DOI: 10.19783/j.cnki.pspc.190625

一种基于 PMU 和 SCADA 单节点互校核的前端数据辨识框架

刘雯静[1]，杨 军[1]，袁 文[2]，唐云红[2]，谭本东[1]，徐 箭[1]

(1. 武汉大学电气与自动化学院，湖北 武汉 430072；2. 国网湖南省常德供电分公司，湖南 常德 415001)

摘要：随着电网自动化技术的发展，数据中心可获取海量多源多时空数据，在此基础上进行多源量测值互校核有利于实现后续大数据高级应用。针对单节点同时存在 PMU 与 SCADA 量测值的情况，提出一种前端不良数据辨识框架。为克服量测值负样本较少的问题，采用基于粒子群优化的改进一分类支持向量机辨识方法，根据两源量测差值识别异常点。对接近向量机边界可能被误判的值利用间隙统计法进行修正，确定不良数据。然后检验其所在时间点的 PMU 量测值，最终确定不良数据位置。基于某省实际电网数据对 PMU 与 SCADA 互校核辨识框架进行了验证与分析。计算结果表明所提方法能够有效地辨识出两数据源的前端不良数据，计算量小、耗时较短，比仅利用单源数据进行校核的结果更加可靠。

关键词：前端数据辨识；数据采集与监视控制系统；同步相量测量单元；改进一分类支持向量机；间隙统计算法

A front-end data identification framework based on single-node mutual checking between PMU and SCADA

LIU Wenjing[1], YANG Jun[1], YUAN Wen[2], TANG Yunhong[2], TAN Bendong[1], XU Jian[1]

(1. School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China;
2. State Grid Hunan Changde Power Supply Company, Changde 415001, China)

Abstract: With the development of power grid automation technology, massive multi-source and multi-spatial-temporal data can be obtained by a data center. On this basis, multi-source measurement data can be checked reciprocally. This is conducive to later high-level big data applications. Given a situation where PMU and SCADA measurements of the same node exist at the same time, a front-end data identification framework is proposed. To overcome a lack of negative samples of data, an improved One-Class Support Vector Machine (OCSVM) identification method based on particle swarm optimization is used to identify outliers according to the difference between two-source measurements. Any data which may be misjudged because of being close to the OCSVM boundary can be corrected by a gap statistical algorithm and the bad data can be found. Then the PMU data at the time point of the bad data is checked to determine the location. With the actual power grid data of a province, the bad data identification framework based on mutual checking between PMU and SCADA is verified and analyzed. The results show that the proposed method can effectively identify front-end bad data of two sources, with less computation and consumption of time, and is more reliable than the results which use single-source data only.

This work is supported by National Key Research and Development Program of China (No. 2017YFB0902900).

Key words: front-end bad data detection; supervisory control and data acquisition; phasor measurement unit; improved one-class support vector machine; gap statistical algorithm

0 引言

电网中多种能源共存，其规模和结构不断发展，数据类型增多的同时数据量也迅速扩大[1]。调度人员对电网进行观测和调控时，需要掌握全面准确的数据，在多个前端数据来源的量测信息上进行不同形式数据的收集整理、挖掘分析等后续应用。在收集数据的过程中，由于通信传输等原因，各来源的前端量测量之间可能存在矛盾，说明数据中存在与真实值偏移较大的量测值，即不良数据，如缺失值

基金项目：国家重点研发计划项目资助(2017YFB0902900)

和离群值。为了避免坏数据给后续分析带来误差，使调度中心做出错误的决策影响电网后续运行，需对原始量测数据进行清洗，找出不良数据。

数据采集与监视控制系统(Supervisory Control and Data Acquisition, SCADA)在电网中广泛应用，在布点数量和采集数据信息量等方面都占据着绝对的优势，加强了电力系统可观测性[2]，但由于量测误差或传输错误可能存在不良数据。传统的不良数据辨识方法主要包括残差搜索法、零残差法和估计辨识法，还有将状态估计后得到的特征与机器学习、数据融合结合的新方法[3-4]，专家学者基于数据挖掘也提出了一些辨识方法[5-7]。目前同步相量测量单元(Phasor Measurement Unit, PMU)在220 kV及以上输电网中应用较广[8-9]，其密集的量测产生的大量数据也给校核带来挑战，有学者针对此问题利用状态估计[10-13]与异常点检测算法，如模式识别[14]、局部离群因子算法[15]对 PMU 量测值进行不良数据辨识，都集中于对单一源数据进行校核。考虑到 SCADA 与 PMU 均分布在电网中采集量测值，同一节点可能存在 SCADA 与 PMU 同时量测，可互为补充。针对这种情况，文献[16]提出将 SCADA 状态估计结果和PMU量测值相结合进行二次线性状态估计，同时利用二次线性状态估计更新残差协方差矩阵的方式，对 SCADA 量测量中的关键量测量是否存在不良数据进行检测，但其没有对 PMU 数据进行校核，且需要多次迭代，较为复杂。文献[17]提出了电网中多源数据互校核的思想及框架，但是并未阐述具体实施方法。文献[18]利用 PMU 对 SCADA 数据校核，应用场景较为简单，没有考虑 SCADA 量测误差未知和 PMU 量测值出现错误的情况。

总的来说，上述校核方法中与状态估计有关的算法需要多个节点的量测与拓扑信息，计算量较大，耗时较长。基于数据挖掘与异常检测的方法侧重于研究单一数据源的数据规律，没有充分利用其他来源的数据，可能存在漏判及误判的情况。

随着电力系统中多源信息的不断丰富，合理利用其进行不良数据检测的识别结果将比单一来源判断更可靠。针对数据平台中多源前端数据辨识问题，本文基于多源数据平台中已对时的两源前端量测值，提出了一种能基于单个节点数据准确校核两源量测值的框架：将两源对同一节点的量测值对时、插值、作差，得到的差值作为检测的特征量；为克服电网中缺少不良数据样本的问题，仅基于正常情况下的两源差值对改进的一分类支持向量机进行训练，使用粒子群算法得到向量机最佳的训练参数，再利用训练好的向量机对待检测数据进行快速初筛；之后使用间隙统计算法对接近支持向量机边界的数据进行核验，确定两源数据是否存在矛盾；辨识出存在矛盾后，对该时刻的 PMU 数据进行不良数据辨识，最终确定不良数据源。

1 PMU 和 SCADA 数据量测现状分析

由于数据传输通道不通畅或数据传输中受到影响，电力系统的实时量测中可能出现噪声数据。对某省级调度中心 D5000 系统中的 PMU 设备和 SCADA 系统历年量测数据进行了分析，发现上传至主站的数据经常会存在少部分的错误数据。SCADA 主要存在的问题是数据不刷新，当电网状态发生变化时，数据未能及时更新。PMU 主要是量测量存在跳变的现象，出现短暂数值尖峰或者一段时间零值，又或在正常值与零值之间跳变。电力系统运行时数据一般不会出现较大变化，但数据有波动并不意味着一定包含不良数据，仅靠单一来源数据难以准确识别出坏数据。

综合数据平台中的电网海量信息由于来自不同的信息系统，同一节点上的信息往往有差异，甚至存在矛盾。SCADA 系统和 WAMS 系统提供的量测量中都包括节点的电压幅值和支路电流幅值，当两源的数据集成到综合数据平台中会成为同一个节点或设备的属性数据，经过对时处理可以直观地看到两个量测的差值，据此对两源量测进行分析。某实际电网 D5000 系统已经可以收集到 127 个 220 kV 及以上变电站与电厂的两源量测信息，给数据分析提供了基础。表 1 是某实际电网 D5000 系统里收集的经过对时后的两源量测值的状况。

表 1 量测值信息

Table 1 Measurement data

量测点	SCADA/kV	PMU/kV	量测差值/kV
1	532.35	532.75	−0.4
2	532.88	532.75	0.13
3	33.30	33.26	0.04
4	533.79	177.68	−356.11

节点 1、2、3 的差值都比较小，此时系统中也未报出数据错误，说明两者都是正确数据，而节点 4 处的数据相差很大，已达到 SCADA 量测值的 66.7%。线路电压等级为 500 kV，而 PMU 测得数据只有 177.68 kV，明显是 PMU 的量测中存在不良数据。电网中节点众多，各测量源的量测误差也存在区别，仅依靠人工检测两源差异必然难以实现。

针对 PMU 与 SCADA 对同一节点存在同一物理量量侧的情况，由于两者分别属于两个独立运行的系统，具有较高的可靠性，可以认为两个量测系

统在系统内同一地点、同一时刻均出现不良数据的概率极小[19]。因此本文不考虑两源同时错误的情况，小概率情况下有一方出现错误。测量设备都存在偶然误差，测量精度也有不同，在正常情况下两个数据源测得的数据也不免存在差值，但不会超出某个范围。当一方出现不良数据时，差值增大且与正常数据表现出差异，说明在这个时刻 PMU 和 SCADA 的量测值存在“非此即彼”的错误。

2　不良数据位置辨识

在综合数据平台中，对同一节点同时存在的 SCADA 和 PMU 量测值进行对时，使 SCADA 量测值与对应时刻的 PMU 数据对准。由于 PMU 的量测尺度比 SCADA 密集，冗余数据可以采用拉格朗日插值的方法补充 SCADA 数据进行数据对应，在相邻两量测值间均匀插值，直到实现两源量测数据的匹配，每一个 PMU 量测值都有与之对应的 SCADA 量测值。利用同一节点两源同一时刻的量测差值可以进行不良数据位置的辨识，当差值不超过 “门槛值”时认为都是正常数据。为了避免人为设置阈值的主观性影响结果的正确率，可使用机器学习算法通过训练学习后进行判断。

2.1 基于 OCSVM 的异常点识别

从电网获取的大量数据中，错误数据的比例很小，无法收集到足够的负样本点的训练集，因此本文利用一分类支持向量机[20](One Class Support Vector Machine, OCSVM)，仅需利用正确数据的差值训练 OCSVM，之后再将收集到的、已进行对时处理的量测差值输入训练好的支持向量机，通过输出值即可判断是否存在坏数据。

在一分类支持向量机进行训练时，将仅存在正样本的输入空间通过核函数映射到高维空间 H，在高维空间寻找以 $\boldsymbol{\omega}$ 为法向量、ρ 为截距的分类超平面，最大化样本点和原点之间的距离，在原点至超平面一侧的数据被划分为一类，超平面外层的被划分为另一类，如图 1 所示。

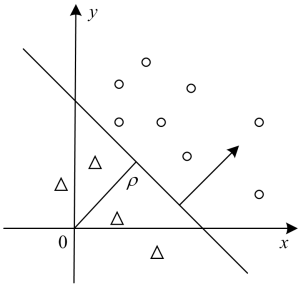


图 1 超平面法示例

Fig. 1 Hyperplane example

为了得到与原点尽可能远的最优超平面，将学习问题转化为优化问题。

$$\begin{aligned}&\min\ \ \frac{1}{2}\|\boldsymbol{\omega}\|^2+\frac{1}{wl}\sum_{i=1}^{l}\xi_i-\rho,\ \ i=1,2,\cdots,l\\&\text{s.t.}\ \ \ \boldsymbol{\omega}\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i)\ge\rho-\xi_i\ ,\ \xi_i\ge0\end{aligned}\tag{1}$$

式中：$\boldsymbol{x}_i$ 是样本的特征向量；l 是训练样本的数量；$w\in(0,1)$，为控制支持向量在训练样本中所占比重的参数，其物理意义是对数据集的不纯净度估计；ξ_i 为松弛变量，避免训练模型过拟合。

利用拉格朗日对偶求解，最小化目标函数。

$$\begin{aligned}&\min\ W(\alpha)=\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&\text{s.t.}\ \ \sum_{i=1}^{l}\alpha_i=1,0\le\alpha_i\le\frac{1}{wl}\end{aligned}\tag{2}$$

式中，$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=<\boldsymbol{\Phi}(\boldsymbol{x}_i)\cdot\boldsymbol{\Phi}(\boldsymbol{x}_j)>$ 为核函数。

最终超平面由少量支持向量决定，其对应拉格朗日因子为 α_i，决策函数为

$$f(x)=\operatorname{sgn}(\sum_i\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x})-\rho)\tag{3}$$

检测时查看输入值与超平面的关系，之后通过输出值判断测试数据是否为离群点。

2.2 基于改进 OCSVM 的不良数据辨识

OCSVM 在使用时有两个重要的参数，分别为不纯净度估计参数 w 与 RBF 核函数的宽度 g，使用粒子群算法对 OCSVM 进行改进[21]，通过迭代寻优确定进行参数的最佳取值，提升 OCSVM 准确度。

粒子群优化(Particle Swarm Optimization, PSO)算法在求解优化问题的解空间内随机初始化规模为 N 的粒子位置，并设置其具有一定的初始速度。根据粒子的位置矢量确定粒子的当前适应度值，通过比较每一代的适应度值确定粒子的当前个体最优值 p 和当前群体最优值 q。根据式(4)与式(5)更新粒子的速度 v 和位置 s。

$$v_i^{t+1}=\omega v_i^t+c_1r_1(p_i-s_i^t)+c_2r_2(q_i-s_i^t)\tag{4}$$

$$s_i^{t+1}=s_i^t+v_i^{t+1}\tag{5}$$

式中：t 为进化代数；c_1、c_2 为进化系数；r_1 和 r_2 为(0,1)之间的随机数；$i=1,2,\cdots,N$；ω为加权系数。粒子不断进行迭代，逐步达到全局最优解。

使用粒子群算法改进一分类支持向量机辨识不良数据的步骤如下。

(1) 将收集到的一定数量的两源正常数据差值 $\Delta x_i(i=1,2,\cdots,n)$ 输入到 OCSVM 中训练。考虑到该向量机仅支持二维及以上的数据，输入特征量设置为 $(\Delta x_i,\Delta x_i)$，设置参数进行训练。

(2) 使用粒子群算法对 OCSVM 的参数进行寻优，每个粒子由两个分量构成，分别代表 OCSVM 参数 w 和高斯核函数参数 g 的位置。设置测试集，以测试集准确率为优化目标函数，达到正确率要求或规定迭代次数时停止，保留识别准确率最高的参数进行后续不良数据辨识。

(3) 将得到的两源量测值进行对时、插值、作差，把要进行检测的样本差值输入训练完成的 OCSVM。

(4) 查看支持向量机的输出，若输出为“1”则认为与之前进行训练的数据为一类，是正常数据；若输出为“–1”，则表明 PMU 和 SCADA 的量测值相差超过了正常范围，为不良数据。

2.3 基于 GSA 的结果修正

使用改进 OCSVM 方法辨识不良数据位置非常迅速，但是当运行状况变化引起数据变化时，经过插值法对应得到连续一段时间的两源量测差值可能会接近向量机超平面的边界，这时正常的 PMU 量测数据可能被误识别为不良数据。为解决这种误识别问题并保证速度，本文引入同节点多时间段数据，对一段时间内的两源差值波动进行考察，利用间隙统计法(Gap Statistic Algorithm, GSA)对识别出的错误数据进行再鉴定，确定不良数据位置。

GSA 算法的本质是将待测数据集与参考数据集对比从而判断出待测集最佳聚类个数的算法[22]。首先构造待测数据集，引入判定为不良数据位置的前 N_g-1 个时间段的两源差值，共 N_g 个数据作为待检测数据，对其进行 k-means 聚类与 *gap* 特征值分析。首先求解其 *gap* 值。

$$gap(k)=E[\ln W_r(k)]-\ln W(k) \tag{6}$$

式中：k 是设定聚类类别个数；$W(k)$ 是待测数据的聚类离散度，为每个聚类中所有数据距离聚类中心的总和，可以反映每类数据间的集中程度；$W_r(k)$ 是 M 个参考数据组的聚类离散度。

$$W(k)=\sum_{i=1}^{k}\frac{1}{|2C_i|}D_i \tag{7}$$

式中：$|C_i|$ 是第 i 组聚类中数据的个数；D_i 是第 i 组聚类中各数据点间的距离。

$$D_i=\sum_{j,j'}\sum(x_j-x_{j'})^2 \tag{8}$$

对数据进行聚类时，由于 k-means 算法的限制，首先单独计算聚类数为 1 的情况，计算数组的平均值，得出数组中每点到均值的距离和，再取自然对数作为 k=1 的特征值 $\ln W(1)$；同样地，计算参考数据集的离散度期望值 $E[\ln W_r(k)]$。聚类数为 2 时，使用 k-means 算法分别对待测序列和参考序列进行聚类，计算原始序列和参考序列的特征值得到 *gap*(2)，按式(9)与式(10)计算标准差 s_2。

$$sd_2=\sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\{\ln W_{r,j}(2)-\mathrm{E}[\ln W_r(2)]\right\}^2} \tag{9}$$

$$s_2=sd_2\sqrt{1+\frac{1}{M}} \tag{10}$$

如 $k=1$ 满足式(11)，说明最佳聚类个数为 1。

$$gap(k)\geq gap(k+1)-s_{k+1} \tag{11}$$

若所得最佳聚类个数为 1，说明这个时间段内两源量测差值分布较平均，没有异常突出的数据存在，故待检测量测值是正常的。如果最佳聚类个数不为 1，意味着检测位置就是不良数据所在的位置。算法流程如图 2 所示。

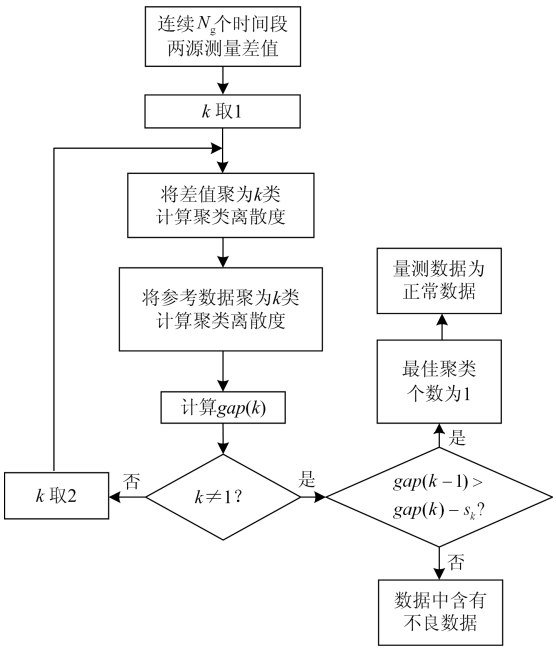


图 2 GSA 算法流程

Fig. 2 Flow diagram of GSA

3 基于线性回归的不良数据源辨识

确定不良数据的位置后，需要判断这一时刻错误数据的来源。SCADA 系统仅能提供稳态的、采样密度较低的电网时间断面信息，而 PMU 能在毫秒级的时间尺度上对电力系统进行同步测量。如图 3 所示，电网稳定运行时 PMU 量测数据在相邻时间内不会产生较大变化且较为集中，检测较为方便，得到的结果也更可靠。

因此，可通过在预设的时间窗内检测 PMU 时间序列的稳定性来辨识不良数据。若此时刻的 PMU 量测被判断为不良数据，则不良数据来源就是 PMU，否则不良数据来源为 SCADA，无需再对

SCADA 量测进行检测。

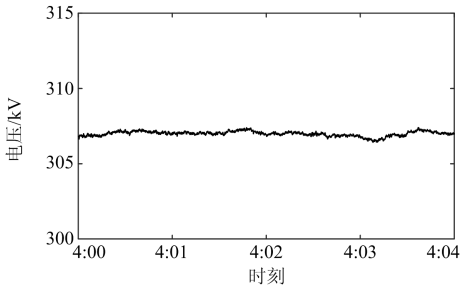


图 3 PMU 量测值

Fig. 3 Measurement value of PMU

得到不良数据位置后，取该时刻的 PMU 源量测真值与相邻的若干个数值，共 N_h 个数值加上时间窗一同进行回归分析。将连续的 N_h 个量测值作为因变量，序号 $i(i=1,2,\cdots,N_h)$ 作为自变量基于最小二乘法得到拟合直线。记为

$$y_d = ai + b \tag{12}$$

式中：i 为时间序列序号；a 和 b 为根据时间窗内量测值确定的参数。按式(13)计算量测值与拟合曲线的平均距离。

$$Dis = \frac{1}{N_h}\sum_{i=1}^{N_h}\mathrm{abs}(y_d(i) - x_d(i)) \tag{13}$$

式中，$x_d(i)$ 为序号为 i 的量测值。

根据 PMU 量测值短时间不会发生突变的特点，可知如果检测数据均为正常数据，那么各点与拟合直线的距离将限制在一个范围内[23]，与所有量测值的距离平均值也相差不大，边界可记为

$$d = k_d \cdot Dis \tag{14}$$

式中，k_d 为检测系数。

若出现不良数据，其与由大部分为正常数据的量测值拟合出的直线距离将变大，与距离平均值差异也会增大。一旦检测点距离超过设置的距离则判定为不良数据。

4　基于两源互校核的不良数据检测框架

通过上述过程，可利用多源数据平台收集的数据互相校核，对量测值进行不良数据检测。该框架的流程如图 4 所示。

将需要检测的、来自同一个节点不同来源的量测值进行对时、插值、作差，输入改进一分类支持向量机进行检测。针对输出结果显示存在较大差值异常点的情况，若判定存在不良数据的位置是经过插值匹配后得到的两源量测差值，此时只需要判断 PMU 数据的正误。由于 SCADA 因采样密度较低没有完全反映数据变化，可能导致其与 PMU 量测值差值变大，对这种情况使用 GSA 算法进行再鉴定，确定是否存在不良数据。若判定错误数据位置的值为直接通过对时匹配的数据，说明两源量测值相差过大，SCADA 或者 PMU 中的一个量测值出现了错误，直接认定其存在不良数据。

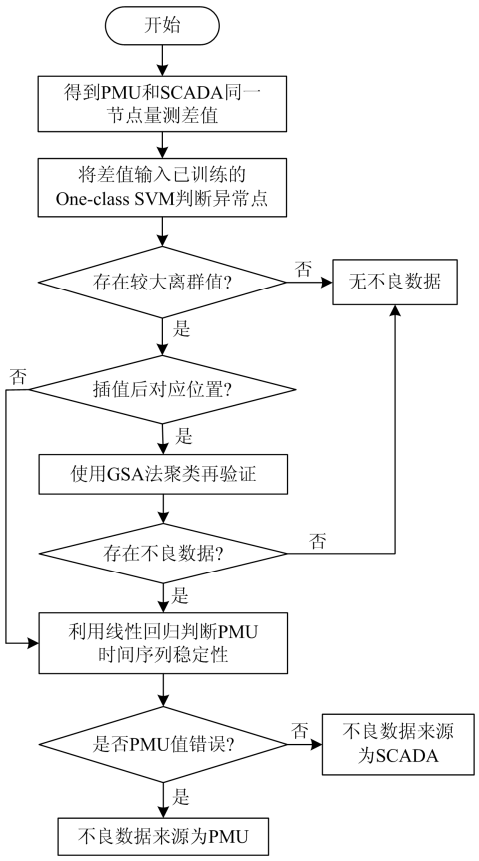


图 4 两源数据互校核流程

Fig. 4 Flow diagram of bad data mutual detection of two sources

针对判定为错误数据所在时刻的 PMU 源量测值进行不良数据检测。如果检测结果为存在错误数据，在这一时刻对应的错误数据源为 PMU；如果判断结果为正常数据，那么这一时刻对应的错误数据源则为 SCADA。设置鉴定辨识方法有效性的判据为

$$P_1 = \frac{n_f}{N_f} \tag{15}$$

$$P_2 = \frac{n_t}{N_t} \tag{16}$$

式中：P_1 为判断正确率；P_2 是误判比例；n_f 为算法正确判断不良数据的个数；n_t 为算法将正常数据误判为不良数据的个数；N_t 和 N_f 分别为测试集中正常数据与不良数据个数。通过这两个指标可以对算

法实际应用效果进行分析和评价。

5　实例验证

5.1 数据预处理

基于实际电网数据对本文提出的 PMU 与 SCADA 互校核不良数据检测方法进行验证。

5.1.1 支持向量机的训练与测试序列

改进一分类支持向量机的训练数据来自某省级电网 D5000 系统，在同一节点同时间段已经过对时处理的 SCADA 与 PMU 正常电压量测值。训练时仅取正常情况下两者差值作为特征值输入 OCSVM 进行训练，共 30 000 组数据作为训练集，从中分出部分数据用于粒子群优化的验证数据。为了避免模型过拟合，基于多次训练的情况，设置验证正确率达到 99%或迭代次数达到 500 时停止[21]，优化训练后 OCSVM 的两个参数 w 与 g 分别为 0.004 9 与 0.487 1。训练集的部分情况如图 5。用于进行不良数据检测的测试集也由同一节点的两源量测值组成。考虑实际情况，SCADA 的数据采集时间间隔为 1 min，而 PMU 量测值则更加密集，为每秒 25 个，对 SCADA 量测值进行拉格朗日插值使两者能够一一匹配。

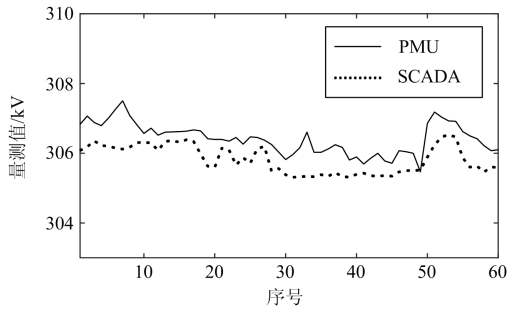


图 5 部分训练集量测值

Fig. 5 Partial measurement value of training set

为了验证文中算法的可行性，在采集数据的各个时段中随机选取 10 段时长 60 min 数据组成测试集，设置比例为 0.3%的不良数据，生成的部分错误量测信息形式如表 2 所示。

随机选取电压量测值产生不良数据的规则如下：

(1) 随机在 PMU 或 SCADA 数据产生不良数据，同一时刻只有一方存在坏数据，不存在两源数据同时出现错误的情况；

(2) 以±20%~±50%的误差在原始数据上产生不良数据，模拟离群值[23]；

(3) 将部分 PMU 量测值置“0”，模拟实际中量测值缺失的情况；

(4) 设置连续若干个 SCADA 量测值不变，模拟数据不刷新使量测值与真实值有偏差的情况。

表 2 不良数据信息(部分)

Table 2 Part of bad measurement data (partial)

序号	错误源	修改后数据/kV	原始数据/kV
1	SCADA	244.310	305.387
2	SCADA	307.309	305.310
3	SCADA	307.309	305.464
4	PMU	244.836	306.045
5	PMU	183.627	306.717
6	PMU	337.577	306.888
7	PMU	367.140	305.950
8	PMU	397.956	306.120
9	PMU	0	305.912
10	PMU	0	306.007
11	PMU	0	305.969

5.1.2 PMU 校核的阈值确定

D5000 中每秒导出 25 个 PMU 数据，本文以秒为单位将故障时刻与前 24 个时刻的量测值一同检测。在对 PMU 量测值进行校核时，需要设置阈值 k_{d} 对时间序列中的离群点进行判断。k_{d} 值控制数据判断的离散水平，值较大时指向更广泛的正常范围，可能导致不良数据的漏检，值较小则仅保留特别密集的量测值为正常数据。鉴于 PMU 的数据波动较小，数值分布相对密集，可以适当地减小 k_{d}。

取时长 15 min 且不含不良数据的 PMU 量测值进行测试，计算时间窗内的各点与其拟合直线的距离，将其与平均距离进行对比，根据最大差距对 k_{d} 的值进行选择。本文考虑一定的裕度设置 $k_{\mathrm{d}}=4$，可以较好地代表原始 PMU 的特征，并且能有效地识别出量测值的离群点，不造成误检。

5.2 不良数据辨识

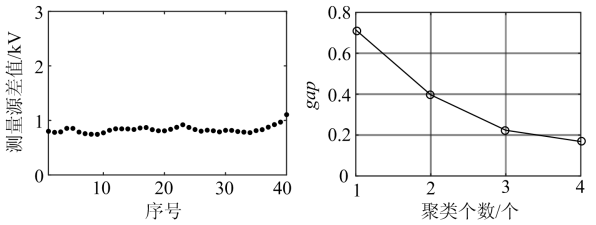
将包含不良数据的两源量测差值输入训练好的支持向量机进行不良数据位置初步检测，再根据情况使用 GSA 进行再判断。

经多次试验，在保证检测灵敏性与可靠性的基础上，引入判定为不良数据位置的前 39 个时间段的两源差值，将 40 个数据作为待检测数据。为了保证参考数据的效果，本文选用 10 组最大值和最小值与待测数据相同的均匀分布作为参考数据集计算[24]。

GSA 算法的修正作用可用图 6 进行解释。对图 6(a)进行分析，序号为 40 的数据实际为正常数据，但由于接近超球体边界，被 OCSVM 识别为不良数据，与前 39 时刻的差值一起进行 GSA 计算，结果如图 6(b)所示，满足 $gap(1)\geq gap(2)-s_2$，最佳聚类个数为 1，判断该数据为正常数据，避免了误判。

对图 7(a)情况进行分析，序号为 40 的数据被一

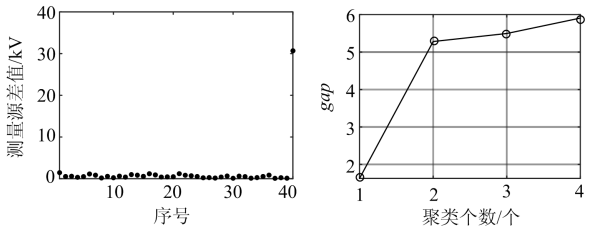
分类支持向量机识别为不良数据，实际上也确实是预处理中制造的坏数据。经过 GSA 计算，结果如图 7(b)所示，最佳聚类个数不为 1，再次确定了不良数据位置。



(a) 量测差值　(b) *gap*值

图 6 两源量测特征

Fig. 6 Measurement feature between two sources



(a) 量测差值　(b) *gap*值

图 7 两源量测特征

Fig. 7 Measurement feature between two sources

单独使用 OCSVM 和单独使用 GSA[22]对测试集进行不良数据检测，所得结果与本文所用方法的对比如表 3。

表 3 不同算法结果对比

Table 3 Result comparison between different algorithms

方法	P_1/%	P_2/%	检测 1 min 数据平均耗时/s
本文方法	100	0.01	0.049 1
GSA	97	0.135	53.956 0
改进 OCSVM	100	0.165	0.013 3

对结果进行分析，可知三种方法均能将错误数据找出。但单独使用 GSA 算法时存在一定的误检，也因为每次检测都需要生成参考序列和特征值计算而耗时最长。单独使用 OCSVM 时，程序运行时间最短，但存在一定的误检率，很大程度上导致 SCADA 数据被判断为不良数据。本文将粒子群改进的 OCSVM 和 GSA 结合使用，综合两种算法的优势，快速辨识不良数据的同时能保证准确性，具有优越性。

对存在不良数据的位置使用回归分析的方法判断该时刻的 PMU 数据是否存在问题，效果如图 8。经过这个步骤，可将人为置入不良数据的数据源与时刻正确找出，完成利用 SCADA 与 PMU 互校核辨识两源不良数据的全过程。

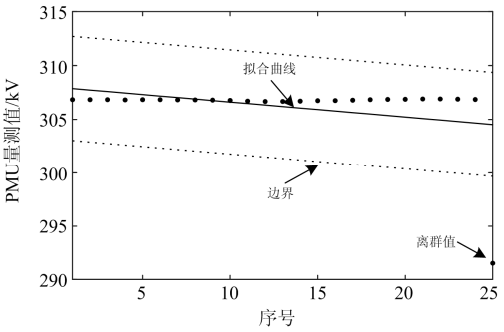


图 8 PMU 量测值辨识

Fig. 8 Identification of PMU measurement value

不良数据辨识算法要经过多个步骤得到故障位置与故障数据源，其辨识效率会受不良数据比例的影响，同时算法的应用对数据的对时准确度也有一定要求。无对时误差下改变测试集的不良数据比例与在 0.3%不良数据比例下改变对时误差[25]的仿真结果见表 4 与表 5。

表 4 不同不良数据比例对比

Table 4 Comparison between different percentages of bad data

错误数据比例	P_1/%	P_2/%	检测 1 min 数据平均耗时/s
0.1%	100	0.01	0.051 7
0.3%	100	0.01	0.055 2
1%	100	0.011	0.068 3
5%	99.05	0.012	0.156 4

表 5 不同对时误差对比

Table 5 Comparison between different timing errors

对时误差/s	P_1/%	P_2/%	检测 1 min 数据平均耗时/s
5	100	0.011 1	0.782
10	100	0.044 6	1.216
20	100	0.050 1	1.346
30	99.02	0.078 0	1.749

表 4、表 5 的结果体现了算法的整体性能，随着不良数据比例的增大与数据对时误差的增加，进入 GSA 计算或回归分析的次数会增加，均可能降低不良数据辨识的正确率或延长检测时间。通过对实际数据的分析发现，不良数据的比例一般不会超过 0.3%，因此本算法应用时仍能达到快速与准确的要求。对时误差在 20 s 内时，识别的正确率也比较高，但若是继续增大则会引起漏判。

5.3 单源校核与多源校核效果对比

现有的利用数据挖掘方法对不良数据辨识时都是利用单一来源的数据，寻找其数据规律，将不符合规律的异常值认定为不良数据。以同样规则给某 10 日的正常 SCADA 电压量测值制造不良数据，将

其作为检测对象。使用本文两源校核算法与仅利用单一来源、以聚类为基础的算法[26]对 SCADA 数据进行不良数据辨识，结果对比如表 6。

表 6 SCADA 检测结果对比

Table 6 Result comparison of SCADA detection

方法	P_1/%	P_2/%	总耗时/s
本文两源互校核	100	0	0.013 6
单源聚类算法	85.71	1.95	1.01×10^{-3}

本文方法虽然在时间上不如聚类算法，但也能够满足检测迅速的要求。同时本文算法准确性更高，也不存在错判的数据。这是因为单源数据提取出来的特征虽具有一般性，但是当电力系统在短时间内负荷变化较大导致电压出现波动时，正常数据也可能呈现与一般规律不一样的数值，同时数据不刷新也可能不会被检测出来。而使用同一节点 PMU 数据作为参考时，同一点的两个量测值应呈现一样的变化，出现差异即可视为异常，检测结果更可靠。在 PMU 校核方面，基于单源数据使用 DBSCAN 算法[23]与模式识别方法[14]对 10 段时长为 1 h、不良数据比例为 0.3%的 PMU 数据进行检测，与本文方法对比的结果如表 7。

表 7 PMU 检测结果对比

Table 7 Result comparison of PMU detection

方法	P_1/%	P_2/%	检测 1 min 数据平均耗时/s
本文方法	100	0	0.058
DBSCAN	100	0	1.137
模式识别	96.37	0	0.032

互校核方法通过一分类支持向量机保证检测速度的同时，两源互校核的检测更具有可靠性，检测 PMU 不良数据的准确率也比模式识别方法更高。

6　结论

本文应用电网中数据平台获取的海量多源数据，提出了一种基于 PMU 和 SCADA 单节点互校核的前端数据辨识框架，利用 SCADA 和 PMU 在同一节点、同一段时间的量测值相互校核，结论如下。

(1) 将基于粒子群优化的改进一分类支持向量机与 GSA 算法结合，利用两源量测差值快速准确地确定异常点时刻，再对不良数据所在时刻的 PMU 量测值进行检测最终确定前端不良数据源。本文方法在实际电网节点量测数据中的应用验证了其有效性，改进后的一分类支持向量机辨识准确度更高，与 GSA 的结合应用相比单一算法更具优越性。

(2) 本文算法计算量小，算例证明了该方法在保证检测速度的同时避免了误检的情况，能在一定对时误差下准确辨识出两个数据源的不良数据，同时比单一源数据校正更具可靠性。

(3) 本文针对同一个节点两源量测值进行互校核，对于如何利用一个节点设置的 PMU 检测其周围节点的 SCADA 不良数据，未来可以进行更深入的研究。

参考文献

[1] 林静怀. 基于大数据平台的电网运行指标统一管控方案[J]. 电力系统保护与控制, 2018, 46(4): 165-170.
LIN Jinghuai. A unified scheme of grid operation index control based on big data platform[J]. Power System Protection and Control, 2018, 46(4): 165-170.

[2] 丁明, 李晓静, 张晶晶. 面向 SCADA 的网络攻击对电力系统可靠性的影响[J]. 电力系统保护与控制, 2018, 46(11): 37-45.
DING Ming, LI Xiaojing, ZHANG Jingjing. Effect of SCADA-oriented cyber attack on power system reliability[J]. Power System Protection and Control, 2018, 46(11): 37-45.

[3] D'ANTONA G, PERFETTO L. Bad data detection and identification in power system state estimation with network parameters uncertainty[C] // International Conference on Knowledge-Based Engineering and Innovation, November 5-6, 2015, Tehran, Iran: 26-31.

[4] 卢志刚, 程慧琳, 冯磊, 等. 基于证据融合理论的多不良数据辨识[J]. 电网技术, 2012, 36(1): 123-128.
LU Zhigang, CHENG Huilin, FENG Lei, et al. Multi bad data identification based on evidence fusion theory[J]. Power System Technology, 2012, 36(1): 123-128.

[5] 孔祥玉, 胡启安, 董旭柱, 等. 引入改进模糊 C 均值聚类的负荷数据辨识及修复方法[J]. 电力系统自动化, 2017, 41(9): 90-95.
KONG Xiangyu, HU Qi'an, DONG Xuzhu, et al. Load data identification and correction method with improved fuzzy C-means clustering algorithm[J]. Automation of Electric Power Systems, 2017, 41(9): 90-95.

[6] 吴越赢. 基于数据挖掘的电力系统不良数据检测与辨识算法研究[D]. 南京: 南京理工大学, 2017.
WU Yueying. Data mining based algorithm for detecting and identifying bad data in power system[D]. Nanjing: Nanjing University of Science and Technology, 2017.

[7] KIRANMAI S A, LAXMI A J. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy[J]. Protection and Control of Modern Power Systems, 2018, (3): 303-314. DOI: 10.1186/s41601-018-0103-3

[8] 李珏, 刘灏, 田建南, 等. 适用于主动配电网 PMU 的数据传输协议与通信性能分析[J]. 电力科学与技术学报, 2019, 34(2): 3-10.
LI Jue, LIU Hao, TIAN Jiannan, et al. Communication

protocol and performance analysis for the PMU of distribution network[J]. Journal of Electric Power Science and Technology, 2019, 34(2): 3-10.

[9] GOPAKUMAR P, MALLIKAJUNA B, JAYA BHARATA REDDY M, et al. Remote monitoring system for real time detection and classification of transmission line faults in a power grid using PMU measurements[J]. Protection and Control of Modern Power Systems, 2018, (2): 159-168. DOI: 10.1186/s41601-018-0089-x.

[10] SHI D, TYLAVSKY D J, LOGIC N. An adaptive method for detection and correction of errors in PMU measurements[J]. IEEE Transactions on Smart Grid, 2012, 3(4): 1575-1583.

[11] THOMAS P, SKARIAH E N, VARGHESE S T. Detection of bad data in phasor measurement units using distributed approach[C] // Innovations in Power and Advanced Computing Technologies, April 21-22, 2017, Vellore, India: 1-8.

[12] YASINZADEH M, AKHBARI M. Detection of PMU spoofing in power grid based on phasor measurement analysis[J]. IET Generation, Transmission & Distribution, 2018, 12(9): 1980-1987.

[13] GOU B, KAVASSERI R G. Unified PMU placement for observability and bad data detection in state estimation[J]. IEEE Transactions on Power Systems, 2014, 29(6): 2573-2580.

[14] 陈亦平, 陈伟彪, 姚伟, 等. WAMS 错误数据的快速辨识及恢复方法[J]. 电力自动化设备, 2016, 36(12): 95-101.
CHEN Yiping, CHEN Weibiao, YAO Wei, et al. Rapid identification and recovery of wrong WAMS data[J]. Electric Power Automation Equipment, 2016, 36(12): 95-101.

[15] WU M, XIE L. Online identification of bad synchrophasor measurements via spatio-temporal correlations[C] // Power Systems Computation Conference, June 20-24, 2016, Genoa, Italy: 1-7.

[16] 许勇. 基于 PMU/SCADA 混合量测状态估计及不良数据检测方法[J]. 四川电力技术, 2015, 38(4): 51-55.
XU Yong. Hybrid state estimation and bad data detection based on PMU/SCADA measurement[J]. Sichuan Electric Power Technology, 2015, 38(4): 51-55.

[17] 刘科研, 盛万兴, 张东霞, 等. 智能配电网大数据应用需求和场景分析研究[J]. 中国电机工程学报, 2015, 35(2): 287-293.
LIU Keyan, SHENG Wanxing, ZHANG Dongxia, et al. Big data application requirements and scenario analysis in smart distribution network[J]. Proceedings of the CSEE, 2015, 35(2): 287-293.

[18] 刘科研, 张剑, 陶顺, 等. 基于多源多时空信息的配电网 SCADA 系统电压数据质量检测与评估方法[J]. 电网技术, 2015, 39(11): 3169-3175.
LIU Keyan, ZHANG Jian, TAO Shun, et al. Detection and evaluation of SCADA voltage data quality in distribution network based on multi temporal and spatial information of multi data sources[J]. Power System Technology, 2015, 39(11): 3169-3175.

[19] 丁军策, 蔡泽祥, 王克英. 基于广域测量系统的状态估计研究综述[J]. 电力系统自动化, 2006, 30(7): 98-103.
DING Junce, CAI Zexiang, WANG Keying. An overview of state estimation based on wide-area measurement system[J]. Automation of Electric Power Systems, 2006, 30(7): 98-103.

[20] SCHOLKOPF B, PLATT J C, SHAWE-TAYLOR J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.

[21] 李琳. 基于 OCSVM 的工业控制系统入侵检测算法研究[D]. 沈阳: 沈阳理工大学, 2016.
LI Lin. Research on intrusion detection algorithm of industrial control systems based on OCSVM[D]. Shenyang: Shenyang Ligong University, 2016.

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

[23] ZHOU M, WANG Y, SRIVASTAVA A K, et al. Ensemble- based algorithm for synchrophasor data anomaly detection[J]. IEEE Transactions on Smart Grid, 2019, 10(3): 2979-2988.

[24] KOGLIN H J, NEISIUS T, BEISSLER G, et al. Bad data detection and identification[J]. International Journal of Electrical Power & Energy Systems, 1990, 12(2): 94-103.

[25] 郭骏. 基于大电网混合量测的状态估计算法应用研究[D]. 北京: 华北电力大学, 2013.
GUO Jun. Research and application of hybrid measurement based state estimation algorithm in large power grid[D]. Beijing: North China Electric Power University, 2013.

[26] 刘辉舟, 周开乐, 胡小建. 基于模糊负荷聚类的不良负荷数据辨识与修正[J]. 中国电力, 2013, 46(10): 29-34.
LIU Huizhou, ZHOU Kaile, HU Xiaojian. Bad data identification and correction based on load clustering by FCM algorithm[J]. Electric Power, 2013, 46(10): 29-34.

收稿日期：2019-05-30；　修回日期：2019-12-18

作者简介：

刘雯静(1996—)，女，硕士研究生，研究方向为机器学习在电力系统中的应用；E-mail: liuwj1996@126.com

杨　军(1977—)，男，通信作者，博士，教授，博士生导师，研究方向为基于数据驱动的电网安全分析、电动汽车与电网互动等。E-mail; jyang@whu.edu.cn

（编辑　许　威）