

引文格式:郭贺宏,任宇路,姚俊峰,等. 基于特征指标降维与改进密度峰值聚类算法的异常用电行为辨识[J]. 太原理工大学学报, 2026, 57(2): 309-320.

GUO Hehong, REN Yulu, YAO Junfeng, et al. Identification of abnormal power consumption behavior based on dimensionality reduction of characteristic index and improved density peak clustering algorithm [J]. Journal of Taiyuan University of Technology, 2026, 57(2): 309-320.

基于特征指标降维与改进密度峰值聚类算法的 异常用电行为辨识

郭贺宏¹, 任宇路², 姚俊峰², 肖春^{2,3}

(1. 国网山西省电力有限公司, 山西 太原; 2. 国网山西省电力有限公司营销服务中心, 山西 太原;
3. 太原理工大学 机械工程学院, 山西 太原)

摘要:【目的】为准确辨识用户的异常用电行为以降低电网的非技术性损失, 提出一种基于线性判别分析(LDA)和改进密度峰值聚类(IDPC)算法的异常用电行为辨识模型。【方法】首先从用电数据中构造能反映用户用电行为的特征集; 其次使用LDA对提取的特征集进行降维处理; 然后使用IDPC算法对降维后的特征集进行聚类分析, 将具有不同用电行为特征的用户聚类后再进行异常辨识; 最后定义离群异常指数来描述降维后的特征集中用户的离群程度, 最终模型输出所有用户用电行为的异常程度排序。针对密度峰值聚类(DPC)算法中截断距离需人为设定的不足, 定义邦费罗尼指数, 并将改进的鲸鱼优化算法(IWOA)应用于DPC截断距离参数的优化。【结果】实验结果表明, 文章所提出的模型只需检测异常程度较大的少数用户即可稽查出大部分异常用户, 且相比其他模型具有更高的召回率与精确率。

关键词:数据挖掘; 异常用电行为; 特征降维; 密度峰值聚类; 鲸鱼优化算法

中图分类号: TM73 **文献标识码:** A

DOI: 10.16355/j.tyut.1007-9432.20240037

文章编号: 1007-9432(2026)02-0309-12

Identification of Abnormal Power Consumption Behavior Based on Dimensionality Reduction of Characteristic Index and Improved Density Peak Clustering Algorithm

GUO Hehong¹, REN Yulu², YAO Junfeng², XIAO Chun^{2,3}

(1. State Grid Shanxi Electric Power Company, Taiyuan, Shanxi, China; 2. Marketing Service Center of State Grid Shanxi Electric Power Co. Ltd., Taiyuan, Shanxi, China; 3. College of Mechanical Engineering, Taiyuan University of Technology, Taiyuan, Shanxi, China)

Abstract: 【Purposes】In order to accurately identify the abnormal power consumption behavior of power users and reduce the non-technical loss of power grid, an abnormal power consumption behavior identification model based on linear discriminant analysis (LDA) and improved density peak clustering (IDPC) algorithm is proposed. 【Methods】First, a feature set was constructed from the power consumption data to reflect the power consumption behavior of users. Second, LDA was used to re-

收稿日期: 2024-01-11; 修回日期: 2024-06-07

基金项目: 国网山西省电力公司科技项目(52051L230003)

第一作者: 郭贺宏, 硕士, 正高级工程师, 主要从事电力系统运行与管理研究, (E-mail) guohh1199@163.com

通信作者: 肖春, 博士研究生, 高级工程师, 主要从事电力营销技术、电力系统运行与控制研究等, (E-mail) tyutxiaochun@163.com

duce the dimension of the extracted feature set. Then, IDPC algorithm was used to cluster the feature set after dimensionality reduction, and the users with different electricity behavior characteristics were clustered before abnormality identification. Finally, the outlier index was defined to describe the degree of user outlier in the feature set after dimensionality reduction, and the final model outputs the order of the degree of abnormality of all users' electricity behavior. In this paper, in view of the deficiency that the truncated distance in the density peak clustering (DPC) algorithm needs to be set manually, the Bonferroni index was defined, and the improved whale optimization algorithm was applied to the optimization process of DPC truncated distance parameters. **【Results】**Experimental results show that the proposed model can detect most abnormal users only by detecting a small number of users with large degree of abnormality, and has higher recall rate and accuracy than those of other models.

Keywords: data mining; abnormal power consumption behavior; feature dimensionality reduction; density peak clustering; whale optimization algorithm

电能的非技术性损失(non-technical loss, NTL)是指在电力系统中由于非技术因素造成的未能计入正常收入的损失^[1],这些因素与电力供应系统的技术运行和电力传输无关,而是源自于非法行为、计量不准确、欺诈或错误的运营管理等问题。其中窃电行为与计量装置故障是造成NTL的两大主要原因^[2]。异常用电行为辨识是窃电检测、计量装置运行状态评价的重要内容,准确的辨识结果对于NTL的防控尤为重要,直接影响电力公司的经济利益和电力系统的供电质量。传统的异常用电行为辨识工作主要依赖于人工现场筛查,耗费人力且效率低下。近年来,人工智能的迅猛发展与智能电表的大规模普及应用为基于数据驱动的异常用电行为辨识提供了技术支持与数据来源^[3]。

有监督学习和无监督学习是基于数据驱动的两中学习方法。其中有监督学习需要有明确的数据标签(正常、异常)来指导模型学习,并根据这些标签进行训练和优化,基于分类与回归的异常检测方法属于典型的有监督学习。基于分类的异常检测方法主要通过大量带标签的数据集来训练分类器的分类精度;基于回归的异常检测方法研究实际用电量与模型预测用电量之间的差值来进行异常判别。无监督学习则是在未标注数据上寻找数据的内在结构和模式,由于训练数据缺乏标签信息,模型必须在没有指导的情况下自行学习数据的结构,典型的无监督学习包括聚类^[4]、局部离群因子^[5](local outlier factor, LOF)、孤立森林^[6]等。分类模型先使用有监督学习算法对已标记的训练数据进行训练,然后利用训练好的分类器来对新的未标记数据进行预测,将其归为正常类别或异常类别。常用的异常用电检测分类模型有极限学习机^[7]、支持

向量机^[8]、K近邻^[9]和神经网络^[10]等。分类模型非常依赖已知的异常标签,然而在实际场景中,很多异常是未知的且没有在训练数据中出现过,这使得基于分类的模型无法直接适用于未知异常检测。回归模型一般用于负荷预测、电价预测等。在异常检测领域,回归模型的目标是建立一个回归函数,通过对输入的用电量进行拟合来预测未来的用电量,通过比较实际值和预测值之间的差异来确定是否存在异常^[11]。然而,基于回归的方法需要对每个电能表分别建立预测模型并调整超参数,会极大地占用计算资源,模型建立也十分耗时。

在实际的异常检测问题中,异常样本通常较正常样本比例小,导致了数据类别的不平衡性^[12]。这样的不平衡分布会使得分类模型在训练过程中倾向于优化对多数类别的分类,使得少数类别的辨识准确率较低。而回归模型面对海量智能电表时,为每个电表建立预测模型将变得非常耗时,还会面临巨大的计算和存储资源挑战^[13]。因此,本文重点研究不需要训练过程的无监督学习模型。基于聚类的异常辨识方法属于最典型的无监督学习,聚类模型首先对数据进行聚类操作,将数据样本划分为不同的群集,然后检测不属于任何群集或属于小群集的数据点作为异常值。K-means作为一种简单高效的聚类算法,其用于异常检测的效果很大程度上取决于聚类中心的选取。文献[14]通过均匀选择数据紧密区域并避免离群点区域,对K-means聚类算法的初始聚类中心选择过程进行了优化。近几年,基于密度聚类的方法逐渐成为异常检测问题的研究热点,文献[15]采用基于密度的带噪声空间聚类方法(density-based spatial clustering of applications with noise, DBSCAN)对三次指数平滑模型的

预测电量与真实电量之间的残差进行聚类,从而检测出有异常的用电量值,实验结果表明其检测率和误报率等评价指标均优于LOF算法和支持向量机算法。然而DBSCAN方法的检测结果对参数取值十分敏感,而且需要确定邻域半径Eps和最小邻居数目minPts双参数,现阶段DBSCAN的参数主要根据人工经验进行选取,缺乏相应的参数选取理论。文献[16]采用LOF算法进行电力用户异常用电行为的挖掘,并采用网格技术提高算法效率,然而LOF算法并不适合检测包含多种用电行为的电力负荷。综上所述,考虑到用电行为的多样性和算法参数选取的便捷性,文章采用密度峰值聚类(density peak clustering, DPC)作为异常辨识模型的核心算法,基于改进鲸鱼优化算法对其参数进行寻优,利用离群异常指数判别数据中的离群点。文章按照特征集构造、特征降维、聚类分析、离群点检测的流程进行异常用电行为的辨识。

1 异常用电行为辨识流程

文章采用基于DPC的无监督学习方法进行异常用电行为辨识,其本质可以看作是高维特征经过降维处理后在低维空间的离群点检测问题^[17]。具体的辨识流程如图1所示。

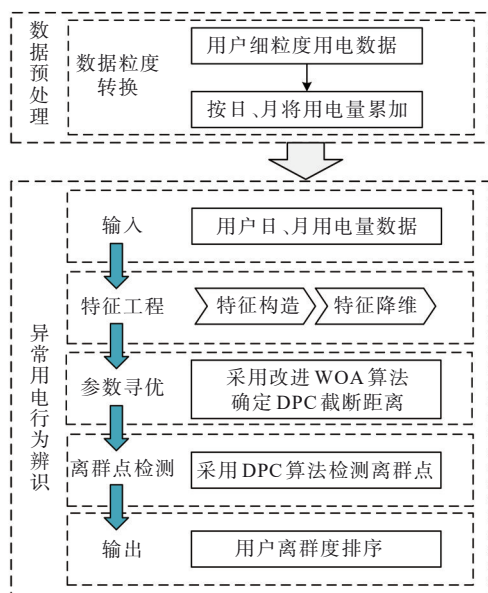


图1 异常用电行为辨识流程图

Fig.1 Flow chart of abnormal power consumption behavior identification

1) 数据预处理

首先将智能电表采集到的细粒度数据分别按日、按月进行累加,为接下来电力负荷数据的特征提取做好准备工作。

2) 特征工程

根据日、月用电量数据,从统计性指标、波动性指标、趋势性指标和相似性度量指标4个角度出发,构造能反应用户用电行为的特征指标集。为了在低维空间直观地展示用户的用电行为,采用特征降维方法对特征集进行降维处理,并通过评价指标对比了三种特征降维效果的优劣。

3) 参数寻优

根据DPC算法中的局部密度和相对距离定义邦费罗尼指数函数(bonferroni index, BI),并利用改进后的鲸鱼优化算法(whale optimization algorithm, WOA)求解其最大值,从而得到全局最优的截断距离参数。

4) 离群点检测

将降维后的特征作为DPC算法的输入,并引入平均向心距离和向心相对距离来计算数据对象的离群异常指数,输出用户离群度排序,并根据设定的阈值输出异常用户编号。

2 改进鲸鱼优化算法

在用密度峰值聚类算法进行离群点检测之前,需要利用优化算法对其截断距离参数进行寻优,从而避免人工确定参数的主观性。

2.1 鲸鱼优化算法原理

鲸鱼优化算法包含3个阶段:包围猎物,螺旋更新和随机搜索。首先,初始化鲸鱼种群大小 N 、最大迭代次数 $T_{\max}$ 、以及种群中每个鲸鱼个体的位置信息 X_i ,定义决策系数 p 为介于 $[0,1]$ 的随机数。系数向量 A 、 C 计算方式如下:

$$\begin{cases} A = 2a \cdot r_1 - a \cdot \frac{r_1}{|r_1|} \\ C = 2r_2 \end{cases} \quad (1)$$

其中, r_1 、 r_2 为介于 $[0,1]$ 的随机向量; a 为收缩包围机制中的线性收敛因子,随着迭代次数的增加从2线性减小到0,可表示为:

$$a = 2 \times (1 - t/T_{\max}) \quad (2)$$

其中: t 表示当前迭代次数; $T_{\max}$ 表示最大迭代次数。

1) 包围猎物阶段。模拟座头鲸发现猎物并收缩包围的过程,此阶段 $p < 0.5$ 且 $|A| < 1$ 。数学模型表示如下:

$$D_1 = |C \cdot X^*(t) - X(t)| \quad (3)$$

$$X(t+1) = X^*(t) - AD_1 \quad (4)$$

式中: D_1 为鲸鱼个体与猎物间的距离; $X(t)$ 为鲸鱼

个体的位置向量; $X^*(t)$ 为当前全局最优鲸鱼的位置向量。

2)螺旋更新阶段。通过收缩包围机制和螺旋更新位置机制模拟座头鲸的气泡网攻击行为,此阶段 $p \geq 0.5$ 。数学模型表示如下:

$$D_2 = |X^*(t) - X(t)|. \quad (5)$$

$$X(t+1) = D_2 \cdot e^{bl} \cos(2\pi l) + X^*(t). \quad (6)$$

式中: D_2 表示鲸鱼个体到猎物的距离; b 是一个用于定义对数螺线形状的常数; l 为介于 $[-1, 1]$ 的随机数。

3)随机搜索阶段。鲸鱼种群增加搜索范围,在全局空间探索最优解。在此阶段, $p < 0.5$ 且 $|A| \geq 1$ 。数学模型表示如下:

$$D_{\text{rand}} = |C \cdot X_{\text{rand}} - X(t)|. \quad (7)$$

$$X(t+1) = X_{\text{rand}} - A \cdot D_{\text{rand}}. \quad (8)$$

式中: D_{rand} 表示随机选取的鲸鱼到猎物的距离; X_{rand} 表示从种群中随机选择的个体位置向量。

2.2 改进鲸鱼优化算法

为提升 WOA 对复杂函数求解寻优的能力,文章提出一种改进的鲸鱼优化算法(improved WOA, IWOA)。

1)非线性变化收敛因子

公式(2)中的线性收敛因子影响着算法的全局搜索能力与局部开发能力。为此引入一种非线性收敛因子:

$$a^* = 2 - 2\ln\left[1 + \frac{(e-1)t}{T_{\max}}\right]. \quad (9)$$

2)自适应概率阈值

为了模拟鲸鱼的气泡网攻击模式,鲸鱼优化算法设置搜索包围机制与螺旋更新位置机制等概率同时进行。即根据决策系数 p 的取值,算法分别有 50% 的概率选择式(4)中的模型和式(6)中的模型。然而这种等概率的选择方式会导致算法收敛速度慢且容易陷入局部最优解。为此引入自适应概率阈值代替等概率模型,数学模型表示如下:

$$p^* = 1 - \left[\frac{1}{\alpha + \beta} \cdot \left(\alpha \left(\frac{t}{T_{\max}} \right)^{\alpha} + \beta \cdot \left(\frac{t}{T_{\max}} \right)^{\beta} \right) \right]. \quad (10)$$

其中: α 和 β 为控制参数, $\alpha = 3, \beta = 2$ 。

3)自适应位置权重

WOA 算法没有考虑到随着迭代的进行,猎物对鲸鱼位置的更新的影响力存在差异。为此,引入一种自适应位置权重参数,对 WOA 中鲸鱼位置进行更新并赋予不同的权重,数学模型表示如下:

$$\omega = \frac{1}{\alpha + \beta} \cdot \left(\alpha \left(\frac{t}{T_{\max}} \right)^{\alpha} + \beta \cdot \left(\frac{t}{T_{\max}} \right)^{\beta} \right). \quad (11)$$

其中: $\alpha = 3, \beta = 2, \omega \in [0, 1]$ 。

综上,改进后的 WOA 数学模型表示如下:

$$\begin{cases} X(t+1) = \omega X^*(t) - AD_1 & |A| < 1, p < p^*, \\ X(t+1) = \omega X_{\text{rand}}(t) - AD_{\text{rand}} & |A| \geq 1, p < p^*, \\ X(t+1) = D_2 \cdot e^{bl} \cos(2\pi l) + (1 - \omega)X^*(t) & p \geq p^*. \end{cases} \quad (12)$$

3 异常辨识模型

异常辨识模型包括负荷特征集的构造与维度规约、密度峰值聚类算法及其参数寻优、离群异常指数的定义、模型评价指标等部分。

3.1 特征集构造

负荷特征构造的目的是将复杂的用电行为转化为更易于分析和理解的形式,从而为后续的分析 and 检测提供基础。文章基于用户的日、月用电量数据构造了统计性指标、波动性指标、趋势性指标和相似性度量指标。

3.1.1 统计性指标

基于日用电量数据构造其均值、标准差、偏度、峰度 4 种特征指标。

3.1.2 波动性指标

1)每季度日用电量变异系数^[18],即每季度的日用电量标准差与日用电量均值之比。

2)月用电量率,即月平均用电量与月最大用电量之比。

3)月用电量峰谷差率^[19],即最大、最小月用电量之差与最大月用电量之比。

3.1.3 趋势性指标

1)月用电量数据的线性拟合斜率 k 。

2)由日用电量数据的零滞后异同移动平均线(zero lag moving average convergence/divergence, ZLMACD)^[20]计算得到的趋势指标。

ZLMACD 指标克服了 MACD 具有滞后性的缺点,能够更及时地捕捉用电量数据的趋势强度和方向,其计算方式如下:

$$\text{EMA}_{(S,n)}^i = \frac{2}{n+1} \cdot S^i + \left(1 - \frac{2}{n+1} \right) \cdot \text{EMA}_{(S,n)}^{i-1}. \quad (13)$$

$$\text{ZLMA}_{(S,n)} = 2\text{EMA}_{(S,n)} - \text{EMA}_{(\text{EMA}_{(S,n)},n)}. \quad (14)$$

$$\text{ZLDIF}_{(S,(p,q))} = \text{ZLMA}_{(S,p)} - \text{ZLMA}_{(S,q)}. \quad (15)$$

$$\text{ZLMACD}_{(S,(p,q),a)} = \text{EMA}_{(\text{ZLDIF}_{(S,(p,q),a)},a)}. \quad (16)$$

$$\begin{cases} T_{\text{up}} = \frac{1}{C_1} \sum_{i=1}^m \max(\text{ZLMACD}_{(S, (p, q, a))}^i, 0), \\ T_{\text{down}} = \frac{1}{C_2} \sum_{i=1}^m \left| \min(\text{ZLMACD}_{(S, (p, q, a))}^i, 0) \right|. \end{cases} \quad (17)$$

其中: $\text{EMA}_{(S, n)} = [\text{EMA}_{(S, n)}^i]$ 表示时间序列 S 的 n 日指数移动平均值序列; $\text{ZLMA}_{(S, n)}$ 表示时间序列 S 的 n 日零滞后移动平均序列, 设 $p < q$, 则 $\text{ZLMA}_{(S, p)}$ 和 $\text{ZLMA}_{(S, q)}$ 分别表示快慢速零滞后移动平均序列; $\text{ZLDIF}_{(S, (p, q))}$ 为差离值序列; $\text{ZLMACD}_{(S, (p, q, a))}$ 为零滞后异同移动平均线指标; C_1 和 C_2 分别表示 ZLMACD 序列中大于零和小于零的数据点数; m 表示用电量序列长度; T_{up} 和 T_{down} 分别表示用电量序列的上升趋势指标和下降趋势指标。

3.1.4 相似性度量指标

用户月用电量数据与典型月用电量数据(所有用户月用电量的均值)的皮尔逊相关系数和欧氏距离。

3.2 特征集维度规约

由于构造的特征数量较多且某些特征之间包含重叠信息, 具有很强的相关性。为了在低维空间直观地展示用户的用电行为并且提升算法效率, 有必要对特征集进行降维处理。常用的降维方法包括主成分分析(principal component analysis, PCA)、因子分析(factor analysis, FA)和线性判别分析(linear discriminant analysis, LDA)。就模型的监督性而言, 其中 PCA 和 FA 属于无监督学习模型, 即不需要数据标签来学习。PCA 主要关注数据的方差分布, 使得投影后的数据具有最大的方差, 由于不需要类别标签, 在某些情况下可能无法很好地区分不同类别。FA 专注于揭示观测数据背后的潜在结构, 它假设数据是由潜在因素和随机误差组成的, 其数学模型可以表示为 $X = AF + \epsilon$, 其中 X 为原始数据矩阵, A 为因子载荷矩阵, F 为公因子, AF 为潜在因子, ϵ 为残差矩阵。FA 的目标是估计 A 、 F 以及 ϵ , 使得 X 能够被最好地重建。与 PCA 和 FA 不同, LDA 是一种有监督学习模型, 即需要明确数据的类别信息标签, 需要明确指出的是, 这里的标签是数据固有的类别标签, 而非异常辨识的异常用户标签。LDA 可以最大化不同类别之间的间距, 同时最小化同一类别内部的差异。就模型可保留新特征的最大维度而言, PCA 和 FA 降维后的新特征维度可以达到原数据的维度, 而 LDA 降维后新特征维度最多为类别标签数减 1。

特征值的贡献率(explained variance ratio,

EVR)和累积贡献率(cumulative explained variance ratio, CEVR)是用来评估降维后的新特征对原始数据方差解释程度的指标。EVR 衡量单个新特征的重要性, 而 CEVR 则衡量多个新特征的综合重要性。一般 CEVR 达到 85% 即可较为全面地表达原始信息。文中使用这两个参数来确定模型保留新特征的维度。

3.3 密度峰值聚类算法原理

DPC 算法是由 Rodriguez 等人于 2014 年提出的一种基于密度的聚类算法^[21], 能够识别任意形状的簇且仅需确定唯一参数。DPC 的核心思想基于两点假设: ①在数据集中, 聚类中心被局部密度较低的数据点所包围; ②聚类中心与其他更高局部密度数据点的距离相对较远。基于这两点假设, 根据以下几点定义来建立对 DPC 算法及其离群点检测功能的详细描述。

定义 1 局部密度 ρ_i 。 ρ_i 表示数据点 i 在截断距离内包含的其他数据点的个数。其计算方式如下:

$$\rho_i = \sum_{j=1}^n \chi(d_{ij} - d_c). \quad (18)$$

其中, n 为数据点总数; $\chi(x)$ 为指示函数, $x < 0$ 时, $\chi(x) = 1$, $x \geq 0$ 时, $\chi(x) = 0$; d_{ij} 表示数据点 i 和 j 之间的欧氏距离; d_c 为截断距离, 是 DPC 算法的唯一参数。

定义 2 相对距离 δ_i 。 δ_i 表示数据点 i 与局部密度比它大的数据点之间欧式距离的最小值。其计算方法如下:

$$\delta_i = \min_{j: \rho_j > \rho_i} d_{ij}. \quad (19)$$

对于整个数据集中 ρ_i 最大的点, 将其 δ_i 设置为与其最远的数据点之间的欧式距离。

定义 3 密度峰值点决策指标 γ_i 。 γ_i 表示数据点 i 的局部密度与相对距离的乘积。其计算方法如下:

$$\gamma_i = \rho_i \cdot \delta_i. \quad (20)$$

计算出每个数据点的局部密度和相对距离后, 可以画出其 δ - ρ 决策图, 密度峰值点具有更大的 ρ 和 δ , 处于决策图的右上角, 可以通过决策指标 γ 的大小来确定密度峰值点(即聚类中心)。

根据上述定义, 可以得到经典的 DPC 算法流程如下:

1) 计算数据集中每个样本点和其余样本点的欧式距离矩阵 D ;

2) 根据 D 中的数据人工设置截断距离 d_c 后, 由式(18)、(19)求每个样本点的局部密度 ρ_i 和相对

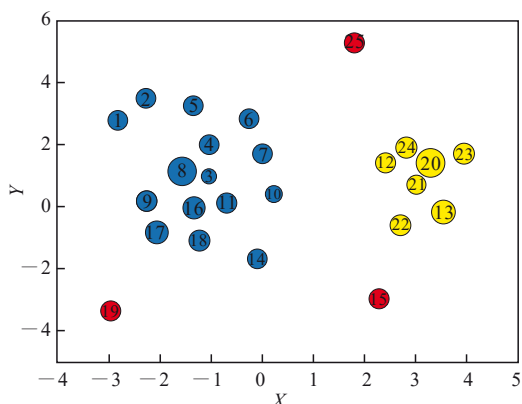
距离 δ_i ;

3) 画出数据集的 δ - ρ 决策图, 并根据决策指标 γ_i 的大小选择聚类中心;

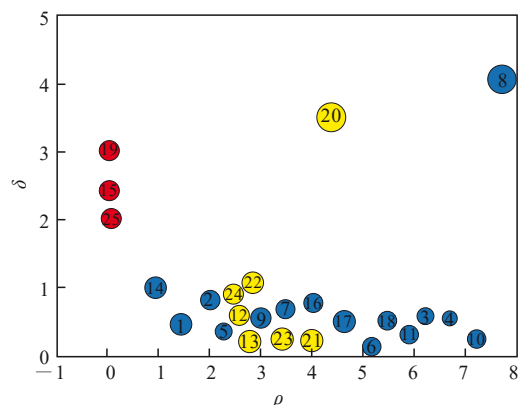
4) 其余样本点被划分给最靠近的聚类中心簇;

5) ρ 较小且 δ 较大的样本点可视为离群点。

图 2 为 DPC 算法原理图, 包含一组二维数据分布图以及对应的 δ - ρ 决策图。黄蓝数据点分别为两个聚类簇, 其中 8 号数据点和 20 号数据点分别为蓝色簇和黄色簇的聚类中心, 即密度峰值点, 其 ρ 和 δ 都很大, 分布在决策图的右上角; 红色点为离群点, 其 ρ 较小 δ 较大, 分布在决策图的左上方; 其余点为正常数据点, 分布在决策图的下方, δ 较小。



(a) 二维数据分布图



(b) 决策图

图 2 DPC 算法原理图

Fig.2 Schematic diagram of DPC algorithm

3.4 改进密度峰值聚类算法

针对 DPC 算法中截断距离 d_c 需要人工设置的不足, 引入邦费罗尼指数^[22]来度量决策指标集合 $\{\gamma_1, \gamma_2, \dots, \gamma_N\}$ 的分布情况, 其中 $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_N$, 数学表达式如下:

$$BI = \frac{1}{N-1} \sum_{i=1}^N \frac{\sum_{j=1}^i \gamma_j / \sum_{j=1}^N \gamma_j - i/N}{i/N}. \quad (21)$$

BI 的大小主要受前几项密度峰值点的影响。

将集合 $\{\gamma_1, \gamma_2, \dots, \gamma_N\}$ 中前 m 项对应的样本数据设置为密度峰值点, $i \in [1, m]$ 表示前 i 个聚类中心的累积决策指标相对于总体累积决策指标的贡献程度, 其值越大, 表示聚类中心包含着越多的信息量, 聚类效果越好, 此时 BI 的值也会越大。由式 (18)~(21) 可知, BI 为 d_c 的函数, 求取 BI 取值最大时的 d_c 即为最佳截断距离 $d_{c, \text{best}}$ 。其求解过程属于非确定多项式问题, 文中采用改进鲸鱼优化算法进行迭代求解。

在离群点检测方面, 可以初步将处于决策图左上方的数据点视为离群点。文献[23]根据 δ - ρ 决策图提出一种 DPC 算法的异常指数来度量数据点的离群程度, 其计算方式如下:

$$\zeta_i = \frac{\delta_i}{\rho_i + 1}. \quad (22)$$

其中异常指数 ζ_i 的设定综合考虑了密度和距离两种因素。 ζ_i 越大, 认为样本点离群程度越高, 对应的电力用户发生异常用电行为的概率越大。然而, 用户用电行为的多样性使得样本数据点经过聚类后被划分为几个聚类中心密度大小各异的簇, 若在整个数据集上使用式 (22) 中的 ζ_i 来进行离群点检测, 则高密度区域的局部离群点易被低密度区域的正常数据点淹没, 从而影响离群点的检出率。通过改进 ζ_i 中的距离因素, 使之更好地适用于检测包含多种不同密度簇的数据集。

定义 4 平均向心距离 d_{avg}^j 。 d_{avg}^j 表示簇 C_j 中所有数据点与聚类中心距离之和的平均值。其计算方式如下:

$$d_{\text{avg}}^j = \frac{1}{n} \sum_{x_i \in C_j} \text{dist}(x_\mu, x_i). \quad (23)$$

其中: n 表示簇 C_j 中数据点的总数; dist 表示求两点间欧式距离的函数; x_μ 表示所属簇的聚类中心; x_i 表示簇 C_j 内的数据点。

定义 5 向心相对距离 d_{rel}^j 。 d_{rel}^j 表示簇 C_j 内数据点的相对距离 δ_i 和平均向心距离 d_{avg}^j 的比值。其计算方式如下:

$$d_{\text{rel}}^j = \frac{\delta_i}{d_{\text{avg}}^j}. \quad (24)$$

d_{rel}^j 放大了离群点在所属聚类簇中的离群特征, 使得数据集中离群点的检测不受簇密度差异的影响, 比 δ_i 更好地刻画了数据点的局部离群特性。

定义 6 离群异常指数 ζ_i^* 。 ζ_i^* 表示数据点向心相对距离 d_{rel}^j 与数据点局部密度 ρ_i 的比值。其计算方式如下:

$$\zeta_i^* = \frac{d_{\text{rel}}^j}{\rho_i}. \quad (25)$$

一般情况下,对于每一个簇的局部离群点而言,其向心相对距离会远大于正常点的向心相对距离。因此离群异常指数可以很好地描述样本数据的离群程度。

3.5 模型评价指标

异常用电行为辨识本质上是一个类别不平衡的二元分类问题,不平衡数据集使得模型存在高准确率误导问题,因此需要建立更为综合的模型评价体系来评估其性能的优劣。

3.5.1 混淆矩阵

混淆矩阵是评估二元分类模型性能的基础,基于混淆矩阵可以计算出各种评价指标,如召回率、精确率等,从而更全面地理解和分析模型的分类能力。混淆矩阵中包含了二分类问题的所有可能情况,将实际异常与实际正常用户分别设置为正例(Positive)和负例(Negative),如表1所示。

表1 混淆矩阵
Table 1 Confusion matrix

用户	实际异常用户	实际正常用户
检测异常用户	TP	FP
检测正常用户	FN	TN

3.5.2 召回率和精确率

1) 召回率 Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%. \quad (26)$$

式中:TP表示模型预测和实际均为异常的用户数量;FN表示模型预测为正常,实际为异常的用户数量。召回率越大,表明模型正确检测出所有正例的能力越强。

2) 精确率 Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%. \quad (27)$$

式中:FP表示模型预测为异常,实际为正常的用户数量。精确率越高,表明模型预测为正例的样本中的误报率越低。

3.5.3 ROC曲线和AUC指标

受试者工作特性曲线(receiver operating characteristic curve, ROC)展示了模型在不同阈值下真正例率(true positive rate, TPR)与假正例率(false positive rate, FPR)之间的关系。ROC曲线的横轴表示FPR,纵轴表示TPR。在二维坐标上,每个阈值对应于ROC曲线上的一个点。模型的预测结果

与实际情况不同时,阈值不同,从而影响TPR和FPR的值。

曲线下面积(area under curve, AUC)为一个数值指标,表示ROC曲线下方面积的大小,其范围在[0,1]之间,AUC值越接近1,说明模型在不同阈值下具有更好的能力区分正例和负例,AUC=1对应理想分类器。

4 算例分析

文章使用爱尔兰智能电表数据集进行实验分析,选取了6400个用户535d的用电量数据进行实验,采样间隔为30min。用户数据分别来自5个台区(标记为1—5号)。该数据集用户数量较多,测量周期较长,用电行为多样化,适合用于异常用电行为辨识的研究。

4.1 异常用电模式构造

针对电力用户日用电量数据 $x = \{x_1, x_2, \dots, x_{48}\}$ 构造了6种异常用电模式,具体如式(28):

$$\begin{cases} f_1(x_t) = \alpha x_t & \alpha = \text{random}(0.1, 0.8) \\ f_2(x_t) = \xi_t x_t & \xi_t = \text{random}(0.1, 0.8) \\ f_3(x_t) = \gamma_t \text{mean}(x) & \gamma_t = \text{random}(0.1, 0.8) \\ f_4(x_t) = \delta_t x_t & \delta_t = \begin{cases} 0 & t \in [m, n] \\ 1 & t \in [1, m] \cup [n, 48] \end{cases} \\ \text{其中: } m \in \text{random}[1, 16], n \in \text{random}[16, 48] \\ f_5(x_t) = \max((x_t - \Delta), 0) \\ f_6(x_t) = 0 \end{cases} \quad (28)$$

式中: t 表示每日异常时刻; Δ 表示统一的削减量; $f_1(\cdot)$ 表示每日按固定比例因子减小用电量; $f_2(\cdot)$ 表示按随机的比例因子减小用电量,显然 $f_1(\cdot)$ 为 $f_2(\cdot)$ 的一种特例; $f_3(\cdot)$ 表示计算用电量平均值后按随机比例因子减小用电量; $f_4(\cdot)$ 表示智能电表在一天某段时间内因故障无法上传量测数据的行为; $f_5(\cdot)$ 表示使一天中每一时刻的用电量都降低相同的数值,最小降低到零; $f_6(\cdot)$ 表示智能电表全天不上传量测数据的行为, $f_6(\cdot)$ 为 $f_4(\cdot)$ 的一种特例。

文章根据各个台区的用户数量进行了异常用户的设定,以保证在每个台区内异常用户的比例大致相同。每个异常用户的异常天数设定为区间[200, 400]内的随机值。算例包含200个异常用户,在每个台区中都包含了之前提到的所有异常用电模式。

4.2 特征提取与维度规约

根据用户日、月用电量数据构造的反映用户用电行为的15维特征如表2所示。

表 2 特征提取指标
Table 2 Feature extraction indices

序号	指标
V1—V4	日电量的均值、标准差、偏度、峰度
V5—V8	每个季度的日用电量变异系数
V9—V10	月用电量率、月用电量峰谷差率
V11—V12	日用电量上升、下降趋势指标
V13	月用电量线性拟合斜率 <i>k</i>
V14—V15	与典型月用电量曲线的皮尔逊相关系数、欧式距离

由于提取的特征间存在较大的尺度差异,为了使得各个特征对模型辨识结果影响均衡,对提取的特征进行归一化处理,计算公式如下:

$$X_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}.$$
 (29)

式中: x_i 表示第*i*个特征归一化之前的数值, X_i 表示该特征归一化后的数值; $x_{\max}$ 、 $x_{\min}$ 分别表示该特征的最大、最小值。

特征 V5—V8 表示全年 4 季度的用电量波动情况,将其权重系数设为 0.25。

对提取的 15 维特征进行相关性分析,得到的相关性矩阵如图 3 所示。部分特征之间存在着信息重叠,具有明显的相关性。考虑到异常辨识模型的效率以及算法的可视化,需对特征集进行降维。文章分别采用 PCA、FA、LDA 三种降维模型对提取的 15 个特征集进行降维处理,并将得到的新特征按照贡献率大小进行排序。其中 LDA 模型的标签类别为台区号(标号 1—5),故其新特征最多为 4 个。

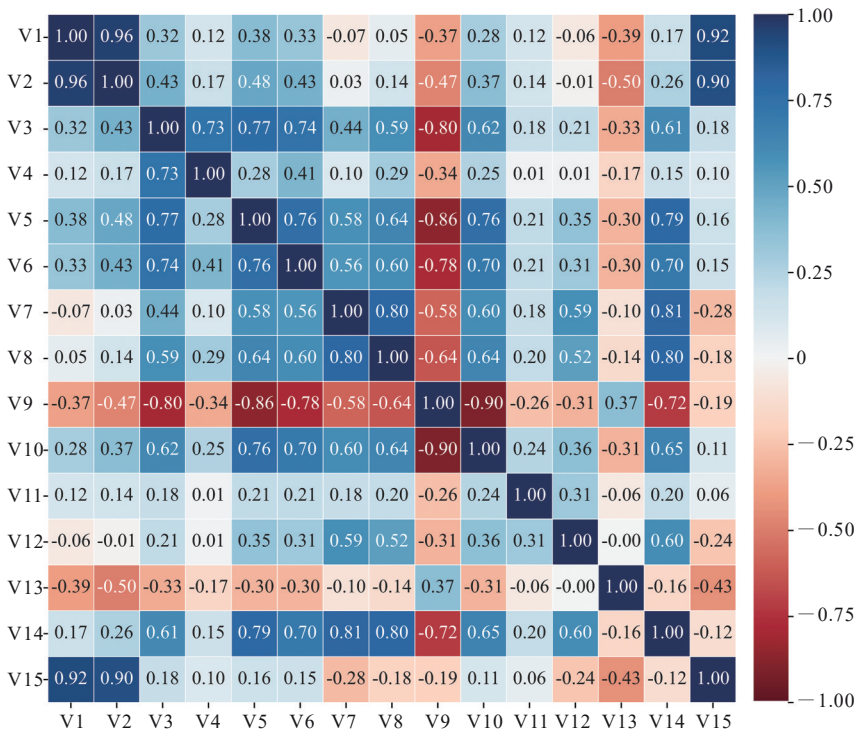


图 3 特征集的相关性矩阵

Fig.3 Correlation matrix of feature set

表 3 截取了前 8 个贡献率较高的新特征。当新特征保留到 3 维时,PCA 与 LDA 的 CEVR 分别为 91.56% 和 99.89%,均能很好地表达原特征集的信息,而 FA 的 CEVR 仅为 64.52%,通常 CEVR 需要达到 85% 才能较为全面地表达原始信息,所以 FA 只保留 3 维特征时会丢失原特征集的许多信息,不利于异常点的辨识。综合考虑特征的可视化与各个模型的 CEVR,文章保留特征贡献率最高的三维特征。

4.3 实验结果分析

4.3.1 降维模型对比

采用降维后前三个贡献率最高的新特征来表

表 3 三种降维模型所得新特征的贡献率与累积贡献率
Table 3 EVR and CEVR of new features obtained from three dimensionality reduction models %

新特征	PCA		FA		LDA	
	EVR	CEVR	EVR	CEVR	EVR	CEVR
1	71.77	71.77	25.62	25.62	94.27	94.27
2	14.19	85.96	24.23	49.85	5.27	99.54
3	5.60	91.56	14.67	64.52	0.35	99.89
4	3.91	95.47	11.81	76.33	0.11	100
5	2.07	97.54	6.12	82.45	—	—
6	1.21	98.75	2.04	84.49	—	—
7	0.72	99.47	1.86	86.35	—	—
8	0.22	99.69	0.44	86.79	—	—

示用户的用电行为,可以将所有用户的用电量曲线映射在三维空间中。根据图4中的决策图((d)、(e)、(f))可知,通过改进密度峰值聚类算法聚类后,三种降维模型的三维空间散点图均被分为三类。橙、绿、蓝三种颜色的点表示三类用户中的正常用户,红色点表示构造的异常用户。异常用电行为辨

识的核心思想就是根据三维空间中数据点的局部密度和向心相对距离来判断每个用户的异常程度。根据决策图中数据点的异常概率来绘制ROC曲线并计算AUC指标,异常概率 P_i 的计算方法如下:

$$P_i = \frac{\delta_i}{\delta_{\max}} \cdot \frac{\rho_{\max} - \rho_i}{\rho_{\max}} \quad (30)$$

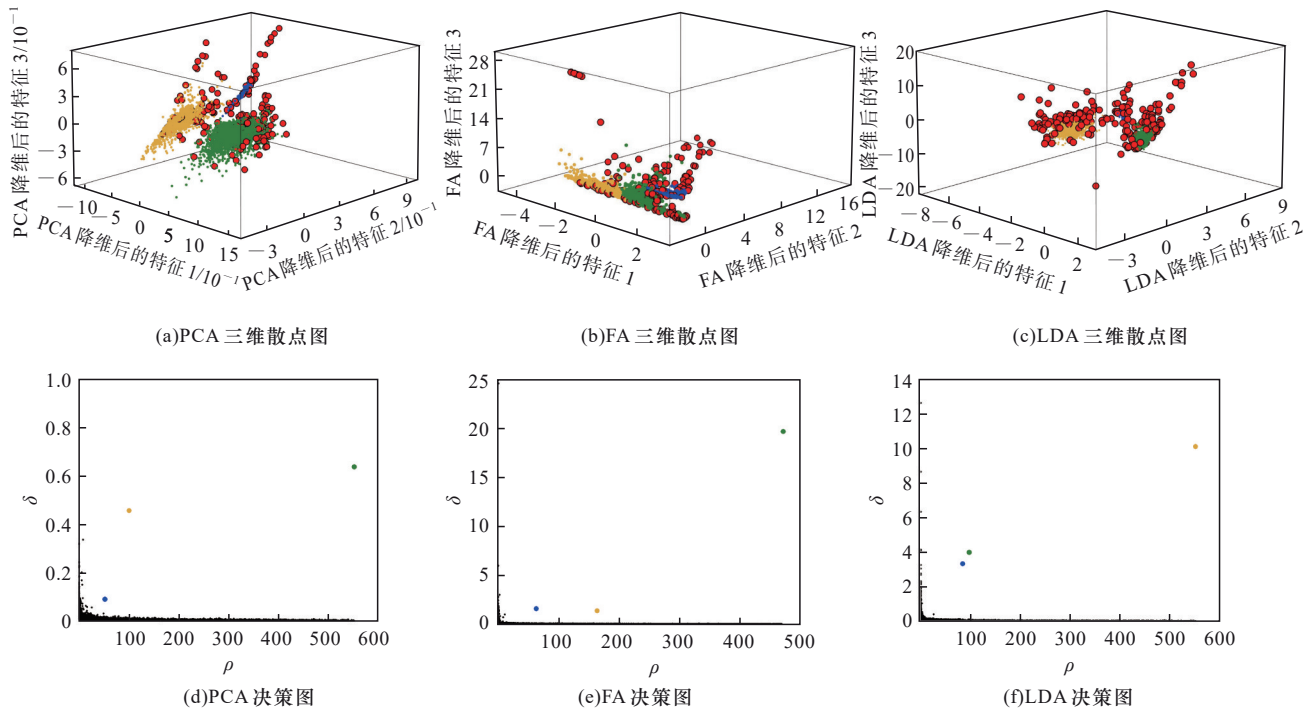


图4 所有用户映射到三维空间的散点图和相应决策图

Fig.4 Scatter diagrams and corresponding decision diagrams that are mapped to 3D space for all users

式中: $\rho_{\max}$ 表示密度最大的聚类中心点的密度; $\delta_{\max}$ 表示密度最大的聚类中心点的相对距离。异常概率 P_i 在用户间对比才有意义,其仅用于ROC曲线的绘制以及AUC指标的计算,并不作为判断用户异常程度的依据。图5表示评估三种降维模型的ROC曲线。其中LDA、PCA、FA对应的AUC指标分别为0.936 5、0.889 5、0.871 4。由此可见采用LDA模型将用电量数据映射到三维空间的效果最优。因此选用LDA模型进行异常辨识。

4.3.2 辨识结果分析

文中模型采用IWOA优化截断距离参数的DPC算法对LDA降维后的三维空间数据点进行聚类分析,并根据式(20)计算每个电力用户的离群异常指数 ζ^* 。图6为表示用户 ζ^* 分布的帕累托图,黑色曲线为异常指数的累积比率,其中绝大多数用户的 ζ^* 的值分布在 $[0, 0.1]$,占到所有用户比例的近94%。很明显 ζ^* 符合长尾分布,即异常用户分布在占比很小的长尾部,正常用户分布在占比很大的头部,故所提异常检测模型只需检测异常指数 ζ^* 较大

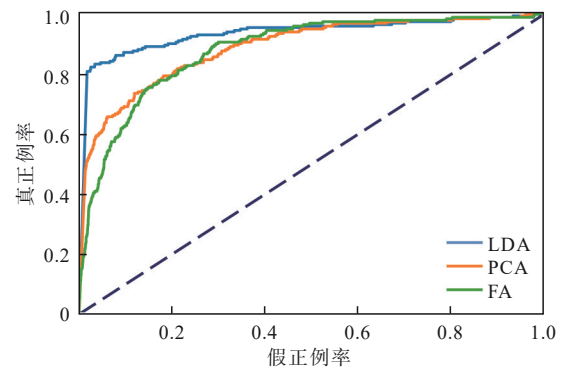


图5 三种降维模型的ROC曲线

Fig.5 ROC curves of three dimensionality reduction models

的用户即可稽查出大部分异常用户。长尾分布的拟合公式如下:

$$f(x) = cx^{-\alpha-1} \quad (31)$$

图6中红色曲线为 ζ^* 分布的拟合曲线,其中参数值为: $c = 0.39904$, $\alpha = -2.21034$ 。设定 ζ^* 的异常辨识阈值为长尾分布的97%分位数,如果 ζ_i^* 大于阈值,则判定该用户异常;如果 ζ_i^* 小于阈值,则判定该用户正常。

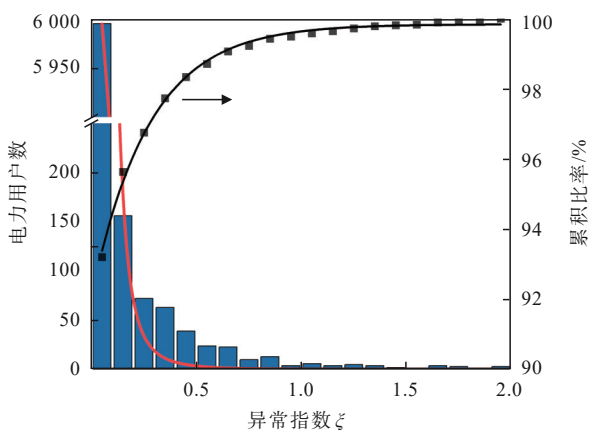


图6 电力用户异常指数的帕累托图

Fig.6 Pareto diagram of abnormality index for power users

为了验证模型中各个模块的有效性,对优化算法模块、特征降维模块和离群异常指数做部分消融实验。其中未使用IWOA算法确定截断距离参数的模型,其DPC截断距离取距离矩阵 D 中所有距离由小到大排列的2%位数。消融实验结果如表4所示,本文模型对每一聚类簇进行异常辨识的结果如图7所示。

表4 不同模型性能对比

Table 4 Performance comparison of different models

模型	召回率/%	精确率/%
IWOA + LDA + 本文离群异常指数	76.00	78.76
IWOA + PCA + 本文离群异常指数	46.00	45.08
IWOA + FA + 本文离群异常指数	31.00	28.49
IWOA + LDA + 文献[23]离群异常指数	59.50	63.21
LDA + 本文离群异常指数	67.50	72.02
LDA + 文献[23]离群异常指数	56.50	61.66

可以发现,在特征降维方面,LDA在本文的异常辨识中相比PCA和FA有着绝对的优势,这是因为LDA作为一种有监督学习方法考虑了类别信息,在降维的过程中最大化类间差异,同时最小化类内差异,这是LDA模型在本文异常检测模型中表现优异的主要原因。在离群异常指数方面,其选择对模型的检测结果影响很大,实验结果表明所提的离群异常指数相比于文献[23]中的异常指数在离群点检测方面更具优越性,召回率和准确率都得到很大的提升。在DPC截断距离参数选择方面,使用IWOA寻优的参数相比于DPC传统参数设置方法具有更好的检测效果。

4.3.3 优化算法性能比较

为了验证对传统WOA改进的有效性,分别利用传统的WOA和文章提出的IWOA求解式(21)中的BI和对应的最佳截断距离 d_{c_best} 。其中算法参数:

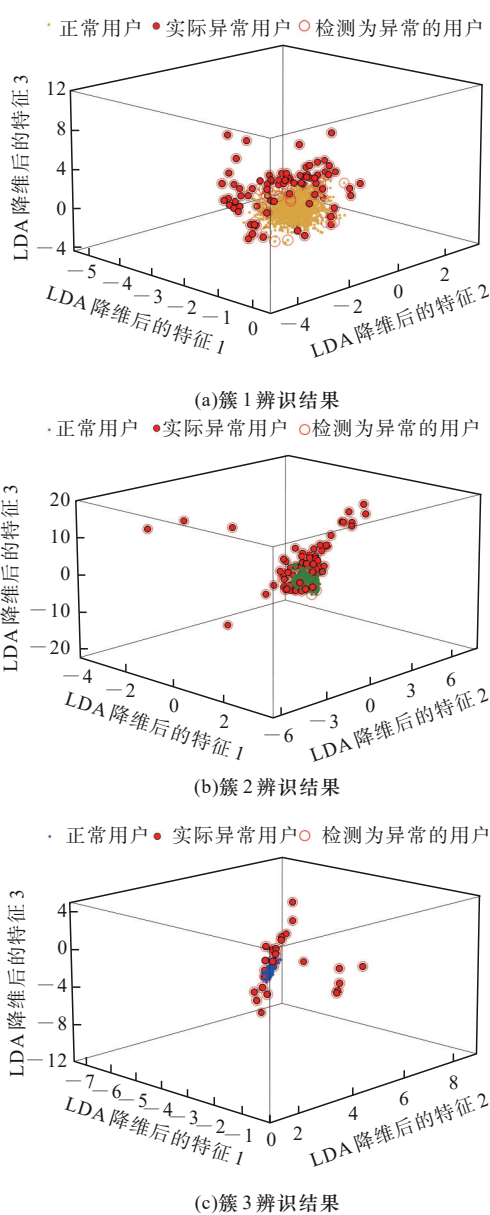


图7 本文模型分簇异常辨识结果

Fig.7 Clustering abnormality identification results of the proposed model

最大迭代次数 $T_{max}=100$,种群规模 $N=20$,求解对比效果如图8所示。

在第2次迭代之后,WOA算法陷入了局部最优解。而本文提出的IWOA算法克服了传统WOA算法易陷入局部最优解的问题,在进行了31次迭代后达到了收敛状态。此时,BI值为3.6389,截断距离为0.0826。实验结果表明文章所提的IWOA算法在寻优过程中表现出色,具有更高的稳定性和求解精度。

5 结语

文章提出的异常用电行为辨识模型包括特征

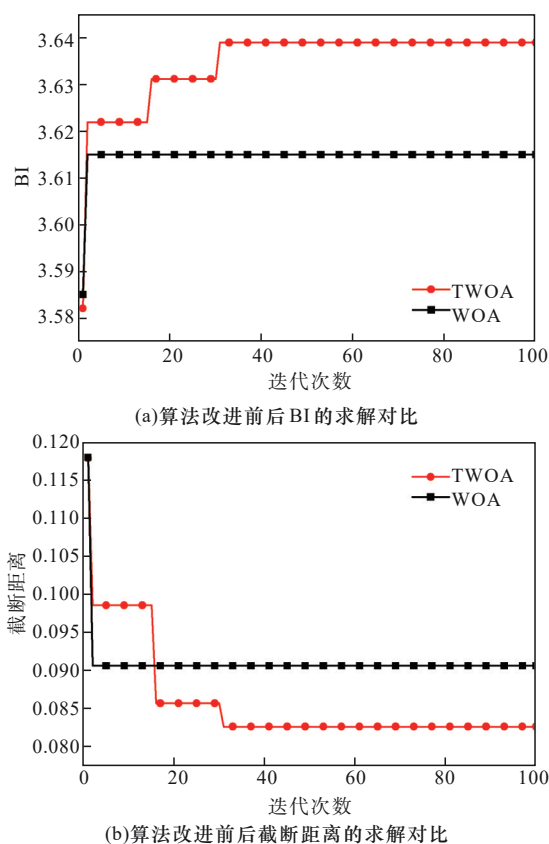


图8 算法改进前后BI与截断距离求解对比

Fig.8 Comparison of BI and truncation distance solutions before and after algorithm improvement

提取、特征降维、密度峰值聚类、离群异常指数计算等步骤。首先从用户用电量数据集中提取可以表征用户用电行为的多维特征;然后对比三种特征降维方法的效果;其次采用DPC算法将所有用户降维后的特征依据密度进行聚类,得到代表不同用电行为的多个密度峰值点;最后计算所有数据点的异常指数,并设定阈值判定异常用户。文章的主要贡献总结如下:

1)针对DPC算法中截断距离需要人工设置的不足,根据决策图中数据点的局部密度和相对距离定义邦费罗尼指数函数,并采用改进鲸鱼优化算法求解,以获取最优截断距离参数,使得优化算法具有更高的稳定性和求解精度。

2)针对高密度区域局部离群点容易被低密度区域正常数据淹没的问题,引入平均向心距离和向心相对距离来计算数据点的离群异常指数,更好地描述了局部离群点的离群特性,从而提升模型检出异常用户的性能。

参考文献:

- [1] YORUKOGLU S, NASIBOV F, MUNGAN M, et al. The effect of the types of network topologies on nontechnical losses in secondary electricity distribution systems[J]. IEEE Transactions on Industry Applications, 2016, 52(5): 3631-3643.
- [2] YIP S C, WONG K, HEW W P, et al. Detection of energy theft and defective smart meters in smart grids using linear regression[J]. International Journal of Electrical Power and Energy Systems, 2017, 91: 230-240.
- [3] 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.
- [4] ZHU T Y, AI Q, HE X, et al. An overview of data driven electricity consumption behavior analysis method and application[J]. Power System Technology, 2020, 44(9): 3497-3507.
- [5] 张忠平, 李森, 刘伟雄, 等. 基于快速密度峰值聚类离群因子的离群点检测算法[J]. 通信学报, 2022, 43(10): 186-195.
- [6] ZHANG Z P, LI S, LIU W X, et al. Outlier detection algorithm based on fast density peak clustering outlier factor[J]. Journal on Communications, 2022, 43(10): 186-195.
- [7] 林之岸, 刘晟源, 金伟超, 等. 基于改进局部离群因子的低压用户用电隐患检测方法[J]. 电力系统自动化, 2022, 46(1): 130-138.
- [8] LIN Z A, LIU S Y, JIN W C, et al. Detection method of hidden danger for power utilization of low-voltage users based on local outlier factor-boosting[J]. Automation of Electric Power Systems, 2022, 46(1): 130-138.
- [9] 李国成, 陆俊, 王赞, 等. 基于Bagging二次加权集成的孤立森林窃电检测算法[J]. 电力系统自动化, 2022, 46(2): 92-100.
- [10] LI G C, LU J, WANG Y, et al. Isolated-forest electricity theft detection algorithm based on Bagging secondary weighted ensemble[J]. Automation of Electric Power Systems, 2022, 46(2): 92-100.
- [11] LI Y C, QIU R X, JING S T. Intrusion detection system using Online Sequence Extreme Learning Machine (OS-ELM) in advanced metering infrastructure of smart grid[J]. PloS One, 2018, 13(2): e0192216.
- [12] 武玉坤, 李伟, 倪敏雅, 等. 单类支持向量机融合深度自编码器的异常检测模型[J]. 计算机科学, 2022, 49(3): 144-151.
- [13] WU Y K, LI W, NI M Y, et al. Anomaly detection model based on one-class support vector machine fused deep auto-encode[J]. Computer Science, 2022, 49(3): 144-151.
- [14] ZOU D X, XIANG Y J, ZHOU T, et al. Outlier detection and data filling based on KNN and LOF for power transformer opera-

- tion data classification[J]. Energy Reports, 2023, 9(S7): 9698-9711.
- [10] 王炜韬, 赵健, 王小宇. 基于对冲对抗机制与图注意力网络的异常用电检测[J]. 电力系统自动化, 2022, 46(22): 120-128.
WANG W T, ZHAO J, WANG X Y. Abnormal electricity consumption detection based on hedge-antagonism mechanism and graph attention network[J]. Automation of Electric Power Systems, 2022, 46(22): 120-128.
- [11] BIAN J H, WANG L, SCHERER R, et al. Abnormal detection of electricity consumption of user based on particle swarm optimization and long short term memory with the attention mechanism[J]. IEEE Access, 2021, 9: 47252-47265.
- [12] 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 J Y, ZHANG S F. 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.
- [13] TIAN Y, SEHOVAC L, GROLINGER K. Similarity-based chained transfer learning for energy forecasting with big data [J]. IEEE Access, 2019, 7: 139895-139908.
- [14] 左进, 陈泽茂. 基于改进 K 均值聚类的异常检测算法[J]. 计算机科学, 2016, 43(8): 258-261.
ZUO J, CHEN Z M. Anomaly detection algorithm based on improved K-means clustering[J]. Computer Science, 2016, 43(8): 258-261.
- [15] 肖勇, 郑楷洪, 余忠忠, 等. 基于三次指数平滑模型与 DBSCAN 聚类的电量数据异常检测[J]. 电网技术, 2020, 44(3): 1099-1104.
XIAO Y, ZHENG K H, YU Z Z, et al. Power data anomaly detection based on Holt-winters model and DBSCAN clustering [J]. Power System Technology, 2020, 44(3): 1099-1104.
- [16] 庄池杰, 张斌, 胡军, 等. 基于无监督学习的电力用户异常用电模式检测[J]. 中国电机工程学报, 2016, 36(2): 379-387.
ZHUANG C J, ZHANG B, HU J, et al. Anomaly detection for power consumption patterns based on unsupervised learning[J]. Proceedings of the CSEE, 2016, 36(2): 379-387.
- [17] 孙毅, 李世豪, 崔灿, 等. 基于高斯核函数改进的电力用户用电数据离群点检测方法[J]. 电网技术, 2018, 42(5): 1595-1606.
SUN Y, LI S H, CUI C, et al. Improved outlier detection method of power consumer data based on gaussian kernel function[J]. Power System Technology, 2018, 42(5): 1595-1606.
- [18] 邓明斌, 徐志森, 邓志飞, 等. 基于多特征融合的窃电识别算法研究[J]. 计算机与数字工程, 2017, 45(12): 2398-2401, 2414.
DENG M B, XU Z M, DENG Z F, et al. Study on stealing recognition model based on multi feature fusion[J]. Computer & Digital Engineering, 2017, 45(12): 2398-2401, 2414.
- [19] 李军徽, 张嘉辉, 穆钢, 等. 计及负荷峰谷特性的储能调峰日前优化调度策略[J]. 电力自动化设备, 2020, 40(7): 128-136, 140.
LI J H, ZHANG J H, MU G, et al. Day-ahead optimal scheduling strategy of peak regulation for energy storage considering peak and valley characteristics of load[J]. Electric Power Automation Equipment, 2020, 40(7): 128-136, 140.
- [20] 杨安, 蒋群, 孙钢, 等. 基于金融技术指标的用电数据分析[J]. 计算机应用, 2022, 42(3): 904-910.
YANG A, JIANG Q, SUN G, et al. Power data analysis based on financial technical indicators[J]. Journal of Computer Applications, 2022, 42(3): 904-910.
- [21] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.
- [22] 狄曙光, 刘峰, 孙建宇, 等. 基于改进 ABC 和 IDPC-MKELM 的短期电力负荷预测[J]. 智慧电力, 2022, 50(9): 74-81.
DI S G, LIU F, SUN J Y, et al. Short term power load forecasting based on improved ABC and IDPC-MKELM[J]. Smart Power, 2022, 50(9): 74-81.
- [23] ZHENG K D, CHEN Q X, WANG Y, et al. A novel combined data-driven approach for electricity theft detection [J]. IEEE Transactions on Industrial Informatics, 2019, 15(3): 1809-1819.

(编辑: 杨 鹏)