DOI:10.13878/j.cnki.jnuist.20230108001

丁建立[1]　龚子恒[1]

基于组合注意力模型 EAAT 的云 KPI 数据预测方法

摘要

为了准确分析云计算集群日常监控中 KPI(Key Performance Indicator)数据的动态和变化趋势,并预测后续发展,达到提高云计算集群高可用性的目标,本文提出三分频的基于组合注意力模型的 EWT-ARIMA-Auto-TPA (EAAT)云 KPI 数据预测方法.首先基于经验小波变换(Empirical Wavelet Transform,EWT)得到云 KPI 数据低中高频的内在模态变量(Intrinsic Mode Functions,IMFs)降低数据预测的复杂程度.其次,根据分解得到的低中高频 IMFs 信息特征,分别运用 ARIMA、Autoformer、TPA-BiLSTM 模型对每类 IMFs 进行预测.最后,将分类预测后结果经过逆变换 IEWT 加以合并得出预测结果.本文预测方法在谷歌和亚马逊的 4 个数据集上得到了验证,无论数据是否具有周期性或者稳定性,本文预测方法都有较好的结果,综合效果比对比模型有较大提升.

关键词

云 KPI 数据;时序预测方法;经验小波变换;组合注意力模型;双向长短时记忆网络

中图分类号 TP391

文献标志码 A

收稿日期 2023-01-08

资助项目 国家自然科学基金民航联合基金项目(U2033205,U1833114)

作者简介

丁建立,男,博士,教授,主要研究方向为智能仿生算法、智能信息处理及民航应用.jlding@ cauc.edu.cn

龚子恒(通信作者),男,硕士生,主要研究数据中心主动容灾.2013952468@ qq.com

1 中国民航大学 计算机科学与技术学院,天津,300300

0 引言

近十年来,基于预测和变换的时间序列的机器学习方法引起了科学界和工业界的广泛关注,其中广受关注的时间序列应用之一是云计算集群的 KPI(Key Performance Indicator,KPI)指标的预测.云集群的 KPI 指标指云计算集群中一些关键的监控指标[1],比如承担云计算集群主要计算功能和业务供给的服务器、集群的 CPU 和内存资源,本质上属于时间序列数据.随着数据中心对云计算集群环境安全性的要求日益增加,对这些 KPI 数据的预测准确度的要求也随之提高,如何准确预测云计算集群中的 KPI 数据的动向和变化,从而提升云计算集群的高可用性,是云计算集群高效运维需面对的问题[2].云计算集群的 KPI 数据变化多样,特征各不相同,线性、平稳性以及周期性常常针对不同的数据集单独分析,但综合分析后的预测准确性并不理想[3].传统模型存在各种欠缺(表 1),急需一种能提高预测准确性的通用型综合分析方法.

2013 年,Gilles 等[11]在降低高噪声电信号数据的复杂度时首次提出经验小波变换 EWT(Empirical Wavelet Transform,EWT),将数据分解成各个内在模态函数 IMFs(Intrinsic Mode Functions,IMFs),大大提高了数据处理能力.为提高云计算集群中 CPU 和内存使用率等关键时序数据的预测精度,本文基于经验小波变换进一步处理得到云 KPI 数据的低中高频 IMFs,根据分解得到的低中高频 IMFs 特征,分别运用 ARIMA[4]、Autoformer[9]、TPA-BiLSTM[10](Temporal Pattern Attention-Bidirectional Long Short-Term Memory)模型对不同 IMFs 进行预测,建立一种应用性更强、预测精度更高的时序数据预测算法 EWT-ARIMA-Auto-TPA(EAAT)方法.本文创新点如下:

1)在 EWT 基础上提出 3 分类模型,使用 EWT 方法将原始序列数据分解为低中高频 3 类 IMFs,根据 3 类 IMFs 不同的特点,选用最合适的预测模型进行预测.

2)将中高频 IMFs 分成中频和高频分开讨论:针对高频 IMFs 中噪声较多的问题,引入 TPA 模型,增强模型对特征的处理能力,并与 BiLSTM 相结合,在噪声较多的数据中也能有更好的结果;在中频 IMFs 中引入最新的基于深度分解架构和自相关机制的 Autoformer 模型进行分析.

3)将注意力机制模型 Autoformer 加入组合模型,规避了其在应对

表 1　传统模型优缺点对比

Table 1　Comparison of advantages and disadvantages of traditional models

分解模型	结构	功能	优点	缺点
ARIMA[4]	5 阶差分	有效提取时序数据特征	训练速度快	未考虑负荷特征之间的时间依赖性
LSTM[5]	1 层 CNN, 2 层 LSTM	有效提取时序数据特征,实现时间依赖性的建模	有效学习到特征之间的时间依赖性	对较长的云 KPI 时序数据的记忆能力受限
注意力机制[6-9]	2 层 LSTM, 多头注意力机制	有效提取时序数据特性,实现时间依赖性的建模	加强 LSTM 对较长的云 KPI 数据的记忆能力	在噪声较多、周期和趋势较差的数据中效果差
TPA-BiLSTM[10]	多个一维滤波器,2 层 LSTM,概率自注意机制,时间模式注意力机制	增加 LSTM 对历史信息的记忆能力及模型对时间依赖性的建模;考虑局部时序数据的特性,防止梯度消失与爆炸	应用空洞卷积提取更长时间跨度的负荷特征;应用时间注意力机制增加 LSTM 的记忆能力;在噪声和随机性因素较多的数据集上表现良好	结构稍复杂

噪声较多、周期较差的数据集上的不利情况,将其长处应用于有一定周期和规律的中频 IMFs 中,进一步提高了预测精度.

本文内容总体安排如下:首先分析已有分类集成模型的优缺点,说明 2 分类组合模型的局限性,以及 3 分类等组合模型的必要性;然后提出基于 EAAT 方法的模型架构;最后在谷歌和亚马逊的 4 个数据集上验证了结果.无论数据是否具有周期性或者稳定性,EAAT 比单个模型在效果上均有较大提升,比 EWT-IF-LSTM 模型在均方根误差上最大降低了 11.26%,验证了 EAAT 效果确实好于传统的 2 分类模型,如 EWT-ARIMA-LSTM、EWT-ARIMA-TPA.

1　组合注意力模型 EAAT 预测方法

1.1　EAAT 的构建背景

预测建模技术已被广泛用于云计算集群 KPI 数据的预测之中,如自回归移动集成平均模型(ARIMA)[4]、循环神经网络(RNN)[12]、长短期记忆网络(LSTM)[5]等.文献[6-7]将 Transformer 类注意力模型引入时间序列预测;Zhou 等[8]提出稀疏注意力机制 Informer,在长时序数据预测中取得了较好的效果;Wu 等[9]提出自相关机制 Autoformer,结合数据分解机制,在周期性和趋势性较好的长时序数据预测上于 2021 年取得了最佳效果.但单个模型常常难以从非线性和非平稳数据中提取特征,很难适应复杂多变的时序数据,容易发生局部优化和过拟合等现象,其准确性往往得不到保证.因此,很多研究利用集成学习的方法对单个模型进行组合,从而提高对云计算集群 KPI 数据预测的准确率.Baig 等[13]将工作负载分为线性和非线性两种类型,并应用 ARIMA 和 RNN 的组合模型进行分类预测;Bi 等[14]专注于识别和预测云负载中的模式问题,提出一种基于资源使用情况的聚类方法来识别周期性任务和非周期性任务.但是上述工作只考虑数据的线性或者周期性,并没有进一步对数据分解和分析以挖掘更深层的特征,对于非线性、非平稳、噪声较高的数据,ARIMA 和 LSTM 等 2 分类组合模型的准确率难以进一步提高.为了对数据进一步分解和研究,文献[15]提出 EMD-CNN-BiLSTM 的混合预测模型用于风力数据的预测,文献[16]提出 EMD-ARIMA 的混合预测模型用于水流数据的预测,文献[17]用 EWT 将数据分解为低频和中高频的 IMFs,用 LSTM 或 GRU 模型对 IMFs 进行预测,并最终合成,从而提高了预测效果.上述模型使用 EMD(Empirical Mode Decomposition,EMD)或 EWT 对数据分解,降低了数据复杂程度以提高效果,区分了低频和中高频 IMFs,但没有用多种模型进行组合和尝试.同时,低中高频 IMFs 本身具有不同的特点,中频 IMFs 没有低频 IMFs 那么大的波动和趋势性,也没有高频 IMFs 中那么多噪声和不确定信息,可以单独对其进行预测.因此,用 EWT 进行数据分解,并运用多种模型分类处理低中高频 IMFs 的方法具有一定的可行性.本文用 EWT 对时间序列数据进行处理,分解为低中高频 3 类 IMFs,然后用最合适的模型进行预测.低频 IMFs 噪声较小、波动较大、趋势较好,适宜用传统的 ARIMA 模型进行预测;中频 IMFs 可以看出原数据中短期的变化规律,有一定的周期性和规律性,适宜用针对周期规律的 Autoformer 进行提取和预测;高频 IMFs 有大量的噪声和随机性因素,适宜用 TPA-BiLSTM 进行处理和预测.

1.2 基于经验小波变换 EWT 的云 KPI 提取和分析

云集群 KPI 数据如 CPU 和内存等,缺少周期性和规律性并存在大量噪声.为了准确地预测非平稳数据,必须首先对非平稳数据进行分解.在已有研究[6]中,经验模态分解(EMD)在处理非线性和非平稳时间序列方面得到了广泛的应用.但 EMD 等方法分解得到的分量过多且伴随模态混乱问题.与其他方法相比,EWT 的主要优点是能够自适应地分割时序数据,傅里叶频谱将信号直接分解成频率分量,并在一段时间内以频域分布模式表示数据.EWT 方法能通过序列分解方式,有效地获取云平台中各种类型的 KPI 时间序列的内在模态变量 IMFs,因此具备了通用性和普适性.在应用预测模型之前,EWT 通过分解和提取云服务机群 KPI 时序数据的信息来获得更高的预测精度.

EWT 具体工作方法[11]如下:对输入的云 KPI 时序数据 $f(t)$ 做傅里叶变换得到傅里叶频谱 $f(\omega)$,在 $[0,\pi]$ 范围内寻找 $f(\omega)$ 的极大值,降序排列后得到 $N-1$ 个边界 $\omega_n(n=1,2,\cdots,N-1)$、$N$ 个连续分段.根据这些分段频谱可构造经验尺度函数 $\hat{\phi}_n(\omega)$ 和经验小波 $\hat{\psi}_n(\omega)$:

$$\hat{\phi}_n(\omega)=\begin{cases}1, & |\omega|\leqslant(1-\gamma)\omega_n,\\ \cos\left(\frac{\pi}{2}\alpha(\gamma,\omega_n)\right), & (1-\gamma)\omega_n\leqslant|\omega|\leqslant(1+\gamma)\omega_n,\\ 0, & \text{其他},\end{cases}\tag{1}$$

$$\hat{\psi}_n(\omega)=\begin{cases}1, & (1+\gamma)\omega_n\leqslant|\omega|\leqslant(1-\gamma)\omega_{n+1},\\ \cos\left(\frac{\pi}{2}\alpha(\gamma,\omega_{n+1})\right), & (1-\gamma)\omega_{n+1}\leqslant|\omega|\leqslant(1+\gamma)\omega_{n+1},\\ \sin\left(\frac{\pi}{2}\alpha(\gamma,\omega_n)\right), & (1-\gamma)\omega_n\leqslant|\omega|\leqslant(1+\gamma)\omega_n,\\ 0, & \text{其他},\end{cases}\tag{2}$$

其中,函数 $\alpha(\gamma,\omega_n)=\beta((1/2\gamma\omega_n)(|\omega|-(1-\gamma)\omega_n))$,函数 $\beta(x)$ 一般需要符合以下条件:

$$\beta(x)=\begin{cases}0, & x\leqslant 0,\\ \beta(x)+\beta(1-x)=1, & x\in(0,1),\\ 1, & x\geqslant 1.\end{cases}\tag{3}$$

一般取 $\beta(x)=x^4(35-84x+70x^2-20x^3)$.

式(1)、(2)中 γ 是为了确保 2 个连续的变换中不存在重叠的分量,取

$$\gamma<\min_n\left(\frac{\omega_{n+1}+\omega_n}{\omega_{n+1}-\omega_n}\right),\tag{4}$$

$$W_f^e(0,t)=<f,\phi_1>=F^{-1}(\hat{f}(\omega)\,\overline{\hat{\phi}_1}(\omega)),\tag{5}$$

$$W_f^e(n,t)=<f,\psi_n>=F^{-1}(\hat{f}(\omega)\,\overline{\hat{\psi}_n}(\omega)),\tag{6}$$

其中,$W_f^e(0,t)$ 至 $W_f^e(n,t)$ 是通过经验尺度函数 $\hat{\phi}_n(\omega)$ 的内积计算得到的近似系数,$F^{-1}(\cdot)$ 表示傅里叶逆变换,$\bar{\cdot}$ 表示复共轭,$\hat{\cdot}$ 表示傅里叶变换.根据式(4)、(5)得到低频分量 $f_0(t)$ 和中高频分量 $f_k(t)$:

$$f_0(t)=W_f^e(0,t)*\hat{\phi}_1(t),\tag{7}$$

$$f_k(t)=W_f^e(k,t)*\hat{\psi}_k(t).\tag{8}$$

对处理后的信号进行重构得到逆经验小波变换 IEWT 的函数:

$$f(t)=f_0(t)+\sum_{k+1}^{N}f_k(t)=W_f^e(0,t)*\hat{\phi}_1(t)+\sum_{k+1}^{N}W_f^e(k,t)*\hat{\psi}_k(t).\tag{9}$$

实验中对每个分量的预测结果记为低频预测结果 $f_0'(t)$ 和中高频分量 $f_k'(t)$,最终预测结果为 $f'(t)$.如果 IMF 数量 k 取较大值时,几个中频信号会存在数据近乎相似、趋势相仿而信息较少的情况.本文实验中 k 取值为 3,分解出的 3 个 IMF 分量 $f_0(t)$、$f_1(t)$、$f_2(t)$ 分别记为低频、中频、高频分量.

如图 1 中的 Google 集群 2011 数据集中某合成的时序数据(编号为 mean CPU usage rate),通过对其进行 EWT 分解可以得到 3 个 IMF 分量,其中第 1 个分量 IMF1 反映原信号的低频部分,包含原信号的长期变化趋势等信息;第 2 个 IMF2 分量反映原信号的中频部分,可以看出原信号中短期的变化规律;噪声及某些突变部分在 IMF3 的高频分量中被体现出来.

1.3 BiLSTM 结合 TPA 模型的云 KPI 高频信息处理

长短时记忆 LSTM(Long Short-Term Memory)网络是常用的改进的 RNN 网络[5],通过网络的记忆能力可将信息保存很长一段时间.LSTM 在一个独立重复单元中存在 4 个运算单元.对于每一个输入的 $\boldsymbol{X}$,LSTM 需要决定来自之前细胞状态的多少信息应该被移除.这一功能由 Sigmoid 函数实现,也被称为“遗忘门”,表示当前重复模块的细胞状态,作为输入传递给下一个重复模块.对于每个细胞状态,接受当前

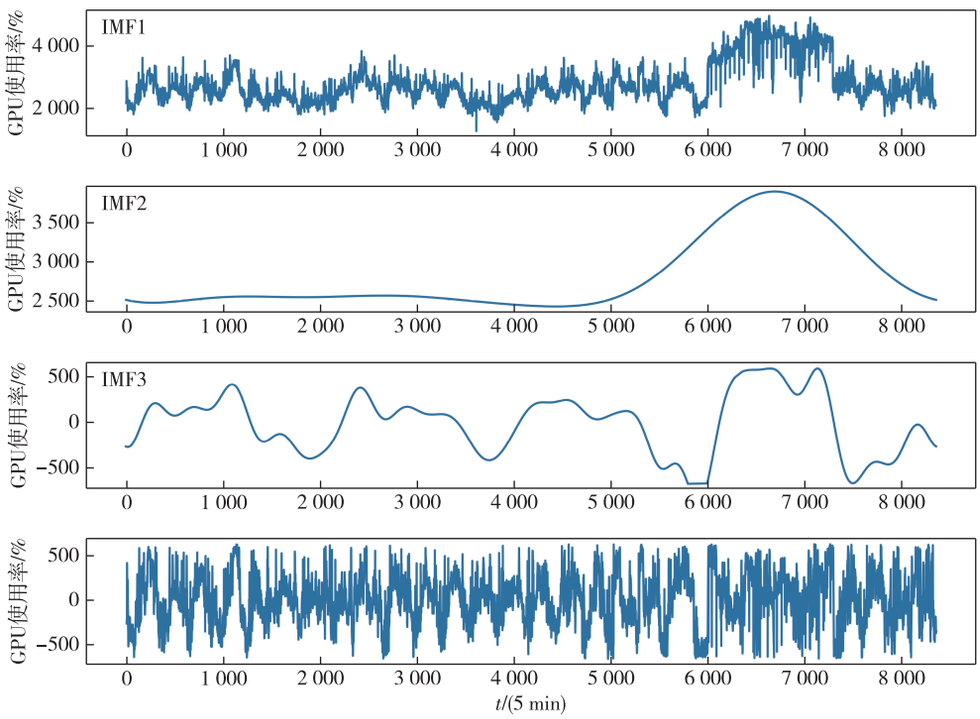

图 1　谷歌某数据集进行 EWT 变换的结果

Fig. 1　Results of empirical wavelet transform for a data set in Google

细胞,遗忘层输入一个介于 0 和 1 之间的值,表示将要从细胞状态中移除的信息量.LSTM 网络决定需要包含在细胞状态中的新信息的数量.将 Sigmoid 层和 tanh 层相乘得到的值与遗忘层计算得到的细胞状态相加,最后每一个单元格都要输出一个值,通过使用另外一个 Sigmoid 层和 tanh 层得到.

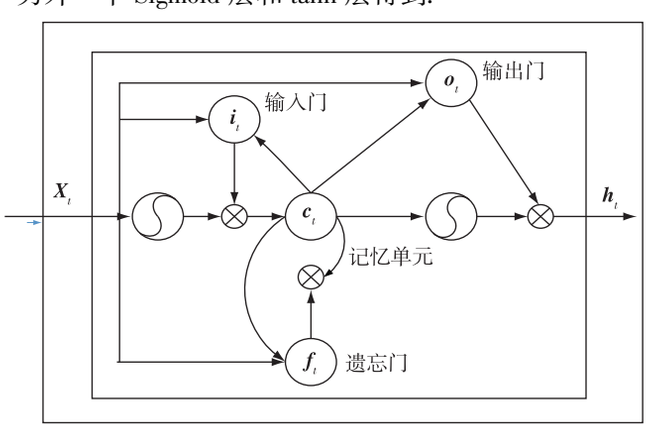

图 2　LSTM 原理

Fig. 2　Schematic diagram of LSTM

LSTM 的最终输出由输出门和单元状态决定:

$$\boldsymbol{h}_t = \boldsymbol{o}_t \tanh(\boldsymbol{c}_t), \tag{10}$$

其中,初始化时 $\boldsymbol{c}_0 = 0$,且 $\boldsymbol{h}_0 = 0$.LSTM 的输入单元为 $\boldsymbol{X}_t$,输出单元为 $\boldsymbol{h}_t$.

传统的 LSTM 有一个缺点,它只能学习时间序列数据的前一个上下文,不能学习同一序列数据的后向上下文信息,也不能编码从后向前的信息.而 BiLSTM 通过将时间序列反向,由正反双向的 LSTM 组成,能够更有效地捕捉时间序列双向的变化.BiLSTM 输出表达式为

$$\boldsymbol{h}_t = \text{concat}(\boldsymbol{h}_{tf}, \boldsymbol{h}_{tb}), \tag{11}$$

式中:$\boldsymbol{h}_t$ 表示 BiLSTM 的隐藏状态向量;concat 表示在输出维度进行拼接操作;$\boldsymbol{h}_{tf}$,$\boldsymbol{h}_{tb}$ 分别表示前向和后向 *LSTM* 的隐藏状态向量.

对于 LSTM 网络而言,每个时刻的输入都是当前时间所有行为组成的向量,使用 BiLSTM 的目的在于能够捕获不同序列方向的更多的特征信息.通过 2 个 LSTM 层以相反的方向处理数据,使得 BiLSTM 可以同时捕捉正向序列信息和反向序列信息.本文采用 BiLSTM 结合 TPA 模型对云 KPI 数据的高频信息进行预测.

EWT 分解得到的高频信号中有较多噪声以及与时间无关的干扰信息,需要一种在数据没有较好的周期性和趋势性时依然能获得较好效果的模型.

传统注意力模型如 CNN+LSTM+Attention[18] 等直接对原始数据的时间序列运用 CNN 完成特征提取,或者仅仅对单个序列的时序特性加以提取,常常

无法兼顾不同序列之间各种错综复杂的关系.而在 TPA-BiLSTM[17] 中使用 BiLSTM 隐状态矩阵的行变量，则涵盖了在各个时间步各个时序下的复杂关系，比传统 CNN-LSTM 多用一层一维卷积层，从而对隐状态矩阵的行向量做特征提取，获得时序关系和各个变量间更深层的联系.在最后注意力加权的过程中，注意力向量是对每个有时间特征信号的时间模式矩阵的行向量加权和，使得 TPA 能从各个时间步中选取最关键的信息.在处理信噪比较高且时间步与不同序列间还存在着非平稳非线性的复杂关系的问题时，TPA 比传统模型有更好的效果. TPA-BiLSTM 的工作流程如图 3 所示.

对输入序列 $\{X_t\}$ 用 BiLSTM 处理，所得向量 $\boldsymbol{h}_{t-w}$ - $\boldsymbol{h}_t$ 表示 BiLSTM 对不同时间步输入所得到的隐藏状态向量，w 即为时序数据长度.定义隐状态矩阵 $\boldsymbol{H}=(\boldsymbol{h}_{t-w},h_{t-w+1},\cdots,\boldsymbol{h}_{t-1})$，其行向量代表每个变量所有时间步下的状况.图 3 中隐状态矩阵 $\boldsymbol{H}$ 的框表示不同的一维卷积核，利用一维卷积沿着 $\boldsymbol{H}$ 的 m 个特征卷积，提取可变时间模式矩阵 $\boldsymbol{H}^C$：

$$\boldsymbol{H}^C=\sum_{l=1}^{w}\boldsymbol{H}_{i,(t-w+1)}*\boldsymbol{C}_{j,T-w+l}, \tag{12}$$

式中：$\boldsymbol{C}_j$ 表示第 j 个长度为 T 的滤波器；T 表示最大长度，w 为其通常取值；* 表示卷积运算.然后定义用注意力机制加权从而计算相关性：

$$f(\boldsymbol{H}_i^C,\boldsymbol{h}_t)=(\boldsymbol{H}_i^C)^{\mathrm{T}}\boldsymbol{W}_a\boldsymbol{h}_t, \tag{13}$$

$$\alpha_i=\mathrm{sigmoid}(f(\boldsymbol{H}_i^C,\boldsymbol{h}_t)), \tag{14}$$

式中：$\boldsymbol{H}_i^C$ 是 $\boldsymbol{H}^C$ 的行向量；$\boldsymbol{W}_a$ 为 $m\times k$ 的权重矩阵；a_i 为注意力权重.利用得到的注意力权重 a_i 与 $\boldsymbol{H}^C$ 加权求和，获得注意力向量 $\boldsymbol{v}_t$：

$$\boldsymbol{v}_t=\sum_{i=1}^{n}\alpha_i\boldsymbol{H}_i^C, \tag{15}$$

式中，n 表示输入变量 x 的特征数.

将 $\boldsymbol{v}_t$ 与 $\boldsymbol{h}_t$ 线性映射后相加获得最终预测值：

$$y_{t-1+\Delta}=\boldsymbol{W}_{h'}h'_t=\boldsymbol{W}_{h'}(\boldsymbol{W}_h\boldsymbol{h}_t+\boldsymbol{W}_v\boldsymbol{v}_t), \tag{16}$$

式中：$y_{t-1+\Delta}$ 表示最终预测值；$\boldsymbol{h}'$ 为用于生成最终值的中间变量；Δ 表示任务预测的时间尺度；$\boldsymbol{W}_{h'}$，$\boldsymbol{W}_h$ 和 $\boldsymbol{W}_v$ 为对应变量的不同权重矩阵.

1.4 基于 Autoformer 分解架构的云 KPI 中频信息处理

Autoformer 结合自相关机制和独特的时序分解机制[10]，对传统注意力类模型在时序预测模型中进行了升级并取得 2021 年最好的效果.一方面，自相关机制计算原始序列和滑动各个步长后序列的自相关系数并做出选择，基于周期性找到子序列之间的相关性，从传统注意力的点到点的相关度上升到子序列和子序列的相关度，从而提高计算效率和信息利用率，且具有较低复杂度 $O(L\log L)$ 的输出.另一方面，Autoformer 在编码器部分的主要目的是对复杂的周期项进行建模，通过多层的时序分解模块，基于滑动平均思想，不断从原始序列中提取周期项和趋势项，用于解码器预测未来的信息.编码器部分输入原始时间序列数据，解码器部分的输入包括趋势项和周期项两个部分.趋势项由两部分组成：一部分是通过原始数据经过时序分解出得到的趋势项，等同于用原始数据近期的趋势项作为解码器的初始化；另一部分是 0 填充的，即尚未得到的未来序列的趋势项.周期项的组成和趋势项类似.

Autoformer 的解码器中对趋势项和周期项分别处理：针对周期项，通过自相关机制与序列的周期特性结合从而聚合有相似特征的子序列；针对趋势项，通过信息累积的方法逐步提取出趋势项信息.

图 4 为 Autoformer 模型工作流程.因为对时序数据周期和趋势的强行提取和加入，使得 Autoformer

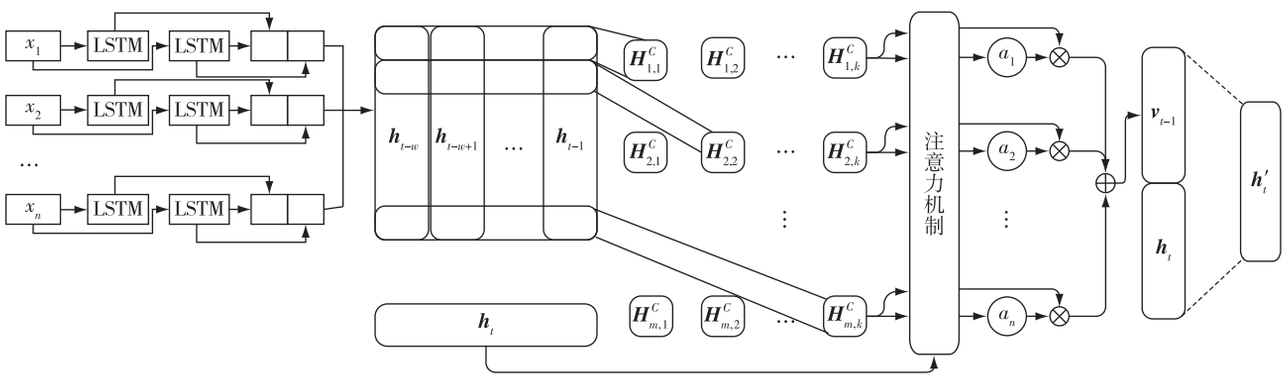

图 3 TPA-BiLSTM 模型架构

Fig. 3 Architecture of TPA-BiLSTM model

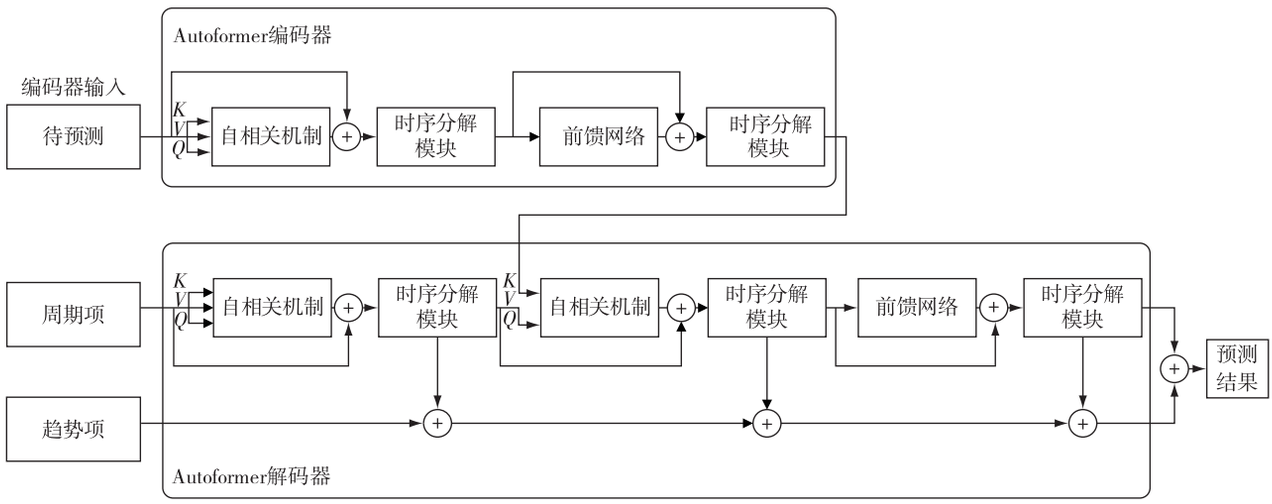

图 4　Autoformer 的流程和机制

Fig. 4　Processes and mechanism of Autoformer

在数据量大,且有季节性和周期性的长时序数据预测中效果较好,而在没有周期和规律的数据中效果较差.EWT 分解得到的中频信号具有一定的周期性和规律性,因此,本文对中频部分采用 Autoformer 的序列分解架构进行预测.

1.5　EAAT 总体预测流程

根据 EWT 与 ARIMA、Autoformer、TPA-BiLSTM 等模型各自的优势,将这 3 个预测模型相结合,以进一步提高数据集数据处理的准确性.该方法的具体操作如下:

第 1 步:输入云计算集群中某段待预测的 KPI 关键时序信息,对其进行 EWT 变换和分解,得到 3 个 IMFs,分别涵盖了低中高频信息.噪声和随机信息主要集中在高频 IMFs 中,中频 IMFs 体现短期有周期和规律性的变化,低频 IMFs 反映原始数据的长期变化和主要趋势.

第 2 步:对噪声较多的中高频 IMFs 使用 iForest 和线性插值等方法清除异常值,并对清除后的空缺部位进行数据填充,从而确保中高频 IMFs 的数据位数保持相对恒定.

第 3 步:对上面预处理后的低中高频 IMFs 分别采用 ARIMA、Autoformer、TPA-BiLSTM 模型进行模拟训练和检测.依据实际状态调节模拟中的各项参数,以获得最佳预期状态.

第 4 步:将各 IMFs 计算后的预测结果,经过对 EWT 的逆变换 IEWT 加以合并,最后得出该 KPI 时序的预测结果.

EAAT 预测方法的整体框图如图 5 所示.

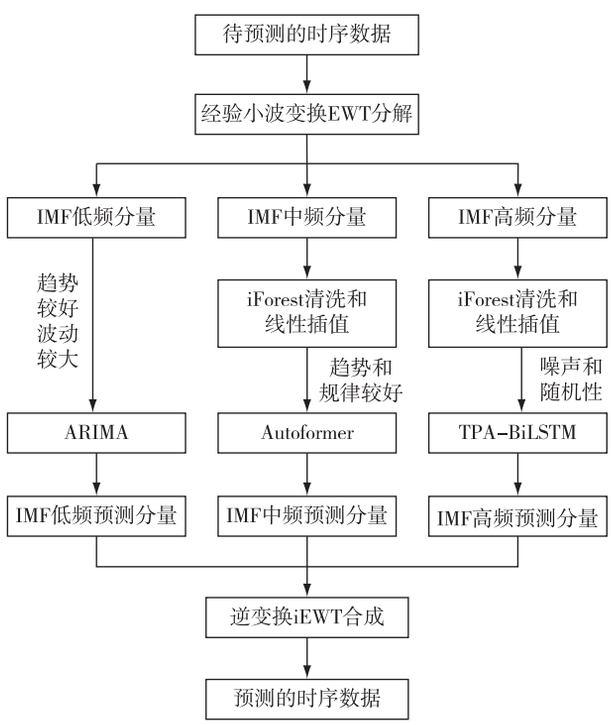

图 5　基于 EAAT 的组合模型预测方法

Fig. 5　Combined model forecasting method of EAAT

2　实验与评价

2.1　数据集准备和参数设置

实验数据集选用谷歌和亚马逊两大云平台所监控采集的 KPI 时序数据集.

谷歌集群数据集 2011 (https://github.com/google/cluster-data)从 2011 年 5 月 1 日 19 时开始记

录 29 d 的 CPU 资源利用率和每个任务在大约 12 500 台计算机上的内存使用情况.取其中 task usage 数据集,分别提取 mean CPU usage 和 assigned memory usage 这两个 KPI 指标构成新的数据集,按 5 min 间隔进行采样,按照任务开始时间进行聚合,数据体现一个数据中心一段时间内总体 CPU 和内存使用率的变换情况,分别记为第 1 组和第 2 组数据集.

亚马逊 KPI 监控数据集(https://github.com/numenta/NAB/tree/master/data/realAWSCloudwatch)来自于美国 AWS 的云监控 CloudWatch 所采集的云服务器的基础资源指标数据,共计 17 种不同类别 KPI 的监控数据,涵盖了云服务器的 CPU 使用率、磁盘 I/O 以及弹性负载均衡(ELB)的请求数等关键监控指标.为了和文献[15]进行对比,本文选用 KPI 编号为 ec2_cpu_utilization_53ea38 和 ec2_cpu_utilization_5f5533 的监控 CPU 利用率的数据集,分别记为第 3 组和第 4 组数据集.

对上述 4 组实验数据集进行切割,其中前 80%作为训练集,后 20%作为测试集.数据集长度代表有多少个时刻的数据,CPU 和内存的使用率数据都是百分比数据,切割后的数据集长度如表 2 所示.

表 2　实验用数据集信息

Table 2　Experimental dataset information

数据集编号	数据集名称	数据集长度/个	训练集长度/个	测试集长度/个
1(谷歌)	mean CPU usage	8 353	6 682	1 671
2(谷歌)	assigned memory	8 353	6 682	1 671
3(亚马逊)	ec2_cpu_53ea38	4 032	3 225	807
4(亚马逊)	ec2_cpu_5f5533	4 032	3 225	807

EAAT 模型选择 Adam 作为优化器,ReLU 函数作为各个神经网络的激活函数,学习率设置为 0.001,滑动窗口长度为 5.经过多次实验和调整后,ARIMA 选择五阶差分回归,TPA-BiLSTM 神经网络层设置 32 个隐层节点,Autoformer 和 TPA-BiLSTM 选择 batch_size 为 32,输入长度为 80 个历史数据,输出长度为 20 个数据,即每次用前 80 个时刻的数据对后 20 个时刻的数据进行预测,然后对每个重复预测的时刻取平均值作为预测结果.

2.2　评价指标

本文使用以下 3 个模型预测量化指标:均方根误差(RMSE)、平均绝对误差(MAE)以及平均绝对百分比误差(MAPE)评估算法性能:

$$e_{\mathrm{RMSE}}=\sqrt{\frac{1}{m}\sum_{j=1}^{m}(X_j-\hat{X}_j)^2},\tag{17}$$

$$e_{\mathrm{MAE}}=\frac{1}{m}\sum_{j=1}^{m}|X_j-\hat{X}_j|,\tag{18}$$

$$e_{\mathrm{MAPE}}=\frac{1}{m}\sum_{j=1}^{m}\left|\frac{X_j-\hat{X}_j}{X_j}\right|,\tag{19}$$

其中,m 是序列长度,X_j 是数据的真实值,$\hat{X}_j$ 是数据的预测值.

2.3　实验结果与分析

对 4 个数据集进行 ADF 校验,一般当 p 值大于 0.05 时,可以认为序列是不平稳的.实际校验数值如表 3 所示.数据集 1、4 不太平稳,数据集 2、3 相对平稳;低频 IMFs 中多数都不平稳,但具有较好的趋势性,在中高频 IMFs 中数据基本都是平稳的.

表 3　数据集 ADF 检测的 p 值

Table 3　The p-value in ADF detection of datasets

数据集	原始数据	IMF1	IMF2	IMF3
1	0.059	0.957	0.005	6.13×10^{-15}
2	3.24×10^{-8}	0.458	2.43×10^{-12}	1.43×10^{-17}
3	4.72×10^{-17}	0.873	1.37×10^{-23}	3.10×10^{-21}
4	0.838	0.958	0.026	2.90×10^{-21}

将本文模型和传统的单一预测模型 LSTM、TPA-LSTM、Autoformer 以及 EWT-LSTM、EWT-ARIMA-LSTM、EWT-ARIMA-TPA 作为实验的对比对象,其中 EWT-ARIMA-LSTM 和 EWT-ARIMA-TPA 指中低频 IMFs 上使用 ARIMA 模型、高频 IMFs 分别使用 LSTM 或 TPA 的模型.测试集上的各模型预测结果如图 6 和图 7 所示,评价指标结果如表 4—7 所示,最优结果以加粗说明.

表 4　各模型在数据集 1 上的效果

Table 4　Performance of each model on dataset 1

模型	RMSE	MAE	MAPE
LSTM	711.072	530.004	0.149
TPA-LSTM	604.552	450.411	0.133
Autoformer	546.635	402.866	0.143
EWT-LSTM	382.682	289.749	0.096
EWT-ARIMA-LSTM	363.922	269.498	0.090
EWT-ARIMA-TPA	362.893	269.734	0.091
EAAT	**357.578**	**264.287**	**0.088**

2.3.1　谷歌数据集的预测结果

处理后的谷歌两个数据集包含一个集群中所有

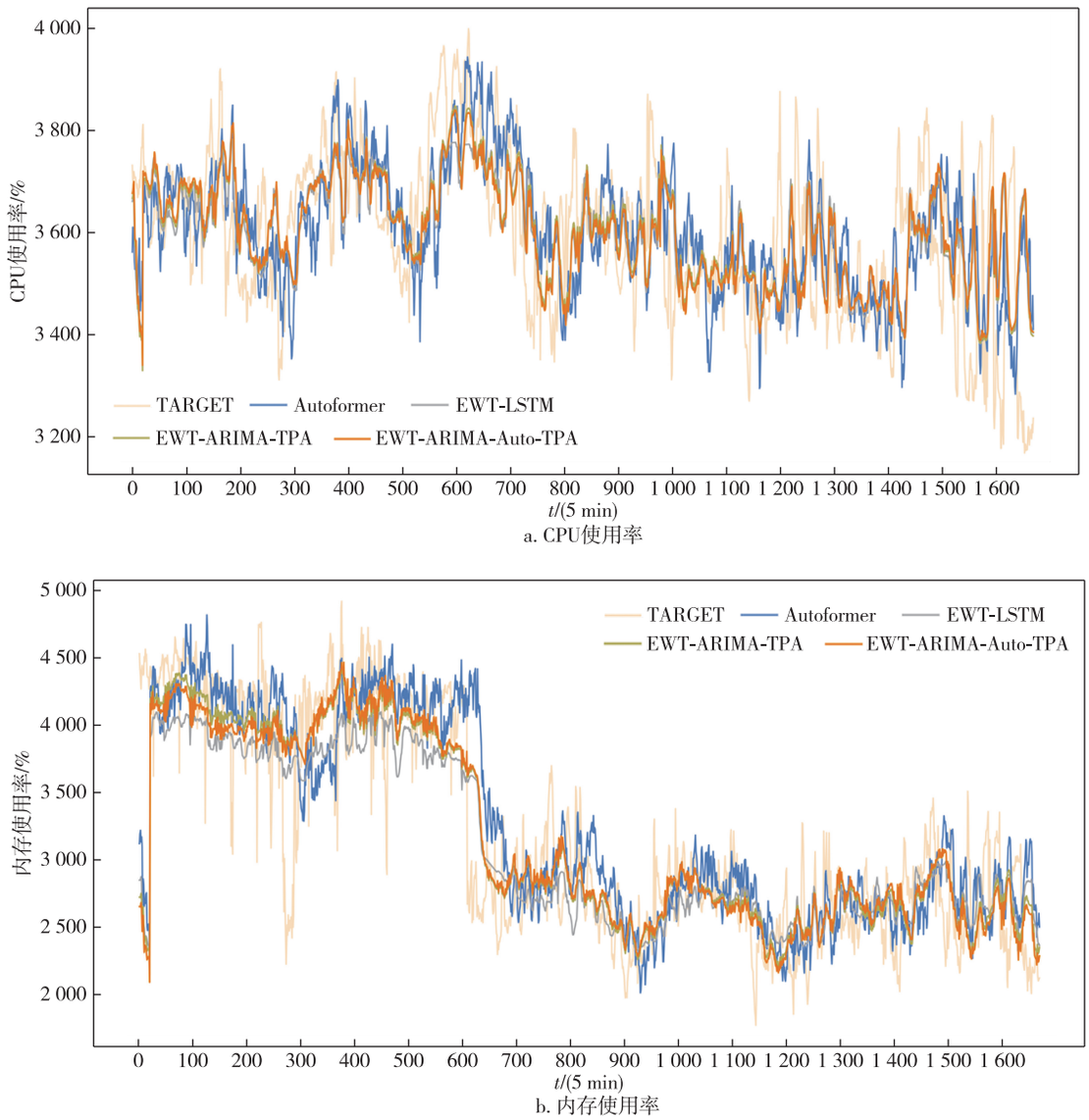

图 6　两个谷歌数据集的测试集上的预测结果

Fig. 6　Prediction results on the test set of two Google datasets

机器的 CPU 和内存使用率的变化情况.由图 6 所示可知,在谷歌数据集的测试集上,EAAT 方法整体拟合程度和稳定程度都略优于其他组合模型,远胜于单一的非组合模型.

两个谷歌数据集聚合了一个集群中所有机器的信息,数据更不平稳,整体误差较大.由表 4、5 可知:在 3 个评价指标上,EAAT 方法都优于 EWT-LSTM 模型;对比 EWT-ARIMA-TPA 模型,EAAT 方法在两个数据集上 RMSE 和 MAPE 指标中效果更优.

2.3.2　亚马逊数据集的预测结果

处理后的亚马逊两个数据集包含一个集群中所有机器的 CPU 和内存使用率的变化情况.由图 7 所示可知:在相对平稳的数据集 3(53ea38)中,EAAT 方法在稳定性和趋势性把握上相对最好;在数据集 4(5f5533)中,因为该数据集的特殊情况,测试集数据

表 5　各模型在数据集 2 上的效果

Table 5　Performance of each model on dataset 2

模型	RMSE	MAE	MAPE
LSTM	125.472	91.525	0.025
TPA-LSTM	124.551	91.251	0.024
Autoformer	136.534	106.560	0.029
EWT-LSTM	123.172	91.230	0.024
EWT-ARIMA-LSTM	122.490	91.496	0.025
EWT-ARIMA-TPA	123.992	90.433	0.024
EAAT	**122.200**	**90.961**	**0.023**

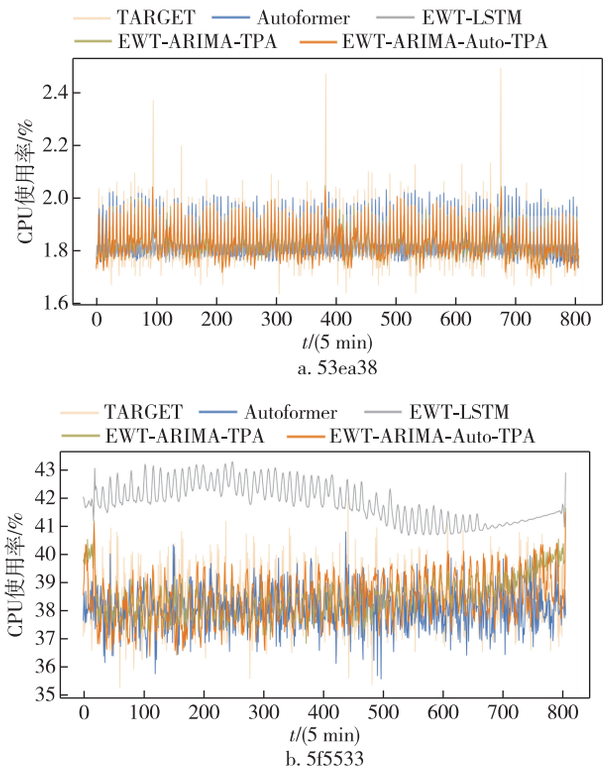

图 7　两个亚马逊数据集的测试集上的预测结果

Fig. 7　Predicted results on test set of two Amazon datasets

表 6　各模型在数据集 3 上的效果

Table 6　Performance of each model on dataset 3

模型	RMSE	MAE	MAPE
LSTM	0.073	0.048	0.026
TPA-LSTM	0.063	0.040	0.021
Autoformer	0.069	0.046	0.025
EWT-LSTM	0.071	0.052	0.028
EWT-ARIMA-LSTM	0.069	0.050	0.027
EWT-ARIMA-TPA	0.066	0.044	0.024
EAAT	**0.063**	**0.045**	**0.023**

表 7　各模型在数据集 4 上的效果

Table 7　Performance of each model on dataset 4

模型	RMSE	MAE	MAPE
LSTM	5.047	4.876	0.128
TPA-LSTM	2.681	1.943	0.051
Autoformer	1.389	1.091	0.028
EWT-LSTM	3.804	3.636	0.065
EWT-ARIMA-LSTM	1.221	0.932	0.024
EWT-ARIMA-TPA	1.075	0.852	0.022
EAAT	**1.049**	**0.840**	**0.023**

比起训练集中的数据有一段陡升,所以实际值远远高于预测值,所有模型的误差都相对较大,此时 EAAT 依旧能在整体上优于其他模型,并在 3 个评价指标上取得了最好的结果(表 6、7).

由实验结果可见:EAAT 方法在 4 个特征各不相同的数据集中,RMSE 和 MAPE 指标均表现最佳,MAE 指标也表现较好;在相对不平缓、误差较大的数据集 1 和 4 中,EAAT 方法效果提升更加明显.因此,EAAT 方法可以有效提高对时间序列的预测的准确率,且具有广泛的适用性.

3　结束语

本文通过 EWT 将到云 KPI 数据分解成低中高频 3 类 IMFs,分别运用 ARIMA、Autoformer、TPA-BiLSTM 模型对低中高频 IMFs 进行预测,再经过逆变换 IEWT 加以合并,最后得出该 KPI 时序的预测结果,即 EAAT 预测方法.本文运用 EAAT 预测方法对从谷歌和亚马逊服务器资源负载数据提取出的 CPU 负载时间序列数据进行了预测,结果表明,与单一 ARIMA、Autoformer 和 TPA-LSTM 模型预测相比,EAAT 在 3 个评价指标上均有显著提升,预测效果更佳,在同一数据集中,比 EWT-IF-LSTM 模型性能更优,验证了 EAAT 方法的先进性.

参考文献

References

[1] Gao J C,Wang H Y,Shen H Y.Machine learning based workload prediction in cloud computing[C]//2020 29th International Conference on Computer Communications and Networks (ICCCN).August 3-6,2020,Honolulu,HI,USA.IEEE,2020:1-9

[2] Muteeh A,Sardaraz M,Tahir M.MrLBA:multi-resource load balancing algorithm for cloud computing using ant colony optimization[J].Cluster Computing,2021,24(4):3135-3145

[3] Kumar J,Singh A K.Performance evaluation of metaheuristics algorithms for workload prediction in cloud environment[J].Applied Soft Computing,2021,113:107895

[4] Wen Q,Wang Y P,Zhang H D,et al.Application of ARIMA and SVM mixed model in agricultural management under the background of intellectual agriculture[J].Cluster Computing,2019,22(6):14349-14358

[5] Fan C,Wang J Y,Gang W J,et al.Assessment of deep recurrent neural network-based strategies for short-term building energy predictions[J].Applied Energy,2019,236:700-710

[6] Wen Q,Zhou T,Zhang C,et al.Transformers in time series:a survey[J].arXiv e-print,2022,arXiv:2202.07125

[7] Zhang Z Y,Tang X H,Han J Z,et al.Sibyl:host load prediction with an efficient deep learning model in cloud computing[C]//International Conference on Algorithms and Architectures for Parallel Processing.Cham:

Springer,2018:226-237
[8] Zhou H Y,Zhang S H,Peng J Q,et al.Informer:beyond efficient transformer for long sequence time-series forecasting[J].Proceedings of the AAAI Conference on Artificial Intelligence,2021,35(12):11106-11115
[9] 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
[10] Shih S Y,Sun F K,Lee H Y.Temporal pattern attention for multivariate time series forecasting[J].Machine Learning,2019,108(8):1421-1441
[11] Gilles J. Empirical wavelet transform[J]. IEEE Transactions on Signal Processing, 2013, 61(16): 3999-4010
[12] Saud A S, Shakya S. Analysis of look back period for stock price prediction with RNN variants:a case study on banking sector of NEPSE[J]. Procedia Computer Science,2020,167:788-798
[13] Baig S U R, Iqbal W, Berral J L, et al. Adaptive prediction models for data center resources utilization estimation[J].IEEE Transactions on Network and Service Management,2019,16(4):1681-1693
[14] Bi J, Li S, Yuan H T, et al. Integrated deep learning method for workload and resource prediction in cloud systems[J].Neurocomputing,2021,424:35-48
[15] Yang Z M, Peng X S, Wei P J, et al. Short-term wind power prediction based on CEEMDAN and parallel CNN-LSTM[C]//2022 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia). July 8-11, 2022, Shanghai,China.IEEE,2022:1166-1172
[16] Wang Z Y, Qiu J, Li F F. Hybrid models combining EMD/EEMD and ARIMA for long-term streamflow forecasting[J].Water,2018,10(7):853
[17] 王超.云环境中时序数据的预测和异常检测算法的研究[D].南京:南京大学,2019
WANG Chao. Research on prediction and anomaly detection algorithm of time series in cloud environment[D].Nanjing:Nanjing University,2019
[18] 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
LI Mei, NING Dejun, GUO Jiacheng. Attention mechanism-based CNN-LSTM model and its application[J]. Computer Engineering and Applications, 2019, 55(13):20-27

Cloud KPI data prediction based on combined attention model EAAT

DING Jianli[1] GONG Ziheng[1]
1 College of Computer Science and Technology,Civil Aviation University of China,Tianjin 300300,China

Abstract To accurately analyze the dynamics and changing trends of KPI (Key Performance Indicator) data in the daily monitoring of cloud computing clusters and predict its subsequent development to achieve high availability of cloud computing clusters,we propose a three-frequency cloud KPI data prediction approach based on combined attention model of EWT-ARIMA-Auto-TPA (EAAT for short).First,low,medium and high frequency Intrinsic Mode Variables (IMFs) of cloud KPI data are obtained via Empirical Wavelet Transform (EWT) to reduce the complexity of data prediction.Second,according to the information characteristics of low,medium and high frequency IMFs obtained from the decomposition,models of ARIMA,Autoformer,and TPA-BiLSTM are used to predict each type of IMFs.Finally,the classification prediction results are combined through the Inverse EWT (IEWT) to obtain the prediction result of the KPI.The proposed prediction approach has been verified on four datasets from Google and Amazon.Whether the data is periodic and stable or not,the proposed approach outperforms comparison models.

Key words cloud KPI data; time series prediction; empirical wavelet transform (EWT); combined attention model; bidirectional long short-term memory (BiLTSM) network