



# 计算机科学

COMPUTER SCIENCE

## 基于EMPC-BCGRU的云虚拟机CPU负载分析预测

谢同磊, 邓莉, 尤文龙, 李锐龙

引用本文

谢同磊, 邓莉, 尤文龙, 李锐龙. 基于EMPC-BCGRU的云虚拟机CPU负载分析预测[J]. 计算机科学, 2023, 50(8): 243-250.

XIE Tonglei, DENG Li, YOU Wenlong, LI Ruilong. [Analysis and Prediction of Cloud VM CPU Load Based on EMPC-BCGRU](#) [J]. Computer Science, 2023, 50(8): 243-250.

---

## 相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

### [基于时间卷积网络的云平台负载预测方法](#)

Cloud Platform Load Prediction Method Based on Temporal Convolutional Network  
计算机科学, 2023, 50(7): 254-260. <https://doi.org/10.11896/jsjx.220500036>

### [基于CEEMDAN-ConvLSTM组合模型的云计算负载预测方法](#)

Cloud Computing Load Prediction Method Based on Hybrid Model of CEEMDAN-ConvLSTM  
计算机科学, 2023, 50(6A): 220300272-9. <https://doi.org/10.11896/jsjx.220300272>

### [一种高精度LSTM-FC大气污染物浓度预测模型](#)

A Kind of High-precision LSTM-FC Atmospheric Contaminant Concentrations Forecasting Model  
计算机科学, 2021, 48(6A): 184-189. <https://doi.org/10.11896/jsjx.200600090>

### [基于形态学图像增强和PCNN的脑部CT与MRI图像融合](#)

Brain CT and MRI Image Fusion Based on Morphological Image Enhancement and PCNN  
计算机科学, 2020, 47(10): 194-199. <https://doi.org/10.11896/jsjx.190700185>

### [基于神经网络和NLP的软件需求安全分析研究](#)

Analysis Research of Software Requirement Safety Based on Neural Network and NLP  
计算机科学, 2019, 46(6A): 348-352.

# 基于 EMPC-BCGRU 的云虚拟机 CPU 负载分析预测

谢同磊 邓莉 尤文龙 李锐龙

武汉科技大学计算机科学与技术学院 武汉 430065

智能信息处理与实时工业系统湖北省重点实验室 武汉 430065

(1139328853@qq.com)

**摘要** 云平台资源预测对于云资源管理和节能具有非常重要的意义。云虚拟机技术是云平台为了充分利用物理资源而实施的一种虚拟化手段,但是有效的云虚拟机负载预测仍具有挑战性,因为云虚拟机负载具有周期性和非周期性的变化模式以及突变的负载峰值,云虚拟机负载受到用户随机提交作业的影响。为了准确分析云虚拟机负载的变化模式,提升云虚拟机 CPU 负载预测性能,提出了一种基于分解-预测的云虚拟机负载预测方法。通过经验模态分解和主成分分析的云虚拟机负载模式分解,得到不同尺度的特征波动序列;预测模型的卷积层能够充分提取分解后的特征,并通过双向门控循环神经网络双向学习序列的前向和后向依赖关系,提高了预测模型学习云虚拟机负载变化模式的能力。最后,在真实云环境微软 Azure 产生的 2019 VM 数据集上进行单步和多步预测实验,验证了所提预测方法的有效性。

**关键词**:云虚拟机;分解模式;本征模函数;负载预测;神经网络模型

中图分类号 TP391

## Analysis and Prediction of Cloud VM CPU Load Based on EMPC-BCGRU

XIE Tonglei, DENG Li, YOU Wenlong and LI Ruilong

College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China

Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China

**Abstract** Cloud platform resource prediction is of great significance for resource management and energy saving. Cloud VM technology is a virtualization method implemented by the cloud to make full use of physical resources, but effective cloud VM load prediction is still challenging, because the cloud VM load has periodic and aperiodic change patterns and sudden load peaks, and the cloud VM load is affected by the random submission of jobs by users. In order to accurately analyze the change mode of VM load and improve the performance of VM CPU load prediction, a cloud VM load prediction method based on decomposition-prediction is proposed. Through EMD and PCA of cloud VM load mode decomposition, the characteristic fluctuation sequences of different time scales are obtained. The convolution layer of the prediction model can fully extract the decomposed features, and learn the forward and backward dependencies of the sequence through the bidirectional gated cyclic neural network, which improves the ability of the prediction model to learn the load change mode of the VM. Finally, single-step and multi-step prediction experiments are performed on the 2019 VM data sets generated by Microsoft Azure in the real cloud environment, which verifies the effectiveness of the prediction method.

**Keywords** Cloud VM, Decomposition mode, Intrinsic mode function, Load prediction, Neural network model

## 1 引言

由于强大的计算能力和按需提供服务的便捷性,云计算在医疗保健、金融服务和学术领域等很多行业发挥着重要作用,如微软 Azure、谷歌云平台(Goog Cloud Platform, GCP)和亚马逊网络服务(Amazon Web Services, AWS)等。研究表明<sup>[1]</sup>,根据微软 Azure 虚拟机(VM)的日志数据分析,有 60% 的 VM 平均 CPU 利用率低于 20%,有 40% 的 VM 第 95 个

百分位数的最大 CPU 利用率低于 50%,谷歌云平台主机 CPU 的平均利用率在 10%~45% 之间波动<sup>[2]</sup>。低资源利用率会使数据中心的效率变得低下,造成不必要的资源浪费,增加企业对数据中心的投资成本。有效的负载预测能帮助系统进行动态的决策和相应的资源调节<sup>[3]</sup>,将虚拟机进行迁移和整合并重新定位到少量活动的物理服务器中,减少物理结点的数量,降低数据中心的能耗。

云虚拟机负载呈现出周期性和非周期性的变化模式以及

到稿日期:2022-06-29 返修日期:2023-02-28

基金项目:新一代信息技术创新项目(2020ITA01005)

This work was supported by the New Generation Information Technology Innovation Project(2020ITA01005).

通信作者:邓莉(dengli@wust.edu.cn)

负载突变的峰值,有效地预测云虚拟机负载存在困难。微软研究人员在文献[1]中指出,云虚拟机负载行为在多个生命周期中表现出相似性,可以利用历史资源使用情况来预测未来的工作负载,并且傅里叶变换可以用于检测虚拟机负载在不同时间尺度的周期性。为了分析虚拟机负载的变化特点,文献[4]基于 Apache Spark 把虚拟机平均 CPU 利用率数据分解为多种变化模式并使用 ARIMA 模型进行预测,但平均 CPU 利用率的预测难以满足服务水平协议(Service Level Agreement, SLA)。文献[5]首先对 VM CPU 负载进行了不同频率的下采样,分解出 VM 最大 CPU 利用率的局部性、周期性和趋势性,然后提出了时间感知残差网络模型分别对负载序列进行建模预测。由于文献[5]只是粗粒度地基于时间间隔 5 min, 1h, 1d 对负载进行下采样,因此难以消除负载噪声的影响。文献[6]提出了一种结合长短期记忆网络和多级小波分解的时间序列预测模型,为时间序列分解预测提供了一种思路,但是小波分解方法需要提前给定小波的分解尺度与基函数,信号中所包含的时间尺度难以被完整地分解出来,需要人为地设置模型的参数。

本文对云虚拟机最大 CPU 利用率进行预测建模,提出了基于经验模态分解和主成分分析的云虚拟机负载分解模式(Empirical Mode Decomposition and Principal Component Analysis, EMPC)以及基于卷积的双向门控循环神经网络(Bidirectional Convolution Gated Recurrent Unit, BCGRU)模型。预测云虚拟机最大 CPU 利用率能够满足云平台用户最大资源请求的需要,有助于保障云平台服务水平协议 SLA。通过经验模态分解将虚拟机负载序列进行多尺度分解,得到一组不同尺度的本征模函数(Intrinsic Mode Function, IMF),然后基于 PCA 算法去除 IMF 中的噪声,降低输入模型的维数和冗余度。基于 EMPC 的分解模式能够分析出虚拟机负载不同变化模式的 IMF,最后得到一组 PCA 序列作为预测模型的输入。预测模型 BCGRU 的卷积层能够充分提取虚拟机负载 PCA 序列的特征,其内核门控循环单元比 LSTM 门控结构少、效率更高,双向门控循环神经网络可以双向学习序列的前向和后向依赖关系,从而提高预测模型学习云虚拟机负载长期依赖关系的能力。

## 2 相关工作

云平台负载预测属于时间序列预测的范畴,因为在云与网络系统中收集的负载数据是按时间来组织的。主机负载是从更加宏观的角度描述物理机的负载情况,而虚拟机负载则是在物理机上细粒度的负载表现。本文总结了主机负载预测和虚拟机负载预测的相关研究。

(1)主机负载预测的相关工作。Di 等<sup>[7]</sup>手动设计了新的特征来表示历史值的波动情况,并使用 Bayes 方法进行主机负载预测。为了提高负载在线预测的精度,文献[8]基于深度循环神经网络编码器-解码器设计了一种主机负载在线预测模型。Song 等<sup>[9]</sup>使用长短期记忆网络 LSTM 模型进行主机负载预测,但随着隐藏层的增加会出现模型过拟合等问题。文献[10]提出了一种基于深度学习的主机负载预测算法,利用顶级稀疏自编码器(Top-Sparse Auto-encoder, TSA)从

原始高维负载数据中提取负载的基本表示,并将 TSA 和门控循环单元 GRU 块集成到循环神经网络 RNN 中,以实现对工作负载进行自适应。由于虚拟化技术,主机被虚拟化为多台虚拟机,主机负载在宏观层面表现物理机的负载情况,而虚拟机负载则在更微观的层面表示物理机的负载情况。虚拟机比主机更具有操作性,如进行虚拟机迁移和整合等。

(2)虚拟机负载预测的相关工作。Abubakar 等<sup>[4]</sup>将虚拟机负载数据分解为观测、趋势、季节性和残差 4 种变化模式,分析了虚拟机负载的变化特点和可预测性,然后基于 ARIMA 模型进行预测。文献[5]提出了时间感知残差网络模型,分别对下采样分解后的 VM 最大 CPU 利用率的局部性、周期性、和趋势性进行建模预测。为了选择合适的 VM 进行迁移,文献[11]利用卷积神经网络 CNN 和门控循环单元 GRU 建立虚拟机负载分类模型,虚拟机被划分为延迟敏感和不敏感类别。为了进行有效的虚拟机负载预测,从而促进虚拟机迁移和整合, Karim 等<sup>[12]</sup>提出了一种新颖的 Bi-LSTM 混合循环神经网络预测模型,用于预测云虚拟机 CPU 工作负载,增强了模型的非线性负载学习能力,与其他统计学模型相比具有更高的准确性。文献[13]将门控循环单元和长短期记忆神经网络进行组合,提出了混合神经网络模型 LSRU(Long Short Term Memory and Gate Recurrent Unit)进行虚拟机工作负载预测。由于组合的神经网络模型过于复杂以及存在梯度消失问题,预测的准确性和时间性能难以达到较好的效果。

为了充分分析虚拟机负载的序列特征,本文提出基于 EMPC 对虚拟机负载进行模式分解,降维去噪后的多组 PCA 序列能够刻画虚拟机负载在不同尺度下的变化模式。为了学习虚拟机负载多组 PCA 序列的数据特征,本文建立了 BCGRU 预测模型。预测模型 BCGRU 的卷积层能够充分提取虚拟机负载 PCA 序列的特征,其内核门控循环单元比 LSTM 门控结构少、效率更高,双向门控循环神经网络可以双向学习序列的前向和后向依赖关系。表 1 综合比较了本文工作与其他相关工作的特征处理方法。与以往的方法相比,本文提出的虚拟机负载分解模式能够充分地分析虚拟机最大 CPU 利用率的特征,预测模型能够有效地提取这些特征和高效地学习虚拟机负载序列长期依赖的关系。

表 1 云平台资源预测的特征处理方法的对比

Table 1 Comparison of feature processing methods for cloud resource prediction

预测方法	特征处理
Bayes	依赖人工提取特征,难以准确表征负载的变化情况,无法预测序列具体的变化
ARIMA	对负载序列进行季节性分解,有利于虚拟机负载的特征表示,ARIMA 适合预测较短时间内的虚拟机负载情况,需频繁训练模型
T-ResNet	粗粒度地对虚拟机负载进行下采样,分别表征负载的时间局部性、周期性和趋势性
LSTM	端到端模型,不依赖单独的特征提取,可以提取到时间依赖特征
Bi-LSTM	端到端模型,不依赖单独的特征提取,可以提取双向依赖特征分解,表征负载的多尺度变化模式,基于卷积提取特征,可以提取双向依赖
EMPC-BCGRU	

### 3 云虚拟机数据分析

谷歌、亚马逊、微软 Azure 等云提供商在为客户端提供云服务方面的技术比较成熟。研究这些公司如何使用云指标来有效地安排客户的工作,从而有效地利用共享基础设施,对云平台资源的调度具有重要意义。以前的大量研究工作是关于谷歌公开发布的跟踪 2011 云平台数据<sup>[2]</sup>。微软 Azure 于 2017 年公布了 VM CPU 利用率数据并发表了文献<sup>[1]</sup>,然后于 2019 年公布了新的 VM CPU 利用率数据。除了 VM 的部署规模大小以外,微软 Azure 2017 VM 数据集和微软 Azure 2019 VM 数据集非常相似,本文基于微软 Azure 2019 VM 数据集进行分析和实验。

#### 3.1 云虚拟机资源使用情况分析

微软 Azure 2019 VM 数据集包含了一个月连续 30 天的虚拟机运行日志,每条记录的时间间隔为 5min,主要包括订阅信息、部署信息、虚拟机类型、核数、创建时间、删除时间、最小 CPU 利用率、最大 CPU 利用率、平均 CPU 利用率和第 95 个百分位数的最大 CPU 利用率等指标。图 1 表明,微软 Azure 中虚拟机的 CPU 利用率并不高,88% 的 VM 平均 CPU 利用率低于 40%,36% 的 VM 最大 CPU 利用率低于 40%,45% 的 VM 第 95 个百分位数的最大 CPU 利用率低于 40%。有效的虚拟机 CPU 负载预测能够帮助虚拟机进行迁移和整合,将虚拟机重新定位到少量活动的物理服务器中,关闭空闲物理机,达到节能的目的。

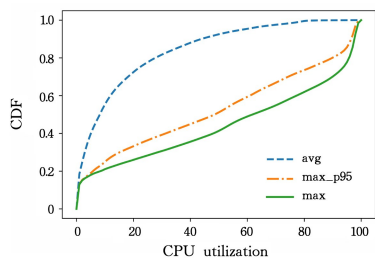


图 1 虚拟机 CPU 资源利用率 CDF 图

Fig.1 CDF of VM CPU resource utilization

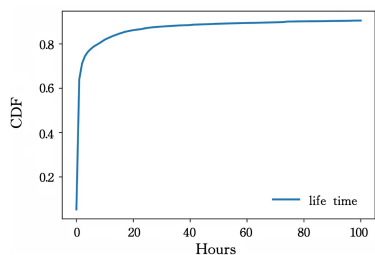


图 2 虚拟机生命周期小时数 CDF 图

Fig.2 CDF of VM life cycle hours

用户创建订阅后可以将虚拟机进行部署,一次部署可以有多个虚拟机来执行指定的任务。由于用户可以在部署中动态地改变虚拟机的数量来执行任务,包括增加和撤销虚拟机,这可能会导致很多虚拟机只执行很短一段时间就被撤销。从图 2 可以看出,大部分虚拟机执行的时间都很短,64% 的虚拟机的生命周期时间不到 1h,82% 的虚拟机生命周期时间没有超过 10h。其中,持续运行了一个月的虚拟机大约占总虚拟机数的 6.65%,此类虚拟机被称为长期运行的虚拟机。

文献<sup>[1]</sup>指出,相对较小比例的长时间运行的虚拟机实际上占总核心小时数的 95% 以上,消耗了云平台主要的资源。因此,本文对长期运行的虚拟机进行预测。

#### 3.2 云虚拟机 CPU 负载行为分析

长期运行的虚拟机会在较长一段时间内占据物理机的资源,分析虚拟机的 CPU 负载行为对于预测建模而言具有重要意义。虚拟机 CPU 负载可能会呈现多种变化模式,将虚拟机 CPU 负载序列拆分为多个组件序列通常很有帮助,每个组件序列代表一个基本模式类别,这是因为分解的方式提供了一种结构化的方法来思考如何更好地捕获每个组件序列特点,从而降低预测模型的复杂性<sup>[4]</sup>。

季节性分解是一种用于时间序列分解的统计学方法,可以将一组原始序列分解为趋势性(Trend)、季节性(Seasonality)和残差(Residual)3 部分。其中,趋势性为时间序列中长期呈现出的增长或者下降的规律,季节性为时间序列中短期的周期变化规律,残差为时间序列中随机扰动的部分,残差值项是一个均值为 0 的随机变量。采用季节性分解的加法模型分解原始序列,此时原始 VM CPU 负载序列  $v(t)$  被分解为如下几个部分:

$$v(t) = T(t) + S(t) + R(t) \quad (1)$$

其中, $T(t)$ , $S(t)$ , $R(t)$  分别代表分解后的趋势项、季节项和残差项。图 3 详细说明了采样后的虚拟机 CPU 利用率序列和其对应的季节性分解部分,依次分别为原始序列、趋势性序列、季节性序列和残差序列。在图中可以发现如下特点:从趋势子图来看,CPU 利用率在某个时间点迅速增长后呈现下降趋势,这种不确定的变化趋势会因为用户对实际任务的需求而变化。从季节子图来看,CPU 利用率呈现出每天的周期变化,这说明虚拟机以天为周期单位消耗物理机资源。从残差子图来看,CPU 利用率随着时间不停扰动,这主要是由短期的不确定性造成的。

由于现实世界的问题是混乱和嘈杂的,季节性分解的加法模型或乘法模型均可能无法将特定时间序列干净或完美地分解。文献<sup>[14]</sup>表明,由于噪声的存在,以往基于傅里叶变换和 ACF 分析以及基于时间间隔的分解难以准确描述虚拟机负载变化模式。

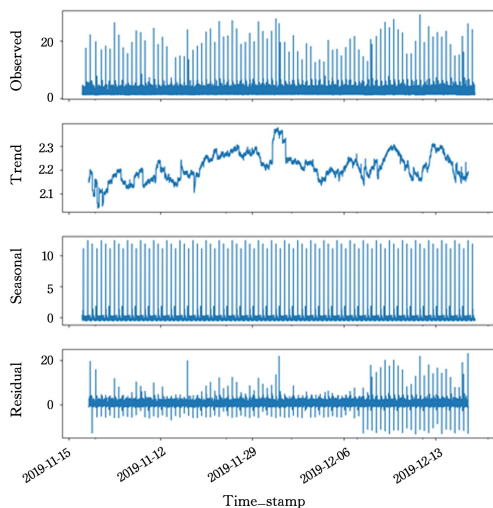


图 3 通过季节性分解的虚拟机 CPU 利用率

Fig.3 CPU utilization of VM by seasonal decomposition

## 4 云虚拟机 CPU 负载预测方法

用户通过订阅申请云平台资源,并动态地增减虚拟机的数量来执行不同的任务,这导致了很多虚拟机只运行很短的一段时间。本文对长期运行的虚拟机最大 CPU 资源利用率进行预测,因为长期运行的虚拟机消耗了云平台大量的物理机资源,最大 CPU 利用率能最大程度地反映出虚拟机对宿主机的资源需求,并可以最大限度地满足用户的需要。通过第 3 节的负载分析,本文建立了基于 EMPC-BCGRU 虚拟机的负载预测方法,如图 4 所示。

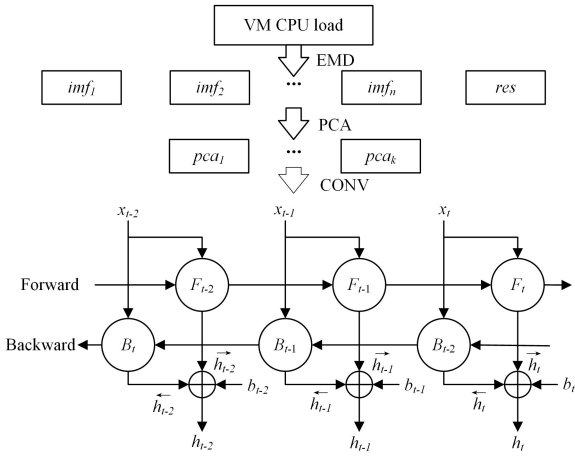


图 4 基于 EMPC-BCGRU 虚拟机的 CPU 负载预测方法

Fig. 4 VM CPU load prediction method based on EMPC-BCGRU

本文先基于经验模态分解将单一负载序列分解为不同尺度的本征模函数,这些本征模函数能更加细节性地表现特征波动序列,表征虚拟机负载的高频周期变化模式和低频的趋势变化模式。由于邻近的本征模函数具有相似的变化,本文基于主成分分析来提取出特征波动序列的关键性因子,降低分解后特征波动序列的冗余性、相关性和输入维度。卷积层能够有效提取经过 PCA 计算的主成分特征信息,再将特征向量以时序数据的形式输入到双向门控单元。双向门控循环神经网络双向学习序列的前向和后向依赖关系,提高了预测模型学习虚拟机负载长期依赖关系的能力。

### 4.1 基于 EMPC 的云虚拟机负载分解模式

经验模态分解 EMD 可以把非平稳非线性的虚拟机负载序列分解为有限个相对平稳的特征波动序列 IMF,使原始虚拟机负载序列的波动特征在不同尺度下表示出来,从而反映虚拟机负载序列的波动性、周期性和变化趋势,这样能够丰富特征并表示虚拟机负载序列的变化模式<sup>[15]</sup>。EMD 是一种自适应的序列分解处理方法<sup>[16]</sup>,在网络流量预测和短期电力负荷预测等方面得到了一定的应用<sup>[17-18]</sup>。相比传统的小波分解,EMD 分解自适应性强,不需要预先设定基函数。由于 EMD 分解后的特征向量维度过高会使数据的冗余度和复杂度都过大,同时存在噪声的影响,此时引入主成分分析 PCA 方法对数据进行降维和精简。主成分分析的主要思想则是选出少量综合性的新指标来表征原来的数据,主成分的选取由特征值的累计贡献率决定。对于虚拟机负载数据分解后的 IMF 函数,应满足如下两个条件:1)极值点与零点的数目

相差不超过 1;2)上下包络函数的均值为 0。

虚拟机负载序列被分解为从高频到低频变化的 IMF,从高频到低频的 IMF 噪声不断减少,特征变化趋势不断增强,直到最后得到一个单调的残差函数  $r_i(t)$ ,然后对得到的特征进行主成分分析,计算累计贡献率。假设某一虚拟机负载序列为  $v(t)$ ,分解模式如下:

(1)找到  $v(t)$  序列中所有的局部极大值点和极小值点,分别用样条曲线连接极值点,给出极大值包络线  $e_{\max}(t)$  和极小值包络线  $e_{\min}(t)$ 。

(2)求极大值包络线和极小值包络线的平均值,得到平均曲线:  $m(t) = (e_{\max}(t) + e_{\min}(t)) / 2$ 。

(3)用  $v(t)$  序列减去上一步中得到的平均值  $m(t)$ ,得到新的函数:  $h_k(t) = v(t) - m(t)$ , ( $k=1, 2, 3, \dots$ )。

(4)用函数  $h_k(t)$  替换原来的  $v(t)$ ,重复步骤(1)一步骤(3),获得  $h_k(t)$ ,直到  $h_k(t)$  满足 IMF 的两个条件为止,定义 EMD 的第  $i$  阶分量  $imf_i = h_k(t)$ , ( $i=1, 2, 3, \dots$ ),定义 EMD 的  $i$  个残差分量:  $r_i(t) = v(t) - imf_i$ 。

(5)判断  $r_i(t)$  是否为单调序列,如果是单调序列则分解过程结束;如果不是,则令  $v(t) = r_i(t)$ ,并回到步骤(1)。

(6)对得到的 VM 特征矩阵进行标准化处理,得到标准化特征矩阵  $\mathbf{X}$ ,计算  $\mathbf{X}$  的相关系数矩阵  $\mathbf{R}$ ,其中  $cov$  为协方差函数,  $var$  为方差函数。

$$X_{ij} = \frac{v_{ij} - \frac{1}{n} \sum_{i=1}^n v_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (v_{ij} - \frac{1}{n} \sum_{i=1}^n v_{ij})^2}} \quad (2)$$

$$R_{ij} = \frac{cov(x_i, x_j)}{\sqrt{var(x_i)} \sqrt{var(x_j)}} \quad (3)$$

(7)设相关系数矩阵  $\mathbf{R}$  的特征值为  $\lambda_k$  ( $k=1, 2, 3, \dots$ ),  $\mathbf{R}$  的特征值所对应特征向量矩阵为  $\mathbf{I}$ 。可得出式(4)所示的特征方程,最终根据  $\lambda_k$  信息量的累计贡献率来确定主成分。

$$|\mathbf{R} - \lambda \mathbf{I}| = 0 \quad (4)$$

### 4.2 云虚拟机负载预测模型

循环神经网络(RNN)神经元的输出可以在下一个时间状态直接作用到自身,比较适合处理时间序列数据。由于 RNN 存在梯度消失问题,因此不能学习时间序列的长期依赖。长短期记忆网络(LSTM)通过引入输入门、遗忘门和输出门的结构,能够学习时间序列的长期依赖,有效缓解了梯度消失问题,但也增加了模型的复杂程度。近些年提出的门控循环单元(GRU)是 LSTM 的一种变体,主要由更新门和重置门构成。相比 LSTM,GRU 具有更少的门控结构,在减少训练参数的同时缓解了 LSTM 收敛慢的问题<sup>[10]</sup>。虚拟机负载表现出了时序数据的变化特点,不仅与历史负载相关联,同时还与未来时期的负载影响因素有关。GRU 只能前向考虑历史时期的负载影响,无法体现虚拟机负载随时间变化的双向相关性<sup>[19]</sup>,不能充分提取经过 EMPC 负载分解模式得到的多维特征波动序列。因此,本文在 EMPC 的基础上集成 BCGRU 作为预测模型,BCGRU 能够提取虚拟机负载的多维特征波动序列,同时学习虚拟机负载的前向和后向时序依赖信息。每个门控单元在时间  $t$  的隐藏状态  $h_t$ 、重置门  $r_t$  和更新门  $z_t$

的计算式如式(5)所示。其中,  $\tilde{h}_t$  表示在时间  $t$  的候选隐藏状态,  $h_{t-1}$  为  $t-1$  时刻的隐藏层输出,  $\mathbf{W}_{xr}$  对应输入层和重置门之间的参数矩阵,  $\mathbf{W}_{hr}$  对应隐藏层和重置门之间的参数矩阵,  $\mathbf{W}_{xz}$  对应输入层和更新门之间的参数矩阵,  $\mathbf{W}_{hz}$  对应隐藏层和更新门之间的参数矩阵,  $b$  为相应的偏置值。

$$\begin{cases} r_t = \sigma(x_t \mathbf{W}_{xr} + h_{t-1} \mathbf{W}_{hr} + b_r) \\ z_t = \sigma(x_t \mathbf{W}_{xz} + h_{t-1} \mathbf{W}_{hz} + b_z) \\ \tilde{h}_t = \tanh(x_t \mathbf{W}_{zh} + (r_t \otimes h_{t-1}) \mathbf{W}_{hh} + b_h) \\ h_t = z_t \otimes \tilde{h}_t + (1 - z_t) \otimes h_{t-1} \end{cases} \quad (5)$$

为了提升神经网络的非线性建模能力,使用  $\sigma(x)$  和  $\tanh(x)$  作为激活函数,计算式如下:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (6)$$

$$\tanh(x) = 1 - \frac{2}{1 + e^{2x}} \quad (7)$$

在预测当前虚拟机的负载值时,当前值与其最近的历史值相关,也与下一时刻的影响因素相联系,表现出了局部相关性。BCGRU 预测模型在卷积层提取特征的基础上,可以同时学习前向依赖和后向依赖,其状态信息的流动如图 4 所示。当前隐藏层状态  $h_t$  由当前的输入  $x_t$  和上一时刻前向隐藏层状态  $\tilde{h}_{t-1}$  以及后向隐藏层状态  $\tilde{h}_{t-1}$  共同决定,计算过程如下:

$$\tilde{h}_t = F(x_t, \tilde{h}_{t-1}) \quad (8)$$

$$\tilde{h}_t = B(x_t, \tilde{h}_{t-1}) \quad (9)$$

$$h_t = \alpha \tilde{h}_t + \beta \tilde{h}_t + b_t \quad (10)$$

其中,  $F(x)$  和  $B(x)$  表示利用前向和后向门控单元对虚拟机负载数据进行的非线性变化,  $\alpha$  和  $\beta$  分别对应前向隐藏层状态  $\tilde{h}_t$  和后向隐藏层状态  $\tilde{h}_t$  的权重,  $b_t$  是隐藏层在  $t$  时刻的状态偏置值。

## 5 方法评估

为了评估本文的方法,使用微软 Azure 2019 虚拟机公开数据集进行实验效果评估。根据 3.1 节的虚拟机资源使用情况进行分析,过滤一个月长期运行的虚拟机作为实验数据。实验提取一个月前 23 天的数据作为训练集,提取第 24 天到第 26 天的数据作为验证集,提取第 27 天到最后的数据为测试集。实验部分本文采用了 Keras2.1.5, Tensorflow1.15.0 深度学习框架,使用 Python3.7 进行编程,IDE 集成开发环境为 PyCharm。硬件环境为 6 核 CPU(Core i5-10400F @ 2.90 GHz),内存容量为 16 GB(威刚 DDR4 2666 MHz)和 Nvidia GeForce GTX 1050Ti(4GB)。将本文提出的虚拟机 CPU 负载预测方法与最近提出的虚拟机 CPU 负载预测方法 ARIMA<sup>[4]</sup>, Bi-LSTM<sup>[12]</sup>, BC-LSTM<sup>[20]</sup> 以及经典的循环神经网络模型 GRU 进行对比实验,使用 5.1 节所述的 4 个指标客观评估本文方法。

### 5.1 评估指标

为了从多角度评估实验效果,本文使用回归任务评估指标,即均方根误差(Root Mean Square Error, RMSE)、平均绝对误差(Mean Absolute Error, MAE)、平均绝对百分比误差

(Mean Absolute Percentage Error, MAPE)和拟合优度( $R^2$ ),来评估实验方法的性能。RMSE 的主要目标是衡量预测值与真实值之间的偏差,受异常值影响大,更容易凸现出预测值与真实值偏离程度较大的点。由于 Azure 云虚拟机 CPU 负载具有突变的峰值,因此 RMSE 适合评估这些峰值的预测效果,其计算过程如式(11)所示。MAE 能较为准确地展现出预测值与真实值的相对误差,具体计算过程如式(12)所示。MAPE 是 MAE 的变形,是一个百分比值,表示预测误差相对于真实值的偏离程度(见式(13))。 $R^2$  表示回归平方和与总离差平方和的比值,表示总离差平方和中可以由回归平方和解释的比例,这一比例越大,模型越精确,拟合效果就越明显(见式(14))。其中,  $o_i$  表示预测值,  $\bar{o}$  表示预测值的平均值,  $y_i$  表示真实值,  $n$  为样本的数量。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - y_i)^2} \quad (11)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |o_i - y_i| \quad (12)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{o_i - y_i}{y_i} \right| \quad (13)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - o_i)^2}{\sum_{i=1}^n (y_i - \bar{o})^2} \quad (14)$$

### 5.2 实验设置

为了精简数据并降低计算成本,实验首先设置样本矩阵特征值的累计贡献率为 85%。实验所得到的前  $n$  个特征的累计贡献率达到或超过 85%,则认为该组特征的前  $n$  个特征最具代表性,可以代表整个虚拟机负载的特征。图 5 给出了原始虚拟机负载序列和基于 EMD 分解以后得到的本征模函数的高频、中频和低频变化。其中高频的频率较高,其变化模式最接近于原始虚拟机负载序列,低频呈现出单调变化的趋势。

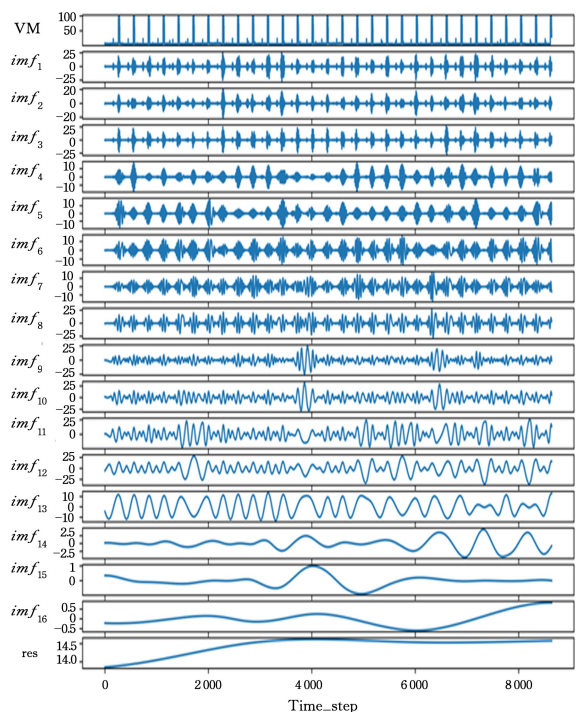


图 5 虚拟机负载序列及相应的 IMF 序列

Fig. 5 VM load sequence and its corresponding IMF sequence

图 6 为特征序列的累计贡献率分布图,在设定累计贡献率上限值为 85% 时,图 5 所示的虚拟机特征序列从 17 维降低到 13 维,进一步消除了噪声的影响和减少了数据维度。不同的参数设置对模型的预测性能有显著的影响,神经元隐藏层的层数一般选择在 1~6 之间,因为隐藏层较少,模型学习参数较少,学习能力较差,隐藏层的不断增加也会导致模型的泛化能力减弱。本文设置隐藏层层数为 2, epochs 设置为 50, batch\_size 的值为 64, dropout 设置为 0.3, 使用 Adam 做为优化器,初始学习率设置为 0.001,历史窗口大小设置为 30。

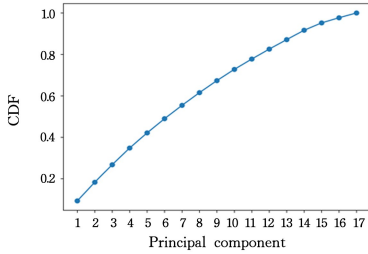


图 6 特征序列的累计贡献率分布图

Fig. 6 CDF of characteristic sequence

### 5.3 实验方法性能对比

文献[4]基于自回归综合移动平均值法 ARIMA 进行了虚拟机平均 CPU 利用率预测实验,ARIMA 是一种应用广泛的时间序列分析模型,结合了移动平均值和自回归方法。Karim 等<sup>[12]</sup>提出了一种新颖的混合循环神经网络预测模型 Bi-LSTM,用于预测云虚拟机 CPU 工作负载。文献[20]采用了一种双向卷积长短期记忆网络模型 BC-LSTM 进行主机负载预测。此外,门控循环神经网络(GRU)作为经典的对比模型,是一种较流行的基于 LSTM 变体的深度学习模型。

如表 2 所列,本文将 ARIMA, GRU, Bi-LSTM, BC-LSTM 以及本文方法进行对比实验,预测虚拟机未来一步最大 CPU 利用率。在 RMSE, MAE 和  $R^2$  这 3 种评估指标中,本文方法均表现出优越的预测性能。本文方法预测的 RMSE 明显小于其他对比方法,与 ARIMA, GRU, Bi-LSTM 和 BC-LSTM 模型相比, RMSE 分别降低了 39.96%, 31.06%, 25.77%, 23.20%; MAE 分别降低了 42.79%, 26.92%, 13.30%, 29.00%。拟合优度( $R^2$ )反映了回归模型的拟合效果,从表 2 可以看出,统计学模型 ARIMA 的非线性拟合能力弱于深度学习模型,另外 3 种深度学习模型  $R^2$  值的差异并不大。图 7 为随机抽的 100 台虚拟机在深度学习模型实验中预测误差 MAPE 的累计分布图,在多台虚拟机的负载预测中,本文预测方法的 CDF 曲线最接近于 1,始终具有较低的预测误差,说明本文方法相比其他预测方法更具有普遍适用性。

表 2 不同方法预测性能对比

Table 2 Comparison of prediction performance of different methods

Model	RMSE	MAE	$R^2$
ARIMA	5.4453	1.7438	0.9448
GRU	4.7425	1.3650	0.9583
Bi-LSTM	4.4043	1.1507	0.9640
BC-LSTM	4.2567	1.4051	0.9664
EMPC-BCGRU	3.2693	0.9976	0.9807

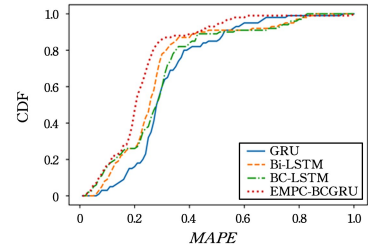


图 7 虚拟机负载预测的 MAPE 比较

Fig. 7 MAPE comparison of VM CPU load prediction

为了进一步评价本文预测方法的性能,本文还在单步预测的基础上进行了未来 3 步(15 min)、6 步(30 min)、9 步(45 min)以及 12 步(60 min)的预测实验。如图 8 所示,随着预测步长的增加,所有方法的预测误差都会逐渐增大,这主要是因为是在微软 Azure 云平台中更长的预测时间步存在较多的不确定因素。但是,在未来 1h 内的多步预测实验结果对比中,所提出的预测模型明显优于其他 4 种对比模型,且随着预测长度的增大这种优越性表现得更加明显,预测性能比较平稳。

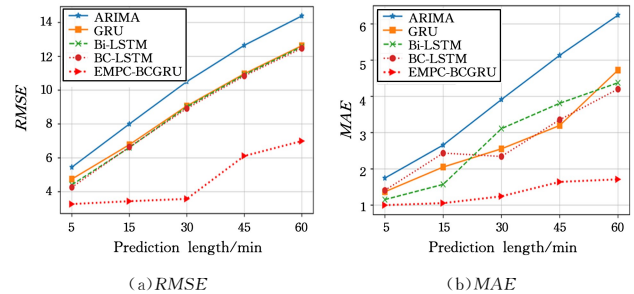


图 8 不同预测步长的 RMSE 和 MAE 比较

Fig. 8 Comparison of RMSE and MAE with different prediction step length

从图 8(a)和图 8(b)可以看出,ARIMA 方法预测的 RMSE 误差值始终保持最大,MAE 误差值也保持相对较高的水平,随着预测步长从 1 步逐渐增大到 12 步,其 RMSE 误差值从 5.4453 增加到了 14.3827, MAE 误差值从 1.7438 增加到了 6.2394。基于深度学习的预测方法 GRU, BC-LSTM 和 Bi-LSTM,其 RMSE 误差差异并不大,其中 BC-LSTM 方法略优于其他两者。本文的 EMPC-BCGRU 方法表现出了较优的预测性能,随着预测步长从 1 步逐渐增加到 12 步, RMSE 误差值从 3.2693 增加到 6.9904, MAE 误差值从 0.9976 增加到 1.7091。即使预测步长增大,本文的 EMPC-BCGRU 方法的预测误差值也保持在较低水平,预测误差增加得较为缓慢。

总的来说,不管是虚拟机 CPU 负载单步预测还是多时间步预测,所提出的基于 EMPC-BCGRU 虚拟机的负载预测方法,相比基准预测方法和最近提出的关于虚拟机负载预测方法,都具有较好的预测性能。这也验证了本文所提方法能够学习到虚拟机 CPU 负载的变化模式和有效提取虚拟机负载序列数据的深层非线性相关特征。ARIMA 模型本质上还是线性模型,难以学习复杂的非线性关系。此外,由于 ARIMA 的非线性学习能力较弱,只能预测较短时间内的 CPU 负载情况,因此需要频繁训练模型。GRU 是简化的 LSTM 模型,比

LSTM 模型少一个门控结构,因此训练的参数较少,训练效率较高。Bi-LSTM 模型可以学习虚拟机负载序列的双向依赖关系,因此预测性能比经典的循环神经网络模型更好,但是特征提取和负载模式学习能力不明显。BC-LSTM 预测方法使用两个独立的隐藏层来学习前向和后向数据的依赖关系,具有一定的特征学习能力,但是难以学习到虚拟机负载局部峰值的变化模式,特征的提取能力仍需要进一步提高。

#### 5.4 消融实验分析

集成多个组件的神经网络模型具有强大的非线性建模能力,通过卷积层对分解后多维特征序列的提取和双向的时序学习,最终提高了预测模型的特征提取能力和对负载变化模式的学习能力。为了进一步分析本文预测方法的有效性,在虚拟机 CPU 负载单步预测的基础上进行了多次消融实验。表 3 详细列出了消融实验的结果。

参与消融实验的模型分别是 BiGRU, BCGRU 和 EM-BCGRU,这些模型依旧以高效的 GRU 作为内核单元,以验证本文所提预测方法各个部分的有效性。结合表 2 和表 3 可以看出, BiGRU 预测方法比 GRU 预测方法的 RMSE 要低 0.3586, MAE 降低了 0.2268。BCGRU 方法在 BiGRU 的基础上加入了卷积层来提取特征,在 BiGRU 的基础上预测误差 RMSE 降低了 0.0459,但是 MAE 增加了 0.7353。可见,单纯融合卷积层对于虚拟机 CPU 负载预测性能的提升并不大,

这主要是因为虚拟机负载在整个生命周期的局部和全局都具有非周期模式的峰值变化,特别是峰值预测的不准确性对于预测结果的 RMSE 会产生较大的影响,卷积层难以充分提取出这些特征。EM-BCGRU 模型在整合了前者优势的基础上,进行了虚拟机负载模式分解,将虚拟机负载分解为不同尺度的本征模函数,这些本征模函数能更加细节性地表现特征波动,丰富了特征,有利于模型对虚拟机负载变化模式和特征的学习。由表 3 可知, EM-BCGRU 模型在 BCGRU 模型的基础上,预测误差 RMSE 减少了 0.9444,预测误差 MAE 减少了 0.5979。EMPC-BCGRU 模型表现出了最好的预测效果,在 EM-BCGRU 的基础上,通过加入主成分分析降低分解后特征波动序列的冗余性、相关性和输入维度,提高了预测模型对虚拟机负载特征的学习效率。从图 9 可以看到,不管是较大的峰值段还是较小的峰值段,本文方法都能保持最优的预测性能,预测值曲线与真实值曲线都能较好地匹配。

表 3 预测模型各部分组件消融实验结果

Table 3 Ablation experimental results of model components

Model	RMSE	MAE	R <sup>2</sup>
BiGRU	4.3839	1.1382	0.9643
BCGRU	4.3380	1.8735	0.9651
EM-BCGRU	3.3936	1.2756	0.9792
EMPC-BCGRU	3.2693	0.9976	0.9807

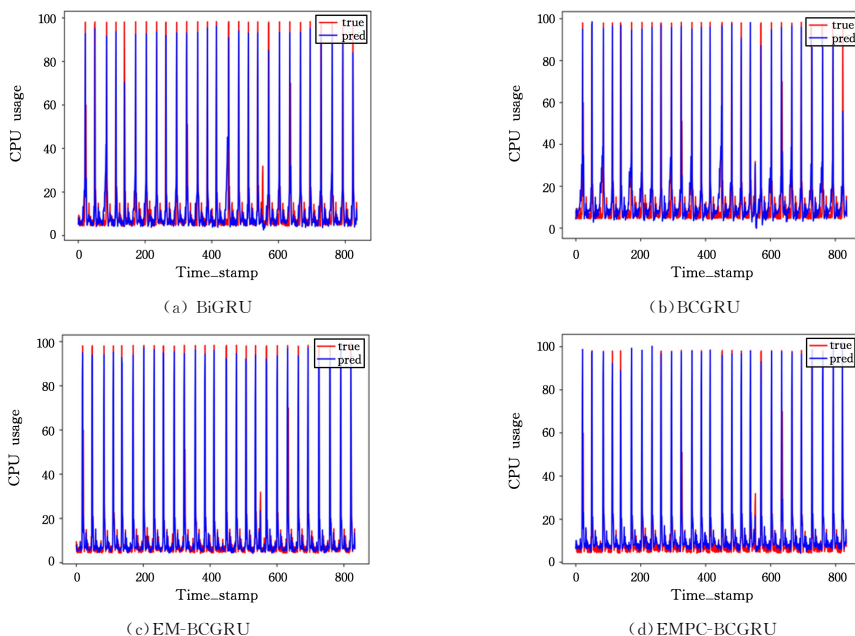


图 9 预测值与真实值的对比

Fig. 9 Comparison of predicted values and true values

**结束语** 本文对云平台虚拟机资源使用情况和负载行为进行了分析,针对云平台虚拟机资源利用率低和负载变化模式的特点,提出了一种基于 EMPC-BCGRU 的云虚拟机 CPU 负载预测方法。由于云虚拟机 CPU 负载具有非线性和峰值多变性等特点,本文提出了基于经验模态分解和主成分分析的虚拟机负载分解模式,通过经验模态分解对虚拟机负载序列进行多尺度分解,然后基于主成分分析提取本征模函数的高频、中频和低频的关键因子,并将其作为预测模型的输入。

EMPC-BCGRU 模型由经验模态分解、主成分分析、卷积层和双向的内核单元 GRU 集成,能够充分提取模式分解后的特征,并学习虚拟机负载序列的长期依赖关系。在对比实验中,本文提出的虚拟机 CPU 负载预测方法相比经典的预测模型和最新提出的虚拟机负载预测模型,都具有更优的预测性能,且在多步预测实验中,预测性能表现得较为稳定。

本文只对长期运行的虚拟机进行 CPU 负载预测,在微软 Azure 中,同一部署中的多台虚拟机为了完成用户的任务

可能协同工作,这就使得多台虚拟机工作负载具有相似性<sup>[1]</sup>。本文在训练模型和预测虚拟机负载时并没有考虑这些协同工作的虚拟机,多源数据的探索仍有待进一步研究。对于接下来的工作,可以基于虚拟机负载的未来预测值对云平台虚拟机进行调度和整合,这将有利于节约资源,提高云平台的资源利用率。

## 参 考 文 献

- [1] CORTEZ E, BONDE A, MUZIO A, et al. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms[C]// Proceedings of the 26th Symposium on Operating Systems Principles. 2017: 153-167.
- [2] FAN Z W, HUANG P J, HUANG P S, et al. A feature generation framework for google trace analysis[C]// Proceedings of the 2015 International Conference on Machine Learning and Cybernetics(ICMLC). IEEE, 2015, 1: 229-234.
- [3] LIAN J D, LIU H L, XIE H B, et al. Hierarchical load balancing algorithm based on prediction mechanism [J]. Computer Engineering and Applications, 2015, 51(11): 67-71, 98.
- [4] ABUBAKAR A, BARBHUIYA S, KILPATRICK P, et al. Fast analysis and prediction in large scale virtual machines resource utilisation[C]// Proceedings of the International Conference on Cloud Computing and Services Science(CLOSER). 2020: 115-126.
- [5] CHEN S, SHEN Y, ZHU Y. Modeling conceptual characteristics of virtual machines for CPU utilization prediction[C]// Proceedings of the International Conference on Conceptual Modeling. Cham: Springer, 2018: 319-333.
- [6] WANG J, WANG Z, LI J, et al. Multilevel wavelet decomposition network for interpretable time series analysis[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2437-2446.
- [7] DI S, KONDO D, CIRNE W. Google hostload prediction based on bayesian model with optimized feature combination[J]. Journal of Parallel and Distributed Computing, 2014, 74(1): 1820-1832.
- [8] QIAN S P, YU Y, ZHAI T Y, et al. Research on online prediction model of host load based on deep learning[J]. Computer Engineering, 2021, 47(9): 84-89.
- [9] SONG B, YU Y, ZHOU Y, et al. Host load prediction with long short-term memory in cloud computing[J]. The Journal of Supercomputing, 2018, 74(12): 6554-6568.
- [10] CHEN Z, HU J, MIN G, et al. Towards accurate prediction for high-dimensional and highly-variable cloud workloads with deep learning[J]. IEEE Transactions on Parallel and Distributed Systems, 2019, 31(4): 923-934.
- [11] KHODAVERDIAN Z, SADR H, EDALATPANAH S A, et al. Combination of convolutional neural network and gated recurrent unit for energy aware resource allocation[J]. arXiv: 2106.12178, 2021.
- [12] 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.
- [13] SHUVO M N H, MASWOOD M M S, ALHARBI A G. LsrU: A novel deep learning based hybrid method to predict the workload of virtual machines in cloud data center[C]// 2020 IEEE Region 10 Symposium(TENSYMP). IEEE, 2020: 1604-1607.
- [14] KUMBHARE A G, AZIMI R, MANOUSAKIS I, et al. {Prediction-Based} Power oversubscription in cloud platforms[C]// 2021 USENIX Annual Technical Conference(USENIX ATC 21). 2021: 473-487.
- [15] ALSALAH A, HOLLOWAY D, MOUSAVI M, et al. Identification of wave impacts and separation of responses using EMD [J]. Mechanical Systems and Signal Processing, 2021, 151: 107385.
- [16] LUO Y X, LI Z H, LIANG X, et al. Multifractal detrended fluctuation analysis method for nonstationary time series based on EMD-LS[J]. Acta Electronic Sinica, 2021, 49(12): 2323-2329.
- [17] YAO L S, LIU D, PEI Z F, et al. Real-time network traffic prediction model based on EMD clustering [J]. Computer Science, 2020, 47(S2): 316-320.
- [18] DENG A, JIN M. Time-scale feature extraction method based on EMD and its application in short-term power load forecasting [J]. Application Research of Computers, 2018, 35(10): 2952-2955.
- [19] LIANG R, CHANG X, JIA P, et al. Mine gas concentration forecasting model based on an optimized BiGRU network[J]. ACS Omega, 2020, 5(44): 28579-28586.
- [20] ZHU J C, DENG L, YAN M, et al. Research on cloud platform host resource load forecast analysis[J]. Journal of Chinese Computer Systems, 2021, 42(12): 2538-2544.



**XIE Tonglei**, born in 1997, postgraduate. His main research interests include cloud computing, data analysis and deep learning.



**DENG Li**, born in 1972, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. Her main research interests include cloud computing and distributed computing.