·热点与综述·　　文章编号：1000-3428(2022)05-0011-07　　文献标志码：A　　中图分类号：TP391

基于GRU-LSTM组合模型的云计算资源负载预测研究

贺小伟[1,2]，徐靖杰[1]，王　宾[2]，吴　昊[1]，张博文[2]

（1.西北大学 网络和数据中心，西安 710127；2.西北大学 信息科学与技术学院，西安 710127）

摘　要：日益增多的应用部署在云端使得云数据中心的功耗波动剧烈，从而导致云数据中心资源利用率不平衡，高效的负载预测是解决该问题的关键技术。针对目前负载预测模型预测精度低、预测时间长的问题，建立一种基于门控循环单元（GRU）与长短期记忆（LSTM）网络的组合预测模型GRU-LSTM。该模型的网络结构包括3层，第一层采用GRU，利用GRU参数少、易收敛的特点减少模型训练时间，第二、第三层采用LSTM，结合LSTM参数多的优势提高模型的预测精度。在此基础上，对数据集作缺失值处理和标准化处理，使用随机森林算法对原始序列进行特征选择后得到一组新的序列值，将该序列值作为GRU-LSTM组合预测模型的输入，以对云计算资源进行高效预测。在集群公开数据集Cluster-trace-v2018上进行实验，结果表明，与传统的单一预测模型ARIMA、LSTM、GRU以及现有的组合预测模型ARIMA-LSTM、Refined LSTM等相比，GRU-LSTM模型预测结果的均方误差减少6~9，预测时间平均缩短约10%。

关键词：云计算；负载预测；预测模型；门控循环单元；长短期记忆网络

开放科学（资源服务）标志码（OSID）：

中文引用格式：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.

英文引用格式：HE X W，XU J J，WANG B，et al.Research on cloud computing resource load forecasting based on GRU-LSTM combination model[J].Computer Engineering，2022，48(5)：11-17，34.

Research on Cloud Computing Resource Load Forecasting Based on GRU-LSTM Combination Model

HE Xiaowei[1,2]，XU Jingjie[1]，WANG Bin[2]，WU Hao[1]，ZHANG Bowen[2]

（1.Network and Data Center，Northwest University，Xi'an 710127，China；
2.School of Information Science and Technology，Northwest University，Xi'an 710127，China）

【Abstract】Increasingly more applications are deployed in the cloud，causing the violent power fluctuation of cloud data consumption and unbalancing resource usage in cloud data centers.Efficient load forecasting is an essential technology for solving these problems.Targeting the low prediction accuracy and long prediction time of current load prediction models，a combined prediction GRU-LSTM model based on Gated Recurrent Unit（GRU）and Long-Short Term Memory（LSTM）network is established.The network structure of the model includes three layers.The first layer adopts GRU，which reduces the training time of the model using the advantages of having few GRU parameters and easy convergence.The second and third layers adopt LSTM，which improves the prediction accuracy of the model by combining the advantages of many LSTM parameters.On this basis，the missing values and standardization of the dataset are processed.After the feature selection of the original sequence using the random forest algorithm，a set of new sequence values are obtained. The sequence values are used as the input of the GRU-LSTM combined prediction model to efficiently predict the cloud computing resources. Experiments are performed on the Cluster-trace-v2018 public dataset.Using the proposed GRU-LSTM model，the Mean Square Error（MSE）of the prediction results and average prediction time reduced by 6~9 and shortened by approximately 10%，respectively，compared with the traditional single prediction ARIMA，LSTM，and GRU models and existing combined prediction ARIMA-LSTM and Refined LSTM models.

【Key words】cloud computing；load forecasting；forecasting model；Gated Recurrent Unit（GRU）；Long-Short Term Memory（LSTM）network

DOI：10. 19678/j. issn. 1000-3428. 0062452

基金项目：国家重点研发计划"智慧博物馆关键技术研发与示范"（2019YFC1521100）；教育部第二批新工科研究与实践项目"面向智慧教学的新工科教育教学资源平台建设"（E-XTYR20200665）；西安市科技计划项目（2020KJRC0117）。

作者简介：贺小伟（1977—），男，教授、博士，主研方向为大数据分析；徐靖杰，硕士研究生；王　宾，讲师、博士；吴　昊，高级工程师、博士；张博文，硕士研究生。

收稿日期：2021-08-23　　**修回日期：**2021-10-15　　**E-mail：**wbin@nwu.edu.cn

0 概述

云计算[1]是基于虚拟化技术和网格计算而发展起来的一种新兴计算模式，与网格计算相比，云计算的任务更具复杂性，在这种模式中，应用、数据和IT资源以服务的形式经过网络提供给用户使用，给用户带来诸多方便。近年来，越来越多的公司将应用部署在云端，这使得云数据中心的功耗波动更加剧烈，从而导致云数据中心资源利用率不平衡问题[2]。在负载快速增加时，从分配主机资源到使用主机资源的过程会产生时间延迟，在负载变化后分配资源时服务水平协议（Service Level Agreement，SLA）将遭到破坏。

为了给用户提供高性能的云服务，在云分布式控制系统中进行资源管理非常重要[3]，它可以降低云数据中心能耗成本和二氧化碳排放量[4-5]。负载预测技术[6]是对云分布式控制系统进行资源管理的关键技术，其可以在不破坏SLA和不影响数据中心运行的前提下，通过对云数据中心的历史数据进行分析，以掌握负载数据的变化规律并准确预测下一时期的负载值，从而通过合理地分配云数据中心的资源来提高其资源利用率。

目前，针对时间序列的负载预测大多基于深度学习中的循环神经网络（Recurrent Neural Network，RNN）来进行建模，但是RNN在预测时间序列时存在梯度消失和梯度爆炸的问题。本文针对负载预测模型中存在的预测精度低、预测时间长的问题，提出一种组合预测模型GRU-LSTM，该模型结合门控循环单元（Gate Recurrent Unit，GRU）预测时间短、长短期记忆（Long-Short Term Memory，LSTM）预测精度高的优点，对云计算资源进行高效预测，从而提高云数据中心的资源利用率。

1 相关工作

对云计算资源负载进行预测是一个典型的时间序列预测问题。目前，针对时间序列的预测方法主要分为基于机器学习的方法和基于深度学习的方法两类。

在基于机器学习的方法中，KHAIRALLA[7]提出一种混合模型，其将自回归移动平均模型（Auto-Regression and Moving Average Model，ARIMA）与支持向量机（Support Vector Machine，SVM）进行组合，然后对金融时间序列进行预测，并与人工神经网络（Artificial Neural Network，ANN）和SVM进行对比。BI等[8]提出一种小波分解与ARIMA的混合模型以对未来负荷进行预测，该模型通过SavitzkyGolay滤波平滑任务时间序列，并通过小波分解将其分解为多个分量，通过小波分解重构它们的预测结果，以估计到达任务数。但是，该模型可以通过使用更好的数据平滑算法来进一步提高其预测精度。LIU[9]利用支持向量回归（Support Vector Regression，SVR）建立0-1整数规划模型，将工作负载分类问题转化为任务分配问题，并提供在线解决方案。ZHONG[10]提出一种基于WSVM的预测算法，其利用小波分解对输入信号进行分解然后使用SVM完成预测，最后利用Google提供的集群公开数据集进行实验验证。GUPTA[11]提出一种基于分数差分的方法来捕捉时间序列数据的长相关性，并在Google提供的集群公开数据集上进行实验验证，结果表明，与非分数差分的方法相比，分数差分的方法具有更好的预测结果。

在基于深度学习的方法中，ZHANG[12]引入一种深度学习模型，利用规范多元分解来预测云负荷，并使用深度学习模型来学习虚拟机中复杂负载数据的重要特征，最后应用规范多元分解来压缩模型参数以提高训练效率。目前，针对时间序列的预测大多基于RNN来建模。文献[13]利用RNN对云计算资源负载进行预测，并使用Google提供的集群公开数据集进行实验来验证该方法的准确性，RNN的循环自回归结构能对时间序列进行很好地表示，但它在对时间序列进行预测时存在梯度消失和梯度爆炸的问题。因此，由RNN衍生出的LSTM网络被广泛应用于时间序列预测任务。SUDHAKAR[14]将LSTM与RNN进行组合，以对服务器未来的工作负载进行预测，并与ARIMA模型进行比较，结果表明，LSTM-RNN模型的预测精度高于ARIMA，同时也验证了组合预测模型的预测精度高于单一预测模型，但由于LSTM与RNN都含有较多参数，因此预测时间较长。文献[15]提出一种改进的LSTM预测模型GSO-LSTM，利用该模型对云主机负载进行预测，并通过实验验证了利用萤火虫智能优化（Glowworm Swarm Optimization，GSO）算法对LSTM进行优化具有可行性。文献[16]提出一种组合预测模型ARIMA-LSTM，以对云平台资源进行预测，并利用CRITIC数据融合对误差值进行预测，实验结果表明，该预测模型的预测精度优于单一预测模型。CHEN[17]对LSTM采用集成学习的思想，通过丰富模型的输入维度来解决由于缺乏历史数据而无法建模的问题，但是，由于模型的输入维度增加，导致该方法训练时间较长。在使用LSTM对时间序列进行预测时，增加网络层数也可以取得更好的预测结果。文献[18-19]分别提出Refined LSTM与Stacked LSTM预测模型（Refined LSTM、Stacked LSTM均为多层网络结构），对电荷数据进行预测后得出结论，Refined LSTM与Stacked LSTM预测精度都优于LSTM。在对时间序列进行预测时，虽然LSTM在预测精度上优于其他单一预测模型，但是LSTM参数过多，导致预测时间较长。

作为LSTM的变体，GRU在对时间序列进行预测时，由于减少了一个门而使得预测时间变短。GUO等[20]提出GRU和自相关分析相结合的多步预测方

法，仿真分析结果表明，所提GRU预测模型在进行负载预测时预测时间得到优化。对于一些特定的时间序列问题，采用时间卷积网络（Temporal Convolutional Network，TCN）进行建模也可以取得很好的效果。文献[21]利用TCN对几个真实世界的数据集进行实验，结果表明，TCN对于点预测和概率预测具有较好的效果，但是对于长时间预测的效果较差。

现有的预测模型通常是单一预测模型，或者是一些基于集成学习的组合预测模型[22]。单一预测模型虽然在预测时间上优于组合预测模型，但预测精度明显低于组合预测模型。而现有的一些组合预测模型预测精度虽然高于单一预测模型，但是预测精度相对也较低且没有考虑预测时间的问题。为提高云平台负载的预测精度和预测效率，本文结合GRU与LSTM各自的预测优势，提出一种基于GRU与LSTM的组合预测模型，以对云平台资源负载进行预测。

2　负载预测模型

2.1　问题描述

本文主要对云计算资源的负载情况进行预测。假设云计算原始资源负载序列值为 $\boldsymbol{x}_t=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中，$\boldsymbol{x}_t$为$t$时刻集群的负载情况。将CPU利用率和内存（mem）利用率的负载值进行组合后作为新的输入，原因是这2个指标可以最直接地体现集群在某个时刻的负载情况，也是使用随机森林算法[23]进行特征选择后得分最高的2个指标。设置步长为k，从原始资源负载序列值$\boldsymbol{x}_t$中选取前k个时刻的数据作为负载预测模型的输入向量 $\boldsymbol{z}=\{\boldsymbol{z}_{n-k},\cdots,\boldsymbol{z}_{n-2},\boldsymbol{z}_{n-1}\}$，通过负载预测模型训练后得到$k$+1时刻集群的负载情况。

云环境中工作负载存在一定的依赖关系，每个时刻的负载情况都和之前时刻的负载情况有着十分密切的联系，当历史值越接近当前时刻t的值，它们之间的关系就越密切。对于模型选择而言，长距离的依赖信息可以提供趋势信息，不能完全忽略掉，为了更好地利用过去的负载数据，需要对历史数据进行有选择地保留和丢弃。LSTM中的遗忘门可以控制进入当前时刻的历史信息量以及需要被舍弃的信息量。因此，本文模型选择LSTM以及与LSTM具有相似网络结构的GRU。

2.2　LSTM

LSTM作为一种改进的RNN，其继承了RNN模型的优点，并且利用独特的门结构有效解决了RNN中的梯度爆炸和梯度消失问题，因此，LSTM可以有效处理长时间序列问题，并已经成功应用于语音识别、图像描述、自然语言处理等领域。相较RNN和GRU，LSTM模型的拟合和预测精度总体较高，但是，由于LSTM参数过多导致其训练过程耗时较长。

LSTM由多个循环单元组成，通过更新神经元状态信息，使用输入门、输出门、遗忘门来控制历史信息的权重从而存储过去的信息。基于LSTM的移动云主机t时刻负荷预测模型的单元结构如图1所示。

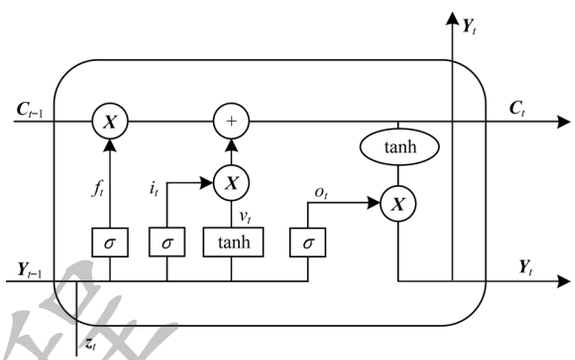

图1　LSTM单元结构

Fig.1　Unit structure of LSTM

在图1中：$\boldsymbol{C}_{t-1}$为前一时刻神经元的状态；$\boldsymbol{Y}_{t-1}$为前一时刻神经元的输出；$\boldsymbol{Z}_t$为当前时刻的输入。

遗忘门决定上一时刻的单元状态$\boldsymbol{C}_{t-1}$有多少可以保留到当前时刻$\boldsymbol{C}_t$，遗忘门的输入包括前一时刻的输出$\boldsymbol{Y}_{t-1}$和当前时刻的负载$\boldsymbol{Z}_t$，最后通过最左边的激活函数sigmoid得到f_t，计算如式（1）所示：

$$f_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{Y}_{t-1},\boldsymbol{Z}_t]+\boldsymbol{b}_f) \tag{1}$$

$$\sigma(z)=\frac{1}{1+\mathrm{e}^{-z}} \tag{2}$$

其中：$\boldsymbol{W}_f$是遗忘门的权重矩阵；$\boldsymbol{b}_f$表示偏差向量；σ表示激活函数sigmoid；f_t表示最后一层神经元被遗忘的概率，取值范围为[0，1]，0表示完全丢弃，1表示完全保留。

输入门决定$\boldsymbol{Z}_t$中哪些新的输入可以存储在神经元中，神经元主要分为i_t和v_t，其计算方法分别如式（3）和式（4）所示：

$$i_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{Y}_{t-1},\boldsymbol{Z}_t]+\boldsymbol{b}_i) \tag{3}$$

$$v_t=\tanh(\boldsymbol{W}_c\cdot[\boldsymbol{Y}_{t-1},\boldsymbol{Z}_t]+\boldsymbol{b}_c) \tag{4}$$

$$\tanh(z)=\frac{\mathrm{e}^{z}-\mathrm{e}^{-z}}{\mathrm{e}^{z}+\mathrm{e}^{-z}} \tag{5}$$

其中：$\boldsymbol{W}_i$和$\boldsymbol{W}_c$是输入门的权重矩阵；$\boldsymbol{b}_i$和$\boldsymbol{b}_c$为偏差向量。i_t和v_t表示当前需要保留的负载信息的比例，其与前一时刻保留的神经元状态相加，生成更新后的神经元状态信息$\boldsymbol{C}_t$，计算方法如式（6）所示：

$$\boldsymbol{C}_t=f_t\cdot\boldsymbol{C}_{t-1}+i_t\cdot v_t \tag{6}$$

输出门控制神经元状态的输出并将状态转移到下一个神经元，输出门o_t的计算方法如式（7）所示：

$$o_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{Y}_{t-1},\boldsymbol{Z}_t]+\boldsymbol{b}_o) \tag{7}$$

其中：$\boldsymbol{W}_o$为输出门的权重矩阵；$\boldsymbol{b}_o$为偏差向量；$\boldsymbol{Y}_t$为当前时刻神经元的输出。通过最后一层的计算得到最终的负载预测值，如式（8）所示：

$$\boldsymbol{Y}_t=o_t\cdot\tanh(\boldsymbol{C}_t) \tag{8}$$

2.3 GRU

GRU作为LSTM的一种变体，也可以有效地解决RNN中的梯度爆炸和梯度消失问题。与LSTM相比，GRU的结构更为简单，其将遗忘门和输入门合并为一个更新门。由于GRU减少了一个门，矩阵乘法变少，因此当训练数据量很大时可以节省大量的时间，GRU网络结构如图2所示。

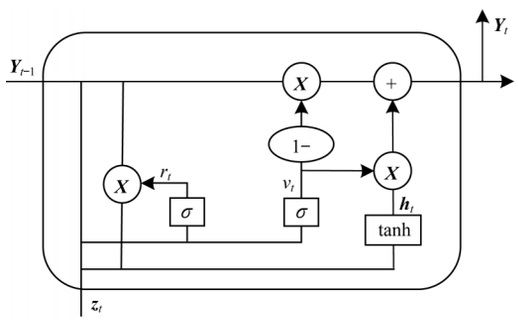

图2 GRU网络结构

Fig.2 GRU network structure

在图2中：$\boldsymbol{Z}_t$为当前时刻的输入；$\boldsymbol{Y}_{t-1}$为上一时刻的输出；$\boldsymbol{Y}_t$为当前时刻的输出。GRU有2个门：

1）第一个门为更新门v_t，其决定有多少历史信息可以继续传递给未来，即更新门v_t决定是否将隐藏状态更新为新状态。从功能上看，GRU的更新门v_t类似于LSTM的输出门，更新门v_t的计算方法如式（9）所示：

$$v_t=\sigma(\boldsymbol{W}_v\cdot[\boldsymbol{Y}_{t-1},\boldsymbol{Z}_t]+\boldsymbol{b}_v) \tag{9}$$

其中：$\boldsymbol{W}_v$为更新门的权重矩阵；$\boldsymbol{b}_v$为偏差向量。

2）第二个门为重置门r_t，其主要功能是确定有多少历史信息不能传递到下一个状态，类似于LSTM中遗忘门和输入门的组合。重置门r_t的计算方法如式（10）所示：

$$r_t=\sigma(\boldsymbol{W}_r\cdot[\boldsymbol{Y}_{t-1},\boldsymbol{Z}_t]+\boldsymbol{b}_r) \tag{10}$$

其中：$\boldsymbol{W}_r$为重置门的权重矩阵；$\boldsymbol{b}_r$为偏差向量。

计算出更新门v_t和重置门r_t后，GRU将计算候选隐藏状态$\boldsymbol{h}_t$。候选隐藏状态$\boldsymbol{h}_t$计算方法如式（11）所示：

$$\boldsymbol{h}_t=\tanh(\boldsymbol{W}_h\cdot[r_t\cdot\boldsymbol{Y}_{t-1},\boldsymbol{Z}_t]+\boldsymbol{b}_h) \tag{11}$$

其中：$\boldsymbol{W}_h$为对应的权重参数；$\boldsymbol{b}_h$为对应的偏差参数。从式（11）可以看出：当重置门r_t的值接近0时，重置门对应的候选隐藏状态值也为0，即丢弃上一时刻的隐藏状态；当重置门r_t的值接近1时，表示保留上一时刻的隐藏状态。因此，重置门的作用是有选择地丢弃与预测无关的历史信息。

最后，t时刻GRU的输出$\boldsymbol{Y}_t$的计算方法如式（12）所示：

$$\boldsymbol{Y}_t=(1-v_t)\cdot\boldsymbol{Y}_{t-1}+v_t\cdot\boldsymbol{h}_t \tag{12}$$

2.4 组合预测模型

综合考虑预测精度和预测时间2个方面的因素，本文建立一种基于GRU与LSTM的组合预测模型，以对云计算资源负载进行预测。组合预测模型网络结构如图3所示。

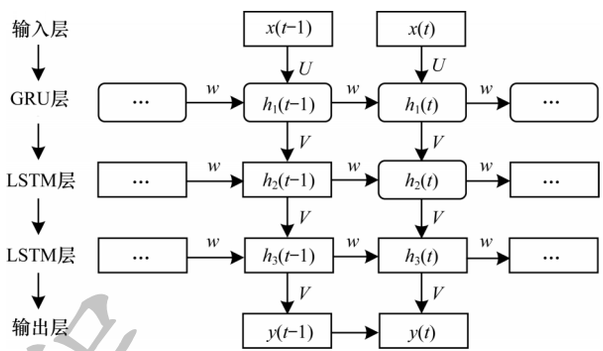

图3 组合预测模型网络结构

Fig.3 Network structure of combined forecasting model

本文组合预测模型的网络结构包括3层：第一层网络结构采用GRU，由于GRU的网络结构简单，参数少，更易收敛，因此在训练数据时GRU训练速度快，可以减少训练时间，但是，GRU的预测精度低于LSTM；第二层和第三层网络结构均采用LSTM，LSTM参数较多，预测精度更高，并且双层LSTM的预测精度优于单层LSTM。

2.5 模型评价标准

本文采用平均绝对误差（Mean Absolute Error，MAE）、平均绝对值百分比误差（Mean Absolute Percentage Error，MAPE）、均方误差（Mean Squared Error，MSE）、均方根误差（Root Mean Square Error，RMSE）、决定系数（R^2）作为模型的性能评估标准。预测时间主要以模型的训练时间作为评价标准。在对模型进行泛化实验时，采用可释方差得分（explanied_variance_score，简写为S）对模型的拟合程度进行评价。各指标的计算方法如下：

$$M_{\mathrm{MAE}}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{13}$$

$$M_{\mathrm{MAPE}}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\% \tag{14}$$

$$M_{\mathrm{MSE}}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{15}$$

$$R_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{16}$$

$$\mathrm{R}^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{17}$$

$$S=1-\frac{\mathrm{Var}\{y_i-\hat{y}_i\}}{\mathrm{Var}\{y_i\}} \tag{18}$$

其中：n为负载预测值个数；y_i为负载的实际值；$\hat{y}_i$为负载的预测值；$\bar{y}$为负载的平均值。

3 实验结果与分析

3.1 数据来源与处理

为了验证本文负载预测模型的预测性能，使用

阿里云2018年发布的集群公开数据集Cluster-trace-v2018[24]进行实验。Cluster-trace-v2018包括大约4 000台机器在8天内的资源使用情况，本文实验使用其中1台机器在8天内的资源使用情况，共3 300条数据记录，选取前80%的数据作为训练集，后20%的数据作为测试集。本文对数据集进行如下处理：

1）缺失值处理。利用均值填充法进行缺失值处理。

2）标准化处理[25]。为了提高神经网络的收敛速度、迭代求解速度以及预测精度，采用式（19）（Min-Max归一化）对训练集的2 600个数据进行标准化处理：

$$X_t^* = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}} \tag{19}$$

其中：X_t^*为归一化处理后的值；X_t为原始数据在t时刻的值；$X_{\min}$为原始数据的最小值；$X_{\max}$为原始数据的最大值。

3）特征选择。特征选择可以从原始数据特征集中选出若干个具有代表性的特征子集，这不仅可以实现数据降维，还可以提升在该特征子集上所构建的回归模型的性能。随机森林算法通过特征随机置换前后误差分析，计算每个特征的重要性得分，得分越高，特征越重要，与其他特证选择算法相比，随机森林算法不仅能体现特征间的相互作用，而且还具有准确性高、鲁棒性好等优点[26]。因此，本文使用随机森林算法进行特征选择，特征选择结果如图4所示，横坐标为特征。

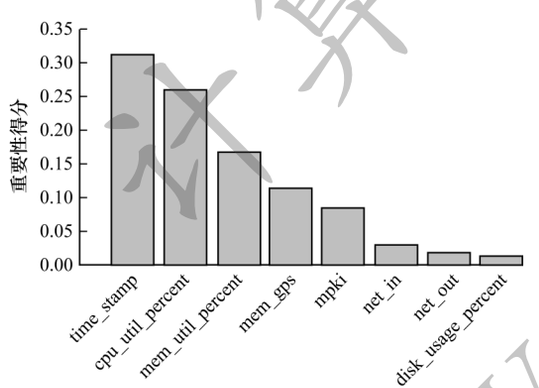

图4　特征选择结果

Fig.4　Feature selection results

从图4可以看出，CPU利用率和mem利用率是除时间特征之外经过特征选择后得分最高的2个特征，因此，本文实验选择CPU利用率和mem利用率的组合值作为云平台资源的负载值。负载值的计算公式如下：

$$\boldsymbol{W} = \boldsymbol{W}_1 \cdot L_1 + \boldsymbol{W}_2 \cdot L_2 \tag{20}$$

其中：$\boldsymbol{W}$为组合后的原始负载值；$\boldsymbol{W}_1$和$\boldsymbol{W}_2$分别代表CPU负载值和mem负载值；L_1和L_2分别代表CPU和mem的权重参数。本文根据特征的得分指数比例，设置L_1为0.6，L_2为0.4。

3.2　实验过程

本文实验是一个单步预测过程，预测流程如图5所示。

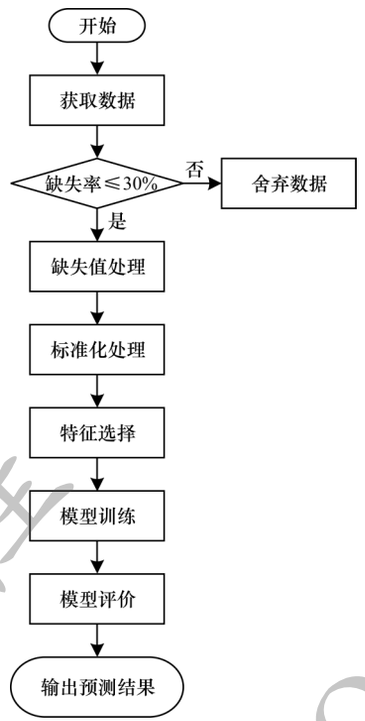

图5　预测流程

Fig.5　Procedure of prediction

在获取数据后，首先对原始数据进行缺失值和标准化处理，当原始数据缺失率大于30%时，舍弃该条数据，若缺失率小于等于30%，则使用均值填充法对缺失值进行填充，并将填充后的数据通过归一化方法进行标准化处理，利用随机森林算法对标准化后的数据进行特征选择，根据特征的重要程度将特征选择后得到的特征数据加入权重参数进行组合，将组合后的负载值输入GRU-LSTM组合预测模型进行训练，设置步长为12（根据前12个数据来预测第13个数据）。最后，使用5个评价标准对模型性能进行评价，同时输出模型预测结果。预测模型参数设置如表1所示。

表1　参数设置

Table 1　Parameters setting

训练参数	参数设置
GRU隐藏神经元个数	24
第一层LSTM隐藏神经元个数	24
第二层LSTM隐藏神经元个数	24
激活函数	tanh
迭代次数	400

模型预测结果（64个样本）如图6所示，从图6可以看出，本文所提GRU-LSTM组合预测模型的预测结果与原始序列的趋势基本一致。从图7可以看出（640个样本），本文GRU-LSTM模型经过训练后得到的预测数据与真实数据的误差大多集中在−5.0~5.0之间，误差较小，因此，该组合预测模型预测精度较高。

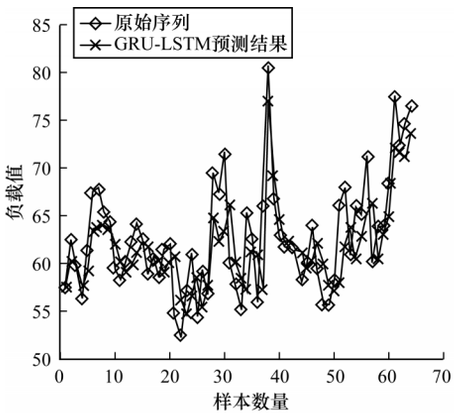

图6 GRU-LSTM模型的预测结果

Fig.6 Prediction results of GRU-LSTM model

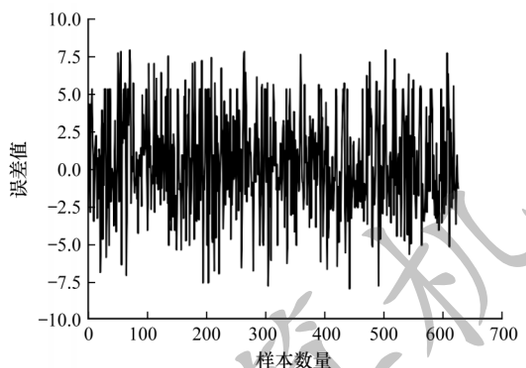

图7 GRU-LSTM预测误差结果

Fig.7 GRU-LSTM prediction error results

3.3 结果分析

将本文模型与传统的单一负载预测模型ARIMA、GRU、LSTM进行对比，同时还与文献[16]提出的ARIMA-LSTM组合预测模型、文献[18]提出的Refined LSTM模型、文献[19]提出的Stacked LSTM模型进行实验对比。各模型的负载预测结果如图8所示(64个样本)，评价指标结果如表2所示，最优结果加粗表示。

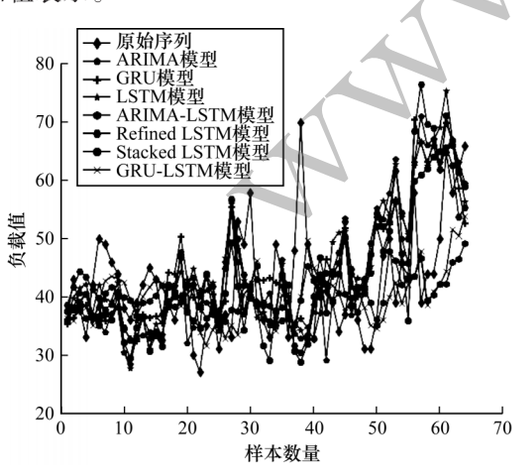

图8 各模型的负载预测结果

Fig.8 Load prediction results of each model

表2 各模型的评价指标结果

Table 2 Evaluation index results of each model

预测模型	MSE	RMSE	MAE	MAPE /%	R^2	预测时间/s
ARIMA	27.316 3	5.226 5	3.757 4	5.893 4	0.403 5	66
GRU	26.536 9	5.151 4	3.691 4	5.655 1	0.415 5	**57**
LSTM	25.587 4	5.058 4	3.655 4	5.412 3	0.433 6	62
Refined LSTM	22.313 3	4.723 7	3.468 6	5.325 4	0.468 1	92
Stacked LSTM	24.595 6	4.959 4	3.538 1	5.397 7	0.453 7	82
ARIMA-LSTM	23.743 2	4.872 7	3.488 5	5.345 3	0.461 2	88
GRU-LSTM	**18.105 0**	**4.255 0**	**3.038 3**	**4.831 1**	**0.523 5**	78

从图8和表2可以看出：相较传统的单一预测模型ARIMA、LSTM以及GRU，本文GRU-LSTM模型的预测精度较高，同时也验证了组合预测模型的预测精度要优于单一预测模型；相较组合预测模型ARIMA-LSTM、Refined LSTM和Stacked LSTM，本文模型的预测精度同样较高。虽然2层网络结构相比3层网络结构减少了参数，但是这在一定程度上削弱了模型的学习能力。本文所提GRU-LSTM模型结合GRU训练速度快、LSTM预测性能好的优点，在预测精度和预测时间上取得性能提升。

为了对模型的泛化能力进行验证，设置不同的步长重复进行实验，使用explanied_variance_score作为评价标准，结果如图9所示。

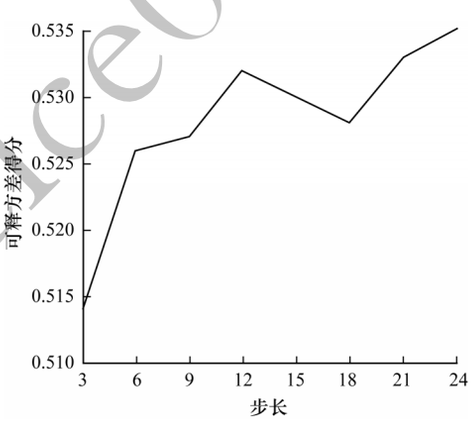

图9 不同步长下的可释方差得分结果

Fig.9 explanied_variance_score results under different steps

通过图9可以看出，GRU-LSTM组合预测模型的可释方差得分介于0.510~0.535之间，即该组合模型在不同的步长情况下均可以有效地对云计算资源负载值进行预测，其具备一定的泛化能力。

4 结束语

针对目前负载预测模型预测精度低且预测时间长的问题，本文提出一种基于LSTM与GRU的组合预测模型GRU-LSTM，以对云计算资源负载情况进行预测。在阿里云平台发布的集群公开数据集上进行实验，结果表明，相较ARIMA、LSTM等模型，该

负载预测模型具有较高的预测精度以及较短的预测时间，同时具有一定的泛化能力，将该模型应用到实际的云平台中，可以解决目前资源利用率不平衡的问题。下一步考虑通过优化算法寻找模型的最优参数，或对数据进行降噪处理，以提升本文模型的预测精度。此外，探索多变量和长时间序列的内在关联性并解决深度神经网络对输入变化不敏感的问题，也是今后的研究方向。

参考文献

[1] BOLIN J，YANG M K. Cloud computing：cost，security，and performance[C]//Proceedings of 2018 ACMSE Conference. New York，USA：ACM Press，2018：1.

[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] MASDARI M，SALEHI F，JALALI M，et al. A survey of PSO-based scheduling algorithms in cloud computing[J]. Journal of Network and Systems Management，2017，25(1)：122-158.

[4] SINGH S，CHANA I. A survey on resource scheduling in cloud computing：issues and challenges[J]. Journal of Grid Computing，2016，14(2)：217-264.

[5] SHAO G L，CHEN J M. A load balancing strategy based on data correlation in cloud computing[C]//Proceedings of the 9th International Conference on Utility and Cloud Computing. New York，USA：ACM Press，2016：364-368.

[6] MASDARI M，KHOSHNEVIS A. A survey and classification of the workload forecasting methods in cloud computing[J]. Cluster Computing，2020，23(4)：2399-2424.

[7] KHAIRALLA M A，NING X. Financial time series forecasting using hybridized support vector machines and ARIMA models[C]//Proceedings of 2017 International Conference on Wireless Communications，Networking and Applications. New York，USA：ACM Press，2017：94-98.

[8] BI J，ZHANG L B，YUAN H T，et al. Hybrid task prediction based on wavelet decomposition and ARIMA model in cloud data center[C]//Proceedings of IEEE International Conference on Networking，Sensing and Control. Washington D. C.，USA：IEEE Press，2018：1-6.

[9] LIU C H，LIU C C，SHANG Y L，et al. An adaptive prediction approach based on workload pattern discrimination in the cloud[J]. Journal of Network and Computer Applications，2017，80：35-44.

[10] ZHONG W，ZHUANG Y，SUN J，et al. The cloud computing load forecasting algorithm based on wavelet support vector machine[C]//Proceedings of the Australasian Computer Science Week Multiconference. New York，USA：ACM Press，2017：1-5.

[11] GUPTA S，DILEEP A D，GONSALVES T A. Fractional difference based hybrid model for resource prediction in cloud network[C]//Proceedings of the 15th International Conference on Network，Communication and Computing. New York，USA：ACM Press，2016：93-97.

[12] ZHANG Q C，YANG L T，YAN Z，et al. An efficient deep learning model to predict cloud workload for industry informatics[J]. IEEE Transactions on Industrial Informatics，2018，14(7)：3170-3178.

[13] ZHANG W S，LI B，ZHAO D H，et al. Workload prediction for cloud cluster using a recurrent neural network[C]//Proceedings of International Conference on Identification，Information and Knowledge in the Internet of Things. Washington D. C.，USA：IEEE Press，2016：104-109.

[14] SUDHAKAR C，KUMAR A R，SIDDARTHA N，et al. Workload prediction using ARIMA statistical model and long short-term memory recurrent neural networks[C]//Proceedings of International Conference on Computing，Power and Communication Technologies. Washington D. C.，USA：IEEE Press，2018：600-604.

[15] ZHANG Z H，ZHU W，ZHONG W，et al. Load forecasting model of mobile cloud computing based on glowworm swarm optimization LSTM network[C]//Proceedings of the 7th International Conference on Information Technology：IoT and Smart City. Washington D. C.，USA：IEEE Press，2019：113-119.

[16] 林涛，冯竞凯，郝章肖，等. 基于组合预测模型的云计算资源负载预测研究[J]. 计算机工程与科学，2020，42(7)：1168-1173.
LIN T，FENG J K，HAO Z X，et al. Cloud computing resource load prediction based on combined prediction model[J]. Computer Engineering & Science，2020，42(7)：1168-1173.（in Chinese）

[17] CHEN L，YU H H，TONG L，et al. Research on load forecasting method of distribution transformer based on deep learning[C]//Proceedings of the 7th IEEE International Conference on Cyber Security and Cloud Computing. Washington D. C.，USA：IEEE Press，2020：228-233.

[18] TANG D D，LI C，JI X H，et al. Power load forecasting using a Refined LSTM[C]//Proceedings of the 11th International Conference on Machine Learning and Computing. Washington D. C.，USA：IEEE Press，2019：104-108.

[19] YUAN C M，XIU T，LOU T Y. Probabilistic long-term load forecasting based on Stacked LSTM[C]//Proceedings of the 4th International Conference on Mathematics and Artificial Intelligence. Washington D. C.，USA：IEEE Press，2019：80-84.

[20] GUO J Y，WANG Z J，CHEN H W. On-line multi-step prediction of short term traffic flow based on GRU neural network[C]//Proceedings of the 2nd International Conference on Intelligent Information Processing. Washington D. C.，USA：IEEE Press，2017：1-6.

[21] CHEN Y T，KANG Y F，CHEN Y X，et al. Probabilistic forecasting with temporal convolutional neural network[J]. Neurocomputing，2020，399：491-501.

[22] SHETTY J，SHOBHA G. An ensemble of automatic algorithms for forecasting resource utilization in cloud[C]//Proceedings of Future Technologies Conference. Washington D. C.，USA：IEEE Press，2016：301-306.

[23] DEVIAENE M，TESTELMANS D，BORZÉE P，et al. Feature selection algorithm based on random forest applied to sleep apnea detection[C]//Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Washington D. C.，USA：IEEE Press，2019：2580-2583.

（下转第34页）

（上接第17页）

[24] CHENG Y, ANWAR A, DUAN X J. Analyzing Alibaba's co-located datacenter workloads[C]//Proceedings of IEEE International Conference on Big Data. Washington D. C., USA: IEEE Press, 2018: 292-297.

[25] 蔡亮，周泓岑，白恒，等. 基于多层BiLSTM和改进粒子群算法的应用负载预测方法[J]. 浙江大学学报(工学版)，2020，54(12)：2414-2422.
CAI L, ZHOU H C, BAI H, et al. Application load forecasting method based on multi-layer bidirectional LSTM and improved PSO algorithm[J]. Journal of Zhejiang University(Engineering Science), 2020, 54(12): 2414-2422. (in Chinese)

[26] 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 L F, ZHANG J T. Research on face age and gender classification based on deep learning and random forest[J]. Computer Engineering, 2018, 44(5): 246-251. (in Chinese)

编辑 吴云芳