

基于两阶段自注意力权重融合模型的 窃电检测缺失数据填补方法

黄钦瑜¹, 杨明发¹, 王万豪¹, 刘宇龙², 陈煌滨¹, 金涛¹

(1. 福州大学电气工程与自动化学院, 福建省 福州市 350108;

2. 北京大学能源研究院, 北京市 海淀区 100871)

A Novel Two-stage Self-attention Weight Fusion Model for Missing Data Imputation in Electricity Theft Detection

HUANG Qinyu¹, YANG Mingfa¹, WANG Wanhao¹, LIU Yulong², CHEN Huangbin¹, JIN Tao¹

(1. College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, Fujian Province, China;

2. Institute of Energy, Peking University, Haidian District, Beijing 100871, China)

ABSTRACT: The missing data in electricity consumption can affect the performance of electricity theft detection models, while traditional interpolation methods is difficult to retain the key features and limit the performance of the model. This paper proposes a two-stage self-attention weight fusion (TS-SAWF) model for missing data imputation in electricity theft detection. This model utilizes a two-stage convolution-enhanced diagonal masked multi-head self-attention module to capture the correlations and features between the data. It further integrates attention weights from both stages to improve imputation accuracy. Experimental results on two real-world electricity meter datasets from the State Grid Corporation of China demonstrate that, compared to the comparison models, the TS-SAWF model achieves higher imputation performance under varying random missing rates and different numbers of consecutive missing days. Moreover, the data imputed by the TS-SAWF model effectively enhances the performance of the electricity theft detection model, further demonstrating its effectiveness.

KEY WORDS: missing data imputation; electricity theft detection; self-attention mechanism; deep learning

摘要: 用电数据缺失会影响窃电检测模型的性能, 而传统插值填补方法难以保留关键特征, 限制模型检测性能。该文提出了一种基于两阶段自注意力权重融合(two-stage self-attention weight fusion, TS-SAWF)模型的窃电检测缺失数据填补方法。该模型利用双阶段卷积增强对角掩码多头自注意力模块, 捕捉数据之间的相关性特征, 并融合两阶段的注意力权重, 提高缺失数据填补的准确性。在中国国家电网公司两个真实

用户电表采集数据集上的实验结果表明, 与对照模型相比, TS-SAWF 模型在不同随机缺失率和不同连续缺失天数下均能提升填补性能。同时, 基于 TS-SAWF 模型填补后的数据有效地提升了窃电检测模型的性能, 进一步证明其有效性。

关键词: 缺失数据填补; 窃电检测; 自注意力机制; 深度学习

DOI: 10.13335/j.1000-3673.pst.2024.2112

0 引言

窃电是造成电网运行中非技术性损失的重要因素之一, 不仅影响电力系统的安全与稳定运行^[1-2], 还给社会和电力公司造成巨大的经济损失^[3-4]。近年来, 随着智能电表的广泛普及和高级量测体系(advanced metering infrastructure, AMI)的快速发展, 它们所提供的海量用电信息为数据驱动的窃电检测奠定了坚实的基础^[5-6], 有效地遏制了窃电行为的发生。数据驱动方法的关键在于数据质量^[7]。然而, 由于用电信息在采集、传输和存储等环节中可能发生设备故障或通讯遭受干扰, 数据丢失的风险不可忽视。这种数据缺失对窃电检测的准确性和可靠性构成了重大挑战。精准的窃电检测依赖于完整、可靠的高质量数据, 因此, 准确地填补缺失用电数据, 确保为窃电检测任务提供高质量的数据, 显得尤为重要。

在窃电检测中, 对于缺失数据的处理方法可以总结为删除法和填补法。在文献[8]中, 对于缺失数据总数超过 600 的用户, 采用删除法进行处理。虽然删除法操作简便, 但它可能会导致数据间重要信息的丢失, 从而破坏数据的完整性。相比之下, 均值填补^[9]、中位数填补和插值填补^[10-11]等基于数学

基金项目: 国家自然科学基金项目(52377088)。

Project Supported by National Natural Science Foundation of China (52377088).

统计的填补法更常用于解决窃电检测中的缺失数据问题。这些方法通过填补缺失数据，保持了数据的完整性。然而，这些方法通常忽略了数据间的相关性和时序特征^[12]，可能导致填补结果不够准确，从而影响窃电检测模型的性能。相较于基于数学统计的填补方法，矩阵补全方法由于不需要先验知识的训练，并具有计算复杂度低的优势，为缺失数据填补提供了新见解。例如，文献[13]针对非侵入式负荷监测与分解的精度下降问题，提出一种基于CP(cnonical polyadic)分解的正则化低秩张量补全的量测数据缺失修复方法。实验结果表明所提出的方法可以提高数据修复精度，在高缺失率和连续缺失情况下仍能较好地补全效果。文献[14]提出一种利用轻量级多通道注意力融合网络的窃电检测边缘计算方法。在该方法中，使用低秩矩阵方法填补缺失值，实验结果表明基于交替方向乘子法的矩阵补全显著改善了数据质量，为后续窃电检测模型的设计及训练奠定基础。然而，基于矩阵补全的方法对于非凸问题采用的优化方法过于复杂且在长时间连续缺失情况下补全性能不高^[15]。为了克服传统缺失数据填补方法的局限性，越来越多的研究学者将人工智能方法应用于时序缺失数据的填补。k最近邻^[16]、极限学习机^[17]和随机森林^[18]等经典机器学习算法在缺失数据填补方面取得了一定成效，具有实现简单、计算成本较低等优点。但是在面对复杂多变的用电模式时，它们往往无法充分捕捉数据的深层次特征。简单的统计特征和相邻数据的相似性不足以全面反映用电行为的变化，从而影响了填补结果的准确性。此外，在AMI体系提供海量用电数据的支撑下，传统机器学习方法在补全大规模数据时可能面临效率低下和泛化能力不足等问题。

近年来，循环神经网络(recurrent neural network, RNN)^[19]和生成对抗网络(generative adversarial network, GAN)^[20]等深度学习模型及其变体因其优异性能，广泛应用于时序缺失数据填补领域，以解决数学统计、矩阵补全和机器学习等填补方法的局限性。基于深度学习的填补方法通过自主学习时序数据间的特征与相关性，能够更准确地预测和填补缺失值。例如，为了解决分布式光伏集群多维量测信息的缺失问题，文献[21]提出了一种基于长短期记忆网络(long short-term memory, LSTM)的双向多阶段循环插补网络。算例结果表明，该方法有效地增强了分布式光伏集群的基础数据质量。文献[22]构建了一种基于增强生成对抗网络模型，用于补全配电网中非实时观测节点的量测

缺失数据。实验结果表明，该方法不仅能够高精度地补全缺失量测数据，还能有效降低电压波动，从而提高配电网运行的稳定性。文献[23]提出了一种基于双通道GAN的城市用电负荷缺失数据补全方法，并在多个不同程度随机和连续缺失场景下验证了所提方法的可行性与优越性。尽管RNN类和GAN类模型在缺失数据填补方面表现出色，但RNN类模型在处理长时间序列数据时计算效率较低，难以充分利用图形处理器(graphics processing unit, GPU)的并行计算优势，而GAN类模型在训练过程中容易出现模式崩溃和收敛困难。为了高效准确地填补缺失数据，文献[24]结合自注意力机制的并行训练能力和全局信息捕捉优势，提出了基于自注意力机制的时序填补(self-attention-based imputation for time series, SAITS)模型。在文献[25]中，SAITS模型被用于补全配电网D-PMU系统中的缺失数据。实验结果表明，SAITS模型在数据补全上表现出色，有效地提高了D-PMU系统的数据质量。SAITS模型通过并行处理数据提高运算效率，并在训练过程中更好地捕捉数据的特征与相关性，从而提高填补精度。凭借这些优势，SAITS模型成为了填补时序缺失数据的新兴方案，同时也为数据驱动的窃电检测缺失数据填补提供了新思路。

综上所述，为进一步提高用电信息缺失数据的填补精度，提供高质量的数据，并提升数据驱动的窃电检测准确率，本文提出了一种基于两阶段自注意力权重融合(two-stage self-attention weight fusion, TS-SAWF)模型的用电缺失数据填补方法。TS-SAWF模型首先通过双阶段卷积增强对角掩码多头自注意力模块，学习用电数据间的特征与相关性，进行缺失数据预填补并获取相应的自注意力权重。然后，通过融合两阶段自注意力权重完成缺失数据的最终补全。为验证TS-SAWF模型填补缺失数据的有效性，本文基于中国国家电网公司公开数据集和国网福建某供电公司数据集进行仿真实验，并加入其他填补模型作为对比。实验结果表明，所提出的缺失数据填补方法具有较高的填补精度。此外，TS-SAWF模型填补后的数据有效提升了窃电检测模型的性能，进一步验证了所提方法的有效性。

1 相关知识

1.1 数据驱动的窃电检测

传统的窃电手段主要包括破坏电表和绕越电表等物理方式。以往的窃电检测通常需要电力公司派遣员工实地核查电表用电情况，这种方法不仅费

时费力,且效率较低^[26]。然而,近年来,随着智能电表的广泛普及,以及其与网络通信和数据管理系统协同组成的AMI,电力公司能够实现对用户用电信息的实时采集和监测。尽管AMI体系的应用在一定程度上降低了窃电的发生率,但窃电者仍在不断探索利用该系统的新方法,使得窃电手段变得愈发隐蔽和复杂。面对这些挑战,在AMI体系提供的海量用电信息的推动下,基于数据驱动的窃电检测研究逐渐成为备受关注的热门领域。通过应用人工智能算法,例如支持向量机^[27]、极端梯度提升(eXtreme gradient boosting, XGBoost)^[28]等机器学习模型,以及卷积神经网络(convolutional neural network, CNN)^[29-30]、LSTM^[31-33]等深度学习模型,可以从海量用电数据中自主学习特征,从而提高窃电检测的准确性和效率。总结该领域的研究可以发现,基于数据驱动的窃电检测流程大致分为3个阶段:用电信息采集、数据预处理和窃电检测。其流程图如附录图A1所示。

由附录图A1可以看出,在用电信息采集阶段,由于智能电表故障或人为蓄意破坏、数据传输线路

故障以及数据云端储存中心遭受网络攻击等问题^[34],可能会导致数据在采集、传输和储存各个环节中发生丢失。用电信息的缺失可能会引发以下问题:首先,人工智能算法模型无法直接处理含有缺失数据的数据集。其次,若缺失数据过多,会削弱数据间的相关性和特征,进而影响模型的训练效果和检测性能。因此,在数据预处理时,必须对缺失数据进行有效、准确的填补,以确保用电数据的高质量,从而提升后续窃电检测任务的性能。

1.2 填补模型联合优化训练策略

为了提升TS-SAWF用电缺失数据填补模型的补全精度,本文针对模型训练引入了一种联合优化训练策略^[24]。该策略包括掩码填补过程(masked imputation process, MIP)和观测重建过程(observed reconstruction process, ORP),图1展示了该训练策略的具体流程。在该策略中,填补模型可根据任务需求进行自行设计,例如以CNN或GAN作为基础模型设计填补模型。MIP旨在补全人为掩盖的数据,从而提高模型预测缺失值的准确性。而ORP旨在使模型尽可能准确地预测未被遮掩的数据。

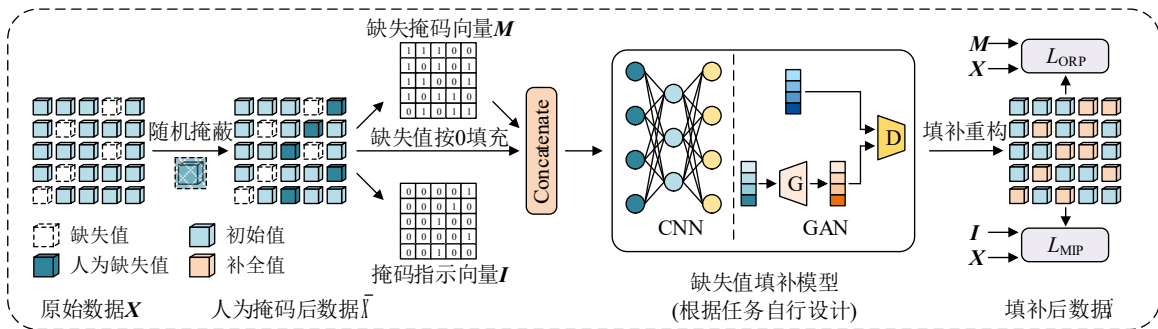


图1 联合优化训练策略流程图

Fig. 1 Flow chart of joint-optimization training strategy

在MIP中,部分初始值被随机掩盖,这些值对于填补模型而言是不可见的,相当于人为缺失值。原始用电数据记为 $X \in \mathbb{R}^{T \times D}$,其表达式如下所示:

$$X = \begin{bmatrix} x_1^1, x_1^2, \dots, x_1^d, \dots, x_1^D \\ x_2^1, x_2^2, \dots, x_2^d, \dots, x_2^D \\ \vdots \\ x_t^1, x_t^2, \dots, x_t^d, \dots, x_t^D \\ \vdots \\ x_T^1, x_T^2, \dots, x_T^d, \dots, x_T^D \end{bmatrix} \quad (1)$$

式中: T 为时间步长; D 为特征维度。

X 经过人为掩码后的数据记为 \bar{X} ,其对应的缺失掩码向量为 $M \in \mathbb{R}^{T \times D}$ 。完成填补后的补全数据记为 \tilde{X} 。为了区分人为缺失值和原始缺失值,该过程

还引入了人为掩码指示向量 $I \in \mathbb{R}^{T \times D}$ 。 M 和 I 的表达式如下所示:

$$M_t^d = \begin{cases} 1, & \text{如果}\bar{X}_t^d\text{非缺失值} \\ 0, & \text{如果}\bar{X}_t^d\text{为缺失值} \end{cases} \quad (2)$$

$$I_t^d = \begin{cases} 1, & \text{如果}\bar{X}_t^d\text{为人为掩盖} \\ 0, & \text{如果}\bar{X}_t^d\text{非人为掩盖} \end{cases} \quad (3)$$

为了评估模型填补人为缺失值的表现,本策略采用平均绝对误差(mean absolute error, MAE)作为损失函数,损失函数 L_{MAE} 如下所示:

$$L_{MAE}(X_{est}, X_{tar}, X_{mask}) =$$

$$\frac{\sum_{d=1}^D \sum_{t=1}^T |(X_{est} - X_{tar}) \odot X_{mask}|_t^d}{\sum_{d=1}^D \sum_{t=1}^T X_{mask,t}^d} \quad (4)$$

式中: X_{est} 为补全值; X_{tar} 为初始值; X_{mask} 为掩码向

量； \odot 表示矩阵哈达玛积运算。

根据 L_{MAE} ，MIP 损失函数 L_{MIP} 以及 ORP 损失函数 L_{ORP} 如下公式所示：

$$\begin{cases} L_{MIP} = L_{MAE}(\tilde{X}, X, I) \\ L_{ORP} = L_{MAE}(\tilde{X}, X, M) \end{cases} \quad (5)$$

在联合优化训练策略中，通过损失函数减小反向传播误差，尽可能缩小用电量数据实际值与预测值间的差距，从而有效提升模型补全缺失值能力。

2 TS-SAWF 缺失数据填补模型

2.1 卷积增强对角掩码多头自注意力机制

时序缺失数据填补本质上是回归任务，通过学习数据间的相关性和潜在特征对缺失值进行补全。自注意力机制不仅能克服 RNN 类、GAN 类等网络在训练效率和稳定性上的限制，又能对重要特征进行加权打分，这有助于提高缺失数据的填补质量和速度。通过联合优化训练策略可知，针对待补全的时间序列，本文对缺失值填充的初始值为 0。而对于缺失值，如果仅考虑当前位置的值，会使得提取的特征具有相似性，从而不能有效地计算出权重值。因此，为防止过多缺失值影响自注意力机制中权重分配的计算，本文提出了卷积增强对角掩码多头自注意力(convolution-enhanced diagonal mask multi-head self-attention, Conv-DMMHSA)机制，通过嵌入卷积层来捕捉每个特征周围的信息，以提高模型填补精度，其结构图如图 2 所示。

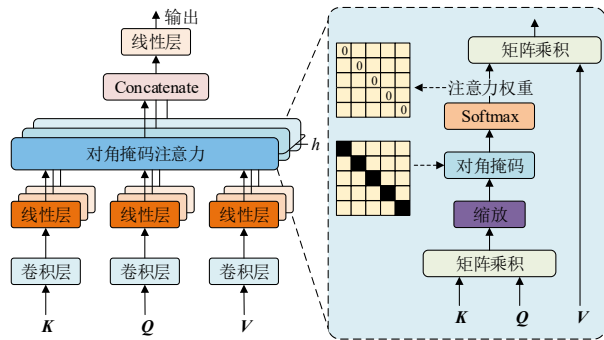


图 2 卷积增强对角掩码多头自注意力机制结构图

Fig. 2 Structure diagram of Conv-DMMHSA

在 Conv-DMMHSA 机制中，自注意力机制首先将用电量数据通过线性层映射为查询向量 Q 、键向量 K 和值向量 V ，3 个向量的维度分别分 d_q 、 d_k 、 d_v 。然后，利用缩放点积求得 Q 和 K^T 之间的注意力分数。最后，通过 Softmax 激活函数求得注意力权重。自注意力机制的表达式如下所示：

$$\text{self-attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

进一步，为了增强补全用电缺失数据的能力，在自注意力机制内部采用了对角掩码(diagonal mask, DM)策略。其中，注意力分数的对角线被设置为负无穷大，经过 Softmax 激活函数之后，注意力分数的对角线权重值趋近于 0。对角掩码自注意力(diagonal mask self-attention, DMSA)可以更好地捕捉用电量数据中的时间依赖性和用电特征，从而提高缺失数据填补精度。DM 公式如下所示：

$$\text{DM}(x)(i, j) = \begin{cases} -\infty, & i = j \\ x(i, j), & i \neq j \end{cases} \quad (7)$$

注意力分数经过 Softmax 函数激活之后，得到注意力权重 A_w ，将 A_w 与 V 相乘得到卷积增强对角掩码单头自注意力的输出值，进而可求得 Conv-DMMHSA 的输出值 Z ，其表达式如下所示：

$$\text{DMSA}(Q, K, V) = \text{Softmax}\left(\text{DM}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)V = A_w V \quad (8)$$

$$Y_i = \text{DMSA}(xW_i^Q, xW_i^K, xW_i^V) \quad (9)$$

$$Z = \text{Conv-DMMHSA}(x) = \text{Concat}(Y_1, Y_2, \dots, Y_n)W^O \quad (10)$$

式中： Y_i 表示 Conv-DMMHSA 中的第 i 个头的输出值； W_i^Q 、 W_i^K 、 W_i^V 为权重矩阵；Concat 表示矩阵拼接函数 Concatenate； W^O 为 Conv-DMMHSA 中输出层的线性变换。

2.2 TS-SAWF 填补模型

基于 SAITS 模型的启发和 Conv-DMMHSA 机制，本文提出了 TS-SAWF 缺失数据填补模型。该模型主要由两个 Conv-DMMHSA 模块以及权重融合模块组成，其整体结构图如附录图 A2 所示。

首先，待填补的用电量数据序列 X 与其对应的缺失值掩码向量 M 通过拼接操作后，输入至线性层映射之后进行位置编码(position encoding, PE)。PE 被用于建立时序位置关系，其由正弦函数和余弦函数表示，表达式如下所示：

$$\begin{cases} \text{PE}(\text{pos}, 2i) = \sin(\text{pos}/10000^{2i/d_{\text{model}}}) \\ \text{PE}(\text{pos}, 2i+1) = \cos(\text{pos}/10000^{2i/d_{\text{model}}}) \end{cases} \quad (11)$$

式中： pos 表示时间步长位置； i 表示维度索引； d_{model} 表示模型中向量维度大小。

完成 PE 之后得到输出结果 E_1 ， E_1 通过 Conv-DMMHSA 模块计算求得 Z_1 ， Z_1 输入至前馈网络(feed-forward network, FFN)计算求得输出 F_1 。FFN 是具有两层线性层的全连接网络，进行“线性-非线性-线性”过程计算。该过程旨在将输入的数据映射到更加高维空间中然后通过非线性计算，以增强 FFN 学习特征的能力，FFN 公式如下所示：

$$\text{FFN}(Z) = \text{ReLU}(ZW_1 + b_1)W_2 + b_2 \quad (12)$$

式中： ReLU 为激活函数； W_1 、 W_2 及 b_1 、 b_2 为全连

接层的权重参数和偏置参数。

紧接着, 经过线性层计算得到补全数据 X_o^1 。将 X 中的缺失值替换为 X_o^1 中对应的值, 得到补全数据 \hat{X} 。该过程的实现可以由下述公式概括:

$$\begin{cases} E_1 = [\text{Concat}(X, M)W_{E1} + b_{E1}] + p \\ F_1 = \{\text{FFN}(\text{Conv-DMMHSA}(E_1))\}^L \\ X_o^1 = F_1W_{F1} + b_{F1} \\ \hat{X} = M \odot X + (1 - M) \odot X_o^1 \end{cases} \quad (13)$$

式中: W_{E1} 、 W_{F1} 和 b_{E1} 、 b_{F1} 分别表示线性层的权重参数和偏置参数; p 表示位置编码; L 表示 Conv-DMMHSA 模块与 FFN 层运算次数。

同样的, 将 \hat{X} 与 M 拼接后, 依次经过 PE、Conv-DMMHSA 模块和 FFN 计算得到 F_2 , 然后经过双线性层得到补全数据 X_o^2 。该过程的实现同样可以由下述公式概括:

$$\begin{cases} E_2 = [\text{Concat}(\hat{X}, M)W_{E2} + b_{E2}] + p \\ F_2 = \{\text{FFN}(\text{Conv-DMMHSA}(E_2))\}^L \\ X_o^2 = \text{ReLU}(F_2W_{F2} + b_{F2})W_{\theta} + b_{\theta} \end{cases} \quad (14)$$

式中: E_2 为 PE 输出; W_{E2} 、 W_{F2} 、 W_{θ} 和 b_{E2} 、 b_{F2} 、 b_{θ} 分别表示线性层的权重和偏置参数。

为了进一步提高模型补全精度, 本文通过自注意力权重融合模块来获得更加准确的补全数据。该模块可以根据时间依赖性和缺失信息动态调整 X_o^1 和 X_o^2 的权重。首先, 求取 Conv-DMMHSA 模块二中最后一层多头注意力权重的输出平均值 \bar{A} :

$$\bar{A} = \frac{1}{h} \sum_{i=1}^h A_w(i) \quad (15)$$

进而, 拼接 \bar{A} 与 M , 求得融合权重系数 η 。基于融合权重系数 η 求得补全数据 X_o^3 , 然后将 X 中缺失值替换成 X_o^3 中相应值, 得到最终补全数据 \tilde{X} :

$$\begin{cases} \eta = \text{Sigmoid}[\text{Concat}(\bar{A}, M)W_{\eta} + b_{\eta}] \\ X_o^3 = (1 - \eta) \odot X_o^1 + \eta \odot X_o^2 \\ \tilde{X} = M \odot X + (1 - M) \odot X_o^3 \end{cases} \quad (16)$$

式中: W_{η} 和 b_{η} 表示线性层的权重参数和偏置参数。

2.3 TS-SAWF 填补模型损失函数

由于 TS-SAWF 缺失数据填补模型采用联合优化训练策略, 结合第 1 节联合优化训练策略损失函数, 可知 TS-SAWF 缺失数据填补模型 ORP 和 MIP 过程的损失函数表达式如下所示:

$$\begin{cases} L_{\text{ORP}} = \frac{1}{3}(L_{\text{MAE}}(X_o^1, X, M) + L_{\text{MAE}}(X_o^2, X, M) + \\ L_{\text{MAE}}(X_o^3, X, M)) \\ L_{\text{MIP}} = L_{\text{MAE}}(\bar{X}, X, I) \end{cases} \quad (18)$$

因此, TS-SAWF 填补模型总损失表达式为

$$L_{\text{TOTAL}} = L_{\text{ORP}} + \lambda L_{\text{MIP}} \quad (19)$$

式中: λ 为权重调整系数。

3 算例分析

在本节中, 采用两个真实数据集对 TS-SAWF 缺失数据填补模型的有效性进行验证。数据集 A 为中国国家电网公司公布的用户电表真实采集数据, 涵盖了 4 万多名用户从 2014 年 1 月 1 日至 2016 年 10 月 31 日共 1034 天的日用电量数据^[30]; 数据集 B 为国网福建某供电公司的用户电表采集数据, 包含了 3 千多名用户 2023 年 11 月 1 日至 2024 年 5 月 20 日共 202 天的日用电量数据。

本节所涉及的模型均基于 Python、Pytorch 和 Pypots 框架搭建, 附录表 A1 列出了仿真实验环境所需的软硬件配置信息。

在模型训练过程中, 采用联合优化训练策略, 提高模型填补精度。并且引入早停机制, 旨在提高模型训练效率并防止模型过拟合。模型的训练回合数被设置为 500, 早停机制中的耐心系数取值为 10, 即当损失值连续 10 个训练回合未下降时, 模型停止训练并保存最优模型。模型的输入为包含缺失数据的用电量数据序列及其对应的缺失掩码向量, 输入维度为(批尺寸, 特征量, 时间步长)。数据集 A 和 B 的时间步长分别为 1034 天和 202 天, 且两个数据集的每个时间步的特征量均为日用电量, 即特征量为 1。因此, 数据集 A 的模型输入维度为(批尺寸, 1, 1034), 数据集 B 的模型输入维度为(批尺寸, 1, 202)。为了全面评估模型的填补性能, 本文选择 MAE 和均方根误差(root mean squared error, RMSE) 作为模型评价指标, 单位均为 kW·h。其中, RMSE 计算公式如下所示:

$$\text{RMSE}(X_{\text{est}}, X_{\text{tar}}, X_{\text{mask}}) = \sqrt{\frac{\sum_{d=1}^D \sum_{t=1}^T \{[(X_{\text{est}} - X_{\text{tar}}) \odot X_{\text{mask}}]^2\}_t^d}{\sum_{d=1}^D \sum_{t=1}^T X_{\text{mask}t}^d}} \quad (20)$$

3.1 模型超参数选择与设置

在模型训练中, 优化器的选择和配置会显著影响模型的收敛速度、稳定性和填补性能。此外, 批尺寸也会对训练效果产生影响。较小的批尺寸可以提高填补精度, 但也可能导致训练时间过长, 影响模型的收敛。相反, 较大的批尺寸可以提升训练效率, 但也可能减弱模型的泛化能力。因此, 合适的优化器和批尺寸对模型最终的填补性能至关重要。

为了确定 TS-SAWF 填补模型的最佳优化器和

批尺寸, 本文进行了不同优化器在不同批尺寸下的仿真实验。实验比较了 5 种常用优化器, 包括 Adam、AdamW、Adagrad、RMSprop 和 SGD。同时, 在每种优化器下, 观察不同批尺寸设置对于模型填补性能的影响。由于实验硬件平台内存的限制, 实验中最大批尺寸为 64, 设置了 16、32 和 64 共 5 种不同的批尺寸。所有优化器的初始学习率设置为 0.0001, 其他参数使用默认配置。实验随机选取数据集 A 中 10000 个用户 1034 天的日用电量数据进行缺失值填补, 以对比不同优化器在各批尺寸下的训练表现, MAE 和 RMSE 的结果如图 3 所示。

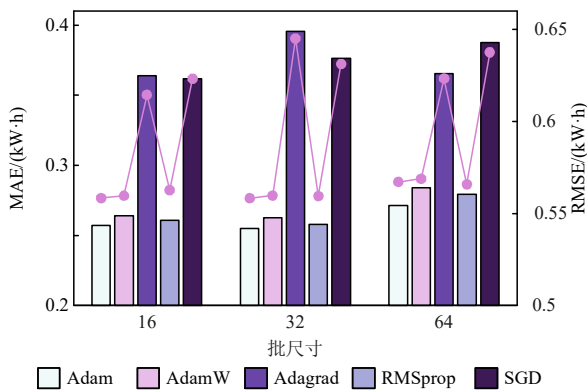


图 3 不同优化器在各批尺寸下的训练表现

Fig. 3 Training performance of different optimizers across various batch sizes

由图 3 可以看出, Adam 优化器表现最佳, 其次是 RMSprop 和 AdamW。特别是在批尺寸为 32 时, Adam 优化器取得最佳效果, MAE 为 0.2765kW·h, RMSE 为 0.5418kW·h。基于以上对比分析, 本文选择 Adam 作为 TS-SAWF 填补模型的优化器, 并将批尺寸设置为 32, 以确保最佳的填补性能和训练效率。

此外, Conv-DMMHSA 模块中的卷积核大小也会影响模型的最终填补性能。因此, 本文设置了 1、3、5 和 7 四种卷积核大小, 用于探索其对模型填补性能的影响。不同卷积核大小下模型的评价指标表现如图 4 所示。这些实验结果将有助于优化卷积核配置, 从而进一步提升模型填补精度和整体性能。

从图 4 可以看出, 适当的卷积核大小能够更有效地捕捉用电数据中的特征以提升模型的填补精度。但过大的卷积核可能会导致模型过拟合, 降低填补性能。当卷积核大小为 3 时, MAE 和 RMSE 均达到最低值, 分别为 0.2765 和 0.5418kW·h, 此时模型的填补性能最佳。因此, 本文将 Conv-DMMHSA 模块中的卷积核尺寸设置为 3。

权重调整系数 λ 用于平衡 ORP 损失值 L_{ORP} 与

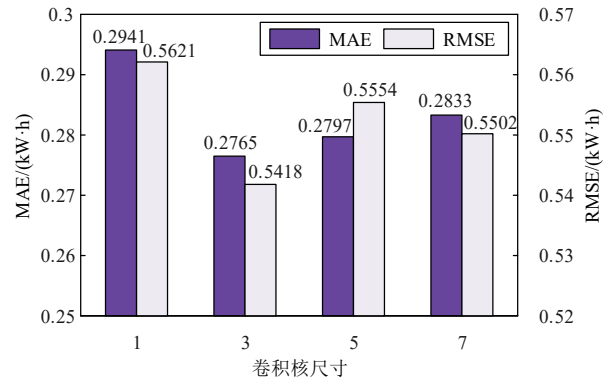


图 4 卷积核大小对于评价指标的影响

Fig. 4 Effect of the size of the convolution kernel on the evaluation metric

MIP 损失值 L_{MIP} 的优化权重, 其取值直接影响填补模型对两种过程的侧重, 因此对模型填补性能至关重要。为全面分析 λ 对模型填补性能的影响, 本节在数据集 A 上进行了实验, 不同权重调整系数 λ 下填补模型的性能表现如表 1 所示。

表 1 不同权重调整系数 λ 下填补模型的性能表现

Table 1 Performance of the imputation model under different weight adjustment coefficients λ

λ	MAE	RMSE
0.5	0.2812	0.5523
1.0	0.2765	0.5418
1.5	0.2801	0.5492
2.0	0.2847	0.5601

由表 1 可以得出, 当权重调整系数 λ 取值为 1 时, MAE 和 RMSE 分别取得最低值 0.2765kW·h 和 0.5418kW·h, 表明模型填补性能最优, 验证了其在平衡 ORP 与 MIP 目标方面的合理性。当 $\lambda < 1$ 时, 填补模型更侧重保留实际观测数据的准确性, 可能提升对已知数据的拟合效果。然而, 若 λ 过小, 模型的缺失数据填补能力可能受限, 尤其在连续缺失场景下表现较差。当 $\lambda > 1$ 时, 填补模型更关注补全人为掩盖的缺失值, 可能提升模型对缺失值的填补能力。但若 λ 过大, 则可能削弱对实际观测数据的拟合能力, 降低泛化性能。综合实验结果以及文献[24]的建议, 本文将 λ 取值为 1, 以获得更小的训练总损失值并进一步提升模型填补性能。

本文通过多次的模型训练, 基于最优模型对应的参数设置 TS-SAWF 缺失数据填补模型。具体参数设置见附录表 A2。

3.2 填补模型性能对比及分析

为验证 TS-SAWF 模型缺失数据填补的有效性, 本文选择均值(Mean)填补、中位数(Median)填补、末次观测值填补(last observation carried forward, LOCF)、双向循环时序填补模型(bidirectional recurrent imputation for time series,

BRITS^[19]、半监督生成对抗网络(semi-supervised generative adversarial network, SSGAN)^[20]和 SAITS 共 6 种填补方法作为对比实验模型。

实验随机选取数据集 A 中 10000 名用户 1034 天的日用电量数据和数据集 B 中 2000 名用户 202 天的日用电量数据,用于评估模型的填补性能。所有模型均在相同的训练环境和数据集上进行训练,初始人为随机缺失率设定为 10%。图 5 展示了各填补模型在两个数据集上的评价指标表现。

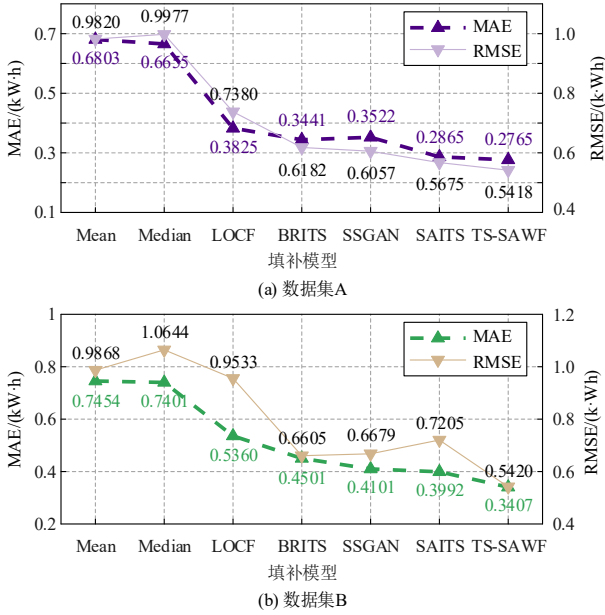


图 5 填补模型在 10%人为随机缺失率下的评价指标表现
Fig. 5 Evaluation metrics of the imputation models at the 10% artificial random missing rate

从图 5 可以明显看出,在两个数据集上,本文所提的填补模型在各项评价指标上均优于其他对比模型,其次为 SAITS。在数据集 A 中,TS-SAWF 模型的 MAE 比 Mean 和 SAITS 分别提升了 59.4% 和 3.49%,RMSE 相较于 Median 和 SAITS 分别提升了 45.7%和 4.53%。在数据集 B 中,TS-SAWF 模型的 MAE 相较于 BRITS、SSGAN、SAITS 分别提升了 24.3%、16.9%和 14.7%,RMSE 相较于 SAITS、

SSGAN、BRITS 分别提升了 24.8%、18.9%和 17.9%。此外,为了直观展示 TS-SAWF 模型填补缺失数据的效果,本文在 10%人为随机缺失率下,随机选取了两个数据集中一名用户的日用电量数据,并绘制其原始数据与填补后的数据对比图,如图 6 所示。

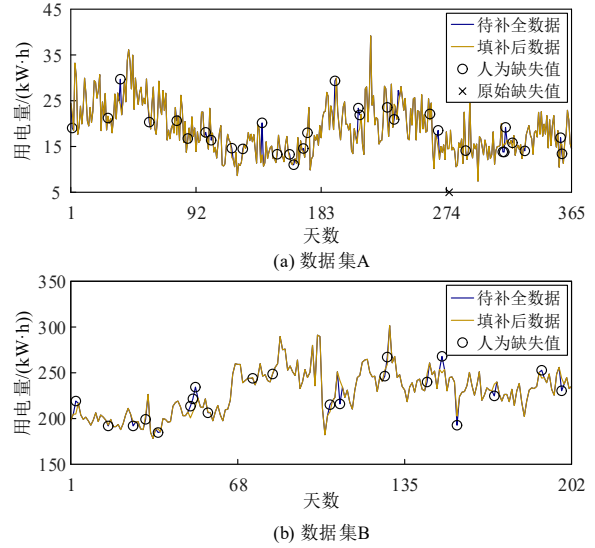


图 6 TS-SAWF 模型的填补效果
Fig. 6 Imputation performance of the TS-SAWF model

由图 6 可以看出,TS-SAWF 模型填补后的数据在整体趋势和局部细节上与待补全数据高度一致。然而,在个别点上,填补结果与真实值之间仍存在一定的误差,这种误差通常是由复杂的时序波动或极端数据点引起的。少量的局部误差并不显著影响模型的全局填补效果,这表明 TS-SAWF 模型具有较强的鲁棒性和有效性。

为进一步评估各填补模型在不同人为随机缺失率下的表现,本文设置了 30%、50%、70%和 90%的缺失率,旨在验证模型在人为随机缺失率增加时是否仍能保持较好的性能。由于 Mean 填补和 Median 填补方法的误差较大,所以这两种方法未被纳入人为随机缺失率的进一步对比。各模型在不同人为随机缺失率下的评价指标表现由表 2 列出。

表 2 填补模型在不同人为随机缺失率下的评价指标表现

Table 2 Performance of the evaluation metrics of the imputation model under different artificial random missing rates

数据集	填补模型	不同人为随机缺失率下的评价指标表现							
		30%		50%		70%		90%	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
数据集 A	LOCF	0.3960	0.7627	0.4157	0.8000	0.4458	0.8428	0.5223	0.9577
	BRITS	0.3382	0.6178	0.3825	0.6603	0.4282	0.7266	0.4668	0.7855
	SSGAN	0.4134	0.6536	0.3998	0.6585	0.3919	0.6817	0.5068	0.8309
	SAITS	0.2976	0.5793	0.3124	0.6067	0.3322	0.6377	0.3816	0.7032
	TS-SAWF	0.2862	0.5578	0.3091	0.5946	0.3294	0.6259	0.3800	0.7010
数据集 B	LOCF	0.4673	0.8137	0.5102	0.8640	0.5602	0.9528	0.7044	1.1528
	BRITS	0.4476	0.7206	0.5037	0.7404	0.5460	0.7779	0.6956	0.9317
	SSGAN	0.4698	0.6759	0.4786	0.7248	0.5594	0.7795	0.6984	0.9364
	SAITS	0.4339	0.6848	0.4523	0.7327	0.5391	0.8596	0.6881	0.9513
	TS-SAWF	0.3847	0.6465	0.4384	0.7022	0.5276	0.7858	0.6792	0.9339

从表 2 可以看出,随着缺失率的增加,各模型的填补误差均有所上升。在数据集 A 中,TS-SAWF 模型在所有人为随机缺失率下的 MAE 和 RMSE 均表现最佳。在 90%人为随机缺失率的情况下,该模型的 MAE 相比于 LOCF 和 SAITS 分别提升了 27.2%和 0.42%,RMSE 相比于 LOCF 和 SAITS 分别提升了 26.8%和 0.31%。在数据集 B 中,TS-SAWF 模型在所有缺失率下的 MAE 均为最佳,而在 70%和 90%缺失率下,其在 RMSE 上的表现略逊于 BRITS 和 SSGAN。

由于智能电表故障、数据传输路线故障或数据云端储存中心遭受网络攻击等原因,无法在短时间内恢复,因此可能会导致用电数据出现连续缺失的情况。为了验证本文模型对于填补连续缺失用电数据方面的有效性,本文在 10%人为随机缺失率的基础上,设置了用电数据连续缺失 3 天、5 天、7 天和 15 天的 4 种场景。图 7 展示了各填补模型在不同连续缺失天数下的性能表现。

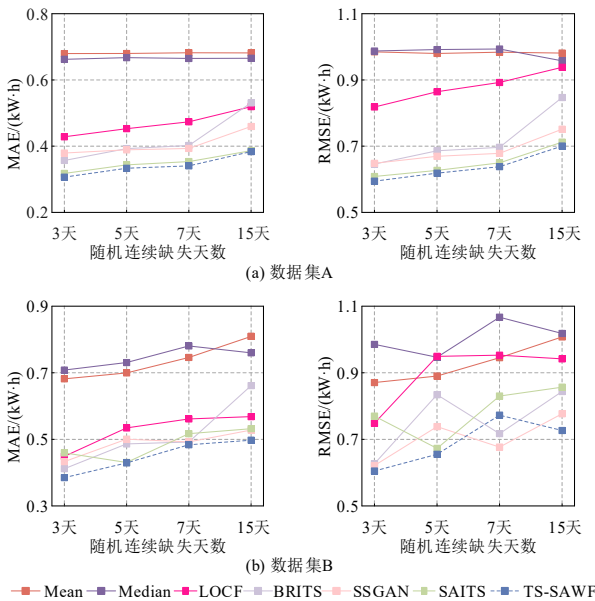


图 7 填补模型随机连续缺失天数下的评价指标表现
Fig. 7 Evaluation metric performance of the imputation model under random continuous missing days

通过图 7 可以看出,传统填补方法的补全精度较低,并且对于连续缺失天数的增加较为敏感,这表明它们对于缺失数据的填补能力有限。相比之下,基于深度学习的填补方法在补全精度上具有更好的表现,且本文所提的 TS-SAWF 模型在两个数据集上的各项评价指标表现最为优异。

为了直观体现 TS-SAWF 模型在连续缺失数据填补中的效果,本文随机选取两个数据集中一名用户的日用电量数据,绘制其原始数据与填补后的数据对比图,如图 8 所示。

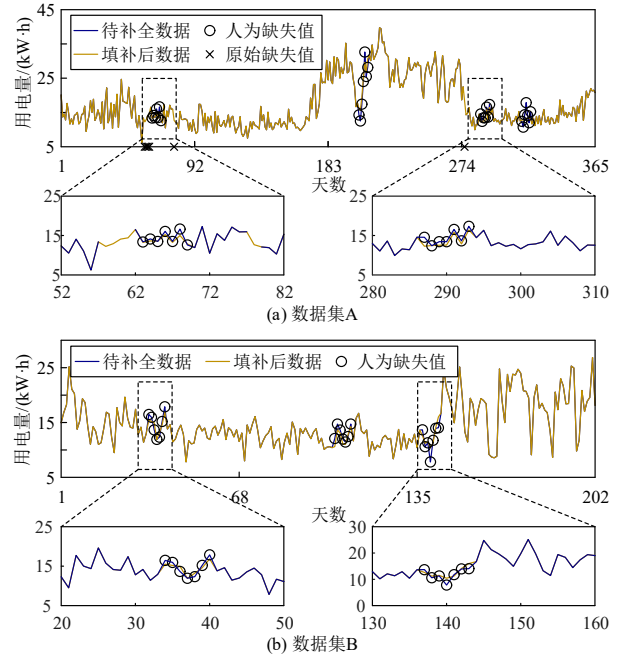


图 8 TS-SAWF 模型在连续缺失 7 天数据下的填补效果
Fig. 8 Imputation performance of the TS-SAWF model under 7 consecutive days of missing data

从图 8 可以看出,尽管在较高连续缺失天数的场景下,本文所提模型仍能有效地拟合缺失的数据。这表明该模型能够充分学习历史及未来的用电数据特征,从而实现有效的缺失数据填补。

综上所述可知,相比较于传统的填补方法,基于深度学习的填补方法在各项指标上表现优异。这表明,深度学习模型能够更有效地学习用电量数据序列之间的相关性与潜在的用电特征,从而显著提升模型的填补性能。在深度学习模型中,本文所提出的 TS-SAWF 缺失数据填补模型在不同数据集的人为随机缺失率和连续缺失天数的场景下均展现出优秀的填补性能,具有良好的鲁棒性和泛化能力。SAITS 模型紧随其后,也表现出稳定的填补性能。这得益于 MIP 和 ORP 联合优化训练策略的协同作用,提升了模型在人为随机缺失和连续缺失场景下的鲁棒性和泛化能力。进一步分析 TS-SAWF 模型与 SAITS 模型的指标表现可以发现,在对角掩码多头自注意力机制中嵌入卷积层,有助于模型更好地捕捉用电特征与相关性。通过权重融合模块,动态调整两阶段 Conv-DMMHSA 模块所学习的自注意力权重,优化填补权重分配,从而增强 TS-SAWF 模型的填补精度,实现更优的补全效果。

3.3 数据补全对窃电检测性能提升验证

为验证 TS-SAWF 模型填补后的用电量数据对于窃电检测模型性能提升的有效性,本文选择了 XGBoost^[28]、HCED^[29]、WDCNN^[30]、CNN-LSTM^[31]、

CAEs-LSTM^[32]和 LSTM-TCN-DCNN^[33]6 种窃电检测模型,分别在 60%、70%和 80%三种训练集比例下进行可行性实验验证。所有模型的参数设置均遵循相关文献的要求,实验使用数据集 A 进行仿真。数据预处理部分,包括异常值处理、归一化处理及数据类不平衡处理,方法与文献[33]一致。本文采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值(F1-Score)、ROC 曲线下面积(AUC)、误检率(false positive rate, FPR)和检出率(true positive rate, TPR)作为窃电检测模型的评价指标。通过比较窃电检测模型在不同缺失值填补方法下的指标表现,验证 TS-SAWF 模型对于窃电检测模型性能提升的有效性。所比较的缺失值填补方法包括插值填补法和 TS-SAWF 模型填补法,在两种缺失值填补方法下,各窃电检测模型在不同训练集比例下的评价指标表现如附录图 A3 所示。

由附录图 A3 可以看出,经过 TS-SAWF 模型填补后数据训练的窃电检测模型,在不同的训练集比例下的各项指标均具有最优异的表现,这也间接证明了 TS-SAWF 模型填补缺失数据对窃电检测模型性能提升方面的有效性。为了更直观地展示 TS-SAWF 模型填补法和插值填补法在性能提升上的差异,两者指标间的差值用 S 表示,以表格形式呈现,如附录表 A3 所示。

结合附录图 A3 和附录表 A3 可知,与传统插值法相比,TS-SAWF 模型填补后的用电量数据有效提升了窃电检测模型在不同训练集比例下的各项评价指标表现。为进一步验证 TS-SAWF 填补模型的有效性,本文选取 Accuracy 和 FPR 两项关键指标进行量化分析。其中,Accuracy 反映了窃电检测模型的总体分类正确率,而 FPR 表示的是正常用户被误报为窃电用户占全部正常用户的比例。从附录表 A3 可以得出,TS-SAWF 模型填补方法对于 CNN-LSTM 窃电检测模型的性能提升效果最为显著,Accuracy 平均提升了 11.13%,FPR 平均降低了 11.93%。即使是在训练数据量较少的情况下,Accuracy 提升了 13.3%,FPR 降低了 15.4%。此外,当训练集比例为 80%时,该填补方法对其它 5 种窃电检测模型的 Accuracy 平均提升了 0.88%,FPR 平均降低了 1.1%。即使是在训练集比例为 60%时,Accuracy 平均提升了 0.74%,FPR 平均降低了 1.0%。综上所述,相较于传统插值法,TS-SAWF 模型依托两阶段 Conv-DMMHSA 模块和权重融合模块,能够更好地捕捉用电特征和相关性,并通过 MIP 和 ORP 联合优化训练,显著提升了用电数据的补全质

量,从而有效提高了窃电检测模型的性能。此外,该结果表明,不同的缺失数据填补方法会影响窃电检测模型的性能,也间接验证了 TS-SAWF 模型在缺失数据填补中的有效性。

4 结论

由于 AMI 体系在采集、传输和存储各环节中存在用电信息数据缺失的风险,从而影响数据质量并对窃电检测任务产生负面影响。为解决这一问题,本文提出了一种基于两阶段自注意力权重融合模型的窃电检测缺失数据填补方法。通过对中国国家电网的两个真实数据集进行仿真实验,结果表明,TS-SAWF 模型在 Conv-DMMHSA 模块的帮助下,能够在不同数据集的随机缺失率和连续缺失天数下表现出色,展现出较强的鲁棒性和良好的泛化能力。此外,TS-SAWF 模型填补后的用电数据显著提升了窃电检测模型的性能,进一步验证了其有效性。值得一提的是,该方法不仅在窃电检测缺失数据填补中表现出色,还可以广泛应用于负荷预测、电力设备健康监测以及故障检测等数据驱动场景中的缺失值填补任务。

窃电行为通常伴随窃电者故意隐藏或篡改用电数据,这种行为可能会对数据缺失填补过程造成对抗性干扰,并为数据填补和窃电检测带来更大挑战。虽然本文的工作聚焦于常规缺失数据场景下的填补性能,但未来研究将重点解决窃电者主动干扰数据修复的对抗性问题,从而为用电缺失数据填补和窃电检测提供更加坚实的技术支持。

附录见本刊网络版(<http://www.dwjs.com.cn/CN/1000-3673/current.shtml>)。

参考文献

- [1] XIA Xiaofang, XIAO Yang, LIANG Wei, et al. Detection methods in smart meters for electricity thefts: a survey[J]. Proceedings of the IEEE, 2022, 110(2): 273-319.
- [2] YANG Kaixiang, CHEN Wuxing, BI Jichao, et al. Multi-view broad learning system for electricity theft detection[J]. Applied Energy, 2023, 352: 121914.
- [3] 陈启鑫, 郑可迪, 康重庆, 等. 异常用电的检测方法: 评述与展望[J]. 电力系统自动化, 2018, 42(17): 189-199.
CHEN Qixin, ZHENG Kedi, KANG Chongqing, et al. Detection methods of abnormal electricity consumption behaviors: review and prospect[J]. Automation of Electric Power Systems, 2018, 42(17): 189-199(in Chinese).
- [4] 林振智, 崔雪原, 金伟超, 等. 用户侧窃电检测关键技术[J]. 电力系统自动化, 2022, 46(5): 188-199.
LIN Zhenzhi, CUI Xueyuan, JIN Weichao, et al. Key technologies of electricity theft detection at consumer side[J]. Automation of Electric Power Systems, 2022, 46(5): 188-199(in Chinese).

- [5] 游文霞, 梁皓, 杨楠, 等. 基于重采样和混合集成学习的不平衡窃电检测[J]. 电网技术, 2024, 48(2): 730-739.
YOU Wenxia, LIANG Hao, YANG Nan, et al. Class imbalanced electricity theft detection based on resampling and hybrid ensemble learning[J]. Power System Technology, 2024, 48(2): 730-739(in Chinese).
- [6] 金晟, 苏盛, 薛阳, 等. 数据驱动窃电检测方法综述与低误报率研究展望[J]. 电力系统自动化, 2022, 46(1): 3-14.
JIN Sheng, SU Sheng, XUE Yang, et al. Review on data-driven based electricity theft detection method and research prospect for low false positive rate[J]. Automation of Electric Power Systems, 2022, 46(1): 3-14(in Chinese).
- [7] 杨挺, 孙兆帅, 季浩, 等. 基于矩阵范数优化理论的用电数据质量提升算法[J]. 中国电机工程学报, 2022, 42(10): 3501-3511.
YANG Ting, SUN Zhaoshuai, JI Hao, et al. Electricity consumption data quality improvement algorithm based on matrix norm optimization theory[J]. Proceedings of the CSEE, 2022, 42(10): 3501-3511(in Chinese).
- [8] KHAN I U, JAVEID N, TAYLOR C J, et al. A stacked machine and deep learning-based approach for analysing electricity theft in smart grids[J]. IEEE Transactions on Smart Grid, 2022, 13(2): 1633-1644.
- [9] BUZAU M M, TEJEDOR-AGUILERA J, CRUZ-ROMERO P, et al. Hybrid deep neural networks for detection of non-technical losses in electricity smart meters[J]. IEEE Transactions on Power Systems, 2020, 35(2): 1254-1263.
- [10] KHAN I U, JAVAID N, TAYLOR C J, et al. Robust data driven analysis for electricity theft attack-resilient power grid[J]. IEEE Transactions on Power Systems, 2023, 38(1): 537-548.
- [11] CUI Lei, GUO Lei, GAO Longxiang, et al. A covert electricity-theft cyberattack against machine learning-based detection models[J]. IEEE Transactions on Industrial Informatics, 2022, 18(11): 7824-7833.
- [12] 刘科研, 周方泽, 周晖, 等. 基于改进生成对抗网络的台区采集数据修复[J]. 电网技术, 2022, 46(8): 3231-3239.
LIU Keyan, ZHOU Fangze, ZHOU Hui, et al. Missing data imputation in transformer district based on improved generative adversarial network[J]. Power System Technology, 2022, 46(8): 3231-3239(in Chinese).
- [13] 杨挺, 叶芷杉, 徐嘉成, 等. 基于低秩张量补全的非侵入式负荷监测缺失数据修复方法[J]. 电网技术, 2024, 48(1): 394-404.
YANG Ting, YE Zhishan, XU Jiacheng, et al. Non-intrusive load monitoring missing data recovery based on low-rank tensor completion[J]. Power System Technology, 2024, 48(1): 394-404(in Chinese).
- [14] 金涛, 王万豪, 黄钦瑜, 等. 一种利用轻量级多通道注意力融合网络的窃电检测边缘计算方法[J/OL]. 中国电机工程学报, 2025: 1-13[2025-01-26]. <http://kns.cnki.net/kcms/detail/11.2107.TM.20250123.1535.010.html>.
JIN Tao, WANG Wanhao, HUANG Qinyu, et al. An edge computing approach for electricity theft detection utilizing lightweight multi-channel attention fusion network[J/OL]. Proceedings of the CSEE, 2025: 1-13[2025-01-26]. <http://kns.cnki.net/kcms/detail/11.2107.TM.20250123.1535.010.html>(in Chinese).
- [15] 崔阳阳, 赵洪山, 曲岳晗, 等. 基于残差 U 型网络的低压台区电力缺失数据补全方法[J]. 电力系统自动化, 2022, 46(9): 83-90.
CUI Yangyang, ZHAO Hongshan, QU Yuehan, et al. Residual U-Net based complementation method for missing electricity data of low-voltage stations[J]. Automation of Electric Power Systems, 2022, 46(9): 83-90(in Chinese).
- [16] 冯磊, 王石刚, 梁庆华. 基于 GAKNN 方法的配电站时间序列缺失数据补全方法[J]. 电力自动化设备, 2021, 41(12): 187-192.
FENG Lei, WANG Shigang, LIANG Qinghua. Completion method for missing time series data of distribution station based on GAKNN method[J]. Electric Power Automation Equipment, 2021, 41(12): 187-192(in Chinese).
- [17] LU Chengbo, MEI Ying. An imputation method for missing data based on an extreme learning machine auto-encoder[J]. IEEE Access, 2018, 6: 52930-52935.
- [18] 梅玉杰, 李勇, 周王峰, 等. 基于机器学习的配电网异常缺失数据动态清洗方法[J]. 电力系统保护与控制, 2023, 51(7): 158-169.
MEI Yujie, LI Yong, ZHOU Wangfeng, et al. Dynamic data cleaning method of abnormal and missing data in a distribution network based on machine learning[J]. Power System Protection and Control, 2023, 51(7): 158-169(in Chinese).
- [19] CAO Wei, WANG Dong, LI Jian, et al. BRITS: bidirectional recurrent imputation for time series[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal: Bytedance AI Lab, 2018.
- [20] MIAO Xiaoye, WU Yangyang, WANG Jun, et al. Generative semi-supervised learning for multivariate time series imputation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2021.
- [21] 廖若愚, 刘友波, 沈晓东, 等. 基于双向循环插补网络的分布式光伏集群时序数据耦合增强方法[J]. 电网技术, 2024, 48(7): 2784-2794.
LIAO Ruoyu, LIU Youbo, SHEN Xiaodong, et al. Time series data coupling enhancement method of distributed photovoltaic cluster based on bidirectional recurrent imputation network[J]. Power System Technology, 2024, 48(7): 2784-2794(in Chinese).
- [22] 张科鑫, 窦晓波, 李炜祺, 等. 基于量测数据补全的有源配电网电压优化技术[J]. 电力自动化设备, 2023, 43(11): 67-74.
ZHANG Kexin, DOU Xiaobo, LI Weiqi, et al. Voltage optimization technology of active distribution network based on measurement data completion[J]. Electric Power Automation Equipment, 2023, 43(11): 67-74(in Chinese).
- [23] 刘志坚, 陶韵旭, 刘航, 等. 基于双通道生成对抗网络的城市用电负荷缺失数据补全方法[J]. 电力系统自动化, 2024, 48(17): 161-170.
LIU Zhijian, TAO Yunxu, LIU Hang, et al. Completion method for missing urban power load data based on double-channel generative adversarial networks[J]. Automation of Electric Power Systems, 2024, 48(17): 161-170(in Chinese).
- [24] DU Wenjie, CÔTÉ D, LIU Yan. SAITS: self-attention-based imputation for time series[J]. Expert Systems with Applications, 2023, 219: 119619.
- [25] NAYAK S, DWIVEDID, BABU K V S M, et al. Data imputation using self attention based model for enhancing distribution grid monitoring and protection systems[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 2510711.
- [26] 李云峰, 高云鹏, 张蓬鹤, 等. 多目标优化时域卷积神经网络的窃电行为高准确检测算法[J]. 电网技术, 2024, 48(8): 3449-3458.
LI Yunfeng, GAO Yunpeng, ZHANG Penghe, et al. High-accuracy detection algorithm of electricity theft behavior based on multi-objective optimization time-domain convolutional neural network[J]. Power System Technology, 2024, 48(8): 3449-3458(in Chinese).

- [27] JOKAR P, ARIANPOO N, LEUNG V C M. Electricity theft detection in AMI using customers' consumption patterns[J]. IEEE Transactions on Smart Grid, 2016, 7(1): 216-226.
- [28] YAN Zhongzong, WEN He. Electricity theft detection base on extreme gradient boosting in AMI[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 2504909.
- [29] 周赣, 华济民, 李铭钧, 等. 基于图转换和混合卷积神经网络的窃电检测方法[J]. 电力系统自动化, 2022, 46(19): 78-86.
ZHOU Gan, HUA Jimin, LI Mingjun, et al. Electricity theft detection method based on graph transformation and hybrid convolutional neural network[J]. Automation of Electric Power Systems, 2022, 46(19): 78-86(in Chinese).
- [30] ZHENG Zhibin, YANG Yatao, NIU Xiangdong, et al. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids[J]. IEEE Transactions on Industrial Informatics, 2018, 14(4): 1606-1615.
- [31] HASAN M N, TOMA R N, NAHID A A, et al. Electricity theft detection in smart grid systems: a CNN-LSTM based approach[J]. Energies, 2019, 12(17): 3310.
- [32] 董立红, 肖纯朗, 叶鸥, 等. 一种基于 CAEs-LSTM 融合模型的窃电检测方法[J]. 电力系统保护与控制, 2022, 50(21): 118-127.
DONG Lihong, XIAO Chunlang, YE Ou, et al. Electricity theft detection method based on a CAEs-LSTM fusion model[J]. Power System Protection and Control, 2022, 50(21): 118-127(in Chinese).
- [33] HUANG Qinyu, TANG Zhenli, WENG Xiaofeng, et al. A novel electricity theft detection strategy based on dual-time feature fusion and deep learning methods[J]. Energies, 2024, 17(2): 275.
- [34] 刘增稷, 王琦, 薛彤, 等. 电力系统中数据驱动算法安全威胁分

析及应对方法研究[J]. 中国电机工程学报, 2023, 43(12): 4538-4553.
LIU Zengji, WANG Qi, XUE Tong, et al. Research on security risks and defense methods of data-driven algorithms in power systems[J]. Proceedings of the CSEE, 2023, 43(12): 4538-4553(in Chinese).



黄钦瑜

在线出版日期: 2025-04-16。

收稿日期: 2024-11-28。

作者简介:

黄钦瑜(1998), 男, 硕士研究生, 研究方向为人工智能在电力系统中的应用, E-mail: 220127039@fzu.edu.cn;

杨明发(1977), 男, 博士, 副教授, 研究方向为电力电子与电力传动、电机控制系统等, E-mail: yangmf@fzu.edu.cn;

王万豪(1999), 男, 硕士研究生, 研究方向为人工智能在电力系统中的应用, E-mail: 220127097@fzu.edu.cn;

刘宇龙(1992), 男, 助理研究员, 研究方向为电气信号检测技术及电能质量扰动分析研究, E-mail: liuyul@pku.edu.cn;

陈焯滨(2000), 男, 硕士研究生, 研究方向为电能质量扰动分析研究, E-mail: 220120046@fzu.edu.cn;

金涛(1976), 男, 教授, 博士生导师, 通信作者, 研究方向为电力系统稳定性分析、在线测量与信号处理及新能源技术, E-mail: jintly@fzu.edu.cn。

(责任编辑 李健一)

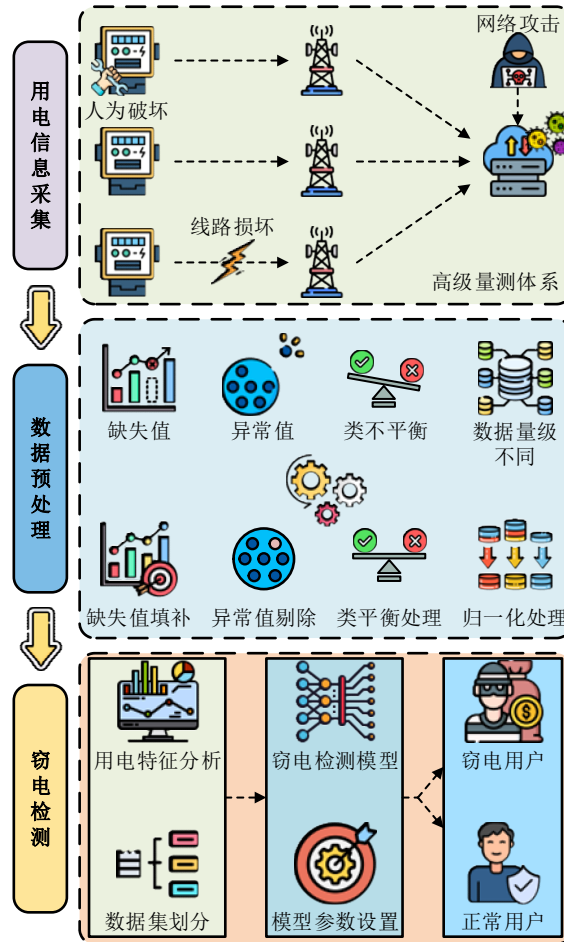


图 A1 基于数据驱动的窃电检测方法流程图

Fig. A1 Flow chart of data-driven electricity theft detection method

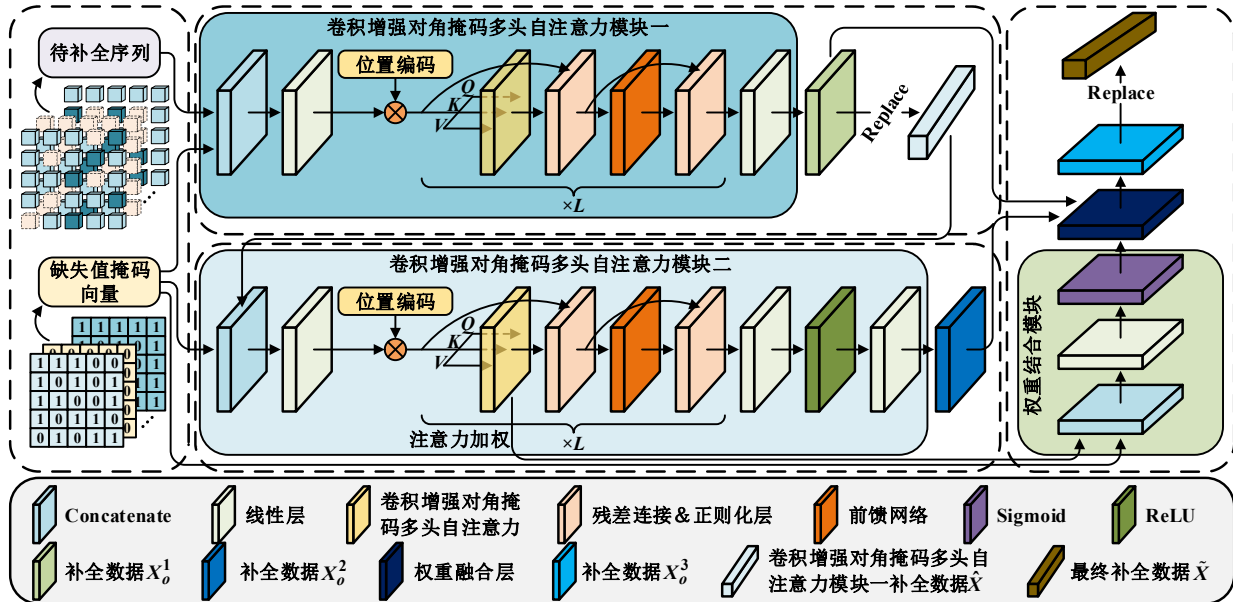
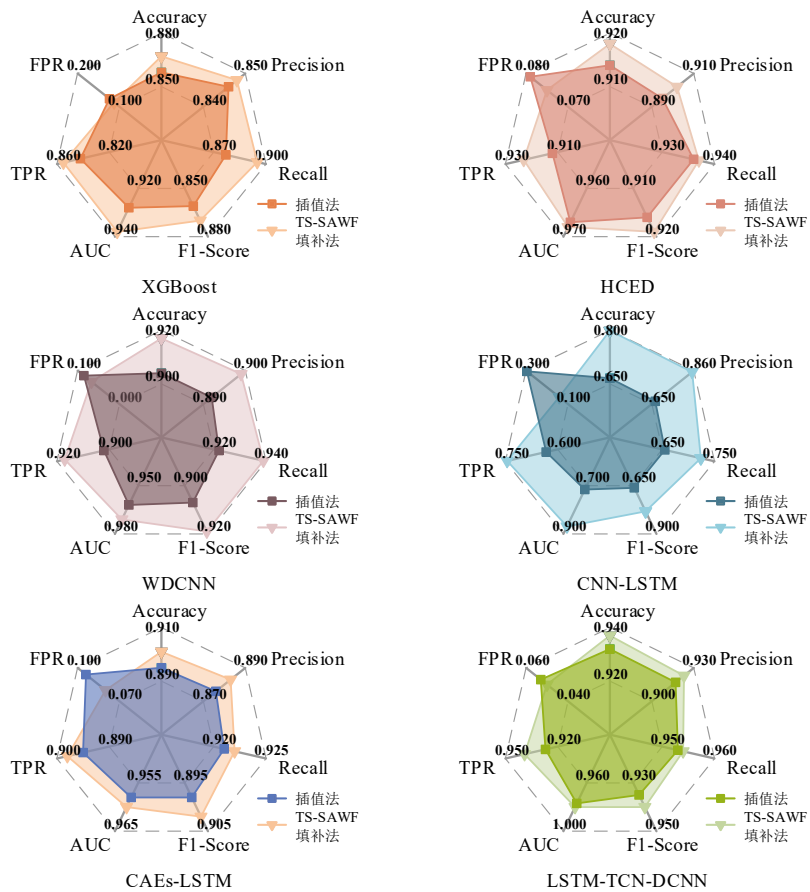
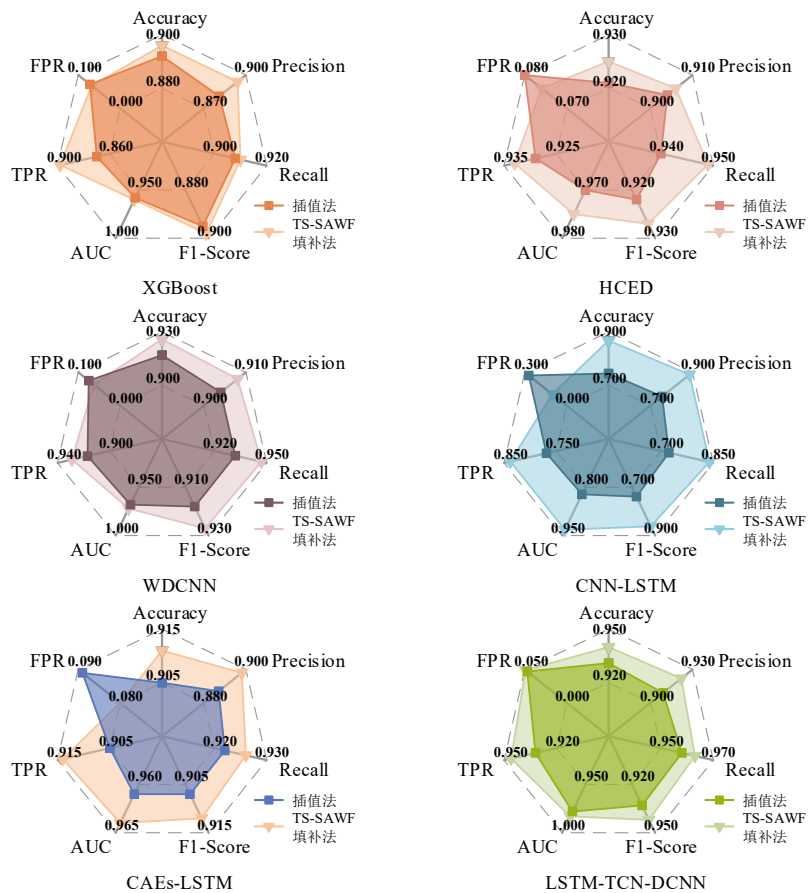


图 A2 两阶段自注意力权重融合缺失数据补全模型结构图

Fig. A2 Structure diagram of the two-stage self-attention weight fusion missing data imputation model



(a) 训练集比例=60%



(b) 训练集比例=70%

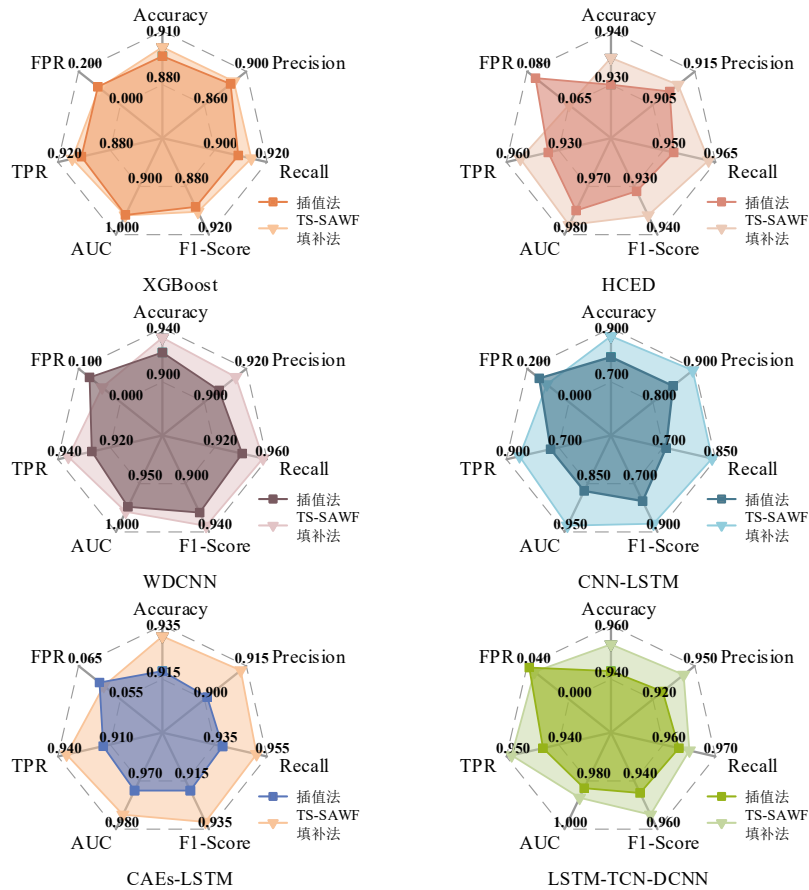


图 A3 不同训练集比例下的窃电检测模型在插补法与 TS-SAWF 填补法上的评价指标对比
Fig. A3 Comparison of the evaluation metrics of electricity theft detection model on the interpolation method and the TS-SAWF imputation method under different training set ratios

表 A1 软硬件配置

Table A1 Hardware models and software versions

硬件	型号	软件	版本
OS	Win10(64bit)	Python	3.10
CPU	Intel Core i9-14900k	Pytorch	2.2
GPU	NVIDIA GeForce RTX 4090	Pypots	0.6
RAM	DDR4 24GB	Tensorboard	2.16
HDD	SSD 1TB	CUDA	12.1

表 A2 TS-SAWF 模型参数设置

Table A2 Parameter settings of TS-SAWF model

参数	数值	参数	数值
Conv-DMMHSA 模块层数	2	查询向量 Q 维度	64
Conv-DMMHSA 层输入维度	512	键向量 K 维度	64
FFN 中隐藏层维度	256	值向量 V 维度	64
Conv-DMMHSA 层多头个数	4	丢弃率	0
Conv-DMMHSA 卷积核大小	3	学习率	1×10^{-4}
MIP 损失函数权重	1	优化器	Adam
ORP 损失函数权重	1	批尺寸	32

表 A3 窃电检测模型评价指标提升对比

Table A3 Comparison of the improvement in evaluation metric for electricity theft detection model

模型	$S_{Accuracy}$	$S_{Precision}$	S_{Recall}	$S_{F1-Score}$	S_{AUC}	S_{TPR}	S_{FPR}
训练集比例=60%							
XGBoost	0.9%	0.2%	1.8%	0.9%	1.0%	1.3%	-1.1%
HCED	0.4%	0.7%	0.1%	0.3%	0.1%	1.1%	-0.4%
WDCNN	1.3%	0.7%	1.7%	1.2%	0.9%	1.5%	-1.8%

CNN-LSTM	13.3%	18.5%	6.9%	12.2%	15.3%	11.2%	-15.4%
CAEs-LSTM	0.6%	0.7%	0.1%	0.4%	0.2%	0.3%	-1.4%
LSTM-TCN-DCNN	0.5%	0.6%	0.1%	0.5%	0.3%	1.2%	-0.3%
训练集比例=70%							
XGBoost	0.4%	1.3%	0.2%	0.3%	0.2%	1.8%	-0.4%
HCED	0.4%	0.2%	0.9%	0.5%	0.5%	0.4%	-0.4%
WDCNN	0.9%	0.4%	1.5%	0.9%	0.4%	1.2%	-1.2%
CNN-LSTM	12.3%	13.1%	11.6%	12.3%	11.0%	6.9%	-16.8%
CAEs-LSTM	0.6%	1.1%	0.4%	0.5%	0.3%	0.9%	-0.9%
LSTM-TCN-DCNN	0.9%	1.3%	0.5%	0.9%	0.5%	1.4%	-0.2%
训练集比例=80%							
XGBoost	0.5%	0.3%	0.5%	0.4%	0.1%	0.7%	-0.7%
HCED	0.5%	0.2%	1.0%	0.5%	0.3%	1.6%	-1.2%
WDCNN	1.1%	0.8%	1.6%	1.1%	0.5%	0.9%	-3.0%
CNN-LSTM	7.8%	4.7%	13.2%	9.4%	7.2%	11.9%	-3.6%
CAEs-LSTM	1.3%	1.2%	1.3%	1.3%	0.5%	2.1%	-0.1%
LSTM-TCN-DCNN	1.0%	1.5%	0.2%	0.9%	0.4%	0.6%	-0.5%