_i-m(L_1))(y_i-m(L_2))}{\sum_{i=1}^{n}w_i} \tag{10}$$

$$\rho=\frac{\operatorname{cov}(L_1,L_2,W)}{\sqrt{\operatorname{cov}(L_1,L_1,W)\operatorname{cov}(L_2,L_2,W)}} \tag{11}$$

$$D(L_1,L_2)=1-\rho(L_1,L_2) \tag{12}$$

皮尔逊距离不同于几何平均距离,它强调两变量之间的变化情况,因此在一定程度上可以反映负荷曲线之间变化趋势的相似性,其值越大,则说明曲线形态越相似。

2.2　密度峰值法

密度峰值法(clustering by fast search and find of density peaks, CFSFDP)[16]很好的解决了传统聚类算法的不足,适用于大规模数据的聚类,对用电负荷数据可以快速实现聚类分析[17-19]。

该算法主要有两个需要计算的量,局部密度 ρ_i 及与高密度点之间的距离 δ_i。

1)局部密度 ρ_i

为了减小截断距离 d_c 对聚类结果的影响,采用高斯核函数对局部密度进行改进[16]:

$$\rho_i=\sum_{j\neq i}\exp\left(-\left(\frac{d_{ij}}{d_c}\right)^2\right) \tag{13}$$

式中: d_{ij} 为两点间的距离,这里用皮尔逊距离及加权皮尔逊距离,计算公式如式(7)~(12)。选取的截断距离 d_c 应保证每个数据点的平均邻居个数约为数据点总数的 1%~2%[16]。原算法中的局部密度 ρ_i 表示数据集 X 中与点 x_i 的距离小于截断距离 d_c 的样本点的个数,是离散

的。改进算法中采用的高斯核是一个连续的值,可以减小不同数据点具有相同局部密度值的可能性。

2)与高密度点间的距离 δ_i

定义每个样本点 x_i 到更高密度点间的最小距离 δ_i:

$$\delta_i = \min_{j:\rho_j>\rho_i}(d_{ij}) \tag{14}$$

式(14)中,对于数据集中局部密度最大的样本点 x_i,距离 δ_i 令为:

$$\delta_i = \max_j(d_{ij}) \tag{15}$$

通过选择加权皮尔逊距离,再结合 CFSFDP 算法,可以在原有聚类算法的基础上,获得更为精确的聚类结果,并有助于提高用户用电预测的质量。

3 基于改进 LSTM 的居民负荷预测

本文采用 LSTM 进行负荷预测。为了改进 LSTM 网络中的长时间梯度消失问题,本文提出了融合激活函数对 LSTM 网络进行改进。

3.1 融合激活函数

激活函数的主要作用是提供网络的非线性建模能力。如果没有激活函数,那么该网络仅能够表达线性映射,此时即便有再多的隐藏层,其整个网络跟单层神经网络也是等价的。因此也可以认为,只有加入了激活函数之后,深度神经网络才具备了分层的非线性映射学习能力。而改善梯度消失主要方法即为改善激活函数的饱和特性。

ReLU 函数是一种后来才出现的激活函数,如下:

$$f(x)=\begin{cases}x, x\geqslant 0\\ 0, x<0\end{cases} \tag{16}$$

可以看到,当 $x<0$ 时,ReLU 硬饱和,而当 $x>0$ 时,则不存在饱和问题。所以,ReLU 能够在 $x>0$ 时保持梯度不衰减,从而缓解梯度消失问题。这能够直接以监督的方式训练深度神经网络,而无需依赖无监督的逐层预训练。然而,随着训练的推进,部分输入会落入硬饱和区,即 $x<0$ 时,导致对应权重无法更新,这会给模型的预测精度带来很大的误差。此外 ReLU 函数的输出均大于零,使得输出不是零均值输出,这会导致后一层的神经元将得到上一层输出的非零均值信号作为输入,即零点漂移。零点漂移和硬饱和误差会共同影响网络的收敛性。

$$f(n)=\begin{cases}x, x\geqslant 0\\ \tanh x, x<0\end{cases} \tag{17}$$

式(17)融合激活函数融合了双曲正弦函数函数和 ReLU 函数的各自优点,左侧具有软饱和性,右侧无饱和性。左侧软饱和能够让融合激活函数对输入更鲁棒,而右侧线性部分使得其能够缓解梯度消失问题。

$$f'(n)=\begin{cases}1, x\geqslant 0\\ 1-\tanh x^2, x<0\end{cases} \tag{18}$$

式(18)为式(17)融合激活函数的导函数。可以看出,当 $x>0$ 时,导函数为人工神经网络保持了梯度的逐渐递增;而 $x<0$ 时,又对外部不同的输入进行筛选,具有自适应的特点。且融合激活函数的输出均值接近于 0,可有效防止零点漂移,所以收敛速度更快。

3.2 预测评价指标

预测误差采用平均绝对百分误差(mean absolute percent error, MAPE)和均方根误差(root mean square error, RMSE)两项指标评价,定义如下:

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i}\times 100\% \tag{19}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{20}$$

式中:n 为测试样本的数量,y_i 表示第 i 个测试样本的负荷真实值,$\hat{y}_i$ 表示第 i 个样本的负荷预测值。

4 居民用电负荷分析

本文选用某市一居民小区 1 年内的用电负荷及天气数据进行分析。其中,居民用电负荷每 1 h 采集 1 次,1 d 共采集 24 个数据点。气象数据中包括最高温、最低温、平均温、压强、湿度、风向、雨水和风速。本文算例在单台 CPU 为 2.6 GHz,内存为 16 GB,操作系统 64 位的个人计算机上完成,使用 MATLAB R2018a 进行算法测试。

4.1 特征选择

首先对最高温、最低温、平均温、压强、湿度、风向、风速和降雨量等 8 个气象特征进行特征筛选。根据第 1 节提出的方法,分别计算得到各个气象特征的 BIC 值,如表 1 所示。同时作为对比,也采用相关性分析的方法计算了各个气象特征的相关系数。

由表 1 可以确定,通过 BIC 算法计算得到的各指标重要性排序为最高温>最低温>平均温>压强>湿度>风向>雨水>风速。而用相关性分析得到的各指标的重要性排序为最高温>平均温>最低温>压强>风向>湿度>雨水>风速。按照相关系数>0.15 进行特征选择的话,会

表 1 特征指标的 BIC 值及相关系数计算结果

Table 1 BIC value and correlation coefficient calculation results of characteristic indicators

	最高温	最低温	平均温	压强	湿度	风向	降雨量	风速
BIC	6.873 0	6.874 4	7.205 6	7.442 0	7.716 5	8.189 3	9.694 7	9.702 8
相关系数	0.281 7	0.277 9	0.280 0	0.213 3	0.051 0	0.181 3	0.042 7	0.005 7

选择最高温、平均温、最低温、压强和风向作为气象特征。由表 1 可知,最低温的 BIC 值比平均温的 BIC 值更低,应该优先选择。这更符合负荷分析中的实际情况,即最高温和最低温比平均温对居民日负荷影响更大。根据 $BIC \leqslant 8$ 的筛选原则,最终选择了最高温、最低温、平均温、压强和湿度作为居民用电分析的气象特征。显然,用 BIC 值更优的湿度特征替换了相关系数更佳的风向特征更加能够体现对负荷的影响。

4.2　模式聚类

通过选取的特征值,接下来基于加权皮尔逊距离的密度峰值法对居民用电日负荷进行分类。本文一共选取了某居民小区 300 d 的居民用电日负荷曲线进行聚类。为了对比聚类效果,还采用了文献[9]基于欧氏距离的 K-means 聚类方法、文献[15]基于皮尔逊距离的 K-means 聚类方法和未加权皮尔逊距离的密度峰值法进行聚类。为了更好地展现聚类效果,本文选取了 8 条日负荷曲线,如图 2 所示。对图 2 的 8 条曲线依次采用 1~8 进行编号。

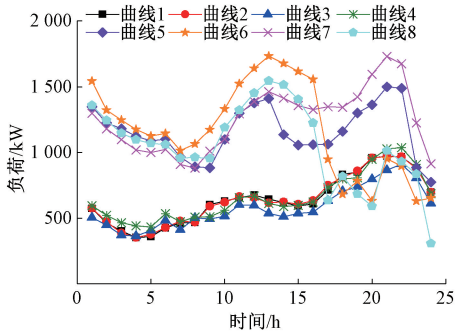

图 2　八条日负荷曲线

Fig. 2　Eight daily load curves

首先,采用文献[9]基于欧氏距离的 K-means 聚类方法对以上日负荷曲线进行分类,分类结果如表 2 所示。

表 2　基于欧氏距离的 K-means 法聚类结果

Table 2　K-means clustering results based on Euclidean distance

曲线编号	各曲线到聚类中心的距离		所属类别
	1	2	
1	**108.555**	3 072.97	1
2	**112.287**	3 066.73	1
3	**581.998**	2 719.87	1
4	**540.259**	2 755.39	1
5	2 692.24	**839.395**	2
6	3 410.19	**2 037.82**	2
7	3 046.92	**100.479**	2
8	2 830.44	**1 983.56**	2

从表 2 可以看出,本文选择的 8 条曲线根据到聚类中心的距离远近明显地分成了两类。具体分类结果如图 3 所示。从图 3 可以看到,类别 1 的 4 条曲线形态基本一致。而类别 2 的曲线虽然通过欧氏距离计算分在了一起,但在形态上则具有明显的差异性。进一步观察,发现类别 1 的曲线虽然形态上基本一致,具有明显的单峰特征,但不同曲线峰值处对应的峰值仍然有差异。

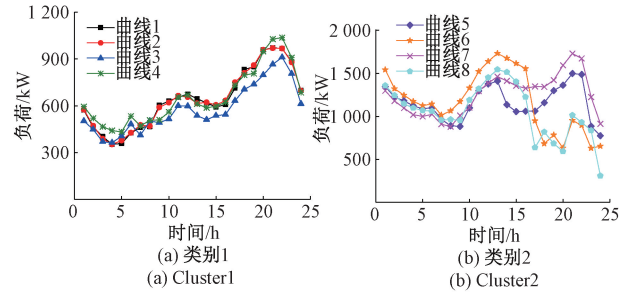

图 3　基于欧氏距离的 K-means 法聚类曲线

Fig. 3　K-means clustering curves based on Euclidean distance

由此可见,采用欧氏距离的 K-means 聚类方法不仅不能将具有明显差异的日用电负荷区分,对于不同负荷大小的同一形态用电负荷曲线也无法区分。因此需要更加精细的居民用电模式聚类方法。

然后采用文献[15]基于皮尔逊距离的 K-means 聚类方法同样对以上日负荷曲线进行分类,分类结果如表 3 所示。

表 3　基于皮尔逊距离的 K-means 法聚类结果

Table 3　K-means clustering results based on Pearson distance

曲线编号	各曲线到聚类中心的距离			所属类别
	1	2	3	
1	**92.380 4**	554.240	3 070.41	1
2	**84.063 1**	542.855	3 064.15	1
3	**83.791 0**	582.129	2 715.85	1
4	**537.520**	2 755.39	2 752.30	1
5	2 709.95	**839.978**	2 433.52	2
6	3 415.50	3 373.91	**2 040.75**	3
7	3 047.41	**100.876**	2 745.67	2
8	2 842.17	2 810.77	**1 983.91**	3

从表 3 可以看出,此时这 8 条曲线根据到聚类中心的距离远近被分成了 3 类。具体分类结果如图 4 所示。从图 4 可以看到,之前属于第 2 类的 4 条曲线被区分开来,此时曲线 5 和曲线 7 分为了第 2 类,曲线 6 和曲线 8 分为了第 3 类。

由此可见,采用皮尔逊距离的 K-means 聚类方法对于不同形态用电曲线的识别要优于欧氏距离,但是对于第 1 类中形态差别不太明显的 4 条曲线无法区分出来,还需要进一步的识别。

接下来,采用基于皮尔逊距离的密度峰值法聚类,聚类结果如表 4 所示。从表 4 可以看到,这次的聚类结果

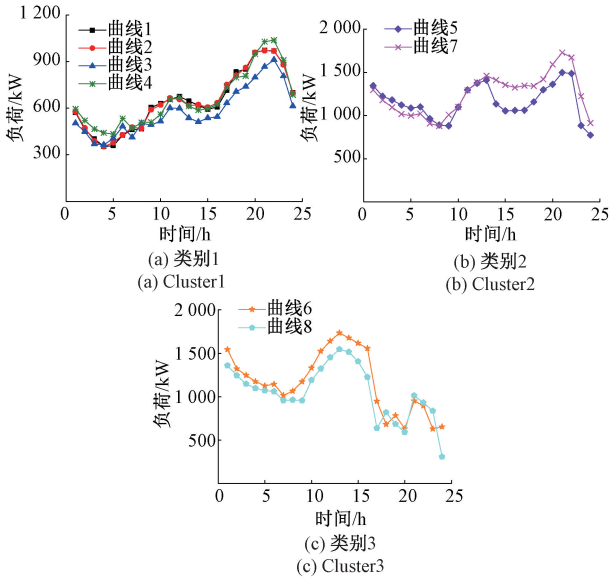

图 4 基于皮尔逊距离的 K-means 法聚类曲线

Fig. 4 K-means clustering curves based on Pearson distance

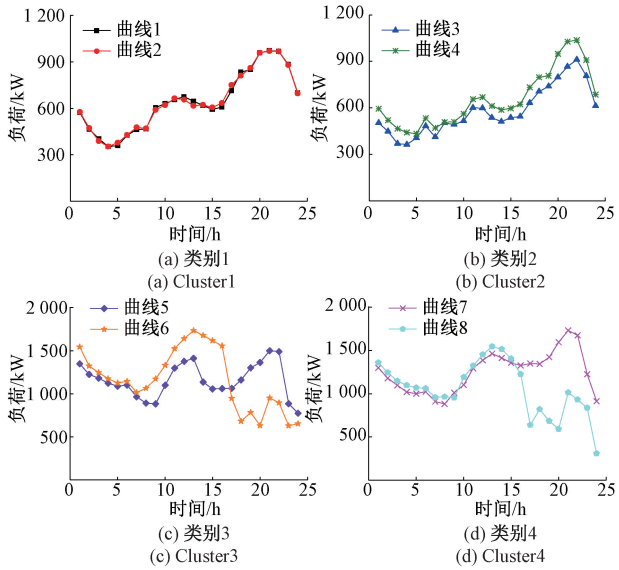

图 5 基于皮尔逊距离的密度峰值聚类曲线

Fig. 5 CFSFDP clustering curves based on Pearson distance

不同于基于欧氏距离的 K-means 聚类结果,之前的 8 条曲线被分成了 4 类。

表 4 基于皮尔逊距离的密度峰值聚类结果

Table 4 CFSFDP clustering results based on pearson distance

曲线编号	各曲线到聚类中心的距离				所属类别
	1	2	3	4	
1	**0.004 9**	0.056 6	0.240 5	1.247 0	1
2	**0.004 1**	0.051 5	0.240 8	1.270 7	1
3	0.068 1	**0.002 3**	0.287 1	1.343 3	2
4	0.059 6	**0.004 1**	0.279 9	1.327 1	2
5	0.227 5	0.280 7	**0.003 8**	0.772 5	3
6	1.252 6	1.344 9	**0.775 4**	0.956 7	3
7	1.371 7	1.484 8	0.852 2	**0.238 8**	4
8	1.438 5	1.602 3	0.982 7	**0.340 7**	4

之前属于第 1 类的曲线 1 和曲线 2 被分为了新的第 1 类,之前属于第 1 类的曲线 3 和曲线 4 被分为了新的第 2 类,之前属于第 2 类的曲线 5 和曲线 6 被分为了新的第 3 类,之前属于第 2 类的曲线 7 和曲线 8 被分为了新的第 4 类。进一步,将新的 4 类曲线如图 5 所示。

从图 5 可以看出,类别 1 和类别 2 虽然在 21:00 左右都为用电负荷高峰,5:00 左右都为用电负荷低谷。但在 5:00~15:00 这一阶段,类别 1 总体上保持了一种缓慢上升的趋势。与类别 1 不同的是,类别 2 在 6:00 左右出现了一个小的尖峰,然后用电负荷先降后升。虽然这一差异不是很大,但如果能够区分这种差异的话,对于后续的负荷预测,可以提高其预测精度。同时,这也说明了采用皮尔逊距离的密度峰值法具有更好的聚类效果,能够对居民日用电负荷进行更精细的聚类。

但是从图 5 的类别 3 和类别 4 可以看到,虽然根据曲线到聚类中心的距离,将曲线 5 和曲线 6 分成了类别 3,将曲线 7 和曲线 8 分成了类别 4。但从形态上来看,无论是曲线 5 和曲线 6,还是曲线 7 和曲线 8,均存在一定的差异。而且从表 4 可以看到,曲线 6 虽然到聚类中心 3 的距离更近,为 0.775 4,但其到聚类中心 4 的距离为 0.956 7,两者相差不大。

由此可知,采用基于皮尔逊距离的密度峰值法对居民日负荷进行聚类,虽然在一定程度上可以得到更加精确的分类结果,但在聚类曲线的形态相似性上还需要进一步改进。

最后,采用基于加权皮尔逊距离的密度峰值法聚类,8 条曲线也分成了 4 类。聚类结果如表 5 所示。从表 5 可以看到,曲线 1~4 仍然和采用基于皮尔逊距离的密度峰值法聚类时的结果一样,分别属于类别 1 和类别 2。

表 5 基于加权皮尔逊距离的密度聚类结果

Table 5 CFSFDP clustering results based on weighted Pearson distance

曲线编号	各曲线到聚类中心的距离				所属类别
	1	2	3	4	
1	**0.004 9**	0.056 6	0.242 0	1.458 6	1
2	**0.004 1**	0.051 5	0.240 5	1.466 6	1
3	0.068 1	**0.002 3**	0.287 1	1.585 3	2
4	0.059 6	**0.004 1**	0.279 9	1.578 3	2
5	0.620 8	0.577 9	**0.213 6**	0.772 0	3
6	1.435 1	1.586 2	0.949 6	**0.101 7**	4
7	0.227 5	0.280 7	**0.003 8**	0.956 7	3
8	1.371 7	1.484 8	0.852 2	**0.073 7**	4

但在基于皮尔逊距离的密度峰值法聚类中，被分为第 3 类的曲线 6，利用本文提出的聚类方法，被分到了第 4 类。类似的，在上一种聚类方法中，被分到第 4 类的曲线 7，在利用本文方法进行距离计算后，被分到了第 3 类。具体的聚类结果如图 6 所示。

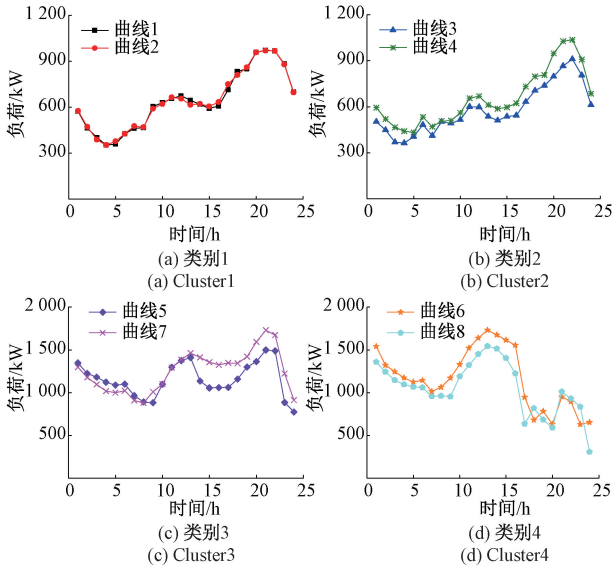

图 6　基于加权皮尔逊距离的密度聚类曲线

Fig. 6　CFSFDP clustering curves based on weighted Pearson distance

从图 6 可以看到，利用本文提出方法得到的类别 3 和类别 4 中的两条曲线，在形态上相比于图 5 的类别 3 和类别 4 两条曲线，已经更加接近。其中，类别 3 主要呈现出先下降后上升的形态，并在 21:00 左右达到负荷峰值。类别 4 的曲线 6 和曲线 8，则在 12:00 左右为一天中的用电最高峰，然后用电负荷逐渐下降，之后在 21:00 左右出现一个晚间的用电小高峰。由此可见，采用改进皮尔逊距离的密度峰值聚类后，可以使形态上相似的居民日用电负荷曲线被准确的分为同一类别，进一步提升了聚类的准确性，也为后续的居民负荷预测打下了良好的基础。

4.3　负荷预测

在利用加权皮尔逊距离的密度峰值法得到了准确的居民用电模式之后，可以采用本文提出的改进 LSTM 网络分别对不同用电模式下的居民用电负荷进行预测。这里采用每一模式内的居民日负荷数据进行预测网络训练，然后对该模式下的一周居民用电负荷进行预测。为了进行比较，也加入了基于欧氏距离的 K-means 算法聚类结果、基于皮尔逊距离的 K-means 算法聚类结果和基于皮尔逊距离的密度峰值法聚类结果的居民负荷预测。其中，对类别 1 的预测结果如图 7 所示。

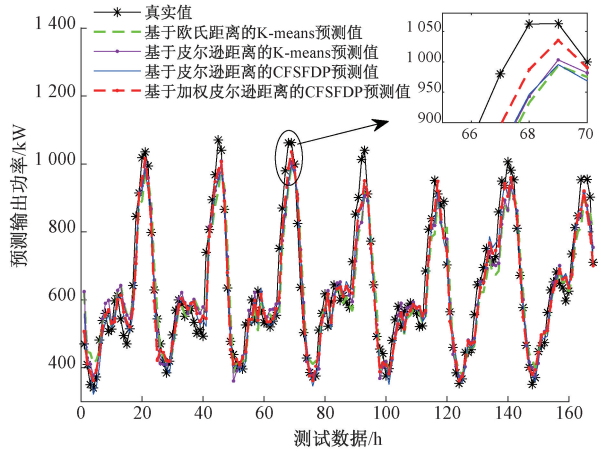

图 7　一周居民负荷预测结果

Fig. 7　Residents load forecast results of one week

图 7 中，真实值和基于 4 种聚类结果进行预测得到的预测值比较接近，所以本文对局部进行了放大。从放大部分可以看到，采用加权皮尔逊距离的密度峰值法聚类结果的预测值（深色虚线加实心圆）是最接近真实值的，要好于基于皮尔逊距离的 K-means 聚类法的预测值（浅色实线加实心圆）、基于皮尔逊距离的密度峰值法聚类结果的预测值（深色实线）和基于欧氏距离的 K-means 聚类法的预测结果（浅色虚线）。其中，浅色虚线的预测误差指标是 $MAPE = 9.065\ 8\%$，$RMSE = 73.142\ 1$；深色实线的预测误差指标是 $MAPE = 7.555\ 5\%$，$RMSE = 62.322\ 7$，浅色实线加实心圆曲线的预测误差指标是 $MAPE = 7.007\ 8\%$，$RMSE = 60.379\ 7$，深色虚线加实心圆曲线的预测误差指标 $MAPE = 6.679\ 2\%$，$RMSE = 56.008\ 5$。

从指标对比上可以看出，采用基于加权皮尔逊距离的密度峰值聚类结果来进行分类预测结果，其预测误差得到了很好的改善，两项指标都得到了提高。其中，平均绝对百分误差相比于基于欧氏距离的 K-means 聚类结果进行分类预测的误差减小了近 2.4%，相比于基于皮尔逊距离的密度峰值法聚类结果进行分类预测的误差减小了 0.88%，相比于基于皮尔逊距离的 K-means 聚类结果进行分类预测的误差也减小了 0.33%。

采用基于加权皮尔逊距离的密度峰值聚类结果来进行分类预测神经网络训练时，训练时间是 1 516.644 s。采用基于皮尔逊距离的密度峰值聚类结果来进行分类预测神经网络训练时，训练时间是 1 521.739 s。采用基于欧式距离的 K-means 聚类结果来进行分类预测神经网络训练时，训练时间是 1 644.724 s。采用基于皮尔逊距离的 K-means 聚类结果来进行分类预测神经网络训练时，训练时间是 1 682.820 s。可见，采用本文提出方法进行居民日负荷曲线聚类后，不仅可以减小预测误差，还可以

减少预测网络的训练时间。

综上可知,通过细分居民用电负荷,可以有效提高居民用电负荷的预测精度。

5 结 论

在聚类特征选择上,扩展了特征选择降维模型,提出了基于 BIC 准则的优化算法,实现了居民用电负荷中的天气特征筛选。

在聚类方法选择上,采用了基于加权皮尔逊距离的密度峰值聚类方法,弥补了特征聚类无法分辨相同电量不同用电模式的缺点,提高了聚类的细粒度,实现负荷曲线更精细化的分类识别。

在预测方法选择上,利用融合激活函数对 LSTM 深度神经网络进行了改进,以适应负荷预测中的长时间梯度消失问题。实验结果表明,在实际预测中通过得到的精细化居民用电模式,结合改进 LSTM 预测网络的方式是一种有效的居民用电负荷预测方法。

参考文献

[1] ZHONG S, TAM K S. Hierarchical classification of load profiles based on their characteristic attributes in frequency domain [J]. IEEE Transactions on Power Systems, 2014, 30(5): 2434-2441.

[2] 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 SH, DU X H, YAO H M, et al. Load curve clustering research for multiple user types [J]. Power grid technology, 2018, 42(10): 3401-3412.

[3] Quilumba F L, Lee W J, Huang H, et al. Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities[J]. IEEE Transactions on Smart Grid, 2015, 6 (1): 911-918.

[4] 龚钢军, 陈志敏, 陆俊, 等. 智能用电用户行为分析的聚类优选策略[J]. 电力系统自动化, 2018, 42(2): 58-63.
GONG G J, CHEN ZH M, LU J, et al. Clustering optimization strategy for intelligent power user behavior analysis[J]. Automation of Electric Power Systems, 2018, 42 (2): 58-63.

[5] 朱炎平. 基于聚类的用户用电行为分析研究[D]. 北京:华北电力大学, 2017.
ZHU Y P. Research on user behavior analysis based on clustering[D]. Beijing: North China Electric Power University, 2017.

[6] 李春燕, 蔡文悦, 赵溶生, 等. 基于优化 SAX 和带权负荷特性指标的 AP 聚类用户用电行为分析[J]. 电工技术学报, 2019, 34(S1): 368-377.
LI CH Y, CAI W Y, ZHAO R SH, et al. AP cluster user electricity behavior analysis based on optimizing SAX and weighted load characteristic index [J]. Journal of Electrical Technology, 2019, 34 (S1): 368-377.

[7] 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 B, ZHUANG CH J, HU J, et al. Integrated clustering algorithm for power load curve based on dimension reduction technology [J]. Chinese Journal of Electrical Engineering, 2015, 35 (15): 3741-3749.

[8] 曾兴东, 林荣恒, 邹华, 等. 面向配电网故障数据的 BIC 评估后向选择方法[J]. 北京邮电大学学报, 2017, 40(3): 104-109.
ZENG X D, LIN R H, ZOU H, et al. Backward selection method for BIC evaluation of distribution network fault data [J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40 (3): 104-109.

[9] 卜凡鹏, 陈俊艺, 张琪祁, 等. 一种基于双层迭代聚类分析的负荷模式可控精细化识别方法[J]. 电网技术, 2018, 42(3): 903-913.
BU F P, CHEN J Y, ZHANG Q Q, et al. A fine and controllable load pattern recognition method based on double-layer iterative clustering analysis [J]. Power grid technology, 2018, 42 (3): 903-913.

[10] 周世波, 徐维祥, 徐良坤. 融合密度峰值和空间邻域信息的 FCM 聚类算法[J]. 仪器仪表学报, 2019, 40(4): 137-144.
ZHOU SH B, XU W X, XU L K. FCM clustering algorithm combining density peak value and spatial neighborhood information [J]. Chinese Journal of Scientific Instrument, 2019, 40 (4): 137-144.

[11] 彭勃, 张逸, 熊军, 等. 结合负荷形态指标的电力负荷曲线两步聚类算法[J]. 电力建设, 2016, 37(6): 96-102.
PENG B, ZHANG Y, XIONG J, et al. Two-step clustering algorithm for power load curve combined with load morphology index [J]. Power Construction, 2016, 37 (6): 96-102.

[12] 陈奕延, 李晔, 李存金. 一种基于密度峰值的针对模糊混合数据的聚类算法[J]. 计算机工程与科学, 2020, 42(2): 317-324.
CHEN Y Y, LI Y, LI C J. A clustering algorithm for fuzzy mixed data based on density peaks[J]. Computer Engineering and Science, 2020, 42 (2): 317-324.

[13] 洪翠, 付宇泽, 郭谋发, 等. 改进多分类支持向量机的配电网故障识别方法[J]. 电子测量与仪器学报, 2019, 33(1): 7-15.
HONG C, FU Y Z, GUO M F, et al. Fault identification method of distribution network with improved multi-class support vector machine [J]. Journal of Electronic Measurement and Instrumentation, 2019, 33 (1): 7-15.

[14] 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 Y T, HOU F, ZHOU Q, et al. A combination analysis method of user load patterns for demand side management [J]. Power system protection and control, 2013, 41 (13): 20-25.

[15] 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 X H, CHEN ZH Y, PENG X G. A method of load pattern combination recognition based on double-layer clustering analysis [J]. Power grid technology, 2016, 40 (5): 1495-1501.

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

[17] 陈俊艺, 丁坚勇, 田世明, 等. 基于改进快速密度峰值算法的电力负荷曲线聚类分析[J]. 电力系统保护与控制, 2018, 46(20): 85-93.
CHEN J Y, DING J Y, TIAN SH M, et al. Cluster analysis of power load curve based on improved fast density peak algorithm [J]. Power system protection and control, 2018, 46 (20): 85-93.

[18] 王鹏飞, 杨余旺, 柯亚琪. 密度峰值快速聚类算法优化研究[J]. 计算机工程与科学, 2018, 40(8): 1503-1510.
WANG P F, YANG Y W, KE Y Q. Research on optimization of fast clustering algorithm for peak density [J]. Computer Engineering and Science, 2018, 40 (8): 1503-1510.

[19] 周世波, 徐维祥. 密度峰值快速搜索与聚类算法及其在船舶位置数据分析中的应用[J]. 仪器仪表学报, 2018, 39(7): 152-163.
ZHOU SH B, XU W X. Fast search and clustering algorithm for peak density and its application in ship position data analysis [J]. Journal of Instruments and Instruments, 2018, 39(7): 152-163.

[20] 谢娟英, 高红超, 谢维信. K 近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学: 信息科学, 2016, 46(2): 258-280.
XIE J Y, GAO H CH, XIE W X. K-nearest neighbor optimized density peak fast search clustering algorithms [J]. Chinese Science: Information Science, 2016, 46 (2): 258-280.

[21] AKAIKE H. A new look at the statistical model identification [J]. IEEE Transactions on Automatic Control, 1974, 19(6): 716-723.

[22] BURNHAM K P, ANDERSON D R. Multimodel inference: Understanding AIC and BIC in model selection [J]. Sociological Methods & Research, 2004, 33 (33): 261-304.

[23] 夏飞, 袁博, 彭道刚, 等. 基于信息量准则的锂离子电池变阶 RC 等效电路模型建模及优化方法[J]. 中国电机工程学报, 2018, 38(21): 6441-6451,6506.
XIA F, YUAN B, PENG D G, et al. Modeling and optimization method of variable-order RC equivalent circuit model for lithium-ion batteries based on information criterion [J]. China Journal of Electrical Engineering, 2018, 38 (21): 6441-6451,6506.

[24] 胡勇. 聚类分析结果评价方法研究[D]. 呼和浩特: 内蒙古科技大学, 2014.
HU Y. Research on evaluation method of cluster analysis results[D]. Huhot: Inner Mongolia University of Science and Technology, 2014.

作者简介

夏飞, 2000 年于沈阳理工大学获得学士学位, 2003 年于(法国)普瓦提埃大学获得硕士学位, 2017 年于同济大学获得博士学位, 现为上海电力大学副教授, 主要研究方向为电力数据分析、图像处理与电力物联网等。
E-mail: xiafeiblue@163.com

Xia Fei received his B. Sc. degree in 2000 from Shenyang University of Technology, received his M. Sc. degree in 2003 from University of Poitiers in France, received his Ph. D. degree in 2017 from Tongji University. Now he is an associate professor at Shanghai University of Electric Power. His main research interest includes power data analysis, image processing, and power IoT, etc.

张洁, 2017 年于山东农业大学获得学士学位, 现为上海电力大学硕士研究生, 主要研究方向为电力数据分析。
E-mail: 569775157@qq.com

Zhang Jie received her B. Sc. degree in 2017 from Shandong Agricultural University. Now she is a M. Sc. candidate at Shanghai University of Electric Power. Her main research interest is power data analysis.

张浩,1990 年于上海交通大学获得博士学位,现为同济大学教授,企业数字化技术教育部工程中心主任,主要研究方向为智能制造、系统工程、电力企业信息化等。
E-mail:hzhangk@ 163. com

Zhang Hao received his Ph. D. degree in 1990 from Shanghai Jiaotong University, now he is a professor in Tongji University and director of Engineering Center of Education Ministry for Enterprise Digital Technology. His main research interests include intelligent manufacturing, systems engineering, and power enterprise informatization.

陆剑峰(通信作者),2001 年于同济大学获得博士学位,现为同济大学副教授,CIMS 研究中心副主任,主要研究方向为系统工程,数字化工厂建模及优化等。
E-mail: lujianfeng@ tongji. edu. cn

Lu Jianfeng (Corresponding author) received his Ph. D. degree in 2001 from Tongji University. Now he is an associate professor at Tongji University and vice director of CIMS Research Center. His main research interests include systems engineering, modeling and optimization of digital factory.