基于去噪自编码器网络特征降维与改进小批优化K均值算法的海量用户用电行为聚类及分析

汪 颖,杨 维,肖先勇,张 姝
(四川大学 电气工程学院,四川 成都 610065)

摘要:海量用户用电特性的挖掘与分析对实现电网与用户间的双向互动具有十分重要的意义。提出一种适用于海量用户用电行为聚类及分析的算法,以降低算法时间复杂度,提升海量用户负荷数据分析效率。提取用户用电行为特征,构建多层去噪自编码网络,实现多维特征的降维;利用小批优化K均值算法进行聚类分析,并对算法进行初始聚类质心优化与超参数优化的改进以提升算法收敛速度与效果,其中超参数优化利用基于高斯过程的贝叶斯优化算法进行;利用类间分离度和类内内聚度的相关指标对聚类效果进行评价;通过互信息筛选有效聚类特征,实现用户画像。算例结果表明,所提方法在特征优化、聚类效果与收敛速度上均有较好的表现。

关键词:用电行为;特征降维;聚类分析;互信息;小批优化K均值算法;超参数优化;贝叶斯优化
中图分类号:TM 73 **文献标志码**:A DOI:10.16081/j.epae.202203017

0 引言

随着我国清洁能源渗透率的不断提高以及新型负荷增长速度的不断加快,用户侧的用电监测与调控愈发重要[1-2]。随着配电网高级量测体系AMI(Advanced Metering Infrastructure)的持续推进与建设,用户用电信息测量、存储、分析与应用的完整体系被构建,这使得基于电力大数据分析来实现用户侧用电调控成为可能。准确进行用户用电行为聚类分析与用户画像是开展用电调控的必要前提,可为优化峰谷差、平衡供需缺口提供数据支撑,并且根据用户画像的结果可明确用户的用电需求与价值,这是进行用户细分、实施精准营销的基础。

用于用户用电行为挖掘与分析的技术主要包括非侵入式负荷监测NILM(Non-Intrusive Load Monitoring)技术和大数据驱动的负荷聚类技术[3-4]。前者通过对用户总线数据的监测与分解,实现各用电设备投切与运行的监测,实时分析用户的用电行为,对该技术的研究较为成熟,但该技术属于设备级的监测技术,受监测终端改进、用户隐私等问题局限,尚未广泛应用。后者是典型的无监督式机器学习的应用,属于用户群的监测技术,适用于分布广泛的海量用户数据的实时分析。

近年来,随着国家电网公司不断推进配电网的智能化和自动化,各类监测终端与计量装置被广泛应用,形成了营销系统、计量系统、配电网自动化系统、配电网生产系统等,多源、海量的电力数据给数据挖掘与分析工作带来了巨大挑战。根据测算,我国智能监测终端每日生成几百亿条数据,每年产生的数据量超过70 TB[5]。在实际工程中,用电特性聚类的各项应用均面临着用户类型多样、体量庞大、数据通信制约等问题,如何高效地实现海量用户用电行为的挖掘与分析,是当前面临的重要问题。

针对上述问题,学者们主要从算法优化、大数据处理技术应用等方面开展研究。算法优化主要体现在数据降维以及聚类方法的选择与优化2个方面。文献[6]定义用电行为指标,对负荷数据进行降维,提出基于聚类有效性修正的德尔菲法,对日负荷特性指标进行权重配置,并以加权欧氏距离为相似性判据,实现日负荷曲线的分类;文献[7]结合推土机距离EMD(Earth Mover's Distance)和欧氏距离度量不同用户用电行为的差异程度,通过统计电力用户在多日同一时刻的负荷分布情况,从横向和纵向2个角度全面表征用户的用电行为,提出一种考虑负荷纵向随机性的基于EMD的用户用电行为识别新方法;文献[8]选取负荷曲线多维度特征,利用最大相关最小冗余mRMR(maximal Relevance and Minimal Redundancy)原则优选出特征子集,实现用户画像;文献[9]通过点积的方式构造核矩阵,将数据映射到高维空间中进行聚类,进而加大数据的可分性,并采用基于核方法的聚类算法提高负荷曲线聚类的准确性。

在大数据处理技术方面,Hadoop、Spark、Storm等大数据框架能有效解决海量数据存储、处理及分析等问题[5,10],可将算法并行化,降低算法的时空复杂度。此外,分布式计算、云计算、边缘计算等为电

收稿日期:2021-05-14;**修回日期**:2022-01-22
在线出版日期:2022-03-15
基金项目:国家自然科学基金资助项目(52077145,52007126)
Project supported by the National Natural Science Foundation of China(52077145,52007126)

力大数据分析提供了有效的解决方案。文献[11]将样本密度、类内样本平均距离的倒数和类间距离三者的乘积定义为权值积，以最大权值积法改进K-means算法，以MapReduce模型实现算法并行化，提高聚类的有效性与算法收敛速度。

传统的K-means算法需要每个样本参与质心计算以及对所有候选中心计算相似度才能进行归类。当每次迭代都需要全体样本参与计算时，就会对硬件的存储、读写速度提出较高要求，无法实现有效的实时监测、分析与应用。为进一步提高计算效率，本文提出一种适用于海量用电数据分析的方法。首先，利用系统抽样方法从海量用户数据中筛选典型负荷数据样本，提取用户行为多维特征；其次，构建多层去噪自编码器DAE（Denoising AutoEncoder）模型，利用典型样本训练网络，实现特征优化与降维；然后，将小批优化K均值MBKM（Mini-Bitch K-Means）算法作为聚类算法，通过优化质心选择与超参数优化对算法进行改进，其中超参数优化是基于高斯过程的贝叶斯优化算法GPBOA（Gaussian Process for Bayesian Optimization Algorithm）通过构建轮廓系数Sil（Silhouette coefficient）指标与超参数之间的优化关系来实现的；最后，利用互信息筛选用户用电行为关键特征，分析各类用户的用电行为与特点，并应用爱尔兰智能电表计量数据验证所提方法的有效性。

1　用电行为特征提取与归一化处理

与工业负荷相比，居民及商业负荷的波动性更强，受工作时间、温度、季节、电价、用户心理等主客观因素影响较大，本文参考国内外文献[6,8,12-13]定义常用的用户用电行为特征，特征的定义与物理意义见附录A表A1。用电行为特征是根据日负荷曲线计算得到的，本文提取的特征分为直观描述型指标（包括日最大负荷时刻、日最小负荷时刻以及峰谷相距时间3个指标）与比值描述型指标（包括日最小负荷率、日峰谷差率、日负荷率、峰期负载率、谷期负载率以及平期负载率6个指标）2类，按照时间尺度可以从全天与峰、谷、平4个角度对用电行为特征进行划分。

由于不同的用户用电属性差别较大，因此在特征降维前对特征进行归一化处理以提升模型的收敛速度与精度，本文采用零均值归一化，即：

$$\boldsymbol{X}=\frac{\boldsymbol{X}_{\text{init}}-\boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{1}$$

式中：$\boldsymbol{X}$为归一化处理后的样本数据；$\boldsymbol{X}_{\text{init}}$为原始特征样本数据；$\boldsymbol{\mu}$为所有样本数据的均值矩阵；$\boldsymbol{\sigma}$为所有样本数据的标准差。处理后的样本数据服从均值为0、标准差为1的标准正态分布。

2　基于DAE网络的多维特征优化

特征选择与降维技术可去除冗余特征以及提取有效信息，本文通过构建DAE网络来实现高维特征的优化，提升聚类速度与效果。

2.1　自动编码器的特征降维与重构

2.1.1　自动编码器基本原理

典型的特征降维方法包括特征选择[13]和特征变换。特征选择包括过滤式、包裹式与嵌入式3种，其通过一定的规则选出分类或聚类特征，难以保留全局信息；特征变换包括线性方法和非线性方法，这2种方法保留全局特征的能力存在差异。

自动编码器AE（AutoEncoder）是一种无监督式的特征降维与特征表达方法，由编码器与解码器构成，是一种输入和训练目标相同的神经网络。AE的参数通过重构损失训练得到，重构损失为：

$$R_{\text{loss}}(f,g)=\underset{f,g}{\operatorname{argmin}}\left\{\frac{1}{n}\sum_{i=1}^{n}L(\boldsymbol{X}^i,\hat{\boldsymbol{X}}^i)\right\} \tag{2}$$

式中：$R_{\text{loss}}(f,g)$为重构损失，f为AE的编码过程，g为AE的解码过程；n为样本数，将每个样本的维度记为m，则$\boldsymbol{X}$的大小为$n\times m$；$\boldsymbol{X}^i$为$\boldsymbol{X}$的第i行，表示第i个样本；$\hat{\boldsymbol{X}}^i$为第i个重构样本，所有重构样本构成$\hat{\boldsymbol{X}}$；argmin{·}表示获取特定AE网络参数使得重构损失达到最小值。

AE的编码与重构过程$r(\boldsymbol{X})$表示为：

$$r(\boldsymbol{X})=g\left[f(\boldsymbol{X})\right]=\hat{\boldsymbol{X}} \tag{3}$$

AE的训练过程就是利用随机梯度下降法调整网络参数（权重$\boldsymbol{w}$与偏置$\boldsymbol{b}$），使重构信号与输入信号误差最小，本文选用交叉熵作为AE的损失函数$\xi(\boldsymbol{X})$，即：

$$\xi(\boldsymbol{X})=\frac{1}{m}\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}\left[\ln\left(1-\hat{X}(i,j)\right)-X(i,j)\ln\hat{X}(i,j)\right] \tag{4}$$

式中：$X(i,j)$为$\boldsymbol{X}$的第i行第j列元素；$\hat{X}(i,j)$为$\hat{\boldsymbol{X}}$的第i行第j列元素。

2.1.2　AE降维的维度约束

特征降维的目的主要有3个：降低输入向量维度，从而降低聚类算法的复杂度；实现数据的可视化，以便于观察数据的分布特点；减轻数据传输压力，只需传递训练好的网络参数与降维后的数据即可实现数据重构。根据上述目的，特征压缩维度越低越好，但该维度受信息保留率[14]和信号重构误差限制，且信息保留率越高，信号重构误差越小。信息保留率η与重构损失函数e_η分别为：

$$\eta=1-\frac{E_{\text{ASPE}}}{T_{\text{V}}}=1-\frac{\sum_{j=1}^{d_{\text{en}}}\left\|\boldsymbol{X}_j-\hat{\boldsymbol{X}}_j\right\|^2/d_{\text{en}}}{\sum_{j=1}^{m}\left\|\boldsymbol{X}_j\right\|^2/m} \tag{5}$$

$$e_\eta=\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(X(i,j)-\hat{X}(i,j)\right)^2}{mn}} \tag{6}$$

式中：E_{ASPE}为平均平方映射误差；T_{V}为总变差；d_{en}为降维后的特征维度，其值根据η与e_η的大小变化进行选取；$\boldsymbol{X}_j$为$\boldsymbol{X}$的第j列，表示第j维特征，$\boldsymbol{X}_1$—$\boldsymbol{X}_9$分别为附录A表A1中日峰谷差率、日最小负荷率、日负荷率、日最大负荷时刻、日最小负荷时刻、峰期负载率、谷期负载率、平期负载率、峰谷相距时间；$\hat{\boldsymbol{X}}_j$为$\hat{\boldsymbol{X}}$的第j列，表示重构后的第j维特征。

2.2 DAE网络

传统AE通过几十次迭代训练即可达到较好的效果，但易出现过拟合现象，通过随机失活（Dropout）正则化、增加输入样本噪声[15]等方法可提高模型泛化能力，因此，本文对训练样本增加噪声，如式（7）所示，在输入层间进行Dropout处理，并在训练阶段减弱神经元的联合适应性，增强模型的泛化能力。

$$\begin{cases}\boldsymbol{X}_{\text{train-N}}=\boldsymbol{X}_{\text{train}}+N_{\mathrm{F}}\boldsymbol{X}_{\mathrm{N}}\\ \boldsymbol{X}_{\mathrm{N}}\sim\mathrm{N}(0,1)\end{cases} \tag{7}$$

式中：$\boldsymbol{X}_{\text{train}}$为训练样本；$\boldsymbol{X}_{\text{train-N}}$为$\boldsymbol{X}_{\text{train}}$通过式（7）产生的损坏数据；$N_{\mathrm{F}}$为噪声因子；$\boldsymbol{X}_{\mathrm{N}}$为服从均值为0、标准差为1的正态分布的数据。

2.3 基于DAE网络降维的特征互信息分析

DAE网络降维后的特征与初始特征不同，本文通过计算降维前、后特征的互信息MI（Mutual Information）得到用户行为的关键特征，如式（8）所示。

$$M_{\mathrm{I}}\left(\boldsymbol{X}_j;\boldsymbol{Y}_s\right)=\sum_{\lambda\in\boldsymbol{Y}_s}\sum_{\tau\in\boldsymbol{X}_j}p(\lambda,\tau)\lg\frac{p(\lambda,\tau)}{p(\lambda)p(\tau)} \tag{8}$$

式中：$\boldsymbol{Y}_s$（$s=1,2$）为降至2维后AE的特征，表示$\boldsymbol{Y}$的第s列；$M_{\mathrm{I}}(\boldsymbol{X}_j;\boldsymbol{Y}_s)$为降维前、后特征的互信息；$\lambda\in\boldsymbol{Y}_s$、$\tau\in\boldsymbol{X}_j$表示$\lambda$、$\tau$分别为$\boldsymbol{Y}_s$、$\boldsymbol{X}_j$的元素；$p(\lambda,\tau)$为$\lambda$和$\tau$的联合分布概率；$p(\lambda)$、$p(\tau)$分别为$\lambda$和$\tau$的概率密度。互信息值越大表示两变量相关性越高；互信息值为0时，表示两变量相互独立。

3 基于改进MBKM算法的海量用电数据聚类

3.1 MBKM算法原理

聚类属于无监督学习范畴，是将无标签的数据按照同属性进行聚合。常见的聚类算法包括基于划分的算法、基于密度的算法、基于层次的算法、基于网络的算法和基于模型的算法五大类。*K*-means算法是典型的基于划分的算法，该算法计算过程简单，时间复杂度低，但其对初始值的设置较为敏感，不能识别非球形类，并且当聚类样本量非常大时，即使考虑了距离优化，算法仍然较为耗时，且聚类效果不佳。在大数据背景下，文献[16]提出MBKM算法以解决大样本聚类问题，该算法使用小批量样本优化*K*-means算法[17]，即每次采用随机产生的子集训练算法，以缩短计算时间，该算法的优势在于小批量的随机噪声往往比整体的低[16]，当数据集随着冗余样本的增加而变大时，不会增加计算成本。

MBKM算法主要是通过取样本的流平均值以及之前分配给质心的所有样本来更新聚类质心，达到降低聚类质心变化率的效果，如式（9）所示。

$$\begin{cases}c_k\leftarrow(1-\eta_{\mathrm{s}})c_k+\eta_{\mathrm{s}}x_{\text{sample}}\\ \eta_{\mathrm{s}}=\dfrac{1}{v(c_k)}\end{cases} \tag{9}$$

式中：c_k（$k=1,2,\cdots$）为第k个聚类质心；η_{s}为学习率；x_{sample}为小样本中的一条数据；$v(c_k)$为小样本第k个类的计数。在达到一定迭代次数后，小样本的收敛特性与整体样本收敛特性接近。

3.2 MBKM算法的改进

为进一步提升MBKM算法的聚类速度和效果，本文从2个方面对该算法进行改进：优选初始聚类质心；基于贝叶斯优化理论进行MBKM算法初始超参数的优化。

在大数据背景下，*K*-means++算法[18]与本文改进MBKM算法的效果相差极小，但是*K*-means++算法需要全部样本进行迭代，算法收敛时间会随着冗余样本的增加而增长。

3.2.1 初始聚类质心优化

本文在MBKM算法的基础上，采用初始优化方法确定初始聚类质心，进一步提升算法的收敛性能。首先随机选取一个初始聚类质心，根据式（10）计算每个样本与已选出的聚类质心的最短距离$D_{\mathrm{e}}(\boldsymbol{X}^i)$，然后选取最短距离最大的点作为新的聚类质心，直到选出K个聚类质心，其中K为聚类数。

$$D_{\mathrm{e}}\left(\boldsymbol{X}^i\right)=\operatorname{argmin}\left\{\left|\boldsymbol{X}^i-\boldsymbol{u}_r\right|^2\right\}\quad r=1,2,\cdots,k_{\text{selected}} \tag{10}$$

式中：$\boldsymbol{u}_r$为已选出的第r个聚类质心；k_{selected}为已选出的聚类质心数。

3.2.2 基于高斯过程贝叶斯优化的超参数优化

超参数是独立于建模过程的自由参数，超参数优化可极大提升计算效率，常见的优化方法包括随机搜索、遗传优化等。在大数据背景下，上述方法的测试成本高，而贝叶斯优化根据已有测试数据决定下一次的测试参数，可大幅提高搜索效率。

MBKM算法的主要参数见附录A表A2。其中，初始化质心运行算法的次数N_{t}、质心被重新赋值的最大次数比例ε、连续采样包个数β决定了算法的整体运行时间，一般将N_{t}与β设置为默认值，对ε值进行寻优；采样包大小b默认值为100，如果发现数据集的类别较多或者噪声点较多，则对b值进行优

化以提升聚类效果。本文选用Sil作为MBKM算法超参数（K、ε与b）的优化指标，K值通过肘部法则与Sil变化曲线拐点综合确定[19]。肘部法则通过计算误差平方和SSE（the Sum of Squares due to Error）确定，即：

$$E_{\mathrm{SSE}}=\sum_{s=1}^{K}\sum_{q\in C_s}\left|q-m_s\right|^2 \tag{11}$$

式中：E_{SSE}为SSE；C_s为第s个类；$q\in C_s$表示q是第s个类的样本；m_s为第s个类的聚类质心。

本文利用GPBOA对MBKM算法的关键超参数（ε与b）进行寻优。记已获取的观测样本为$B_{1:t}=\left\{\left(\boldsymbol{\gamma}_1,y_1\right),\left(\boldsymbol{\gamma}_2,y_2\right),\cdots,\left(\boldsymbol{\gamma}_t,y_t\right)\right\}$，其中$t$为观测样本数，$\boldsymbol{\gamma}_i(i=1,2,\cdots,t)$为第$i$个样本的超参数寻优向量，所有超参数寻优向量构成矩阵$\boldsymbol{\gamma}=\left[\boldsymbol{\gamma}_1\ \boldsymbol{\gamma}_2\ \cdots\ \boldsymbol{\gamma}_t\right]$，$y_i(i=1,2,\cdots,t)$为第$i$个样本的观测值，所有样本的观测值构成向量$\boldsymbol{y}=\left[y_1\ y_2\ \cdots\ y_t\right]$。以Sil值最大化作为优化目标，建立Sil与超参数之间的优化关系为：

$$\boldsymbol{\gamma}_{\mathrm{option}(\varepsilon,b)}=\underset{\boldsymbol{\gamma}_{\mathrm{option}(\varepsilon,b)}\in\boldsymbol{\chi}\subset\boldsymbol{R}^D}{\operatorname{argmax}}\left\{f_{\mathrm{sil}}\left(\boldsymbol{\gamma}\right)\right\} \tag{12}$$

式中：$\boldsymbol{\gamma}_{\mathrm{option}(\varepsilon,b)}$为待寻优超参数向量；$\boldsymbol{\chi}$为以$\varepsilon$与$b$为变量的超参数寻优空间；$\boldsymbol{R}^D$为以全体超参数为变量的寻优空间；argmax{·}表示最大值寻优过程；$f_{\mathrm{sil}}(\cdot)$为超参数与Sil的函数关系。GPBOA主要包括以下3个步骤。

1）构建样本的高斯过程回归模型$g_{\mathrm{D}}\left(\boldsymbol{\gamma},\boldsymbol{v},\boldsymbol{\Sigma}\right)$，其中$\boldsymbol{v}$为均值矩阵，$\boldsymbol{\Sigma}$为协方差矩阵。

2）求采样函数$u\left(\boldsymbol{\gamma},g_{\mathrm{D}}\left(\boldsymbol{\gamma},\boldsymbol{v},\boldsymbol{\Sigma}\right)\right)$的极值点$\boldsymbol{\gamma}_*$，即：

$$\boldsymbol{\gamma}_*=\operatorname{argmax}\left\{u\left(\boldsymbol{\gamma},g_{\mathrm{D}}\left(\boldsymbol{\gamma},\boldsymbol{v},\boldsymbol{\Sigma}\right)\right)\right\} \tag{13}$$

3）通过测试得到新样本$\left(\boldsymbol{\gamma}_*,\boldsymbol{\gamma}_*\right)$，其中$\boldsymbol{\gamma}_*=f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_*\right)+\boldsymbol{\varepsilon}_*$为新样本的观测值，$f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_*\right)$为新样本的估计值，$\boldsymbol{\varepsilon}_*$为新样本的观测误差。更新总样本以及高斯过程回归模型，进入下一次迭代，依此循环，直至迭代结束。

步骤1）中的高斯模型是对给定样本估计出的概率分布，$g_{\mathrm{D}}\left(\boldsymbol{\gamma},\boldsymbol{v},\boldsymbol{\Sigma}\right)$服从如下多维正态分布：

$$g_{\mathrm{D}}\left(\boldsymbol{\gamma},\boldsymbol{v},\boldsymbol{\Sigma}\right)=\frac{1}{\sqrt{(2\pi)^t\left|\boldsymbol{\Sigma}\right|}}\exp\left[-\frac{1}{2}\left(\boldsymbol{\gamma}-\boldsymbol{v}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\gamma}-\boldsymbol{v}\right)\right] \tag{14}$$

$$\boldsymbol{v}=\begin{bmatrix}m\left(\boldsymbol{\gamma}_1\right)\\m\left(\boldsymbol{\gamma}_2\right)\\\vdots\\m\left(\boldsymbol{\gamma}_t\right)\end{bmatrix}=\frac{1}{t}\begin{bmatrix}\sum_{i=1}^{t}f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_i\right)\\\sum_{i=1}^{t}f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_i\right)\\\vdots\\\sum_{i=1}^{t}f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_i\right)\end{bmatrix} \tag{15}$$

$$\boldsymbol{\Sigma}=\begin{bmatrix}\psi\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_1\right)&\psi\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\right)&\psi\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_t\right)\\\psi\left(\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_1\right)&\psi\left(\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_2\right)&\psi\left(\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_t\right)\\\vdots&\vdots&\vdots\\\psi\left(\boldsymbol{\gamma}_t,\boldsymbol{\gamma}_1\right)&\psi\left(\boldsymbol{\gamma}_t,\boldsymbol{\gamma}_2\right)&\psi\left(\boldsymbol{\gamma}_t,\boldsymbol{\gamma}_t\right)\end{bmatrix} \tag{16}$$

式中：$m(\cdot)$为均值函数；$f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_i\right)(i=1,2,\cdots,t)$为第$i$个样本的估计值；$\psi(\cdot,\cdot)$为高斯核函数，本文选择平方指数SE（Squared Exponential）作为核函数。根据任意有限个随机变量都满足一个联合高斯分布的性质，并考虑观测值$\boldsymbol{y}$的噪声误差，可得观测值$\boldsymbol{y}$与超参数$\boldsymbol{\gamma}$的边际似然分布$\vartheta\left(\boldsymbol{y}\middle|\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_t,\sigma_{\mathrm{noise}}^2\right)$为：

$$\vartheta\left(\boldsymbol{y}\middle|\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_t,\sigma_{\mathrm{noise}}^2\right)\sim\mathrm{N}\left(0,\boldsymbol{\Sigma}+\sigma_{\mathrm{noise}}^2\boldsymbol{I}\right) \tag{17}$$

$$\boldsymbol{y}=f_{\mathrm{sil}}\left(\boldsymbol{\gamma}\right)+\boldsymbol{\varepsilon} \tag{18}$$

式中：$\boldsymbol{\varepsilon}$为满足高斯独立同分布的噪声，噪声均值为0，标准差为σ_{noise}；$\boldsymbol{I}$为单位矩阵。当出现测试新样本$\left(\boldsymbol{\gamma}_*,\boldsymbol{\gamma}_*\right)$时，根据样本观测值$\boldsymbol{y}$与$f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_*\right)$的联合分布得到预测分布$\vartheta\left(f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_*\right)\middle|\boldsymbol{\gamma}_*,B_{1:t}\right)$为：

$$\vartheta\left(f_{\mathrm{sil}}\left(\boldsymbol{\gamma}_*\right)\middle|\boldsymbol{\gamma}_*,B_{1:t}\right)\sim\mathrm{N}\left(m\left(\boldsymbol{\gamma}_*\right),\mathrm{cov}\left(m\left(\boldsymbol{\gamma}_*\right)\right)\right) \tag{19}$$

$$\begin{cases}m\left(\boldsymbol{\gamma}_*\right)=\boldsymbol{\psi}_*^{\mathrm{T}}\left(\boldsymbol{\Sigma}+\sigma_{\mathrm{noise}}^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}\\\mathrm{cov}\left(m\left(\boldsymbol{\gamma}_*\right)\right)=\psi_{**}-\boldsymbol{\psi}_*^{\mathrm{T}}\left(\boldsymbol{\Sigma}+\sigma_{\mathrm{noise}}^2\boldsymbol{I}\right)^{-1}\boldsymbol{\psi}_*\\\boldsymbol{\psi}_*=\left[\psi\left(\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_*\right)\ \psi\left(\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_*\right)\ \cdots\ \psi\left(\boldsymbol{\gamma}_t,\boldsymbol{\gamma}_*\right)\right]^{\mathrm{T}}\\\psi_{**}=\psi\left(\boldsymbol{\gamma}_*,\boldsymbol{\gamma}_*\right)\end{cases} \tag{20}$$

式中：$m\left(\boldsymbol{\gamma}_*\right)$为预测均值；$\mathrm{cov}\left(m\left(\boldsymbol{\gamma}_*\right)\right)$为预测协方差。

4　聚类有效性的评价指标

在实际工程中，大量数据是无标签的，常采用Sil、方差比标准指数CHI（Calinski-Harabaz Index）、邓恩指数DVI（Dunn Validity Index）、戴维森堡丁指数DBI（Davies-Bouldin Index）等[20]进行评价。本文选用Sil、CHI及DBI这3类指数进行聚类有效性的评价。

第i个样本的Sil值$I_{\mathrm{sil}}(i)$通过结合类内内聚度和类间分离度进行计算，即：

$$I_{\mathrm{sil}}(i)=\frac{g(i)-a(i)}{\max\left\{a(i),g(i)\right\}} \tag{21}$$

式中：$a(i)$为第i个样本的类内内聚度；$g(i)$为第i个样本的类间分离度。

$$\begin{cases}a(i)=\dfrac{1}{n_s}\sum\limits_{i,j\in C_s}D(i,j)\quad i\neq j\\g(i)=\min\left\{\dfrac{1}{n_s}\sum\limits_{i\notin C_s,j\in C_s}D(i,j)\right\}\end{cases}\quad s=1,2,\cdots,K \tag{22}$$

式中：n_s为第s个类中样本的数量；$D(i,j)$为第i个样本与第j个样本之间的欧氏距离，表征不相似度。对所有样本的Sil值求平均值，以用于表示整体聚类效果，Sil值取值范围为[-1,1]，其越趋近于1表示类内内聚度和类间分离度越优。

CHI是类间色散平均值与类内色散的比值，如式(23)所示，其值越大，则聚类效果越好。

$$s(K)=\frac{T_r(\boldsymbol{B}_K)}{T_r(\boldsymbol{W}_K)}\frac{N-K}{K-1} \tag{23}$$

式中：$s(K)$为K个聚类数的总得分；$T_r(\cdot)$为矩阵的迹；N为数据总数；$\boldsymbol{B}_K$为类间色散矩阵，$\boldsymbol{W}_K$为类内色散矩阵，两矩阵定义见文献[21]。CHI的计算速度快，但凸类的CHI较高。

DBI是类内距离之和与类间距离之比，即：

$$I_{\mathrm{DBI}}=\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K}\max_{i\neq j}\left\{\frac{\overline{S_i}+\overline{S_j}}{\left\|\boldsymbol{\omega}_i-\boldsymbol{\omega}_j\right\|_2}\right\} \tag{24}$$

式中：I_{DBI}为DBI；$\overline{S_i}$、$\overline{S_j}$分别为第i个和第j个类内数据到类质心的平均距离；$\boldsymbol{\omega}_i$、$\boldsymbol{\omega}_j$分别为第i个和第j个类的类向量，其欧氏距离表示类间距离。DBI越小，则聚类效果越好。

5 本文方法流程

本文方法流程如图1所示，主要包括典型样本训练、特征提取与预处理、DAE网络特征降维、基于改进的MBKM算法聚类以及用户行为分析与画像。

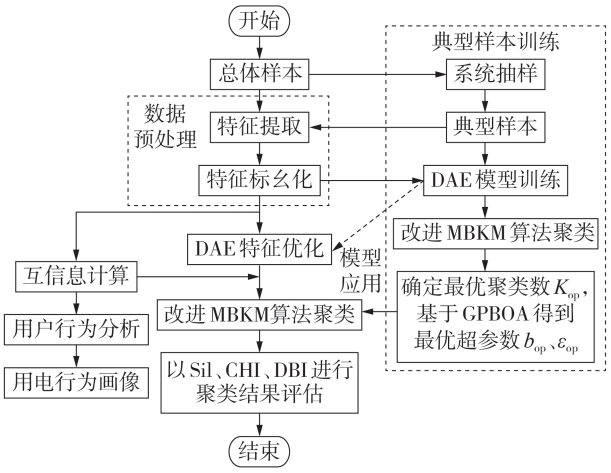


图1 本文方法流程图

Fig.1 Flowchart of proposed method

典型样本训练是本文所提方法的基础，训练好DAE模型应用于总体样本的特征优化后，本文采用系统抽样方法对典型样本进行抽取，即等距抽样。需要注意的是，在确定抽样方法前，需要校验样本的均衡性，当样本不均衡时，需要采取数据增强、分层抽样等策略。

6 算例分析

6.1 数据来源与实验平台

本文利用爱尔兰智能电表的计量数据[19]对所提方法进行验证，该数据来源于爱尔兰电力和天然气行业监管机构CER(Commission for Energy Regulation)每隔半小时记录一次的用电量数据。实验硬件平台是64位Windows系统，该系统采用Intel core(i7)，3.40 GHz处理器，8 GB RAM。深度学习框架基于Keras(基于TensorFlow1.2)实现。

6.2 用户用电行为特征聚类

6.2.1 多层DAE网络的构建

本文随机抽取总样本的10%作为典型样本，共提取9种用户用电行为特征，以用于训练网络。所构建的多层DAE网络参数见附录A表A3。网络共7层，采用全连接层，在全连接层Dense_1与全连接层Dense_2之间增加Dropout正则化处理(神经元失活率为0.5)；从输入层到全连接层Dense_3为编码结构，其将特征压缩至2维；从全连接层Dense_3到输出层为解码结构。网络训练共迭代200次。图2为DAE网络训练误差曲线，由图可见，在迭代约30次后DAE网络有效收敛。

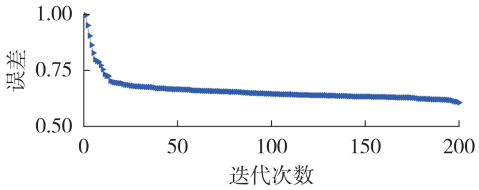


图2 DAE网络训练误差曲线

Fig.2 Training error curve of DAE network

降维维度是通过不同数据压缩比例下数据的信号重构误差和信息保留率来确定的。当维度为2时，信号重构误差e_η=0.86，信息保留率为90%；当维度为5时，信号重构误差e_η=0.82，信息保留率为95%。可见，维度降低并未损失过多信息，因此可令维度为2。

6.2.2 多层DAE网络的特征优化效果测试

为了证明所构建的DAE网络的优化效果，本文从聚类指标、计算时长等方面对比本文所构建的多层DAE网络、标准主成分分析PCA(Principal Component Analysis)、截断奇异值分解TSVD(Truncated Singular Value Decomposition)、单层AE以及多层AE的特征降维效果。

从数据集中随机抽取1 000条用户数据作为测试样本，提取用电特征并将其维度降至2、采用K-means++算法进行聚类，结果对比如附录A表A4所示。由表可见：与未降维的结果相比，降维后聚类效果得到明显提升，计算时间明显缩短；本文方法与PCA和TSVD的计算时间相近，但聚类效果差别明

显；单层AE以及多层AE的效果不如本文的多层DAE，这说明DAE的模型泛化能力较好。

6.2.3　基于改进MBKM算法与典型样本的超参数优化

本节对典型样本进行训练，利用训练完成的DAE网络与改进MBKM算法，以抽取的典型样本作为输入，确定全局样本最优聚类数K_{op}、改进MBKM算法最优超参数（ε_{op}与b_{op}）。

选取典型训练样本，利用改进MBKM算法得到图3，并进行最优聚类数K_{op}的判定。由图可见：当聚类数为2和3时，Sil较大，当聚类数增至4时，Sil大幅下降，这说明聚类数增至4之后聚类效果不佳；在聚类数增至3后，SSE减小的幅度变缓，这说明再增加聚类数的效果提升不明显，即聚类数为3时是肘点。综上，可以确定最优聚类数$K_{op}=3$。

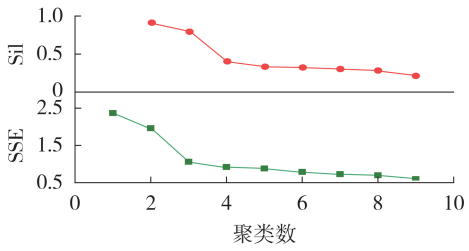


图3　聚类数与Sil、SSE的关系图

Fig.3　Relationship diagram of clustering number vs. Sil and SSE

选取典型样本的用户数据作为超参数寻优样本，设置ε的取值范围为[0，0.04]，间隔为0.005；设置b的取值范围为[50，400]，间隔为50。附录A图A1为在ε与b初始测试样本下的Sil分数热图，由图可见，不同的参数组合具有不同的聚类效果。设置GPBOA的迭代次数为30，迭代过程中超参数与优化目标Sil之间的对应关系如附录A表A5所示。由表可见，当迭代次数为24时，Sil出现最大值，迭代24次之后的Sil呈现下降趋势，最终得到最优超参数b_{op}=122、ε_{op}=0.03096。

6.2.4　聚类效果与计算时间对比

在不同数据集大小下，将改进MBKM算法、传统MBKM算法、K-means++算法、利用层次方法的平衡迭代规约和聚类BIRCH（Balanced Iterative Reducing and Clustering using Hierarchies）算法、基于高斯混合模型的期望最大值聚类EM-GMM（Expected Maximum clustering based on Gaussian Mixture Model）算法、谱聚类SPC（SPectral Clustering）算法进行比较，各算法的聚类效果如附录A表A6所示。不同算法的收敛时间如附录A图A2所示。由表A6可知：BIRCH算法、K-means++算法、传统MBKM算法与改进MBKM算法的聚类指标均明显优于EM-GMM算法与SPC算法，其中K-means++算法、传统MBKM算法与改进MBKM算法的效果优于BIRCH算法；未经超参数优化与质心优化的传统MBKM算法的聚类效果较差；相较于K-means++算法，以Sil指数为超参数优化目标的改进MBKM算法的性能得到明显提升，不仅在聚类效果的整体指标上与K-means++算法十分接近，而且Sil指数更优。由图A2(a)可知，SPC算法收敛时间最长，呈现指数增长。由图A2(b)可知，数据集大小在0~10 000范围内时，K-means++算法与改进MBKM算法的收敛时间差别不大，但随着数据集继续扩大，K-means++算法的收敛时间增长明显，而改进MBKM算法的收敛时间呈现缓慢的线性增长趋势，在分析更大的数据集时，改进MBKM算法的优势将更明显。

6.3　基于特征互信息计算的用电行为分析

本节应用所提方法对1 000个不同用户在同一天的用电数据进行聚类。图4为经过DAE降维后的聚类情况，共分为3类用户，第1类用户的数量最多。

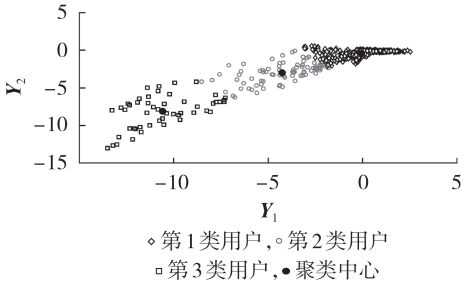


图4　降维特征的聚类结果

Fig.4　Clustering results of dimension reduction features

聚类中心与行为特征的相关图如附录A图A3所示。用户行为特征的互信息计算结果如表1所示。由表可知：$\boldsymbol{X}_1$—$\boldsymbol{X}_3$与降维特征间的互信息相对较大，平均值在6以上；$\boldsymbol{X}_4$、$\boldsymbol{X}_5$、$\boldsymbol{X}_7$—$\boldsymbol{X}_9$与降维特征间的互信息相对较小，平均值在4以下。由图A3(a)可知，$\boldsymbol{X}_1$—$\boldsymbol{X}_3$与各类用户的统计值区分明显。综上可知，在初始特征空间中，$\boldsymbol{X}_1$—$\boldsymbol{X}_3$是主要分类特征。由图A3(b)可知，$\boldsymbol{Y}_1$与各类用户的统计值区分明显，而$\boldsymbol{Y}_2$与各类用户的统计值区分不明显。

表1　初始特征与降维特征的互信息

Table 1　Mutual information between initial features and dimension reduction features

特征	互信息								
	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	$\boldsymbol{X}_7$	$\boldsymbol{X}_8$	$\boldsymbol{X}_9$
$\boldsymbol{Y}_1$	6.32	5.32	5.73	2.12	2.54	4.77	4.10	3.95	2.98
$\boldsymbol{Y}_2$	6.75	6.75	6.83	3.60	3.68	4.83	4.43	4.83	3.48

3类用户的初始特征统计如附录A表A7所示，3类用户的用电属性特点如附录A图A4所示。第1类用户的$\boldsymbol{X}_1$—$\boldsymbol{X}_3$中心值与另外2类用户的差别明显，第1类用户的$\boldsymbol{X}_1$中心值为94.715 9%，约为第2类用户的4.8倍，约为第3类用户的1.7倍；3类用户的用电高峰约出现在11:00—14:00，这与降温负荷

关系密切;第2类用户的峰谷相距时间为10.379 3 h,第1类与第3类用户的负荷启停规律较相似。对比3类用户的指标可知,不同类型用户在不同维度的用电特征上存在较大差异:第1类用户用电波动最大,峰期负载率$\boldsymbol{X}_6$中心值最高,而谷期负载率$\boldsymbol{X}_7$中心值最低,这类用户多为商业等用户;第2类用户的用电最平稳,多为轻工业等用户;第3类用户的负荷有一定的波动性,整体负载率也较高,这类用户多为居民用户。

7 结论

本文提出一种适用于海量用户用电特征聚类与分析的方法,可为电力需求管理、电力营销方案制定等工作提供支撑,并应用爱尔兰智能电表的计量数据对所提方法进行了验证,得到以下结论:

1)构建DAE网络进行特征降维,与典型特征降维方法相比,所提方法在特征可视化、保留全局信息、信号重构等方面更优;

2)将MBKM算法应用于海量电力负荷数据的聚类,并对算法在质心优化与超参数优化两方面进行改进,实验结果表明,与其他算法相比,本文算法在收敛速度与聚类效果上表现更优;

3)通过计算降维前、后特征间的互信息可有效筛选关键特征,实现用户用电行为画像。

附录见本刊网络版(http://www.epae.cn)。

参考文献:

[1] 彭小圣,邓迪元,程时杰,等. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报,2015,35(3):503-511.
PENG Xiaosheng, DENG Diyuan, CHENG Shijie, et al. Key technologies of electric power big data and its application prospects in smart grid[J]. Proceedings of the CSEE, 2015, 35(3):503-511.

[2] 高赐威,李倩玉,李慧星,等. 基于负荷聚合商业务的需求响应资源整合方法与运营机制[J]. 电力系统自动化,2013,37(17):78-86.
GAO Ciwei, LI Qianyu, LI Huixing, et al. Methodology and operation mechanism of demand response resources integration based on load aggregator[J]. Automation of Electric Power Systems, 2013, 37(17):78-86.

[3] HART G W. Nonintrusive appliance load monitoring[J]. Proceedings of the IEEE, 1992, 80(12):1870-1891.

[4] 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.

[5] MATTHEWS S J, ST LEGER A. Leveraging MapReduce and synchrophasors for real-time anomaly detection in the smart grid[J]. IEEE Transactions on Emerging Topics in Computing, 2019, 7(3):392-403.

[6] 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.

[7] 冯志颖,唐文虎,吴青华,等. 考虑负荷纵向随机性的用户用电行为聚类方法[J]. 电力自动化设备,2018,38(9):39-44,53.
FENG Zhiying, TANG Wenhu, WU Qinghua, et al. Users' consumption behavior clustering method considering longitudinal randomness of load[J]. Electric Power Automation Equipment, 2018, 38(9):39-44, 53.

[8] 赵晋泉,夏雪,刘子文,等. 电力用户用电特征选择与行为画像[J]. 电网技术,2020,44(9):3488-3496.
ZHAO Jinquan, XIA Xue, LIU Ziwen, et al. User electricity consumption feature selection and behavioral portrait[J]. Power System Technology, 2020, 44(9):3488-3496.

[9] 赵文清,龚亚强. 基于Kernel *K*-means的负荷曲线聚类[J]. 电力自动化设备,2016,36(6):203-207.
ZHAO Wenqing, GONG Yaqiang. Load curve clustering based on Kernel *K*-means[J]. Electric Power Automation Equipment, 2016, 36(6):203-207.

[10] 朱文俊,王毅,罗敏,等. 面向海量用户用电特性感知的分布式聚类算法[J]. 电力系统自动化,2016,40(12):21-27.
ZHU Wenjun, WANG Yi, LUO Min, et al. Distributed clustering algorithm for awareness of electricity consumption characteristics of massive consumers[J]. Automation of Electric Power Systems, 2016, 40(12):21-27.

[11] 张承畅,张华誉,罗建昌,等. 基于云计算和改进*K*-means算法的海量用电数据分析方法[J]. 计算机应用,2018,38(1):159-164.
ZHANG Chengchang, ZHANG Huayu, LUO Jianchang, et al. Massive data analysis of power utilization based on improved *K*-means algorithm and cloud computing[J]. Journal of Computer Applications, 2018, 38(1):159-164.

[12] 国家发展改革委员会,国家电网公司. 电力需求侧管理工作指南[M]. 北京:中国电力出版社,2007:237-245.

[13] WANG Z C, WU H, JIANG Z B, et al. Singular value decomposition-based load indexes for load profiles clustering[J]. IET Generation, Transmission & Distribution, 2020, 14(19):4164-4172.

[14] VALLE S, LI W, QIN S J. Selection of the number of principal components[J]. Industrial & Engineering Chemistry Research, 1999, 38(11):4389-4401.

[15] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the 25th international conference on Machine learning-ICML'08. New York, USA: ACM Press, 2008:1096-1103.

[16] SCULLEY D. Web-scale *k*-means clustering[C]//Proceedings of the 19th international conference on World wide web-WWW'10. New York, USA: ACM Press, 2010:1177-1178.

[17] BOTTOU L, BENGIO Y. Convergence properties of the *K*-means algorithms[C]//Advances in Neural Information Processing Systems 7, NIPS Conference. Denver, USA: MIT Press, 1994:585-592.

[18] ARTHUR D, VASSILVITSKII S. *K*-means++: the advantages of careful seeding[C]//Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'07). New Orleans, USA: SIAM, 2007:1027-1035.

[19] Sustainable Energy Authority of Ireland. CER_electricity_data [EB/OL]. [2022-01-10]. http://www.ucd.ie/issda/.

[20] CALINSKI T, HARABASZ J. A dendrite method for cluster analysis[J]. Communications in Statistics-Simulation and Computation, 1974, 3(1):1-27.

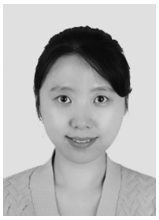
[21] 周世兵,徐振源,唐旭清. *K*-means算法最佳聚类数确定方法[J]. 计算机应用,2010,30(8):1995-1998.

ZHOU Shibing, XU Zhenyuan, TANG Xuqing. Method for determining optimal number of clusters in *K*-means clustering algorithm[J]. Journal of Computer Applications, 2010, 30(8): 1995-1998.

作者简介：

汪　颖（1981—），女，教授，博士研究生导师，博士，研究方向为电能质量与优质供电（**E-mail**：20312028@qq.com）；

杨　维（1996—），男，硕士研究生，研究方向为负荷特性分析与负荷辨识等；



汪　颖

肖先勇（1968—），男，教授，博士研究生导师，博士，研究方向为电能质量与优质供电等（**E-mail**：xiaoxianyong@163.com）；

张　妹（1988—），女，助理研究员，博士，通信作者，研究方向为配电网保护与定位、电能质量与优质供电（**E-mail**：zs20061621@163.com）。

（**编辑**　王锦秀）

Clustering and analysis of electricity consumption behavior of massive users based on network feature dimension reduction of denoising autoencoder and improved mini-batch *K*-means algorithm

WANG Ying, YANG Wei, XIAO Xianyong, ZHANG Shu

(College of Electrical Engineering, Sichuan University, Chengdu 610065, China)

Abstract: The mining and analysis of electricity consumption features of massive users is of great significance for realizing bi-directional interaction between power grid and users. An algorithm suitable for clustering and analysis of electricity consumption behavior of massive users is proposed to reduce the time complexity of the algorithm and improve the load data analysis efficiency of massive users. The electricity consumption behavior features of the users are extracted, and a multi-layer denoising autoencoder network is built to realize dimension reduction of multi-dimensional features. The mini-batch *K*-means algorithm is used for clustering analysis, and the improvements of initial clustering centroid optimization and hyperparameter optimization on the algorithm are carried out to improve the convergence speed and effect of the algorithm, in which, the hyperparameter optimization is carried out with Bayesian optimization algorithm based on Gaussian process. The related indexes of separation degree between the clusters and the cohesion degree within the clusters are used to evaluate the clustering effect. The effective clustering features are screened through mutual information to realize user portraits. The case results show that the proposed method has good performance in feature optimization, clustering effect and convergence speed.

Key words: electricity consumption behavior; feature dimension reduction; clustering analysis; mutual information; mini-batch *K*-means algorithm; hyperparameter optimization; Bayesian optimization

（上接第 137 页　continued from page 137）

Online entry and exit control strategy for single converter station based on DC high speed switch

HUANG Ruhai[1], QIU Defeng[1], LU Jiang[1], DONG Yunlong[1], GAN Zongyue[2]

(1. Nanjing Nari-Relays Electric Co., Ltd., Nanjing 211102, China;
2. EHV Power Transmission Company of China Southern Power Grid Co., Ltd., Guangzhou 510663, China)

Abstract: In multi-terminal UHVDC (Ultra High Voltage Direct Current) projects, it is a good choice to realize DC fault clearing and online entry and exit of converter station by using DC HSS (High Speed Switch) combined with full / half-bridge hybrid-module converter valves. However, the weak DC current breaking capacity of HSS puts forward strict requirements for the control and protection equipment, and the online entry and exit strategy of converter station under this mode has not been reported yet. Based on the electrical characteristics of HSS and the ability of hybrid-module converter valves, the configuration principle of HSS is analyzed and the typical circuit of HSS is designed. Then, the online entry and exit strategy for converter station based on HSS and the cooperative charging scheme of multiple VSC (Voltage Source Converter) converter stations are studied. Finally, in order to ensure the safety of system and equipment, the detailed protection strategies of HSS and their action results are proposed. All of the proposed strategies have been applied in KLL hybrid DC project, and the research results can provide reference for single station entry and exit scheme of multi-terminal DC projects.

Key words: DC high speed switch; online entry and exit; operation mode; VSC-HVDC power transmission; hybrid HVDC power transmission

附录 A：

表 A1 用电行为特征

Table A1 Characteristics of electricity consumption behavior

时段	日负荷指标与定义	指标意义
全天	日最大负荷时刻 T_{Pmax}	反映全天峰值时间
	日最小负荷时刻 T_{Pmin}	反映全天低谷时间
	峰谷相距时间 T_{Pave}	反映峰谷时间跨度
	日最小负荷率 P_{min}/P_{max}	反映负荷波动大小
	日峰谷差率$(P_{max} - P_{min})/ P_{max}$	反映负荷波动程度与电网调峰能力
	日负荷率 P_{ave}/P_{max}	反映负荷变化程度
峰期（08:00—11:00，18:00—21:00）	峰期负载率 $P_{av.peak}/P_{ave}$	反映用电峰期用电的负荷变化波动程度
谷期（00:00—06:00，22:00—24:00）	谷期负载率 $P_{av.val}/P_{ave}$	反映用电谷期用电的负荷变化波动程度
平期（06:00—08:00，11:00—18:00，21:00—22:00）	平期负载率 $P_{av.sh}/P_{ave}$	反映用电平期用电的负荷变化波动程度

注：P_{min} 为日最小负荷，P_{max} 为日最大负荷，P_{ave} 为日平均负荷，$P_{av.peak}$ 为日峰期平均负荷，$P_{av.val}$ 为日谷期平均负荷，$P_{av.sh}$ 为日平期平均负荷。

表 A2 MBKM 算法的超参数

Table A2 Super parameters of MBKM algorithm

参数名称	取值	参数意义	寻优方式
K	≥2	聚类数	Sil，见第 5 节
D_{max_iter}	默认 100	最大迭代次数	设置为默认
N_t	3 次	初始化质心运行算法的次数	设置为默认
b	默认 100	采样包大小	超参数优化
ε	默认 0.01	质心被重新赋值的最大次数比例	超参数优化
β	默认 10	连续采样包个数	设为默认

表 A3 用户行为特征的多层 DAE 网络

Table A3 Multi-layer DAE network of users' behavior characteristics

层名称	输出形状	激活函数	参数个数
输入层(Input)	(None，9)	ReLu	0
全连接层 1($Dense_1$)	(None，9)	ReLu	90
随机失活(Dropout)	(None，9)	—	0
全连接层 2($Dense_2$)	(None，4)	ReLu	40
全连接层 3($Dense_3$)	(None，2)	ReLu	10
全连接层 4($Dense_4$)	(None，4)	ReLu	12
全连接层 5($Dense_5$)	(None，9)	ReLu	45
输出层($Dense_6$)	(None，9)	Tanh	90

注：None 表示输入维度待定；总参数为 287；训练参数为 287；非训练参数为 0。

表 A4 特征降维前、后的聚类效果统计

Table A4 Statistics of clustering effect before and after feature dimension reduction

降维方法	降维后聚类结果与降维特征的指标评估			
	CHI	Sil	DBI	计算时间/s
未降维(标准)	346.652	0.570	1.339	0.048 6
标准 PCA	615.633	0.767	1.026	0.026 9
TSVD	617.545	0.776	0.985	0.289 0
单层 AE	1 036.168	0.814	0.777	0.046 8
多层 AE	2 494.742	0.850	0.459	0.029 9
本文方法(DAE)	3 424.195	0.908	0.460	0.028 9

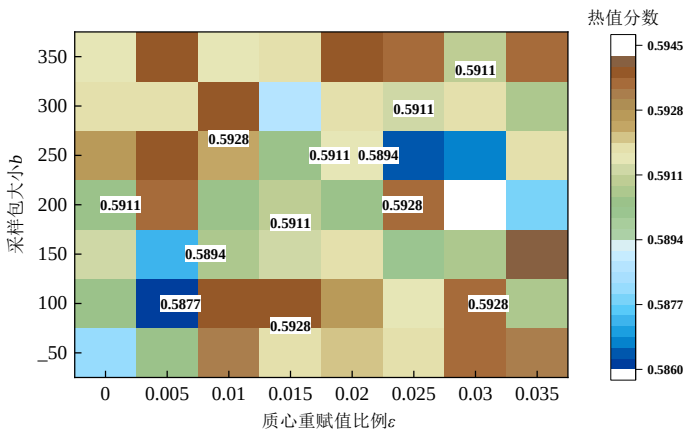


图 A1 不同超参数下 Sil 分数的热图

Fig.A1 Heat map of Sil fraction with different super parameters

表 A5 超参数寻优的部分迭代结果

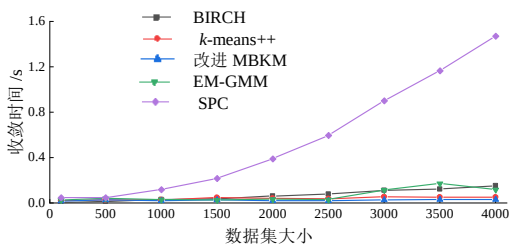
Table A5 Partial iterative results of hyperparameter optimization

迭代次数	优化目标 Sil	b	ε
1	0.682 2	386	0.014 37
2	0.708 9	204	0.039 09
3	0.667 3	131	0.039 11
4	0.681 6	130	0.035 00
⋮	⋮	⋮	⋮
24	0.751 2	122	0.030 96
25	0.685 7	232	0.352 80
26	0.667 9	497	0.114 30
⋮	⋮	⋮	⋮

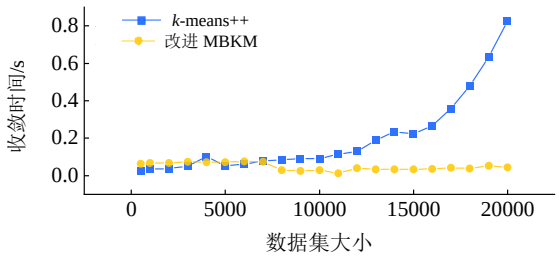
表 A6 改进 MBKM 算法与其他算法的聚类效果对比

Table A6 Comparison of clustering effect between improved MBKM algorithm and other algorithms

聚类方法	样本大小为 500			样本大小为 1 000			样本大小为 2 000			样本大小为 3 000			样本大小为 4 000		
	CHI	Sil	DBI	CHI	Sil	DBI	CHI	Sil	DBI	CHI	Sil	DBI	CHI	Sil	DBI
EM-GMM	333.76	0.374	0.795	563.73	0.378	0.858	855.89	0.381	1.120	1 207.56	0.270 0	1.008	3 618.44	0.280	1.902
SPC	718.10	0.462	0.680	1 217.42	0.379	0.696	2 912.92	0.503	0.585	3 575.03	0.405 0	0.686	4 350.01	0.351	0.742
BIRCH	415.16	0.751	0.462	1 252.83	0.747	0.517	2 196.37	0.551	0.746	3 291.95	0.779 0	0.362	5 446.44	0.692	0.573
k-means++	850.14	0.633	0.621	1 610.86	0.654	0.651	3 110.78	0.516	0.607	4 477.57	0.743 0	0.698	5 953.21	0.665	0.611
传统 MBKM	610.25	0.491	0.680	1 300.11	0.430	0.703	3 035.30	0.497	0.643	4 036.23	0.715 0	0.507	5 470.91	0.404	0.762
改进 MBKM	843.64	0.646	0.559	1 604.67	0.663	0.641	3 077.74	0.5211	0.588	4 148.00	0.737 2	0.618	5 930.57	0.670	0.604



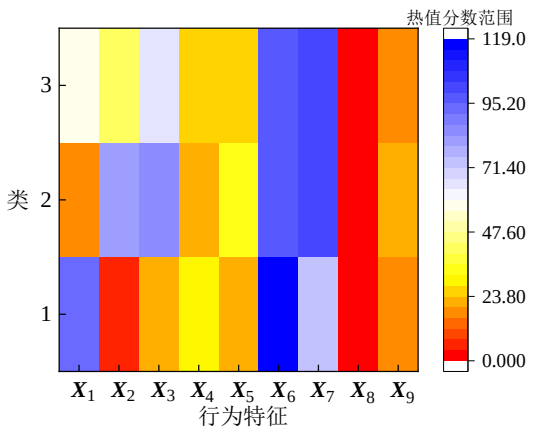
(a)各类算法



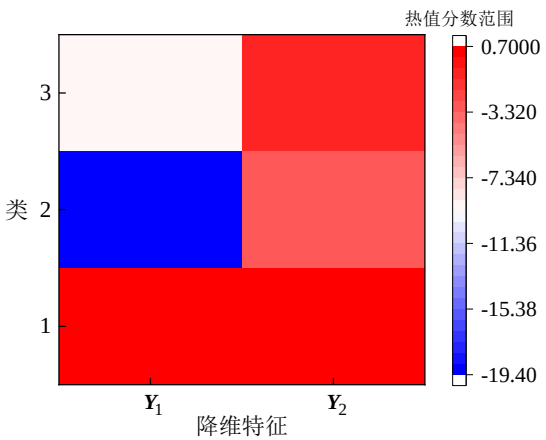
(b)改进 MBKM 算法与 *k*-means++ 算法

图 A2 各类算法在不同数据集大小下的收敛时间

Fig.A2 Convergence time of various algorithms under different sizes of dataset



(a)行为特征



(b)降维特征

图 A3 聚类中心与降维前、后特征的相关图

Fig.A3 Correlation graph of clustering center and features before and after dimension reduction

表 A7 3 类用户的初始特征统计

Table A7 Initial features statistics of three types of users

初始特征	第 1 类用户	第 2 类用户	第 3 类用户
$\boldsymbol{X}_1$ 中心值/%	94.715 9	19.678 3	56.372 1
$\boldsymbol{X}_2$ 中心值/%	5.2840 9	80.321 6	43.627 8
$\boldsymbol{X}_3$ 中心值/%	21.737 9	87.256 3	64.764 4
$\boldsymbol{X}_4$ 中心值	14:43	10:51	13:46
$\boldsymbol{X}_5$ 中心值	02:18	18:28	04:12
$\boldsymbol{X}_6$ 中心值/%	118.876 3	99.118 3	97.981 6
$\boldsymbol{X}_7$ 中心值/%	73.844 1	101.012 3	101.805 4
$\boldsymbol{X}_8$ 中心值/%	0.946 3	0.253 7	0.523 8
$\boldsymbol{X}_9$ 中心值/h	7.944 6	10.379 3	8.732 1

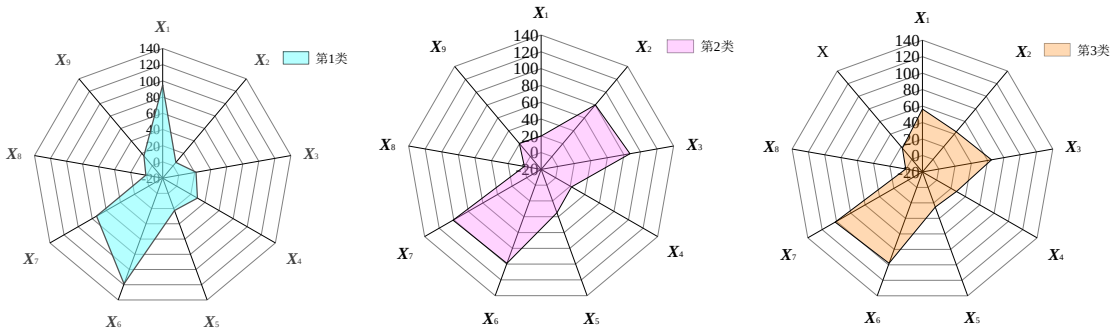


图 A4　3 类用户的用电属性特点

Fig.A4 Electricity consumption attributes of three types of users