

DOI:10.19651/j.cnki.emt.2520147

基于改进型 Informer 的应用性能多步预测与 在线异常监控系统*

陈棋东¹ 钱家昊¹ 吴 豪² 叶广飞¹ 张松洋³(1. 南京信息工程大学计算机学院 南京 210044; 2. 无锡学院物联网工程学院 无锡 214105;
3. 苏州科技大学电子与信息工程学院 苏州 215009)

摘 要: 面向云原生场景下多指标、强波动的应用性能时序,提出一种基于改进型 Informer 的预测—监控一体化方法与系统。在编码端引入上下文感知可逆归一化(CA-RevIN)以抑制跨时段分布漂移,结合多频段卷积增强(MFBCEM)与长时依赖聚合(LTA)实现多尺度建模,采用生成式解码器输出多步预测。系统侧构建数据采集、特征治理、推理服务与可视化告警闭环,利用峰超阈(POT)与自适应残差阈值实现分级异常检测与提前预警。在公开数据集与真实业务数据上,所提方案较长时间记忆(LSTM)、门控循环单元(GRU)及原始 Informer 在多步误差与稳定性上均有提升,并显著降低误报与漏报。

关键词: 长序列预测;Informer;归一化;时频增强;异常检测;云原生监控

中图分类号: TP391;TN915 **文献标识码:** A **国家标准学科分类代码:** 520.20

Multi-step prediction and online anomaly monitoring system for application performance based on an improved Informer

Chen Qidong¹ Qian Jiahao¹ Wu Hao² Ye Guangfei¹ Zhang Songyang³

(1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China;

2. School of Internet of Things Engineering, Wuxi University, Wuxi 214105, China;

3. School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China)

Abstract: For volatile multi-metric application performance time series in cloud-native environments, we propose an integrated forecasting and monitoring method and system based on an enhanced Informer. At the encoder side, context-aware reversible normalization (CA-RevIN) is introduced to mitigate distribution shifts, while multi-frequency band convolutional enhancement (MFBCEM) and long-term dependency aggregation (LTA) are used for multi-scale modeling. A generative decoder produces multi-step forecasts. On the system side, a closed-loop architecture is built for data acquisition, feature governance, inference and alerting. Peaks-over-threshold (POT) and adaptive residual thresholds enable hierarchical anomaly detection and early warning. Experiments on public datasets and real production data show that the proposed approach outperforms LSTM, GRU and the vanilla Informer in multi-step prediction accuracy and stability, while reducing false positives and missed detections.

Keywords: long-sequence time-series forecasting; Informer; normalization; time-frequency enhancement; anomaly detection; cloud-native monitoring

0 引 言

云原生与微服务架构的快速普及,使得生产环境中的应用运行呈现出前所未有的高动态性与强异构性^[1]。实例的弹性伸缩、短周期发布、容器化与微服务间复杂的调用

链,使单一服务的运行状态受制于整个拓扑与协同策略的实时变化。同时,多租户、资源共享与异构硬件进一步加剧了观测噪声与干扰。正因为这些特性,业务负载往往呈现多变量耦合、长时依赖与显著的非平稳性,既包含稳定的趋势与周期性波动,也常伴随由业务峰值、发布事件或网络抖

收稿日期:2025-10-31

* 基金项目:江苏省高等学校基础科学(自然科学)研究面上项目(25KJB520046)资助

动触发的厚尾尖峰行为。在此背景下,单纯依赖静态规则阈值或简单滑动平均的监控策略,既容易导致漏报又会产生大量误报,难以满足运维对可解释性预警与中长期容量规划的双重需求^[2]。

另一方面,云原生场景下的数据分布并非静态——不同时间窗口、不同服务阶段或不同部署拓扑都会引入显著的均值与方差漂移,这使得在离线训练阶段获得的模型在线推理时可能失去原有泛化性能。特别是在面对节假日流量、业务促销或流量突变等跨时段分布迁移事件时,模型若不能实现样本级或上下文感知的归一化处理,预测偏差与不确定性将被放大,进而影响告警的可靠性与响应效率^[3]。此外,云环境中观测的高频噪声与低频趋势并存,要求预测方法既能捕捉局部突发,又能稳健把握长期依赖,这对模型的结构设计、训练策略与在线校准提出了更高要求。

近年来,随着深度学习在时间序列预测中的广泛应用,各类基于卷积、循环与自注意力的混合结构被提出以应对非线性与多尺度特征的挑战。早期工作多依赖于卷积与循环的组合以兼顾局部模式与短期依赖,例如基于时间卷积网络与长短期记忆(long short-term memory, LSTM)的混合模型在不规则采样与短期波动建模上取得了良好效果^[4];而 CNN-GLSTM^[5]等通过向量自回归与卷积预处理以滤除多元序列间的线性耦合,提升了多变量预测的稳定性。与此类似,面向在线学习场景的研究也开始将图结构建模与序列网络结合,以同时刻画学生之间的关系特征与随时间演化的学习行为。然而,这类以卷积、循环结构为主的时序模型总体上更擅长在有限时间窗口内提取局部模式和短期依赖,难以在多租户、多服务并发的云环境中同时处理多指标强耦合、长时依赖与厚尾尖峰等复杂行为,长序列多步外推时易出现误差累积与稳定性下降。崔立志等^[6]将关系图卷积(relational graph convolutional network, R-GCN)、注意力机制与门控循环单元(gated recurrent unit, GRU)结合,提出了 R-GCN-GRU 模型,在图结构约束下建模在线学习场景中学生与教学实体的时序交互,但这类方法通常假定图结构相对静态或变化缓慢,难以适应云服务拓扑和资源依赖关系的高频动态变化。王海亮等^[7]提出以 BiLSTM 的双向时序建模能力为骨架的 INGO_BiLSTM-SA,通过单头注意力机制强化对关键时间片与长程依赖的聚焦,但对基于历史相关性筛选特征与离线寻优的时序建模难以快速适配,因而对跨阶段变化的泛化与预警能力仍不足。

随着长序列时间序列预测研究的深入,基于 Informer 的模型家族逐渐成为主流方案。Zhou 等^[8]提出的 Informer 通过稀疏自注意力与层间蒸馏大幅降低了长序列的计算复杂度并支持一次性多步生成。随后,大量改进模型围绕 Informer 展开,如 Crossformer^[9]通过交叉维度注意力显式建模多变量时间序列各通道之间的相关性,在保持长序列建模能力的同时强化了跨维度依赖的刻画,从而提升了多

变量长序列预测的整体精度与稳定性;FEDformer^[10]通过在 Transformer 中融入傅里叶、小波等频域算子,增强了对低频趋势与周期成分的建模能力;TA-Informer^[11]、PatchTST^[12]等则从时间编码、patch 切片、通道独立建模等角度出发,进一步提升了长序列多变量预测的性能。王国辉等^[13]提出的基于 Informer-BiGRU-Global Attention 的深度预测框架给出了新的思路,利用 Informer 建模长时依赖,引入双向 GRU 强化局部时序特征提取,并通过全局注意力机制对多源输入特征进行重加权,兼顾了点预测精度与区间不确定性刻画。总体来看,现有改进型 Informer 研究多以提升通用基准数据集上的预测精度为主要目标,其模型设计与实验验证主要集中于电力负荷、气象、交通等结构相对规整的时间序列,特征维度较少、噪声水平可控且跨时段分布变化相对平缓。尽管 Crossformer、FEDformer 等工作在时域分解和频域增强方面进行了有益探索,但对于同时包含高频噪声、局部尖峰与厚尾极值以及低频趋势和长期依赖的复杂序列,其时频一体化建模能力仍显不足,尤其是在多指标强耦合的云原生场景中,对局部突发行为与整体趋势的统筹刻画能力有限。此外,上述研究大多聚焦于预测模型结构本身的改进,较少从系统视角讨论与在线监控、异常检测和分级告警的深度耦合,使得模型输出的不确定性估计和尾部分布刻画难以与实际告警策略形成完整的工程闭环。

针对这些问题,面向在线与工程化部署的研究更关注模型的轻量化、在线自适应与跨场景迁移能力。诸如 Wen 等^[14]实现的基于 OneNet 的在线集成与迁移学习方法,以及 TCformer^[15]、CauDformer^[16]等为工业场景定制的 Transformer 变体,均强调在保留预测能力的同时降低延迟并支持在线更新。卢翔等^[17]面向航空发动机剩余寿命预测场景,提出了 Informer 与特征优选结合的深度回归框架,验证了长序列 Transformer 在设备寿命预测任务中的有效性。CL-Informer^[18]则针对长序列多变量时间序列的特点,对编码器结构和注意力权重分配方式进行了改造,在保持计算效率的前提下进一步提升了长期预测的精度与稳定性。然而现有面向在线与工程化部署的研究多将预测模型视作相对独立的功能模块,主要关注模型结构的轻量化与对概念漂移的自适应能力,对模型如何与云原生监控栈中的数据采集、特征治理、时序存储及告警策略实现端到端解耦与闭环集成关注相对不足,在这类背景下负载优化问题愈发凸显^[19]。同时,此类方法的设计对象多为单一业务流或少量关键指标,对多租户、多服务并发的应用性能监控场景中如何支持插件化部署、弹性扩缩容以及由多步预测直接驱动的分级告警考虑有限,难以充分满足云原生环境对于低时延、高可用性和易运维等综合工程需求^[20]。本文在此基础上结合上下文感知可逆归一化(context-aware reversible normalization, CA-RevIN)、多频段卷积增强模块(multi-frequency band convolution enhancement

module, MFBCEM) 与长时依赖聚合模块 (long-term aggregation, LTA) 的设计, 旨在实现面向云原生场景的多步稳健预测与分级告警体系, 并通过离线与在线实验验证其有效性。

综上, 现有研究在局部以及全局时序建模、分布漂移缓解以及在线部署等方面虽已取得诸多进展, 但面向云原生应用性能监控 (application performance monitoring, APM) 场景, 多指标强耦合、跨时段分布迁移与预测-告警工程闭环等问题仍未得到系统解决, 亟需一套兼顾预测精度、稳健性与可部署性的统一框架^[21]。为此, 本文围绕性能时序的长序列多步预测与在线异常监控, 开展了如下工作:

1) 为提升模型在分布漂移、业务场景切换等复杂条件下的预测稳健性, 在改进型 Informer 编码端引入 CA-RevIN, 实现跨场景分布对齐; 在骨干中通过叠加 MFBCEM 与 LTA, 以兼顾局部-全局、多尺度时序特征的建模能力, 并采用生成式解码器输出未来多步预测结果。

2) 为降低在线监控中的误报与漏报, 构建预测—监控—告警闭环: 在预测不确定性建模中结合 MC-Dropout 与模型集成获取区间置信度, 再利用峰超阈方法 (peak over threshold, POT) 与自适应残差阈值实现分级告警与提前预警, 从而支持对性能异常的精细化识别与响应。

3) 在公开数据集与真实业务性能数据上开展系统性对比与消融实验, 验证所提改进型 Informer 架构及闭环监控策略在长序列多步预测稳定性、告警准确性与工程可集成性方面的有效性与通用性。

1 框架设计

1.1 框架设计思路

本系统采用分层、模块化的架构设计, 在满足工程部署需求的同时兼顾研究场景下的可复现性, 整体结构如图 1 所示。按功能划分, 系统由数据平面、控制平面和展示层 3 部分组成。数据平面与展示层在模块划分上与典型云原生监控平台, 采用基于 Prometheus 与 Grafana 的方案, 但在数据流编排以及预测结果与告警逻辑的耦合方式上针对云原生应用性能监控场景进行了重新设计与增强。

数据平面由多路数据采集插件构成 (如 Prometheus、pidstat 等), 负责按统一格式采集主机、容器及进程级的时序指标, 并将其送入预处理流水线。预处理阶段包括缺失值与噪声处理、重采样、滑动窗口构造等步骤, 以生成用于模型训练和在线推理的标准化样本^[22]。经处理后的数据写入时序数据库 InfluxDB, 并据此驱动离线训练与在线评估两个模块: 离线训练模块用于模型开发、交叉验证及开放神经网络交换格式 (open neural network exchange, ONNX) 模型导出; 在线推理模块以已部署的 ONNX 模型为核心, 通过推理服务实时消费时序数据, 输出点预测及其不确定性区间, 结果同样回写至时序库以支持后续查询与可视化分析。与仅基于固定阈值或简单统计量构建告警规

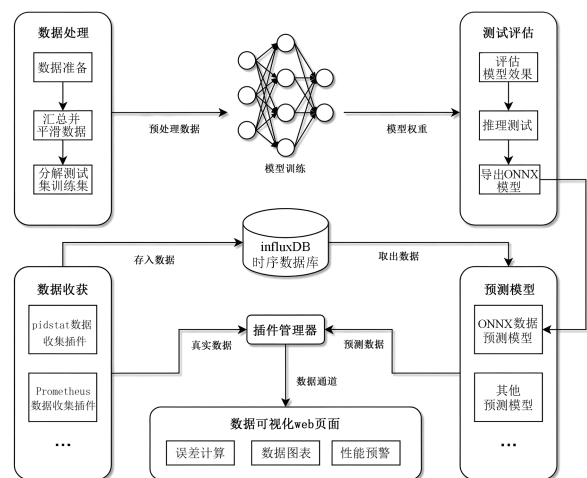


图 1 框架设计

Fig. 1 Framework architecture

则的传统方案不同, 本系统将预测值及其不确定性统一纳入时序数据存储, 用于后续的残差分析、极值建模与分级告警策略设计, 从而为预测—监控—告警闭环的构建提供完整的数据基础。

控制平面以插件管理器为核心, 负责插件的注册、配置下发、生命周期管理与数据管道编排。插件管理器将数据采集插件、数据预测插件与数据库插件解耦, 使得新增采集源或替换预测模型仅需在管理器中完成配置而不改动整体流程, 从而实现高内聚低耦合的工程化扩展性。该设计也便于在多租户或多指标场景下横向扩展与灰度部署, 并支持按需隔离推理资源以保证线上延迟约束^[23], 具体设计如图 2 所示。本系统在控制平面中显式引入预测插件: 告警子系统所依赖的残差序列、预测区间和 POT 阈值均由预测插件和特征治理流水线产出, 告警逻辑与具体模型解耦但仍以预测为核心驱动, 从而在规则层保持兼容性的同时, 将传统的静态阈值告警升级为“基于预测残差与尾部分布的分级告警”。

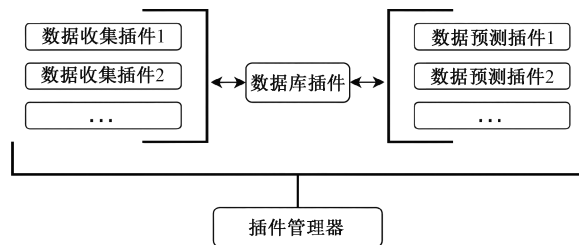


图 2 插件管理器示意图

Fig. 2 Plugin manager schematic

展示层通过前后端分离的 Web 界面直接读取时序库中的原始观测、预测均值与预测区间, 并提供总体态势视图、插件配置面板、采集指标列表及单指标的时序对比、误差面板与分级告警面板。可视化与告警子系统与插件管理器协同工作, 当预测残差、POT 极值或自适应残差阈值触

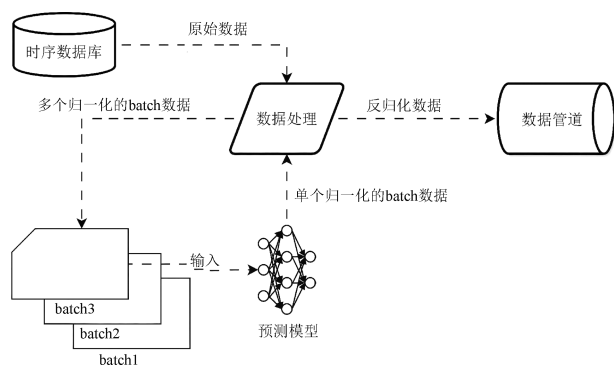


图 4 数据预测插件的工作流程

Fig. 4 Data prediction plugin work flow

信区间。输出结果经反归一化还原至原量纲后,与对应时间窗口的实测值一并写回数据管道,用于后续的误差计算、POT 阈值拟合及可视化展示。与仅输出下一时刻点预测并直接与固定阈值比较的传统告警策略相比,本文将多步预测轨迹及其不确定性区间共同作为告警判据:一方面利用多步残差序列刻画异常的持续性与提前量,另一方面在残差基础上引入 POT 极值分析与自适应残差阈值,从而构建与预测模型输出紧密耦合的分级告警机制。

此外,数据库插件不仅承担数据的持久化存储与索引查询功能,还支持多线程并发读写,以确保在高频监控场

景下的数据吞吐能力。通过参数化接口,预测插件可灵活指定查询时间范围、采样粒度与预测步长,从而实现动态自适应的预测策略。整体流程在保证计算效率的同时,兼顾了数据一致性与结果可追溯性,为后续的模式评估与在线优化提供了坚实的工程基础^[25]。

2 模型结构

2.1 模型设计思路

为应对云计算环境下长期负载预测中存在的分布偏移、局部特征建模不足以及长期依赖捕获不充分等问题,本文构建了一种基于改进型 Informer 框架的多步预测模型——Informer-RevMF。该模型以“分布自适应归一化—多频特征增强—长期依赖聚合”为总体设计思路,通过在输入、特征增强与主干建模 3 个层次上的协同改进,实现对复杂负载序列的稳定建模与高精度预测,Informer 主体模块如图 5 所示。与 Crossformer、FEDformer、TA-Informer 等主要从分解结构或注意力机制形态对 Transformer 进行改进的工作不同,Informer-RevMF 并不重新设计自注意力结构,而是更侧重于面向云原生场景提升长序列多步预测的稳健性与工程可用性,围绕分布漂移、时频多尺度特征以及长期依赖等在 APM 负载中尤为突出的问题进行针对性增强。

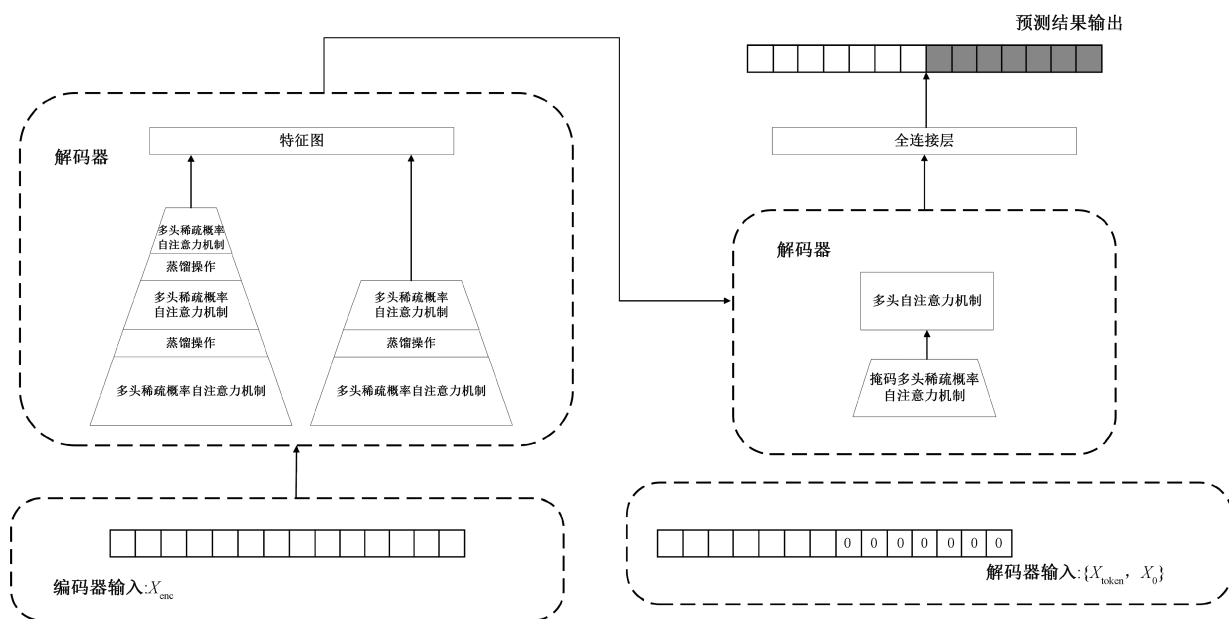


图 5 Informer 模块结构

Fig. 5 Informer module structure

在输入阶段,模型引入 CA-RevIN 模块,以解决云计算负载序列中普遍存在的非平稳性与分布漂移问题。CA-RevIN 通过对每个样本、每个通道沿时间维进行独立归一化,并结合轻量上下文编码器动态生成时间相关的仿射参数 γ_i, β_i , 从而自适应地调整不同时间片段的尺度映射,使特征在跨时间段预测时保持统计一致性与数值稳定性,为

后续特征提取提供稳健输入。

在特征增强阶段,设计 MFBCEM 以同时捕获负载信号中的低频趋势与高频波动。该模块基于多层 Haar 小波分解(discrete wavelet transform, DWT),将输入序列分解为不同频带的子序列,并对各子带分别施加深度可分离卷积与非线性激活,以提取不同频率下的局部与全局特征;

随后通过逆小波重构实现频域信息的融合与回流,从而实现时频特征的联合建模,强化模型对周期性模式与突变的感知能力。

在主干网络中,用 LTA 从多尺度角度提取全局趋势与长期依赖关系。该模块结合多尺度卷积与频域低通滤波操作,在抑制高频噪声的同时保留长期变化趋势,并通过门控机制与短期卷积分支自适应融合多层时序特征,从而在低计算复杂度下实现对长期依赖与短期动态的平衡建模。

整体而言,Informer-RevMF 通过“时变归一化—多频增强—长期聚合”的层次化协同机制,实现了从输入分布对齐到多尺度时空建模的统一框架。该模型能够在负载模式多变、分布动态变化的复杂云计算场景中有效提升预测的稳定性、精度与跨时间段泛化性能。

2.2 可逆归一化

在云计算负载预测任务中,数据常常表现出显著的非平稳性与分布偏移,例如不同时段的资源使用模式可能截

然不同,甚至在同一时间窗口内也可能存在大幅波动。这类问题不仅会降低深度预测模型的泛化能力,还会导致训练过程中出现收敛缓慢、预测结果偏移等现象。传统的批量归一化(batch normalization)或者层归一化(layer normalization)通常依赖于全局统计量,难以在非平稳序列数据上发挥理想效果。针对分布漂移的问题,Kim 等^[26]提出的 RevIN 被证明能有效缓解训练-推理不一致,从而提高在跨时段预测任务中的稳健性。但原始 RevIN 更多作为一个“即插即用”的预处理模块应用在通用时间序列上,通常采用固定形式的仿射变换,未显式引入业务阶段、时间位置等上下文信息,也未考虑云原生在线推理场景下的算子约束与时延开销。若直接套用在长序列 Informer 上,在强场景切换和严格延迟约束下容易暴露出适应性与可部署性不足的问题。

为此,本文在 RevIN 的基础上提出 CA-RevIN 模块,具体如图 6 所示。

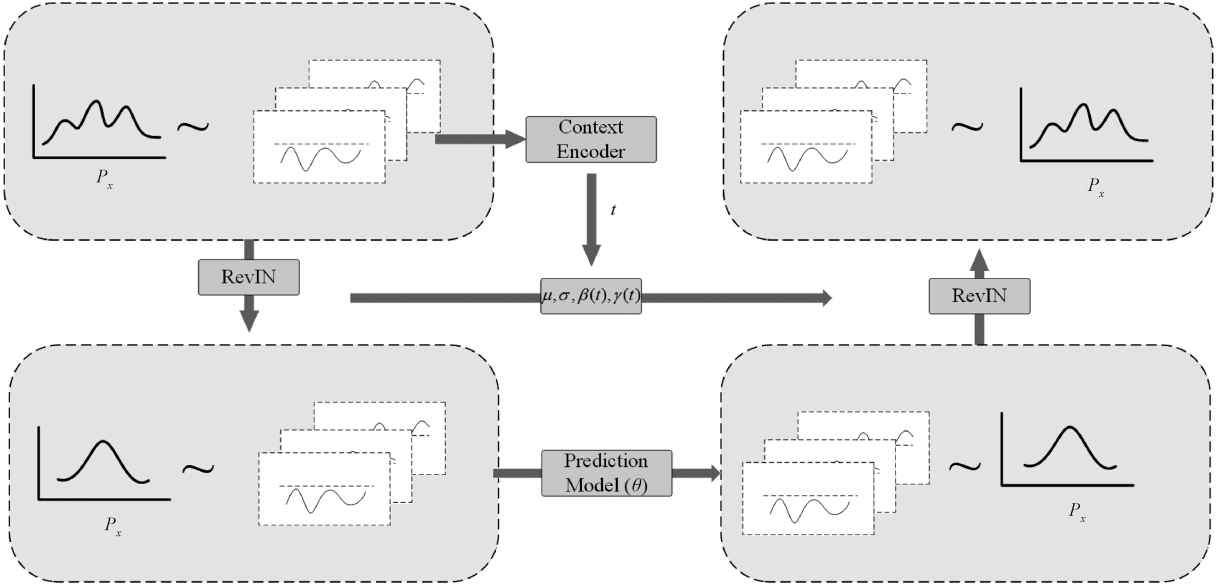


图 6 可逆归一化模块结构

Fig. 6 Reversible normalization module structure

设历史输入片段为 $\mathbf{X} \in \mathbb{R}^{B \times L \times C}$ (批大小 B 、长度 L 、通道数 C)，首先在编码端对每个样本与通道沿时间维统计均值与标准差：

$$\mu_{b,c} = \frac{1}{L} \sum_{t=1}^L \mathbf{X}_{b,t,c} \quad (1)$$

$$\sigma_{b,c} = \sqrt{\frac{1}{L} \sum_{t=1}^L (\mathbf{X}_{b,t,c} - \mu_{b,c})^2 + \epsilon} \quad (2)$$

其中, ϵ 为数值稳定项。随后对输入作实例级标准化并施加逐通道可学习仿射变换：

$$\mathbf{Z}_{b,t,c} = \alpha_c \cdot \frac{\mathbf{X}_{b,t,c} - \mu_{b,c}}{\sigma_{b,c}} + \beta_c \quad (3)$$

得到分布对齐后的序列 $\mathbf{Z}$ 送入主干网络 f_θ 以获得标准化域输出 $\hat{\mathbf{Z}} = f_\theta(\mathbf{Z})$ 。在 RevIN 框架中引入上下文信

息时,需要同时满足建模约束和工程约束两方面的要求。一方面,归一化与反归一化过程必须严格基于当前时间窗口的可观测信息,以保持因果性和可逆性,避免通过跨窗口聚合引入任何形式的未来信息泄露;另一方面,还需符合 ONNX 及云原生推理引擎对算子类型和推理时延的限制,不宜采用复杂控制流或大规模深层网络结构。基于上述考虑,本文将上下文编码器设计为仅接收当前窗口的均值、方差及时间位置等局部统计量作为输入,采用由一维归约运算与小规模全连接卷积算子构成的轻量结构来生成仿射参数。该设计在不破坏 RevIN 可逆属性和因果性的前提下,实现了对不同时段负载分布的自适应对齐,同时保证整体计算图可以顺利导出为 ONNX,并在云原生推理引擎中以较低开销稳定运行。

$$\hat{\mathbf{Y}}_{b,\tau,c} = \left(\frac{\hat{\mathbf{Z}}_{b,\tau,c} - \beta_c}{\alpha_c} \right) \cdot \sigma_{b,c} + \mu_{b,c}, \tau = 1, \dots, T \quad (4)$$

从而将预测恢复到原分布空间。训练阶段在原量纲上最小化目标函数为:

$$L = \frac{1}{BTC} \sum_{b,\tau,c} \ell(\hat{\mathbf{Y}}_{b,\tau,c}, \mathbf{Y}_{b,\tau,c}^*) \quad (5)$$

其中, ℓ 可取 MAE、MSE、SMAPE 等。推理阶段同样仅基于当前上下文 $\mathbf{X}$ 计算 (μ, σ) 并完成正逆映射, 避免信息泄露。式(4)和(5)体现了 RevIN 的两点关键性质: 1) 标准化与反标准化均在“样本 $\times$ 通道”的粒度上、沿时间维进行, 不依赖批间或全局统计, 因而能够在不破坏时序依赖的前提下有效抑制不同时间段、不同节点带来的均值与方差漂移; 2) 仿射参数 α, β 为逐通道可学习量, 初值 $\alpha = 1, \beta = 0$, 模型可根据任务自动调整保留或放大有用尺度, 从而避免固定标准化造成的信息损失。将 CA-RevIN 嵌入到包含 LTA 与 MFBCEM 的预测框架时, 整体计算流程可表为:

$$\hat{\mathbf{Y}} = \text{RevIN}^{-1}(f_{\theta}(\text{RevIN}(\mathbf{X}))) \quad (6)$$

内部的 LTA 负责通过快速傅里叶变换 (fast Fourier transform, FFT)-低通-逆快速傅里叶变换 (inverse fast Fourier transform, iFFT) 提取长期成分并与短期卷积分支门控融合, MFBCEM 通过时域小波多频段增强刻画突发

细节; CA-RevIN 则在输入与输出两端动态对齐分布, 使主干网络专注于形状与相关结构而非幅值漂移。整体实现仅依赖一维归约和逐元素运算等基础算子, 时间与空间复杂度近似为 $O(BLC)$ 且仅需缓存 $(\mu, \sigma) \in \mathbb{R}^{B \times 1 \times C}$, 工程代价可忽略。

2.3 多频段卷积增强

云计算负载信号通常由多种模式叠加而成, 既包含长期趋势和周期性模式, 也存在由突发请求、发布操作或资源争用引起的短时峰值与高频噪声。如果仅依赖单一卷积或注意力机制进行特征提取, 往往难以同时兼顾高频与低频信息。大卷积核或全局注意力更偏向于平滑趋势, 对局部尖峰和快速震荡的敏感性不足, 仅针对局部的卷积或窗口化注意力又难以捕捉跨窗口的长期依赖。已有工作如 FEDformer、WFTNet 等将傅里叶或小波变换引入 Transformer, 在频域增强全局模式建模能力, 但多采用单次或固定层数的频域变换, 频率分辨率较粗, 对云原生场景中多尺度不规则波动和厚尾尖峰的刻画仍不够充分, 同时频域算子和复杂结构也会增加在线推理的时延与实现难度。

为更精细地刻画云负载中的多尺度模式, 并在复杂度可控的前提下增强模型对频域结构的感知, 本文提出了 MFBCEM。其详细结构如图 7 所示。

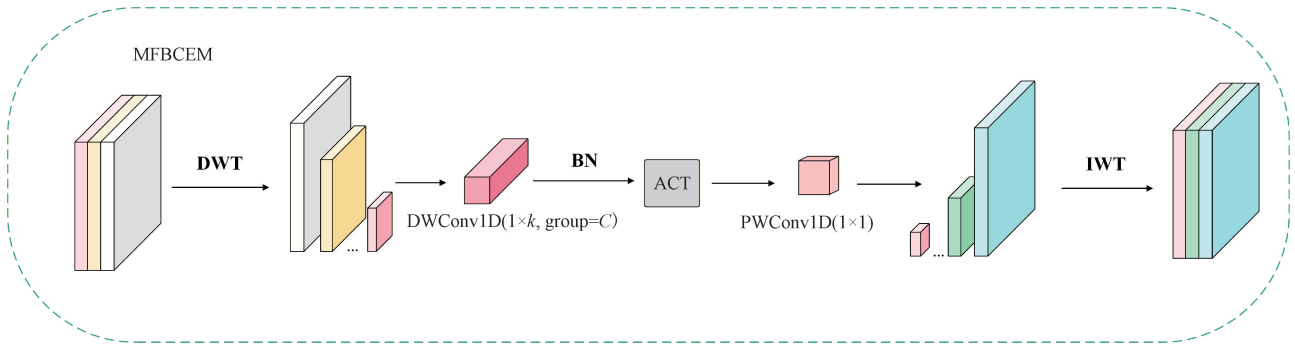


图 7 多频段卷积增强模块结构

Fig. 7 MFBCEM module structure

该模块首先利用多层 DWT 将输入序列分解为不同频段的低频与高频分量。低频部分能够反映全局趋势与长期依赖, 高频部分则包含负载变化的细节与突发模式, 具体过程如式(7)、(8)所示。

$$\begin{cases} \mathbf{L}_1 = (\mathbf{X} * \mathbf{g}) \downarrow 2 \\ \mathbf{H}_1 = (\mathbf{X} * \mathbf{h}) \downarrow 2 \end{cases} \quad (7)$$

$$\mathbf{L}_k, \mathbf{H}_k = \text{DWT}(\mathbf{L}_{k-1}), k = 2, \dots, K \quad (8)$$

其中, $\mathbf{g}, \mathbf{h}$ 为 Haar 分解滤波器, $\downarrow 2$ 表示 2 倍下采样, 与在整段序列上一次性做频谱变换不同, 多层 DWT 构成了一个可逆的小波金字塔, 可以在多个尺度上逐级分解与重构时间序列, 这是 MFBCEM 在结构上的核心设计。

随后, 对每一层的每个子带 $\mathbf{S} \in \{\mathbf{L}_k, \mathbf{H}_k\}$ 施加深度可

分离卷积以在该频带上进行轻量增强:

$$\tilde{\mathbf{S}} = \underbrace{\mathbf{W}_{\text{pw}}^{(k)} * \phi(\text{BN}(\mathbf{W}_{\text{dw}}^{(k)} * \mathbf{S}))}_{\text{DWConv}(1 \times k_s, \text{group}=C) \rightarrow \text{BN} \rightarrow \text{Act} \rightarrow \text{PWConv}(1 \times 1)} \quad (9)$$

其中, $\mathbf{W}_{\text{dw}}^{(k)}$ 为逐通道一维卷积核 (可带膨胀 d_s), $\mathbf{W}_{\text{pw}}^{(k)}$ 为 1×1 逐点卷积, $\phi(\cdot)$ 为非线性激活 (RELU), BN 为批量归一化, 参数可按层共享或分带独立。

该设计方法旨在增强对局部尖峰和不同频率成分的表达能力的同时, 控制住卷积核大小和通道维度, 以保证整体计算量与原始 Informer 同一量级。为此本文采用深度可分离卷积与逐点卷积的结构, 在不牺牲表达能力的前提下显著降低参数量和计算开销, 使 MFBCEM 可以安全地嵌入在线推理链路。

增强后的子带按自底向上逐层重构以恢复原长度:

$$\mathbf{L}_{K-1}^* = \text{IWT}(\tilde{\mathbf{L}}_K, \tilde{\mathbf{H}}_K) = (\uparrow 2\tilde{\mathbf{L}}_K) * \tilde{\mathbf{g}} + (\uparrow 2\tilde{\mathbf{H}}_K) * \tilde{\mathbf{h}} \quad (10)$$

$$\mathbf{L}_{k-1}^* = \text{IWT}(\mathbf{L}_k^*, \tilde{\mathbf{H}}_k), k = K-1, \dots, 1, \tilde{\mathbf{X}} = \mathbf{L}_0^* \quad (11)$$

为进一步融合多频段信息, $\tilde{\mathbf{X}}$ 通过一维线性映射回特征空间并与输入残差相加、归一化得到模块输出:

$$\mathbf{U} = \mathbf{W}_{\text{map}} * \tilde{\mathbf{X}} + \mathbf{b} \quad (12)$$

$$\mathbf{X}_{\text{out}} = \text{LayerNorm}(\mathbf{X} + \mathbf{U}) \quad (13)$$

上述流程以可逆的小波金字塔在不同频段内进行时域模式建模(DWConv)与通道交互(PWConv),既保留了 $\mathbf{L}_k$ 中的长期趋势与周期性结构,又在 $\mathbf{H}_k$ 中强化了峰值与快速震荡等细节;逐层IWT保证信息不失真地回流至原长度,残差与归一化提高了训练稳定性与跨场景泛化。实验表明,相比仅依赖时域卷积或单次频域变换的方案,MFBCEM对多尺度波动和厚尾尖峰具有更好的刻画能力。

2.4 长时依赖聚合

长期依赖信息的建模能力直接决定了多步外推的准确性与稳定性。传统的基于自注意力(Self-Attention)的

模型在理论上可以捕获长程依赖,但在长序列上的计算复杂度较高,对噪声和局部异常也较为敏感。例如Autoformer等部分工作通过自相关分解或深度分解结构在一定程度上缓解了这一问题,但往往需要引入较为复杂的结构设计或额外分解分支,在云原生场景的在线推理中会带来额外的实现和时延成本。

针对上述不足,本文设计的长时依赖聚合模块,在不改变Informer主干自注意力结构的前提下,通过多尺度卷积与频域低通滤波的结合,对第2.3节多频段卷积增强后的特征进行进一步的长期建模,如图8所示。具体而言,LTA首先采用不同感受野的一维卷积提取局部与中长期依赖特征,得到多尺度时域表示;随后在特征层面引入FFT,对其中一条长期分支执行频域低通滤波,将高频噪声和局部尖峰从长期通道中显式压制,仅保留缓慢变化的趋势成分。与直接在原始序列上做频域操作不同,这种“在特征空间做低通”的方式可以减少对相位的破坏,同时将FFT、iFFT的计算复杂度控制在与注意力同一量级,避免对整体推理时延造成明显影响,这是在长序列场景下引入频域算子时需要重点平衡的实现问题之一。

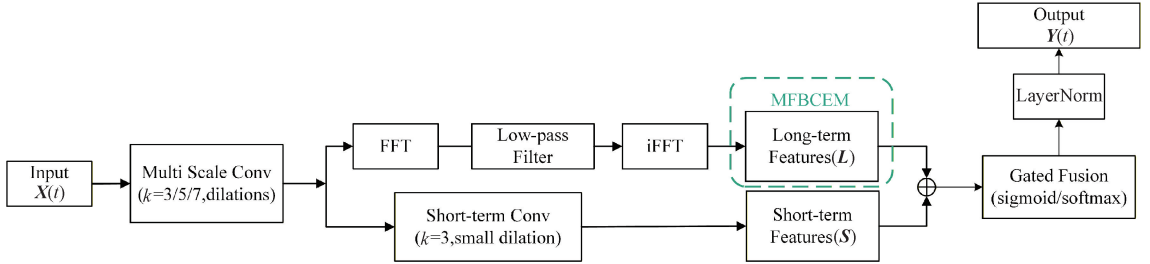


图8 长时依赖聚合模块结构

Fig. 8 Long-term aggregation module structure

考虑到完全低通会损失对中高频负载变化的刻画能力,LTA在主干路径中保留一条短期卷积分支,与长期低通分支并行,对局部动态和快速波动进行建模。随后通过门控机制自适应地融合短期特征与多尺度长期特征;门控权重由当前特征自适应生成,用于决定在不同时间位置下“更依赖长期趋势还是短期变化”,从而在频域平滑与时域敏感性之间取得折中。最终,融合后的输出经过层归一化处理,进一步提升网络训练的稳定性。

实验结果表明,引入LTA后,Informer-RevMF在长步多步预测场景下的误差增长显著放缓,体现出更好的长期外推能力和数值稳定性。

3 实验验证

本研究首先基于GWA-T-12(Bitbrains)数据集内的两类轨迹FastStorage与Rnd开展离线评测,以验证所提Informer-RevMF在典型数据中心负载上的有效性与鲁棒性。该数据集包含1750台虚拟机(virtual machine, VM)

的多维性能指标,原始采样间隔为300 ms。对全部VM进行统一预处理:缺失修复与异常点剔除,按1 s重采样(每3点取均值以平滑噪声),并构造CPU、内存、磁盘与网络四大类核心特征。数据按时间顺序7:1:1划分为训练集、验证集和测试集,确保无信息泄露。离线阶段采用MAE、MSE、RMSE等指标,这些指标采用以下形式:

$$E_{\text{MAE}} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (14)$$

$$E_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (15)$$

$$E_{\text{RMSE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (16)$$

在完成离线验证后,将模型以ONNX部署至所构建的预测—监控系统中,通过采集插件实时拉取目标主机、容器、进程的指标流,按与离线一致的预处理管线在线推理,输出点预测与区间置信度,并写回时序库用于可视化与告警^[26]。系统层面以预测准确率与误差指标为主,同时

统计分级告警的 Precision、Recall、提前量与端到端延迟，并对不同场景进行分桶评测。通过“FastStorage 和 Rnd 的离线结果”与“线上闭环的系统指标”两条证据链相互印证，确保实验的有效性与完整性^[27]。

3.1 实验环境及数据

实验在 6 核 CPU(型号 Intel i5-12400F)、16 GB 的 DDR4 内存、NVIDIA GeForce RTX4060 12 GB GPU、win10 操作系统。在 Bitbrains GWA-T 的两条典型轨迹上评测：Rnd 与 FastStorage。Rnd 的特点是波动高、多尖峰，但是 FastStorage 则更规整。预测步长设为 24/48/96，评价指标为 MAE/MSE/RMSE。对比基线包含 ARIMA、LSTM、GRU、Informer^[8]、Crossformer^[9]、FEDformer^[10]以及 CL-Informer^[18]，本文方法为 Informer-RevMF。具体实验结果如表 1、2 所示。

表 1 Rnd 数据集的预测结果对比

Table 1 Comparison of prediction results on the Rnd dataset

模型	指标	预测步长	预测步长	预测步长
		为 24	为 48	为 96
ARIMA	MAE	0.890	0.962	1.103
	MSE	1.288	1.415	1.632
	RMSE	1.135	1.189	1.278
LSTM	MAE	0.743	0.801	0.905
	MSE	1.021	1.132	1.282
	RMSE	1.010	1.064	1.133
GRU	MAE	0.718	0.779	0.872
	MSE	0.986	1.098	1.256
	RMSE	0.993	1.048	1.121
Informer	MAE	0.682	0.747	0.843
	MSE	0.902	1.012	1.189
	RMSE	0.949	1.006	1.091
Crossformer	MAE	0.564	0.667	0.811
	MSE	0.589	0.742	1.048
	RMSE	0.768	0.861	1.024
FEDformer	MAE	0.528	0.604	0.751
	MSE	0.545	0.695	0.992
	RMSE	0.738	0.834	0.996
CL-Informer	MAE	0.512	0.634	0.825
	MSE	0.520	0.769	1.181
	RMSE	0.721	0.877	1.087
Informer-RevMF	MAE	0.489	0.601	0.725
	MSE	0.511	0.671	0.911
	RMSE	0.715	0.819	0.953

基于 BitbrainsGWA-T 数据集的 Rnd 与 FastStorage 两条典型轨迹，本文在 24、48 和 96 步 3 种预测步长下，对 ARIMA、LSTM、GRU、Informer、Crossformer、FEDformer、

表 2 FastStorage 数据集的预测结果对比
Table 2 Comparison of prediction results on the FastStorage dataset

模型	指标	预测步长	预测步长	预测步长
		为 24	为 48	为 96
ARIMA	MAE	0.801	0.866	0.993
	MSE	1.159	1.274	1.469
	RMSE	1.077	1.129	1.212
LSTM	MAE	0.654	0.705	0.796
	MSE	0.898	0.996	1.128
	RMSE	0.948	0.998	1.062
GRU	MAE	0.632	0.686	0.767
	MSE	0.868	0.996	1.105
	RMSE	0.932	0.983	1.051
Informer	MAE	0.587	0.642	0.725
	MSE	0.776	0.870	1.023
	RMSE	0.881	0.933	1.011
Crossformer	MAE	0.479	0.567	0.689
	MSE	0.501	0.631	0.891
	RMSE	0.708	0.794	0.944
FEDformer	MAE	0.449	0.513	0.638
	MSE	0.463	0.591	0.843
	RMSE	0.680	0.769	0.918
CL-Informer	MAE	0.412	0.783	0.632
	MSE	0.389	1.452	0.769
	RMSE	0.624	1.205	0.877
Informer-RevMF	MAE	0.401	0.493	0.594
	MSE	0.419	0.550	0.747
	RMSE	0.647	0.742	0.864

CL-Informer 与本文改进模型 Informer-RevMF 进行了系统评测，评价采用反归一化后的 MAE、MSE 与 RMSE。结果表明，随预测步长增加，各模型误差呈单调上升，且 3 项指标之间的关系符合度量一致性；在所有设置中，整体性能排序保持稳定，依次为 Informer-RevMF、CL-Informer、FEDformer、Crossformer、Informer、GRU 与 LSTM(水平相近)，最后为 ARIMA。相较于原始 Informer，Informer-RevMF 在 Rnd 轨迹上将 MAE 分别降低 13.3%、9.9% 和 10.6%，MSE 分别降低 13.2%、9.6% 和 13.1%；三步平均 MAE 由 0.681 降至 0.605，三步平均 MSE 由 0.793 降至 0.698。在 FastStorage 轨迹上，Informer-RevMF 的 MAE 降幅分别为 16.3%、13.1% 和 13.8%，MSE 降幅分别为 16.4%、12.8% 和 16.2%；三步平均 MAE 由 0.578 降至 0.496，三步平均 MSE 由 0.674 降至 0.572。

与 3 种代表性的长序列 Transformer 基线 Crossformer、FEDformer 和 CL-Informer 相比，Informer-RevMF 在 Rnd 与 FastStorage 两条轨迹上整体保持最优。Crossformer

与 FEDformer 分别代表了跨维度依赖建模和频域增强两条主流改进路径,在两组数据上的 MSE 普遍明显优于原始 Informer,但在长步预测(如 96 步)时误差仍有较快累积;CL-Informer 通过小波多尺度编码在 FastStorage 24 步上取得了当前最优的短步 MSE(0.389),说明时频增强对规整负载的短期预测具有一定优势,但在高波动的 Rnd 轨迹及较长预测步长下稳定性不足,48/96 步 MSE 明显高于 FEDformer 和 Informer-RevMF。上述证据表明,基于 CA-RevIN 的分布对齐、MFBCEM 的多频段增强与 LTA 的长期依赖聚合能够协同提升长序列多步预测的稳定性与准确性,具有良好的工程应用价值。

3.2 消融实验

在前述对比实验部分中 Informer-RevMF 在各步长下均实现了显著性能提升,尤其在长步预测中表现出更优的稳定性与鲁棒性。为了进一步验证模型中各模块设计的独立贡献与协同效果,本节在相同实验条件下开展了消融实验。通过逐步引入 CA-RevIN、MFBCEM 与 LTA 这 3 个核心改进模块,来量化它们对模型预测精度与外推能力的影响,揭示 Informer-RevMF 性能提升的内在机理,实验结果如表 3、4 所示。

表 3 Rnd 数据集的消融结果对比

Table 3 Ablation study results on the Rnd dataset

模型	指标	预测步长为 24	预测步长为 48	预测步长为 96
Informer	MAE	0.564	0.667	0.811
	MSE	0.589	0.742	1.048
	RMSE	0.767	0.861	1.024
+CA-RevIN	MAE	0.538	0.641	0.783
	MSE	0.565	0.711	1.012
	RMSE	0.752	0.843	1.006
+CA-RevIN+MFBCEM	MAE	0.516	0.620	0.761
	MSE	0.545	0.690	0.980
	RMSE	0.738	0.831	0.990
+CA-RevIN+MFBCEM+LTA	MAE	0.489	0.601	0.725
	MSE	0.511	0.671	0.911
	RMSE	0.715	0.819	0.954

基于两组消融结果可以得出一致结论:在 Rnd 与 FastStorage 两条轨迹上,从基线 Informer 出发,依次引入 CA-RevIN、MFBCEM 与 LTA 将误差稳定、单调地降低,且这一趋势在 3 种步长上均成立。首先,CA-RevIN 显著缓解了跨时段分布与尺度漂移,对短中步长最为直接:以 Rnd 为例,24 步 MAE 由 0.564 降至 0.538, RMSE 由 0.767 降至 0.752;在 FastStorage 上,24 步 MAE 由 0.479 降至 0.452、RMSE 由 0.708 降至 0.688。其次,加入 MFBCEM 后对高频扰动与局部突发的刻画进一步增强,

表 4 FastStorage 数据集的消融结果对比

Table 4 Ablation study results on the FastStorage dataset

模型	指标	预测步长为 24	预测步长为 48	预测步长为 96
Informer	MAE	0.479	0.567	0.689
	MSE	0.501	0.631	0.891
	RMSE	0.708	0.794	0.944
+CA-RevIN	MAE	0.452	0.536	0.650
	MSE	0.473	0.596	0.842
	RMSE	0.688	0.772	0.918
+CA-RevIN+MFBCEM	MAE	0.435	0.514	0.620
	MSE	0.458	0.575	0.806
	RMSE	0.677	0.758	0.898
+CA-RevIN+MFBCEM+LTA	MAE	0.401	0.493	0.594
	MSE	0.419	0.550	0.747
	RMSE	0.647	0.742	0.864

误差继续下降:Rnd 的 48 步 MAE 由 0.641 降至 0.620, FastStorage 的 96 步 RMSE 由 0.918 降至 0.898。最后,引入 LTA 形成完整模型时,对长步外推的稳健性改善最为明显:Rnd 的 96 步 RMSE 由 0.990 进一步降至 0.954, FastStorage 的 96 步 RMSE 由 0.898 降至 0.864。总体来看,3 个模块在功能上具有互补性:CA-RevIN 保障训练与推理的一致性,MFBCEM 强化时频多尺度信息,LTA 聚合长期依赖并抑制长步误差累积;三者叠加带来跨数据分布与跨步长的协同收益,验证了所提架构在真实复杂负载下的有效性与可推广性。

3.3 预测结果

在经过对比实验与消融实验的系统分析后,模型各组成模块(CA-RevIN、MFBCEM、LTA)的独立贡献及协同增益已得到充分验证。实验结果表明,改进后的 Informer-RevMF 能够在不同数据分布与步长下实现误差的稳定下降与预测精度的持续提升,具备较强的跨场景适应性与长步鲁棒性。然而,公开数据集在业务复杂度与噪声结构上仍存在一定理想化特征,难以完全反映真实云原生环境中的动态负载特性。

因此,为进一步验证模型在实际监控场景下的可用性与泛化能力,本文将改进后的预测模块部署于自研的云原生性能监测系统中,利用系统实时采集的 CPU 使用率数据开展实证分析。通过对 Rnd 3 个月和 FastStorage 的预测结果进行可视化对比与误差评估,验证模型在真实业务数据上的稳定性、响应性及不确定性刻画能力,为后续的容量规划与异常告警提供工程依据,具体结果如图 9 所示。

从 4 个窗口化对比可知,模型在各时段均能较好跟踪 CPU 使用率的基线与中等幅度波动:在长时间的平稳区间,预测曲线与实测曲线基本重合,仅在尖峰前后出现可

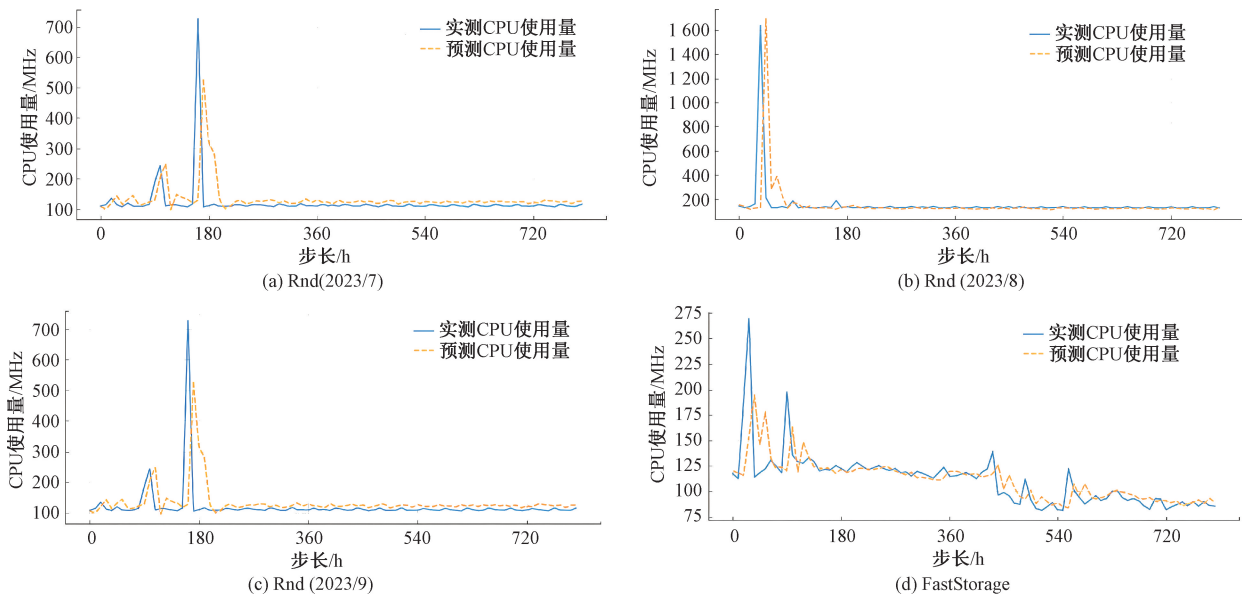


图 9 预测结果

Fig. 9 Prediction results

见偏差。对于 Rnd 的 3 个月如图 9(a)~(c) 所示,模型能较准确地捕捉到尖峰出现的时间位置与回落形态,但普遍存在峰值幅度低估与上升沿轻微滞后和提前的现象,体现了高波动、厚尾分布下对极端值的保守性;峰后衰减段的拟合较为稳定,噪声被有效平滑。FastStorage 的 1 个月如图 9(d),则表现更为紧凑,预测与实测在峰值之外几乎同轨,峰值低估程度相对 Rnd 明显减轻,说明在负载更规整的场景中模型具有更好的可拟合性与稳健性。总体而言,方法对趋势与周期成分的刻画准确,主要误差集中在突发尖峰的幅值而非时序,符合多步回归在极端波动下的典型偏差特征。

3.4 实机验证

在窗口化实验中,模型在不同时间窗口与预测步长下均表现出良好的趋势跟踪与局部波动刻画能力,能够在短中期范围内准确捕捉负载变化的节律特征与突发响应。为验证在 BitbrainsGWA-T 上得到的模型在真实场景中的可用性,将 Informer-RevMF 部署至自研监控系统的在线推理服务,通过系统采集的实际运行数据进行独立测试。本文对监控系统中连续采样得到的 CPU 使用率全时域数据进行了预测分析,通过将预测结果与实际观测值在全时间跨度上进行对比,可以更全面地评估模型在高波动负载下的全局适应性、基线拟合度以及对极端峰值的响应能力。相关结果如图 10 所示。

监控端采样获得应用 CPU 使用率等指标,经与离线训练一致的清洗与标准化流程处理后,按滑动起点生成 48、96 步多步预测;评价时将每个起点的预测与目标时刻的真实观测对齐,计算 MAE、MSE、RMSE 等误差指标。由于系统侧同时输出置信区间,用以服务告警与容量决

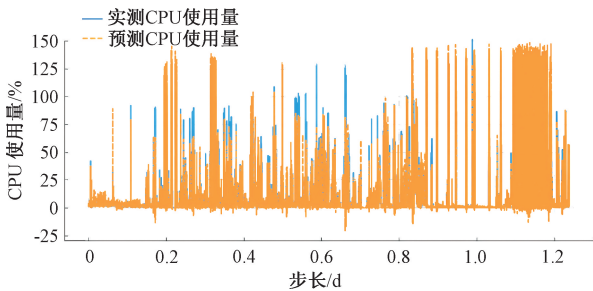


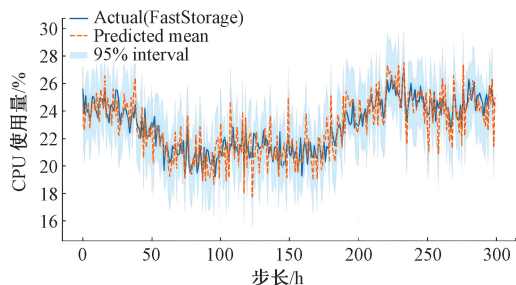
图 10 实际 CPU 使用率与预测结果全时域对比

Fig. 10 Actual CPU utilization vs predictions—fulltime-domain comparison

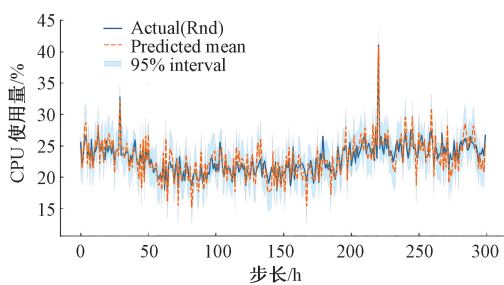
策,本文在窗口化示例中给出 95% 区间的覆盖状况作为辅助证据,如图 11 所示。

由图 11(a)可知,在中短期外推下,预测均值与实测曲线基本重合,周期性起伏与局部小峰得到准确复现。置信区间较为收敛且覆盖稳定,样本点大多落入区间,仅个别尖峰顶部出现外泄。该现象说明在 48 步这一业务常用的滚动决策跨度内,模型的不确定性刻画与校准较好,适用于容量规划、弹性伸缩与短期告警。由图 11(b)可知,当外推步长扩展到较长时域比如 96 步,区间宽度随之自适应增大,预测仍能追随基线与多数波动,但对稀有尖峰表现出更明显的保守低估,以及转折处的轻微相位误差。总体覆盖率仍处于可接受水平,越步长导致的误差主要来源于极端事件。这表明在长步外推场景下,模型保持“趋势可用、极值偏保守”的特性,适合与阈值或峰超阈方法联用,形成分级预警与提前干预。

真实数据验证与离线基准的结论一致:模型对趋势与周期成分具有稳定刻画能力,对高波动与厚尾尖峰保持保



(a) 预测结果输出 $H=48$
(a) The prediction result output horizon is set to $H=48$



(b) 预测结果输出 $H=96$
(b) The prediction result output horizon is set to $H=96$

图 11 窗口化示例

Fig. 11 Windowing example

守估计;中短期外推的区间校准更为可靠,长步外推需配合异常检测与策略缓冲共同使用。该结果说明,基于 CA-RevIN 的分布对齐、MFBCEM 的多频段增强与 LTA 的长期依赖聚合能够在训练数据之外的生产负载中保持良好的泛化与工程可用性。

4 结 论

本文面向云原生场景下多指标、强波动、长依赖的应用性能时序,提出了基于改进型 Informer 的预测—监控一体化方案 Informer-RevMF 与工程系统。方法层面,在编码端引入 CA-RevIN 以抑制跨时段分布漂移,在主干叠加 MFBCEM 与 LTA 以实现时频多尺度联合建模,并采用生成式多步解码;系统层面,构建了覆盖数据采集—特征治理—在线推理—可视化与告警的闭环,实现与现有监控平台的低成本集成。

本文的局限性在于对厚尾极端事件仍偏保守,磁盘和网络等高度随机指标的误差相对较大。后续工作将引入分位数、重尾似然与尖峰感知损失,改进极值刻画与区间校准;结合外生变量与在线和迁移学习,提升跨场景与跨时段鲁棒性;扩展到多主机、多服务的时空联合预测与分级异常检测,将预测结果直接闭环到自动扩缩容与 SLA 保障策略。总体而言,Informer-RevMF 在准确性、稳定性与可部署性上取得了兼顾,能够为复杂生产环境中的性能监控与资源调度提供有效支撑。

参考文献

- [1] 朱建军, 方琰崑. 面向服务的 5G 云原生核心网及关键技术研究[J]. 数字通信世界, 2018(2): 111.
Zhu Jianjun, Fang Yanwei. Research on service-oriented 5G cloud-native core network and key technologies[J]. Digital Communication World, 2018(2): 111.
- [2] 田兵, 王玮, 苏琦, 等. 基于微服务架构的应用性能监控平台研究[J]. 信息技术与信息化, 2018(1): 125-128.
Tian Bing, Wang Wei, Su Qi, et al. Research on application performance monitoring platform based on

microservices architecture[J]. Information Technology and Informatization, 2018(1): 125-128.

- [3] 付楠, 程光, 戴广晔, 等. 面向云原生网络的微服务系统故障诊断方法[J]. 中国科学: 信息科学, 2025, 55(8): 1888-1905.
Fu Nan, Cheng Guang, Dai Guangye, et al. Fault diagnosis method for microservice systems oriented to cloud-native networks[J]. Science China: Information Sciences, 2025, 55(8): 1888-1905.
- [4] 周浩, 冯浩, 周晨曦, 等. 基于 SSA-LSTM-Attention 的典型液压系统压力数据高精度预测方法[J]. 电子测量与仪器学报, 2025, 39(8): 241-249.
Zhou Hao, Feng Hao, Zhou Chenxi, et al. High-accuracy prediction method for pressure data of typical hydraulic systems based on SSA-LSTM-Attention[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(8): 241-249.
- [5] Shoorkand H D, Nourel F M, Hajji A. A hybrid CNN-GLSTM model for joint optimization of production and imperfect predictive maintenance planning[J]. Reliability Engineering & System Safety, 2024, 241: 109707.
- [6] 崔立志, 何泽彬, 李璇. 基于注意力的 R-GCN-GRU 的在线学生绩效预测[J]. 电子测量技术, 2021, 44(19): 69-75.
Cui Lizhi, He Zebin, Li Xuan. Online student performance prediction using attention-based R-GCN-GRU[J]. Electronic Measurement Technology, 2021, 44(19): 69-75.
- [7] 王海亮, 李敏, 刘亚翊, 等. 基于 INGO_BiLSTM_SA 的短期风电功率预测[J]. 电子测量技术, 2026, 49(7): 74-82.
Wang Hailiang, Li Min, Liu Yahong, et al. Short-term wind power forecasting based on INGO_BiLSTM_SA[J]. Electronic Measurement Technology, 2026, 49(7): 74-82.

- [8] Zhou H Y, Zhang S H, Peng J Q, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//AAAI Conference on Artificial Intelligence, 2021, 35(12): 11106-11115.
- [9] Zhang Y H, Yan J C. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting [C]//11th International Conference on Learning Representations(ICLR), 2023.
- [10] Zhou T, Ma Z Q, Wen Q S, et al. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting [C]//39th International Conference on Machine Learning (ICML), 2022: 27268-27286.
- [11] 王新科, 梅红岩, 赵勤, 等. 基于 TA-Informer 模型的多元长期时间序列预测研究[J]. 计算机工程与应用, 2026, 62(8): 366-379.
- Wang Xinke, Mei Hongyan, Zhao Qin, et al. Multivariate long-term time series forecasting based on TA-Informer model[J]. Computer Engineering and Applications, 2026, 62(8): 366-379.
- [12] Qu Z J, Meng Y, Hou X X, et al. Integrated energy short-term multivariate load forecasting based on PatchTST secondary decoupling reconstruction for progressive layered extraction multi-task learning network [J]. Expert Systems with Applications, 2025, 269: 126446.
- [13] 王国辉, 张靖, 肖龙, 等. 基于 Informer-BiGRU-Global Attention 的综合能源系统多元负荷预测[J/OL]. 电网技术, 2025-12-04. <https://doi.org/10.13335/j.1000-3673.pst.2025.1488>.
- Wang Guohui, Zhang Jing, Xiao Long, et al. Multivariate load forecasting for integrated energy systems based on Informer-BiGRU-Global Attention [J/OL]. Power System Technology, 2025-12-04. <https://doi.org/10.13335/j.1000-3673.pst.2025.1488>.
- [14] Zhang Y F, Wen Q S, Wang X, et al. OneNet: Enhancing time series forecasting models under concept drift by online ensembling [PP/OL]. V1. (2023-09-22) [2025-10-31]. <https://doi.org/10.48550/arXiv.2309.12659>.
- [15] 汪红霞, 郑诚, 黄性芳. TCformer: 一种基于 Transformer 框架下改进的在线时间序列预测模型[J/OL]. 系统科学与数学, 2025-06-10. <https://link.cnki.net/urlid/11.2019.O1.20250609.1517.002>.
- Wang Hongxia, Zheng Cheng, Huang Xingfang. TCformer: An improved online time series forecasting model under the Transformer framework [J/OL]. Systems Science and Mathematics, 2025-06-10. <https://link.cnki.net/urlid/11.2019.O1.20250609.1517.002>.
- [16] 徐楷彭, 王艳, 张欣, 等. 基于 CauDformer 模型的工业浓缩参数预测[J/OL]. 系统仿真学报, 2025-10-20. <https://doi.org/10.16182/j.issn1004731x.joss.25-0785>.
- Xu Kaipeng, Wang Yan, Zhang Xin, et al. Industrial concentration parameter prediction based on CauDformer model [J/OL]. Journal of System Simulation, 2025-10-20. <https://doi.org/10.16182/j.issn1004731x.joss.25-0785>.
- [17] 卢翔, 高新越, 王杜, 等. 基于特征优选及 Informer 的航空发动机剩余寿命预测[J]. 电子测量技术, 2025, 48(22): 10-19.
- Lu Xiang, Gao Xinyue, Wang Du, et al. Remaining useful life prediction of aero-engines based on feature selection and Informer [J]. Electronic Measurement Technology, 2025, 48(22): 10-19.
- [18] Liu B J, Li Z M, Li Z L, et al. CL-Informer: Long time series prediction model based on continuous wavelet transform[J]. PLoS One, 2024, 19(9): e0310245.
- [19] 耿宇宇, 陈华国, 王涛. 无线电能传输系统最优负载优化与最大效率跟踪控制方法[J]. 仪器仪表学报, 2025, 46(4): 23-34.
- Geng Yuyu, Chen Huaguo, Wang Tao. Optimal load optimization and maximum efficiency tracking control method for wireless power transfer systems [J]. Chinese Journal of Scientific Instrument, 2025, 46(4): 23-34.
- [20] Tang L, Yuan L, Zheng G Q, et al. DTSN: No-reference image quality assessment via deformable transformer and semantic network [C]//2024 IEEE International Conference on Image Processing (ICIP), 2024: 1207-1211.
- [21] Wu H X, Xu J H, Wang J M, et al. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting[J]. Advances in Neural Information Processing Systems, 2021, 34: 22419-22430.
- [22] Ding S Y, Wang Z H, Yang Y F, et al. Multiscale deformable transformer meets lightweight design: Enabling real-time quality detection for automobile body stamping parts under production-line constraints [J]. Signal, Image and Video Processing, 2025, 19(9): 704.
- [23] 赵伟, 王蓓, 张士祁, 等. 基于 Prometheus 的 openGauss 监控系统的关键技术及验证[J]. 郑州大学学报(理学版), 2022, 54(6): 74-81.
- Zhao Wei, Wang Bei, Zhang Shiqi, et al. Key

technologies and verification of an openGauss monitoring system based on Prometheus[J]. Journal of Zhengzhou University (Natural Science Edition), 2022, 54(6): 74-81.

[24] 刘健,程建勋,闫高峰,等. 深入理解 IoTDB 在时序数据工作负载下的性能表现[J]. 小型微型计算机系统, 2025, 46(4): 1014-1024.

Liu Jian, Cheng Jianxun, Yan Gaofeng, et al. In-depth understanding of IoTDB performance under time-series workloads [J]. Journal of Chinese Computer Systems, 2025, 46(4): 1014-1024.

[25] 段立娟,高文,王伟强. 时序数据库中相似序列的挖掘[J]. 计算机科学, 2000(5): 39-44.

Duan Lijuan, Gao Wen, Wang Weiqiang. Mining similar sequences in time series databases [J]. Computer Science, 2000(5): 39-44.

[26] Kim T, Kim J, Tae Y, et al. Reversible instance normalization for accurate time-series forecasting against distribution shift[C]//International Conference on Learning Representations(ICLR), 2022.

[27] 王昌达,李柔,张治平. 服务功能链部署的能耗与负载均衡联合优化[J]. 计算机应用研究, 2025, 42(5): 1500-1506.

Wang Changda, Li Rou, Zhang Zhiping. Joint optimization of energy consumption and load balancing for service function chain deployment[J]. Application Research of Computers, 2025, 42(5): 1500-1506.

作者简介

陈祺东,副教授,主要研究方向为智能优化。
E-mail:cqd_jnu@hotmail.com

钱家昊,硕士研究生,主要研究方向为时序预测。
E-mail:qjh165195627@163.com

吴豪(通信作者),讲师,主要研究方向为图像识别。
E-mail:860415@cw Xu.edu.cn

叶广飞,硕士研究生,主要研究方向为图像识别。
E-mail:ygf.Luck@outlook.com

张松洋,硕士研究生,主要研究方向为时序预测。
E-mail:JamesZXY0127@outlook.com