基于 xLSTM 的数据中心服务器能耗预测

刘 敏[1], 弓 昊[1], 戴文海[1], 莎日娜[1], 范永胜[1]
刘 波[1], 白玉洁[1], 刘 杨[1], 丁 博[2]
(1. 内蒙古电力(集团)有限责任公司 电能计量分公司, 呼和浩特 010010;
2. 哈尔滨理工大学 计算机科学与技术学院, 哈尔滨 150080)

摘 要:服务器能耗数据作为电力数据资产的重要组成部分,对其进行准确预测,可以优化电力使用,提高电力数据资产的经济效益。现有的能耗预测方法难以准确捕捉数据的周期性和长期变化特征及数据间的长期依赖关系。为了解决以上问题,提出了基于 xLSTM 的数据中心服务器能耗预测方法。首先,将服务器的能耗数据分为趋势性和季节性数据,有效的提取时序数据的长期变化趋势及周期内的重复模式。然后,采用 xLSTM 提取时间点之间的时序相关性及时间段之间的长期依赖关系,提高了能耗预测的准确率。将本文方法在自建数据集和4个公开数据集上与现有方法进行比较,在评价指标均方误差(mean square error, MSE)和平均绝对误差(mean absolute error, MAE)上取得了最优效果。

关键词:数据中心;能耗预测;长期时间序列;xLSTM;时间序列分解
DOI:10.15938/j.jhust.2026.04.009
中图分类号: TP316 **文献标志码**: A **文章编号**: 1007-2683(2026)04-0094-09

Data Center Server Energy Consumption Prediction Based on xLSTM

LIU Min[1], GONG Hao[1], DAI Wenhai[1], SHA Rina[1], FAN Yongsheng[1],
LIU Bo[1], BAI Yujie[1], LIU Yang[1], DING Bo[2]
(1. Electric Energy Measurement Branch, Inner Mongolia Electric Power (Group) Co., Ltd., Hohhot 010010, China;
2. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: Server energy consumption data is a crucial component of power data assets. Accurate prediction of this data can optimize power usage and enhance the economic benefits derived from power data assets. Existing prediction methods struggle to accurately capture the periodicity, long-term trends, and dependencies in energy consumption data. To address these issues, this paper proposes a server energy consumption prediction method based on xLSTM. This method first decomposes server energy consumption data into trend and seasonality components, effectively extracting long-term trends and repetitive patterns. Then xLSTM is used to capture the temporal correlations between time points and the long-term dependencies across time periods, enhancing the accuracy of energy consumption predictions. The proposed method is compared with existing methods on a self-built dataset and four public datasets, achieving the best results in terms of the evaluation indicators MAE and MSE.

Keywords: data center; energy consumption prediction; long-term time series; xLSTM; time series decomposition

收稿日期: 2024-11-07
基金项目: 国家自然科学基金面上项目(62473111);内蒙古电力公司2024年第81号科技项目(内电科创[2024]5号).
作者简介: 弓 昊(1982—),男,工程师;
戴文海(1990—),男,工程师;
丁 博(1983—),女,副教授,硕士研究生导师.
通信作者: 刘 敏(1977—),女,硕士,高级工程师,E-mail:951026811@qq.com.

0　引　言

随着大数据时代的到来,数据中心的产业规模迅速发展,能耗也逐渐增高,预计到 2025 年,全球数据中心的总能耗将占全球总能耗的 4.5%[1]。此外,数据中心带来的环境问题也不容忽视。2019 年,数据中心碳排放量约占全球碳排放总量的 2%,预计到 2030 年,数据中心的能源消耗将增加到全国电力总消耗的 3.7%,这将间接产生超过9 000万吨的二氧化碳排放量[2]。因此,数据中心能耗预测至关重要。同时,数据中心能耗数据是电力数据资产的核心组成部分,通过对能耗数据的准确预测和深入分析,数据中心可优化电力使用,降低成本,提高运营效率,还能减少碳排放,促进环境保护,并开发新的业务模式,从而显著提升电力数据资产的经济价值。

服务器、存储驱动器、网络、冷却和供电系统是数据中心的主要组成部分,其中服务器在数据中心用电量占比最大,其次是冷却和供电系统[3]。研究发现,约 15% 的服务器处于空闲状态,而处于忙碌状态的服务器中的 CPU 使用率不足 20%,但这些服务器却消耗了数据中心 43% 的电力。目前,服务器的高能耗低能效问题受到学术界和工业界的广泛关注,提高服务器能源效率、降低其能耗已成为亟待解决的关键问题[4]。

数据中心的服务器能耗数据本质上是长期时间序列数据,长期时间序列预测是根据过去预测未来的指标或事件的发展情况[5]。长期时间序列预测面临着数据更加复杂、时间跨度更大的问题。目前,基于深度学习的长期时间序列预测已成为主流方法,其中最具代表性的有循环神经网络(recurrent neural network, RNN)[6]、时序卷积网络(temporal convolutional network, TCN)[7]、多层感知器(multi-layer perceptron, MLP)[8]等。但是以上时间序列预测方法存在预测不准确、时间范围短的问题。

建立高精度能耗模型有助于实现对服务器能耗的全面评估和定量分析,有助于人们更好地理解服务器能效特性及其相关因素的影响,为节能措施的设计和优化提供有效支持,并可提高电力数据资产的经济价值。为此,本文提出了基于 xLSTM 的数据中心服务器能耗预测方法,该方法首先将服务器能耗时序数据分解为趋势性和季节性数据。趋势性是指随着时间的增加,时间序列的整体变化趋势;季节性是指相同时间间隔不断重复出现的周期性变化[9]。将时序数据分解成趋势性和季节性数据有助于分别捕捉长期变化和周期性模式,使模型可更准确地反映数据的内在规律,并且通过这种分解方式还可以简化模型结构,便于训练和优化,增强数据可解释性。此外,分解后的时序数据可过滤掉随机噪声,提高预测的稳定性和可靠性,从而提供更精确的预测结果。其次,本文采用 xLSTM 提取时间点之间的时序相关性及时间段之间的长期依赖关系,提高了能耗预测的准确率。同时,由于 xLSTM 可以更高效地关注重要的时间步长和特征,减少了对无关信息的计算,从而大大提高了长序列预测的推理速度。利用真实服务器能耗数据进行多次仿真实验,证明了本文方法的有效性,并且在 4 个公开时序数据集中进行了实验,证明了本文方法的通用性。

1　相关工作

近年来,人工智能在能耗预测中的应用已成为研究热点。能源消耗预测方法可以分为两类:传统的线性和非线性回归方法和基于深度学习的方法。

1.1　基于线性和非线性回归的能耗预测方法

早期的能耗预测方法基于线性回归[10]。随后的研究表明非线性回归模型可以更准确地描述服务器系统性能和能耗之间的关系[11]。LUO 等[12]结合多种回归技术,包括多元线性回归、逐步回归、非线性回归,总结分析了各类参数对能耗模型的影响,证明了多项式模型预测能力的优越性。该模型考虑了工作负载的变化和特征参数的多样性,大大提高了预测精度。ZHOU 等[13]利用主成分分析筛选特征参数,并将其用于多元线性和非线性回归方法的能耗建模。该模型还考虑了工作量的变化和特征参数的多样性,大大提高了预测精度。但是基于线性和非线性回归的能耗预测方法无法充分捕捉复杂系统中的高维特征和非线性关系,模型的数据拟合能力较差,预测精度较低。

1.2　基于深度学习的能耗预测方法

此前主流的能耗预测方法将能耗视为时间序列数据,采用建立时间序列模型的方法进行能耗预测[14]。近年来,由于深度学习的迅猛发展,预测的准确性和时间范围都有大幅提升,最具代表性的有 RNN、TCN、MLP 和 Transformer 等。RNN 方法利用序列数据之间的时间递延关系,改进时间序列的预测精度。DA-RNN[15]通过引入时间注意力机制和输

入注意力机制,实现对数据的有效预测。TPA-LSTM[16]首先使用滤波器提取出时间不变的模式,然后通过 1 种注意力机制和频域信息进行多元预测。TCN 方法通过卷积操作有效地捕捉时序依赖性,具体的方法有:FRANCESCHI 等[17]通过结合基于因果扩张卷积的编码器和 1 种基于时间的负采样三元组损失,学习时间序列的通用嵌入。MLP 方法是 1 种前馈神经网络,至少包括 3 层结构:输入层、隐藏层和输出层。LightTS[18]基于 MLP 结构,使用了间隔采样和连续采样的降采样策略。N-HiTS[19]通过在全连接块前加入下采样层和层次化插值进行时序依赖性建模。

Transformer 方法采用注意力机制,按输入数据各部分重要性来分配不同的权重对数据进行预测。FEDformer[20]提出了 1 种频率增强的分解块,用以更好地捕获时间序列的全局特征。Crossformer[21]是利用跨维度依赖性进行多元时间序列预测的方法。Non-stationary Transformers[22]提出了非平稳 Transformer 作为通用框架,包括序列平稳化和去平稳注意力。序列平稳化统一了每个输入的统计数据,并使用恢复的统计数据转换输出以获得更好的可预测性;去平稳注意力通过近似从原始序列中学习到的可区分注意力,将内在的非平稳信息恢复到时间依赖关系中。虽然 Transformer 是当前最流行的能耗预测方法之一,但是由于其采用自注意力机制,计算量非常大,计算效率低下,尤其是对于长期时间序列预测。

本文采用扩展长短期记忆(extended long short-term memor, xLSTM)[23]作为能耗预测模型。与 LSTM 相比,xLSTM 引入了动态门控机制并组合了多个门控单元,根据输入和当前状态自适应地调整门控的开关状态,增强了模型对不同时间步长依赖的灵活处理,并且可更细粒度地控制信息流动,提高了模型捕捉复杂模式的能力。同时,xLSTM 包含 sLSTM 和 mLSTM 两种结构,通过引入混合更新策略,增强了模型的记忆能力和对长时间依赖的捕捉能力。因此,xLSTM 具有捕捉提取时间点之间的时序相关性及时间段之间的长期依赖关系的能力,可提高能耗预测的准确性。此外,xLSTM 采用了矩阵优化和并行计算,与 Transformer 相比,减少了计算复杂度和时间,显著提升了计算效率。

2 本文方法

2.1 总体框架

服务器能耗数据可反映出服务器工作负载的实时波动,并且随着如日夜变化、周末与工作日差异等,呈现出周期性模式。除了短期波动外,服务器能耗数据还可能显示出长期趋势,如随着业务增长而逐渐增加的能耗。此外,服务器能耗数据可能会受到突发事件(如故障、升级、网络攻击等)的影响,表现出较大的波动性和不稳定性。因此,服务器能耗数据存在周期性和突变性。针对服务器能耗数据的特点,本文提出了基于 xLSTM 的数据中心服务器能耗预测方法,该方法的整体框架如图 1 所示。服务器能耗数据属于长期时序数据,本文首先采用分解模块将其分为趋势性和季节性数据,目的是有效地捕捉能耗数据的长期变化和周期性模式,使模型可更准确地反映数据的内在规律。其次,本文采用 xLSTM 提取能耗数据时间点之间的时序相关性及时间段之间的长期依赖关系,提高能耗预测的准确率和推理速度。

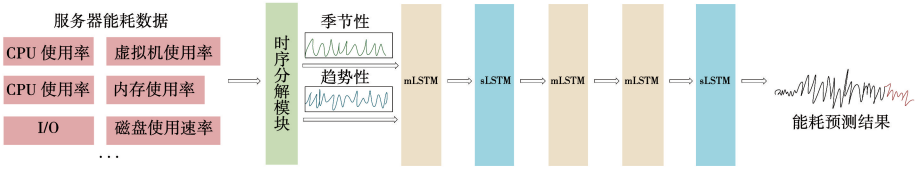


图 1 总体框架

Fig. 1 Overall framework

2.2 长期时间序列分解

在长期时间序列预测中,识别时间序列数据的潜在模式至关重要。时间序列分解技术可将时间序列分解为多个组成部分,包括趋势、季节性和余数,提高时间序列预测准确率。

为了准确地预测服务器能耗,本文采用分解模型,将服务器能耗数据分解为趋势性和周期性两类,在复杂情况下提取更多可预测的成分。现有的时间序列分解方法将时间序列分解为几个部分,每个部分代表可预测的底层模式类别之一,主要用于探索

历史随时间的变化。对于预测任务,分解总是用作预测未来序列之前对历史序列的预处理[24-25]。然而,在预测环境下,这种预处理受到历史序列简单分解效果的限制,忽略了长期未来序列底层模式之间的层次相互作用。

针对以上问题,本文改进现有分解方法的预处理方式,采用 1 种通用架构,使深度预测模型具有内在的渐进分解能力。此外,这种分解可缓解时间模式,突出时间序列的固有属性。本文提出的分解模块 SerisesDecomp 可从预测的中间隐藏变量中逐步提取长期平稳趋势。具体来说,我们调整移动平均线以平滑周期性波动并突出长期趋势。对于长度为 I 的时间序列 $\boldsymbol{\chi} \in \mathbb{R}^{I\times d}$,处理过程如下:

$$\boldsymbol{\chi}_t = \mathrm{AvgPool}(\mathrm{Padding}(\boldsymbol{\chi})) \tag{1}$$

$$\boldsymbol{\chi}_s = \boldsymbol{\chi} - \boldsymbol{\chi}_t \tag{2}$$

式中:$\boldsymbol{\chi}_s, \boldsymbol{\chi}_t \in \mathbb{R}^{I\times d}$ 分别为提取的季节性和趋势性部分。采用 AvgPool(·)进行移动平均,并进行填充操作以保持序列长度不变。因此,$\boldsymbol{\chi}_s$ 和 $\boldsymbol{\chi}_t$ 可表示为 $\boldsymbol{\chi}_s, \boldsymbol{\chi}_t = \mathrm{SeriesDecomp}(\boldsymbol{\chi})$。

接下来,本文将分解模型得到的服务器能耗预测时间序列的季节性和趋势性数据输入至 xLSTM 中。输入部分是过去 I 个时间步 $\boldsymbol{\chi}_{\mathrm{en}} \in \mathbb{R}^{I\times d}$。通过分解架构可得季节性部分 $\boldsymbol{\chi}_{\mathrm{des}} \in \mathbb{R}^{(\frac{I}{2}+O)\times d}$ 和趋势周期部分 $\boldsymbol{\chi}_{\mathrm{det}} \in \mathbb{R}^{(\frac{I}{2}+O)\times d}$,每次初始化由两部分组成:从输入 $\boldsymbol{\chi}_{\mathrm{en}}$ 的后半部分分解出的长度为 $\frac{I}{2}$ 的部分,用于提供最新信息;长度为 O 的占位符,由标量填充。

其公式如下:

$$\boldsymbol{\chi}_{\mathrm{ens}}, \boldsymbol{\chi}_{\mathrm{ent}} = \mathrm{SeriesDecomp}\left(\boldsymbol{\chi}_{\mathrm{en}\,\frac{I}{2}:I}\right) \tag{3}$$

$$\boldsymbol{\chi}_{\mathrm{des}} = \mathrm{Concat}(\boldsymbol{\chi}_{\mathrm{ens}}, \boldsymbol{\chi}_0) \tag{4}$$

$$\boldsymbol{\chi}_{\mathrm{det}} = \mathrm{Concat}(\boldsymbol{\chi}_{\mathrm{ent}}, \boldsymbol{\chi}_{\mathrm{Mean}}) \tag{5}$$

式中:$\boldsymbol{\chi}_{\mathrm{ens}}, \boldsymbol{\chi}_{\mathrm{ent}} \in \mathbb{R}^{\frac{I}{2}\times d}$ 分别为 $\boldsymbol{\chi}_{\mathrm{en}}$ 的季节部分和趋势周期部分;$\boldsymbol{\chi}_0, \boldsymbol{\chi}_{\mathrm{Mean}} \in \mathbb{R}^{O\times d}$ 分别为用零填充的占位符和 $\boldsymbol{\chi}_{\mathrm{en}}$ 的平均值。

2.3　xLSTM

20 世纪 90 年代,长短期记忆(long short-term memory, LSTM)[26]通过引入恒定误差轮播和门控,克服了循环神经网络的梯度消失和梯度爆炸问题。LSTM 的记忆单元包含 3 个门:输入门、输出门和遗忘门。输入门控制新的信息能否进入记忆单元;遗忘门决定记忆单元中哪些信息需要被遗忘;输出门控制从记忆单元输出的信息量。目前,LSTM 已被广泛应用于时间序列预测,并取得了显著效果。但 LSTM 存在 3 个主要局限性:1)无法修改存储决策:当找到更相似的向量时,LSTM 很难修改存储的值;2)存储容量有限:信息必须压缩成标量单元状态。由于存储容量有限,LSTM 在稀有标记上的表现较差;3)由于内存混合而缺乏可并行性:从一个时间步骤到下一个时间步骤的隐藏状态之间的隐藏 - 隐藏连接强制进行顺序处理。这是 LSTM 的最大问题,因为它无法充分利用 GPU 的可并行性。

服务器能耗数据具有数据量大、存在周期性和突变性等特点。大量研究表明,同一时间序列的不同部分及同一数据集内的不同时间序列都表现出周期性特征,且变化明显。周期性时间序列通常表现出更多的全局依赖性,其中某一时间点的序列值会受到特定间隔内其他时间点的值的影响。另一方面,非周期性时间序列表现出更多的局部依赖性。在这种情况下,某一时间点的值可能会受到附近时间值的影响,但在较长时期内可能不存在一致且可预测的关系。因此,采用 LSTM 难以提取服务器能耗数据的局部和全局信息。

为了克服 LSTM 的局限性,本文采用扩展长短期记忆(extended long short-term memory, xLSTM)提取服务器能耗数据的特征。xLSTM 可以有效提取时间点之间的时序相关性及时间段之间的长期依赖关系,捕获服务器能耗数据的局部和全局信息,提高预测准确率。xLSTM 引入了 2 项主要修改,分别是指数门控和新型记忆结构,这 2 个修改为 LSTM 系列增添了 2 个成员:1)具有标量记忆、标量更新和记忆混合的新 sLSTM;2)具有矩阵记忆和协方差(外积)更新规则的新 mLSTM,完全可并行化。此外,xLSTM 采用多层堆叠增强了特征提取能力、Attention机制和多头注意力。这些机制帮助模型更高效地关注重要的时间步长和特征,减少对无关信息的计算,并且跳跃连接使得梯度可更有效地传递,减少了训练时间。xLSTM 的结构如图 2 所示。

本文采用 xLSTM 提取服务器能耗数据的特征,目的在于更好地捕捉时间点和时间段间的长期依赖关系,提高预测准确率。通过 xLSTM 的并行处理能力,也可提高预测效率。

2.3.1　sLSTM

sLSTM 通过指数门控增强了 LSTM。门控机制

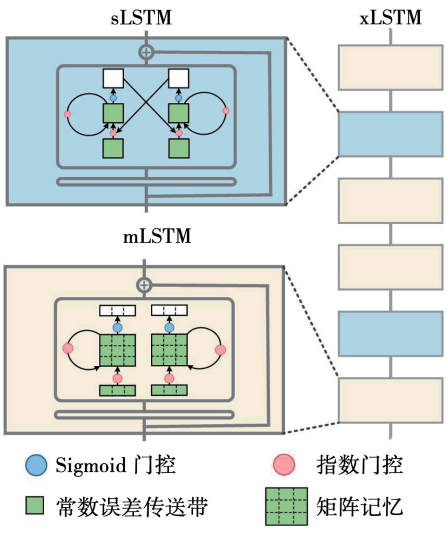


图 2 xLSTM 结构

Fig. 2 The structure of xLSTM

中的指数函数允许更动态、更灵活的更新。与将输入压缩到 (0, 1) 范围的 S 型函数不同,指数函数可以处理更大范围的值,有助于捕捉输入数据中更细微的变化,从而提高模型学习复杂模式的能力。为了实现并行化,sLSTM 可以扩展到多个记忆单元,并且 sLSTM 具有跨单元记忆混合的特点。此外,sLSTM 可以有多个头部,头部之间不进行记忆混合,但每个头部内的单元之间只进行记忆混合。为 sLSTM 引入头部并结合指数门控,建立了全新的记忆混合方式。

为了使 sLSTM 能够修改存储决策,引入了指数门及规范化和稳定化状态。特别地,输入门和遗忘门可以具有指数激活函数。输入门、遗忘门和输出门如式(6) ~ (8)所示:

$$\boldsymbol{i}_t = \exp(\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{x}_t + r_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{6}$$

$$\boldsymbol{f}_t = \exp(\boldsymbol{\omega}_f^{\mathrm{T}}\boldsymbol{x}_t + r_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{7}$$

$$\boldsymbol{o}_t = \exp(\boldsymbol{\omega}_o^{\mathrm{T}}\boldsymbol{x}_t + r_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{8}$$

式中:$\boldsymbol{\omega}_i$、$\boldsymbol{\omega}_f$ 和 $\boldsymbol{\omega}_o$ 分别为输入 $\boldsymbol{x}_t$ 和输入门、遗忘门及输出门之间的输入权重向量;r_i、r_f 和 r_o 分别为隐藏状态 $\boldsymbol{h}_{t-1}$ 和输入门、遗忘门和输出门之间的循环权重;$\boldsymbol{b}_i$、$\boldsymbol{b}_f$ 和 $\boldsymbol{b}_o$ 为偏置;$\sigma(x) = 1/(1+\exp(-x))$。

为了实现规范化,引入了规范化状态 $\boldsymbol{n}_t$,该状态等于上一状态经遗忘门衰减后,加上当前输入。

$$\boldsymbol{n}_t = \boldsymbol{f}_t\boldsymbol{n}_{t-1} + \boldsymbol{i}_t \tag{9}$$

指数激活函数可能导致较大的值,从而导致溢出。因此,通过稳定状态 $\boldsymbol{m}_t$ 来稳定门控,如式(10)所示。通过稳定状态的输入门和遗忘门如式(11) ~ (12)所示。

$$\boldsymbol{m}_t = \max(\log(\boldsymbol{f}_t) + \boldsymbol{m}_{t-1}, \log(\boldsymbol{i}_t)) \tag{10}$$

$$\boldsymbol{i}_t' = \exp(\log(\boldsymbol{i}_t) - \boldsymbol{m}_t) = \exp(\tilde{\boldsymbol{i}}_t - \boldsymbol{m}_t) \tag{11}$$

$$\boldsymbol{f}_t' = \exp(\log(\boldsymbol{f}_t) + \boldsymbol{m}_{t-1} - \boldsymbol{m}_t) \tag{12}$$

式中:$\tilde{\boldsymbol{i}}_t$ 为原始输入门。

在前向传递中采用 $\boldsymbol{i}_t'$ 和 $\boldsymbol{f}_t'$ 分别代替 $\boldsymbol{i}_t$ 和 $\boldsymbol{f}_t$ 不会改变整个网络的输出,也不会改变损失关于参数的导数。

2.3.2 mLSTM

mLSTM 具有矩阵内存和协方差更新规则,可完全并行化。mLSTM 可以与原始 LSTM 一样拥有多个记忆单元,由于没有记忆混合,mLSTM 的多个头和多个单元是等效的。为了提高存储容量,将 LSTM 存储单元从标量 $c \in \mathbb{R}$ 增加到矩阵 $\boldsymbol{C} \in \mathbb{R}^{d\times d}$。因此,检索是通过矩阵乘法执行的。在时间 t,想要存储 1 对向量,即键 $\boldsymbol{k}_t \in \mathbb{R}^d$ 和值 $\boldsymbol{v}_t \in \mathbb{R}^d$,存储键值对的协方差更新规则如下:

$$\boldsymbol{C}_t = \boldsymbol{f}_t\boldsymbol{C}_{t-1} + \boldsymbol{i}_t\boldsymbol{v}_t\boldsymbol{k}_t^{\mathrm{T}} \tag{13}$$

mLSTM 的输入门、遗忘门和输出门分别如式(14) ~ (16)所示。

$$\boldsymbol{i}_t = \exp(\boldsymbol{\omega}_i^{\mathrm{T}}\boldsymbol{x}_t + \boldsymbol{b}_i) \tag{14}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{\omega}_f^{\mathrm{T}}\boldsymbol{x}_t + \boldsymbol{b}_f) \text{ 或 } \exp(\boldsymbol{\omega}_f^{\mathrm{T}}\boldsymbol{x}_t + \boldsymbol{b}_f) \tag{15}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o\boldsymbol{x}_t + \boldsymbol{b}_o) \tag{16}$$

3 实验结果与分析

本文方法基于 PyTorch 框架,所有方法在配备了 24 GB 内存的 NVIDIA GeForce RTX 3090Ti 显卡上运行。采用了 Adam 优化器进行训练,初始学习率为0.000 1,训练时设定批量大小为 32。在训练中,如果在 10 个轮次之内没有明显提升,提前终止训练。为了得到稳定的结果,实验均重复 3 次。本文采用的时序预测长度为 96 和 192。

3.1 数据集

本文构建仿真平台收集与服务器能耗相关的数据。同时,为了评估本文方法的有效性和通用性,本文选用时序预测领域的 4 个开源数据集进行实验,包括电力变压器、交通、电力和天气[27]。其中,电力变压器选取的是间隔为 1 h 的 ETTh1 数据集,记录了 2016 年 7 月至 2018 年 7 月期间 7 种石油和电力变压器的负载特征。交通领域数据集(Traffic)记录了 2015 年至 2016 年旧金山高速公路每小时道路占

用率。电力领域数据集(Electricity)记录了 2012 年到 2014 年 321 位客户每小时电力消耗数据。天气数据集(Weather)记录了 2020 年每 10 min 的 21 项天气指标。

单一的监测工具只能监测服务器集群的基本性能指标,导致系统开销过大。因此,本文采用由 Zabbix 和 Ganglia 组成的联合监测方法,图 3 为数据采样流程。设备包括服务器、记录设备、智能电表和电源:使用智能电表获取服务器的实时能耗数据,监测服务器性能,并通过联合监测方法收集数据;记录设备用于从服务器和智能电表收集实验数据。

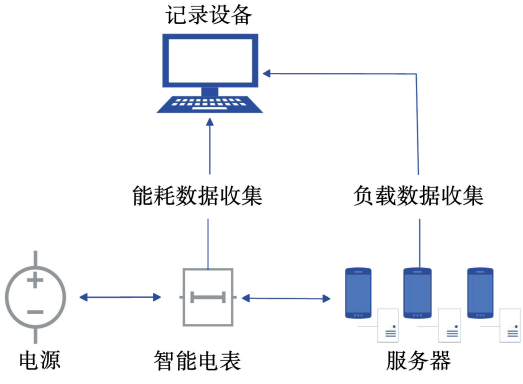


图 3　数据采样流程

Fig. 3　Data sampling process

本文自建了服务器能耗数据集,服务器拥有 4 张Nvidia 2080Ti 显卡、2 块 Intel E5-2678 处理器,4 块8 T 硬盘。服务器能耗数据集记录了服务器的能耗变化,采集的特征包含 CPU、GPU、硬盘及内存等。本文对采集到的原始服务器能耗数据进行了预处理:首先,剔除原始数据集中对服务器功耗没有影响的特征参数,并对数据集进一步清洗,即去除异常值或填充空值;然后,对特征数据进行归一化,将预处理后的数据限制在一定范围内,有助于加速基于学习的算法的收敛,本文使用 MinMax 归一化对数据进行归一化;最后,得到 12 个特征参数,服务器能耗数据集的特征如表 1 所示。

表 1　服务器能耗数据集特征

Tab. 1　Server energy consumption dataset features

特征序号	特征	描述
1	W	服务器总功耗
2	Power (GPU0)	0 号显卡功耗
3	Power (GPU1)	1 号显卡功耗
4	Power (GPU2)	2 号显卡功耗
5	Power (GPU3)	3 号显卡功耗
6	Memory Utilization	内存占用率
7	Disk0_Utilization	0 号磁盘占用率
8	Disk1_Utilization	1 号磁盘占用率
9	Disk2_Utilization	2 号磁盘占用率
10	Disk3_Utilization	3 号磁盘占用率
11	Power:package - 0,0	0 号处理器总功耗
12	Power:package - 1,0	1 号处理器总功耗

3.2　对比实验

将本文方法与 LSTM[26]、Autoformer[27]、FEDformer[20]、Crossformer[21]和 ATFNet[28]进行对比,选择均方误差(mean square error, MSE)和平均绝对误差(mean absolute error, MAE)作为评价指标,对比实验结果如表 2 所示。

表 2　不同数据集对比结果

Tab. 2　Comparison results of different datasets

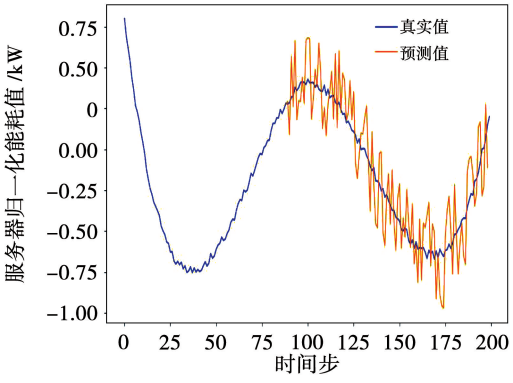
方法	时序预测长度	ETTh1		Traffic		Electricity		Weather		自建	
		α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}
LSTM	96	1.044	0.773	1.453	1.049	0.611	0.397	0.447	0.489	0.369	0.406
	192	1.217	0.832	1.846	1.179	0.617	0.393	0.453	0.497	0.416	0.435
Autoformer	96	0.505	0.475	0.670	0.438	0.346	0.388	0.266	0.336	0.297	0.323
	192	0.553	0.496	0.756	0.466	0.456	0.452	0.307	0.367	0.288	0.269

续表 2

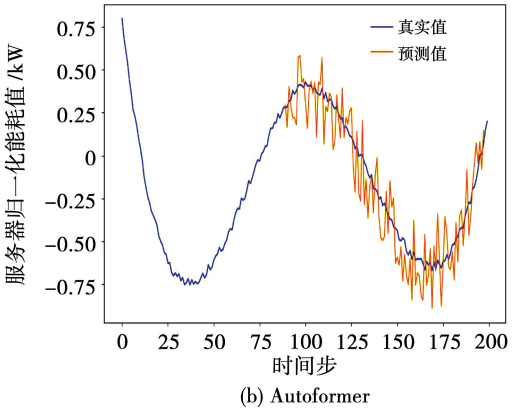
方法	时序预测长度	ETTh1		Traffic		Electricity		Weather		自建	
		α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}	α_{MSE}	α_{MAE}
FEDformer	96	0.379	0.419	0.648	0.403	0.423	0.448	0.217	0.296	0.228	0.263
	192	0.426	0.441	0.731	0.444	0.471	0.474	0.276	0.336	0.242	0.285
Crossformer	96	0.404	0.426	0.577	0.369	0.358	0.397	0.158	0.230	0.184	0.213
	192	0.450	0.451	0.616	0.392	0.429	0.439	0.206	0.277	0.217	0.251
ATFNet	96	0.374	0.398	0.322	0.265	0.210	0.305	0.237	0.265	0.201	0.258
	192	0.382	0.403	0.393	0.319	0.251	0.326	0.251	0.283	0.227	0.251
本文方法	96	0.335	0.346	0.334	0.263	0.205	0.281	0.144	0.153	0.139	0.151
	192	0.339	0.350	0.344	0.358	0.236	0.312	0.157	0.166	0.134	0.142

由表 2 可见，本文方法在自建数据集及开源数据集 ETTh1、Electricity 和 Weather 上均取得最佳结果。只有在 Traffic 数据集上，时序预测长度为 96 时，α_{MSE}位居第二。LSTM 由于难以获取时序数据间的长期依赖关系，效果最差。基于 Transformer 的 3 种方法 Autoformer、FEDformer 和 Crossformer 由于采用了多头注意力机制，可有效提取时序数据的重要信息，效果优于 LSTM。ATFNet 将时序数据分别映射到时域和频域空间，然后提取和融合二者的特征信息，在 Traffic 数据集上，时序预测长度为 96 时，α_{MSE}最优。本文方法首先将长期时序数据分解为趋势性和周期性数据，然后采用 xLSTM 提取时间点之间的时序相关性及时间段之间的长期依赖关系，有效提高了能耗预测的效果。

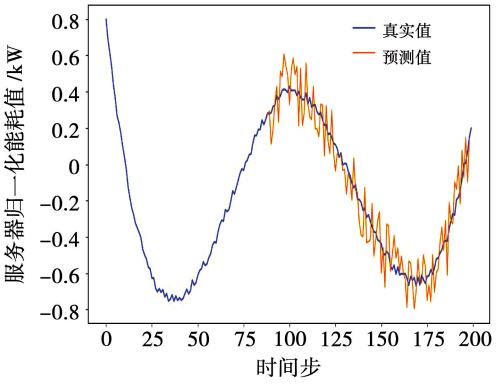
图 4 为本文方法和对比方法在服务器能耗数据集上的预测可视化图，模型预测长度为 96。由图可见，本文方法预测值和实际值的拟合程度最好，ATFNet 方法次之。



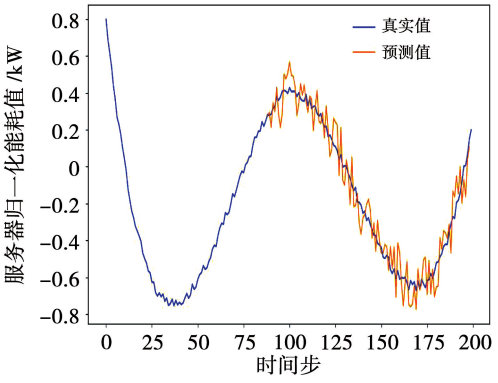
(a) LSTM



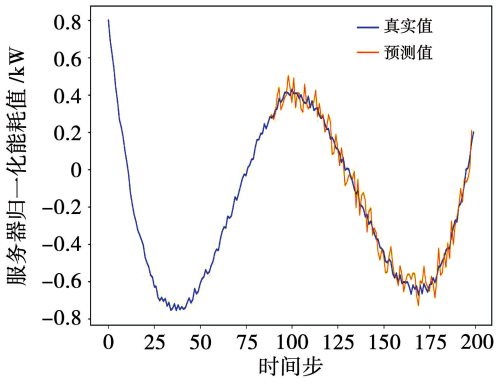
(b) Autoformer



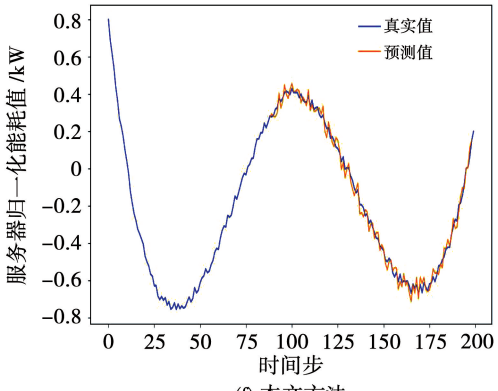
(c) FEDformer



(d) Crossformer



(e) ATFNet



(f) 本文方法

图 4　可视化结果

Fig. 4　Visualization results

3.3　消融实验

为了证明本文方法的有效性,在预测长度为 96 时进行了消融实验,实验结果如表 3 所示。

表 3　消融实验结果

Tab. 3　Ablation experiment results

模块	α_{MSE}	α_{MAE}
1 块 sLSTM	0.295	0.321
1 块 mLSTM	0.291	0.267
1 块 sLSTM + 1 块 mLSTM	0.258	0.260
3 块 sLSTM + 2 块 mLSTM	0.202	0.195
本文方法	0.139	0.151

当仅采用 1 块 sLSTM 时,α_{MSE} 和 α_{MAE} 分别为 0.295 和 0.321。当仅采用 1 块 mLSTM 时,α_{MSE} 和 α_{MAE} 分别为 0.291 和 0.267。当同时使用 1 块 sLSTM和 1 块 mLSTM 时,α_{MSE} 和 α_{MAE} 分别为 0.258 和 0.260,性能均有所提升,说明二者的共同使用提升了特征提取能力。当同时使用 3 块 sLSTM 和 2 块 mLSTM 时,α_{MSE} 和 α_{MAE} 达到了 0.202 和 0.195,说明 sLSTM 和 mLSTM 的堆叠进一步加强了模型提取服务器能耗数据时间点之间的时序相关性及时间段之间的长期依赖关系的能力。本文方法为分解模块 +3 块 sLSTM + 2 块 mLSTM。当增加了分解模块后,效果达到最优,α_{MSE} 和 α_{MAE} 分别为 0.139 和 0.151。说明分解模块有效地将服务器能耗数据分解为趋势性和季节性数据,便于 xLSTM 提取服务器能耗数据的周期性特征和突变特征,提高了能耗预测的准确性。

4　结　语

本文提出了基于 xLSTM 的数据中心服务器能耗预测方法。该方法首先采用分解模块将服务器能耗数据分解为趋势性和季节性数据,有效提取了能耗数据的长期趋势和周期模式。然后通过 sLSTM 和 mLSTM 的堆叠捕捉时间点和时间段之间的相关性和依赖关系,提高了能耗预测的准确率。与现有方法相比,本文方法在自建数据集和 4 个公开数据集上的评价指标均取得了较优效果。准确预测数据中心能耗可提升电力数据资产的管理效率和经济价值,优化电力资源的利用和分配,还可通过智能调度减少浪费,推动绿色能源应用,进一步增强电力数据资产的长期可持续性。

参 考 文 献:

[1] ZHANG L, LI K, LI C, et al. Bi-objective Workflow Scheduling of the Energy Consumption and Reliability in Heterogeneous Computing Systems [J]. Information Sciences, 2017, 379: 241.

[2] ZHOU Z, SHOJAFAR M, ALAZAB M, et al. AFED-EF: An Energy-efficient VM Allocation Algorithm for IoT Applications in a Cloud Data Center [J]. IEEE Transactions on Green Communications and Networking, 2021, 5(2): 658.

[3] ISMAIL L, MATERWALA H. Computing Server Power Modeling in a Data Center: Survey, Taxonomy, and Performance Evaluation [J]. ACM Computing Surveys, 2021, 53(3): 1.

[4] SHEN Z, ZHANG X, XIA B, et al. Multi-Granularity Power Prediction for Data Center Operations via Long Short-term Memory Network [C]//2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking, 2019: 194.

[5] LIU M, ZENG A, CHEN M, et al. Scinet: Time Series

Modeling and Forecasting with Sample Convolution and Interaction [J]. Advances in Neural Information Processing Systems, 2022, 35: 5816.

[6] PETNEHAZI G. Recurrent Neural Networks for Time Series Forecasting [J]. arXiv preprint arXiv: 1901.00069, 2019.

[7] FAN J, ZHANG K, HUANG Y, et al. Parallel Spatio-temporal Attention-based TCN for Multivariate Time Series Prediction [J]. Neural Computing and Applications, 2023, 35(18): 13109.

[8] YI K, ZHANG Q, FAN W, et al. Frequency-domain MLPs are More Effective Learners in Time Series Forecasting [J]. Advances in Neural Information Processing Systems, 2023, 36: 76656.

[9] BANDAR K, HYNDMAN R J, BERGMEIR C. MSTL: A Seasonal-trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns[J]. International Journal of Operational Research, 2025, 52(1): 79.

[10] 程华, 陈左宁, 孙凝晖, 等. 一种基于细粒度性能计数器建立系统实时功耗模型的方法[J]. 计算机学报, 2014, 37(3): 687.
CHENG Hua, CHEN Zuoning, SUN Ninghui, et al. Build Real-time Power Consumption Model of a Computer System Based on Fine-grained Performance Counters[J]. Chinese Journal of Computers, 2014, 37(3): 687.

[11] ZHANG X, LU J J, QIN X, et al. A High-level Energy Consumption Model for Heterogeneous Data Centers[J]. Simulation Modelling Practice and Theory, 2013, 39: 41.

[12] LUO L, WU W J, ZHANG F. Energy Modeling Based on Cloud Data Center[J]. Journal of Software, 2014, 25(7): 1371.

[13] ZHOU Z, SHOJAFAR M, ABAWAJY J, et al. ECMS: An Edge Intelligent Energy Efficient Model in Mobile Edge Computing [J]. IEEE Transactions on Green Communications and Networking, 2022, 6(1): 238.

[14] LI Y, HU H, WEN Y, et al. Learning-based Power Prediction for Data Centre Operations via Deep Neural Networks [C]//Proceedings of the 5th International Workshop on Energy Efficient Data Centres, 2016: 1.

[15] QIN Y, SONG D, CHEN H, et al. A Dual-stage Attention-based Recurrent Neural Network for Time Series Prediction[J]. arXiv preprint arXiv:1704.02971, 2017.

[16] SHIH S Y, SUN F K, LEE H Y. Temporal Pattern Attention for Multivariate Time Series Forecasting[J]. Machine Learning, 2019, 108(8): 1421.

[17] FRANCESCHI J Y, DIEULEVEUT A, JAGGI M. Unsupervised Scalable Representation Learning for Multivariate Time Series [J]. Advances in Neural Information Processing Systems, 2019, 32:1.

[18] ZHANG T P, ZHANG Y Z, CAO W, et al. Less Is More: Fast Multivariate Time Series Forecasting with Light Sampling-oriented MLP Structures [J]. arXiv preprint arXiv:2207.01186, 2022.

[19] CHALLU C, OLIVARES K G, ORESHKIN B N, et al. N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 6989.

[20] ZHOU T, MA Z, WEN Q, et al. Fedformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting[C]//International Conference on Machine Learning, 2022: 27268.

[21] ZHANG Y H, YAN J C. Crossformer: Transformer Utilizing Cross-dimension Dependency for Multivariate Time Series Forecasting [C]//The Eleventh International Conference on Learning Representations, 2023:1.

[22] LIU Y, LONG M, WANG J, et al. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting [C]//Advances in Neural Information Processing Systems , 2022: 9881.

[23] BECK M, PÖPPEL K, SPANRING M, et al. xLSTM: Extended Long Short-Term Memory [J]. Advances in Neural Information Processing Systems, 2024, 37: 107547.

[24] ASADI R, REGAN A C. A Spatio-temporal Decomposition Based Deep Neural Network for Time Series Forecasting[J]. Applied Soft Computing, 2020, 87: 105963.

[25] SEN R, YU H F, DHILLON I S. Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting[C]//Advances in Neural Information Processing Systems, 2019: 1.

[26] SHERSTINSKY A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-term Memory (LSTM) Network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[27] WU H, XU J, 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.

[28] YE H, CHEN J, GONG S, et al. ATFNet: Adaptive Time-Frequency Ensembled Network for Long-term Time Series Forecasting [J]. arXiv preprint arXiv: 2404.05192, 2024.

(编辑:温泽宇)