文章编号:1671-4598(2026)03-0009-09 DOI:10.16526/j.cnki.11-4762/tp.2026.03.002 中图分类号:TP183 文献标识码:A

融合分解算法和注意力机制的云负载预测模型

沈煦悦，刘 景，李风彪
(河海大学 信息科学与工程学院，江苏 常州 213000)

摘要：为了提高云平台资源分配的均衡性，针对云资源负载数据的非线性、高噪声以及动态性特点，提出一种融合CEEMDAN分解算法及注意力机制的Transformer-BiLSTM负载预测模型；该模型使用CEEMDAN分解算法对负载序列数据进行分解，得到不同频率的分量，降低数据复杂度；通过Transformer编码层构成的编码器对各分量进行编码，获取数据的全局信息，并把得到的编码输出通过注意力模块进行权重的自适应分配；采用BiLSTM构成的解码器解码得到预测结果；实验结果表明，相较于主流模型，所提出的模型在不同预测步长的误差均有降低，验证了预测方法的有效性。

关键词：云计算；负载预测；CEEMDAN；Transformer；注意力机制；BiLSTM

Cloud Load Forecasting Model Integrating Decomposition Algorithm and Attention Mechanism

SHEN Xuyue，LIU Jing，LI Fengbiao
(College of Information Science and Engineering，Hohai University，Changzhou 213000，China)

Abstract：To improve the balance of cloud platform resource allocation and address the nonlinear，high noise and dynamic characteristics of cloud resource load data，a load forecasting model integrating Transformer and bi-directional long short-term memory network (Transformer-BiLSTM) is proposed，which integrates complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) decomposition algorithm and attention mechanism. In this model，the CEEMDAN decomposition algorithm is used to decompose the load sequence data to obtain the components of different frequencies，thus reducing data complexity. An encoder with the Transformer coding layer is used to encode each component，obtaining the global information of the data，and the obtained encoded output conducts adaptive weight allocation through the attention module. The decoder with the BiLSTMP is used to decode and obtain prediction results. Experimental results show that，compared with the mainstream model，the proposed model reduces the error of different prediction steps，which verifies the effectiveness of the prediction method.

Keywords：cloud computing；CEEMDAN；Transformer；attention mechanism；BiLSTM

0 引言

云计算通过互联网提供可扩展、按需分配的计算资源和服务[1]，它大大提高了企业的计算能力和灵活性，同时显著降低了IT成本。但是，随着云计算应用的不断扩大，资源管理面临的挑战也日益突出，特别是资源分配不均的问题[2]。由于缺乏准确的预测机制，云服务提供商通常会预留过多的资源以应对突发需求，导致资源浪费[3]；而在高峰期或突发事件中，资源不足则导致服务性能下降[4]，影响用户体验和企业运营。因此，提高对云计算资源使用情况的预测准确性非常重要。

云资源负载预测本质上是一种时间序列预测问题，其早期的方法主要通过对历史数据的统计分析来预测资源需求趋势，如自回归、差分整合移动平均自回归模型(ARIMA，auto regressive integrated moving average model)[5]和指数平滑[6]等。但是，传统的统计方法往往局限于线性模式捕捉，无法获取时间序列负载数据的非线性特征，在负载预测中的效果不佳。

收稿日期:2025-02-19； **修回日期**:2025-04-08。
作者简介:沈煦悦(2000-)，男，硕士研究生。
通讯作者:刘 景(1973-)，男，博士，副教授。
引用格式:沈煦悦，刘 景，李风彪，等.融合分解算法和注意力机制的云负载预测模型[J].计算机测量与控制，2026，34(3)：9-17.

在深度学习领域，针对时间序列预测和云计算资源预测问题，研究人员构建了神经网络结构，以学习数据的多层次特征表示。文献［7］使用了长短期记忆网络进行主机负载预测，捕捉了时间序列数据中的长期依赖关系，提高了预测精度。然而，LSTM 本身序列依赖的结构使其很难具备高效的并行计算的能力。文献［8］比较多个单一模型的时序预测性能，证明双向的长短期记忆网络（BiLSTM，bi-directional long short-term memory networks）比 LSTM 有更好的预测性能。文献［9］将 Transformer 模型应用于时间序列预测和下一帧预测任务，通过调整超参数、数据预处理等方法来提高模型上下文感知能力和降低空间复杂度。Transformer 模型在时序预测中不需要考虑时间和距离，在处理远程信息和远距离依赖方面更有潜力。

随着研究的深入，一些学者们发现相比于单一的神经网络模型，融合多种模型的方法在捕捉云资源序列及时间序列数据的非线性特性方面表现更佳。文献［10］提出了一种基于 BiLS-TM 的混合循环神经网络预测模型，对云虚拟机 CPU 工作负载进行预测，增强了模型的非线性数据特征的理解。文献［11］提出了一种深度融合 CNN 和 LSTM 特征提取能力的资源负载预测模型，实现了多角度时间序列特征的精准提取。文献［12］提出了一种结合 CNN 的特征提取和 LSTM 的时间序列分析的组合模型来预测股票价格，实验结果表明该模型在时间序列预测方面优于单独使用长短期记忆网络（LSTM，long short-term memory）的模型。

此外，面对负载序列数据的非平稳性和局部峰值特性，时间序列预测研究愈发重视在有效提取全局特征的基础上，加强对局部特征的捕捉与利用。文献［13］提出的 EMD -LSTM 模型通过对原始序列进行 EMD 滑动窗口分解并进行去噪，再对所有序列重构输入 LSTM。文献［14］提出的模型通过 CEEMDAN 分解算法控制噪声，让预测模型更容易捕捉非线性特征，结合 Transformer 提高了电力负荷预测精度。文献［15］在标准的 CNN 网络上增加注意力分支以提取重要的细粒度特征，再使用 LSTM 提取粗粒度特征，有效的区分时序特征的重要程度。文献［16］提出的频率增强通道注意力机制通过自适应调整注意力权重来捕捉时序数据通道之间的相互依赖性，增强了模型在频率域的特征提取能力，提高了时间序列预测的准确性。这些工作通过引入新的数据处理方法和模型架构，增强了时间序列预测中模型对局部特征的捕捉能力。

现有研究在时间序列预测方面取得了许多进展，但云资源预测仍存在以下主要问题：一是复杂特征数据处理不足，云资源原始数据的非线性、高噪声使得提取和利用这些特征进行预测时较为困难；二是局部特征的提取和处理能力有限，对于云资源数据的局部特征的精确提取以及去除极端峰值的影响方面仍有不足。本文提出了一种融合分解算法和注意力机制的云负载预测模型。首先鉴于云资源数据的不确定性和非线性，采用完整集合经验模态分解自适应噪声（CEEMDAN，complete ensemble empirical mode decomposition with adaptive noise）分解算法对云资源时间序列进行分解；然后，利用 Transformer 模型的编码器和 BiLS-TM 构成编解码的组合模型，获取序列的全局信息和双向时序依赖；随后，基于 SE 注意力机制和软阈值化提出一种多尺度去噪注意力机制，以强化模型对时序数据的重要局部特征的捕捉能力；最后，通过消融实验及与其他模型的对比实验，验证了模型的有效性。

1 基于分解算法和注意力机制的云负载预测模型

1.1 总体框架

云计算资源负载的任务目标是使用历史负载序列数据预测未来一段时间的负载序列数据。其中是历史序列长度，代表预测序列长度，是特征维度，表示不同的负载变量。

为完成云负载预测任务，且鉴于云计算资源负载序列数据的非线性、高噪声和非平稳特性，本文提出了一种基于分解算法和注意力机制的云资源预测模型-CEEMDAN-Transformer-BiLSTM-Attention。模型整体框架如图 1 所示。该模型主要由分解模块、编解码模型和注意力机制组成。

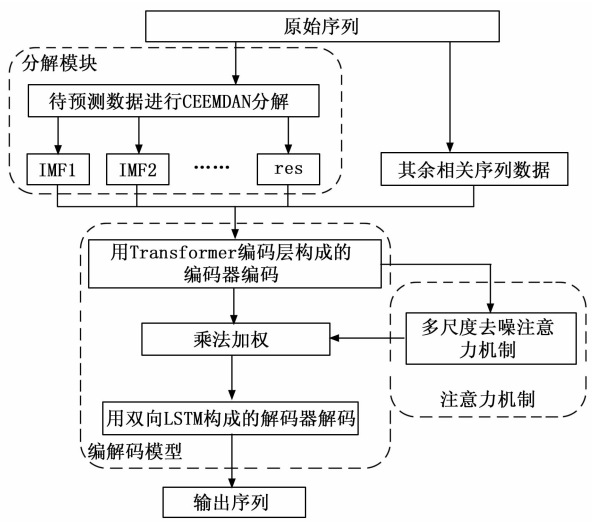

图 1 模型整体框架

首先，利用 CEEMDAN 算法将待预测的云资源负载序列数据分解为多个不同频率的子序列，以降低数据复杂度并便于后续处理，这些子序列与相关资源数据整合后，能更全面地反映负载特征；其次，在编解码模型中，通过 Transformer 编码层构成的编码器对整合后的

数据进行编码，以获取全局信息并提取深层次特征；然后，引入多尺度去噪注意力机制，对编码后的输出进行权重分配，从而获得不同频率分量的重要性，即局部特征的提取，并实现去除冗余特征效果；最终，利用 BiLSTM 解码器对经过注意力机制调整后的数据进行解码，以捕捉双向时序依赖关系，得到准确预测输出序列。

1.2　CEEMDAN 分解算法

云数据中心负载数据的特性，尤其是其高度的动态性和不确定性，加之突发事件所引发的请求峰值，极大地增加了原始序列数据特征学习的复杂度，进而对资源预测的准确性构成了挑战[17]。为了精准捕捉这些复杂变化并优化预测能力，对预测数据的预处理显得尤为重要。

在处理这类非平稳时间序列数据时，经验模态分解（EMD，empirical mode decomposition）技术以其自适应分解的优势脱颖而出，能够将原始数据拆解为多个具有不同特征尺度的序列，从而有助于提升预测方法的精度。然而，EMD 在分解过程中潜在的模态混叠问题，即单个本征模态分量（IMF，intrinsic mode function）中混入多个不同模态的现象，成为制约其应用效果的一大瓶颈。

为了有效克服这一局限，本文选择了完整集合经验模态分解自适应噪声算法[18]作为解决方案。CEEMDAN 算法继承了 EMD 的自适应分解能力，并且为了降低噪声对数据分解过程的影响，CEEMDAN 算法在分解过程中引入了惩罚系数 σ，该系数用于控制分解出的模态数量及其稳定性。在 CEEMDAN 算法的分解中，M 次对当前残余分量 $r(t)$ 与噪声 $\omega(t)$ 重新组合和分解处理。对于原始云资源数据 $X(t)$，CEEMDAN 算法通过以下步骤将其分解为 N 个本征模态分量和一个最终残余分量 R，如式（1）所示：

$$X(t)=\sum_{j=1}^{N}\frac{1}{M}\sum_{i=1}^{M}E_j\{r_j(t)+\sigma_j E_j[\omega_i(t)]\}+R(t) \tag{1}$$

其中：$r_1(t)$ 为 EMD 分解 $X(t)+\sigma_0\omega_i(t)$ 获得的第一个 IMF，即 $r_1(t)=X(t)-\frac{1}{M}\sum_{i=1}^{M}E_1[S(t)+\sigma_0\omega_i(t)]$；运算符 E_j（）表示由 EMD 分解序列获得的第 j 个 IMF。

1.3　编解码模型

面对原始负载数据分解得到的多个模态分量，云资源预测模型需有效提取跨尺度的全局依赖和局部时序特征，为此，本文设计了一个编解码组合模型。如图 1 所示，该模型包括两个部分：一是基于 Transformer 编码层所构建的编码器部分；二是采用多层 BiLSTM 构建的解码器部分。Transformer 编码器捕获云资源数据的全局信息，BiLSTM 解码器利用其强大的时序建模能力，全面获取正向与反向的时序依赖信息，然后通过编解码方式融合两个过程的特征，最终输出预测结果，提升了模型在处理云资源序列数据复杂特征时的性能与准确性。编解码模型如图 2 所示。

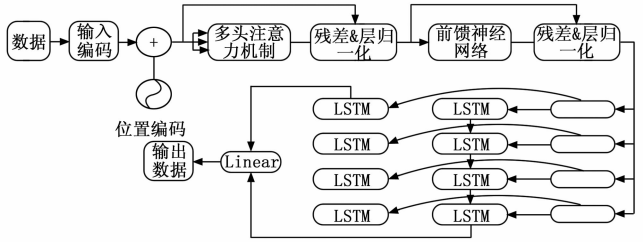

图 2　Transformer-BiLSTM 模型图

在编解码模型中，首先对负载序列数据输入编码和位置编码等预处理操作，将位置相关信息添加到序列中，接着使用多头注意力、层规范化和残差连接对序列进行进一步的处理，生成高维的特征表示，这些特征表示能够捕捉序列中的全局依赖关系和局部特征。将这些特征向量输入到 BiLSTM 中，通过捕捉过去和未来的上下文信息，进一步提取时序特征。最后，BiLSTM 的输出被传递至线性层，生成最终的预测结果。整个模型结合 Transformer 和 BiLSTM，在负载序列预测中优化了特征提取过程，并且能够更精准的识别时序模式，提升了模型的预测精度。

1.3.1　Transformer 编码器

Transformer 模型[19]专为处理序列到序列任务而设计，以解决长期依赖问题。该模型采用自注意力机制进行特征提取，具备快速的运算速度，有效克服了 RNN 在处理长期依赖时的局限。

本文选择 Transformer 中的编码层构建编解码模型中的编码器，Transformer 编码层为模型提供经注意力加权的数据，获取全局信息，增强了模型在长期依赖预测任务中的特征捕捉能力[20]。序列数据的注意力权重由编码层的点积运算生成，并由多个并行的 Transformer 编码层进行处理，实现序列预测。编码层主要由多头注意力机制和前馈层组成，每个模块后进行层规范化操作，以此缓解梯度爆炸问题，增强网络稳定性。Transformer 编码层的结构如图 3 所示。

1）自注意力机制：

Transformer 模型提出了自注意力机制，能够注意输入序列不同位置的信息以此计算该序列的表示能力。自注意力机制通过计算查询（Q）、键（K）和值（V）之间的注意力权重来将序列的不同位置联系起来，评估各个位置的相对重要性。计算公式为：

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)*V \tag{2}$$

式中，d_K 表示键向量的维度。

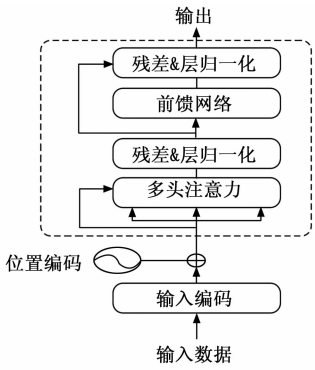

图 3 Transformer 编码器

2）多头注意力机制：

在 Transformer 模型中，多头注意力机制由自注意力构成，其计算方式与自注意力相同，每个头在不同的表示子空间中捕获信息。计算公式为：

$$MultiHead(Q,K,V) = Concat(head_1,\cdots,head_h)W^0 \tag{3}$$

$$head_i = Attention(QW_i^Q,KW_i^K,VW_i^V) \tag{4}$$

式中，W_i^Q、W_i^K、W_i^V 分别表示第 i 个不同的特征空间上的注意力机制运算参数；W^0 用来将合并后的注意力矩阵压缩变换。

3）位置编码：

由于 Transformer 模型本身不包含序列顺序信息，位置编码被用来向输入序列的每个位置注入额外的信息，从而使模型能够区分不同位置的标记：

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000\dfrac{2i}{d_{model}}}\right) \tag{5}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000\dfrac{2i}{d_{model}}}\right) \tag{6}$$

其中：pos 表示序列位置，i 表示位置序号，d_{model} 表示嵌入空间维度的大小。

4）前馈网络：

除了自注意力层之外，Transformer 模型的前馈网络含全连接层，其输出经过残差归一化层之后进入解码器。前馈网络层的计算过程表示为：

$$FFN(x) = max(0,x\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{7}$$

式中，x 表示前馈网络层的输入；$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 表示两个参数矩阵；$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 表示偏置向量。

1.3.2 BiLSTM 解码器

梯度消失和短期记忆问题是传统循环神经网络（RNN，recurrent neural network）中最普遍的问题，这些问题可以通过 LSTM 这一种基于 RNN 改进的模型来解决。LSTM 在学习过程存在长期记忆和短期记忆，可以学习长期的依赖关系，适用于时间序列预测问题。LSTM 中的神经单元具有 3 个门控单元，即输入门、遗忘门和输出门。所有的门都有一个 σ 函数，其作为一个过滤器，决定保留和遗忘哪些信息，σ 是 Sigmoid 激活函数。x_t 是当前时间步 t 的输入数据，h_{t-1} 是前一时间步的隐藏状态，它在 LSTM 中充当短期记忆，C_{t-1} 是上一时刻的细胞状态。将 x_t、h_{t-1} 和 C_{t-1} 共同输入，经过门控单元的处理，得到 LSTM 神经单元的输出。LSTM 网络结构如图 4 所示。

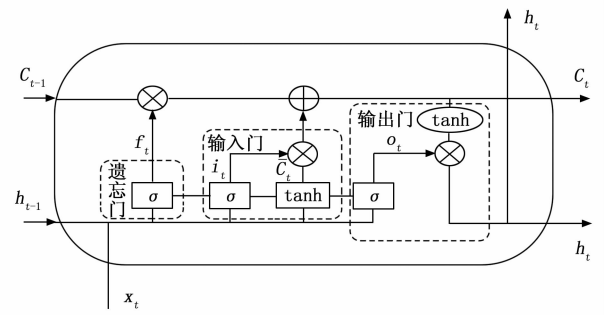

图 4 LSTM 网络结构

遗忘门中数据的计算公式如式（8）所示：

$$f_t = \sigma(W_f[h_{t-1},x_t] + b_f) \tag{8}$$

输入门中数据的计算公式如式（9）、（10）所示：

$$i_t = \sigma(W_i[h_{t-1},x_t] + b_i) \tag{9}$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1},x_t] + b_C) \tag{10}$$

输出门中数据的计算公式如式（11）所示：

$$o_t = \sigma(W_o[h_{t-1},x_t] + b_o) \tag{11}$$

每个时间步长的 LSTM 单元输出为：

$$C_t = f_tC_{t-1} + i_t\tilde{C}_t \tag{12}$$

$$h_t = o_t\tanh(C_t) \tag{13}$$

BiLSTM 是一种结合前向和后向 LSTM 的双向 LSTM 模型，它通过两次不同方向的 LSTM 提取过去和未来的数据信息，这使得网络可获取的信息大量增加。因此，BiLSTM 能够更好的理解上下文，且能够有效捕的获序列的双向的依赖关系和潜在特征[21]。因此，本文选择 BiLSTM 构建编解码模型的解码器，网络结构如图 5 所示。

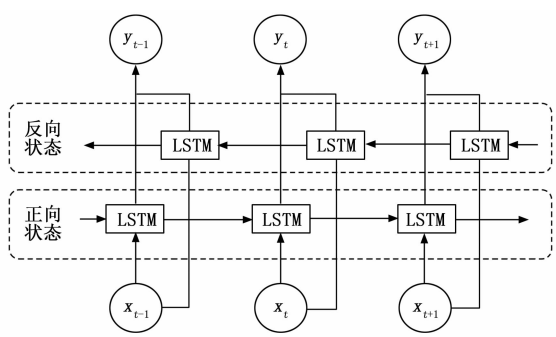

图 5 BiLSTM 结构

1.4 多尺度去噪注意力机制

尽管 CEEMDAN 分解已分解得到了不同频率的分

量，但不同频率分量的局部突变（如突发负载峰值）仍可能引入预测偏差。为此，本文针对云资源负载数据局部特征提取受限及极端峰值噪声问题，提出一种多尺度去噪注意力机制。该机制通过采用 3 个不同尺度的一维池化压缩操作替代 SE 注意力机制中的压缩步骤，有效捕获了云资源数据中多尺度的局部信息。在权重计算与加权过程中，引入软阈值化操作，去除冗余特征的影响。增强了模型在提取局部特征及降噪方面的能力。

1.4.1　SE 注意力机制

SE（Squeeze and Excitation）注意力机制[22]利用全局信息对特征映射的通道之间的相互依赖关系进行建模，并重新校准特征映射以提高表示能力。它由压缩和激励两个步骤组成，如图 6 所示。对于时间序列信号 $X \in R^{C\times L}$，C 是通道数，L 是时间序列的长度，这种类型的张量 X 可以在时间序列模型中的任何位置。

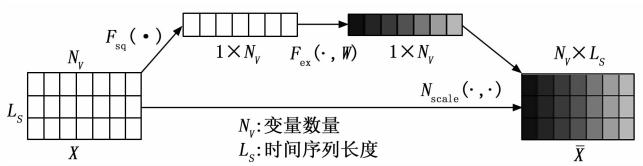

图 6　SE 注意力

对于时间信号，压缩步骤 F_{sq} 在时间维度上应用全局平均池化（GAP）进行降维，提取全局信息。其中，全局特征 $Z_c \in R^C$ 是通过缩减 x 的时间维度 L_S 来生成的，这样就可以通过以下方式计算 Z 的第 c 项：

$$Z_c = GAP(x_c) = \frac{1}{L_S}\sum_{i=1}^{L_S} x_c(i) \tag{14}$$

其中：c，L_S 分别表示通道和时间维度。标量 Z_c 是 Z 的第 c 个元素，然后激励步骤 F_{ex} 旨在通过使用两个全连接层 W_1 和 W_2 来激活通道相关性：

$$att = \sigma[W_2\delta(W_1 Z)] \tag{15}$$

其中：$att \in R^C$ 是学习到的注意力向量，它与原始特征映射相乘以重新缩放每个通道，σ 和 δ 分别指 ReLU 和 Sigmoid 激活函数。

1.4.2　软阈值化

软阈值化[23]是一种信号去噪技术，通过收缩输入信号中的小幅度值至零并保留较大幅度值，实现信号去噪和稀疏化。软阈值函数及其导数如公式（16），（17）所示：

$$f(x) = \begin{cases} x-\tau, & x > \tau \\ 0, & -\tau \leqslant x \leqslant \tau \\ x+\tau, & x < -\tau \end{cases} \tag{16}$$

$$\frac{\partial f(x)}{\partial x} = \begin{cases} 1, & x > \tau \\ 0, & -\tau \leqslant x \leqslant \tau \\ 1, & x < -\tau \end{cases} \tag{17}$$

式中，x 代表输入值；τ 为阈值；$f(x)$ 为软阈值函数输出。

软阈值化通过将绝对值低于阈值的特征值设为零，以消除冗余，同时其求导后的梯度只有 0 或 1，有效防止梯度爆炸和消失。为降低人为设定参数的不确定性，需结合注意力机制以自适应地确定阈值。

1.4.3　改进注意力机制

本文提出的多尺度去噪注意力机制，基于 SE 注意力机制并融合多尺度池化、软阈值化技术，来强化时间序列特征的分析与去噪能力。其整体结构如图 7。该机制首先利用 3 个不同尺度的池化层对输入序列执行全局池化操作，以此捕捉跨越不同时间尺度的关键特征。其次，通过计算这些池化结果拼接后的均值，并将其作为权重因子与原始拼接值相乘，实现对通道特征的初步加权调整，从而增强或减弱特定通道的影响。随后，该机制引入了一个由两层全连接网络和激活函数构成的小型子网络，该子网络能够学习到更为精确的注意力权重，这些权重经由 Sigmoid 激活函数处理后，被用于对原始序列数据进行加权，以进一步优化模型对时间序列特征的响应能力。最后，为了消除冗余特征并有效提取关键特征，采用了软阈值化技术进行去噪处理。此注意力机制结合分解算法应用在模型中，加强了对云资源数据关键局部特征的提取。

在多尺度去噪注意力机制中，采用 3 个不同尺度的一维池化层替代 SE 注意力机制中的单一全局平均池化，旨在捕捉云资源负载数据在不同时间跨度上的局部特征和动态变化。云资源负载数据通常表现出多时间尺度的特性，单一尺度的池化操作仅能提取全局信息，忽略了这些局部突变和多尺度依赖关系，而多尺度池化则能够更全面地建模这些特性。其中，较小的输出尺寸代表较大的压缩比例，如一维的自适应平均池化，可以捕

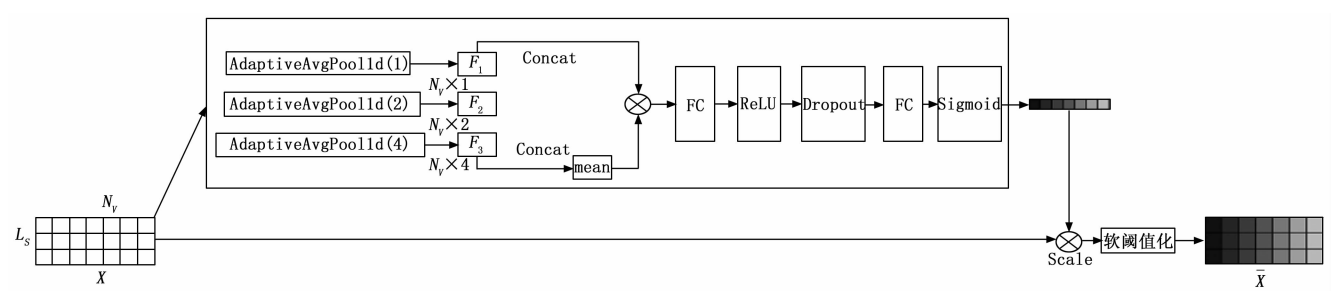

图 7　多尺度去噪注意力

捉序列的整体趋势；中等输出尺寸则反映中间层次的趋势变化；较大的输出则用于捕捉局部特征。

其中 AdaptiveAvgPool1d 表示一维的自适应平均池化，函数参数表示池化后的输出尺寸；F_1、F_2 和 F_3 表示池化后的数据；Concat 表示数据进行拼接；mean 表示求均值；Scale 表示加权相乘；FC 表示全连接层；ReLU 为激活函数；Dropout 随机失活用来丢弃网络部分权重；Sigmoid 表示激活函数。

2 实验结果与分析

2.1 实验环境

本文实验环境设置如表 1 所示。

表 1 实验环境设置

实验环境	设置
运行环境	Python 3.9
深度学习框架	PyTorch 2.1
操作系统	Windows 10
内存/GB	32.0
CPU	Intel（R）Xeon（R）Bronze 3206R CPU@1.90 GHz
GPU	NVIDIA Quadro P2200

2.2 评价指标

本文采用均方误差（MSE，mean squared error）、平均绝对误差（MAE，mean absolute error）和均方根误差（RMSE，root mean squared error）作为评价指标，以客观衡量不同 CPU 负载预测方法的性能。各个指标的计算公式如下：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{18}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{19}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{20}$$

其中：$\hat{y}_i$ 为对应的预测值。

2.3 数据预处理

2.3.1 实验数据集

为验证本文所提出的模型在云资源预测方面的性能及适用性，本文选取了 2 个开源数据集，如下：

1）Serverless-trace-dataset[24]（以下称 ST）实验数据集。该数据集来自于华为云数据湖探索，该数据集涵盖了多种真实应用场景下的工作负载数据，详细记录了诸如采样时间点、CPU 使用率、内存占用率、出入口流量等一系列云资源的关键负载指标。

2）谷歌 Google Cloud Trace 2019（以下称 GC19）实验数据集。该数据集由 8 个不同集群单元的 29 天数据组成，每个集群单元约有 10 000 台机器，分布在世界各地的不同地理区域。

鉴于 CPU 使用率在反映云计算集群工作负载状态方面具有较高的代表性，本研究在实验过程中特选取了 CPU 使用率作为主要预测目标。对这两个数据集，选择其中 8 000 条左右的数据作为实验数据，将其分为训练集、测试集、验证集，比例为 8∶1∶1。

2.3.2 数据处理

对于缺失值的处理，采用均值填充法。将原始数据中的待预测指标使用 CEEMDAN 分解算法进行分解，得到多个本征模态分量，将其拼接到原数据集。为得到稳定的模态，CEE-MDAN 分解算法的集成次数 trials 设置为 100，信噪比 epsilon 设置为 0.005。分解结果如图 8 所示。

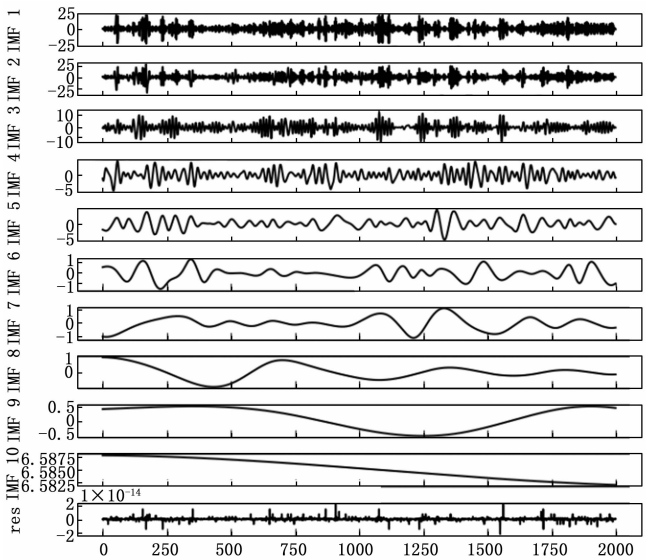

图 8 分解结果

随后进行归一化处理，再使用滑动窗口制作数据集，其中按照上述 8∶1∶1 比例划分训练集、测试集和验证集。

2.4 实验结果

本节进行了对比实验和消融实验，并对实验结果分别进行分析。CEEMDAN-Transformer-BiLSTM-Attention 的超参数设置如表 2 所示。在 ST 和 GC19 两个数据集上，分别选取 24、36 和 48 个时间步作为预测长度，以此探究本文模型在不同预测步长的性能表现，且契合长期预测场景的需求。

表 2 超参数设置

超参数	设置
输入步长	96
预测步长	24、36、48
编码器层数/层	2
解码器层数/层	2
批尺寸	64
初始学习率	0.000 3
丢弃率	0.5
损失函数	*MSE*

2.4.1　对比实验结果分析

表 3 提供了 CNN、LSTM、等预测模型在不同预测步长下的误差指标对比，其中最优指标值用加粗字体表示，下同。由表 3 可以看出，在预测步长为 24、36、48 时，本文提出的基于分解算法和注意力机制的云资源预测模型的各项预测性能指标均明显优于其他模型，表明该模型在云资源预测方面具有更高的精度。具体表现为：

1）相较于单一深度学习模型，本文提出的模型预测精度均有提升。与传统 Transformer 相比，本文提出的模型在预测步长为 24、36、48 时，*MAE* 在 ST 数据集中分别降低了 54.0%、52.9%、50.9%，在 GC19 数据集中分别降低了 50.3%、51.0%、56.7%。本文提出的模型与 LSTM 相比，在预测步长为 24、36、48 时，*RMSE* 在 ST 数据集中分别降低了 25.5%、22.8%、23.1%，在 GC19 数据集中分别降低了 30.3%、31.7%、36.7%。这表明本文模型通过结合 Transformer 和 BiLSTM 的编解码结构，同时获取全局信息和双向时序依赖，提升了云资源数据复杂特征的提取能力，从而实现了预测精度的提升。

2）本文提出的模型与对比模型中预测效果最好的 CNN-LSTM 相比，当预测步长分别为 24、36、48 时，ST 数据集中 *RMSE* 分别降低了 20.1%、6.6%、7.5%，在 GC19 数据集中分别降低了 27.7%、35.3%、32.8%。这表明尽管 CNN-LSTM 通过二维卷积神经网络提取了更多输入数据的特征信息，提高了预测精度，但效果有限。本文提出的模型使用 CEEMDAN 分解算法降低数据复杂度，为后续特征提取创造了有利条件，再利用注意力机制增强对局部关键特征的提取能力，解决了云资源预测中对局部特征提取不足的问题。

3）CEEMDAN-Transformer-BiLSTM-Attention 模型在不同预测步长下的 *MAE*、*MSE* 和 *RMSE* 比 CNN、LSTM、Transformer、CNN-LSTM 等预测模型在 ST 数据集上降低了 2.7%～54.0%、12.0%～77.1%、6.7%～52.1%，在 GC1-9 数据集上降低了 29.9%～63.1%、45.2%～79.7%、27.8%～55.0%。

2.4.2　消融实验结果分析

为了研究本文提出的模型中各个部分对预测性能的影响，以该模型为基础进行消融实验。在实验过程中，移除了注意力机制中的多尺度池化和软阈值化的模型命名为 CEEMDAN-Tra-nsformer-BiLSTMSE，移除了 SE 注意力机制的模型命名为 CEEMDAN-Transformer-BiLSTM，移除了分解算法和注意力机制的模型命名为 Transformer-BiLSTM。本文提出的模型与各个模型在不同条件下进行性能对比分析，实验结果如表 4 所示。

由表 4 可以看出，在两个数据集中，本文提出的基于分解算法和注意力机制的云资源预测模型 CEEMDAN-Transformer-BiLSTM-Attention 与各个模型相比预测性能均有显著提升。这表明：

1）集成 BiLSTM 到 Transformer 模型中，可以在获取全局信息的基础上捕捉时间序列数据中的双向依赖关系，提升模型的复杂特征提取能力。

2）在 Transformer-BiLSTM 中加入 CEEMDAN 分解可以从时间序列数据中提取不同频率的本征模态分量，

表 3　对比实验结果

模型	指标	ST			GC19		
		24	36	48	24	36	48
CNN	*MAE*	3.867 3	3.806 4	3.817 7	5.785 2	8.007 0	8.163 3
	MSE	23.344 8	23.155 8	22.891 2	54.605 1	98.945 8	115.567 7
	RMSE	4.831 7	4.812 0	4.784 5	7.389 5	9.947 1	10.750 2
LSTM	*MAE*	3.022 1	3.114 4	3.073 9	5.380 2	6.232 6	7.137 3
	MSE	23.069 3	23.430 8	22.944 6	48.658 1	60.696 3	81.312 7
	RMSE	4.803 0	4.840 5	4.790 0	6.975 5	7.790 8	9.017 4
Transformer	*MAE*	5.355 9	5.315 7	5.329 2	7.609 1	8.592 0	10.266 9
	MSE	55.921 9	55.872 2	56.049 1	102.507 0	115.185 4	160.693 7
	RMSE	7.478 1	7.474 8	7.486 6	10.124 6	10.732 4	12.676 5
TCN	*MAE*	5.189 4	5.168 7	5.122 8	6.353 3	6.376 4	6.947 8
	MSE	53.574 7	53.134 4	53.513 1	60.861 9	68.239 7	77.945 6
	RMSE	7.319 5	7.289 3	7.315 3	7.801 4	8.260 7	8.828 7
CNN-LSTM	*MAE*	2.962 1	2.737 8	2.691 3	5.389 4	6.536 8	6.732 1
	MSE	20.055 5	16.026 9	15.863 0	45.202 1	67.516 0	72.153 1
	RMSE	4.478 3	4.003 4	3.982 8	6.723 2	8.216 8	8.494 3
CNN-Transformer	*MAE*	3.676 1	3.527 8	3.484 3	6.254 1	6.711 4	7.263 9
	MSE	27.558 1	26.394 5	26.076 3	64.637 6	76.977 4	89.543 8
	RMSE	5.249 6	5.137 6	5.106 5	8.039 8	8.773 7	9.462 8
CEEMDAN-Transformer-BiLSTM-Attention	*MAE*	2.461 3	2.502 2	2.618 0	3.781 4	4.211 6	4.441 7
	MSE	12.792 5	13.968 3	13.566 7	23.646 0	28.287 1	32.602 6
	RMSE	3.576 7	3.737 4	3.683 3	4.862 7	5.318 6	5.709 9

表 4 消融实验结果

模型	指标	ST			GC19		
		24	36	48	24	36	48
Transformer	*MAE*	5.355 9	5.315 7	5.329 2	7.609 1	8.592 0	10.266 9
	MSE	55.921 9	55.872 2	56.049 1	102.507 0	115.185 4	160.693 7
	RMSE	7.478 1	7.474 8	7.486 6	10.124 6	10.732 4	12.676 5
Transformer-BiLSTM	*MAE*	3.315 2	3.413 8	3.169 6	5.141 1	6.040 7	6.825 1
	MSE	20.837 8	21.403 5	18.524 9	40.104 0	54.281 3	68.739 5
	RMSE	4.564 8	4.626 4	4.304 1	6.332 8	7.367 6	8.290 9
CEEMDAN-Transformer-BiLSTM	*MAE*	3.338 4	3.357 2	3.048 0	4.077 8	4.485 5	4.987 6
	MSE	23.715 1	23.126 7	20.295 1	27.015 7	34.393 0	41.269 9
	RMSE	4.869 8	4.809 0	4.505 0	5.197 7	5.864 6	6.424 2
CEEMDAN-Transformer-BiLSTM-SE	*MAE*	2.820 4	2.727 8	2.723 5	3.886 1	4.354 0	5.005 5
	MSE	18.366 9	16.739 7	16.565 9	24.522 0	30.884 5	40.875 9
	RMSE	4.285 7	4.091 4	4.070 1	4.952 0	5.557 4	6.393 4
CEEMDAN-Transformer-BiLSTM-Attention	*MAE*	2.461 3	2.502 2	2.618 0	3.781 4	4.211 6	4.441 7
	MSE	12.792 5	13.968 3	13.566 7	23.646 0	28.287 1	32.602 6
	RMSE	3.576 7	3.737 4	3.683 3	4.862 7	5.318 6	5.709 9

降低云资源数据复杂度，有利于模型的特征提取。

3）在 CEEMDAN-Transformer-BiLSTM 模型中添加 SE 注意力机制，则对不同频率的分量进行权重计算，可以有效地提高模型对不同频率特征的关注度，提升预测性能。

4）在 CEEMDAN-Transformer-BiLSTMSE 中注意力机制的基础上加上多尺度特征提取和软阈值化的去噪技术，提取多尺度特征，增强模型对局部特征的提取能力，同时降低冗余特征的影响，使模型具有更好的预测结果。

由此可见，将上述网络结构融合到模型中，并与其他模型在不同预测步长下进行对比，本文提出的模型 CEEMDAN-Transformer-BiLS TM-Attention 均取得了较好的效果，证明了模型各个组件的有效性，达到了预期的效果。

3 结束语

针对云资源预测问题，本文构建了融合分解算法和注意力机制的预测模型。使用 CEEMDAN 方法分解云资源负载数据为不同频率的子序列，降低数据复杂度并便于后续特征提取。通过 Transformer 编码器和 BiLSTM 的编解码模型获取全局信息及双向时序依赖。结合提出的多尺度去噪注意力机制，对不同频率的子序列特征进行加权与去噪，增强了模型对局部关键信息的捕捉能力。实验结果表明，本文提出的模型在云资源负载预测任务中表现出色，相较于传统方法，有效的提高了预测精度，在多个评估指标上（MAE、MSE、RMSE）均优于其他模型，充分验证了模型的有效性和稳定性。后续将进一步优化模型结构，探索更高效的特征提取与融合策略，以应对更加复杂多变的云资源负载场景，推动云计算技术的持续发展与创新。

参考文献：

[1] MASDARI M，VALIKARDAN S，SHAHI Z，et al. Towards workflow scheduling in cloud computing：a comprehensive analysis [J]. Journal of Network and Computer Applications，2016，66：64－82.

[2] MASDARI M，NABAVI S S，AHMADI V. An overview of virtual machine placement schemes in cloud computing [J]. Journal of Network and Computer Applications，2016，66：106－127.

[3] WANG B，LIU F，LIN W. Energy-efficient VM scheduling based on deep reinforcement learning [J]. Future Generation Computer Systems，2021，125：616－628.

[4] XIE Y，JIN M，ZOU Z，et al. Real-time prediction of docker container resource loa-d based on a hybrid model of ARIMA and triple exponential smoothing [J]. IEEE Transactions on Cloud Computing，2020，10 (2)：1386－1401.

[5] CALHEIROS R N，MASOUMI E，RANJAN R，et al. Workload prediction using ARIMA model and its impact on cloud applications'QoS [J]. IEEE Transactions on Cloud Computing，2014，3 (4)：449－458.

[6] HUANG J，LI C，YU J. Resource prediction based on double exponential smoothing in cloud computing [C] // 2012 2nd International Conference on Consumer Electronics，Communications and Networks (CECNet)，IEEE，2012：2056－2060.

[7] SONG B，YU Y，ZHOU Y，et al. Host load prediction with long short-term memory in cloud computing [J]. The Journal of Superco-mputing，2018，74：6554－6568.

[8] SIAMI-NAMINI S，TAVAKOLI N，NAMIN A S. A comparative analysis of forecasting financial time series using arima，lstm，and bilstm [J]. ArXiv Preprint ArXiv，2019：1911.09512.

[9] CHOLAKOV R，KOLEV T. Transformers predicting the future. Applying attention in next-frame and time series forecasting [J]. ArXiv Preprint ArXiv，2021：2108.08224.

[10] KARIM M E，MASWOOD M M S，DAS S，et al. BHyPreC：a novel Bi-LSTM based hybrid recurrent neural network model to predict the CPU workload of cloud virtual machine [J]. IEEE Access，2021，9：131476－131495.

[11] SONG J，ZHANG L，XUE G，et al. Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model [J]. Energy and Buildings，2021，243：110998.

[12] KIM T，KIM H Y. Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data [J]. PloS one，2019，14 (2)：e0212320.

[13] 姚洪刚，沐年国. EMD-LSTM 模型对金融时间序列的预测 [J]. 计算机工程与应用，2021，57 (5)：239－244.

[14] RAN P，DONG K，LIU X，et al. Short-term load forecasting based on CEEMDAN and Transformer [J]. Electric Power Systems Research，2023，214：108885.

[15] 李　梅，宁德军，郭佳程. 基于注意力机制的 CNN-LSTM 模型及其应用 [J]. 计算机工程与应用，2019，55 (13)：20－27.

[16] JIANG M，ZENG P，WANG K，et al. FECAM：Frequency enhanced channel attention mech- anism for time series forecasting [J]. Advanced Engineering Informatics，2023，58：102158.

[17] 李浩阳，贺小伟，王宾，等. 基于改进 Informer 的云计算资源负载预测 [J]. 计算机工程，2024，50 (2)：43－50.

[18] TORRES M E，COLOMINAS M A，SCHLOTTHAUER G，et al. A complete ensemble empirical mode decomposition with adaptive noise [C] //2011 IEEE international conference on acoustics，speech and signal processing (ICASSP)，IEEE，2011：4144－4147.

[19] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [J]. Advances in Neural Information Processing Systems，2017，14 (30)：374－383.

[20] 张　帆，姚德臣，姚圣卓，等. 基于 Transformer-LSTM 网络的轴承寿命预测 [J]. 振动与冲击，2024，43 (6)：320－328.

[21] MA W，YU H，ZHAO K，et al. Tibetan location name recognition based on bilstm-crf model [C] //2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)，IEEE，2019：412－416.

[22] HU J，SHEN L，SUN G. Squeeze-and-exc-itation networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：7132－7141.

[23] DONOHO D L. De-noising by soft-threshol-ding [J]. IEEE Transactions on Information Theory，2002，41 (3)：613－627.

[24] SERVERLESS-TRACE-DATASET [EB/OL]. (2021－4－13). https://www.datafountain.cn/datasets/5874.

(上接第 8 页)

[8] 胡坦能，高鸿波，崔凯歌，等. 显微 CT 涡轮叶片重建质量优化方法 [J]. 中国测试，2023，49 (12)：41－46.

[9] 杨　磊，窦　唯，龚杰峰，等. 基于制造加工因素诱导的涡轮泵叶片故障与可靠性分析 [J]. 厦门大学学报 (自然科学版)，2024，63 (2)：252－258.

[10] 石多奇，刘长奇，程　震，等. SiC/SiC 复合材料涡轮叶片结构设计及静强度评价 [J]. 航空动力学报，2023，38 (1)：1－12.

[11] 王　楠，吕　东，王晓放，等. 涡轮叶片蜂巢式冷却结构减阻能力数值研究 [J]. 热科学与技术，2023，22 (2)：165－173.

[12] 石多奇，王振宇，刘长奇，等. 典型涡扇发动机陶瓷基复合材料涡轮叶片概念设计 [J]. 航空动力学报，2023，38 (2)：431－444.

[13] 王乾坤，王　威，迟庆新，等. 定向凝固合金涡轮叶片服役后组织研究 [J]. 航空材料学报，2023，43 (2)：9－16.

[14] 聂卫健，邓旺群，杨　刚，等. 航空发动机动力涡轮叶片断裂试验 [J]. 科学技术与工程，2024，24 (19)：8332－8338.

[15] 高玲玲，李瑞勤. 航空发动机涡轮叶片热障涂层的现状与发展 [J]. 兵器材料科学与工程，2024，47 (4)：121－129.

[16] 范博超，张小栋，熊逸伟，等. 涡轮叶片裂纹方位角的三维叶尖间隙动态响应特性研究 [J]. 西安交通大学学报，2024，58 (7)：170－178.

[17] 秦仁耀，曲致奇，陈冰清，等. 航空发动机单晶高温合金涡轮转子叶片增材修复技术研究进展 [J]. 材料工程，2024，52 (12)：1－14.

[18] 付星豪，黄玉娟，马志乐，等. 尾缘切除对低压涡轮叶栅气动性能的影响 [J]. 大连海事大学学报，2024，50 (1)：158－166.

[19] 焦　健，孙世杰，焦春荣，等. SiCf/SiC 复合材料涡轮导向叶片研究进展 [J]. 复合材料学报，2023，40 (8)：4342－4354.

[20] 吕玥萱，罗稼昊，饶　宇. 涡轮叶片不同肋倾角交错肋冷却流动与传热研究 [J]. 工程热物理学报，2023，44 (5)：1341－1347.