DOI:10.12158/j.2096-3203.2023.04.018

基于流聚类的 PMU 异常数据辨识算法

邓小玉[1]，王向兵[1]，曹华珍[1]，王流火[1]，严洪峰[2]，王宏宇[2]
（1. 广东电网有限责任公司，广东 广州 510699；
2. 江苏金智科技股份有限公司，江苏 南京 211100）

摘 要：为保证同步相量测量装置（phasor measurement unit，PMU）采集数据的准确应用，须排除其量测值中的异常数据。现有 PMU 异常数据辨识算法存在算法复杂度高、难以在线更新、多源数据难以校准、依赖多源数据应用难度大等不足。为此，文中从 PMU 事件数据和异常数据模型及 PMU 异常数据判别信息熵定义出发，提出基于该信息熵的异常数据辨识框架。在此框架基础上，基于利用层次方法的平衡迭代规约和聚类（balanced iterative reducing and clustering using hierarchies，BIRCH）算法提出 PMU 异常数据辨识算法；然后，对所提出的算法进行原型实现，并针对某变电站的 PMU 采集数据集进行算法实验验证。实验结果表明，与一类支持向量机（one-class support vector machine，OCSVM）算法与间隙统计算法相比，文中算法的准确度及实时性均具有较强的优势。

关键词：同步相量测量装置（PMU）；异常数据；事件数据；辨识框架；信息熵；流聚类
中图分类号：TM63 **文献标志码**：A **文章编号**：2096-3203（2023）04-0167-08

0 引言

目前，电网中大量应用的动态数据采集设备为同步相量测量装置（phasor measurement unit，PMU）[1]。PMU 量测在电力系统内部状态切换时可能产生跳变的事件数据，也会因互感器误差、PMU 设备故障、时间同步异常、通信系统中断等诸多因素[2]产生异常跳变，即量测值中出现异常数据。智能电网的电网安全评估、预防控制和运行分析均以准确的电力系统状态估计为基础，若量测值中存在误差较大的异常数据，将导致系统状态估计准确度降低，影响系统实时监测及控制[3-7]。而基于异常数据所作的决策判断可能威胁到整个电网安全。因此，为保障电网安全可靠运行，研究 PMU 异常数据及其辨识方法有重要意义。

国内外学者已经针对 PMU 异常数据辨识方法展开了深入研究[8-15]。文献[12]提出了一种基于谱聚类的 PMU 异常数据检测算法，采用决策树方法分辨出事件数据，再通过谱聚类进行正常数据与异常数据的辨识，但其决策依赖异常数据维持的时间长度定义，易将事件数据误辨识为异常数据；文献[13]提出了基于核心微簇与离群微簇的异常数据辨识算法，先在线更新潜在核心微簇和离群微簇，再通过基于密度的噪声应用空间聚类（density-based spatial clustering of applications with noise，DBSCAN）算法重新对所有微簇进行离线聚类，更新核心微簇，算法复杂度较高；文献[14]提出了一种基于 PMU 和数据采集与监视控制系统的单一点互校核算法，其准确率高，速度较快；文献[15]提出了一种基于多维特征向量和阈值的异常数据辨识算法。上述算法应用时都要先训练模式分类器再进行异常数据区分，无法实现在线更新，应用难度大。

为了降低辨识算法的复杂度，可以采用流聚类[16-18]算法进行数据的在线训练和聚类。该算法通过提取有效的聚类特征（cluster feature，CF），动态地对数据进行聚类分析，依据少量的特征数据集存储对新数据进行快速处理和分类，因此具有良好的在线更新和实时应用特性，适用于连续采集等具有大数据特征的应用场景。目前，尚未有相关工作将流聚类算法应用于 PMU 异常数据辨识。

信息熵[19]是信息论的基本概念，可描述信息源各可能事件发生的不确定性。采用信息熵对 PMU 量测值进行筛选，可以减少训练集的样本数以用于配电网安全态势感知要素分析[20]。文献[21]通过小波熵对故障状态下的电气量相角特征进行分析，从而获得准确的故障元件。信息熵同样也可用于聚类算法以获得更加准确的分类结果。文献[22]将信息熵作为加权依据，在聚类过程中弱化低质量的簇，使聚类结果更加准确；文献[23]将信息熵应用于混合数据类型的聚类算法研究中，以确定不同数据类型的数据权重；文献[24]在计算聚类距离时引入信息熵对距离进行加权，提升了分类的准确性。文献[25]提出了样本稳定性的概念，利用信息

收稿日期：2023-01-15；修回日期：2023-04-06
基金项目：中国南方电网有限责任公司科技项目（037700KK52190023）

熵描述二元信源的确定性，优先筛选稳定性高的样本簇，提升后续的聚类准确度。综上所述，信息熵用于流聚类主要是从度量和聚类距离方面对聚类进行良性的干预，使算法更为准确。

文中基于信息熵理论，研究 PMU 事件数据和异常数据描述，定义 PMU 异常数据判别熵（PMU abnormal data identification entropy，PADIE），提出基于信息熵的 PMU 异常数据和事件数据描述和辨识框架；将 PADIE 与流聚类算法相结合，提出基于流聚类的在线 PMU 异常数据辨识算法。该算法实现了对异常数据与事件数据的在线、准确、实时辨识。基于文中理论研究所实现的变电站 PMU 数据校核装置可以在站内实时、就地完成 PMU 异常数据识别，改进了传统 PMU 数据上送主站后要从海量数据中依赖多源数据校核识别出异常数据的方法，提升了 PMU 异常数据识别的实时性，降低了运算量，为变电站内保护、测控各装置准确实时应用 PMU 数据提供了更好的技术支撑。

1 基于信息熵的 PMU 异常数据辨识框架

PMU 事件数据是指由于电力系统内部状态切换导致的 PMU 量测值跳变；PMU 异常数据是指电力系统内部状态并未发生变化，而是由于数据采集误差、通信异常等导致的 PMU 量测值跳变。

1.1 PMU 异常数据和事件数据定义

文中参考了文献[12]和[15]中对异常数据的定义。异常数据在偏离正常值后会回到正常值，即其值围绕正常值上下波动；事件数据在偏离正常值后不会回到正常值或需要较长时间再回到正常值。

定义 α 为偏离因子；t_a 为发生数据偏离前的时刻，P_{t_a} 为该时刻的 PMU 数据；t_b 为发生数据偏离后恢复到正常数据的时刻，P_{t_b} 为该时刻的 PMU 数据；P_t 为 t 时刻（$t_a<t<t_b$）跳变过程中的 PMU 数据；$\bar{P}_{t_a}$、$\bar{P}_t$、$\bar{P}_{t_b}$ 分别为 t_a、t 和 t_b 时刻的正常值；t_e 为事件判断时长阈值。当 P_t 满足式（1）时为异常数据。

$$\begin{cases} |P_{t_a}-\bar{P}_{t_a}|/\bar{P}_{t_a} \leqslant \alpha \\ |P_t-\bar{P}_t|/\bar{P}_t > \alpha \quad t_a < t < t_b \\ |P_{t_b}-\bar{P}_{t_b}|/\bar{P}_{t_b} \leqslant \alpha \\ t_b - t_a \leqslant t_e \end{cases} \tag{1}$$

事件数据的特性满足：

$$\begin{cases} |P_{t_a}-\bar{P}_{t_a}|/\bar{P}_{t_a} \leqslant \alpha \\ |P_t-\bar{P}_t|/\bar{P}_t > \alpha \quad t_a < t < t_b \\ |P_{t_b}-\bar{P}_{t_b}|/\bar{P}_{t_b} > \alpha \\ t_b - t_a > t_e \end{cases} \tag{2}$$

由式（2）可知，当发生数据偏离正常值后超过 t_e 时间仍未恢复到正常值时，该数据为事件数据。

1.2 PMU 异常数据判别信息熵定义

信息熵用于描述事件发生的不确定性，定义如式（3）所示。

$$H(D) = -\sum_{i=1}^{n} P(C_i)\log_2(P(C_i)) \tag{3}$$

式中：D 为整个数据集；$H(D)$ 为数据集 D 的信息熵；n 为数据集 D 中的类别个数；C_i 为数据集 D 中第 i 个分类；$P(C_i)$ 为数据集 D 中第 i 个分类的占比。

在异常数据辨识中引入信息熵的概念，定义 PADIE，用于描述一段数据中出现异常数据后的数据不确定度。首先对 PMU 数据的信息进行定义。定义一个样本 C 如下：

$$C = \{P_1, P_2, \cdots, P_N\} \tag{4}$$

式中：P_1—P_N 为 PMU 数据；N 为样本中数据的个数。假如该样本中存在 c 个异常数据，定义 p_c 为样本中异常数据的占比，如式（5）所示。

$$p_c = c/N \tag{5}$$

定义 $1-p_c$ 为样本中正常数据的占比。基于 PMU 数据信息 p_c 和 $1-p_c$，采用信息熵公式构造样本 C 的 PADIE 值 $Z(C)$：

$$Z(C) = -k_1 p_c \log_2(p_c) - k_2(1-p_c)\log_2(1-p_c) \tag{6}$$

式中：k_1、k_2 为权重系数。

1.3 基于 PMU 异常数据判别信息熵的辨识框架

PADIE 反映了 PMU 数据异常的不确定性。按照事件判断时长阈值 t_e 内 PMU 数据点的数量进行样本 C 的选取。

假如 k_1 和 k_2 取值相同：

（1）当 $p_c=0.5$ 时，跳变数据与正常数据比例相同，PMU 数据源的不确定度最大。

（2）当 $p_c<0.5$ 时，p_c 越接近 0，$Z(C)$ 越接近 0，样本 C 的不确定度越小；当 p_c 为 0 时，不确定度为 0，样本 C 中均为正常数据。

（3）当 $p_c>0.5$ 时，p_c 越接近 1，$Z(C)$ 越接近 0，样本 C 的不确定度越小；当 p_c 为 1 时，不确定度为 0，样本 C 中数据偏离持续时间超过了 t_e，所以样本 C 中均为事件数据。

因此，结合 p_c 与 $Z(C)$ 可对数据进行正常数据、异常数据与事件数据的判断。定义连续的样本 C_1、C_2、$\cdots$、C_i，对应的异常数据占比为 p_{c1}、p_{c2}、$\cdots$、p_{ci}，信息熵为 $Z(C_1)$、$Z(C_2)$、$\cdots$、$Z(C_i)$。

当 $Z(C_i)=0$ 时，若 $p_{ci}=0$，则样本 C_i 均为正常数据；若 $p_{ci}=1$，则样本 C_i 均为事件数据。当

$Z(C_i)>0$时,若下一个样本 C_{i+1} 的 $Z(C_{i+1})=0$ 且 $p_{c(i+1)}=1$,即样本 C_{i+1} 为事件数据,则说明事件是在样本 C_i 的时间内发生的,所以 C_i 中的跳变值也为事件数据,否则 C_i 中的跳变值为异常数据。

2 异常数据辨识算法

文中将 PADIE 值作为度量,结合流聚类算法实现 PMU 异常数据辨识。流聚类算法通过对数据序列迭代处理,不断更新 CF,并对不断更新的窗口数据进行迭代聚类,从而实现对连续数据流的快速聚类。与传统聚类方法相比,流聚类方法更适用于动态扩展的数据集,其通过 CF 维护和窗口定义降低了聚类处理的计算性能要求,且不再需要存储全部样本数据,提升了算法的实时性。

2.1 PMU 异常数据判别信息熵计算

PADIE 计算中首先要找出 PMU 数据中的跳变值。按照式(1)和式(2)计算跳变值时,$\bar{P}_t$ 采用滑窗求取均值的方式计算。假设滑窗的样本数量为 m,则在 t 时刻 $\bar{P}_t$ 的计算公式如式(7)和式(8)所示,其中 k 为权重系数。式(8)用于设置数据点的上下限。

$$\bar{P}_t=(m\bar{P}_{t-1}-Q_{t-m}+Q_t)/m \tag{7}$$

$$Q_t=\begin{cases}P_t & -k\alpha\leqslant(P_t-\bar{P}_{t-1})/\bar{P}_{t-1}\leqslant k\alpha\\ \bar{P}_{t-1}(1-k\alpha) & (P_t-\bar{P}_{t-1})/\bar{P}_{t-1}<-k\alpha\\ \bar{P}_{t-1}(1+k\alpha) & (P_t-\bar{P}_{t-1})/\bar{P}_{t-1}>-k\alpha\end{cases} \tag{8}$$

得到正常值后,根据式(1)—式(6)即可完成对 PMU 异常数据判别信息熵的计算。

2.2 目标函数定义

聚类的优化准则为组内距离最小化,组间距离最大化。以信息熵为 CF,对样本及对应的簇内、簇间距离进行定义,定义如下。

将聚类特征簇作为数据集 D 进行信息熵的计算。设 C_i、C_j 为聚类特征簇 D 中的 2 个样本,对应的信息熵指标分别为 $Z(C_i)$、$Z(C_j)$,反映 2 个样本内数据类型的不确定度,两者之间的距离可定义为:

$$L(C_i,C_j)=|Z(C_i)-Z(C_j)| \tag{9}$$

簇 D 的质心 $\bar{Z}_0$ 为:

$$\bar{Z}_0=\sum_{i=1}^{N}Z(C_i)\Big/N \tag{10}$$

任意样本 C_i 到簇 D 的距离定义为样本 C_i 到簇的质心的距离:

$$L(C_i,D)=|Z(C_i)-\bar{Z}_0| \tag{11}$$

设 N_m、N_n 分别为 2 个聚类特征簇 D_m 和 D_n 的样本个数,簇 D_m 和簇 D_n 之间的距离定义为:

$$L(D_m,D_n)=\left[\frac{\sum_{i=1}^{N_m}\sum_{j=1}^{N_n}(Z(C_i)-Z(C_j))^2}{N_mN_n}\right]^{\frac{1}{2}} \tag{12}$$

若存在样本空间$\mathbb{R}^M$,有 M 个簇,每个簇内有 N_i 个样本,令 D_m、$D_n\in\mathbb{R}^M$,则对 $\forall D_m$、D_n,有:

$$\begin{cases}\min\sum_{i=1}^{M}\sum_{\substack{j=1,\\ j\in D_i}}^{N_i}|Z(C_j)-\bar{Z}_{0,i}|\\ \max\sum_{\substack{1\leqslant m,n\leqslant M,\\ m\neq n}}^{M}L(D_m,D_n)\end{cases} \tag{13}$$

式中:$\bar{Z}_{0,i}$ 为簇 D_i 的质心;N_i 为簇 D_i 的样本个数。

2.3 基于 BIRCH 的 PMU 异常数据辨识算法

文中采用利用层次方法的平衡迭代规约和聚类(balanced iterative reducing and clustering using hierarchies,BIRCH)[18]算法实现流聚类。对数据建立分层结构 CF 树,其中每个节点包含一组 CF。这些 CF 包含了描述数据集中一组点的充分统计信息以及子节点指向的 CF 的所有信息。该 CF 树的规模由参数 B(每个中间节点最大的 CF 数)、L(每个叶子节点最大的 CF 数)、T(每个 CF 的聚类半径)决定。具体 CF 树的结构如图 1 所示。

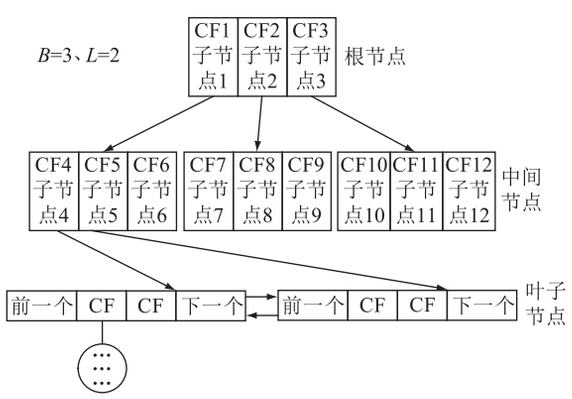

图 1 CF 树

Fig.1 CF tree

CF 结构是一个三元组$(N,\boldsymbol{S}_{\mathrm{LS}},S_{\mathrm{SS}})$,用于存储一组点的充分统计信息,其中 $\boldsymbol{S}_{\mathrm{LS}}$ 为 N 个样本特征维度的向量和;S_{SS} 为 N 个样本特征维度的平方和。CF 三元组满足线性关系,因此可以高效地更新 CF 树。每插入一个样本时,从根开始向下遍历当前树,计算与新样本最近的叶子 CF,新样本到簇的距离以及簇间的距离通过式(11)和式(12)进行计算。找到合适的叶子 CF 后,基于 CF 树中三元组的线性可加特性,可快速更新 CF 值,完成实时在线聚类。

完成聚类后，每个样本可以根据聚类结果实现对正常数据、异常数据与事件数据的辨识。

基于流聚类的 PMU 异常数据辨识的具体流程如图 2 所示。以样本 $C_i=\{P_{i1},P_{i2},\cdots,P_{iN}\}$ 为例，首先计算样本的 PADIE 值 $Z(C_i)$，然后计算样本的 CF，再按照 BIRCH 参数 B、L、T 进行聚类，得到其聚类的类别标识 $S(C_i)$，最后根据聚类结果完成对正常数据、异常数据与事件数据的辨识。

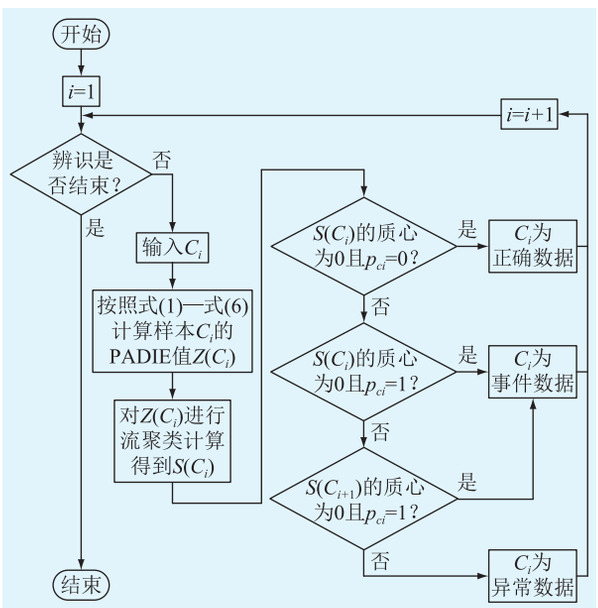

图 2　基于流聚类的 PMU 异常数据辨识流程

Fig.2　Flow chart of PMU abnormal data identification based on stream clustering

3　算法实验与结果分析

文中采用站域多态数据融合测控原型系统实现基于流聚类的 PMU 异常数据辨识算法。中央处理器（central processing unit，CPU）采用 Intel Core i7-8665U 处理器（主频 1.9 GHz），其通过以太网口按照 GB/T 26865.2 协议接收同步相量数据集中器的 PMU 数据，然后进行异常数据辨识。

3.1　异常数据辨识实验

某变电站的 PMU 数据的传输频率为 100 点/s，每个波形持续时间为 60 s，每个波形总计 6 000 个点。对 43 个采集对象的 282 万个 PMU 采集数据进行统计分析，选取图 3 与图 4 为典型波形进行说明。图 3 为采样传感器受到干扰后引起 PMU 数据异常突变与波动的典型波形，图 4 为典型的由负荷切换引起的事件数据波形。

图 3 异常数据特征为：短时突降，维持片刻，随后恢复；短时突升，维持片刻，随后恢复；数据异常波动。图 4 事件数据特性为：偏离正常值后不再返回。

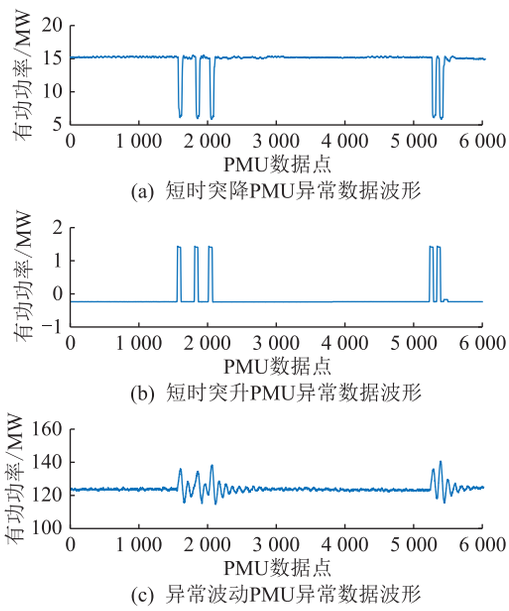

图 3　现场采集的 PMU 异常数据波形

Fig.3　Abnormal data waveforms of PMU collected on site

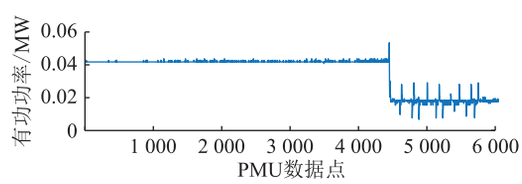

图 4　现场采集的 PMU 事件数据波形

Fig.4　Event data waveform of PMU collected on site

针对以上 PMU 数据样本，采用基于流聚类的 PMU 异常数据辨识算法，其参数取值如下：式（1）和式（2）中的 t_e 取 1 s，α 取 0.05，即正常数据的偏差不超过±5%，α 的取值与实际应用中被辨识数据的正常波动范围有关；式（4）中的 N 取 100；式（6）的 k_1 和 k_2 均取 30，因为文中同时关注正常数据与异常数据，如果取 $k_1>k_2$，则正常数据的比例对信息熵影响更大，因此 k_1 和 k_2 的取值原则是根据应用对正常数据和异常数据的关注程度取值，两者幅值大小只影响聚类时半径的取值；式（7）中的 m 取 300，即以 3 s 为窗口计算参考正常值，如果增大 m，则滑窗数据更为平滑，但响应速度更慢，因此 m 的取值原则是根据辨识数据的变化速率与应用的响应速度要求取值；式（8）中的 k 取 1，即正常值计算时上下限偏差为±5%。BIRCH 算法中 B 取 3，L 取 2，半径 T 取 0.5。由于 p_c 为 0.5 时信息熵达到峰值，对 p_c 取中间值 0.25 时，计算 p_c（即 0.25 与 0.24）的信息熵的差值为 0.487，所以在半径 T 取 0.5 时，$p_c\leqslant 0.25$ 的样本都会单独分为一类，而 $p_c>0.25$ 的样本则会逐

渐合并。

对图 3 中 3 种典型 PMU 异常数据波形进行归一化处理后,得到异常数据辨识结果,如图 5—图 7 所示,纵坐标“聚类结果”是指各样本的信息熵通过流聚类分类后的类别编号,其中类别 0 是质心为 0,即 $Z(C_i)$ 为 0 的点,其他类别是 $Z(C_i)$ 大于 0 的点。按照 1.3 节,可通过 $Z(C_i)$ 与 p_{ci} 对正常数据、异常数据与事件数据进行判别。

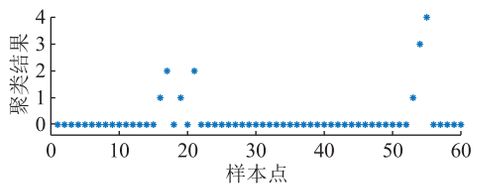

图 5 短时突降辨识结果

Fig.5 Identification results of short-time sudden drop

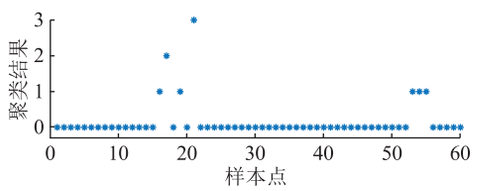

图 6 短时突升辨识结果

Fig.6 Identification results of short-time sudden rise

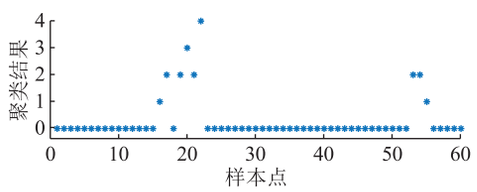

图 7 数据波动辨识结果

Fig.7 Identification results of data fluctuation

图 5—图 7 的聚类结果中未出现聚类类别为 0 且 p_c 为 1 的事件数据样本,可知图中聚类类别不为 0 的样本中的跳变值均为异常数据。

对图 4 的 PMU 事件数据波形进行归一化处理后,得到的异常数据辨识结果如图 8 所示。

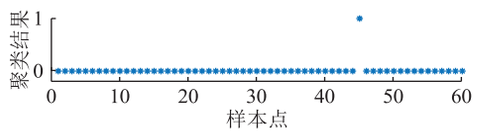

图 8 事件数据辨识结果

Fig.8 Identification results of event data

图 8 中样本 1~44 的聚类类别均为 0,且 p_c 为 0,因此均为正常数据;样本 46~60 的聚类类别均为 0,且 p_c 为 1,因此均为事件数据;由于样本 45 聚类类别为 1,但样本 46 为事件数据,说明事件是在样本 45 的时间内发生,所以样本 45 中的跳变值为事件数据。

基于流聚类的异常数据辨识结果如表 1 所示。辨识结果中“0”为正常数据,“1”为异常数据,“2”为事件数据。该方法对以上样本的异常数据和事件数据的辨识准确率为 100%。

表 1 基于流聚类的异常数据辨识结果

Table 1 Identification results of abnormal data based on stream clustering

数据类型		类别	质心	样本分布	辨识结果
异常数据	短时突降	0	0(p_c=0)	1~15、18、20、22~52、56~60	0
		1	28.97	16、19、53	1
		2	24.10	17、21	1
		3	26.44	54	1
		4	9.82	55	1
	短时突升	0	0(p_c=0)	1~15、18、20、22~52、56~60	0
		1	29.06	16、19、53~55	1
		2	21.04	17	1
		3	27.74	21	1
	异常波动	0	0(p_c=0)	1~15、18、23~52、56~60	0
		1	23.08	16、55	1
		2	29.79	17、19、21、53、54	1
		3	21.04	20	1
		4	25.66	22	1
事件数据		0	0(p_c=0)	1~44	0
		0	0(p_c=1)	46~60	2
		1	29.57	45	2

原型系统中基于流聚类的异常数据辨识算法由 x86 架构 Intel Core i7-8665U 实现。由于 p_c 为 0 或 1 时进行信息熵计算均无须计算对数,因此按照计算最复杂情况,即 0< p_c <1 时统计计算耗时与算法总耗时。信息熵计算包括加减 600 次,乘法 7 次,除法 1 次,对数 2 次,比较 300 次;流聚类计算包括加减 10 次,乘法 2 次,除法 2 次,平方 1 次,比较 5 次;共计时钟周期 1 147 个,耗时 604 ns,实测计算耗时 692 μs。

实验结果表明,对于每个 10 ms PMU 数据通过 692 μs 的时间即可计算完成,满足实时性要求。1 min 6 000 个 PMU 数据作为 1 组实验数据,从 282 万个 PMU 数据中提取异常数据 18 组,事件数据 85 组,经文中算法辨识后,异常数据辨识准确率为 100%。对比某变电站 2022 年采集实际数据的波动特性与文献[15]中其他同行所获取的 2015 年的实际数据,发现其波动特性基本一致。可见在电网运行数据波动特征不发生剧烈变化的前提下,文中算法具有较好的适用性。

3.2 同类算法对比

采用一类支持向量机[26](one-class support vector machine,OCSVM)算法与间隙统计算法[27](gap statistic algorithm,GSA)对 PMU 异常数据进行辨识。OCSVM 算法设置为:使用高斯核,训练数据的异常

点比例为0.01,训练停止的公差标准为0.001。辨识结果为1表明PMU数据点为正常数据,辨识结果为-1表明PMU数据点为异常数据。GSA设置聚类参数 k 的遍历范围为1~10。辨识结果为0表明PMU数据点为正常数据,为其他值表明PMU数据点为异常数据。对图3中的异常数据波形进行归一化处理,采用2种算法所得的辨识结果如图9—图11所示。

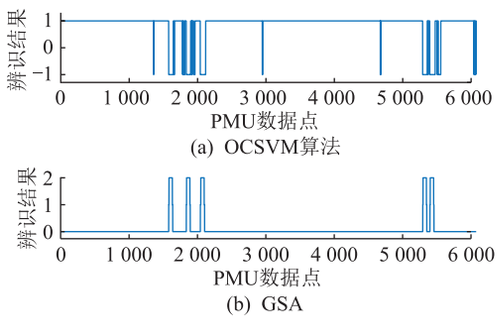

图9 OCSVM算法和GSA的短时突降辨识结果

Fig.9 Identification results of short-time sudden drop using OCSVM algorithm and GSA

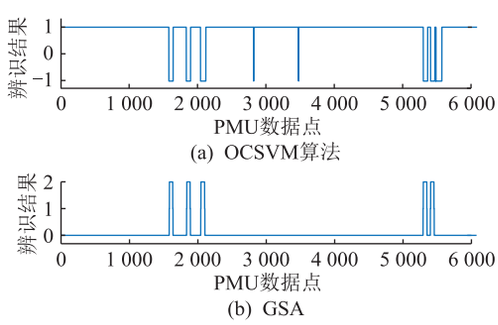

图10 OCSVM算法和GSA的短时突升辨识结果

Fig.10 Identification results of short-time sudden rise using OCSVM algorithm and GSA

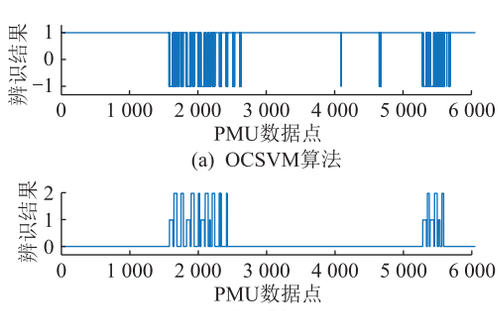

图11 OCSVM算法和GSA的数据波动辨识结果

Fig.11 Identification results of data fluctuation using OCSVM algorithm and GSA

文中算法与OCSVM算法和GSA的比较结果如表2所示。其中 P_1 为异常数据辨识率;P_2 为误辨识率。P_1 和 P_2 的计算分别见式(14)、式(15)。

$$P_1 = n_f/N_f \tag{14}$$

表2 与同类算法辨识指标的比较

Table 2 Comparison of identification indexes with similar algorithms

算法	异常数据类型	P_1/%	P_2/%	辨识100个PMU数据最大耗时/μs
文中算法	短时突降	100	0	692
	短时突升	100	0	
	异常波动	100	0	
OCSVM算法	短时突降	100	2.45	663
	短时突升	100	0.85	
	异常波动	79.26	5.27	
GSA	短时突降	86.39	0	8 732
	短时突升	76.49	0	
	异常波动	100	5.46	

$$P_2 = n_t/N_t \tag{15}$$

式中:n_f 为辨识出的异常数据点数;N_f 为异常数据总数;n_t 为将正常数据误辨识为异常数据的点数;N_t 为正常数据总数。

文中18组异常数据中短时突降、短时突升和数据波动3种异常数据类型占比为7∶8∶3,与文献[15]中的统计数据相当,将3种异常数据类型的辨识结果按照发生比率加权后得到OCSVM算法的 $P_1=96.54\%$、$P_2=2.21\%$;GSA的 $P_1=84.26\%$、$P_2=0.91\%$;文中算法的 $P_1=100\%$、$P_2=0$。

根据以上实验结果可知,OCSVM算法辨识耗时较低,但是辨识准确度不够:部分样本辨识时 P_1 指标能达到100%,但 P_2 指标不稳定,异常波动样本辨识的 P_1 和 P_2 指标均较差。GSA耗时长,辨识准确度不够:部分样本辨识时 P_2 指标能达到0,但 P_1 指标较差,异常波动样本辨识的 P_1 指标虽然达到100%,但 P_2 指标较差。文中算法对于不同样本类型均实现了 P_1 指标达到100%,P_2 指标达到0的辨识效果,且可实现事件数据与异常数据的区分。文中算法的最大耗时也仅比OCSVM算法大4.4%。

4 结语

文中针对PMU异常数据辨识问题,从信息熵理论出发,研究提出PADIE定义和基于该信息熵的PMU异常数据辨识框架。将PADIE和流聚类算法结合,提出一种基于流聚类的PMU异常数据辨识算法。结合原型实现,对从实际运行电网中所获取的PMU量测值数据集进行了所提出辨识算法的实验,实验结果表明文中算法的辨识率和误辨识率均优于同类算法。

文中研究可为今后PMU异常数据的修正、异常数据溯源与智能告警等应用提供参考。如何优化异常数据信息熵的聚类效果和进一步对异常数据进行修复、溯源和告警,是未来要继续研究的方向。

参考文献:

[1] 黄子蒙,余娟,向明旭,等. 基于改进动态时间弯曲的 PMU 频率异常检测及类型识别[J]. 电力系统自动化,2022,46(24):104-112.
HUANG Zimeng,YU Juan,XIANG Mingxu,et al. Frequency anomaly detection and type identification of PMU based on improved dynamic time warping[J]. Automation of Electric Power Systems,2022,46(24):104-112.

[2] 徐飞阳,薛安成,常乃超,等. 电力系统同步相量异常数据检测与修复研究现状与展望[J]. 中国电机工程学报,2021,41(20):6869-6886.
XU Feiyang,XUE Ancheng,CHANG Naichao,et al. Research status and prospects of detection,correction and recovery for abnormal synchrophasor data in power system[J]. Proceedings of the CSEE,2021,41(20):6869-6886.

[3] 周婧怡,李红娇. 针对 PMU 测量的虚假数据注入攻击检测方法[J]. 信息网络安全,2022,22(5):75-83.
ZHOU Jingyi,LI Hongjiao. False data injection attack detection method against PMU measurements[J]. Netinfo Security,2022,22(5):75-83.

[4] 常鹏,吴泽群,孙文仲,等. 基于 PMU 优化部署的电网 CPS 线下攻击保护[J]. 智慧电力,2021,49(6):60-66.
CHANG Peng,WU Zequn,SUN Wenzhong,et al. Offline attack protection of power grid CPS based on PMU optimized deployment[J]. Smart Power,2021,49(6):60-66.

[5] 王海波,吴升涛,周文海,等. 基于不同量测数据融合的配电网状态估计研究[J]. 机电信息,2023(4):12-16.
WANG Haibo,WU Shengtao,ZHOU Wenhai,et al. Research on distribution network state estimation based on different measurement data fusion[J]. Mechanical and Electrical Information,2023(4):12-16.

[6] CUI M J,WANG J H,TAN J,et al. A novel event detection method using PMU data with high precision[J]. IEEE Transactions on Power Systems,2019,34(1):454-466.

[7] 曹鹏,刘敏,杭鲁庆. 基于改进磷虾群算法的配电网 PMU 优化配置研究[J]. 电网与清洁能源,2022,38(4):61-67.
CAO Peng,LIU Min,HANG Luqing. Research on PMU optimal configuration of distribution networks based on improved krill herd algorithm[J]. Power System and Clean Energy,2022,38(4):61-67.

[8] 王磊. 配电网同步相量与电能质量同步监测一体机[D]. 济南:山东大学,2018.
WANG Lei. Integrated machine for synchronous phasor and power quality monitoring of distribution network[D]. Jinan:Shandong University,2018.

[9] 朱志敏. 基于 Linux 的广域测量系统相量数据集中器的研发[D]. 合肥:合肥工业大学,2019.
ZHU Zhimin. Research and development of phasor data concentrator for wide-area measurement system based on Linux[D]. Hefei:Hefei University of Technology,2019.

[10] IDEHEN I,OVERBYE T J. PMU time error detection using second-order phase angle derivative measurements[C]//2019 IEEE Texas Power and Energy Conference (TPEC). College Station,TX,USA. IEEE,2019:1-6.

[11] 张江南,雷江龙,贺勇,等. 基于 PMU 误差校正的输电线路参数在线辨识方法[J]. 电力系统保护与控制,2022,50(19):130-137.
ZHANG Jiangnan,LEI Jianglong,HE Yong,et al. Transmission line parameter identification method based on PMU error correction[J]. Power System Protection and Control,2022,50(19):130-137.

[12] YANG Z W,LIU H,BI T S,et al. Bad data detection algorithm for PMU based on spectral clustering[J]. Journal of Modern Power Systems and Clean Energy,2020,8(3):473-483.

[13] CAO F,ESTERT M,QIAN W N,et al. Density-based clustering over an evolving data stream with noise[C]//Proceedings of the 2006 SIAM International Conference on Data Mining. Philadelphia,PA:Society for Industrial and Applied Mathematics,2006.

[14] 刘雯静,杨军,袁文,等. 一种基于 PMU 和 SCADA 单节点互校核的前端数据辨识框架[J]. 电力系统保护与控制,2020,48(8):1-9.
LIU Wenjing,YANG Jun,YUAN Wen,et al. A front-end data identification framework based on single-node mutual checking between PMU and SCADA[J]. Power System Protection and Control,2020,48(8):1-9.

[15] 万楚林,陈皓勇,郭曼兰. 基于模式识别的 WAMS 有功功率错误数据处理[J]. 电网技术,2017,41(3):922-930.
WAN Chulin,CHEN Haoyong,GUO Manlan. Wrong active power data identification and correction for WAMS based on pattern recognition[J]. Power System Technology,2017,41(3):922-930.

[16] BEZDEK J C,KELLER J M. Streaming data analysis:clustering or classification? [J]. IEEE Transactions on Systems,Man,and Cybernetics:Systems,2021,51(1):91-102.

[17] CARNEIN M,TRAUTMANN H. Optimizing data stream representation:an extensive survey on stream clustering algorithms[J]. Business & Information Systems Engineering,2019,61(3):277-297.

[18] ZHANG T,RAMAKRISHNAN R,LIVNY M. BIRCH:an efficient data clustering method for very large databases[J]. ACM SIGMOD Conference,1996.

[19] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal,1948,27(3):379-423.

[20] 田书欣,李昆鹏,魏书荣,等. 基于同步相量测量装置的配电网安全态势感知方法[J]. 中国电机工程学报,2021,41(2):617-632.
TIAN Shuxin,LI Kunpeng,WEI Shurong,et al. Security situation awareness approach for distribution network based on synchronous phasor measurement unit[J]. Proceedings of the CSEE,2021,41(2):617-632.

[21] 杨欢. 基于 WAMS/SCADA 数据的配电网故障分析方法[D]. 沈阳:东北大学,2018.
YANG Huan. Fault analysis method of distribution network

based on WAMS/SCADA data[D]. Shenyang:Northeastern University,2018.

[22] 邵长龙,孙统风,丁世飞. 基于信息熵加权的聚类集成算法[J]. 南京大学学报(自然科学),2021,57(2):189-196.
SHAO Changlong,SUN Tongfeng,DING Shifei. Ensemble clustering based on information entropy weighted[J]. Journal of Nanjing University (Natural Science),2021,57(2):189-196.

[23] 崔文秀. 基于信息熵定义属性权重的混合数据聚类算法研究[D]. 太原:山西大学,2021.
CUI Wenxiu. Research on mixed data clustering algorithm based on information entropy to define attribute weight[D]. Taiyuan:Shanxi University,2021.

[24] 张安勤,吴蕊,张挺. 基于信息熵的异常检测算法[J]. 上海电力大学学报,2020,36(4):386-390.
ZHANG Anqin,WU Rui,ZHANG Ting. Anomaly detection algorithm based on information entropy[J]. Journal of Shanghai University of Electric Power,2020,36(4):386-390.

[25] 李飞江,钱宇华,王婕婷,等. 基于样本稳定性的聚类方法[J]. 中国科学:信息科学,2020,50(8):1239-1254.
LI Feijiang,QIAN Yuhua,WANG Jieting,et al. Clustering method based on sample's stability[J]. Scientia Sinica (Informationis),2020,50(8):1239-1254.

[26] 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.

[27] BRODINOVÁ Š,FILZMOSER P,ORTNER T,et al. Robust and sparse k-means clustering for high-dimensional data[J]. Advances in Data Analysis and Classification,2019,13(4):905-932.

作者简介:


邓小玉

邓小玉(1975),女,学士,高级工程师,从事电力系统及自动化工作(E-mail:13622837-326@139.com);

王向兵(1981),男,硕士,高级工程师,从事电力系统及自动化工作;

曹华珍(1974),女,硕士,高级工程师,从事电力系统及自动化工作。

PMU abnormal data identification algorithm based on stream clustering

DENG Xiaoyu[1], WANG Xiangbing[1], CAO Huazhen[1], WANG Liuhuo[1], YAN Hongfeng[2], WANG Hongyu[2]

(1. Guangdong Power Grid Co.,Ltd.,Guangzhou 510699,China;
2. Wiscom System Co.,Ltd.,Nanjing 211100,China)

Abstract:In order to ensure the accurate application of the data collected by the phasor measurement unit (PMU),it is necessary to eliminate the abnormal data in its measured values. The existing PMU abnormal data identification algorithm has the disadvantages of high algorithm complexity,difficulty in online updating,difficulty in the calibration of multi-source data,and difficulty in application relying on multi-source data. In this paper,an abnormal data identification framework is proposed based on the PMU event data and abnormal data model and the definition of PMU abnormal data identification information entropy. On the basis of the framework,a PMU abnormal data identification algorithm is proposed based on the balanced iterative reducing and clustering using hierarchies (BIRCH) algorithm. The proposed algorithm is implemented,and an algorithm experiment is carried out for the PMU dataset of a substation. The experimental results show that the proposed algorithm has better accuracy and real-time performance than one-class support vector machine (OCSVM) algorithm and gap statistic algorithm.

Keywords:phasor measurement unit (PMU);abnormal data;event data;identification framework;information entropy;stream clustering

(编辑 陆海霞)