doi: 10.7690/bgzdh.2024.02.010

基于粗糙熵加权密度的电网调控系统异常检测

田　江，吕　洋，赵　奇，徐秀之，赵　慧
（国网江苏省电力有限公司苏州供电分公司电力调度控制中心，江苏 苏州 215004）

摘要：针对调控系统运行过程中出现的异常状态，提出基于粗糙熵加权密度的智能电网调控系统运行异常数据检测方法。采用粗糙集对电网调控系统运行数据进行分析和推理；利用隶属度的割关系方法，将复杂的不确定关系转化为布尔数据并排序；基于对象加权密度对智能电网调控系统运行中出现的异常数据进行检测，实现对调控系统各种功能异常状态数据准确识别。采用真实电网调控系统数据对所提方法进行验证，结果表明：该方法与传统异常状态识别方法相比，具有更高准确率和更低漏判率。

关键词：粗糙熵；割关系；加权密度；电网调控数据；异常检测
中图分类号：TM743　　**文献标志码：**A

Anomaly Detection of Power Grid Control System Based on Weighted Density of Rough Entropy

Tian Jiang, LYU Yang, Zhao Qi, Xu Xiuzhi, Zhao Hui
(*Power Dispatching Control Center*, *State Grid Suzhou Power Supply Company*, *Suzhou* 215004, *China*)

Abstract: Aiming at the abnormal state in the operation process of the control system, a detection method of abnormal data in the operation of smart grid control system based on rough entropy weighted density is proposed. The rough set is used to analyze and reason the operation data of power grid control system. The complex uncertain relationship is transformed into Boolean data and sorted by using the cut relationship method of membership degree. The abnormal data in the operation of smart grid control system are detected based on the weighted density of objects, and the abnormal data of various functions of the control system are accurately identified. The proposed method is verified by the real power grid control system data, and the results show that the proposed method has higher accuracy and lower omission rate compared with the traditional abnormal state identification method.

Keywords: rough entropy; cut relation; weighted density; power grid control data; anomaly detection

0　引言

电网调度控制系统作为电网运行控制和调度生产管理的核心支撑，其健康状况是保障电网安全、稳定运行的关键[1]。随着大规模分布式发电并网以及大范围随机性多元负荷接入，智能电网调控系统监控设备范围不断扩大，直采直控厂站数量和种类快速增加，导致了由于干扰源、数据采集和网络传输等环节造成的数据质量下降风险[2]。实时检测并识别智能电网调控系统中的异常数据，是保证系统稳定安全运行、协助运维人员实时掌握系统健康状况的关键因素之一。

目前，电力行业内相关系统异常数据检测工作的研究主要集中在电力设备运行、用电量计量等应用范畴。文献[3]通过使用瞬态能量函数作为机器学习算法的输入特征，识别和预测电网设备故障；文献[4]基于电网态势感知理论，提出了适用于低信噪比环境的基于样本协方差矩阵最大特征值的电网设备异常数据检测方法；文献[5]提出基于密度聚类技术的电力系统用电量异常分析算法，高效识别用电信息异常数据；文献[6]提出了一种基于熵序列的智能电网数据流异常状态监测方法。

现有算法未涉及电网调控系统运行数据异常识别领域。由于智能电网调控系统构成复杂、服务器数量多[7]，异常数据不确定性强，为评价系统运行状况，需采集并分析服务器硬件设备、软件进程业务等多种类型的数据[8]，异常数据评价准则多样化。故现有方法无法适应区域智能电网调控系统运行状态中异常数据的辨识需求。

离群点是显著不同于其他数据分布的数据对象，通过分析离群点数据分布特征可以从海量数据中挖掘异常信息、提取兴趣模式[9]。基于离群点监测的异常识别方法在电力行业中已有应用[10-13]。在

收稿日期：2023-10-23；**修回日期：**2023-11-25
基金项目：2021 年江苏省电力有限公司科技项目(J2021046)
第一作者：田　江(1981—)，男，内蒙古人，硕士。

多样化大数据环境下，系统运行数据不仅体量巨大而且种类繁多[14]。针对快速、准确发现调控系统运行异常数据检测需求，笔者提出一种基于粗糙熵加权密度的智能电网调控系统运行异常数据检测方法。

1　智能电网调控系统运行异常数据检测方法

以电力为代表的工业控制领域大多通过设置状态参量指标越限阈值发现异常数据。尽管这种基于阈值的方法简便易行，但指标阀值设定标准多依赖专家经验，且未超出阈值的异常数据无法被及时发现。另外，硬件工况状态与软件进程执行状态关联因素多，单一阀值算法局限性大，判断准确率低。离群点检测目的是通过数据挖掘方法找出不同于大规模数据中的异常点，并发现潜在的、有意义的信息量。实时辨识和分析离群点数据，对于正确识别智能电网调控系统运行态势、提高系统运行质量具有积极和广阔的应用前景。

1.1　粗糙集

粗糙集是一种刻画不完整性和不确定性的数学工具，不仅能有效地分析不精确、不一致、不完整等各种不完备的信息，还可以对数据进行分析和推理，从中发现隐含的知识，揭示潜在的规律。粗糙集已被广泛应用于知识发现、机器学习、决策支持、模式识别、专家系统及归纳推理等领域。

粗糙集理论遵循上近似和下近似[15]2 个概念。在处理不确定数据时，粗糙集主要通过从数据中提取规则来参与决策过程[16]。它不需要任何先验信息，可以直接处理数据以获得解决方案。

在二元组 $G=(P,B)$中，对任意 $F\subseteq P$ 和 B 的不可区分关系 $B\in \mathrm{IND}(G)$，可以定义关于 F 的上近似 $\overline{B}F$ 和下近似 $\underline{B}F$，且上下近似的差异 $(\underline{B}F-\overline{B}F)$ 为 F 的边界域。

$$\overline{B}F=\{r\in P:[r]_B\cap F\neq\varnothing\}; \tag{1}$$

$$\underline{B}F=\{r\in P:[r]_B\subseteq F\}。 \tag{2}$$

m 的下近似基于完全确定性分类，包含于集合 F，集合 F 与属性集合 $B(\underline{B}F)$是相关的；f 的上近似基于包含于集合 F和属性集合 $B(\overline{B}F)$的成员对象分类。

电网调控系统的运行数据中，异常数据同样存在模糊性和不确定性。利用粗糙集概念处理不确定数据。通过对单个聚类应用所提出的基于粗糙熵的加权密度方法来识别离群对象，计算对象和条件属性(不包括决策属性)的加权密度值，可以实现现有方法无法完成的异常值识别，从电网调控系统运行数据中发现异常状态。

1.2　中性集

在同时处理不相容、不准确和不完整的信息时，Smarandache 提出了比粗糙集更广泛的中性集[17]。设 Y 为点或对象组成的空间，令 y 为 Y 中的一般元素。则 Y 的中性集 S 可用隶属函数表示：真实隶属函数 TS、不确定性隶属函数 IS 和虚假隶属函数 FS。如果函数 $TS(y)$、$IS(y)$和 $FS(y)$是真实标准[0, 1]中的单区间或单子集，则：

$TS(y)\rightarrow[0,1]$；$IS(y)\rightarrow[0,1]$；$FS(y)\rightarrow[0,1]$。

则对象 y 被表示为中性集 S，$y=y(T, I, F)\in S$。其中：T 表示真实隶属度；I 表示不确定性隶属度；F 表示虚假值隶属度。从分析的角度，中性集推广了经典模糊集、区间模糊集和直觉模糊集的概念。中性集在工程和科学领域的应用中，需要做出一定限定[18]，导致传统中性集方法在处理如智能电网调控系统等实时应用时会存在不适用的情况。

模糊集近似是基于一种清晰的近似，从而产生模糊粗糙集的概念。将粗糙集的概念引入到中性集中，可以提供来自各种信息系统的知识[19]。为此，将中性集与粗糙集相结合，以确定粗糙度及其区间，作为中性粗糙集，可以处理具有近似值(如下限和上限)的模糊数据。

1.3　基于粗糙熵的调控数据加权密度离群点检测

离群对象是单个对象的异常行为或形成的与其他对象不一致的小簇。它们在空间或时间位置出现异常，形成一个称为异常或离群值的集群。离群点检测作为数据挖掘到一个重要研究方向，在生成模式和大型数据库信息检索等领域都有较广泛的应用。目前针对参数和非参数、单变量和多变量均已提出了多种离群点检测算法，常用的离群点检测方法主要有基于统计、基于深度、基于密度、基于距离和基于聚类等类型[20]。如果数据集中存在离群值，则可以通过提供合适的估计值，重点关注离群值的变化情况以保持系统的稳健性。

在智能电网调控系统中，离群数据主要由硬件设备故障、软件进程阻塞和交互数据异常等造成，这些异常数据存在不确定性；但传统离群点检测方法对此类数据的处理效果不够理想[21]。针对电网调

控系统运行中产生的异常数据类型不一致，且具有高度不确定性的特点，笔者提出一种基于粗糙熵加权密度的电网调控系统运行数据异常检测算法，其实现过程如图 1 所示。

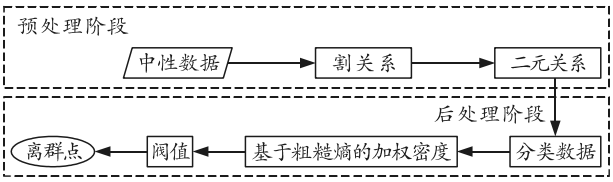

图 1 基于粗糙熵加权密度的电网调控系统运行数据异常检测过程

在预处理阶段，将实时采集的智能电网调控系统运行原始数据作为中性集以(T, I, F)的形式输入。通过应用割关系(α, β, γ)值，将数据集的二元关系由真实、不确定和虚假隶属度值实现。在后处理阶段，通过排序将布尔数据集转换为分类数据，基于粗糙熵方法计算数据点的加权密度离，根据设置的阀值，检测出离群点异常数据。

1.3.1 预处理阶段

电网调控系统运行数据主要包括硬件设备状态、业务进程状态和交互数据状态 3 种类型的数据，对于采集的这 3 类数据，用真实隶属度 α、不确定隶属度 β 和虚假隶属度 γ 表示中性集，其中(α, β, γ)为通过对其应用割关系将各类输入数据转换成的布尔数据。判断条件为$\geqslant\alpha$ 且$\leqslant\beta$ 和 γ，如果条件为真，则应标记为 1，否则标记为 0，表示形式为(α, β, γ)。

定义数据集信息表为一个三元组 $DS=(U, O, A)$，其中：U 表示全域，O 表示对象集，A 表示属性集。应用割关系，获得信息表中各元素的布尔值，然后将布尔值转换为分类数据。

对于给定信息表 $DS=(U, O, A)$，对于任意 $AT\subseteq A$，全域数据中由 AT 所决定的不可分辨的关系，表示为：

$$\{U|\mathrm{IND}(AT)\}=\{[o_i]_{AT}|o_i\in U;[a_j]_{AT}|a_j\in U\}。\quad(3)$$

式中：o_i 为对象集中 O 中的任意第 i 个元素；a_j 为属性集 A 中的任意第 j 个元素。通过识别每个属性中的相似数据可获得不可分辨的关系。

1.3.2 后处理阶段

在后处理阶段，将预处理阶段获得的不可分辨关系，通过排序获得分类数据，然后定义不可分辨性、熵、属性和对象的平均权重，以识别异常值。

设 $DS=(U, O, A)$，对于任意 $AT\subseteq A$，令 $U/\mathrm{IND}(AT)=\{A_1, A_2, \cdots, A_n\}$，定义基于 AT 的补码熵 CE 为：

$$CE(AT)=\sum_{j=1}^{n}|o_j|/|U|\left(1-|o_j|/|U|\right)。\quad(4)$$

A_j^k 为 A_j 的补码集：

$$A_j^k=U-A_j。\quad(5)$$

设 $DS=(U, O, A)$，基于 A 的属性权重 W 定义为：

$$W(A)=(1-CE(AT))\Big/\sum_{j=1}^{n}A_j。\quad(6)$$

对于每个属性，计算其加权密度 Den，

$$\mathrm{Den}(A_j)=|[A_j]|A/|U|。\quad(7)$$

然后，计算每个对象的加权密度 $W\mathrm{Den}$：

$$W\mathrm{Den}(O)=\sum_{o_j\in O}(\mathrm{AvgDen}(o_j)\cdot U(A))。\quad(8)$$

式中 $\mathrm{AvgDen}(A_j)$为属性 A_j 的平均加权密度。

最后得到的平均属性加权密度，可以直观地反映各对象数据的合理性。对于电网调控系统数据集 $DS=(U, O, A)$，从对象加权密度值来看，正常工作数据的加权密度值应该在合理范围内。设∈为固定阈值，电网调控系统正常工作数据对象的加权密度 $W\mathrm{Den}(O)$应该＞∈；否则，数据对象 O 将被视为离群异常值。

2 实验和结果分析

智能电网调控系统运行数据规模大、类型多，通常以系统日志数据的方式存储。笔者提出的方法在对日志数据进行分类，预处理后，基于粗糙熵加权密度检测智能电网调控系统离群数据，判断系统工作状态。离群点判断流程如图 2 所示。

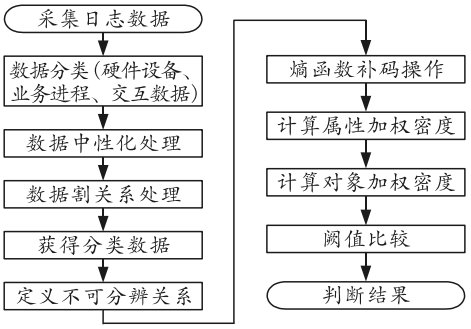

图 2 基于粗糙熵加权密度检测区域电网调控系统离群数据判断流程

采集电网调控系统日志数据时，将数据分为硬件设备状态、业务进程状态和交互数据状体 3 种类型，分别对应调控系统 3 方面的运行状态。每种类型的数据具体指标在笔者提出的方法中作为属性处理，属性分配如表 1 所示。

表 1　种数据状态类型属性分配

状态类型	属性
硬件设备	O_1：CPU 利用率；O_2：内存利用率； O_3：硬盘利用率；O_4：网络连接速率
业务进程	O_1：进程占用内存比；O_2：线程个数； O_3：工作时间；O_4：告警个数；O_5：网络连接个数； O_6：前置处理；O_7：稳态监控；O_8：数据服务； O_9：公共基础应用
交互数据	O_1：前置通道工况；O_2：重要实采数据不变； O_3：误码率；O_4：越限；O_5：跳变；O_6：异常波动； O_7：状态估计合格率；O_8：CPS 指标

对于采集的各类数据，首先进行中性化处理，将各类型数据转化为满足输入要求的具有(U, O, A)值的电网调控系统数据集。然后应用割关系(α, β, γ)获得布尔值，将电网调控系统数据集转化为布尔形式，通过对布尔值排序获得分类数据。接着对于每个属性 $O_i \in U$，定义不可分辨关，并对数据集的熵函数进行补码。然后就可以获得每个属性的加权密度，进而计算得到每个对象的加权密度值。最后将每个对象的加权密度值与规定阈值做比对，小于阈值的加权密度是异常值，该对象被判断为异常数据对象。

2.1　硬件设备状态数据异常检测

以硬件设备状态数据为例，进行离群点检测。硬件设备状态的属性有 CPU 利用率、内存利用率、硬盘利用率和网络连接速率 4 个，表示为 $O=\{O_1, O_2, O_3, O_4\}$。测试数据选取华东某市电网调控系统日志数据，随机采样了 100 组硬件设备数据进行离群点检测，以其中 8 组数据为例，进行检测过程描述。

8 组测试数据即为 8 个具有 4 个属性的对象，表示为 $O=\{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$，属性表示为 $A=\{A_1, A_2, A_3, A_4\}$，其中：A_1 表示 CPU 利用率，A_2 表示内存利用率，A_3 表示硬盘利用率，A_4 表示网络连接速率。将 8 个对象数的属性数据进行类型转换，得到满足输入要求的电网调控系统数据集，如表 2 所示。

表 2　测试数据集

O	A_1	A_2	A_3	A_4
O_1	(0.05,0.85,0.05)	(0.10,0.25,0.65)	(0.50,0.40,0.10)	(0.10,0.05,0.85)
O_2	(0.05,0.90,0.05)	(0.10,0.25,0.65)	(0.50,0.40,0.10)	(0.10,0.05,0.85)
O_3	(0.05,0.85,0.10)	(0.05,0.15,0.80)	(0.10,0.85,0.05)	(0.40,0.05,0.55)
O_4	(0,0.10,0)	(0.70,0,0)	(0.80,0,0)	(0.30,0,0.70)
O_5	(0.05,0.90,0.05)	(0.05,0.90,0.05)	(0.10,0.80,0.10)	(0.15,0.15,0.70)
O_6	(0.10,0.20,0.70)	(0.90,0.05,0.05)	(0.10,0.75,0.75)	(0.30,0.20,0.50)
O_7	(0.10,0.80,0.10)	(0.20,0.80,0.10)	(0.10,0.80,0.10)	(0.10,0.10,0.90)
O_8	(0.50,0.25,0.25)	(0.70,0.10,0.20)	(0.50,0.25,0.25)	(0.50,0.25,0.25)

用 α、β 和 γ 分别表示真、不确定和假的隶属度，文中将(α, β, γ)的割关系固定为$(0.05, 0.90, 0.85)$，即真值≥0.05，β 和 γ 值≤0.90 和 0.85。如果条件满足，则表示数据为 1 或标记为 0，如表 3 所示。

表 3　应用割关系布尔化处理后数据集

O	A_1	A_2	A_3	A_4
O_1	0	1	1	1
O_2	1	1	1	1
O_3	1	1	1	1
O_4	0	1	1	1
O_5	1	0	0	0
O_6	1	0	1	1
O_7	1	1	1	1
O_8	1	1	0	0

根据上表，将 CPU 利用率、内存利用率、硬盘利用率和网络连接速率的属性排序为高或低，布尔化处理为 0 的指标属性设置为低，布尔化处理为 1 的指标属性设置为高。

根据笔者提出的基于粗糙熵的加权密度离群点检测算法，首先计算每个属性的不可分辨关系：

$\mathrm{IND}(A_1)=\{O_1, O_4\}\{O_2, O_3, O_5, O_6, O_7, O_8\}$；

$\mathrm{IND}(A_2)=\{O_1, O_2, O_3, O_4, O_7, O_8\}\{O_5, O_6\}$；

$\mathrm{IND}(A_3)=\{O_1, O_2, O_3, O_4, O_6, O_7\}\{O_5, O_8\}$；

$\mathrm{IND}(A_4)=\{O_1, O_2, O_3, O_4, O_6, O_7\}\{O_5, O_8\}$。

接着计算每个属性不可分辨值的补码熵：

$$CE(A_1)=\frac{2}{8}\left(1-\frac{2}{8}\right)+\frac{6}{8}\left(1-\frac{6}{8}\right)=\frac{3}{8};$$

$$CE(A_2)=\frac{6}{8}\left(1-\frac{6}{8}\right)+\frac{2}{8}\left(1-\frac{2}{8}\right)=\frac{3}{8};$$

$$CE(A_3)=\frac{3}{8};\quad CE(A_4)=\frac{3}{8}。$$

然后根据补码粗糙熵计算每个属性的加权密度：

$$\mathrm{Den}(A_1)=\mathrm{Den}(A_2)=\mathrm{Den}(A_3)=\mathrm{Den}(A_4)=5/12。$$

最后，计算每个对象的加权密度值：

$$W\mathrm{Den}(O_1)=\frac{2}{8}*\frac{5}{12}+\frac{6}{8}*\frac{5}{12}+\frac{2}{8}*\frac{5}{12}+\frac{2}{8}*\frac{5}{12}=1.04;$$

$$W\mathrm{Den}(O_2)=W\mathrm{Den}(O_3)=W\mathrm{Den}(O_4)=W\mathrm{Den}(O_7)=1.4;$$

$$W\mathrm{Den}(O_5)=1.6;\quad W\mathrm{Den}(O_6)=1.04;\quad W\mathrm{Den}(O_8)=1.2。$$

根据对象加权密度值，基于历史经验设置固定阈值以识别离群点异常值。在本次实验中，算例数据选取华东某市电网 2022 年前 2 个季度数据，从 261 200 组数据样本中，选取 3 480 条异常状态数据进行分析，确定固定阈值为 1.4。由于对象 O_5 的加权密度值高于设定阈值，因此该状态检测结果为异常。

2.2　电网调控系统离群点数据检测

笔者设计的智能电网调控系统离群点数据检测

方法不仅适用于硬件设备状态数据的分析与挖掘，也适用于业务进程状态数据和交互数据状体数据异常识别，进而保证区域电网电控系统安全稳定运行。采集华东某市电网 2021 年后 2 个季度和 2022 年前 2 个季度电网调控系统日志数据，随机选择 250 组日志数据，分别采用基于欧式距离的方法(EM)、局部离群因子算法(LOF)和分类离群因子算法(COF)检测日志数据中的异常状态，并与笔者提出的方法(RDen)进行对比，测试结果如表 4—7 所示。

表 4 测试对象数 21 时

测试方法	离群点数	准确率/%	漏判率/%
EM	4	95.2	0
LOF	0	85.7	100
COF	4	95.2	0
RDen	3	100	0

表 5 测试对象数 66 时

测试方法	离群点数	准确率/%	漏判率/%
EM	8	100	0
LOF	5	85.5	37.5
COF	4	93.9	50
RDen	8	100	0

表 6 测试对象数 121 时

测试方法	离群点数	准确率/%	漏判率/%
EM	10	100	0
LOF	3	94.2	70
COF	7	98.3	30
RDen	10	100	0

表 7 测试对象数 247 时

测试方法	离群点数	准确率/%	漏判率/%
EM	28	97.1	9.5
LOF	12	96.3	42.8
COF	9	95.1	57.1
RDen	17	98.4	0

从 250 组日志数据中抽取不同数量的测试对象，测试对象中均包含硬件设备状态数据、业务进程状态数据和交互数据状体数据，且各类测试对象的数量随机。经人工比对确认，分别得到异常状态判断正确率和异常漏判率 2 个指标，并以此进行异常识别效果评价。

从对比结果可知，电网调控系统运行过程中的异常状态并不会随着测试数据的增加而大幅上升，各种方法均能对异常状态进行较为有效的识别，特别是测试数据量较大时，识别准确率会进一步提高，而笔者提出的方法具有更高的判断准确率。同时，异常状态漏判率对于电网调控系统也尤为重要，若出现误判，可由人工操作解决；而漏判则有可能造成无法挽回的损失。从表 4—7 可见，笔者提出的方法未发生漏判情况，该方面性能明显优于其他 3 种方法。

3 结论

笔者提出一种基于粗糙熵加权密度的智能电网调控系统运行异常数据检测方法，主要结论如下：

1) 采用粗糙集对不精确、不一致、不完整的电网调控系统运行数据进行分析和推理，得到数据间的隐含关联性。

2) 利用隶属度的割关系方法，将复杂的不确定关系转化为布尔数据，实现对关系数据的降维，有利于离群点数据的处理与发现。

3) 提出基于对象加权密度的智能电网调控系统运行异常数据检测方法，提高了各类功能数据异常识别的准确性，降低了漏判率。相比传统方法，在异常状态识别方面具有明显优势。

参考文献：

[1] OURAHOU M, AYRIR W, HASSOUNI B E, et al. Review on smart grid control and reliability in presence of renewable energies: Challenges and prospects[J]. Mathematics and computers in simulation, 2020, 167(1): 19-31.

[2] 谈林涛, 李军良, 任冕, 等. 基于 RB-XGBoost 算法的智能电网调度控制系统健康度评价模型[J]. 电力自动化设备, 2020, 40(2): 189-195.

[3] GEEGANAGE J, ANNAKKAGE U D, WEEKES M A, et al. Application of energy-based power system features for dynamic security assessment[C]//2015 IEEE Power & Energy Society General Meeting. IEEE, 2015: 1.

[4] 周忠强, 韩松. 基于样本协方差矩阵最大特征值的低信噪比环境电网异常状态检测[J]. 电力系统保护与控制, 2019, 47(8): 113-119.

[5] 田力, 向敏. 基于密度聚类技术的电力系统用电量异常分析算法[J]. 电力系统自动化, 2017, 41(5): 64-70.

[6] 刘庆连, 王雪平. 智能电网大数据异常状态实时监测仿真[J]. 计算机仿真, 2019, 36(3): 364-367.

[7] 黄晴晴. 电力调度自动化系统中健康度评价系统的设计与开发[D]. 北京: 北京邮电大学, 2018.

[8] 杨洁. 基于机器学习的智能电网调度控制系统在线健康度评价研究[D]. 北京: 北京邮电大学, 2019.

[9] 梅林, 张凤荔, 高强. 离群点检测技术综述[J]. 计算机应用研究, 2020, 37(12): 7-13.

[10] 耿俊成, 张小斐, 周庆捷, 等. 基于局部离群点检测的低压台区用户窃电识别[J]. 电网与清洁能源, 2019, 35(11): 30-36.

[11] 郭丽娟, 张玉波, 尹立群, 等. 基于离群点检测的变电主设备异常辨识与规律分析[J]. 南方电网技术, 2018, 12(9): 14-21.

（下转第 57 页）