文章编号:1009-1831(2023)05-0053-06　　中图分类号:TM73;TP311.13　　文献标志码:A

DOI:10.3969/j.issn.1009-1831.2023.05.009

基于正则自编码器及Optuna寻优的异常用电数据清洗研究

陈　慧[1],陈　适[1],郭银婷[1],连淑婷[2],王　康[2],韦先灿[2]

(1. 国网福建省电力有限公司 营销服务中心,福州　350001;2. 福州大学 电气工程与自动化学院,福州　350108)

Abnormal power consumption data cleaning based on regular self-encoding and Optuna optimization

CHEN Hui[1], *CHEN Shi*[1], *GUO Yinting*[1], *LIAN Shuting*[2], *WANG Kang*[2], *WEI Xiancan*[2]

(1. Marketing Service Center, State Grid Fujian Electric Power Co., Ltd., Fuzhou 350001, China;2. College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China)

摘要:为有效解决用电信息采集系统中电量数据丢失问题,提出基于正则自编码器的缺失数据填补方法。首先,根据正则自编码器学习到的特征重构电量数据,实现缺失数据的修复。然后,通过对损失函数增加L21范数及正交约束实现正则化,提升模型的泛化能力,并采用Optuna实现超参数的自动寻优。最后,实际数据集的测试结果表明:与其他自编码器相比,正则自编码器能够较为准确地补齐缺失数据。

关键词:异常数据清洗;自编码器;正则化;Optuna寻优

Abstract: In order to effectively solve the problem of consumption loss in the electric energy information acquisition system, a method of filling missing data based on regular self-encoders is proposed. Firstly, the energy data according to the characteristics learned by the regular autoencoder is reconstructed, and the repair of the missing data is realized. Then, regularization by adding the L21-norm is realized and orthogonal constraints to the loss function, the generalization ability of the model and uses Optuna to realize the automatic optimization of hyperparameters is improved. Finally, the test results of the actual data set show that compared with other autoencoders, the regular autoencoder can accurately fill in the missing data.

Key words: abnormal data cleaning; self-encoder; regularization; Optuna optimization

0　引言

智能电能表可实时采样、精确计量原始能量数据。所采集电量数据可用于为用户画像,助力配电系统的规划及优化[1],对用户短期电量预测配电网状态估计等研究实施亦有重要意义[2-3]。但电表本体、通信信道以及采集软件等环节都可能发生故障,导致计量失准,数据异常、数据缺失。基于电量数据研究的准确性,对智能电能表获取的原始数据进行数据清洗显得尤为重要。

数据清洗通常包括识别异常数据及填补缺失数据两部分[4]。针对缺失数据,文献[5]采用均值填补法,但算法简单对于突变数据填补效果差;文献[6]将皮尔逊相关系数与回归模型相结合,对缺失数据进行插补,但对非连续和稀疏特征适应性较差;文献[7]通过函数估计法将原有数据映射到新的函数空间,利用相似用户用电特征对缺失值进行修复。

异常数据检测,常见的有统计量分析法[8]、3sigma法[9]、四分位算法[10]等。其中,统计量分析法对变量做描述性统计,剔除不合理数据适合处理有明确范围的数据;3sigma法则将超过3倍标准差数据标为异常量;四分位算法通过四分位数间距获得异常数据范围。近年来,机器学习算法为识别异常数据提供新的思路,如采用K-means聚类算法[11]、支持向量机回归算法[12]等实现正常和异常数据分离。

考虑到智能电能表数据可信度较大,本文的数据清洗侧重在填补缺失数据,设计基于正则自编码器的电量数据补全方案。该方案利用正则自编码器学习原始数据特征再重构原始数据,通过实例数据验证所提方案可行性与准确性。正则自编码器在损失函数中增加L21范数、正交约束实现正则化,防止模型过拟合,提升泛化能力,并利用Optuna实现模型超参数的自动寻优,避免手动调参的繁琐。

收稿日期:2023-02-09;**修回日期**:2023-06-30

基金项目:国网福建省电力有限公司科技项目(52130X21001A)

1 数据分析与处理

用采系统每日采集用户的智能电能表数据。工业用户2020年用电量如图1所示，居民用户电量曲线如图2所示。

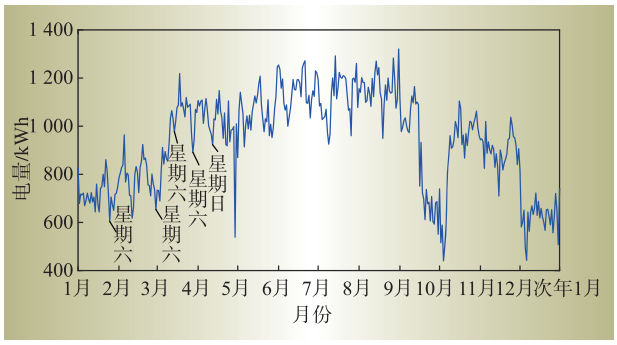

图1 某工业用户2020年用电量

Fig. 1 Electricity consumption of an industrial user in 2020

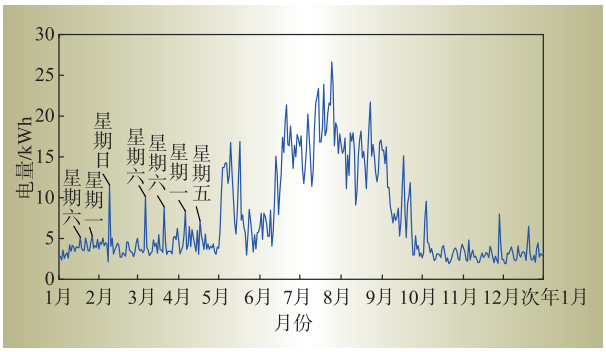

图2 某居民用户2020年用电量

Fig. 2 Electricity consumption of a resident user in 2020

观察图1可知，工业用户在节假日的用电量明显下降；非节假日时，用电谷底大部分是周末，尖峰大部分为星期三、星期四。图2中居民用户的用电行为和工业用户恰好相反，此外，居民用户季节周期明显，夏日的用电量高于其他季节。可见，用户用电消费呈现一定的周期性，这些都为缺失数据补全提供了有利的条件。

对用采系统所采集到原始数据进行预处理，处理过程主要包括电量计算、异常数据处置及数据切片。

（1）日用电量计算

获取台区中分表的数据后，当用户前后两天都存在数值时，将后一天减去前一天的正向有功，即可得到用户的日电量数据；当用户前后两天数据存在空值时，将用户当天的用电量置为Nan，即数据缺失。

（2）数据异常点的处理

原始数据中的正向有功应为单调增加，若计算所得日用电量为负值，则视为数据异常，置为Nan。在不考虑低压侧线损的情况下，若当日所有用户的电量数据均存在，并在一定的误差范围内的求和数值不等于总电量，即认定当日所有用户的电量数据错误，置为Nan。在上述处理完成后，统计用户日用电量的缺失天数，若缺失百分比大于或等于总天数的40%，则不修正填补该用户的电量数据。

（3）数据切片

由于用户的用电行为呈现一定的周期性。本文将符合要求的用户电量数据以28天为一个周期，从星期一开始，星期日结束，进行数据切片工作。

2 方案设计与原理

2.1 数据清洗方案

针对无法直接获得缺失值的情况，本文利用正则自编码器实现缺失数据的填充，完成数据修复工作，所设计数据清洗方案流程如图3所示。

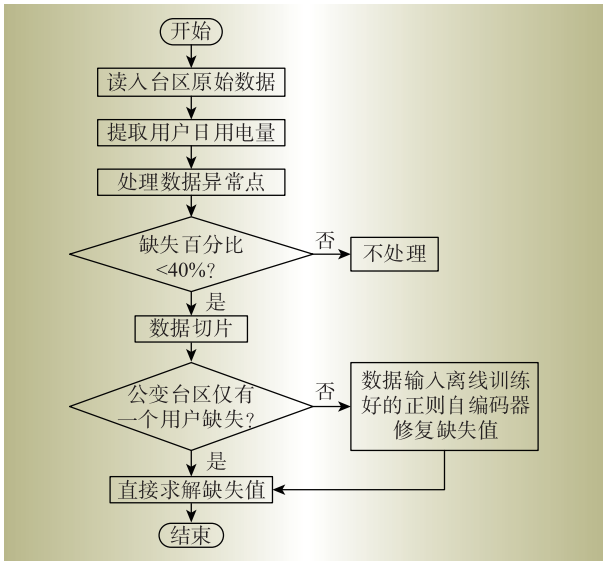

图3 数据清洗流程

Fig. 3 Data cleaning flow

2.2 正则自编码器原理

自编码器原理结构[13]如图4所示。将输入向量 x 转化为隐藏表示 h 的确定性映射称为编码过程，数学表达式为

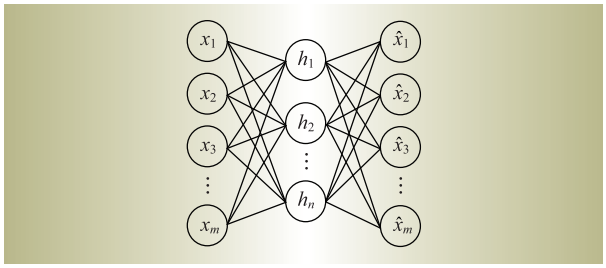

图4 自编码器结构

Fig. 4 Structure of the auto-encoder

$$h = f(\boldsymbol{W}x + \boldsymbol{b}) \tag{1}$$

式中：f 为激活函数；$\boldsymbol{W}$ 为编码网络的权值；$\boldsymbol{b}$ 为偏置。

为了衡量网络学习到的特征的质量，需对隐藏表示 h 进行解码，解码过程为

$$\hat{x}=g(\boldsymbol{W}'h+\boldsymbol{d}) \tag{2}$$

式中：$\boldsymbol{W}'$ 为解码网络的权值；$\boldsymbol{d}$ 为偏置。

利用均方误差（mean square error，MSE）来量化输入数据 x 和重构结果 $\hat{x}$ 之间的误差，表达式为

$$J_{\mathrm{MSE}}(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\left\|\hat{x}_i-x_i\right\|^2\right) \tag{3}$$

式中：N 为样本总数；θ 为所有的参数集合。

自编码器目标是误差 $J_{\mathrm{MSE}}(\theta)$ 越小越好，但一味降低损失函数值可能存在过拟合的风险。因此，需超越训练集中的数据对模型的泛化性能进行提升。

为防止出现过拟合现象，本文考虑对损失函数施加规则约束缩小解的空间。在正则化函数中选择L1范数、L2范数，防止过拟合并稳定快速求解优化值[14]。模型使用L21范数及正交约束作为损失函数的正则项，L21范数数学表达式如下

$$\left\|\boldsymbol{X}\right\|_{2,1}=\sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n}\boldsymbol{X}_{ij}^2} \tag{4}$$

式中：m 为行数；n 为列数；$\boldsymbol{X}$ 为网络的权值。

此外，通过增加正交约束项生成不相关的子空间。模型通过衡量 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}-\boldsymbol{I}$ 的L2范数大小判断网络权重矩阵 $\boldsymbol{W}$ 的好坏。因此，本文求解 $\left\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}-\boldsymbol{I}\right\|_2$ 并将该项加到损失函数中。借助正交约束和L21范数的嵌入，最终待优化的目标函数可以写为

$$J=J_{\mathrm{MSE}}(\theta)+\sigma\sum_{i=1}^{v}\left\|\boldsymbol{W}\right\|_{2,1}+\beta\sum_{i=1}^{v}\left\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}-\boldsymbol{I}\right\|_2 \tag{5}$$

式中：v 为自编码器的层数；σ 为L21范数的系数；β 为正交约束的系数；$\boldsymbol{I}$ 为单位矩阵。

2.3 用于电量数据清洗的正则自编码器设计

本文方案中，经测试后设置编码器和解码器均为4层。采用梯度下降法来更新网络的权值、偏置和参数，减小原始数据和重构数据误差。正则自编码器训练的收敛过程示意如图5。

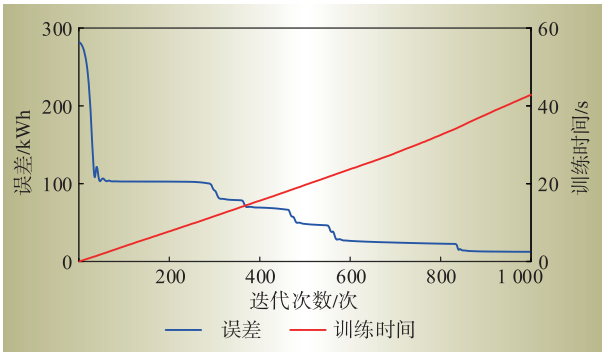

图5 不同迭代次数下网络的误差及训练时间

Fig. 5 Error and time of the network under different trials

由图5可知，迭代次数过小时，模型欠拟合导致误差较大，而迭代次数过大会使模型训练时间过长，因此，本文迭代次数为1 000。

深度学习网络的超参数无法通过训练优化，需在训练前人工设定并进行调整。为测试本文所搭建的自编码器模型对超参数的敏感度，有规律地调整超参数，获得的结果如图6所示。图中：k 为正则自编码器提取到的特征 h 的维数，f_{a}、f_{b}、f_{c} 为每一层神经元的个数。

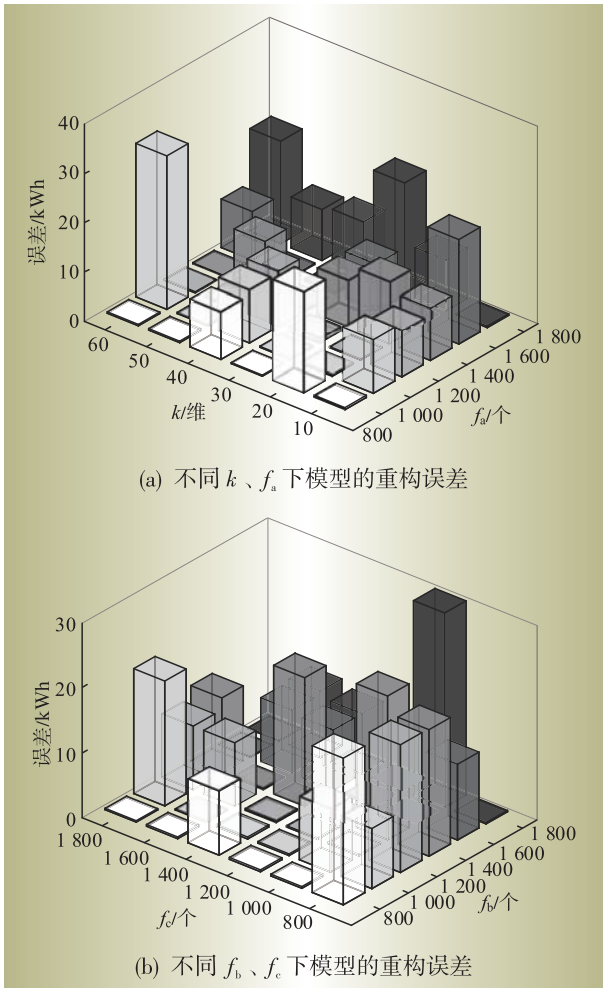

图6 不同超参数对模型误差的影响

Fig. 6 Influence of different hyperparameters on errors

由图6(a)可知，正则自编码器将电量数据映射到高维空间时误差较小。从 f_{a} 的角度来看，神经元个数在800—1 200区间时，模型误差除偶尔的波动外大多较低。根据图6(b)可得，当 f_{b} 取值较低，f_{c} 取值较高时，模型误差较小。其主要原因为不同网络层维度差异较大时，网络能提取出表征能力较强的特征，从而达到重构误差小于阈值的要求。

采用试错法的超参数优化过程依赖大量经验，且时间成本高；目前存在的大部分超参数寻优方法需要静态地为每个模型构造搜索空间，当空间定义

不恰当时，算法的优势可能不复存在，因此本文考虑采用Optuna自动调整超参数。

Optuna将超参数寻优描述为最小化目标函数值的过程，通过该函数动态构建网络结构的搜索空间，不依赖外部定义的静态变量，最后执行协方差矩阵自适应进化策略（covariance matrix adaptation-evolution strategy，CMA-ES）关系采样算法获得超参数的最优解。CMA-ES通过对多元正态分布中采样 z 个独立的样本 $x_z^{(g+1)}$ 生成新搜索点[15]，数学表达式如下

$$x_z^{(g+1)} \sim m^{(g)} + \sigma^{(g)} \mathrm{N}(0, \boldsymbol{C}^{(g)}) \quad z = 1, 2, \cdots, \lambda \tag{6}$$

式中：$m^{(g)}$ 为第g代样本的加权均值；$\sigma^{(g)}$ 为步长；$\mathrm{N}(0, \boldsymbol{C}^{(g)})$ 为均值为0，协方差矩阵为 $\boldsymbol{C}^{(g)}$ 的正态分布。

在本文设计的用于电量数据清洗的自编码器模型中，定义待优化的目标函数为重构误差，参考图6设置超参数的范围，部分训练的收敛过程如图7所示，模型参数如表1所示。

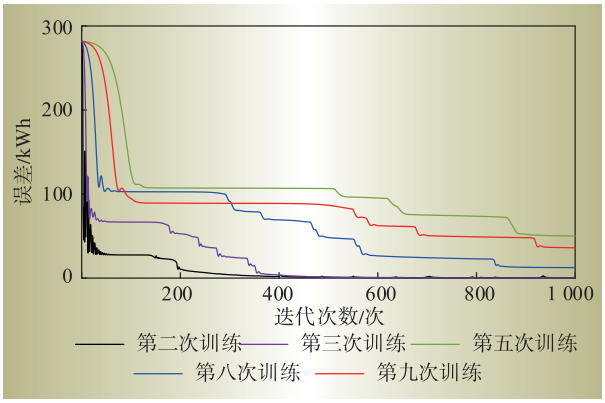

图7 模型的收敛过程

Fig. 7 Convergence process of the model

由图7可知，采用Optuna自动调整超参数，大部分会在700个循环内收敛，在训练过程结束后，保存误差最小的模型，离线训练自编码器的过程就此完成。本文最终保存第三次训练的模型。

表1 不同模型的参数值

Table 1 Parameters of different models

参数	k/维	f_a/个	f_b/个	f_c/个
第二次训练模型	69	917	104	1 539
第三次训练模型	116	594	1 080	1 634
第五次训练模型	123	367	356	577
第八次训练模型	11	130	1 929	889
第九次训练模型	116	275	1 331	295

3 实验验证与对比

3.1 基于正则自编码器的电量数据填补

真实数据进行部分剔除处理后作为测试集，测试集数据涵盖了16个台区的电量数据，并对每条数据随机剔除3个点，包含用户用电量骤增、骤减、逐步上升、波动相对稳定等情况。基于离线训练好的正则自编码器模型实现缺失数据的补齐，填补效果如图8所示。

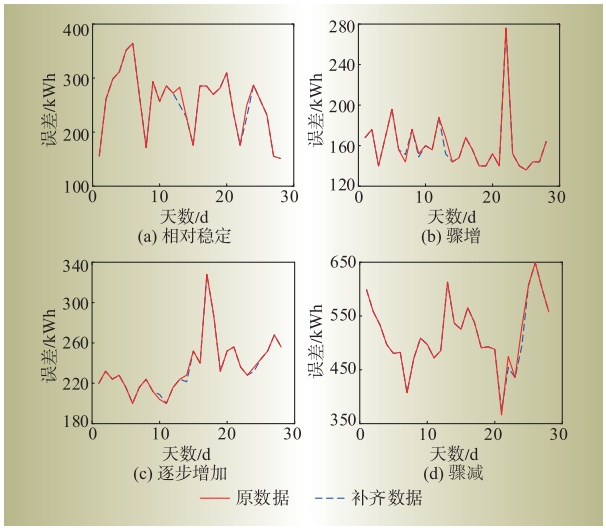

图8 基于正则自编码器的填补效果

Fig. 8 Filling effect based on regular auto-encoder

从图8看出，电量数据波动稳定或是存在突变情况需要填补时，正则自编码器的填补数据与原始数据误差较小。

3.2 实验对比

3.2.1 基于普通自编码器的电量数据填补

普通自编码器的数据修复效果如图9所示。

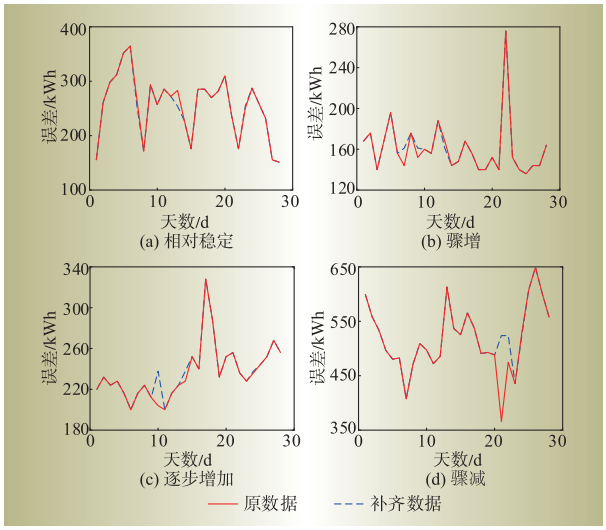

图9 基于普通自编码器的填补效果

Fig. 9 Filling effect based on auto-encoder

由图9可知，电量稳定波动时，基于普通自编码器的数据填补误差较小，但用户用电量突变时，普通自编码器重构数据效果差，并且在数据陡降时，填补结果相反。

3.2.2 基于降噪自编码器的电量数据填补

降噪自编码器通过在输入数据中引入随机噪声扰动，迫使网络在学习的过程中去除随机噪声，从而达到最小化重构误差的目的。利用降噪自编码器修复缺失电量数据的结果如图10所示。

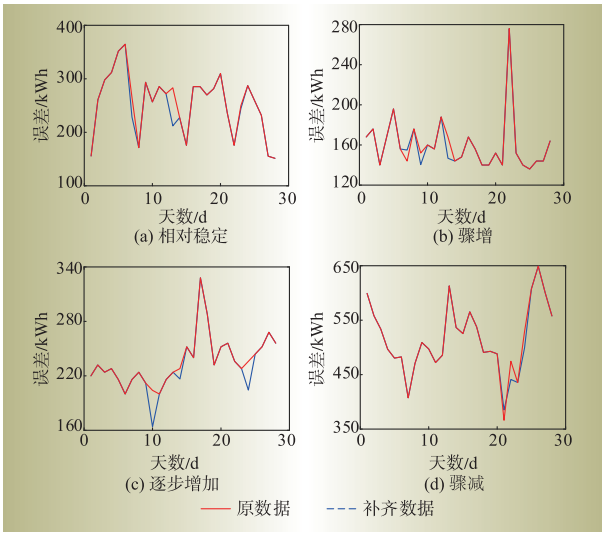

图10 电量数据填补效果

Fig. 10 Electricity data filling effect

相比采用传统的自编码器，基于降噪自编码器填补缺失电量数据的效果较差，尤其对于波动幅度较小的图10(a)、图10(c)，降噪自编码器无法准确学习到该类型的表示，导致重构误差较大，数据修复效果并不理想。

为更直观反映基于正则自编码器、普通自编码器及降噪自编码器的数据填补效果，采用 E 函数衡量补齐数据和原始数据的误差，数学表达式如下

$$E=\frac{\sum_{x_{ij}\in M}\left|x_{ij}-x'_{ij}\right|}{\sum_{x_{ij}\in M}x_{ij}} \tag{7}$$

式中：x_{ij} 为真实值；x'_{ij} 为填补值；M 为所有缺失值的集合。

将测试集中所有的补全数据和已知的真实数据代入式(7)，对填补结果进行误差统计，如表2所示。

表2 不同自编码器误差对比

Table 2 Error comparison of different autoencoders

%

算法	E
正则自编码器	13.86
普通自编码器	16.16
降噪自编码器	34.33

由表3可知，正则自编码器在用户电量数据填补中展现出优异性能，表现优于其他的自编码器形式。为体现本文所提的自编码器正则化函数的优势，当损失函数加入不同的约束时，填补的误差如表3所示。

表3 加入不同约束的自编码器误差对比

Table 3 Error of autoencoders with different constraints

%

约束	E	约束	E
L1范数	22.04	L21范数	14.75
L2范数	15.40	L21范数+正交约束	13.86

由表3可知，采用L21范数和正交约束实现自编码器正则化效果最好。由于电量数据变化较为多样，而L1范数趋向于产生少量的特征，L2范数会选择更多的特征，因此采用L2范数作为约束项重构误差较小；L21范数先对每一列求L2范数，再将得到的结果求L1范数，能充分利用两者的鲁棒性，重构误差更小；而正交约束令模型学习到的特征相互独立，更全面，进一步减小数据的重构误差。综上，文中所提的正则自编码器能融合不同约束的优势，准确地补齐缺失的数据。

4 结束语

本文以实际用户用电量数据作为研究样本，针对智能电能表数据采集过程中易出现的数据缺失问题，提出了基于正则自编码器的数据填补方法。该方法通过对损失函数施以一定的约束实现自编码器的正则化，增强模型的泛化能力，同时采用Optuna自动寻找最优超参数，避免手动调整的繁琐，最后根据正则自编码器学习获取的用户用电量特征解码重构数据，实现对缺失数据的修复。对实际样本进行的测试实验表明，正则自编码器能够较为准确地提取出用户用电量的特征，并据此补齐缺失的数据；并且，该方案修补结果的误差相对较小，对于电量数据相关性研究开展及配电网大数据技术的数据质量改善具有积极作用。

参考文献：

[1] 张鹏，蒯圣宇，刘维，等. 考虑双侧不确定性的负荷聚集商需求响应资源规划[J]. 电力需求侧管理，2021，23(1)：49-54.
ZHANG Peng, KUAI Shengyu, LIU Wei, et al. Demand response resources planning for load aggregators considering bilateral uncertainty[J]. Power Demand Side Management, 2021, 23(1): 49-54.

[2] 陈明帆，宁光涛，李琳玮，等. 基于K-L信息量和ARIMA误差修正的月度电量预测[J]. 电力需求侧管理，2021，23(2)：43-46.
CHEN Mingfan, NING Guangtao, LI Linwei, et al. Monthly electricity forecasting based on K-L information and ARIMA error correction [J]. Power Demand Side Management, 2021, 23(2): 43-46.

[3] 赵永红，张旭，程少华，等. 基于关联度变权的台区理论线损水平评估方法研究[J]. 电力需求侧管理，2020，22(2)：39-43.
ZHAO Yonghong, ZHANG Xun, CHENG Shaohua, et al. Evaluation method of line loss level in low voltage area based oncorrelation degree variable weight[J]. Power De-mand Side Management, 2020, 22(2): 39-43.

[4] 李晓宇，陈炫锴，李嘉栩，等. 基于机器学习的电动汽车续航里程预测[J]. 电器与能效管理技术，2021(10)：78-82，91.
LI Xiaoyu, CHEN Xunkai, LI Jiaxu, et al. Driving range prediction of electric vehicles based on machine learn-ing[J]. Electrical & Energy Management Technology, 2021(10): 78-82, 91.

[5] 陈瑞兴，尹洪苓，安东升，等. 大数据技术在配电网全时序运行效率分析中的应用[J]. 供用电，2021，38(3)：22-30.
CHEN Ruixing, Yin Hongling, AN Dongsheng. Application of big data technology in the analysis of full time sequence operation efficiency of distribution network [J]. Distribution & Utilization, 2021, 38(3): 22-30.

[6] 刘沅昆，栾文鹏，徐岩，等. 针对配电变压器的数据清洗方法[J]. 电网技术，2017，41(3)：1 008-1 014.
LIU Yuankun, LUAN Wenpeng, XU Yan, et al. Data cleaning method for distribution transformer [J]. Power System Technology, 2017, 41(3): 1 008-1 014.

[7] 田英杰，洪子靖，周李. 基于函数型数据分析的工商业居民用户电力数据清洗算法[J]. 电测与仪表，2021，58(1)：11-19.
TIAN Yinjie, HONG Zijing, ZHOU Li. Data cleaning algorithm for industrial and commercial residential users based on functional data analysis [J]. Electrical Measurement & Instrumentation, 2021, 58(1): 11-19.

[8] 杨悦，吴量，叶则韶，等. 用于电费结算的用户电量数据清洗与拟合方法[J]. 电力与能源，2021，42(1)：137-139.
YANG Yue, WU Liang, YE Zeshao, et al. Cleaning and fit-ting method of user electricity consumption data used for electricity bill settlement[J]. Power & Energy, 2021, 42(1): 137-139.

[9] 沈小军，周冲成，吕洪. 基于运行数据的风电机组间风速相关性统计分析[J]. 电工技术学报，2017，32(16)：265-274.
SHEN Xiaojun, ZHOU Chongcheng, LYU Hong. Statistical analysis of wind speed correlation between wind turbines based on operational data[J]. Transactions of China Electrotechnical Society, 2017, 32(16): 265-274.

[10] 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, 2019, 10(1): 46-54.

[11] 王赛一，余建平，孙丰杰，等. 电力大数据的价值密度评价及结合改进k-means的提升方法研究[J]. 智慧电力，2019，47(3)：8-15.
WANG Saiyi, YU Jianping, SUN Fengjie, et al. Evaluation and promotion methods with improved k-means for value density of electric power big data [J]. Smart Power, 2019, 47(3): 8-15.

[12] 王雷，张瑞青，盛伟，等. 基于支持向量机的回归预测和异常数据检测[J]. 中国电机工程学报，2009，29(8)：92-96.
WANG Lei, ZHANG Ruiqing, SHENG Wei, et al. Regression forecast and abnormal data detection based on support vector regression [J]. Proceedings of the CSEE, 2009, 29(8): 92-96.

[13] 朱恒东，马盈仓，张要，等. 基于L21范数和回归正则项的半监督聚类算法[J]. 郑州大学学报(理学版)，2020，52(4)：67-74.
ZHU Hengdong, MA Yingcang, ZHANG Yao, et al. Semi-supervised clustering algorithm based on L21-norm and regression regular term[J]. Journal of Zhengzhou University(Natural Science Edition), 2020, 52(4): 67-74.

[14] 曾飞，杨雄，苏伟等. 基于区块链与数据湖的电力数据存储与共享方法[J]. 电力工程技术，2022，41(3)：48-54.
ZENG Fei, YANG Xiong, SU Wei, et al. Power data storage and sharing method based on blockchain and data lake[J]. Electric Power Engineering Technology, 2022, 41(3): 48-54.

[15] 高冰，周文博，王正平. 配电网多太阳能光伏系统分布式协调控制策略[J]. 电力电容器与无功补偿，2022，43(2)：154-163..
GAO Bing, ZHOU Wenbo, WANG Zhengping. Distributed coordinated control strategy for multi-solar-photovoltaic system in distribution network [J]. Power Capacitor & Reactive Power Compensation, 2022, 43(2): 154-163.

作者简介：

陈慧(1990)，女，福建仙游人，硕士，工程师，主要从事电能计量技术工作；

陈适(1981)，男，福建福州人，学士，工程师，主要从事电能计量及用电信息采集终端检测工作。

(责任编辑 于丽芳)