计算机科学
COMPUTER SCIENCE

基于自编码的改进*K*-means光伏能源数据清洗方法

彭勃, 李耀东, 龚贤夫

引用本文

彭勃, 李耀东, 龚贤夫. 基于自编码的改进*K*-means光伏能源数据清洗方法[J]. 计算机科学, 2024, 51(6A): 230700070-5.

PENG Bo, LI Yaodong, GONG Xianfu. Improved *K*-means Photovoltaic Energy Data Cleaning Method Based on Autoencoder [J]. Computer Science, 2024, 51(6A): 230700070-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700070

基于自编码的改进 *K*-means 光伏能源数据清洗方法

彭 勃 李耀东 龚贤夫
广东电网有限责任公司电网规划研究中心 广州 510080
(1339512151@qq.com)

摘 要 智能电网的发展带来了海量能源数据,数据质量是开展数据价值挖掘等任务的基础。然而,多源海量光伏能源数据的采集与传输过程中不可避免地存在异常数据,因此需要进行数据清洗。目前,基于传统统计机器学习的数据清洗模型存在一定的局限性。文中提出了一种基于 Transformer 自编码结构的改进型 *K*-means 聚类模型,用于能源大数据清洗。该模型通过肘部法则自适应地确定聚类簇数,并利用自编码网络对聚类内数据进行压缩和重构,从而实现异常数据的检测和恢复。同时,模型利用 Transformer 的多头注意力机制学习数据间的相关特征,提高了对异常数据的筛查能力。在光伏发电公开数据集上的实验证明,与其他方法相比,该模型具有更好的异常数据检测效果,筛查准确率可达 96%以上。此外,所提模型能在一定程度上恢复异常数据,为能源大数据应用提供了有效的支持。

关键词:自编码;数据清洗;异常检测;Transformer;*K*-means
中图分类号 TP391

Improved *K*-means Photovoltaic Energy Data Cleaning Method Based on Autoencoder

PENG Bo,LI Yaodong and GONG Xianfu
The Grid Planning and Research Center of Guangdong Power Grid Corporation Limited,Guangzhou 510080,China

Abstract The development of smart grids has brought about a massive amount of energy data,and data quality is the foundation for tasks such as data value mining. However,during the collection and transmission process of large-scale photovoltaic energy data from multiple sources,it is inevitable to encounter abnormal data,thus requiring data cleaning. Currently,traditional statistical machine learning-based data cleaning models have certain limitations. This paper proposes an improved *K*-means clustering model based on the Transformer autoencoder structure for energy big data cleaning. It adaptively determines the number of clusters using the elbow method and utilizes autoencoder networks to compress and reconstruct data within clusters,thereby detecting and recovering abnormal data. Additionally,the proposed model employs the multi-head attention mechanism of Transformer to learn the relevant features among the data,enhancing the screening capability for abnormal data. Experimental results on a publicly available photovoltaic power generation dataset demonstrate that,compared to other methods,the proposed model achieves better performance in detecting abnormal data,with a screening accuracy of over 96%. Moreover,it is capable of recovering abnormal data to a certain extent,providing effective support for the application of energy big data.

Keywords Autoencoder,Data cleaning,Anomaly detection,Transformer,*K*-means

1 引言

信息化时代景背下,智能电网作为大数据应用的重要领域,受到广泛关注[1]。近年来,随着我国信息化、工业化进程不断推进,智能电网快速发展,各类传感技术也得到了广泛应用。电力行业产生了海量来源各异、结构复杂的数据。一方面,挖掘能源大数据蕴含的巨大价值,将为电网管理与运行中的有关决策提供重要支撑;另一方面,多源海量能源数据采集过程中不可避免地存在异常数据污染,影响了数据质量[2-5]。如何构建高效的能源大数据清洗模型,对异常数据进行检测,已成为能源行业亟待解决的重要课题。

本文主要关注数据清洗任务,其主体目标是通过模型实现对一系列电力传感数据中的异常数据进行精准检测,完成数据清洗,从而保证能源大数据的正确性与可靠性,为数据挖掘等下游任务奠定基础。

目前,能源大数据清洗模型研究领域已取得初步进展,但面对当今规模庞大、种类繁多的能源大数据仍存在一定局限性。当前国内外能源大数据清洗领域的研究主要聚焦于分布式系统、传统的聚类分析、关联分析、条件函数依赖等与统计机器学习有关的方法[5-14]。

在分布式文件系统框架下,Meng 等[6]提出了一种基于 hadoop 的异常数据清洗模型,但该模型需要用户人为设定属性集上的条件函数依赖,而能源大数据由于其多源性与复杂性,不宜直接构建异常数据识别规则。Qu 等[7]提出了一种基于 Spark 框架的能源大数据清洗模型,基于改进 CURE 聚类算法获取正常簇,但在处理海量数据时性能较差。

基金项目:中国南方电网有限责任公司科技项目 037700KK52220042(GDKJXM20220906)
This work was supported by the Science and Technology Project of China Southern Power Grid Co. Ltd 037700KK52220042(GDKJXM20220906).
通信作者:彭勃(1339512151@qq.com)

在传统的统计机器学习方面，Lv 等[5]针对能源大数据缺失值提出了一种基于支持向量机的验证算法以实现数据清洗，但该方法仅适用于处理缺失值任务。Xing 等[8]提出了一种基于滑动窗口的数据流清洗方法，用于检测异常数据序列。模型在清洗步骤中选用的 Apriori 算法需多次遍历数据集，在海量能源数据应用场景中效率较低。Xu 等[9]提出了基于 DBSCAN 算法和改进高斯核相关向量机的联合异常参数清理方法，检测异常点并预测修复，改进模型在保证预测精度的同时耗时较短。Zhang 等[12]也在 DBSCAN 算法的基础上进行了改进，以自适应地获取参数，基于密度聚类实现数据清洗。但 DBSCAN 很难在不同密度的数据中识别集群且难以聚类高维数据，在面对复杂的海量能源大数据时性能受限。

更进一步，文献[10]针对电力工业终端数据，使用孤立森林算法改进遗传神经网络实现清洗，在脏数据的筛选和修正方面具有较高准确率。但孤立森林算法仍存在不擅长处理局部稀疏点，且不适用于较高维度数据的缺点。Huang 等[11]提出了一种基于次序依赖的电力数据缺失修复方法，通过 RBF 训练神经网络模型对电力数据的次序属性进行特征提取。该方法受限于能源数据次序信息，难以推广至其他复杂场景。

面对海量能源大数据的迅猛增长与变化趋势，现有清洗模型在计算效率、泛化性能及模型准确性等方面仍有欠缺。因此，针对能源大数据特性，研究更为高效稳定的数据清洗方法对开展高质量能源大数据应用具有重要意义。

本文运用了基于 Transformer 的自编码网络结构，提出了一种改进型 K-means 能源大数据清洗模型。运用肘部法则自适应地确定最佳簇数，并基于数据相关的自编码网络，压缩簇内数据并重构其特征，但无法重构离群数据，从而检测筛查离群点并恢复数据。基于 Transformer 的改进自编码网络运用多头注意力机制学习数据间的相关特征，提高了模型对不相关性异常数据的筛查能力，运用 K-means 和自编码网络的联合寻找出网络最佳分类点，能够最大程度筛查出离群值。在公开光伏发电厂传感数据集上进行实验，结果表明该模型相比传统单一的聚类算法筛查效果更佳，并能够在一定程度上恢复异常数据。

2 相关理论

2.1 肘部法则

肘部法则(Elbow Method)是一种经典的聚类算法评估方法，用于确定聚类算法中最佳聚类数的选择[15]。在 K-means 聚类算法中，选择符合样本分布情况的聚类数量对于有效的数据分析至关重要。

肘部法则通过分析聚类结果的评估指标与聚类数量之间的关系来确定最佳聚类数量。其核心是在不同聚类数量下计算聚类结果的紧密度，通常基于数据点与其对应聚类中心之间的平方距离和，即 MSE 评价指标进行评估。它可以自适应地选择最佳聚类中心数。MSE 评价指标可以表示为：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i) \tag{1}$$

随着聚类数逐渐增加，MSE 会先逐渐下降再上升，其拐点表示聚类效果由显著改进到较小改进的转折点。拐点对应的聚类数量在保持聚类效果的同时避免了过度拟合和捕捉噪声的问题。肘部法则简单直观，可以有效避免过拟合导致对新样本的泛化性能下降，为聚类分析中的决策提供依据。

2.2 自动编码器

自动编码器是一种前馈神经网络，具有特殊的结构和功能[16]。它的输入和输出相同，通过将输入数据压缩为低维特征向量，然后根据该表示重构输出。这个特征向量被称为潜在空间表示，是输入数据的一种紧凑的“压缩”形式。

自动编码器由编码器、特征向量和解码器 3 个组件组成，其结构如图 1 所示。编码器负责将输入数据压缩为特征向量，解码器则利用这个特征向量来重构输入数据。

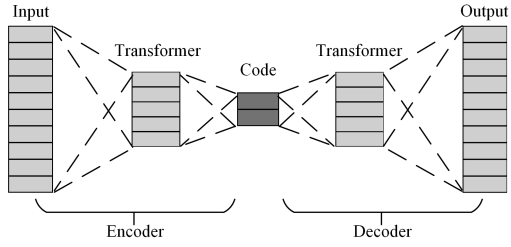

图 1 自动编码器结构

Fig. 1 Autoencoder architecture

自动编码器的结构数学表示如下，其中式(2)为编码器，式(3)为解码器。

$$y=h(x) \tag{2}$$

$$r=f(y)=f(h(x)) \tag{3}$$

自动编码器可以完成特征提取与降维，并发掘数据中的隐藏模式和关联关系。通过学习输入数据的紧凑表示，自动编码器可以去除数据中的噪声和冗余信息，从而提高数据的质量和表征能力。对于能源数据中一些简单的噪声数据，自编码结构本身在编码解码的过程就可以去除噪声数据，达到简易的数据清洗效果。而对于正常数据，降维至升维的过程会很大程度地保留数据原本信息，但噪声数据不含有该结构，在自编码网络后难以恢复到数据本身，从而放大正常数据与异常数据的误差值，便于离群异常检测。

2.3 Transformer 结构

Transformer 是一种基于自注意力机制的序列模型，在自然语言处理和其他序列任务中取得了显著的成果[17-19]。它具有并行计算能力和较短的训练时间。Transformer 结构中最重要的部分是多头注意力结构(Multi-head Attention)。注意力机制被应用于每个层中，用于计算每个输入向量与其他向量之间的关注程度，以更好地捕捉序列中的相关信息。通过对不同数据项进行注意力计算，可以发现它们之间的关联关系。

自注意力机制是 Transformer 的关键组成部分，用于捕捉序列中元素之间的依赖关系。在自注意力机制中，每个输入元素(如词向量)都会与其他元素进行交互，并根据它们的相关性进行加权。这种注意力机制能够自适应地学习不同元素之间的关联程度，而无须依赖于固定的位置关系。

$$Attention(\boldsymbol{Q},\boldsymbol{K}_i,\boldsymbol{V}_i)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

在注意力机制中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 表示 3 个向量均由数据本身转化而来，通过查询、键和值的计算，引入了位置间的关联信息，从而实现了全局的上下文感知能力。

通过计算注意力权重，Transformer 能够根据查询向量衡量每个位置对其他位置的重要性，并利用键和值向量来捕捉位置之间的相关性，构建全局的上下文表示，使 Transformer 模型能够处理长序列并捕捉其内部的依赖关系。

为了增强模型的表达能力,Transformer 采用了多个并行的自注意力头。每个注意力头学习一组不同的注意力权重,从而捕捉不同粒度和方面的关系。最后,多个注意力头的输出被拼接在一起,并通过线性变换进行投影。式(5)为多头注意力的拼接,其中自注意力头与 Attention 计算一一对应。

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_h)W_0 \quad (5)$$

Transformer 引入了编码器和解码器结构。编码器由多个堆叠的相同层组成,用于对输入序列进行编码。解码器也由多个堆叠的相同层组成,用于逐步生成输出序列。

Transformer 结构有助于寻找数据之间的相关信息和关联性,同时增加了 Autoencoder 结构的复杂度,使其能够处理更复杂的异常数据。

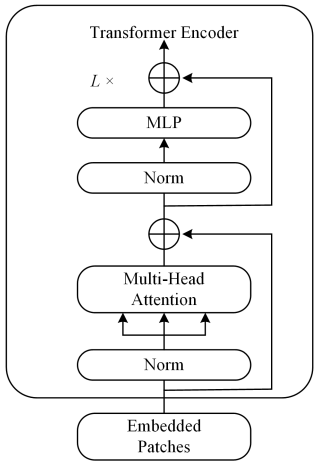

图 2　Transformer 编码器结构

Fig. 2　Structure of Transformer encoder

3　基于自编码的改进 *K*-means 能源大数据清洗方法

本研究提出了一种结合了传统 K-means 聚类和基于 Transformer 的自编码器的数据清洗方法,总体流程如图 3 所示。其中,K-means 的聚类数由肘部法自适应地确定。

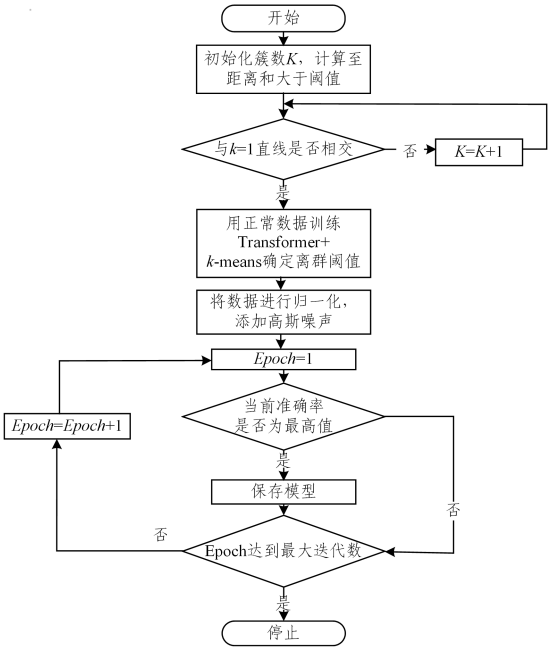

图 3　基于注意力机制改进 K-means 算法流程图

Fig. 3　Process diagram of improved K-means algorithm based on attention mechanism

3.1　聚类簇数选择

对于能源数据,首先对其进行归一化处理,将其转化为固定区间的数据。然后,通过遍历不同的 K-means 聚类中心数,对完整的数据进行分类。每个聚类数与其对应的损失值之和可以在二维平面上表示为一个点。通过遍历得到的 N 个中心点可以连接成一条近似曲线的折线段。这条曲线的下降速率会逐渐减缓,与斜率为 1 的直线一定会有唯一的交点。选取该交点对应的值作为最合适的聚类中心数。

该逻辑的伪代码如算法 1 所示。

算法 1　肘部法

```
输入:数组
输出:K-MEANS聚类类别数
1. for all clusters(K) do
2.    for k=1;k<=K;k++
3.       if loss < eps then
4.          K2← k
5.          break
6.    end for
7.    for k=K2;k>= 1;k--
8.       if intersect(Skk-1,S(k=1)) then
9.          return k
10.   end for
11. end for
```

3.2　基于 Transformer 自编码器的数据编码

自编码器是一种无监督学习模型,通过学习数据的压缩表示和重构能力,从而实现发现数据中的异常值、去除噪声或填补缺失值。

对于微小的噪声干扰,自编码器在压缩过程中通过学习将输入数据映射到对应的低维表示,该低维压缩信息会尽可能地保留数据的原始信息,同时过滤具有随机分布的噪声,使得在低维表示中进行特征选择时不会保留训练中不规律且具有随机性的噪声。

在训练过程中,损失函数衡量压缩重构后的数据与原始数据之间的误差。模型优先拟合样本中大部分合群数据,这些数据具有聚类特征,以最快速度降低损失值。之后,模型会修复离群点数据。在模型训练过程中,必然存在一个阶段,可以区分两类数据并检测出数据中的异常值。

原始数据集为一组不含异常值的数据。我们首先根据肘部法则选择聚类中心值进行聚类,并计算每个样本数据与中心点之间的距离。然后选取距离最大的 5% 的样本数据,对它们与聚类中心的距离求平均值,并取平均距离的 110% 作为 K-means 的离群阈值。

将整体数据添加一定比例的高斯噪声输入 Transformer 自编码网络中。在每次训练后,对 Transformer 的编码结果进行 K-means 检测,最终以筛查噪声离群点的准确率作为模型选择的条件。

在离群值检测准确后,继续训练模型使得损失值降低到稳定。由于自编码器部分的输入输出维度保持一致,我们可以通过 Transformer 网络修复一部分传感器丢失或差异较大的数据异常值。在最终网络结果修复后,将数据输入 K-means 进行聚类,若原有的离群点消失,则说明我们的网络对数据起到了修复作用。

Transformer 模型相较于传统方法具有参数量大、模型训练速度慢的问题。针对此问题,我们对 Transformer 模块

作出了以下调整：

1)逐步下调 Transformer block 的层数以及多注意机制头数，在清洗准确率未出现较大改变时，尽可能简化模型结构。

2)将数据标准化、归一化以及更改 Transformer 网络的初始化参数。

3)使用余弦退火等学习率下调的有关策略，加速模型的收敛。

4)硬件上采用多卡并行训练方式，增大 Batchsize，并通过肘部法则选取中心点，采用 *K*-means 边训练边验证的方式，一但发现有过拟合趋势就立即停止训练，可以精准控制训练迭代数，在一定程度上也减轻了模型压力。

此外，Transformer 结构的自注意力机制可以更有效地学习到数据内部的相关性特征，相较于其他离群点检测算法可以适用于数据维度更小、噪声分布更不明显的数集上，使得模型泛化性能大大提高，因此在时间复杂度上的增加是有意义的。

4 实验与分析

实验选用某光伏发电厂的公开数据集，共含有 4213 条数据，每条数据包含同一观测时间节点对应温度、云量、入射角、发电功率等传感器反馈信息。随机选取其中 70%的数据作为正常数据，对 30%的数据添加不同程度高斯分布的噪声，并在噪声数据中以 7∶3 的比例划分训练和验证集。

对于样本数据，为了防止模型遍历选择过多不必要的中心点，当曲线斜率已经低于一个阈值时，就不进行更多的尝试，从而提高算法效率。在本次实验中，总共遍历到 91 个中心点时停止。选择损失值曲线与斜率为 1 的直线的交点，最终选取中心点数量为 16，选取过程如图 6 所示。

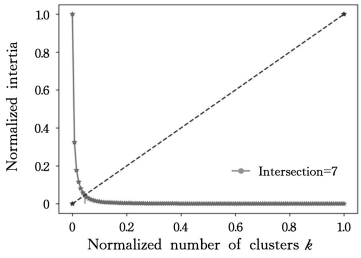

图 4 肘部法则聚类簇数

Fig. 4 Number of clusters by elbow rule

在未添加噪声的 70%原有正常能源数据集上，通过肘部法则选出中心数后，用自编码网络训练正常数据，并记录误差值最高的 5%数据的平均值的阈值，对训练集数据添加一定程度的高斯噪声后，在每个训练步长上都使用 *K*-means 聚类检测离群点，取准确率最高的模型对测试集进行验证。实验结果表明，在测试集上的准确率能达到 96%。

在此基础上进行对比实验，肘部法选取的中心点为 16 个。选取第 15，16，17 这 3 个聚类中心数进行了实验，实验结果表明这几个中心点的效果为最佳。

对于噪声的数值，实验从小到大遍历了多个数值的均值与标准差，当噪声的均值达到 0.5 及以上时，数据清洗模型的效率较高，异常数据筛查任务可达到 98%的准确率。

在相同噪声数据下进行本文方法与传统清洗算法的对比实验，本文方法在准确率、精确度、召回率和 F1 分数等方面均超过了其他方法。实验结果如表 1 所列。

表 1 能源数据清洗对比实验结果

Table 1 Comparative experimental results of energy data cleaning

Methods	*ACC*	*Prec.*	*Rec.*	*F*1
DBSCAN	0.9729	0.8806	0.9423	0.9104
One-Class SVM	0.9262	0.8214	0.9076	0.8623
CURE	0.8637	0.8202	0.8839	0.8508
TransAuto+*K*-means	0.9862	0.9319	0.9600	0.9458

在同一噪声值条件下进行模型的消融实验，与传统的自编码网络和 *K*-means 算法相比，本文模型的准确率最高，实验结果如表 2 所列。

表 2 能源数据清洗消融实验结果

Table 2 Results of ablation experiments on energy data cleaning

Methods	*ACC*	*Prec.*	*Rec.*	*F*1
AutoEncoder	0.9512	0.7605	0.8900	0.8202
TransAuto	0.9762	0.8714	0.9500	0.9090
K-means	0.9750	0.8702	0.9399	0.9038
TransAuto+*K*-means	0.9862	0.9319	0.9600	0.9458

对最终 Loss 值最低的模型进行保存，对所有加噪声的数据进行修复，并重新采用 *K*-means 计算每个噪声数据与中心点间的距离，约有 95%的数据能够被正常修复到阈值以内。

综上所述，结合 Transformer 自编码网络的聚类方法能够在异常数据检测和修复任务中取得较高准确率，可以保障多源海量能源数据的正确性和可用性。

结束语 本文提出了一种基于 Transformer 自编码网络和 *K*-means 聚类的能源大数据清洗方法，针对能源大数据来源多样、结构复杂、数据间存在关联等特点，总结了现有能源数据清洗模型存在的主要局限，创新性地将自编码结构与聚类相结合来训练模型，用以解决能源数据的异常检测与修复问题。

该方法首先由肘部法则自适应地确定聚类簇数，再将原始数据通过初始聚类确定离群阈值，之后通过基于 Transformer 的自编码网络筛选离群点，最后利用训练好的网络对异常数据进行重构。通过与其他聚类方法在异常点检测任务的对比实验证明，该方法在结果上呈现的准确性最高，在此任务上具有优势。

但是，在阈值选取环节，本文提出的方法首先进行正常数据聚类，之后进行人为确定，具有一定局限性。在后续的研究中，我们将考虑针对能源数据特点，采用自适应阈值确定方法，使阈值选取更为合理。

参 考 文 献

[1] WU Y, LIU Y, AHMED S H, et al. Dominant Data Set Selection Algorithms for Electricity Consumption Time-Series Data Analysis Based on Affine Transformation[J]. IEEE Internet of Things Journal, 2020, 7(5): 4347-4360.

[2] KUMAR V, KHOSLA C. Data Cleaning—A thorough analysis and survey on unstructured data[C]// 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018: 305-309.

[3] SHEN X, FU X, ZHOU C. A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm[J]. IEEE Transactions on Sustainable Energy, 2018, 10(1): 46-54.

[4] GUO Z, LV Z, CHEN C. Research ontypical model of network

intrusion and attack in power industrial control system[J]. Information Technology and Network Security，2018，37：37-39.

[5] LV Z，DENG W，ZHANG Z，et al. A Data Fusion and Data Cleaning System for Smart Grids Big Data[C]// 2019 IEEE International Conference on Parallel & Distributed Processing with Applications，Big Data & Cloud Computing，Sustainable Computing & Communications，Social Computing & Networking. IEEE，2019：802-807.

[6] MENG X，ZHOU L，WANG H，et al. Research on MapReduce Data Computing Technology for Intelligent Power Grid Based on Hadoop Cloud Platform [J]. Electrical Measurement & Instrumentation，2015(10)：66-72.

[7] QU Z，ZHANG Y，WANG Y，et al. Big Data cleaning model of energy Internet based on Spark framework [J]. Electrical Measurement and Instrumentation，2018，55(2)：39-44.

[8] XING Y，MENG C，LI C，et al. Lean Operation and Maintenance Evaluation Technology of Power Grid Equipment Based on Improved Big Data Cleaning Method[C]// 2020 IEEE 4th Conference on Energy Internet and Energy System Integration. IEEE，2020：2749-2752.

[9] XU B. Power Station Abnormal Data Cleaning Method Based On Big Data Mining[C]// 2021 IEEE Sustainable Power and Energy Conference. 2021：3809-3814.

[10] LV Z，HU Z，NING B，et al. The Data Cleaning of Electric Industrial Control Terminal Based on the iForest and Genetic BP Neural Network Algorithms[C]// 2019 IEEE 2nd International Conference on Information Communication and Signal Processing. IEEE，2019：490-494.

[11] HUANG F H C. Research on automatic repair method of power data missing based on order dependence [J]. Automation and Instrumentation，2020(12)：233-236.

[12] ZHANG X，LIN R，XU H. An Adaptive Parameters Density Cluster Algorithm for Data Cleaning in Big Data[C]// Artificial Intelligence and Security：6th International Conference. 2020：543-553.

[13] LIN N，WU Y. A Big Data Cleaning Method Based on Improved K-means [J]. Journal of Microcomputer Applications，2021，37(11)：133-136.

[14] CHEN X，ZHANG X. Extract-transform-load of data cleaning method in electric company[C]// 2010 International Conference on Artificial Intelligence and Computational Intelligence. IEEE，2010，3：345-349.

[15] CUI M. Introduction to the k-means clustering algorithm based on the elbow method[J]. Accounting，Auditing and Finance，2020，1(1)：5-8.

[16] ZHAI J，ZHANG S，CHEN J，et al. Autoencoder and its various variants[C]// 2018 IEEE International Conference on Systems，Man，and Cybernetics(SMC). IEEE，2018：415-419.

[17] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J/OL]. Advances in Neural Information Processing Systems，2017，30. https://proceedings. neurips. cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper. pdf.

[18] LIN J，SHENG G，YAN Y，et al. Online monitoring data cleaning of transformer considering time series correlation[C]// 2018 IEEE/PES Transmission and Distribution Conference and Exposition(T&D). IEEE，2018：1-9.

[19] SO D，LE Q，LIANG C. The evolved transformer[C]// International Conference on Machine Learning. PMLR，2019：5877-5886.

PENG Bo，born in 1991，master，engineer. His main research interest is power grid planning.