2025 年 第 25 卷 第 7 期
2025, 25(7):02856-09
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2307597
引用格式:陈基漓, 李海军, 谢晓兰. 基于时间卷积和长短期记忆网络的短期云资源预测模型[J]. 科学技术与工程, 2025, 25(7): 2856-2864.
Chen Jili, Li Haijun, Xie Xiaolan. Short-term cloud resource prediction model based on temporal convolution and long short-term memory [J]. Science Technology and Engineering, 2025, 25(7): 2856-2864.

基于时间卷积和长短期记忆网络的短期云资源预测模型

陈基漓[1,2], 李海军[1,2], 谢晓兰[1,2*]
(1. 桂林理工大学计算机科学与工程学院, 桂林 541004; 2. 广西嵌入式技术与智能系统重点实验室, 桂林 541004)

摘 要 随着容器云技术的不断深入发展,通过预测分析云资源请求的整体趋势及高峰期,对于容器云资源的高效利用和合理分配具有重要意义。利用深度学习技术进行负载预测已经成为解决容器云资源利用率不平衡的关键技术。针对目前负载预测的单一模型和组合模型所存在的预测精度低以及捕获序列特征不充分问题,提出基于时间卷积和长短期记忆网络(temporal convolutional network-long short-term memory,TCN-LSTM)的短期云资源组合预测模型,组合模型中的空洞卷积在不减少特征尺寸的情况下增加感受野获取更长久的时间序列特征,其中残差网络可以跨层传递信息以加快网络的收敛,所获取的时间序列特征可有效提高 LSTM 的预测精度。利用阿里巴巴公开数据集的进行预测,实验表明所提出的模型与单一的预测模型以及其他组合模型进行对比分析,误差指标-平均绝对误差(mean absolute error,MAE)降低 8% ~13.7%,均方根误差(root mean squared error,RMSE)降低 9.8% ~13.1%,证明所提模型的有效性。

关键词 容器云; 云资源预测; 时间卷积网络(TCN); 长短期记忆网络(LSTM)
中图法分类号 TP393; **文献标志码** A

Short-term Cloud Resource Prediction Model Based on Temporal Convolution and Long Short-term Memory

CHEN Ji-li[1,2], LI Hai-jun[1,2], XIE Xiao-lan[1,2*]
(1. College of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China;
2. Guangxi Key Laboratory of Embedded Technology and Intelligent Systems, Guilin 541004, China)

[**Abstract**] With the continuous development of container cloud technology, it is of great significance to predict and analyze the overall trend and peak of cloud resource requests for efficient utilization and reasonable allocation of container cloud resources. Deep learning technology for load prediction has become a key technology to solve the unbalanced utilization of container cloud resources. Aiming at the problems of low prediction accuracy and insufficient capture sequence features existing in the current single model and combination model of load prediction, a cloud resource combination prediction model based on temporal convolutional network-long short-term memory(TCN-LSTM) was proposed. The hollow convolution in the combination model increased the sensitivity field without reducing the feature size to obtain longer time series features. The residual network could transfer information across layers to accelerate the convergence of the network, and the obtained time series features could effectively improve the prediction accuracy of LSTM. Useing Alibaba's publicly available dataset to make predictions, the experiment shows that the proposed model is compared with the single prediction model and other combined models, and the error index-mean absolute error(MAE) is reduced by 8% ~13.7% and root mean squared error (RMSE) by 9.8% ~13.1%, which proves the effectiveness of the proposed model.

[**Keywords**] container cloud; cloud resource prediction; temporal convolutional network; long short-term memory network

容器是一种新的轻量级虚拟化技术,不需要虚拟机监控,与传统云计算相比,容器云开销少,启动时间短的优点[1],但对于资源的管理存在一定的问题,会出现供不应求和过度供应的情况。目前各大厂商致力于为用户提供优质的服务,合理对容器内的资源进行分配和部署可以提升用户的体验。云

收稿日期:2023-09-26 **修订日期**:2024-07-09
基金项目: 国家自然科学基金(62262011);广西自然科学基金(2021JJA170130)
第一作者: 陈基漓(1972—),女,瑶族,广西玉林人,硕士,副教授。研究方向:智能计算及数据挖掘。E-mail:345062001@qq.com。
* **通信作者**: 谢晓兰(1974—),女,汉族,广西桂林人,博士,教授。研究方向:云计算及大数据。E-mail:237290696@qq.com。

计算是信息技术发展的产物,它允许远程访问计算资源,允许多台计算机协作和共享资源,这些资源可以根据工作负载进行动态配置,从而实现资源的合理分配以及利用。目前容器云的兴起使得云计算发展更加快速,使得用户能更加容易和快速获取所需的资源。云计算资源负载预测是通过解析机器的历史负载信息,寻找过去与未来的云计算资源负载之间的关系,预测未来短期时间内的云资源负载,为厂商云计算资源的动态分配、云计算平台的机器部署以及降低运维人力成本提供重要依据。

近年来,国内外的学者对如何提高容器云资源负载预测的准确性和稳定性进行了探索和研究,研究方法主要分为传统模型和深度学习模型,传统的预测模型主要包含自回归积分滑动平均模型(autoregressive integrated moving average mode, ARIMA)、灰色模型以及改进支持向量机(support vector machine, SVM)为主,深度学习预测模型主要以长短期记忆神经网络(long short-term memory, LSTM)或门控循环单元(gate recurrent unit, GRU)等神经网络进行预测。

文献[2]提出基于马尔科夫链模型的云资源状态监控预测机制,尽管该方法在某些情况下表现良好,但在面对大规模数据时,马尔科夫链模型可能无法充分捕捉各种状态之间的转换的关系,可能会面临预测结果误差较大的问题。可在数据预处理阶段采取有效的噪声处理和异常值检测技术,以减少数据中的干扰和误导,提高模型的稳定性。灰色模型在某些情况下非常适合数据量较小、短期的时序预测任务,然而,灰色模型也有局限性,灰色模型通常要求数据是等间隔的,而且需要足够的历史数据点来进行模型的构建,当数据非常有限和不规则时,难以使用灰色模型进行可靠的预测。赵莉[3]采用混沌分析算法对资源负载的时序数据进行分析和重构,构建多维时间序列,设计组合核函数克服当前核函数存在的弊端,提升 SVM 的学习能力,有效提升单步或多步的负载预测精度。传统模型相对于深度学习模型在对时间序列数据进行预测时存在一定的局限性,当数据没有季节性规律以及数据不够平滑时,模型的预测精度较低,而单一的深度学习模型虽然预测精度有一定的提升,但是对于时序数据的变化感知不够准确,不能很好地捕捉时序数据中的关键特征,有学者研究表明,组合模型可以获取充分特征的同时而又能保持较高的预测精度。文献[4-5]提出在神经网络中加入注意力机制用来捕获时间序列的关键时空特征,利用基于注意力的解码器挖掘时间依赖关系,加入注意力机制可有效提升模型的精度和泛化能力。Ren 等[6]将粒子群算法(particle swarm optimization, PSO)与长短期记忆网络相融合,粒子群算法被用于搜索参数空间中的最佳组合,以最小化或最大化预定义的目标函数,粒子群算法将 LSTM 模型的隐层神经元数量、学习率和迭代次数作为优化问题种的变量,通过在参数空间中搜索,找到使目标函数最优的参数组合,有效地克服了传统的人工确定参数的限制,并在减少工作负担的同时,提高了模型的性能。Bai 等[7]提出了时间卷积网络(temporal convolutional network,TCN),并在多种类型的数据上进行了实验验证,研究结果表明,TCN 相对于传统的循环神经网络(recurrent neural network,RNN)架构在不同的序列建模任务中表现更为有效。TCN 模型具有更好的清晰度和简单性,对于时间序列数据的特征提取和捕捉以及时间序列数据的长期依赖性方面具有一定的优势,这些优势从一定程度上提高了序列预测的准确性。有部分研究者从数据的特征方面进行研究,为获得更为有效的特征,王悦悦等[8]的研究工作对深度学习在负载预测领域的应用进行了重要的探索,它的方法结合了卷积神经网络(convolutional neural network, CNN)和长短期记忆网络,并采用了自适应的学习率策略,以提高网络的收敛速度,并避免梯度下降时陷入局部最优点。此外,通过 CNN 捕获特征供 LSTM 学习,加快了模型的收敛速度并提高了负载预测的精度。贺小伟等[9]的研究工作提出了一种创新的组合预测模型,即 GRU-LSTM 模型,将 GRU 和 LSTM 这两种不同类型的循环神经网络结合到一个模型中,充分利用他们各自的优点,GRU 通常具有更少的参数和计算复杂性,而 LSTM 能够更好地捕获长期依赖性,研究者证明该模型相较于单一模型,它能够更准确地捕捉和预测时间序列数据中的趋势。李新飞等[10]提出的 Hot-Winters-LSTM 预测模型是在时间序列预测领域的一项有价值的研究工作,该模型引入了一种创新的残差变异系数赋权方法,用于度量模型预测残差的离散程度指标,以此来避免主观性赋权,使得模型能够客观地分配权重,优化模型的误差,提高了模型的精度和稳定性。王琛等[11]从一个新的角度出一种基于 ResNet-LSTM 的多特征预测模型,该模型采用多层的 ResNet 作为特征提取单元,有助于从多元负荷数据中提取丰富的特征信息,这种多特征融合的策略有助于全面捕捉负荷数据中的空间耦合和交互特征。引入注意力机制,提高对共享特征的关注程度,从而更好地适应多元负荷的联合预测。文献[12-13]将 CNN 与 LSTM 或 GRU 序列预

测网络相结合,CNN 负责提取短期特征 LSTM 或 GRU 提取长期特征,注意力机制可以调整神经元的权值,增加神经元之间的联系,提升模型的准确性和稳定性。

针对单一模型和其他组合模型的特点与不足,提出一种基于 TCN-LSTM 的组合预测模型,通过 TCN 的空洞卷积在不牺牲输入尺寸的同时获取更多有效的时间序列特征,并加入残差连接加快网络的收敛,再依靠 LSTM 较强的非线性拟合能力,可以有效提升模型对于负载预测的精确度和稳定性。

1 负载预测模型

1.1 预测原理

机器学习模型通常是用来建立输入特征和输出目标之间的映射关系,对于时间序列的预测任务,数据的重构和特征工程是非常重要的步骤之一。这些步骤的目标是将原始数据整理成适合机器学习算法的形式,以获得一系列标准的线性和非线性机器学习模型所需的输入特征[14]。滑动窗口是时间序列预测中常用的一种技术,将时间序列数据重构为一个机器学习问题。通过以前一个时间步的值作为输入,来预测下一个时间步的值,从而将时间序列预测问题转换为监督学习问题。将所获得的时间序列原始数据重构为标准的机器学习数据集,通常使用滑动窗口法对数据进行预处理,窗口大小作为一个变量参数,通常称为时间步长。

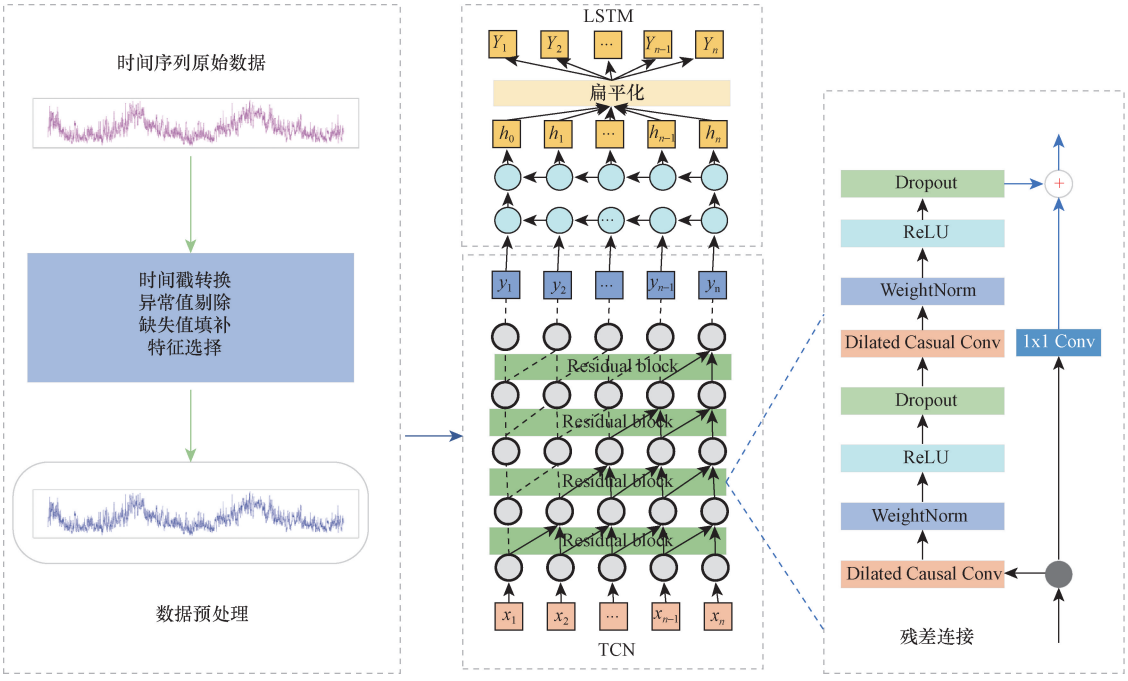
1.2 TCN-LSTM 组合模型

TCN 通过一维的因果卷积对历史序列进行特征提取,确保未来的信息不会泄露以保证预测的准确性,扩张卷积增大卷积核的感受野从而能够有效地捕捉数据的长期依赖性,TCN 结构允许卷积层并行计算,加快计算速度。LSTM 网络作为 RNN 的变种,拥有非常出色的非线性拟合能力,在较大的数据集上其预测精度较同类型的 GRU 更高一点。本文中将 TCN 与 LSTM 结合形成 TCN-LSTM 组合预测模型,模型结构图如图 1 所示。

处理过的数据进入 TCN 在获取到足够的时间序列特征后,数据的时间序列长度由于扩张卷积的填充会增加,通过裁剪维持数据的输出与输入长度一致,所得到的输出相比原始序列的特征更为有效,TCN 的输出作为 LSTM 的输入,提高模型的性能,降低数据处理的难度,同时减少整体模型的复杂度,TCN 和 LSTM 各自具有独特的优点,可以互补地提高时间序列预测模型地准确性和效率。

1.3 时间卷积网络

时间卷积网络的设计使其能够更好地应对时间序列任务中地长期依赖性,同时保持了因果性,确保对过去信息的有效建模,残差连接机制可以加速模型的训练收敛速度,这些特点从使得 TCN 在时间序列预测和其他序列建模任务中表现出色,并成为深度学习领域的重要工具之一。



x 表示输入值;h 表示上一层输出;y、Y 表示输出值

图 1 TCN-LSTM 模型结构图

Fig. 1 The model structure of TCN-LSTM

1.3.1　因果卷积

因果卷积是一种卷积操作,其目标是确保卷积操作不会泄露未来信息,以维持时间序列的因果性。因为在时间序列预测中,模型只能依赖于过去的信息来进行预测,而不能使用未来的信息。通过将因果卷积与一维全卷积网络(fully convolutional networks,FCN)架构结合使用,TCN 能够在保持因果性的同时,有效地进行特征提取和建模。这种结构允许模型在不增加序列长度的情况下获取更多的信息,有助于提高时序预测的准确性。可以将问题转化为:根据 $x_1,x_2,\cdots,x_t$ 去预测 $y_1,y_2,\cdots,y_t$。因果卷积的定义为,滤波器 $F=(f_1,f_2,\cdots,f_K)$,序列 $X=(x_1,x_2,\cdots,x_T)$,在 x_t 处的因果卷积表达式为

$$(FX)_{(x_t)}=\sum_{k=1}^{K}f_k x_{t-K+k} \tag{1}$$

因果卷积原理如图 2 所示,它是一种单向结构,对于上一层 t 时刻的值,只依赖于下一层 t 时刻及其之前的值,信息在网络中只能沿着一个方向传递,从过去到未来,是一种严格约束的时间模型。

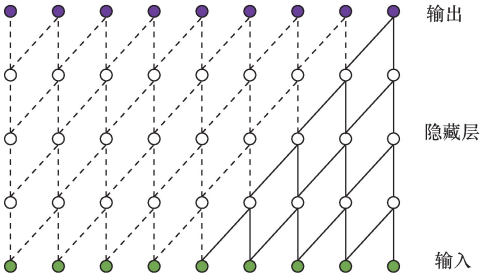


图 2　因果卷积结构图

Fig. 2　The structure of causal convolution

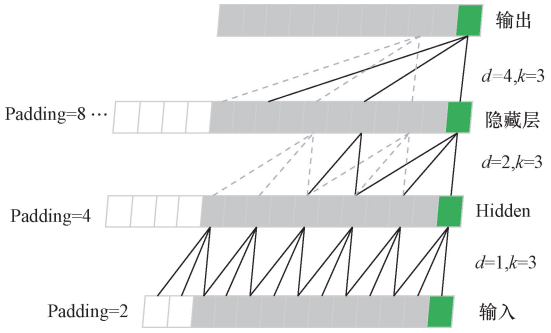
1.3.2　扩张卷积

传统卷积神经网络在处理时间序列时,卷积核的大小限制网络的感受野,如果要获取更长的依赖关系,通常需要堆叠多个卷积层,导致网络变得很深,增加训练难度,容易发生梯度消失和梯度爆炸问题。扩张卷积(dilated convolution)通过引入扩张因子来解决这一问题,扩张因子决定了卷积核内部元素之间的间隔,通过增加扩张因子,卷积核能够在不牺牲特征尺寸的情况下增加感受野,它可以跳过一些时间特征直接访问更早之前的特征。每一层对上一层的信息进行跳跃式提取,有助于网络捕获长期的时间依赖性,扩展卷积原理如图 3 所示。

1.3.3　残差连接

TCN 是结合扩张卷积和残差连接的用于序列建模的神经网络模型,由两层扩张卷积和残差连接方式构成,其残差结构如图 4 所示。

残差连接被证明是训练深层网络的有效方法之一,残差连接允许信息在网络的不同层之间以跨层的方式进行传递,将每个残差块的输入直接添加到块的输出中来实现的。这种机制改善了梯度传播和训练深层网络的稳定性。一个残差块通常包含两个卷积层和一个非线性映射,卷积层用于特征提取,而非线性映射引入非线性性质,增强网络的表示能力。TCN 中引入 WeightNorm 和 Dropout 来正则化网络,WeightNorm 有助于稳定网络训练,而 Dropout 可以减轻拟合问题,这些正则化可以改善模型的泛化性能。残差连接的应用有助于加速网络的收敛素的,通常需要较少的训练迭代就可以达到较好的性能。



d 代表扩张率;k 代表卷积核大小

图 3　扩张卷积结构图

Fig. 3　The structure of dilated convolution

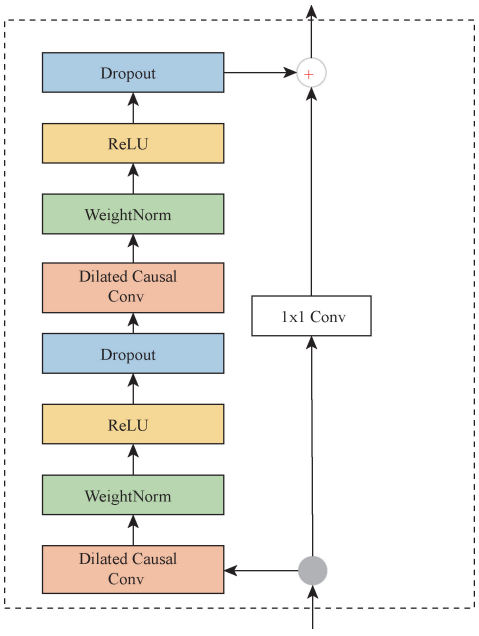


图 4　残差结构图

Fig. 4　The structure of residual network

1.4　长短期记忆网络

长短期记忆网络是一种特殊类型的循环神经网络,它通过引入门控单元来解决了传统 RNN 中的长距离依赖问题。LSTM 网络通常由 4 个主要部分组成:输入层、LSTM 层、全连接层和输出层组。其结构如图 5 所示。

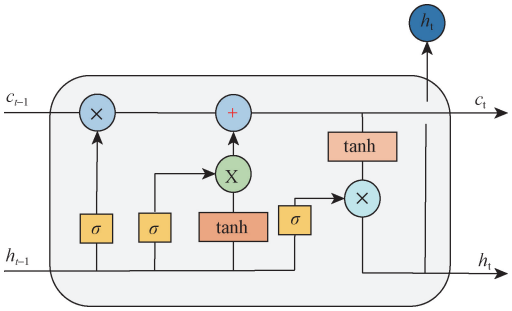


图 5 LSTM 结构图

Fig. 5 The structure of LSTM

在 LSTM 网络中,每个 LSTM 单元针对输入进行函数的计算,公式为

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \quad (2)$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \quad (3)$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \quad (4)$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \quad (5)$$

$$c_t = f_t c_{t-1} + i_t g_t \quad (6)$$

$$h_t = o_t \tanh(c_t) \quad (7)$$

式中:h_t 为 t 时刻的隐藏状态;c_t 为 t 时刻的元组状态;x_t 为 t 时刻的输入;h_{t-1} 为 $t-1$ 时刻的隐藏状态,初始时刻的隐藏状态为 0;i_t、f_t、g_t、o_t 分别为输入门、遗忘门、选择门和输出门;σ 表示激活函数。LSTM 在信息处理上分为 3 个阶段。

(1)遗忘阶段。这个主要针对上一个节点传进来的输入进行选择性忘记,决定哪些旧的信息应该被忘记,使得单元状态保持最新的、最重要的信息。即通过 f_t 的值来控制上一状态 c_{t-1} 中哪些需要记住,哪些需要遗忘,有助于模型处理长期依赖性,减少信息累计造成的误差。

(2)选择记忆阶段。这个阶段将输入 x_t 有选择性地进行“记忆”。当前单元的输入内容是计算得到的 i_t,可以通过 g_t 对其进行有选择地输出。

(3)输出阶段。这个阶段将决定哪些会被当作当前的状态输出。主要通过 o_t 进行控制,并且要对 c_t 使用 tanh 激活函数进行缩放。

2 实验结果与分析

2.1 实验环境

本实验使用的计算机配置为 Windows10 系统,CPU 为 AMD R5-5600,32 G 运行内存,GPU 为 RTX3070,磁盘空间 1 TB;编程环境为基于 python3.8 的 PyCharm 以及基于 PyTorch2.0 搭建的深度学习框架。

2.2 数据集来源及参数

数据集选取来自阿里巴巴公司的 2018 年的公开集群数据集合(cluster-trace-2018),这种数据通常是在真实生产环境中采集的,因此具有很高的真实性和代表性。数据集涵盖了 4 000 台服务器上的在线应用容器和离线计算任务的运行情况,数据集覆盖了长达 8 d 的时间段,这对于分析资源利用情况的趋势和周期性变化非常重要。由于 CPU 是服务器上最关键的资源之一,它的负载最能体现云资源的使用情况,对于应用性能和资源分配至关重要。本次实验采用的数据是其中多台机器 8 d 的 CPU 利用率,特征工程利用随机森林选出具有代表性的内存利用率,提升预测的准确性,根据 CPU 利用率和内存利用率作为输入特征对 CPU 负载进行预测,数据集参数如表 1 所示。

2.2.1 实验过程

该实验是一个单步预测的实验,实验流程图如图 6 所示。

表 1 数据集参数

Table 1 Dataset parameters

列名	说明
machine_id	机器唯一的 id
time_stamp	时间戳
cpu_util_percent	CPU 利用率
mem_util_percent	内存利用率
mem_gps	内存带宽使用率
net_in	传入网络包的数量
net_out	传出网络包的数量
disk_usage_percent	磁盘空间利用率

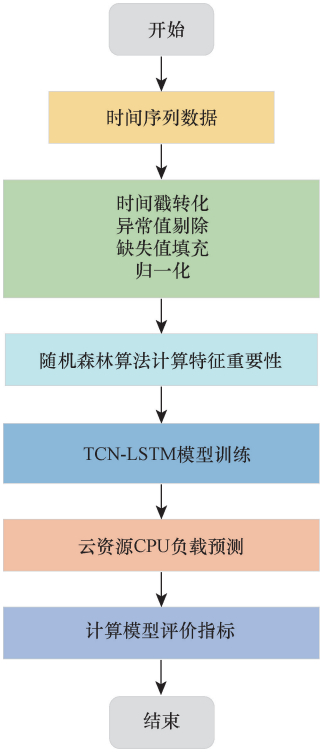


图 6 算法流程图

Fig. 6 Flow chart of the algorithm

从图6中可以发现,获取数据后首先需要针对原始数据进行预处理,转化时间戳可以将其标准化为模型可以处理的形式,将时间戳转化为时期对象。异常值可能是数据中的噪声,会对模型产生负面影响,使用箱型图可以识别和移除异常值。使用插补法来填充部分的缺失值和标准化数据来降低数据之间差异过大造成的影响。再通过随机森林算法选择重要的特征,识别和移除不相关或者冗余的特征,以确保模型不会受到多余特征的干扰。

2.2.2　数据预处理

由于原始数据集时间的数据格式采用时间戳,需要将其转换为标准时间格式,发现原始数据集的采样是每隔1~60 s随机采样,时序预测要保证数据均匀间隔采样才能保证实验的准确性和有效性,将数据处理为等间隔采样,对于缺失值采用相邻值的均值进行填充,并对训练数据集进行Min-Max归一化处理,归一化后的数据加快梯度下降求解的速度。公式为

$$X'_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{8}$$

式(8)中:X'_i为归一化后i时刻的值;X_i为原数据集i时刻的值;$X_{\min}$为原数据集中的最小值;$X_{\max}$为原数据集中的最大值。

负载的观测值通过箱型图来进行体现,其中很少有点被观察到作为当前数据集范围之外的异常值,异常值会影响模型对数据趋势的预测,模型在特征提取的时候若将异常值作为特征则会对模型的精度产生影响,从而导致模型的泛化性降低,如图7所示,此处体现的为部分数据进行异常值剔除。

2.2.3　特征选择

特征选择是一个重要的数据预处理步骤,有助于提升模型的性能和效率,其目的从原始数据的特征集中选择出若干个最具有代表性的特征[15]。随机森林算法是一个强大的特征选择工具,它的重要性得分可以帮助确定哪些特征对于构建高性能的回归模型非常关键。通过对每个特征进行重要性评分来衡量其在模型中的贡献,分数越高的特征被认为是最重要的。随机森林还能够捕获特征之间的相互作用,它不仅可以评估单个特征的重要性,还可以识别多个特征之间的组合对模型性能的影响。因此本文中选择随机森林算法进行特征选择,特征选择的结果如图8所示。

从图8中可以看出,内存利用率和时间戳是得分最高的特征,对目标值CPU利用率影响最大,因此选择内存利用率和时间戳作为特征。

将数据集分为训练集和测试集是一个标准的机器学习实验步骤,通过将数据集的80%用于训练,可以确保模型有足够的数据来学习,并将20%用于测试,用于评估模型的性能。实验使用PyTorch框架来搭建模型,损失函数为MSE,激活函数ReLU,优化器选择Adam,对模型进行调参选取较优的一组参数,TCN-LSTM的部分参数设置如表2所示,对比模型的实验参数如表3所示。

2.3　实验结果与分析

将各个模型的预测值与真实值进行对比,观察各个模型的预测情况,设横坐标为采样点,纵坐标为机器的对应样本点的CPU使用率,为保证实验的稳定性,每个模型重复多次实验取平均值,机器1的模型预测对比图如图9所示。

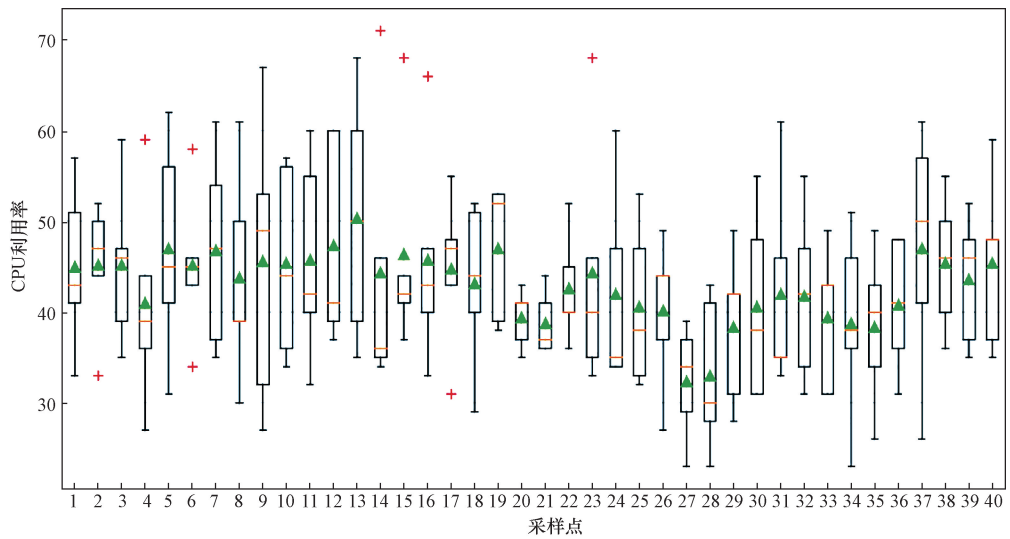


图7　异常值剔除图

Fig. 7　Plot of outlier rejection

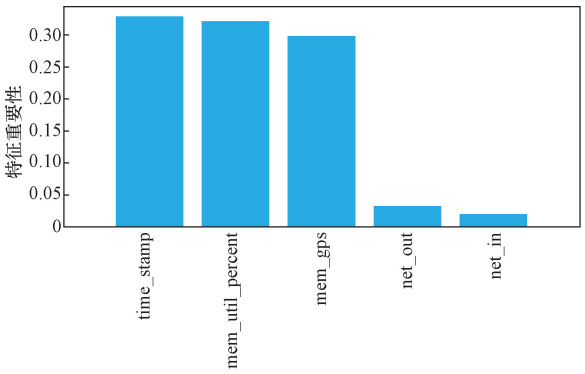


图 8 特征选择结果图

Fig. 8 Plot of feature selection results

表 2 实验参数表

Table 2 Experimental parameters

参数名称	参数取值	参数含义
windows_size	24	滑动窗口
kernel_size	2	卷积核大小
dilations	1,2,4,8	每层空洞因子
batch_size	24	批处理大小
epochs	200	迭代次数
input_size	2	输入特征数
hidden_size	20	隐藏神经元元数
num_layers	5	堆叠层数

表 3 对比模型实验参数表

Table 3 Comparison of model experimental parameters

模型	参数名称	参数取值
LSTM	input_size	2
	hidden_size	24
	num_layers	5
	epochs	100
GRU	input_size	2
	hidden_size	24
	num_layers	5
	epochs	100
TCN	Input_size	2
	num_channels	[10,10,10,10]
	kernel_size	2
	dilations	1,2,4,8
	epochs	50
文献[4]	input_size	2
	kernel_size	3
	hidden_size	64
	num_layers	1
	epochs	100
文献[12]	input_size	2
	kernel_size	3
	hidden_size	64
	num_layers	1
	epochs	100

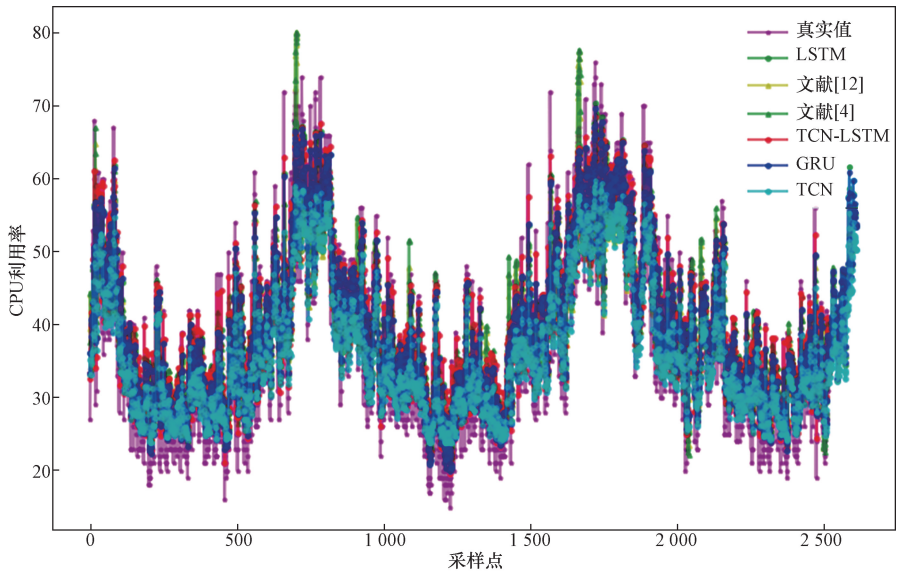


图 9 机器 1 模型的预测值与真实值对比图

Fig. 9 The predicted value of the machine 1 model compared to the true value

为验证模型的有效性,选取其他云资源数据集中的某台机器数据进行测试,机器 2 模型预测结果对比图如图 10 所示。

从图 10 中可以发现,每个模型都能大致预测时间序列数据未来的大致趋势,模型间的差异更多体现在峰值和谷值的预测上,对于数据的突变一直是时序预测的难点,可以看出 TCN-LSTM 组合模型在的对于整体趋势以及峰值的预测要优于其他单一模型和组合模型,所提出模型的拟合度较好。

为评估模型的有效性,实验分别选取 5 种评估指标对实验误差进行分析,分别为均方根误差(root mean squared error,RMSE),平均绝对值百分比误差(mean absolute percentage error,MAPE),平均绝对误差(mean absolute error,MAE),决定系数(R^2)以及运行时间,它们的计算公式分别为

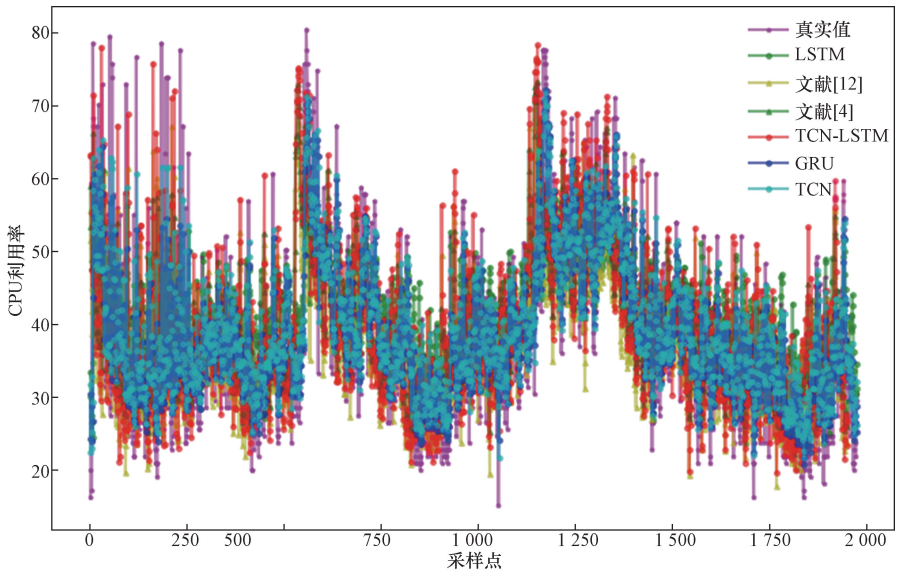


图 10　机器 2 模型的预测值与真实值对比图

Fig. 10　The predicted value of machine 2 model compared with the true value

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i-y_i)^2} \tag{9}$$

$$\mathrm{MAPE}=\frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{10}$$

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i-y_i| \tag{11}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{12}$$

式中:m 代表样本数量;$\hat{y}_i$ 代表模型对于时刻 i 的预测值;y_i 代表时刻 i 的真实值。

随机选择两台机器,对于各个模型在两台机器上的预测误差对比结果如表 4 和表 5 所示。

从表 4、表 5 中可以看出,单一模型的 LSTM 和 GRU 结果相似,TCN 表现略差,R^2 表示了模型的拟合程度,该值越接近 1 表示模型的拟合度越好,组合模型 TCN-LSTM 的 R^2 高于单一模型以及其他组合模型,误差 MAPE 相比单一模型降低 8.1% ~ 13.7%,RMSE 降低 9.8% ~ 13.1%,可以看出所提出的组合模型的效果更好一些,能够预测出大体趋势,且在峰值的预测上要优于其他模型,误差的评价指标低于单一模型。

表 4　机器 1 模型预测误差对比

Table 4　Machine 1 model prediction error comparison

预测模型	评价指标				
	MAE	MAPE	RMSE	R^2	运行时间/s
LSTM	5.051 4	0.116 5	6.681 1	0.512 1	60
GRU	5.265 0	0.120 6	6.808 9	0.505 8	58
TCN	5.449 4	0.129 9	6.896 1	0.484 8	65
文献[4]	4.863 0	0.147 0	7.187 1	0.625 0	82
文献[12]	4.890 9	0.149 9	7.244 6	0.622 0	85
TCN-LSTM	4.542 5	0.120 9	5.956 8	0.696 3	95

表 5　机器 2 模型预测误差对比

Table 5　Machine 2 model prediction error comparison

预测模型	评价指标				
	MAE	MAPE	RMSE	R^2	运行时间/s
LSTM	5.151 4	0.118 5	6.681 1	0.512 1	62
GRU	5.265 0	0.123 6	6.808 9	0.505 8	57
TCN	5.229 3	0.121 9	6.856 1	0.496 7	63
文献[4]	4.632 3	0.115 6	5.856 3	0.688 0	83
文献[12]	4.723 1	0.116 4	5.902 3	0.660 3	88
TCN-LSTM	4.432 5	0.110 9	5.456 8	0.706 3	90

3　结论

提出一种基于时间卷积和长短期记忆网络的组合云资源预测模型,利用时间卷积网络可以通过层数、扩张因子以及过滤器大小来调节感受野的大小,获取多的特征保证长期依赖关系,其中的残差连接使得输入长度很长时,梯度也能更稳定,缩短模型的训练时间。与单一模型相比,所提出的 TCN-LSTM 组合预测模型表现较优于其他单一模型和部分组合模型,可以对容器云资源的分配提供有效的参考,提高容器云资源的利用率,具有一定的实际应用价值。

虽然所提出的模型性能优于单一模型,但是仍存在一定的局限性:对于峰值的预测仍不够准确,因此在未来的工作中将进一步研究提升时间序列数据对于峰值预测的精确度。

参　考　文　献

[1] Xie X, Yuan T, Zhou X, et al. Research on trust model in con-

tainer-based cloud service[J]. Computers, Materials and Continua, 2018, 56(2): 273-283.

[2] Al-Sayed M M, Khattab S, Omara F A. Prediction mechanisms for monitoring state of cloud resources using Markov chain model[J]. Journal of Parallel and Distributed Computing, 2016, 96: 163-171.

[3] 赵莉. 基于支持向量机的云计算资源负载预测模型[J]. 南京理工大学, 2018, 42(6): 687-692.
Zhao Li. Cloud computing resource load prediction model based on support vector machine [J]. Nanjing University of Science and Technology, 2018, 42(6): 687-692.

[4] 廖雪超, 伍杰平, 陈才圣. 结合注意力机制与 LSTM 的短期风电功率预测模型[J]. 计算机工程, 2022, 48(9): 286-297, 304.
Liao Xuechao, Wu Jieping, Chen Caisheng. Short-term wind power prediction model combining attention mechanism and LSTM [J]. Computer Engineering, 2019, 48(9): 286-297, 304.

[5] Lin J, Ma J, Zhu J G, et al. Short-term load forecasting based on LSTM networks consideringattention mechanism, international [J]. Journal of Electrical Power & Energy Systems, 2022, 137: 107818.

[6] Ren X Q, Liu S L, Yu X D, et al. A method for state-of-charge estimation of lithium-ion batteries based on PSO-LSTM[J]. Energy, 2021, 234: 121236.

[7] Bai S J, Zico K J, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. Computing Research Repository, 2018. https://arxiv.org/abs/1803.01271.

[8] 王悦悦, 谢晓兰, 郭杨, 等. 基于自适应神经网络的云资源预测模型[J]. 科学技术与工程, 2021, 21(25): 10814-10819.
Wang Yueyue, Xie Xiaolan, Guo Yang, et al. Cloud resource prediction model based on adaptive neural network[J]. Science Technology and Engineering, 2019, 21(25): 10814-10819.

[9] 贺小伟, 徐靖杰, 王宾, 等. 基于 GRU-LSTM 组合模型的云计算资源负载预测研究[J]. 计算机工程, 2022, 48(5): 11-17, 34.
He Xiaowei, Xu Jingjie, Wang Bin, et al. Research on cloud computing resource load prediction based on GRU-LSTM combination model [J]. Computer Engineering, 2022, 48(5): 11-17, 34.

[10] 李新飞, 谢晓兰. 基于 Holt-Winters 及长短期记忆的云资源组合预测模型[J]. 科学技术与工程, 2022, 22(13): 5306-5311.
Li Xinfei, Xie Xiaolan. Cloud resource portfolioprediction model based on Holt-Winters and long short-term memory [J]. Science Technology and Engineering, 2002, 22(13): 5306-5311.

[11] 王琛, 王颖, 郑涛, 等. 基于 ResNet-LSTM 网络和注意力机制的综合能源系统多元负荷预测[J]. 电工技术学报, 2022, 37(7): 1789-1799.
Wang Chen, Wang Ying, Zheng Tao, et al. Multivariate load forecasting of integrated energy system based on ResNet-LSTM network and attention mechanism [J]. Transactions of China Electrotechnical Society, 2022, 37(7): 1789-1799.

[12] Shen Z P, Fan X C, Zhang L Y, et al. Wind speed prediction of unmanned sailboat based on CNN and LSTM hybrid neural network [J]. Ocean Engineering, 2022, 254: 111352.

[13] Mei P, Li M, Zhang Q, et al. Prediction model of drinking water source quality with potential industrial-agricultural pollution based on CNN-GRU-attention [J]. Journal of Hydrology, 2022, 610: 127934.

[14] Gao S, Huang Y F, Zhang S, et al. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation[J]. Journal of Hydrology, 2020, 589: 125188.

[15] 董兰芳, 张军挺. 基于深度学习与随机森林的人脸年龄与性别分类研究[J]. 计算机工程, 2018, 44(5): 246-251.
Dong Lanfang, Zhang Junting. Research on faceage and gender classification based on deep learning and random forest [J]. Computer Engineering, 2018, 44(5): 246-251.