doi: 10.7690/bgzdh.2025.03.005

基于局部矩阵重构算法的电力用户窃电分析

梁哲辉，李　莹，罗智慧
（广东电网有限责任公司广州供电局，广州 510620）

摘要：针对目前窃电检测模型需要大量样本、训练过程复杂、检测性能有待提高等缺点，提出一种基于局部矩阵重构的检测模型。引入主成分分析用于分析相邻数据样本之间差异的分布特征；利用欧几里德距离和数据分布特性用来检测不同耗电模式之间的差异；引入局部异常值得分，从而确定窃电样本。对某电力公司提供的住宅电力负荷数据进行实验分析，结果表明：所提模型 ROC 曲线下面积为 0.846 3，具有相对稳定的检测阈值和更强的鲁棒性，对配电网窃电检测评估具有一定借鉴作用。

关键词：配电网；窃电检测；主成分分析；局部矩阵；欧氏距离
中图分类号：MT7　　**文献标志码：**A

Analysis of Electricity Stealing by Power Users Based on Local Matrix Reconstruction Algorithm

Liang Zhehui, Li Ying, Luo Zhihui
(*Guangzhou Power Supply Bureau of Guangdong Power Grid Co*., *Ltd*., *Guangzhou* 510620, *China*)

Abstract: In view of the shortcomings of the current electricity theft detection model, such as requiring a large number of samples, complex training process, and the detection performance needs to be improved, a detection model based on local matrix reconstruction is proposed. The principal component analysis is introduced to analyze the distribution characteristics of the differences between adjacent data samples; the Euclidean distance and data distribution characteristics are used to detect the differences between different power consumption modes; the local outlier score is introduced to determine the power theft samples. The residential power load data provided by a power company is used for experimental analysis, and the results show that the area under the ROC curve of the proposed model is 0.846 3, which has a relatively stable detection threshold and stronger robustness, and has a certain reference for the electricity theft detection and evaluation of distribution network.

Keywords: distribution network; electricity theft detection; principal component analysis; local matrix; Euclidean distance

0　引言

近年来，许多自然灾害和人为因素给电力系统[1-2]带来了前所未有的挑战，由此导致的功率损耗问题给电力系统带来了严重的经济损失和安全威胁。一般情况下，功率损耗主要包括技术损耗（technical losses，TL）和非技术损耗（non-technical losses，NTL）。其中技术损耗主要有电力传输过程中固有 功率损失引起；非技术损耗通常指涉及窃电、电表读数错误或有缺陷的电表引起的损耗。随着网络、大数据、物联网、通信等技术[3-5]不断发展，可通过这些先进的技术手段或系统，如高级计量系统（advanced measurement infrastructure，AMI）分析电力用户数据，从而检测电力系统中窃电模式，并识别非法客户的窃电行为。这对于保障配电网的可靠运行和减少供电企业的经济损失具有重要意义。

目前，分析窃电检测方法主要分为 3 类：统计数据分析、有监督机器学习和无监督机器学习。基于统计数据分析[6]的检测方法侧重于配电网状态估计或客户负荷预测。如果估计或预测结果与实际客户数据存在显著差异，则认为观察到的数据异常。窃电行为复杂多变，统计数据分析出的窃电曲线推广性较差。主流的无监督学习方法为聚类分析[7-8]。无监督学习方法模型简单，实现较为容易。与统计数据分析方法类似，训练好的无监督学习方法仅适用于特定场景，模型推广性较差。此外，现有主流方法主要从全局范围检测异常数据，将其中特定样本与所有其他样本进行比较，而不考虑数据样本的局部分布特征。对于具有时间不规律、与正常数据差异小的窃电数据只考虑全局特征是完全不够的。

随着机器学习不断发展，大量学习对有监督机器学习的检测方法进行研究。文献[9]提出了基于密

收稿日期：2024-07-07；**修回日期：**2024-08-11
第一作者：梁哲辉（1985—），男，广东人，硕士。

集卷积神经网络和随机森林模型，充分利用深度学习方法特征提取能力较强的优势，并结合随机森林可有效实现窃电检测。文献[10]提出了基于 SDAE 和双模型联合训练的低压用户窃电检测方法。文献[11]提出基于多头注意力机制的用户窃电行为检测模型，引入多头注意力机制来进一步增强关键特征的区分度，并通过加深网络来提高学习效果。与无标签机器学习方法相比，前述深度网络算法[9−11]在分类性能和 NTL 检测与分析方面都有所提高。部分方法需要制作大量标签数据，这限制有标签监督方法的普适性。深度学习模型较为复杂，需要耗费大量训练时间。在大多数情况下，电力公司并不了解其所在地区的 NTL 具体细节；因此，有必要继续研究基于无监督机器学习的检测方法来解决 NTL 检测问题。

为改善现有窃电检测模型需要大量样本、训练过程复杂、检测性能有待提高等缺点，提出一种基于局部矩阵重构的检测方法。该方法为一种无标签学习方法，综合考虑数据的局部特征，可实现对时间不规律、与正常数据差异小的窃电数据稳定的异常点检测和更强的鲁棒性。

1　基础知识

笔者所提基于局部矩阵重构的检测算法是一种无监督的窃电检测方法。该方法通过分析相邻样本数据分布的差异进行窃电检测。将检测算法中涉及的基本概念进行介绍。

定义 1：k−距离。

假定在用电数据集 C 中，任意数据 $a\in C$ 和 $o\in C$ 的欧几里德距离(模长)定义为 $\mathrm{dist}(a, o)$。对于任何正整数 $\mathrm{dist}(a, o)$，$a\in C$ 的 k 距离表示为 $\mathrm{dist}_k(a)$。因此，当且仅当以下各式成立时，有 $\mathrm{dist}(a, o)=\mathrm{dist}_k(a)$：

1) 至少有 k 个点 $o'\in C\setminus\{a\}$ 在不包括 a 的集合中，且有 $\mathrm{dist}(a, o')\leqslant\mathrm{dist}(a, o)$；

2) 最多有 k−1 个点 $o'\in C\setminus\{a\}$ 在不包括 a 的集合中，且有 $\mathrm{dist}(a, o')\leqslant\mathrm{dist}(a, o)$。

定义 2：　k−邻域距离。

给定任意数据 $a\in C$ 的 k 距离，其 k−邻域距离表示为 $N_k(a)$。$N_k(a)$为与数据 a 的距离不大于 k 距离的所有数据：

$$N_k(a)=\{b\notin C\setminus\{a\}|\mathrm{dist}(a, b)\leqslant\mathrm{dist}_k(a)\}。\tag{1}$$

根据定义，a 的 k−邻域距离中的点数不小于 k，即 $N_k(a)\geqslant k$。

定义 3：主成分分析(principal component analysis，PCA)。

PCA[12]是一种广泛使用的坐标变换技术，可通过降低数据的维数来显示高维数据的特征，即通过正交变换将一组可能相关的变量转换为一组线性无关变量。变换后的变量称为主分量，该值可从协方差矩阵的特征值分解中获得。

通过 PCA 可确定数据分布的主要方向，即变化幅度最大的方向。当样本的分布特征相似时，样本的重建残差较小；当一个或几个异常样本混合到正常样本中时，异常样本的重构残差大于正常样本的重构残差。通过重构可有效检测窃电造成的异常用电样本。令 $\boldsymbol{A}=\{x_1, x_2, \cdots, x_i, \cdots, x_n\}$为 $p\times n$ 维用电样本矩阵，其中 x_i 为 p 维空间中的第 i 个样本，n 为样本总数。则 $\boldsymbol{A}$ 的协方差矩阵为：

$$\boldsymbol{CO}(\boldsymbol{A})=\boldsymbol{V}(\boldsymbol{A})\times\boldsymbol{D}(\boldsymbol{A})\times\boldsymbol{V}(\boldsymbol{A})^{\mathrm{T}}。\tag{2}$$

式中：$\boldsymbol{CO}(\boldsymbol{A})$为 $\boldsymbol{A}$ 的协方差矩阵；$\boldsymbol{V}(\boldsymbol{A})$为 $p\times p$ 的正交矩阵，其每列为 $\boldsymbol{CO}(\boldsymbol{A})$的特征向量；$\boldsymbol{D}(\boldsymbol{A})$为具有特征值 $\lambda_i(i=1, 2, \cdots, p)$的 P 维对角矩阵。同时，第 1 个累积贡献率 $h(h\leqslant p)$的最大特征值可计算如下：

$$\gamma_h=\frac{\sum_{i=1}^{h}\gamma_i}{\sum_{i=1}^{p}\gamma_i}\times 100\%。\tag{3}$$

此外，PCA 中投影和重构 2 个主要部分 $Y_h(A)$ 和 $R^h(A)$分别由以下公式计算：

$$Y_h(\boldsymbol{A})=A\times\boldsymbol{V}^h(\boldsymbol{A})；\tag{4}$$

$$R^h(\boldsymbol{A})=Y^h(\boldsymbol{A})\times(\boldsymbol{V}^h(\boldsymbol{A}))^{\mathrm{T}}。\tag{5}$$

式中 $\boldsymbol{V}^h$ 为一个 $p\times h$ 维矩阵，包含对应于 h 的最大特征值的特征向量，即 top−h 主分量。

在投影和重构之后，重构残差可计算如下：

$$\mathrm{re}(x_i)=x_i-r_i，\ i=1, 2, \cdots, p。\tag{6}$$

式中 r_i 为矩阵 $R^h(\boldsymbol{A})$的第 i 行。

当一个异常样本被添加到一组正常样本中时，数据集的主成分方向发生了变化；因此，可以通过主成分方向角度的变化来判断添加样本的异常程度。当样本的分布特征相似时，样本的重建残差较小；当一个或几个异常样本混合到正常样本中时，异常样本的重构残差大于正常样本的重构残差。基于上述特征，主成分分析可以作为一种有效的用电数据窃电检测工具。

定义 4：日负荷曲线。

日负荷曲线反映了一天的电力消耗方式。一般情况下，采样频率越高，可获得的信息越多。但对于窃电检测，较高的采样率并不一定会获得更好的检测性能。分析原因，主要是因为对于较低的采样率，窃电对每个样本的影响更明显。同时，采样频率越高，计算复杂性越高，且计算时间越长。为平衡上述问题，使用 5 个日负荷特征分析电力用户用电特征，具体如表 1 所示。可以看出：最小荷载系数为最小荷载值与最大荷载值的比值，反映了观察期内荷载变化的范围；负荷率为平均负荷值与最大负荷值之比，反映了观测期间负荷曲线的分散性。同时，不同采样频率的数据可以很容易地转换成上述指标，从而便于不同用电数据之间的比较分析。

表 1 不同优化策略下性能对比结果

描述	时间段	日负荷特征	定义	物理意义
一天	00:00～24:00	负载率 最小载荷系数	$a_1=P_{av}/P_{\max}$ $a_2=P_{\min}/P_{\max}$	反映全天负荷的波动 反映全天的负荷变化范围
高峰时间	08:00～11:00 18:00～21:00	高峰负荷期间的负荷率	$a_3=P_{av}^{P}/P_{\max}^{p}$	反映高峰负荷期间的负荷波动
稳定时间	06:00～08:00 11:00～18:00 21:00～22:00	稳定负荷期间的负荷率	$a_4=P_{av}^{s}/P_{\max}^{s}$	反映稳定负荷期间的负荷波动
谷值时间	22:00～24:00 00:00～06:00	谷负荷期间的负荷率	$a_5=P_{av}^{v}/P_{\max}^{v}$	反映谷负荷期间的负荷波动

综上，每个用电客户的典型用电特征 DLC 可以表示如下：

$$\mathrm{DLC}_j=\{a_{j,1}, a_{j,2}, \cdots, a_{j,5}\}。 \tag{7}$$

式中 $a_{j,i}$ 为第 j 个用户的第 i 个特征。

2 基于局部矩阵重构的孤立点检测方法

根据前述 PCA 以及局部矩阵理论可实现异常点检测，基于前述理论分析提出一种基于局部矩阵重构的孤立点检测方法，快速检测违反相邻数据点统计分布的窃电值。基于局部矩阵重构的孤立点检测方法的流程如图 1 所示。

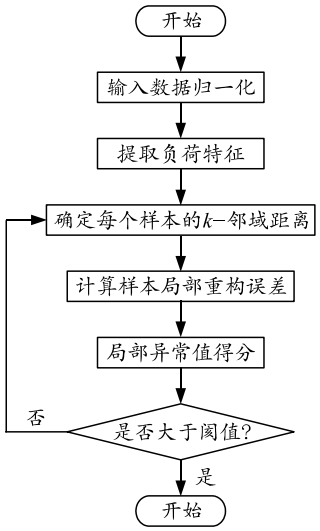

图 1 基于局部矩阵重构的孤立点检测方法的流程

2.1 日负荷特征

每个用电客户的负荷特征矩阵可表示为：

$$\boldsymbol{L}_j=\left\{L_j^1, L_j^2, \cdots, L_j^d, \cdots, L_j^T\right\}； \tag{8}$$

$$L_j^d\left(d=1,2,\cdots,T\right)=\left\{l_{j,1}^d, l_{j,2}^d, \cdots, l_{j,m}^d\right\}。 \tag{9}$$

式中：L_j^d 为客户 j 的第 d 天的日负荷曲线；m 为日负荷曲线的采样频率。进一步，可计算用电客户的平均日负荷特征：

$$L_j'=\frac{1}{T}\left\{\sum_{d=1}^{T}l_{j,1}^d, \sum_{d=1}^{T}l_{j,2}^d, \cdots, \sum_{d=1}^{T}l_{j,m}^d\right\}。 \tag{10}$$

式中 T 为统计周期(通常可取 1 个月或 1 年)。为避免工作日和周末之间用电行为差异的影响，只考虑工作日的日负荷特征。

2.2 局部分布矩阵

根据前述对 k 距离和 k 邻域距离的定义，任意数据 p 的 k 邻域距离定义如下：

$$N_k(p)=\left\{o_1(p), o_2(p), \cdots, o_K(p)\right\}。 \tag{11}$$

式中：$K=|N_k(p)|$；$K\geqslant k$；p 为待检测的用电客户。邻域的作用主要描述如下：将数据的局部重建误差与其邻域距离中的样本进行比较。如果是正常样本，则其局部重构误差小于其邻域距离中其他样本的局部重构误差；同理，如果为异常样本，则其局部重建误差比其邻域距离中其他样本的重建误差大。

同时，p 的局部分布矩阵可以表示为：

$$\boldsymbol{M}(p)=\left\{\mathrm{DLC}_{o_1(p)}, \mathrm{DLC}_{o_2(p)}, \cdots, \mathrm{DLC}_{o_K(p)}\right\}。 \tag{12}$$

式中 $\boldsymbol{M}(p)$为 $K\times5$ 维矩阵。

进一步，可计算 $\boldsymbol{M}(p)$的协方差矩阵 $\boldsymbol{CO}(\boldsymbol{M}(p))$ 并对其进行特征值分解：

$$\boldsymbol{CO}(\boldsymbol{M}(p))=\boldsymbol{V}(\boldsymbol{M}(p))\times\boldsymbol{D}(\boldsymbol{M}(p))\times\boldsymbol{V}(\boldsymbol{M}(p))^{\mathrm{T}}。 \tag{13}$$

式中：$\boldsymbol{V}(\boldsymbol{M}(p))$为一个 5×5 的矩阵；$\boldsymbol{V}(\boldsymbol{M}(p))$的每一

列为 $\boldsymbol{CO}(\boldsymbol{M}(p))$的特征向量；$\boldsymbol{D}(\boldsymbol{M}(p))$为一个 5×5 对角矩阵，其对角元素为 $\boldsymbol{CO}(\boldsymbol{M}(p))$的特征值($\lambda_{p,1}$, $\lambda_{p,2}$, …, $\lambda_{p,5}$)。

令矩阵[$\boldsymbol{M}(p)$; DLC$_p$]为 p 及其 k 邻域距离的典型用电特征，因此矩阵[$\boldsymbol{M}(p)$; DLC$_p$]的主成分空间可计算如下：

$$Y^h(\boldsymbol{M}(p))=[\boldsymbol{M}(p);\ \mathrm{DLC}_p]\times\boldsymbol{V}^h(\boldsymbol{M}(p))。\tag{14}$$

式中：$\boldsymbol{V}^h(\boldsymbol{M}(p))$为矩阵 $\boldsymbol{V}(\boldsymbol{M}(p))$的前 h 列，对应于最大的 h 个特征值，且有 h=1, 2, …, 5；DLC$_p$ 为 p 的典型用电特征。

同时，可重构局部分布矩阵，具体计算如下：

$$\boldsymbol{R}^h(\boldsymbol{M}(p))=\boldsymbol{Y}^h(\boldsymbol{M}(p))\times\boldsymbol{V}^h(\boldsymbol{M}(p))^{\mathrm{T}}。\tag{15}$$

式中 $\boldsymbol{R}^h(\boldsymbol{M}(p))$为使用 top-$h$ 主分量重构后的局部分布矩阵。

在确定重构局部分布矩阵过程中，可计算 p 的局部重构误差，具体计算如下：

$$er(p)=\sum_{h=1}^{5}\mathrm{DLC}_P-R_{K+1}^h(\boldsymbol{M}(p))\times\gamma_h(p)；\tag{16}$$

$$\gamma_h(p)=\frac{\sum_{s=1}^{h}\lambda_{p,s}}{\sum_{s=1}^{5}\lambda_{p,s}}。\tag{17}$$

式中：$R_{K+1}^h(\boldsymbol{M}(p))$为矩阵 $\boldsymbol{R}^h(\boldsymbol{M}(p))$的 K+1 行；$\lambda_{p,s}$为矩阵 $\boldsymbol{CO}(\boldsymbol{M}(p))$的最大特征值；$\gamma_h(p)$为前 h 个主成分在所有主成分中的比例。理论上，h 越小，重构误差计算中考虑的主成分越少，矩阵重构的效果越差。为此，引入 $\gamma_h(p)$作为乘法因子，从而平衡不同 h 对局部重构误差的影响。

对数据集中的所有数据执行上述过程，并将数据 p 的局部重建误差仅与其 k 邻域距离中的样本进行比较。如果 p 是正常样本，则其局部重构误差小于其 k 邻域距离中其他样本的局部重构误差；同理，如果 p 为异常样本，则其局部重建误差比其 k 邻域距离中其他样本的重建误差大。引入局部异常值得分 $S(p)$用来描述数据 p 与其相邻样本之间的差异，具体计算如下：

$$S(p)=\frac{1}{K}\sum_{i=1}^{K}\frac{\mathrm{er}(p)\times\mathrm{dist}(p,o_i(p))}{\mathrm{er}(o_i(p))}。\tag{18}$$

式中：dist(p, $o_i(p)$)为 p 与其 k 邻域距离 $o_i(p)$内的其他数据之间的距离；K 为总样本大小。

进一步，设置阈值 σ，如果 $S(p)>\sigma$，则 p 将视为窃电；同时，阈值 σ 可由电力公司根据 NTL 估计值设定。

3　仿真与分析

3.1　数据集与仿真环境

仿真时所用数据集为某电力公司提供的住宅电力负荷数据。该数据集包含 2018 年 3 月至 2020 年 12 月 5 034 个住宅用户一年的电力负荷数据，采集周期为每 60 min/次。其中，存在窃电行为用户 1 063 个。考虑到数据不完整、缺失等情况，对数据集执行数据清洗操作；同时，为保护客户数据隐私，将带有标识性信息(如姓名、家庭人数等)全部替换为数字 ID，从而有效区分不同用电客户。

仿真软件环境基于 python 搭建检测算法；同时，算法运行硬件环境为酷睿 i7 CPU，内存为 64 G ARM 的联想服务器，操作系统为 Ubuntu 18.04 64 位，显卡为 NVIDIA RTX2080Ti。

3.2　对比与分析

为验证所提检测算法性能，实验时选取 ROC 曲线和 PR 曲线为指标，分别与局部异常因子算法[13](local outlier factor，LOF)和高斯核函数 LOF 算法[14](GKLOF)进行对比。

3.2.1　综合对比与分析

验证不同检测算法在数据集中综合性能，实验时选取 k 邻域距离为 50。

图 2 为不同检测算法在数据集中的 ROC 曲线对比结果。其中：横轴 FPR 表示当前被错误分到正样本类别中真实的负样本所占所有负样本总数的比例；纵轴 TPR 表示当前分到正样本中真实的正样本所占所有正样本的比例。可以看出，LOF 的曲线下面积(AUC)为 0.780 4。GKLOF 的 AUC 为 0.829 7。所提模型 AUC 为 0.846 3。由此可见，所提模型检测性能优于传统 LOF 和 GKLOF 检测算法。

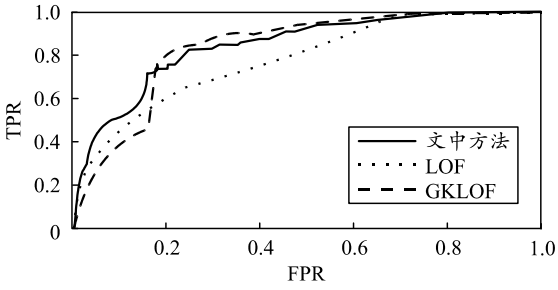

图 2　不同检测算法在数据集中的 ROC 曲线

图 3 为不同检测算法 PR 曲线对比结果。其中横轴为召回率，纵轴为精度。可以看出，所提检测算法的 PR 曲线位于其他 2 种算法的 PR 曲线的右上角；因此，所提检测算法的检测性能优于 LOF 和

GKLOF 检测算法。

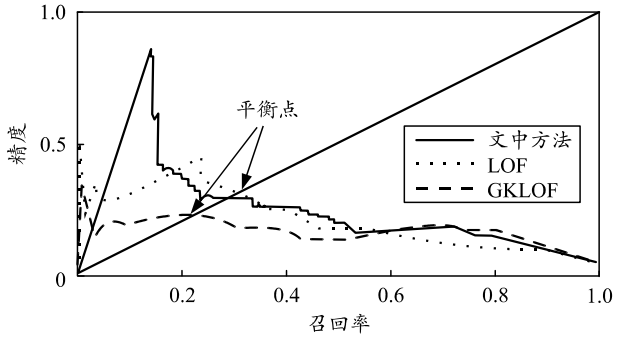

图 3　不同检测算法在数据集中的 PR 曲线

3.2.2　多尺度对比与分析

验证不同 k 值对检测性能影响。表 2 为 ROC 曲线的 AUC 统计结果。可以看出，随着 k 的变化，所提方法的 ROC 曲线和相应的 AUC 略有变化。然而，随着 k 变小，LOF 和 GKLOF 检测算法的 AUC 值急剧下降，其检测性能迅速恶化。结果表明，所提算法具有相对稳定的检测阈值和更强的鲁棒性。

表 2　多尺度 ROC 曲线的 AUC 统计结果

参数 k 取值	AUC		
	所提模型	LOF	GKLOF
20	0.824 8	0.658 9	0.582 1
30	0.840 3	0.701 7	0.640 2
40	0.812 6	0.739 8	0.759 3
50	0.846 3	0.780 4	0.829 7

4　结论

笔者建立一种基于局部矩阵重构的窃电检测方法，可实现对电力用户用电数据的可靠分析与评估。该模型为电力系统智能化、安全化管理提供了一定借鉴作用。

参考文献：

[1] 卢赟, 邓婧, 王渝红, 等. 电力系统受极端天气的影响分析及其适应策略[J]. 发电技术, 2021, 42(6): 751−764.

[2] 李雪, 孙霆锴, 侯恺, 等. 极端天气下电力系统大范围随机设备故障的 N-k 安全分析及筛选方法[J]. 中国电机工程学报, 2020, 40(16): 5113−5126.

[3] 毛龙灿, 杨南. 基于大数据背景的皮革人才培养优化研究[J]. 中国皮革, 2021, 50(9): 38−41.

[4] 杨涛. 互联网时代下皮革行业电商物流体系研究分析[J]. 中国皮革, 2021, 50(8): 82−85.

[5] 钟建栩, 余少锋, 廖崇阳, 等. 基于云计算的电力设备智能监测系统[J]. 云南师范大学学报(自然科学版), 2022, 42(3): 37−41.

[6] 宋少杰, 张长胜, 李英娜, 等. 基于曲线相似度和集成学习的窃电识别[J]. 数据通信, 2022(3): 39−44.

[7] 袁于程, 黄健, 谢晨旸. 基于聚类算法的防窃电监测与辨识[J]. 自动化仪表, 2021, 42(10): 22−26.

[8] 万泉, 袁葆, 刘辉舟, 等. 基于用电行为分析的绕越窃电识别研究[J]. 电力信息与通信技术, 2022, 20(6): 115−121.

[9] 蔡嘉辉, 王琨, 董康, 等. 基于DenseNet和随机森林的电力用户窃电检测[J]. 计算机应用, 2021, 41(S1): 75−80.

[10] 招景明, 唐捷, 潘峰, 等. 基于 SDAE 和双模型联合训练的低压用户窃电检测方法[J]. 电测与仪表, 2021, 58(12): 161−168.

[11] 肖丁, 张玥璠, 纪厚业. 基于多头注意力机制的用户窃电行为检测[J]. 计算机科学, 2022, 49(1): 140−145.

[12] 赖伟平, 林笔星. 基于 PCA-MP-BP 的智能电网数据融合方法[J]. 微型电脑应用, 2022, 38(1): 198−201.

[13] LIU Y, YUAN R, ZHENG S, et al. An abnormal detection of positive active total power based on local outlier factor[C]//2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA). IEEE, 2021: 180−183.

[14] 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.