

王若兮, 谢雪松. 基于优化最小计数概要与深度学习的小样本污水数据特征增强方法[J]. 净水技术, 2026, 45(7): 116-121, 146.

Wang R X, Xie X S. Feature enhancement method for small sample wastewater data based on optimized count min sketch and deep learning models[J].

Water Purification Technology, 2026, 45(7): 116-121, 146.

基于优化最小计数概要与深度学习的小样本污水数据特征增强方法

王若兮, 谢雪松*

(北京工业大学信息科学与技术学院, 北京 100124)

摘要 【目的】 由于污水水质数据的获取成本较高, 现有的基于深度学习技术的污水水质预测模型普遍面临着小样本数据集带来的精度和鲁棒性不足的问题。【方法】 为解决此问题, 本文提出了一种基于衰减因子的优化最小计数概要(CMS)算法与长短期记忆(LSTM)网络-注意力(Attention)机制相结合的污水水质预测模型。首先, 优化的CMS算法通过引入衰减因子, 有效提取频域特征, 增强了数据在频域上的信息表达, 使得数据能够更好地反映出其潜在的规律性。其次, LSTM-Attention网络基于优化后的CMS特征进行软测量建模。LSTM能够捕捉数据的时间序列特征, 而Attention机制则在此基础上对重要特征进行加权, 从而提高了模型在处理小样本数据时的预测精度和鲁棒性。【结果】 为验证这种建模方法的有效性, 本文的试验部分使用了真实的污水水质数据集, 结果表明, 结合优化CMS特征提取的LSTM-Attention模型显著减少了预测误差, 并有效提高了预测精度。优化后的模型在测试集上的表现得到了显著提升, 其中均方误差(RMSE)降低了9.04%, 平均绝对误差(MAE)降低了7.93%。【结论】 试验结果验证了这种数据增强方法在解决小样本数据背景下的有效性, 能够有效解决小样本数据下的预测问题, 显著提高了模型的精度和鲁棒性, 为污水处理过程的优化提供了有效的技术支持。

关键词 最小计数概要; 深度学习; 软测量; 污水处理; 数据增强

中图分类号: TP18; X703 文献标志码: A 文章编号: 1009-0177(2026)07-0116-07

DOI: 10.15890/j.cnki.jsjs.2026.07.013

Feature Enhancement Method for Small Sample Wastewater Data Based on Optimized Count Min Sketch and Deep Learning Models

Wang Ruoxi, Xie Xuesong*

(School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China)

Abstract [Objective] Due to the high cost of acquiring wastewater quality data, existing deep learning-based models for wastewater quality prediction often face issues related to insufficient accuracy and robustness, particularly when dealing with small sample datasets. [Methods] To address this challenge, this paper proposed a novel wastewater quality prediction model that combines an optimized count-min sketch (CMS) algorithm with a long short-term memory (LSTM) network enhanced by an Attention mechanism. First, the optimized CMS algorithm introduced a decay factor to effectively extract frequency domain features, enhancing the representation of data in the frequency domain and enabling better reflection of underlying patterns. Second, the LSTM-Attention network performed soft measurement modeling based on the optimized CMS features. The LSTM captured the temporal characteristics of the data, while the Attention mechanism assigned weights to important features, thereby improving the model's prediction accuracy and robustness in small sample data environments. [Results] To validate the effectiveness of this modeling approach, experiments were conducted using a real-world wastewater quality dataset. The result showed that the LSTM-Attention model, enhanced by CMS feature extraction, significantly reduced prediction errors and improved prediction accuracy. Specifically, the optimized model demonstrated a substantial improvement in performance on the test set, with a reduction of 9.04% in root mean square error (RMSE).

[收稿日期] 2024-11-27

[作者简介] 王若兮(1999—), 女, 硕士研究生, 研究方向为软测量污水水质监测系统, E-mail: wangruoxi_1008@126.com。

[通信作者] 谢雪松, 男, 博士研究生, 研究方向为人工智能在嵌入式系统中的应用, E-mail: xiexuesong@bjut.edu.cn。

and 7.93% in mean absolute error (MAE). [**Conclusion**] These experimental result validate the effectiveness of this data augmentation method in addressing the challenges posed by small sample data, significantly improving the model's accuracy and robustness. This approach provides effective technical support for optimizing wastewater treatment processes.

Keywords count-min sketch (CMS); deep learning; soft sensor; wastewater treatment; data enhancement

随着工业化和城镇化的迅速推进,水污染和水资源短缺问题日益严峻。为了提升水资源利用效率,污水回收与处理显得尤为重要。在污水处理过程中,水质监测数据为处理过程提供了客观的评价标准。然而不同水质指标的监测周期和成本差异较大,这为传统的直接测量方法带来了挑战。基于深度学习的软测量技术,通过引入辅助变量对难以直接测量的水质指标进行预测,显著降低了监测成本。尤其在资源受限的地区,软测量技术为污水处理提供了精准的决策依据。

然而,输入的样本数据量对预测模型的精准性和鲁棒性有着重要影响。由于数据获取往往受到安全性等因素的限制,可应用于模型训练的数据集样本容量普遍较小。因此,如何充分挖掘有限数据集的特征以进行精准的预测是当前亟待解决的问题。为应对此挑战,本文提出了一种结合最小计数概要(CMS)算法与深度学习模型的方法。通过优化后的 CMS 算法提取频域信息作为特征输入长短期记

忆(LSTM)网络-注意力(Attention)模型中,从而提高了模型的预测精度。

本文的组织结构如下:第一部分将简要回顾相关工作;第二部分描述了 CMS 的优化方法及如何应用于模型训练;第三部分介绍了所使用的数据集及试验结果;第四部分总结了本文的主要贡献及未来的研究方向。

1 相关工作

1.1 CMS 算法及其优化

CMS 是一种基于二维数组的数据结构(图 1),其大小由数组的宽度值 w 和深度值 d 决定,由 Cormode 等^[1]提出。CMS 利用 d 个独立的哈希函数($h_0, \dots, h_{d-1}$)将输入数据映射到不同的计数器中,通过计数器的最小值来估计频率并且进行压缩存储。CMS 存在哈希冲突产生的误差,会导致查询结果大于实际值。但由于其结构简单且具有实时处理能力,已经广泛应用于互联网流量分析、大规模数据库监控、聚合查询、重击元素、分位数查询和分布式计算^[2-3]等领域。

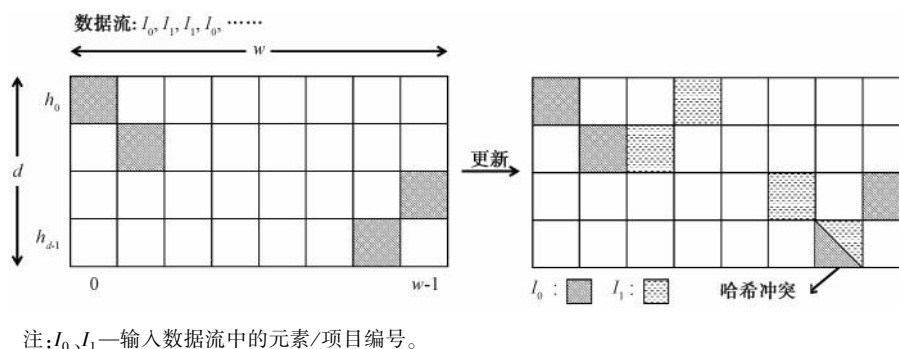


图 1 CMS 概览

Fig. 1 Overview of CMS

目前, CMS 已被广泛应用于大规模数据流的频率估计和数据摘要中,但在工业物联网(IIoT)领域的应用研究仍然有限。随着 IIoT 的快速发展,设备监控和实时数据流处理在工业环境中变得越来越重要。由于 IIoT 环境中存在大量传感器和智能设备,这些设备需要持续生成和传输大量的数据流,而处理这些数据往往面临存储和计算资源有限的挑战。

尽管 CMS 无法提供完全精确的结果,其误差可通过增加计数器的宽度和深度来有效控制。此外, CMS 在空间和时间复杂度方面具有显著优势,适合在内存和计算资源受限的环境中使用。

Bostanci 等^[4]提出了一种使用 CMS 解决存储器高频查询时电荷翻转问题的方法,这一研究在硬件设备监控中具有重要的启示。在 IIoT 环境中,传感

器和嵌入式设备通常在计算资源有限的情况下执行高频查询,而 CMS 通过其压缩存储和快速查询能力,可以有效减少硬件设备的资源消耗,支持设备的实时数据处理和故障检测。这对于提升 IIoT 设备的运行效率,减少设备故障和停机时间具有重要意义。

由于 CMS 使用的是增量计数的方式,因此它并没有为每个数据项提供独立存储且可以删除的记录。此外由于哈希冲突的问题,如果 2 个不同的数据项被映射到相同的位置,则这个位置的计数器的值会被 2 个数据项同时共享。针对 CMS 无法删除数据的问题,本文提出了一种基于衰减因子的优化方法。在保证结构尽可能简化的同时,该方法能够更好地反映动态数据场景中数据的实时性特点。

1.2 小样本问题及解决办法

深度学习在许多领域取得了显著的进展,尤其是在大数据集上。高精度且鲁棒性强的深度学习模型通常需要大量且高质量的训练数据,而数据稀缺或难以获取往往会导致模型无法提供有效的预测,甚至可能出现过拟合、偏差或不稳定的预测结果。在污水水质预测中,数据获取通常需要获得真实污水处理厂的授权,这使得数据的获取变得更加困难和受限。

为解决小样本问题,许多研究者提出了多种数据增强方法,这些方法通过扩充数据集来提升模型的泛化能力。最常见的数据增强方法之一是利用生成对抗网络(GAN)^[5]。GAN 通过生成与原始数据分布相似的虚拟数据来扩充数据集,从而改善模型的训练效果。GAN 在图像生成和某些其他领域中表现出色,但是在时序数据预测中仍面临着不少局限性。首先,GAN 对计算资源的需求较高,尤其在生成复杂数据时,这对于资源受限的工业环境尤为不利。其次,GAN 生成的数据往往缺乏实际的物理背景或领域知识,难以很好地与真实的工业数据相匹配,从而影响模型的训练质量和稳定性。

除了 GAN,加入噪声^[6]等方法也被用来扩充数据并提高模型鲁棒性。尽管这些方法能够缓解小样本带来的问题,但它们通常假设数据是独立同分布的,即样本之间相互独立,并且来自同一个概率分布,而污水数据往往具有时序性或强烈的领域依赖性。

针对这一问题,学者们^[7]提出了专门针对时间序列数据小样本问题的生成模型,如 TimeGAN,它通过生成对抗框架生成时间序列数据,以帮助缓解

数据稀缺的挑战。然而,尽管这些生成模型对时间序列数据有所帮助,但在 IIoT 中应用时,仍然面临计算开销大和特征捕捉不足等问题,尤其是在需要根据数据特征提供更准确预测的应用场景中。

在小样本问题的解决过程中,挖掘现有数据集的频域特征是一种重要方法。常用的频域特征提取方法包括快速傅里叶变换(FFT)和小波变换,但它们在处理复杂动态数据时存在一定局限性。

FFT 通常假设数据是平稳的,适用于周期性或者局部平稳的数据集,而污水数据通常包含明显的非平稳波动,这种波动通常受到工况变化等复杂因素的影响,因此其与污水数据的匹配度较差。此外,FFT 的计算复杂度为 $O(N \log N)$,其中 N 为序列的长度,因此在资源受限的场景中可能面临计算效率和存储需求的挑战。

小波变换在捕捉时频特性方面较为灵活,已有学者^[8]将小波函数与神经网络结合构建为小波网络,用于水质数据的预测。对于长度为 N 的一维序列,离散小波变换的计算复杂度通常达到 $O(N)$,但在处理多变量水质数据或进行多尺度分解时,计算量和存储开销会随变量数量、分解层数及样本规模增加而上升。此外,当样本数量有限,且数据中存在较强的观测噪声、传感器误差或异常波动时,小波特征提取或模型参数学习过程可能受到干扰,从而影响后续预测效果。

相比之下,CMS 的空间复杂度取决于选用的二维数组的大小。若记总计数器数量为 $k=w \times d$,则其空间复杂度为 $O(k)$,远小于 FFT 和小波网络的存储需求。在实时数据流场景下,CMS 能够通过紧凑的存储结构完成频率特征的提取。

因此,本文提出了结合频域信息的优化 CMS 方法,并将其与深度学习模型结合,以增强污水数据特征的表达。CMS 通过引入频域信息,有效提升了污水数据的表达能力,尤其是在小样本条件下,能够更好地提取数据的高维特征并减少噪声干扰,进而改善模型的训练效果。这一方法通过同时充分利用污水数据的时域特征和频域特征解决了小样本问题,从而显著提高了模型的预测精度。

2 方法

2.1 模型整体设计概述

本文所提出的污水水质预测系统的整体框架

如图 2 所示,模型主要包括基于衰减因子的优化 CMS 模块、频域数据处理模块及基于 LSTM-Attention 的软测量预测模块。传感器采集的数据为时序序列,每一个数据通过衰减 CMS 模块统计获得频域信息,作为频域特征的原始数据。数据处理模块包括卡方分箱和独热编码 2 个部分。首

先,卡方分箱通过将连续的频域数据特征划分为若干离散的区间,减少特征维度。之后,独热编码通过将离散特征转化为独立的高维表示,避免了分箱类别之间的顺序影响。作为 LSTM-Attention 模型的频域输入数据和时域数据一同输入以完成水质指标的预测。

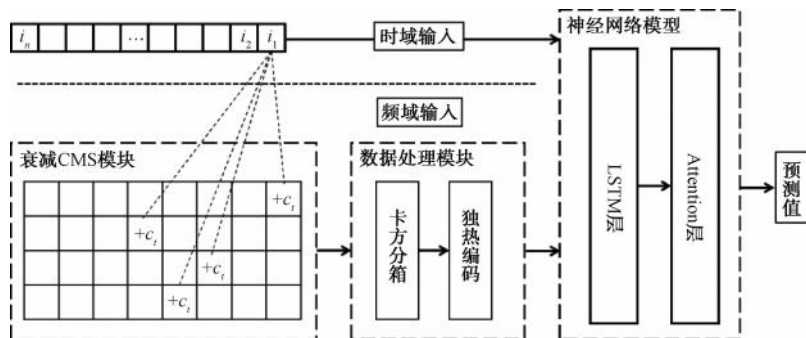


图 2 污水水质预测系统框架

Fig. 2 Framework Overview of Wastewater Quality Prediction System

2.2 基于衰减因子 α 的 CMS 优化

CMS 是一种轻量级的频率估计数据结构,其通过 $w \times d$ 计数器构成的二维数组实现数据的紧凑存储。CMS 通过多个独立的哈希函数将输入数据映射到计数器中,并利用计数器的最小值实现频率的估计。然而,传统 CMS 存在无法删除历史数据的固有缺陷,这使得在动态数据流场景中,例如 IIoT 中容易受到历史数据的干扰,导致预测精度下降。此外,随着时序数据特性和趋势的变化,历史数据的频率分布逐渐失效,使 CMS 难以准确反映当前数据的特征分布。

为了解决上述问题,本文提出了一种衰减 CMS。衰减因子 α 是用于控制历史数据权重衰减速率的重要参数,其合理选择可以平衡模型对历史数据和实时数据的关注程度。虽然本文选取了固定的 α 值,但在不同场景中,可以根据具体的应用需求和数据特性调整取值范围。例如,在实时性要求较高的场景中可以选择较小的 α 值;而在历史数据对预测结果影响更大的场景中,则可适当提高 α 值。在优化后的 CMS 中,通过引入指数衰减因子 α ,可以逐步削弱历史数据对频率估计的影响,从而更关注当前数据的分布特征。对于每个时间周期 t ,CMS 中计数器的更新计算如式(1)。

$$C_i(t+1) = \alpha \cdot C_i(t) + \Delta_i(t) \quad (1)$$

其中: $C_i(t)$ ——第 i 个计数器在时间 t 的值,无量纲;

α ——衰减因子, $0 < \alpha < 1$, 无量纲;

$\Delta_i(t)$ ——当前时序数据的新增量,通过哈希函数更新得到,无量纲。

为了更灵活地适应动态数据场景,该式可扩展为式(2)。

$$C_i(t+1) = \alpha^n \cdot C_i(t) + \sum_{k=0}^{n-1} (\alpha^k \cdot \Delta_{t+n-k}) \quad (2)$$

其中: n ——时间周期的数量,无量纲;

k ——求和索引,表示历史数据增量距离当前时刻的时间步数,无量纲;

Δ_{t+n-k} ——第 $t+n-k$ 时间点的新数据增量,无量纲。

式(2)表明,随着时间周期的推移,历史数据的权重将呈指数衰减,从而更加精准地捕捉当前数据的变化趋势,当到达预设的衰减周期时,衰减 CMS 的工作原理如图 3 所示。

2.3 小样本问题和特征工程

为解决污水数据的小样本问题,衰减 CMS 提取得到的特征通过特征工程的数据处理方法,将频域信息和时域信息进行结合,生成增强的特征集用于预测模型的训练。

衰减 CMS 从时序数据中提取频域特征 X_{freq} , 这

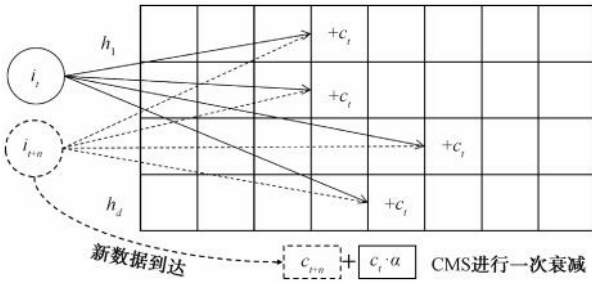


图3 衰减CMS工作原理

Fig. 3 Operation Principle of Decay CMS

些特征能够捕捉时序数据的频率分布和动态变化特性。在提取频域特征后,为了进一步增强其表达能力,利用卡方分箱将数据划分为若干区间,确保不同区间之间的数据特征差异最大,同时使同一区间内的数据特征差异最小,从而更好地体现特征的分布趋势。设卡方分箱生成每个区间代表一个数据类别,频域特征的值将按照对应区间进行映射和转换。

设卡方分箱生成的区间为 $B_i (i = 1, 2, \dots, m)$, 每个区间代表一个数据类别,频域特征 X_{freq} 的值将按照对应区间进行映射和转换。

在卡方分箱之后,将其转化为独热编码的形式。对于每个分箱 B_i ,独热编码生成一个 k 维的向量,其中数据所属区间对应的位置为 1,其余位置均为 0。最终得到的独热编码特征矩阵 X_{onehot} 能够以稀疏的形式表达频率分布。最后,将原始时序数据 X_{time} 与 X_{onehot} 进行拼接,形成增强的特征集 X_{combined} 。最终的特征集构建如式(3)。

$$X_{\text{combined}} = [X_{\text{time}}, X_{\text{onehot}}] \quad (3)$$

其中: X_{time} ——原始时序数据,单位为具体时间序列值。

X_{onehot} ——独热编码的频域特征矩阵,无量纲。

2.4 LSTM-Attention 模型

为提高水质数据预测的精准性,如图4所示,本文采用了结合LSTM与Attention机制的深度学习模型。LSTM网络专为处理时间序列数据而设计,能够有效捕捉数据的长期依赖性和辅助时序模式;Attention机制通过为时间序列的不同时间步分配权重,将更多计算资源聚焦在关键时间步上,从而提高模型对重要信息的提取能力。

网络的输入层包含频域特征和时域特征的拼接

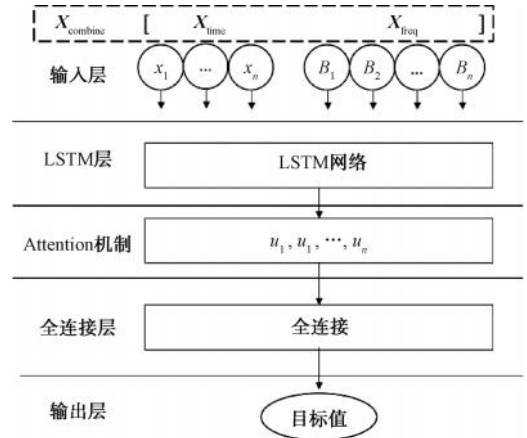


图4 LSTM-Attention 网络结构

Fig. 4 LSTM-Attention Network Structure

数据。LSTM层是模型的核心部分,包含3层LSTM,每层有256个隐藏单元,用于提取时间序列的长期依赖关系。为了增强模型对重要时间步的权重占比,网络中引入了Attention机制。Attention机制通过分配不同的权重,将更多的计算资源聚焦在关键时间步上,从而提高模型对重要信息的提取能力。

3 数据集和试验结果

本文基于优化后的衰减CMS提取的频域特征和原始时序特征构建增强数据集,并使用该增强数据集对目标水质生化需氧量(BOD)的预测进行建模。同时,为验证方法的有效性,本文将所提方法与仅基于时序特征的模型进行对比。这些模型均采用Adam算法进行参数优化。

3.1 数据集

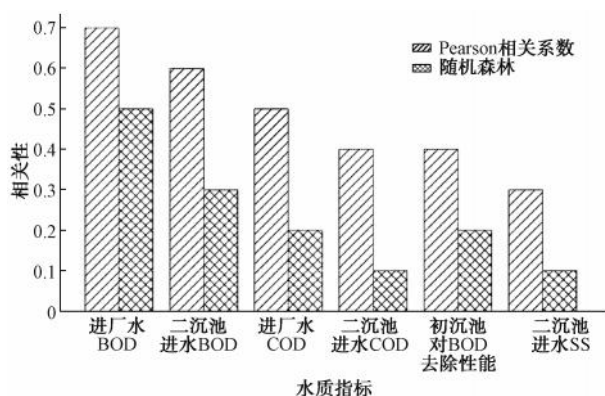
本文所使用的数据集来源于UCI Machine Learning Repository网站上的开源污水水质数据集。该数据集记录了一座污水处理厂1990年1月—1991年8月527 d的不同处理环节的监测数据,每个样本代表1 d的观测结果。数据集包含38个属性,所有属性均为连续型数值。本文选择第2个处理池的BOD作为软测量的目标变量。

在辅助变量的选择上,本文分别采用Pearson相关系数和随机森林算法,筛选出与目标变量相关性最大的6个变量,筛选结果如图5所示。

3.2 数据预处理和划分

为保证试验的准确性,本文对数据集进行了以下预处理。

(1)归一化处理:将数据值映射到[0,1],以消



注:COD—化学需氧量;SS—悬浮物。

图 5 辅助变量选择结果

Fig. 5 Selection of Auxiliary Variables

除不同属性量纲的影响。

(2)缺失值处理:对数据集中缺失值进行线性插值补充,确保模型能够处理完整的输入特征。

数据集按照 8 : 2 的比例随机划分为训练集和测试集,其中 80%用于模型训练,20%用于模型性能评估。

3.3 试验参数设置与评价指标

为验证衰减 CMS 对模型精度提升的有效性,均保持如表 1 所示的超参数设置。

表 1 超参数设置

Tab. 1 Hyperparameter Settings

参数名称	内容
衰减因子 α	0.85
Epoch	500
优化器	Adam
初始学习率	0.000 5
Batch size	64

本文将衰减因子 α 的值固定为 0.85。在实际的应用过程中,可以根据历史数据的重要性调整参数的大小。此外,在训练过程中,所有试验均在相同的硬件配置下运行,确保了结果的可比性。本文使用以下 2 种指标评估水质预测模型的精度。

均方根误差(RMSE)反映了预测误差的平方平均值,是评估预测误差的标准差,其计算如式(4)。

$$R_{\text{MSE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

其中: R_{MSE} ——RMSE 值;

y_i ——样本 i 的真实质量浓度,mg/L;

$\hat{y}_i$ ——样本 i 的真实质量浓度,mg/L。

平均绝对误差(MAE)反映了预测误差的绝对平均值,其计算如式(5)所示。

$$M_{\text{AE}} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

其中: M_{AE} ——MAE 值。

两者的值越小,表示模型的预测精度越高。

3.4 试验结果和对对比分析

为评估本文提出的衰减 CMS 和 LSTM-Attention 网络结构的有效性,本文对比了以下模型在相同数据集上的表现:时间卷积网络(TCN)、LSTM 和本文最终选择的 LSTM-Attention 网络。预测结果如表 2 所示。

表 2 测试集结果

Tab. 2 Results of Test Sets

模型	原始数据模型		衰减 CMS 数据模型	
	NRMSE	NMAE	NRMSE	NMAE
TCN	11.89%	10.23%	1.12%	1.12%
LSTM	9.74%	8.45%	0.94%	0.73%
LSTM-Attention	9.97%	8.65%	0.93%	0.72%

注:NRMSE 为归一化后的 RMSE,NMAE 为归一化后的 MAE。

由表 2 可知,引入频域特征的增强模型的 RMSE 和 MAE 值均显著低于仅使用时域特征的模型,这表明频域特征有效提升了模型的预测精度。RMSE 降低了 9.04%,MAE 降低了 7.93%,验证了优化后的 CMS 特征提取方法对小样本数据的适用性和有效性。LSTM-Attention 模型的误差均低于其他模型,这表明 Attention 机制能够有效捕捉时间序列中的关键信息,从而提升预测精度。

为验证衰减 CMS 的有效性,本文进一步试验对比了 3 种数据集:原始数据、包含原始 CMS 频域信息的数据集和衰减 CMS 频域的数据集的试验结果,如表 3 所示。

表 3 原始数据、原始 CMS 及衰减 CMS 数据集在 LSTM-Attention 模型的测试结果

Tab. 3 Test Results of Raw Data, Raw CMS and Decay CMS Data Sets in the LSTM-Attention Model

参数	原始数据	原始 CMS	衰减 CMS
RMSE	9.97%	5.58%	0.93%
MAE	8.65%	4.39%	0.72%

(下转第 146 页)

- Lasi Sectia Constructii, Arhitectura, 2018, 64(3): 73–82.
- [8] Daneshvar F A, Talebbeydokhti N, Dehghan S M, et al. Evaluation of darcy-weisbach friction factors of fiberglass pipes based on internal surface roughness measurement [J]. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 2023, 47(3): 1749–1762.
- [9] Chen L S. Discussion on the application of fiberglass reinforced plastics in building energy saving [J]. IOP Conference Series: Earth and Environmental Science, 2020, 580(1): 012051. DOI: 10.1088/1755-1315/580/1/012051.
- [10] Trykoz L V, Kamchatnaya S M, Pustovoitova O M, et al. The investigation of prestressed pressure pipes, reinforced with fiberglass plastic [J]. International Journal of Engineering Research in Africa, 2018, 36: 1–11. DOI: 10.4028/www.scientific.net/JERA.36.1.
- [11] Baali B, Benmounah A, Rokbi M. Mechanical characterization and optimum design of wound glass-fiber-reinforced polymer pipes based on the winding angle and the number of plies [J]. Mechanics of Composite Materials, 2020, 56(5): 673–684.
- [12] Avdeeva A, Shlykova I, Perez M, et al. Chemical properties of reinforcing fiberglass in aggressive media [J]. MATEC Web of Conferences, 2016, 53: 01004. DOI: 10.1051/mateconf/20165301004.
- [13] Popielski P, Bednarz B, Siefko R, et al. Monitoring of large diameter sewage collector strengthened with glass-fiber reinforced plastic (GRP) panels by means of distributed fiber optic sensors (DFOS) [J]. Sensors, 2021, 21(19): 6607. DOI: 10.3390/s21196607.
- [14] Wang Q D, Zhang Y B, Lv Q F, et al. Susceptibility assessment of soil erosion in overlaying diluvial fan of shallow underground pipelines [J]. Bulletin of Engineering Geology and the Environment, 2021, 80(3): 2573–2585.

(上接第 121 页)

由表 3 可知,包含衰减 CMS 频域信息的数据集的模型精度最高,验证了衰减因子对小样本模型精度提升的有效性。

4 结论

为解决污水数据集存在的小样本问题,本文提出了一种基于改进 CMS 的特征增强方法:衰减 CMS。本文首先引入衰减因子对 CMS 进行优化,使 CMS 更好地反映当下数据的频率特征,削弱历史数据的影响,之后通过卡方分箱和独热编码等数据预处理方法为软测量模型 LSTM-Attention 的输入做准备。最后将同时包含时域和频域信息的数据集用于污水水质指标的预测。从结果来看,衰减 CMS 引入的频域特征可以有效地提高小样本时间序列模型的预测精度,但是现在仍然存在一些限制,本方法对衰减因子 α 的选择依赖于经验,需要针对不同的数据集和场景进行调优,以取得最佳效果。此外,卡方分箱和独热编码方法尽管在小样本情况下提升了模型性能,但对于特征之间更加复杂的关系的捕捉能力依然有限。在未来的研究中,可以进一步探索更为复杂的特征提取和融合方法,以及将衰减 CMS 与布隆过滤器和压缩感知等技术结合,在保证模型精度的同时进一步降低数据存储成本。

参考文献

- [1] Cormode G, Muthukrishnan S. An improved data stream summary: The count-Min sketch and its applications [J]. Journal of Algorithms, 2005, 55(1): 58–75.
- [2] Arkan K, Palumbo A, Cassano L, et al. Processor security: Detecting microarchitectural attacks *via* count-Min sketches [J]. IEEE Transactions on Very Large Scale Integration: VLSI Systems, 2022, 30(7): 938–951.
- [3] Li G R, Liu Y, Wang Y. Analysis of the count-Min sketch based anomaly detection scheme in WSN [C]//2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications. September 24–26, 2014, Beijing, China. IEEE, 2014: 961–966.
- [4] Bostanci F N, Yüksel I E, Olgun A, et al. CoMeT: Count-Min-sketch-based row tracking to mitigate RowHammer at low cost [C]//2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA). March 2–6, 2024, Edinburgh, United Kingdom. IEEE, 2024: 593–612.
- [5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139–144.
- [6] Žvirblis T, Pikšrys A, Bzinkowski D, et al. Time series data augmentation methods for deep learning models in conveyor belt load classification [C]//2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream). April 25–25, 2024, Vilnius, Lithuania. IEEE, 2024: 1–6.
- [7] Yoon J, Jarrett D, van der Schaar M. Time-series generative adversarial networks [J]. Advances in neural information processing systems. 2019, 32. DOI: 10.5555/3454287.3454781.
- [8] Cong Q M, Yu W. Integrated soft sensor with wavelet neural network and adaptive weighted fusion for water quality estimation in wastewater treatment process [J]. Measurement, 2018, 124: 436–446.