DOI: 10.12158/j.2096-3203.2026.02.008

基于 OCSVM 的行业负荷特征异常辨识方法

陈光宇[1]，杨光[1]，施蔚锦[2]，蔡鑫灿[2]，陈婉清[2]，刘昊[1]
（1. 南京工程学院电力工程学院，江苏 南京 211167；
2. 国网福建省电力有限公司泉州供电公司，福建 泉州 362000）

摘　要：为解决近年来用户行业变化特性加剧导致的难以准确辨识用户档案信息变动的问题，文中提出一种基于数据驱动的负荷特征异常辨识方法。首先，提出一种两阶段行业典型负荷形态构建方法，利用基于层次密度的含噪声应用空间聚类（hierarchical density-based spatial clustering of applications with noise, HDBSCAN）提取用户在不同场景下的典型日负荷曲线，并利用改进的 *K*-means 算法对提取出的典型日负荷曲线进行聚类分析，构建行业的典型负荷形态；其次，提出一种多维场景负荷特征异常智能研判方法，通过构造用户的负荷特征，使用熵权法评估行业典型场景的相对重要性，并采用单分类支持向量机（one-class support vector machine, OCSVM）算法量化每个场景下的用户负荷特征的异常程度，通过加权计算得到用户的综合嫌疑得分并排序，从而实现对负荷特征异常用户的准确辨识。最后，采用某地区实际用户数据进行算例验证。仿真结果表明，所提方法在行业典型负荷场景构建及负荷特征异常辨识方面表现出良好的可行性与实用价值。

关键词：数据驱动；负荷特征异常；基于层次密度的含噪声应用空间聚类（HDBSCAN）-改进 *K*-means 算法；多维场景分析；单分类支持向量机（OCSVM）；综合嫌疑得分

中图分类号：TM714　　**文献标志码：**A　　**文章编号：**2096-3203（2026）02-0070-10

0　引言

近年来，随着地区电网的快速发展和新能源渗透率的不断提高[1]，区域内负荷用电特性的复杂性和随机性呈显著上升趋势。尤其在后疫情时代，用户用电行为的快速变化[2-3]以及所属行业的频繁变动，导致原有的行业分类信息更新滞后甚至停滞。在这一背景下，“源-荷”双重不确定性问题愈发突出[4-6]，给电网的精细化调控和管理带来了巨大挑战。因此，准确识别用户行业特征异常，对于提升电网调控的智能化和精细化具有重要的现实意义。

近年来，国内外学者对电力负荷分类与辨识进行了大量研究。文献[7]先采用多维尺度分析对楼宇负荷数据进行降维，再使用高斯混合模型进行聚类，显著提高了聚类效率；文献[8]提出一种基于灰狼算法优化的模糊 C 均值算法，提高了对初始聚类中心的全局搜索能力；文献[9]提出一种基于空间密度聚类和 *K*-shape 算法的两阶段负荷聚类方法，提高了对城市综合体负荷的聚类精度；文献[10]通过分析用户的日负荷曲线与典型曲线之间的相似度，评估其为养殖行业用户的可能性；文献[11-12]通过阈值方法分别辨识路灯专变异常用户和线损异常用户；文献[13]对学校用户不同时段的负荷进行聚类，并与特定用电行为模式进行对比，检测是否存在用电异常；文献[14]提出一种基于剪枝策略的密度峰值聚类方法，用于划分行业典型用电负荷并识别其中的信息异常用户。上述文献结合用户的细分行业信息，采用曲线相似度或根据特征指标设定阈值等方法对用户异常特征进行检测，取得了较好的效果。然而，这些方法也存在不足，比如特征考虑较少、未充分考虑用户在不同时间段或场景下的用电行为差异，尤其应用于用电行为多样且负荷特征比较复杂的行业时，仅靠单一的负荷曲线或简单的相似度度量会存在一定的局限性。此外，行业异常用户标签的缺失导致参数选择过程难以系统化评估，常依赖经验判断，难以保证参数的最优性。因此，迫切需要开发一种更全面、更精准的结合行业信息的负荷特征异常辨识方法。

针对电网用户行业分类不准确且频繁变动导致的台账信息更新滞后以及运维工作量较大的问题，文中提出一种基于单分类支持向量机（one-class support vector machine, OCSVM）的行业负荷特征异常辨识方法。首先，利用基于层次密度的含噪声应用空间聚类（hierarchical density-based spatial clustering of applications with noise, HDBSCAN）和改进的 *K*-means 算法提取行业用户的典型用电场景和用电模式，构建一个能够反映行业用电特性的特征画像库。其次，为客观评估各电力使用场景的重要性，

收稿日期：2025-06-29；修回日期：2025-10-23
基金项目：国家自然科学基金资助项目（52107098）

文中应用熵权法确定场景权重, 并结合多维场景分析与 OCSVM 算法全面检测各用电场景和维度下的异常特征。最后, 通过计算用户的综合异常评分, 有效识别和定位行业异常用户。通过对某地区橡胶和塑料制品行业用户实际用电数据的算例分析, 验证所提方法在构建行业典型负荷形态及辨识负荷特征异常方面的有效性。该方法不仅能够准确发现用户分类异常, 保障电网收益, 还能显著减轻运行维护人员的劳动强度和时间成本, 为快速筛查特征变动以及识别行业信息异常用户提供强有力的支持。

1 基于 HDBSCAN-改进 *K*-means 的行业典型负荷形态构建

1.1 HDBSCAN 聚类算法

HDBSCAN 是由 Campello 等人开发的一种基于密度的聚类算法[15], 是对密度空间聚类(density-based spatial clustering of applications with noise, DBSCAN)的改进。与传统的 DBSCAN 算法相比, HDBSCAN 在调参上具有显著优势, 只需要选择最小簇大小和最小样本数两个参数即可。同时, HDBSCAN 能够自动识别任意形状和不同密度的集群, 并去除噪声数据, 不需要预设聚类数量和聚类中心, 可自动确定最佳聚类数目。然而, 该算法也存在计算复杂度较高等缺陷。

HDBSCAN 算法的核心原理在于将密度变换与层次聚类技术相结合, 通过基于簇稳定性的方法提取平面聚类, 以有效地扩展传统 DBSCAN 算法的功能, 其具体步骤如下:

(1) 在 HDBSCAN 中, 首先计算每个数据点的核心距离, 即数据点到其邻域内第 g 个最近邻点的距离, 此距离反映了数据点在其邻域内的密集程度。随后, 定义可达距离为 a、b 两个数据点间的欧氏距离与其核心距离的较大值:

$$d_{\text{reach},g}(a,b)=\max\{d_{\text{core},g}(a),d_{\text{core},g}(b),d(a,b)\} \tag{1}$$

式中: $d_{\text{core},g}(a)$、$d_{\text{core},g}(b)$分别为 a 点和 b 点的核心距离; $d(a,b)$为 a、b 点之间的欧氏距离。

(2) 接着, 算法利用可达距离信息通过 Prim 算法[16]高效构建最小生成树(minimum spanning tree, MST)。该树结构映射了数据点间基于密度的紧密连接性。在构建过程中, 算法逐步添加当前 MST 与尚未连接顶点之间的最小可达距离边, 从而形成覆盖所有数据点的 MST, 如图 1 所示。

(3) 构建层次聚类树时, 算法设定最小簇大小作为剪枝阈值, 自上而下遍历聚类树的所有节点。

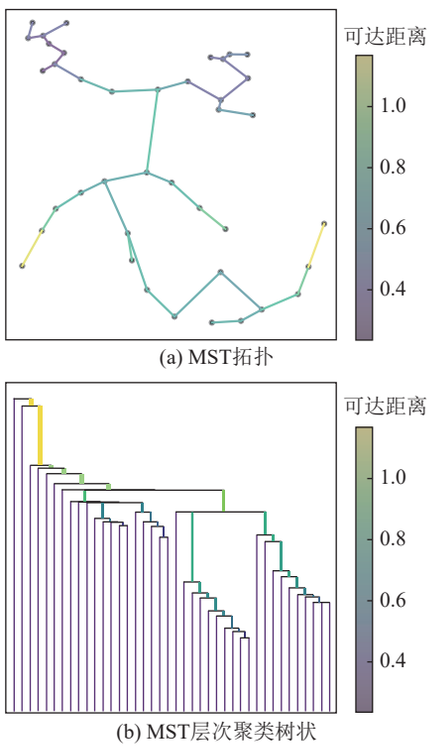

图 1 g 为 5 时的 MST 示意

Fig.1 Schematic diagram of MST with g=5

若子节点数达到或超过剪枝阈值则保留, 未达到则删除, 从而形成一个优化后的压缩聚类树。

(4) 为选择聚类树中合适的簇, 引入可达距离的倒数λ作为密度, 用来计算每个簇的稳定性, 稳定性越大表示该簇结构越稳定, 簇C_i的稳定性$\theta(C_i)$的计算方法如下:

$$\lambda=\frac{1}{d_{\text{reach}}} \tag{2}$$

$$\theta(C_i)=\sum_{x\in C_i}(\lambda_x-\lambda_{\text{birth}}) \tag{3}$$

式中: d_{reach}为可达距离; λ_x为样本点 x 加入簇 C_i 时的λ值; λ_{birth}为该簇形成时的λ值。

(5) 通过对簇稳定性的评估, 算法选择最合适的簇进行划分。此外, 自动识别密度较低的数据点并将其标记为噪声, 最终输出包含聚类标签和噪声点的聚类结果。

1.2 两阶段行业典型负荷形态构建方法

传统 *K*-means 算法使用欧氏距离计算样本与簇中心之间的距离, 这使得其对离群点非常敏感。离群点可能会显著偏移簇中心, 从而影响整个聚类结果。为避免离群点的影响, 文中采用基于 *K*-means 聚类的改进算法, 首先通过 *K*-means 算法对数据进行初步聚类, 计算公式如下:

$$\mu_k=\frac{1}{|C_k|}\sum_{x\in C_k}x \tag{4}$$

式中：μ_k为初始的聚类中心；$|C_k|$为第 k 个聚类中样本点的数量。

然后，计算每个样本点 x 与其对应簇中心μ_k的距离，并设定阈值τ，识别并过滤距离超过该阈值的异常样本。接着，使用二次聚类对过滤后的数据进行聚类中心的更新，如式(5)所示。

$$\mu_k' = \frac{1}{|C_k'|}\sum_{x\in C_k', d(x,\mu_k)\leqslant\tau} x \tag{5}$$

式中：μ_k'为更新后的聚类中心；$|C_k'|$为更新后第 k 个聚类中样本点的数量。

最后，将所有数据匹配到相应的簇中，获得最终的聚类中心和结果，如式(6)所示。该方法能够显著提高聚类中心的鲁棒性和准确性，有效避免异常点对聚类中心的影响。

$$A(x) = \arg\min\ d(x,\mu_k') \tag{6}$$

式中：$A(x)$为将样本点 x 分配到最近的簇中。

文中采用一种两阶段的负荷直接聚类方法，对用户的负荷进行典型负荷形态聚类。第一阶段使用 HDBSCAN 算法提取每个用户在特定时间段内的典型用电模式；第二阶段采用 K-means 算法对提取出的典型场景进行二次聚类，获得行业的典型用电场景。行业典型负荷形态构建的具体步骤如图 2 所示。

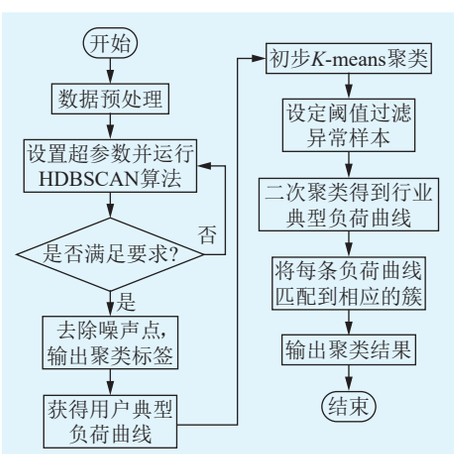

图 2　行业典型负荷形态构建流程

Fig.2　Flow chart of industry typical load profile construction

2　负荷特征构造及场景评估方法

2.1　用户特征构造

特征提取旨在识别和评估不同负荷场景下的用电特性，获取原始数据中隐藏的额外信息，从而提高行业特征异常识别的效率[17]。文中基于提取负荷典型形态构建其峰谷时段、相似性、波动性等相关指标，可以较为全面地体现用户的典型用电特征。将包含 N 个用户的用电数据表示为数据集$X=\{x_1,x_2,\cdots,x_n|n=1,2,\cdots,N\}$。用户 n 在第 i 个典型场景下的负荷曲线表示为$\{x_{1,n,i},x_{2,n,i},\cdots,x_{T,n,i}\}$，其中 T 为负荷序列的长度。

2.1.1　负荷特性指标

为全面刻画用户用电的负荷特性，文中定义表 1 所示的负荷特性指标[18]。

表 1　负荷特性指标

Table 1　Load characteristic indexes

负荷特征	定义	时段
日负荷率	$P_1=x_{av}/x_{max}$	00:00—24:00
尖峰负载率	$P_2=x_{sharp,av}/x_{av}$	11:00—12:00、17:00—18:00、20:00—21:00
高峰负载率	$P_3=x_{peak,av}/x_{av}$	09:00—11:00、15:00—17:00、19:00—20:00
平期负载率	$P_4=x_{shoulder,av}/x_{av}$	07:00—09:00、12:00—15:00、18:00—19:00、21:00—23:00
低谷负载率	$P_5=x_{valley,av}/x_{av}$	23:00—24:00、00:00—07:00

注：x_{av}为日负荷均值；x_{max}为日负荷最大值；$x_{sharp,av}$为尖峰时段的负荷均值；$x_{peak,av}$为高峰时段的负荷均值；$x_{shoulder,av}$为平期时段的负荷均值；$x_{valley,av}$为低谷时段的负荷均值。

2.1.2　相似性指标

（1）用户典型用电负荷与行业典型负荷之间的欧式距离。

（2）用户典型用电负荷与行业典型负荷之间的动态时间弯曲(dynamic time warping, DTW)距离。

2.1.3　用电特性指标

（1）用户的用电倍率，即用户电能表所对应互感器的倍率。

（2）用户典型场景下的用电量。

2.1.4　波动性指标

（1）用户典型日用电负荷曲线的峰值数量。对于每个x_t，检查其是否为峰值需要满足以下要求：$\{x_{t-m},\cdots,x_{t-1},x_t,x_{t+1},\cdots,x_{t+m}\}$，$t>m$且$T-t\geqslant m$，其中$x_t$为用户典型日用电负荷曲线在 t 时刻的负荷值；m 为时间窗口的长度，文中 m 选择 4。

（2）用户典型日用电负荷曲线的排列熵[19]。排列熵用于衡量用户的典型日用电负荷曲线的波动性和复杂度，计算如下：

$$H(X_n) = -\sum_{j=1}^{q} p_j \ln(p_j) \tag{7}$$

式中：$H(X_n)$为用户 n 的典型日用电负荷曲线的排列熵；q 为可能的排列模式数量，对于给定的嵌入维度 l，$q=l!$，文中 l 选择 4；p_j为第 j 种排列模式在时间序列中出现的概率。计算出来的排列熵值越高，说明时间序列的复杂性和随机性越高。

2.2　典型场景评估方法

不同的行业典型场景反映了用户在不同时间

段和条件下的用电情况, 不同场景中的重要性因负荷的不同而异。高峰负荷场景的重要性显然更高, 而低谷负荷场景的重要性相对较低。通过提取多维度特征并应用熵权法, 能够客观地衡量各负荷场景的重要性, 为后续的行业特征异常研判提供有效依据。

熵权法是一种基于信息熵理论的客观加权方法, 常用于多指标综合评价[20-21]。该方法利用信息熵衡量各指标的差异性和不确定性, 以确定各指标的权重, 进而实现对评价对象的综合排序。权重越大, 说明该指标的信息量越大, 对综合评价的贡献也越大。文中使用第 2.1 节构建的典型特征对不同场景的重要性进行评估。

2.3 特征维度规约

主成分分析(principal component analysis, PCA)是一种广泛使用的特征降维方法[22-23]。PCA 通过构建新的特征(主成分)来最大化数据的方差, 同时减少原始数据的维度。首先, 对原始数据进行标准化处理, 然后计算协方差矩阵, 并求解其特征值和特征向量。通过选择特征值较大的前几个主成分作为新特征, 可以有效降低数据维度, 去除特征中的噪声和冗余信息, 同时保留大部分原始信息, 为后续的异常检测提供高质量的新特征。

3 基于 OCSVM 的行业特征异常研判方法

3.1 OCSVM 算法

在众多异常检测方法中, 单类支持向量机因其良好的性能和广泛的应用而备受关注[24-25]。OCSVM 是一种基于支持向量机(support vector machine, SVM)的无监督学习算法, 其目标是通过最大化分类超平面与数据点之间的距离, 将正常数据与异常数据区分开, 使得样本能够包裹尽可能多的正常样本。OCSVM 使用核函数的非线性映射能力, 可以将数据从原始空间$\mathbf{R}^d$映射到高维空间$\mathbf{R}^D$, d、D 为空间的维度。OCSVM寻找最优超平面的过程可以表示为一个二次规划问题, 如式(8)、式(9)所示。

$$\min_{\boldsymbol{w},\rho,\xi_r}\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{1}{\nu n_{\mathrm{s}}}\sum_{r=1}^{n_{\mathrm{s}}}\xi_r-\rho \tag{8}$$

$$\begin{cases}\text{s.t.}\quad \boldsymbol{w}\phi(x_r)\geqslant\rho-\xi_r\\ \quad\ \xi_r\geqslant 0\\ \quad\ r=1,2,\cdots,n_{\mathrm{s}}\end{cases} \tag{9}$$

式中: $\boldsymbol{w}$为权重向量; ρ为偏移量; ξ_r为松弛变量; $\phi(x_r)$为特征映射函数, 文中选择高斯核函数; ν为用于权衡正则化项和惩罚项的超参数; n_{s}为样本数量。优化问题的目标是通过最小化目标函数来找到一个最优的超平面, 同时满足上述约束条件, 从而在高维空间中实现对数据的有效分类。

OCSVM 的决策函数$f(x)$定义如下:

$$f(x)=\mathrm{sign}(g(x))=\begin{cases}1 & g(x)\geqslant 0\\ -1 & g(x)<0\end{cases} \tag{10}$$

式中: $g(x)=\boldsymbol{w}^{\mathrm{T}}K(x)-\rho$表示中间计算的结果, $K(x)$为将输入样本点 x 映射到高维特征空间的核函数。$f(x)$为 1 则被认为是正常数据, $f(x)$为−1 则被认为是异常数据。

为量化每个样本的异常程度, 使用式(11)计算样本的离群程度。该方法主要通过在 OCSVM 中引入归一化处理方法, 提高异常检测的准确性和敏感性, 特别是当$g_{\max}$接近 1.0 时, 能够更精确地区分接近正常值的数据点。

$$f'(x)=\frac{g_{\max}-g(x)}{|g_{\max}|} \tag{11}$$

式中: $f'(x)$为归一化后的函数值; $g_{\max}$为所有训练样本中决策函数的最大值, 如图 3 所示。通过比较输入数据点的决策函数值与最大值之间的差异, 可以定量评估其离群程度。正常点的分数接近 1.0, 而离群点的分数则显著高于 1.0, 越离群的数据异常得分越高。通过这种评分机制, 可以根据数据点的得分动态调整检测阈值, 从而更灵活地适应不同的数据分布和环境变化。这种方法不仅增强了异常检测的灵敏度, 也提高了检测结果的可靠性和适用性。

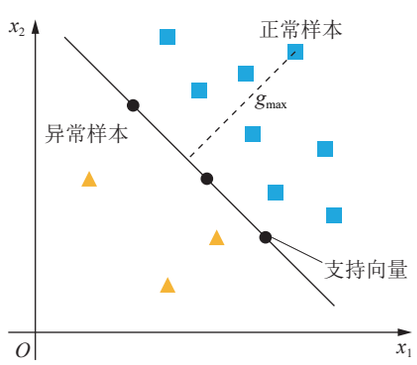

图 3 OCSVM 分类示意

Fig.3 Schematic diagram of OCSVM classification

3.2 基于用户综合评判的异常指标构建

异常得分越高表示用户是行业特征异常用户的嫌疑程度越高。通过综合考虑每个用户在不同典型场景下的负荷特征、场景的重要性以及用户不同场景的比重, 计算每个用户的综合异常评分:

$$s_n=\sum_{i=1}^{I}w_iq_ia_i \tag{12}$$

式中: s_n为用户 n 的综合异常评分; w_i为第 i 个典型场景包含的天数占所有天数的比重; q_i为第 i 个典型场景的评估结果; a_i为用户在第 i 个典型场景下

的异常程度；I 为典型场景数量。根据综合嫌疑评分对用户进行排序，加权得分越高的用户，其用电特征越可能存在异常，从而实现对特征异常用户的精准定位。

3.3 异常研判框架

综上所述，文中算法的流程如下。

（1）数据获取与预处理。获取用户用电数据并进行清洗、去噪、归一化处理。

（2）典型用电场景提取。使用 HDBSCAN-改进 *K*-means 聚类算法提取行业典型用电场景。

（3）用电特征构建。基于提取的典型场景，提取用户在各场景下的用电特征。

（4）异常检测与评分计算。利用 OCSVM 算法对用户的用电特征进行异常检测，并计算每个用户在各场景下的异常得分。

（5）异常用户定位与评估。根据异常得分对用户进行排序，精准定位异常得分较高的用户，最终实现异常用户的综合评估与异常识别。

具体异常研判框架如图 4 所示。

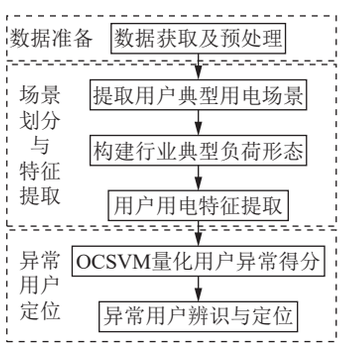

图 4 异常特征研判框架

Fig.4 Framework for anomaly feature analysis

4 算例分析

为验证文中算法的有效性，采用南方某地区 867 个橡胶和塑料制品行业用户 2023 年 3 月 1 日—2023 年 11 月 26 日的用电负荷数据进行算例分析，其中采样频率为 15 min，每天有 96 个采样点。

4.1 行业典型负荷形态构建

对每个用户的用电负荷数据进行预处理，包括数据清洗和归一化处理。随后，将预处理后的数据使用 HDBSCAN 算法进行聚类分析，在初始参数的设定上，参数过大可能会导致聚类中无法形成有效的簇，并导致过多的数据被划分为噪声；参数过小可能因数据随机波动导致创建过多的簇。对不同的参数进行仿真并计算其平均轮廓系数和平均戴维森堡丁指数（Davies-Bouldin index, DBI）[26]，结果如图 5 所示。

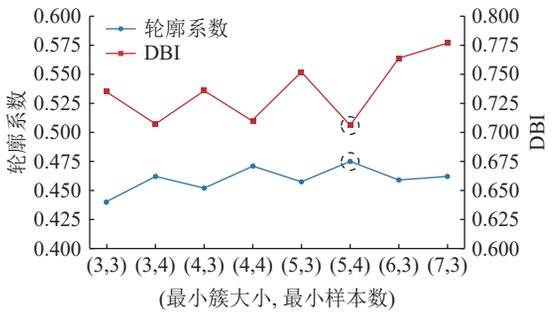

图 5 不同参数下的轮廓系数和 DBI

（噪声点比例<30%）

Fig.5 Silhouette coefficient and DBI under different parameters (noise ratio < 30%)

考虑负荷曲线的空间分布特点，一般用户的异常负荷曲线数量不会超过 30%，故选择噪声点在 30% 以下且平均轮廓系数最大、DBI 最小的参数配置作为 HDBSCAN 算法的初始参数设定。具体参数包括最小簇大小为 5，最小样本数为 4，其他参数保持算法的默认设置。

HDBSCAN 通过构建密度关联的聚类树，识别出每个用户的不同负荷场景。为避免用户聚类得到的簇数过多，相应地增加最小簇大小和最小样本数，使用户的场景数量在 2~6。聚类得到部分用户的典型负荷如图 6 所示。

从图 6 可以看出，每个用户存在多种不同的用电模式。用户 4 表现出明显的双峰特征，白天工作时间用电量较高，而在晚上和凌晨用电量较低，是典型的日间生产型企业。用户 10 为夜间生产型的两班制企业，白天负荷较低而夜间负荷较高。用户 22 则同时包含日间生产和夜间生产的特性，显示出多种不同的用电模式，其用电特性较为复杂。

经过第一阶段 HDBSCAN 算法聚类后，共得到 2 829 条用户典型负荷曲线。在第二阶段，将每条用户典型负荷曲线视为单独的个体，使用改进的 *K*-means 算法进行聚类。首先，采用手肘法来确定第二阶段 *K*-means 的最佳聚类数目为 12。然后，对数据进行初步聚类并计算每条曲线到聚类中心的欧氏距离。设定阈值为平均值加上两倍的标准差，使用小于阈值的负荷曲线对聚类中心进行更新。最后，将所有负荷曲线匹配到相应的聚类中心，得到 12 种不同的行业典型负荷形态，其中部分行业典型形态如图 7 所示。

由图 7 可知，橡胶与塑料制品行业的日负荷曲线清晰地展示了该行业的电力需求特征。该行业中的生产设备，如注塑机[27]、挤出机和压延机通常需要长时间连续运行，该模式使电力需求呈现高稳

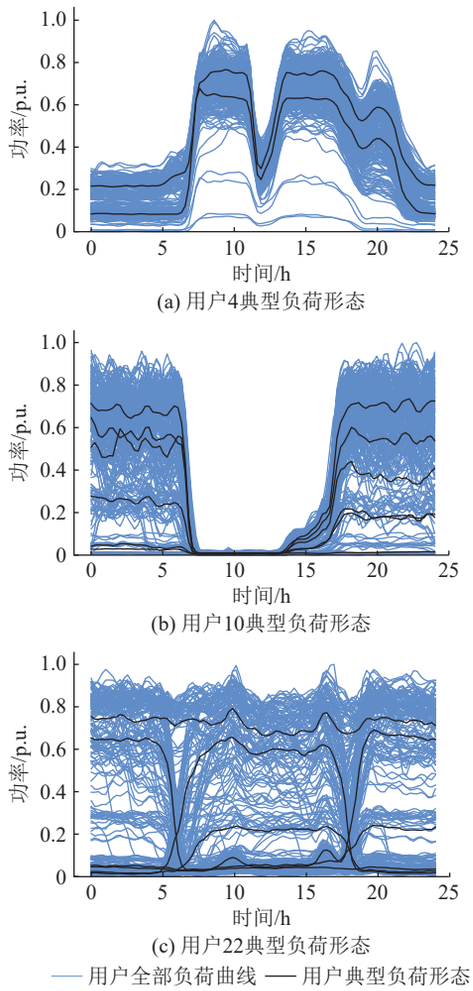

图 6　用户典型负荷形态

Fig.6　Typical customer load curves

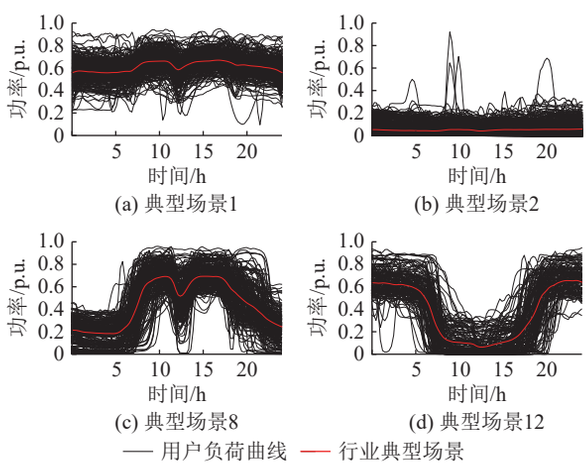

图 7　行业典型负荷形态

Fig.7　Typical industry load curves

定性和持续性。因此，在场景 1 和场景 8 中可以观察到相对较高的负荷水平。相较之下，场景 2 的负荷曲线波动较小，可能是工厂停工或假期中仅保持少量设备运行的低负荷状态。场景 12 的曲线展现出明显的避峰特性，这体现了部分用户对电价波动的高度敏感性，企业可能会选择在电价较低的时段增产以优化运营成本。行业典型负荷形态不仅展示了企业对电力市场变化的适应策略，也反映了用户对峰谷电价政策的积极响应，为电网调度部门精细化管理提供重要依据。

4.2　行业负荷特征异常辨识方法

根据 2.1 节的特征提取方法，提取每个用户典型负荷形态的用电特征情况，得到每个场景 11 维的用电特征。信息熵衡量的是指标信息量的大小和不确定性[28]，熵值越大，表示该指标的信息越分散，不确定性越高；熵值越小，表示该指标的信息越集中，不确定性越低。根据行业典型负荷场景的划分结果，计算每个场景负荷特征的平均值，然后使用熵权法对该场景的重要性进行评估。计算得到的结果如表 2 所示。

表 2　场景综合评分

Table 2　Comprehensive scenario score

场景	综合评分	场景	综合评分
1	0.597	7	0.474
2	0.282	8	0.468
3	0.809	9	0.288
4	0.549	10	0.270
5	0.303	11	0.299
6	0.240	12	0.400

计算结果反映了不同负荷场景在用电特征上的重要性。高得分场景(如场景 1)显示用户在高峰负荷时段具有显著的用电特征，而低得分场景(如场景 2)则对应于低负荷或非关键用电行为。计算结果表明，基于熵权法的综合评价方法可以客观评估不同负荷场景的相对重要性，并为后续的行业特征异常识别提供可靠依据。

为避免特征冗余问题，使用 PCA 对所得特征进行降维处理，选择前两个能解释大部分特征方差并保留尽可能多信息的主成分，将其作为 OCSVM 模型的输入。OCSVM 通常用于无监督学习场景，这意味着没有预先标注的正常或异常数据点。在这种情况下，很难通过传统的监督学习方法(如交叉验证)来优化参数。文献[29]提出一种无监督的方法，利用数据点及其 k 最近邻之间的距离分布来估计 OCSVM 中超参数的值。该方法对不同类型的数据具有鲁棒性，与传统方法相比，显著减少了计算复杂度。在对数据进行 PCA 降维处理后，利用 OCSVM 对降维后的特征进行拟合，以识别不同行业场景中的异常行为和模式，并使用式(11)和式(12)计算用户的异常评分。

以用户 22 为例进行异常得分计算。该用户包

含 6 条典型用电负荷曲线，各个变量的计算结果如表 3 所示。对用户每个场景的异常得分相加，得到该用户的综合异常评分为 0.293。

表 3　用户 22 各场景异常得分情况

Table 3　Anomaly scores for different scenarios of user 22

用户典型场景	包含曲线数量	曲线所占比例	异常程度	所属行业场景	场景评分	综合得分
1	52	0.227	0.566	3	0.809	0.104
2	47	0.205	0.464	12	0.400	0.038
3	27	0.118	0.840	5	0.303	0.030
4	8	0.035	1.001	11	0.299	0.010
5	89	0.389	0.953	2	0.282	0.104
6	6	0.026	0.973	2	0.282	0.007

图 8 展示了场景 5 中用户 22 和用户 87 的典型负荷特征对比。根据表 3 可知，用户 22 的典型曲线 3 在场景 5 中评分为 0.303，说明在该场景下其负荷特征与行业典型特征相近。相对而言，用户 87 的典型曲线 2 在同一场景中的评分为 1.72，显著高于离群数值 1，表明其在多个特征维度上与行业典型模式存在显著偏差，图 8 中的可视化分析进一步证实了该现象。

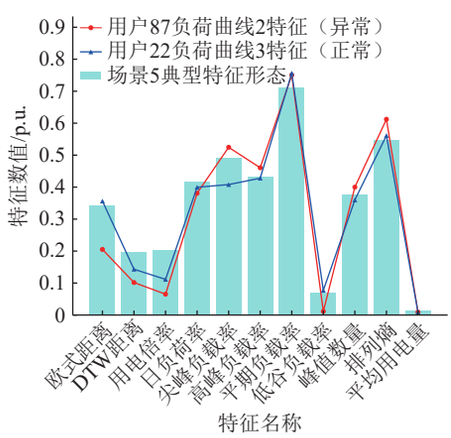

图 8　场景 5 典型特征形态

Fig.8　Typical feature patterns in scenario 5

为证明文中方法的合理性，使用用户单个典型场景（即用户最典型的用电场景，表明用户在该用电场景下的负荷曲线最多，对应用户 22 的场景 5）和多个典型场景（即考虑用户全部用电场景，对应用户 22 的场景 1—6）的负荷曲线进行行业特征异常检测。使用该方法计算所有用户在两种情况下的异常得分，并按照异常得分从大到小的顺序进行排序，前 100 名用户的异常得分情况如图 9 所示。

从图 9 可以发现，随着用户异常得分排序的后移，两条曲线的异常得分开始快速减小，并在综合异常得分接近 0.8 时趋于平缓。这表明大多数用户

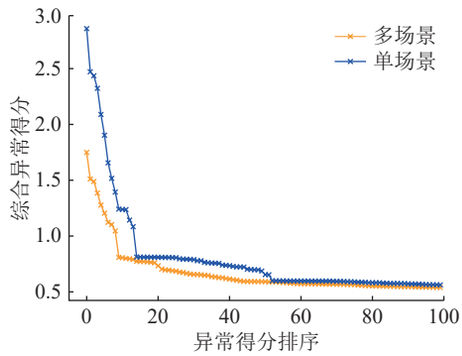

图 9　异常得分排序

Fig.9　Anomaly score ranking

的异常得分较低，只有少数用户的异常特征比较显著。只要对异常得分较高的用户进行就地排查，即可检查出大部分行业特征异常用户[30]。用户在单场景下的异常得分相对较高，最高得分为 2.858，表明某些用户在特定场景下的异常行为非常显著，而在多场景下的得分相对较低，最高得分为 1.749。这是因为多场景方法综合考虑了用户在多个场景下的行为特征，使得异常得分更加综合和全面。

为核实其中行业档案错误的情况，分别对考虑单场景的 OCSVM（single-scenario OCSVM, SS-OCSVM）和考虑多场景的 OCSVM（multi-scenario OCSVM, MS-OCSVM）两种情况下异常得分大于 0.81 的用户进行就地排查，排查结果见表 4 和表 5。

表 4　单场景异常排查结果（按单场景得分排序）

Table 4　Single-scenario anomaly investigation results (sorted by single-scenario score)

用户编号	单场景异常得分	多场景异常得分	是否为档案异常用户
a	2.858	0.568	否
b	2.472	1.103	是
c	2.436	0.549	否
d	2.324	1.749	是
e	2.089	1.385	否
f	1.903	1.511	是
g	1.656	1.277	是
h	1.517	1.489	否
i	1.394	0.397	否
j	1.241	0.694	否
k	1.238	1.205	是
l	1.237	1.123	是
m	1.143	1.046	是
n	1.084	0.414	是

从排查结果可以发现，单场景方法在特定场景下识别异常用户的能力较强，但可能会因为忽略其他场景的信息而导致误报。例如，用户 a 在单场景

表 5 多场景异常排查结果（按多场景得分排序）

Table 5 Multi-scenario anomaly investigation results (sorted by multi-scenario score)

用户编号	单场景异常得分	多场景异常得分	是否为档案异常用户
d	2.324	1.749	是
f	1.903	1.511	是
h	1.517	1.489	否
e	2.089	1.385	否
g	1.656	1.277	是
k	1.238	1.205	是
l	1.237	1.123	是
b	2.472	1.103	是
m	1.143	1.046	是

评估中异常得分为 2.858，但在多场景评估中得分仅为 0.568，其整体行为并未表现出异常，所以未被识别为特征异常用户，排查结果也证明了该用户并不是档案异常用户。同时，大部分使用单场景方法识别出的真实异常用户，也能通过多场景方法进行准确识别。具体来说，表 5 中被单场景方法识别为档案异常用户的 8 个用户中，有 7 个用户在多场景评分中得到相对较高的异常评分，并被识别为特征异常用户。

在不考虑各场景权重及用户负荷曲线的具体异常程度的前提下，若仅利用 OCSVM 模型对用户的每个场景负荷特征进行研判，对所有场景都被标记为异常的 16 个用户进行排查，得到的结果如表 6 所示。

表 6 基于不同方法的 OCSVM 模型精确率对比

Table 6 Comparison of precision rates for OCSVM models based on different methods

检测方法	排查用户数量	实际异常用户数量	精确率/%
OCSVM	16	5	31.25
SS-OCSVM	14	8	57.14
MS-OCSVM	9	7	77.78

表 6 中显示，MS-OCSVM 方法的精确率高达 77.78%，显著优于其他两种方法。较高的精确率证明了 MS-OCSVM 在整合多场景信息后，能够显著降低误报率，并有效地识别出其他单场景方法可能遗漏的异常，进一步验证了该方法在实际应用中的有效性和可靠性。

5 结语

为解决电网用户行业分类不准并且变动频繁导致台账信息更新不及时的问题，文中提出一种基于 HDBSCAN 和 *K*-means 的两阶段聚类方法，构建橡胶和塑料制品行业的典型负荷形态，并通过考虑多维场景和 OCSVM 算法实现用户特征异常的智能辨识和定位，提高了检测的准确性并减少了误判概率，可以极大地减轻运维人员的人力成本和时间成本，为运维人员快速筛查特征异常用户提供有效支撑。

下一步的研究重点将集中在研究更多不同行业用户特征异常检测，并研究自适应方法在行业特征异常检测中的应用，进一步提高行业特征异常检测的效率和精度。

参考文献：

[1] 卓振宇，张宁，谢小荣，等. 高比例可再生能源电力系统关键技术及发展挑战[J]. 电力系统自动化，2021，45(9)：171-191.
ZHUO Zhenyu, ZHANG Ning, XIE Xiaorong, et al. Key technologies and developing challenges of power system with high proportion of renewable energy[J]. Automation of Electric Power Systems, 2021, 45(9): 171-191.

[2] 陈启鑫，吕睿可，郭鸿业，等. 面向需求响应的电力用户行为建模：研究现状与应用[J]. 电力自动化设备，2023，43(10)：23-37.
CHEN Qixin, LÜ Ruike, GUO Hongye, et al. Electricity user behavior modeling for demand response: research status quo and applications[J]. Electric Power Automation Equipment, 2023, 43(10): 23-37.

[3] 张凯瑞，明昊，高赐威. 基于 CPSS 视角的需求响应能力评估综述[J]. 电力科学与技术学报，2024，39(1)：28-46.
ZHANG Kairui, MING Hao, GAO Ciwei. A review of demand response capability assessment based on CPSS perspective[J]. Journal of Electric Power Science and Technology, 2024, 39(1): 28-46.

[4] 董旭柱，华祝虎，尚磊，等. 新型配电系统形态特征与技术展望[J]. 高电压技术，2021，47(9)：3021-3035.
DONG Xuzhu, HUA Zhuhu, SHANG Lei, et al. Morphological characteristics and technology prospect of new distribution system[J]. High Voltage Engineering, 2021, 47(9): 3021-3035.

[5] 代心芸，陈皓勇，肖东亮，等. 电力市场环境下工业需求响应技术的应用与研究综述[J]. 电网技术，2022，46(11)：4169-4186.
DAI Xinyun, CHEN Haoyong, XIAO Dongliang, et al. Review of applications and researches of industrial demand response technology under electricity market environment[J]. Power System Technology, 2022, 46(11): 4169-4186.

[6] 陈湘元，吴公平，龙卓，等. 考虑源荷不确定性及用户侧需求响应的综合能源系统多时间尺度优化调度[J]. 电力科学与技术学报，2024，39(3)：217-227.
CHEN Xiangyuan, WU Gongping, LONG Zhuo, et al. Multi-time scale optimal dispatch of integrated energy systems considering source-load uncertainty and user-side demand response [J]. Journal of Electric Power Science and Technology, 2024,

39(3): 217-227.

[7] 张美霞, 李丽, 杨秀, 等. 基于高斯混合模型聚类和多维尺度分析的负荷分类方法[J]. 电网技术, 2020, 44(11): 4283-4296.
ZHANG Meixia, LI Li, YANG Xiu, et al. A load classification method based on Gaussian mixture model clustering and multi-dimensional scaling analysis[J]. Power System Technology, 2020, 44(11): 4283-4296.

[8] 吴亚雄, 高崇, 曹华珍, 等. 基于灰狼优化聚类算法的日负荷曲线聚类分析[J]. 电力系统保护与控制, 2020, 48(6): 68-76.
WU Yaxiong, GAO Chong, CAO Huazhen, et al. Clustering analysis of daily load curves based on GWO algorithm[J]. Power System Protection and Control, 2020, 48(6): 68-76.

[9] 魏勇, 李学军, 李万伟, 等. 基于空间密度聚类和 *K*-shape 算法的城市综合体负荷模式聚类方法[J]. 电力系统保护与控制, 2021, 49(14): 37-44.
WEI Yong, LI Xuejun, LI Wanwei, et al. Load pattern clustering method of an urban complex based on DBSCAN and *K*-shape algorithm[J]. Power System Protection and Control, 2021, 49(14): 37-44.

[10] 吴郅君, 殷新博, 陈中, 等. 基于模糊聚类曲线相似度的负荷用户识别方法[J]. 电力工程技术, 2019, 38(3): 151-156.
WU Zhijun, YIN Xinbo, CHEN Zhong, et al. Identification method of load customers based on similarity of fuzzy clustering curves[J]. Electric Power Engineering Technology, 2019, 38(3): 151-156.

[11] 付青, 匡文凯, 薛阳, 等. 基于高斯核密度估计法的路灯窃电检测方法[J]. 电力系统及其自动化学报, 2021, 33(10): 18-23.
FU Qing, KUANG Wenkai, XUE Yang, et al. Detection method for street lamp electricity theft based on Gaussian kernel density estimation[J]. Proceedings of the CSU-EPSA, 2021, 33(10): 18-23.

[12] 陈光宇, 徐嘉杰, 卢兆军, 等. 基于相关性度量算法的台区线损异常判断及精准定位[J]. 电力工程技术, 2022, 41(4): 67-74.
CHEN Guangyu, XU Jiajie, LU Zhaojun, et al. Judgment and precise location of abnormal line loss in station area based on correlation measurement algorithm[J]. Electric Power Engineering Technology, 2022, 41(4): 67-74.

[13] CUI W Q, WANG H. Anomaly detection and visualization of school electricity consumption data[C]//2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA). Beijing, China. IEEE, 2017: 606-611.

[14] 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.

[15] MCINNES L, HEALY J, ASTELS S. HDBSCAN: hierarchical density based clustering[J]. The Journal of Open Source Software, 2017, 2(11): 205.

[16] 盛万兴, 段青, 王良, 等. 基于多代理协调机制的能量路由器群组与配电网综合规划[J]. 高电压技术, 2021, 47(1): 1-13.
SHENG Wanxing, DUAN Qing, WANG Liang, et al. Comprehensive planning for energy routers and distribution network based on multi-agent system coordination mechanism[J]. High Voltage Engineering, 2021, 47(1): 1-13.

[17] 王建元, 张少锋. 基于线性判别分析和密度峰值聚类的异常用电模式检测[J]. 电力系统自动化, 2022, 46(5): 87-98.
WANG Jianyuan, ZHANG Shaofeng. Anomaly detection for power consumption patterns based on linear discriminant analysis and density peak clustering[J]. Automation of Electric Power Systems, 2022, 46(5): 87-98.

[18] 唐俊熙, 曹华珍, 高崇, 等. 一种基于时间序列数据挖掘的用户负荷曲线分析方法[J]. 电力系统保护与控制, 2021, 49(5): 140-148.
TANG Junxi, CAO Huazhen, GAO Chong, et al. A new user load curve analysis method based on time series data mining [J]. Power System Protection and Control, 2021, 49(5): 140-148.

[19] 赵凌云, 刘友波, 沈晓东, 等. 基于 CEEMDAN 和改进时间卷积网络的短期风电功率预测模型[J]. 电力系统保护与控制, 2022, 50(1): 42-50.
ZHAO Lingyun, LIU Youbo, SHEN Xiaodong, et al. Short-term wind power prediction model based on CEEMDAN and an improved time convolutional network[J]. Power System Protection and Control, 2022, 50(1): 42-50.

[20] 赵源上, 林伟芳. 基于皮尔逊相关系数融合密度峰值和熵权法典型场景研究[J]. 中国电力, 2023, 56(5): 193-202.
ZHAO Yuanshang, LIN Weifang. Research on typical scenarios based on fusion density peak value and entropy weight method of Pearson's correlation coefficient[J]. Electric Power, 2023, 56(5): 193-202.

[21] 汪敏, 张孟健, 禹洪波, 等. 基于 EEMD 和特征降维的非侵入式负荷分解方法研究[J]. 电测与仪表, 2024, 61(6): 80-86.
WANG Min, ZHANG Mengjian, YU Hongbo, et al. Research on non-intrusive load decomposition method based on EEMD and feature dimensionality reduction[J]. Electrical Measurement & Instrumentation, 2024, 61(6): 80-86.

[22] 杨启帆, 段大卫, 李楠, 等. 基于主成分分析的串联电池组故障诊断实用方法[J]. 电力自动化设备, 2022, 42(12): 210-216.
YANG Qifan, DUAN Dawei, LI Nan, et al. A practical fault diagnosis method for series-connected battery packs based on principle component analysis[J]. Electric Power Automation Equipment, 2022, 42(12): 210-216.

[23] 邹港, 赵斌, 罗强, 等. 基于 PCA-VMD-MVO-SVM 的短期光伏输出功率预测方法[J]. 电力科学与技术学报, 2024, 39(5): 163-171.
ZOU Gang, ZHAO Bin, LUO Qiang, et al. Prediction method of short-term PV output power based on PCA-VMD-MVO-

SVM[J]. Journal of Electric Power Science and Technology, 2024, 39(5): 163-171.

[24] 施雨松，徐青山，郑建. 基于特征选择与增量学习的非侵入式电动自行车充电辨识方法[J]. 电力系统自动化，2021，45(7)：87-94.
SHI Yusong, XU Qingshan, ZHENG Jian. Non-intrusive charging identification method for electric bicycles based on feature selection and incremental learning[J]. Automation of Electric Power Systems, 2021, 45(7): 87-94.

[25] 闫梦秋，杨轶俊，赵舫. 基于改进 OCSVM 的智能变电站数据流异常检测方法研究[J]. 电力系统保护与控制，2022，50(6)：100-106.
YAN Mengqiu, YANG Yijun, ZHAO Fang. A data stream anomaly detection method based on an improved OCSVM smart substation[J]. Power System Protection and Control, 2022, 50(6): 100-106.

[26] 白雅玲，周亚同，刘君. 基于深度卷积嵌入聚类的日负荷曲线聚类分析[J]. 电网技术，2022，46(6)：2104-2113.
BAI Yaling, ZHOU Yatong, LIU Jun. Clustering analysis of daily load curve based on deep convolution embedding clustering[J]. Power System Technology, 2022, 46(6): 2104-2113.

[27] 何和智，高琦，张涛. 国内外大型注塑机技术发展动态综述[J]. 中国塑料，2022，36(11)：140-149.
HE Hezhi, GAO Qi, ZHANG Tao. A review of development trend in large injection molding machine technology at home and abroad[J]. China Plastics, 2022, 36(11): 140-149.

[28] 张广伦，钟海旺. 信息熵在电力系统中的应用综述及展望[J]. 中国电机工程学报，2023，43(16)：6155-6180.
ZHANG Guanglun, ZHONG Haiwang. Review and prospect of information entropy and its applications in power systems[J]. Proceedings of the CSEE, 2023, 43(16): 6155-6180.

[29] GHAFOORI Z, ERFANI S M, RAJASEGARAR S, et al. Efficient unsupervised parameter estimation for one-class support vector machines[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(10): 5057-5070.

[30] 陈启鑫，郑可迪，康重庆，等. 异常用电的检测方法：评述与展望[J]. 电力系统自动化，2018，42(17)：189-199.
CHEN Qixin, ZHENG Kedi, KANG Chongqing, et al. Detection methods of abnormal electricity consumption behaviors: review and prospect[J]. Automation of Electric Power Systems, 2018, 42(17): 189-199.

作者简介：


陈光宇

陈光宇(1980)，男，博士，教授，研究方向为电力系统运行与控制(E-mail：cgyhhu@163.com)；

杨光(1998)，男，硕士，研究方向为电力系统运行与控制；

施蔚锦(1974)，男，硕士，高级工程师，从事电力系统调度自动化工作。

OCSVM-based method for identifying abnormal load characteristics in industry

CHEN Guangyu[1], YANG Guang[1], SHI Weijin[2], CAI Xincan[2], CHEN Wanqing[2], LIU Hao[1]

(1. School of Electric Power Engineering, Nanjing Institute of Technology, Nanjing 211167, China; 2. Quanzhou Power Supply Company of State Grid Fujian Electric Power Co., Ltd., Quanzhou 362000, China)

Abstract: To address the challenge faced by power grid companies in accurately detecting changes in user industry information, which has been complicated by the increasing variability of industry characteristics in recent years, a data-driven approach for identifying anomalies in load characteristics is proposed. Initially, a two-stage methodology for developing typical load patterns for various industries is presented. The hierarchical density-based spatial clustering of applications with noise (HDBSCAN) technique is utilized to extract typical daily load curves for users under different scenarios. Subsequently, these extracted daily load curves are clustered using an improved *K*-means algorithm to establish typical load patterns for the respective industries. In the second phase, a multidimensional intelligent diagnostic method for load characteristic anomalies is introduced. User load characteristics are constructed, and the entropy weight method is employed to evaluate the relative significance of typical industry scenarios. The one-class support vector machine (OCSVM) algorithm is then utilized to quantify the degree of anomaly present in user load characteristics across each scenario. Comprehensive suspicion scores are calculated and ranked to accurately identify users exhibiting abnormal load characteristics. The effectiveness of the proposed method is validated through the analysis of actual user data from a specific region. The results demonstrate that the method is both feasible and practical for constructing typical industry load scenarios and for the identification of load characteristic anomalies.

Keywords: data-driven; load characteristic anomalies; hierarchical density-based spatial clustering of applications with noise (HDBSCAN)-improved *K*-means algorithm; multi-dimensional scenario analysis; one-class support vector machine (OCSVM); comprehensive suspicion score

（编辑　陆海霞）