DOI:10.20033/j.1003-7241.(2025)01-0071-05.

基于边缘计算的企业海量云端动态数据清洗方法研究

徐晓冰[1]，李奇越[1]，陈 艺[2]，赵 龙[2]，秦 琪[2]，汪 玉[2]

（1.合肥工业大学 电气与自动化工程学院，安徽 合肥 230009；
2.国网安徽省电力有限公司电力科学研究院，安徽 合肥 230000）

摘 要：原有清洗方法对缺失数据的筛选标准过于模糊，无法完成未来企业的数据预测。因此，提出基于边缘计算的企业海量云端动态数据清洗方法。在数据采集模块，对动态数据进行预处理。基于边缘计算构建数据筛选模型，将企业数据进行分类。计算企业数据的传输延时和耗能。通过时间序列理论依据，保证企业动态原始数据的真实性，从全局和局部两个方面划分云端的动态数据特征，完成动态数据清洗。实现基于边缘计算的清洗方法设计。实验结果表明：所提方法的准确度在99.8%以上，数据预测结果精度高，更具有应用价值。

关键词：边缘计算；企业数据；云端数据；数据清洗
中图分类号：TP391 文献标识码：A 文章编号：1003-7241(2025)01-0071-05

Research on Dynamic Data Cleaning Method of Enterprise Massive Cloud Based on Edge Computing

XU Xiao-bing[1], LI Qi-yue[1], CHEN Yi[2], ZHAO Long[2], QIN Qi[2], WANG Yu[2]

(1. School of Electrical and Automation Engineering, Hefei University of Technology, Hefei 230009 China;
2. State Grid Anhui Electric Power Co., Ltd., Electric Power Research Institute, Hefei 230000 China)

Abstract: The original cleaning method's screening criteria for missing data are too vague to complete the data prediction of future enterprises. Therefore, an edge computing-based dynamic data cleaning method for enterprise massive cloud is proposed. In the data acquisition module, the dynamic data is preprocessed. It builds a data screening model based on edge computing to classify enterprise data, calculates the transmission delay and energy consumption of enterprise data. Based on the theoretical basis of time series, the authenticity of the dynamic raw data of the enterprise is guaranteed, and the dynamic data characteristics of the cloud are divided into global and local aspects to complete the dynamic data cleaning. It realizes the design of cleaning method based on edge computing. The experimental results show that the accuracy of this method is above 99.8%. It shows that the data prediction results of the method in this paper have high accuracy and have more application value.

Keywords: edge calculation; enterprise data; cloud data; data cleaning

0 引言

随着社会进入“大数据”时代，云计算技术的出现正在逐步改善数据处理能力，该技术以运行成本低和动态扩展性能，得到了广泛的应用和研究[1]。在云计算架构中，数据可通过网络结构进行传输，保存在云端处理器中，使其能够动态和灵活地进行调取和处理。与此同时云计算所需要的数据资源，基本上要通过采集设备才能得到，不同的终端设备，将采集到的数据传送至云端平台中，在集中处理方法下得到想要的数据结果[2]。

在大数据被广泛应用的背景下，以多种算法进行数据分析，可以对云端数据进行处理。例如：文献[3]基于边缘检测与方差变点的风功率数据清洗方法。首先进行数据预清洗，以识别负值点，接着基于边缘检测识别曲线主体，以清洗分散型异常数据，然后通过方差变点分区间获得风速功率点中的方差突变点清洗堆积型异常数据，最后得到分类清洗后数据；文献[4]基于回归模型的动态精细识别算法，在剔除潜在野值的基础上利用前后两段数据的回归值作为参考值，再结合采集参数变化速率限制给出野值判决，基于回归模型数据清洗的处理流程，采用野值粗识别、精细识别、回归估计的步骤得到清洗数据。但是上述方法都存在与增长速度无法进行精准匹配、难以及时地对云端数据进行处理的问题。因此提出基于边缘计算的企业海量云端动态数据清洗方法，为解决数据处理问题提供理论支持。

*基金项目：国家电网公司总部科技项目（5108-202299268A-1-0-ZB）
收稿日期：2023-09-28

1 预处理企业海量云端动态数据

在采集模块对企业数据进行接收和利用时，一方面要对数据进行整理，以提高数据的读取和利用性能，另一方面要对数据进行预处理，避免后续的数据丢失问题，保证企业数据的完整性。具体构成内容如图 1 所示。图 1 中，在数据预处理过程中，包含数据集合和变换等内容，对数据集成主要是，将处理过的企业数据进行整合，而数据变换是直接对原始数据进行处理[5]。

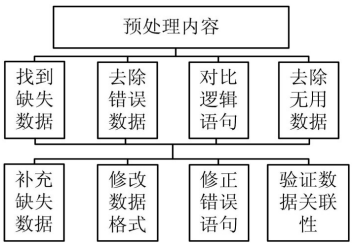


图 1 数据预处理构成

引入数据预测方程，将企业云端动态数据看做一个集合，包含 Q 组实时数据，则数据的状态向量表示为：

$$W=\left[R_1,R_2,\cdots,R_Q,T_1,T_2,T_3,T_4,\cdots,T_Q\right] \tag{1}$$

式中：向量集合为 W。现有向量为 $R_1,R_2,\cdots,R_Q$。丢失向量为 $T_1,T_2,T_3,T_4,\cdots,T_Q$。

在数据传输过程中，引入预测增益矩阵 A_{IP}。判断两种类型向量的预测变量，对应方程为：

$$\begin{aligned}
U_I&=R_I\sum_{P\in Q_I}R_P\left[A_{IP}\cos T_{IP}+S_{IP}\sin T_{IP}\right]\\
O_I&=R_I\sum_{P\in Q_I}R_P\left[A_{IP}\cos T_{IP}+S_{IP}\sin T_{IP}\right]\\
U_{IP}&=R_I^2\left[A_{IP}+S_{IP}\right]-R_IR_P\left[A_{IP}\cos T_{IP}+S_{IP}\sin T_{IP}\right]\\
O_{IP}&=-R_I^2\left[A_{IP}+S_{IP}\right]-R_IR_P\left[A_{IP}\cos T_{IP}-S_{IP}\sin T_{IP}\right]
\end{aligned} \tag{2}$$

式中：现有向量的变量方程为 U_I，对应的预测方程为 U_{IP}。现有数据为 R_I，正在传输的现有数据向量为 R_P。丢失向量的变量方程为 O_I，对应的预测方程为 Q_{IP}。预测指数为 P。预测丢失的数据为 T_{IP}。预测误差为 T_{IP}。

通过对企业数据的预测，对上传数据过程中出现的异常数据，划分成三种模式，并匹配出识别类型，见表 1 所示。

表 1 描述预处理数据类型

类型	对象	标准
数据统计	云端内所有数据	1.检测异常数据 2.正态分布动态数据 3.匹配数据传输通路
数据距离	不同数据之间的距离	1.相邻正常数据之间的距离 2.不同异常数据之间的距离 3.正常与异常数据之间的距离
数据偏差	数据特征	1.异常数据类型 2.偏离设定特征

根据表 1 中内容所示，在数据处理过程中，不仅会包含缺失数据，还需要对数据进行填补，已完成不完整信息的备份和分析[6-7]。为有效获取企业云端中的海量数据，连接互联网边缘设备，作为采集模块，结合网络爬虫技术，在一定的算法基础上，对网页代码中的数据进行预处理。

2 基于边缘计算构建数据筛选模型

当企业需要对数据进行操作时，需要考虑到数据本身的特征，引入边缘计算方式，构建海量动态数据的筛选模型，对企业数据进行分类。

设置数据任务为 D_F，在企业服务器中定义数据传输延时，总延时为 G_K，由上传延时 G_H，和处理延时 G_J 组成。表达式如下：

$$\begin{aligned}
G_H&=\frac{L_F}{\varepsilon_H}\\
G_J&=\frac{Z_F}{\phi_H}\\
G_K&=G_H+G_J=\frac{L_F}{\varepsilon_H}+\frac{Z_F}{\phi_H}
\end{aligned} \tag{3}$$

式中：任务数据量为 L_F。企业数据的传输效率为 ε_H[8]。服务器的计算能力为 ϕ_H。服务器内数据量为 Z_F。

以能耗损失为基础，在不同阶段内计算企业动态数据的消耗能量，公式为：

$$\begin{aligned}
X_H&=C_HG_H\\
V_J&=B_HG_J\\
N_K&=X_H+V_J=C_HG_H+B_HG_J
\end{aligned} \tag{4}$$

式中：传输能耗为 X_H。处理能耗为 V_J。云端设备到服务器的单位能耗为 C_H。处理任务中的单位能耗为 B_H。总能耗为 N_K。

比较服务器接收数据的过程，在边缘计算应用过程中，能够将数据的储量，接近于无限的传输状态[9]。以式(3)为代表，利用边缘计算方式，对服务器云端的数据执行能力，进行分析，如下：

$$\begin{aligned}
G_H^{\,1}&=\frac{L_F}{\varepsilon_{H1}}\\
G_J^{\,1}&=\frac{Z_F}{\phi_Z}\\
G_K^{\,1}&=G_H+G_H^{\,1}+G_J^{\,1}
\end{aligned} \tag{5}$$

式中：数据到云端的传输速率为 ε_{H1}。服务器到云端的延时为 $G_H^{\,1}$。到云端的处理能力为 ϕ_Z。云端的处理延时为 $G_J^{\,1}$。到云端的总延时为 $G_K^{\,1}$。

数据传输需要考虑“服务器－云端”“终端－云端”两个阶段，和式(4)的耗能标准为基础，计算数据到云端的耗能，公式为：

$$\begin{aligned}
X_H^{\,1}&=C_H^{\,1}G_H^{\,1}+G_H\\
V_J^{\,1}&=B_H^{\,1}G_J^{\,1}\\
N_K^{\,1}&=X_H^{\,1}+V_J^{\,1}
\end{aligned} \tag{6}$$

式中：服务器到云端的单位能耗为 $C_H^{\,1}$。数据任务到云端

能耗为X_H^1。云端处理阶段的能耗为V_J^1。处理阶段单位能耗为B_H^1。云端处理总能耗为N_K^1。

通过上述不同类型的动态数据的传输类型，在完成数据延时计算和耗能计算的基础上，在目标状态的基础上对数据进行分类。按照提取特征将数据分属为三种类别。统计其相关的平均高度，最后将数据分类结果导出。

建立数据筛选的具体流程，如图2。

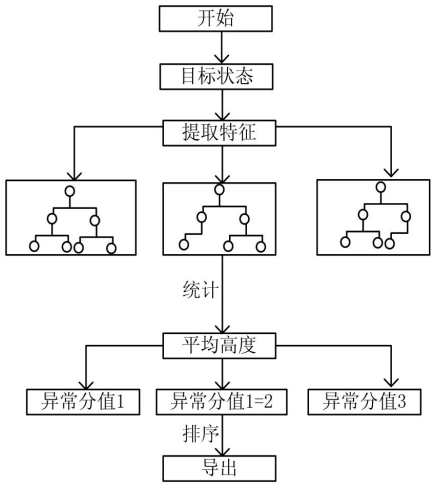


图2 筛选模型的数据分类流程

根据图2中内容所示，当筛选模型构建完成后，对企业内的数据进行清洗。而在此之前，需要对数据进行划分，可以通过动态数据特征类型，组建三种序列树，以不同的数据高度，对应数据特征值[10]。并在时间序列计算方法下，根据特征完成动态数据清洗。

3 时间序列划分特征清洗动态数据

考虑数据的传输时间和效率，引入时序序列理论作为判断标准，设置序列为q，其中q等于$[q_1],[q_2],\cdots,[q_n]$，n为序列量[11]。其中$q[w]$表示第w个数据点的数值。每个数值$q[w]$均含有时间戳$e[w]$，简化写作q_w和e_w。

将清洗过程分为全局和局部两个部分，形式化定义全局清洗条件，具体清洗条件如下：

$$\begin{aligned}&\min\sum_{w=1}^{n}\left|q_w-q'_w\right|\\&s.t.\frac{q'_r-q'_w}{e_r-e_w}\le t_{\max},\begin{cases}e_w<e_r<e_r+u\\1\le w\le n,1\le r\le n\end{cases}\\&s.t.\frac{q'_r-q'_w}{e_r-e_w}\ge t_{\min},\begin{cases}e_w<e_r<e_r+u\\1\le w\le n,1\le r\le n\end{cases}\end{aligned}\tag{7}$$

式中：u表示时间序列窗口大小[12]；r表示标准化清洗数值；$t_{\min}$表示最小速度，$t_{\max}$表示最大速度。t表示窗口速度约束，$t=(t_{\max},t_{\min})$；q_w和q_r分别表示一个窗口内的两个数据点的数值，e_r和e_w分别表示对应的清洗时间，q'_w和q'_r分别表示清洗后的数据值。

其中q'_w可以作为求解问题时的未知变量。将其转化为线性规划问题，则：

$$\begin{aligned}&p_i=\frac{\left|q'_w-q_w\right|+\left(q'_w-q_w\right)}{2}\\&a_i=\frac{\left|q'_w-q_w\right|-\left(q'_w-q_w\right)}{2}\\&\left|q'_w-q_w\right|=p_i+a_i\\&q'_w-q_w=p_i-a_i\end{aligned}\tag{8}$$

式中：求解问题时的变量为p_i和a_i。

则数据的约束条件如下所示：

$$\begin{aligned}&\min\sum_{w=1}^{n}p_i+a_i\\&s.t.\frac{p_r-a_r+p_w-a_w-q_w+q_r}{e_r-e_w}\le t_{\max},\begin{cases}e_w<e_r<e_w+u\\1\le w\le n,1\le r\le n\\e_w<e_r<e_w+u\end{cases}\\&s.t.\frac{p_r-a_r+p_w-a_w-q_w+q_r}{e_r-e_w}\ge t_{\min},\begin{cases}e_w<e_r<e_r+u\\1\le w\le n,1\le r\le n\end{cases}\\&s.t.p_w\ge 0,a_w\ge 0,1\le w\le n\end{aligned}\tag{9}$$

式中：p_r和a_r表示两个数据点的数值未知变量。

在此基础上，对局部云端数据进行清洗，和全部清洗条件较为相似，将q_d作为起始窗口处的每一组数据点，任意点与数据点q_i的需要速度约束条件，则：

$$\begin{aligned}&\min\sum_{w=1}^{n}\left|q_w-q'_w\right|\\&s.t.\frac{q'_d-q'_w}{e_d-e_w}\le t_{\max},\begin{cases}e_d<e_w<e_d+u\\1\le w\le n\end{cases}\\&s.t.\frac{q'_d-q'_w}{e_d-e_w}\ge t_{\min},\begin{cases}e_d<e_w<e_d+u\\1\le w\le n\end{cases}\end{aligned}\tag{10}$$

式中：q'_w仍为求解问题时的未知变量。而数据清洗不仅要完成丢失数据的召回，还要完成错误数据的更改，对清洗后的数据进行评价，基本上以清洗质量为标准。

将全部清洗和局部清洗的结果，按照真正例和真反例，以及假正例和假反例为四个类型，设定企业数据的准确率和召回率。两个标准为一对相互对立的评价指标，当一个指标较高时，另一组指标就越小。至此，通过边缘计算方式，完成企业海量动态数据清洗方法设计。

4 实验测试分析

通过边缘计算方法，重新设计了企业云端海量动态数据的清洗方法，为验证本文方法的有效性，采用对比实验的方式进行论证。以两组传统方法作为对照，比较不同缺失情况下，数据清洗的准确效果。

4.1 选择测试样本

选择电力公司为测试对象，以低度值较为接近的三组电力线路作为样本采集点，反映不同线路负荷下，数据的变化量。此次选择100 kV、200 kV以及300 kV以及

350 kV四组线路，按照每隔30 min为一次采集节点，对正向有功电量进行数据采集，情况见表2所示。

表2 电力数据部分样本/kV

采集点	电压等级			
	100 kV	200 kV	300 kV	350 kV
0:00	51.75	60	20.5	18.4
0:30	51.8	60.2	21	18.5
1:00	51.85	60.3	21.5	18.6
1:30	51.9	60.4	22	18.7
2:00	51.95	60.5	22.5	18.8
2:30	52	60.6	23	18.9
3:00	52.05	60.7	23.5	19
3:30	52.1	60.8	24	19.1
15:30	55.1	66.1	29	28
16:00	55.12	66.2	30	28.1
16:30	55.2	66.3	30.5	28.2
23:00	56.3	68	34.5	30.4
23:30	56.35	68.1	35	30.5

对表中数据进行说明，电压线在不同等级内，单位时间的电力数据变化量，会产生不同的增值，此次正向有功电量，属于电力数据的一个属性域。将所有数据导入至MATLAB测试平台中，分别连接三组清洗方法进行测试。

4.2 单一数据缺失测试

第一阶段测试中，分别将四组路线采集数据中，18点的采集点数据进行空白处理，将其作为缺失数据，以选择的三组清洗方法，分别对缺失的采集点数据进行预测填充。

调取全部采集样本，此时这四点的真实值，在100 kV和100 kV、200 kV以及300 kV线路上，分别为55.1和67以及30.8和30。统计预测结果，见图3所示。

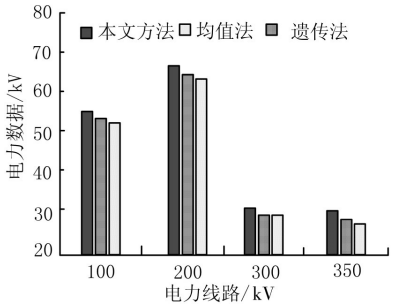


图3 预测结果

图3中，本文方法得到的数据值，基本与真实值相近，而传统方法的预测数据，均低于真实值。以准确度为衡量标准，计算公式如下：

$$c=\frac{v}{b}\times 100\% \tag{11}$$

式中：v为预测值。b为真实值。准确度为c。分别将图中数据带入到式(11)中，准确度见图4。

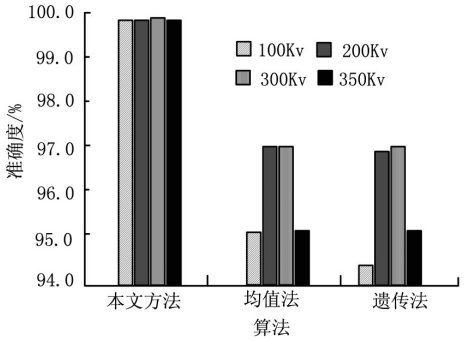
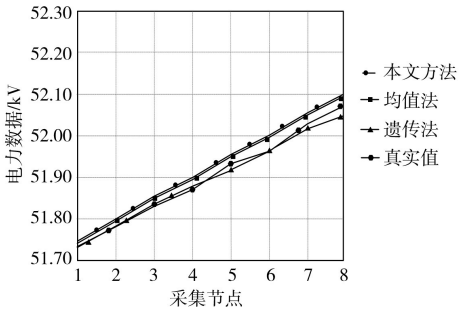
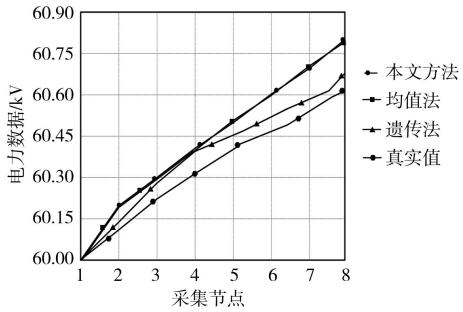


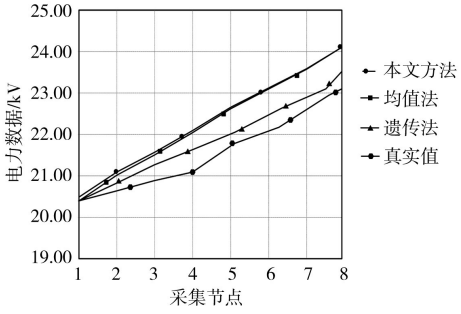
图4 准确度统计



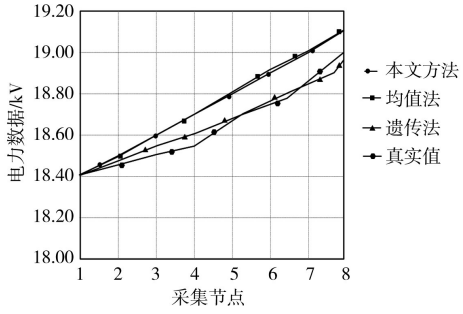
(a) 100 kV



(b) 200 kV



(c) 300 kV



(d) 350 kV

图5 预测结果

图4中，本文方法的准确度均值99.8%以上。均值清洗法和遗传网络神经清洗方法的准确率，分别为96.0%和95.5%，说明本文方法更加有效。

4.3 连续采集节点缺失数据测试

对单一缺失数据的清洗预测较为简单，在此基础上，完成第二阶段测试。从三组线路中分别取8个采集点，作为电力数据缺失值。仍以表2为参数值，即采集点为0:00—3:30的数据，分别采用三组清洗算法进行预测，结果见图5所示。

根据图5中所示。本文方法在三条线路的连续采集点，预测过程中，不会受到电压等级和负荷影响，能够与真实值保持一致。而均值清洗法和遗传网络神经清洗方法，预测的数据值与真实存在一定差距，在连续清洗过程中，能够更加直观地进行观察。综合结果表明，本文方法更加有效。

5 结束语

伴随数据呈现的价值越来越高，对数据的挖掘和清洗也被众多企业所重视，在未来的发展中，数据会呈现多样性和实时性的特点。在此背景下，原有清洗只对原有数据进行清洗，此次利用边缘计算方式，对数据点的重复使用性能，作出了一定的改变，并在实验论证下取得一定的成果。但由于数据中除了缺失数据外，还会包含异常数据，本次由于时间受限，没有进行异常数据的研究。后续研究中会采用不同的算法，对异常数据进行分析，设计出更加适用于企业使用的数据清洗方法。

参考文献：

[1] 卢峰，吴朝文，陈小龙，张柯柯，桂宁.基于云计算的电力能源大数据清洗模型构建[J].自动化仪表，2022，43(1)：72—76.

[2] 谢小川，周绍军，黎力，等.基于时序关联智慧城市边缘数据异常检测算法[J].计算机工程与设计，2022，43(7)：1879—1889.

[3] 苏荣，张斌，沈晨，陈俊生.基于边缘检测与方差变点的风功率数据清洗方法[J].广东电力，2021，34(5)：48—56.

[4] 李洪烈，夏栋，王倩.基于回归模型的采集数据清洗技术[J].电光与控制，2022，29(4)：117—120.

[5] 李海林，张丽萍.时间序列数据挖掘中的聚类研究综述[J].电子科技大学学报，2022，51(3)：416—424.

[6] 王星，于炯，杜旭升，张姗姗，杨少智.边缘计算系统中延时敏感型任务调度策略[J].计算机工程与设计，2022，43(7)：1949—1954.

[7] 卢剑伟.基于神经网络模型的COPD诊断信息敏感数据自动挖掘方法[J].自动化技术与应用，2021，40(11)：80—85.

[8] 刘漳辉，郑鸿强，张建山，等.多无人机使能移动边缘计算系统中的计算卸载与部署优化[J].计算机科学，2022，49(S1)：619—627.

[9] 高菲，宋韶旭，王建民.多区间速度约束下的时序数据清洗方法[J].软件学报，2021，32(3)：689—711.

[10] 鲁树武，伍小龙，郑江，等.基于动态融合LOF的城市污水处理过程数据清洗方法[J].控制与决策，2022，37(5)：1231—1240.

[11] 周建，石浩，秦铭泽，等.基于边缘智能技术的台区线损异常自动检测研究[J].自动化技术与应用，2024，43(8)：46—49，59.

[12] 左建勇，冯富人，丁景贤.基于Super smoother和3σ原理的列车动态测试趋势性异常数据清洗方法与分析[J].仪器仪表学报，2020，41(10)：65—73.

作者简介：徐晓冰(1967—)，男，博士，副教授，研究方向：电气工程。

(上接第48页)

24—31.

[5] 杨丽君，王伟浩，孙秀飞，等.考虑时间尺度的配电网故障恢复方法研究[J].太阳能学报，2021，42(1)：453—459.

[6] 陈春，吴宜桐，曹一家，等.基于网络拓扑有向遍历的配电网故障快速恢复方法[J].电力系统自动化，2021，45(7)：44—52.

[7] 邓成江，李海锋，曾德辉，等.含逆变型分布式电源的不平衡配电网短路电流计算方法研究[J].陕西电力，2021，49(2)：39—46.

[8] 张亚东，茅东华，余洋.基于蚁群算法的分布式配电网局部故障定位方法[J].能源与环保，2021，43(11)：182—187.

[9] 温立彬，山雨，常亮，等.配电网短路故障在线监测及综合诊断研究[J].电气传动，2021，51(14)：73—80.

[10] 卢晓强，李刚，周万竣，等.基于最小二乘双支持向量机的配电网短路故障辨识方法研究[J].武汉大学学报：工学版，2022，55(4)：401—408.

[11] 李武君，宋伟，张作戍.含分布式电源的配电网短路研究浅述[J].工业控制计算机，2021(3)：144—146.

[12] 乔妮，史亚南，田振华.分布式光伏短路特性对配电网保护的影响研究[J].电力勘测设计，2022(2)：39—45，53.

作者简介：任庭昊(1982—)，男，本科，工程师，研究方向：电力系统调度自动化和配网自动化。