文章编号：2095-2163(2024)02-0172-05 中图分类号：TP391.1 文献标志码：A

融合注意力机制的 BiLSTM 集群负载预测模型

罗 邦，张云华
（浙江理工大学 计算机科学与技术学院，杭州 310018）

摘 要：服务器集群的负载预测有助于集群资源的优化配置，以提升集群资源的利用率。针对传统负载预测算法存在预测精度低的问题，本文在双向长短期记忆网络模型的基础上，提出一种融合注意力机制的负载预测模型（Att-BiLSTM）。该模型充分考虑到服务器 CPU、内存、磁盘和网络等因素，利用双向长短期记忆网络前后传递信息的特点，并使用注意力机制关注负载时间序列中的重要信息，从而提高了预测精度。实验结果表明，相对于目前已经提出的负载预测模型 ARIMA 、CNN-LSTM 和 BiLSTM、Att-BiLSTM 模型具有更好的预测性能。

关键词：服务器集群；BiLSTM；负载预测；注意力机制

Cluster load prediction model with BiLSTM incorporating attention mechanism

LUO Bang, ZHANG Yunhua
(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Load forecasting for server clusters contributes to optimizing cluster resource allocation and improving the utilization of cluster resources. Addressing the issue of low prediction accuracy in traditional load forecasting algorithms, this paper proposes a load prediction model, called Att-BiLSTM, which integrates an attention mechanism into the Bidirectional Long Short-Term Memory (BiLSTM) network model. This model takes into full consideration factors such as server CPU, memory, disk, and network. It leverages the bidirectional information flow characteristics of the BiLSTM network and employs an attention mechanism to focus on crucial information within the load time series, ultimately enhancing prediction accuracy. Experimental results indicate that, in comparison to existing load prediction models, including ARIMA, CNN-LSTM, and BiLSTM, the Att-BiLSTM model demonstrates superior predictive performance.

Key words: server cluster; BiLSTM; load prediction; attention mechanism

0 引 言

互联网的飞速发展给人们生活带来了巨大的便利，同时也给网络服务器带来了巨大的访问量，这对于网络服务器来说是一个严峻的考验。由于传统的单一服务器无法支撑起海量的访问请求，于是出现了将多台服务器有效地组织起来共同完成任务的服务器集群[1]。然而，要实现服务器集群的高效运行，需要进行提前规划和合理配置集群资源。在此过程中，建立一个准确的集群负载预测模型具有关键作用，可以为服务器集群资源的规划和配置提供重要参考价值[2]。

传统的负载预测方法通常使用基于统计学理论的模型。如：自回归移动平均模型（ARMA）[3]和差分自回归移动平均模型（ARIMA）[4]。然而，这些模型虽然对数据序列中的线性关系能较好地拟合，但在复杂时间序列数据中的拟合效果较差，存在较大误差。近年来，由于深度学习算法拥有强大特征提取能力，其在负载预测领域中的应用越来越多。Kuo 等[5]采用了卷积神经网络（CNN）预测负载，运用 CNN 对局部特征信息进行提取，并将其聚合成全局信息，使预测准确性得以提高，但无法对时间序列数据的前后依赖关系进行记忆和学习。Nguyen 等[6]提出了一种基于循环神经网络（RNN）的预测模型，使用记忆单元对动态时间特征进行学习，然而在较长的时间序列下，容易导致梯度消失或梯度爆炸的问题，因此对长时间工作的集群负载预测并不适用。Zheng 等[7]提出了一种使用门控单元的 LSTM 网络，以解决梯度消失问题，但并未将服务器工作负载的特性考虑进去。郭杨虎[8]提出了一种

作者简介：罗 邦（1995-），男，硕士研究生，主要研究方向：软件工程技术。
通讯作者：张云华（1965-），男，博士，教授，主要研究方向：软件架构、软件工程、智能信息处理。Email：605498519@qq.com

收稿日期：2023-02-21

改进的 GRU 负载预测算法，虽然考虑到影响服务器负载的各种因素，但并未加入 Attention 机制，因此负载预测精度仍有提高的空间。

由于上述文献中提出的预测方法都存在一定的局限性，为了更好地对集群负载进行预测，本文将基于双向长短期记忆网络、注意力机制研究改进负载预测模型。

1　相关工作

1.1　负载模型

建立一个良好的计算集群负载状况的模型，有助于集群负载预测精度的提高。目前，很多负载模型只考虑了单个因素（如 CPU 或内存）对整体负载的影响，不能全面描述集群负载的特征，因此本文首先对负载模型进行研究，提出一种改进的负载模型，综合了 CPU、磁盘、内存和网络 4 种因素，可以更全面、准确地描述集群负载的状态。集群负载的计算如式（1）、式（2）所示。

$$\begin{cases} L(S_i) = [\alpha_1,\alpha_2,\alpha_3,\alpha_4]\begin{bmatrix} L^i_{\text{cpu}} \\ L^i_{\text{mem}} \\ L^i_{\text{disk}} \\ L^i_{\text{net}} \end{bmatrix} \\ \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1 \end{cases} \tag{1}$$

$$L(S) = \frac{1}{n}\sum_{i=1}^{n} L(S_i) \tag{2}$$

式（1）用于计算单台服务器 S_i 的负载率 $L(S_i)$。其中，L^i_{cpu}、L^i_{disk}、L^i_{mem}、L^i_{net} 分别代表服务器 S_i 的 CPU、磁盘、内存、网络的负载率，与之对应的 α_1、α_2、α_3、α_4 分别表示这 4 个因素在模型中所占的权重，其数值取决于服务器本身。假如服务器 S_i 主要提供 CPU 密集型服务，则应适当提高 α_1 的数值。

式（2）用于计算整个集群的负载率 $L(S)$。由于集群由多台服务器构成，共同对外提供服务，因此仅靠单台服务器的负载率来描述当前集群的整体负载情况并不准确，而 $L(S)$ 综合了所有服务器的负载，能更准确地反映出整个集群的负载情况。

1.2　长短期记忆神经网络

为了解决循环神经网络（RNN）难以学习长期依赖信息、存在梯度消失的问题，Hochreiter 等[9]提出了长短期记忆神经网络（LSTM）模型。相较于其他模型，LSTM 能更好地处理集群负载的时间序列数据，因为集群负载数据采集时间长，属于长序列类型的数据。LSTM 作为 RNN 的一种改进，通过引入门控机制来控制信息的输入和输出，从而实现对长距离信息的记忆或遗忘，LSTM 结构如图 1 所示。

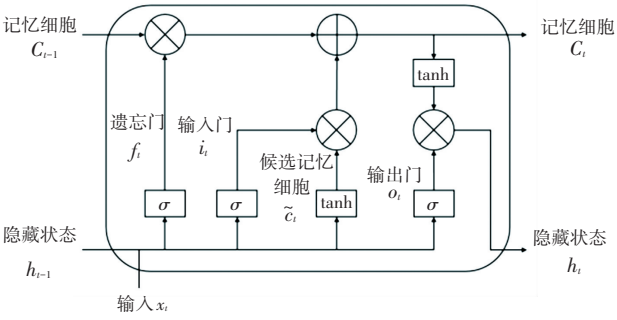

图 1　LSTM 结构

Fig. 1　The structure of LSTM

该模型主要由记忆细胞 C、更新门 i、遗忘门 f 和输出门 o 组成。更新门 i 决定当前时刻的输入信息对输出的影响程度，遗忘门 f 用于控制之前记忆的信息是否需要被遗忘，输出门 o 用于描述当前时刻输出记忆细胞与下一时刻输入信息的相关性。记忆细胞 C 存储着当前时刻所处理的特征信息。LSTM 模型涉及的符号说明见表 1，涉及到的计算公式如式（3）~式（8）所示：

$$\tilde{c}_t = \tanh\left(W_c \cdot [h_{t-1},x_t] + b_c\right) \tag{3}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1},x_t] + b_i\right) \tag{4}$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1},x_t] + b_f\right) \tag{5}$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1},x_t] + b_o\right) \tag{6}$$

$$c_t = i_t * \tilde{c}_t + f_t * c_{t-1} \tag{7}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{8}$$

表 1　符号说明

Table 1　Symbol description

符号	说明
C_{t-1}	$t-1$ 时刻记忆细胞状态
c_t	t 时刻记忆细胞状态
$\tilde{c}_t$	t 时刻记忆细胞的候选值
h_{t-1}	$t-1$ 时刻隐藏状态
h_t	t 时刻隐藏状态
x_t	t 时刻网络输入值
f_t	t 时刻的遗忘门
i_t	t 时刻的更新门
o_t	t 时刻的输出门
tanh	反正切激活函数

1.3 双向长短期记忆神经网络

由前向 LSTM 与后向 LSTM 组合而成的双向长短期记忆神经网络(BiLSTM)同时具有双向循环神经网络和长短时记忆网络的特点。BiLSTM 既可以考虑过去信息,也能够考虑未来信息,同时具有长短时记忆网络的记忆能力。为了充分利用集群负载时间序列前后信息对当前时刻负载进行预测,本文将负载预测模型改进为 BiLSTM,其结构如图 2 所示。

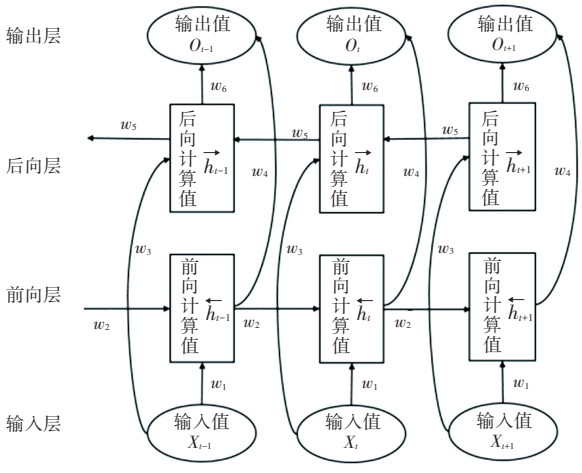

图 2 BiLSTM 结构

Fig. 2 Structure of BiLSTM

其中,$\overrightarrow{h}_t$ 代表前向计算值;$\overleftarrow{h}_t$ 代表后向计算值;o_t 代表由前向与后向值计算得到的最终输出值;w_1、w_3 代表输入到前向和后向隐藏层权重;w_2、w_5 代表隐藏层到隐藏层权重;w_4、w_6 代表前向和后向隐藏层到输出层权重。相关计算公式如下:

$$\overrightarrow{h}_t = f(w_1 X_t + w_2 \overrightarrow{h}_{t-1}) \tag{9}$$

$$\overleftarrow{h}_t = f(w_3 X_t + w_5 \overleftarrow{h}_{t+1}) \tag{10}$$

$$o_t = g(w_4 \overrightarrow{h}_t + w_6 \overleftarrow{h}_t) \tag{11}$$

1.4 注意力机制

由于传统的 LSTM 和 BiLSTM 不能对预测结果影响更大的数据进行重点关注,因此预测效果有待提高。本文在 BiLSTM 模型中引入注意力机制,着重考虑了前面某个时刻对下一时刻负载的影响,以提高预测模型的预测精度。注意力机制是一种筛选信息的方法,其能够从大量的信息中挑选出最为关键的一部分,通过提高相关参数权重的方式将注意力集中在关键信息上,对不重要的信息进行忽略。在网络训练过程中,这种方法可以帮助提高模型对关键信息的识别和利用能力。注意力机制的原理如图 3 所示。

根据图示,注意力机制将数据源中的元素抽象为一系列的键值对。并以目标中的某个元素作为查询,计算查询与每个键之间的相关性,以此为基础为每个键分配权重系数,该系数将用于对值进行加权求和,从而得出最终的注意力值。

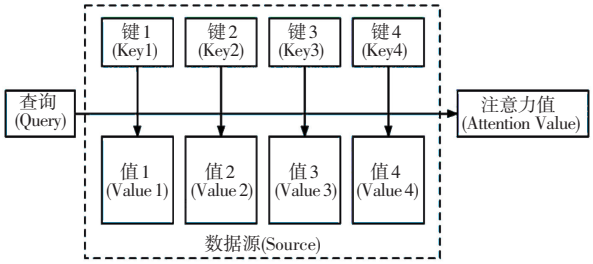

图 3 注意力机制原理

Fig. 3 Principle of attention mechanism

2 Att-BiLSTM 负载预测模型

2.1 数据预处理

为了建立一个可靠的负载预测模型,首先需要对集群中各个服务器的负载原始数据进行采集和处理(其中包括 CPU、内存、磁盘和网络数据);再通过负载模型中的式(1)和式(2)计算获得汇总的负载时间序列数据;然后将数据中 70%划分为训练集,30%划分为测试集。对输入的数据使用滑动窗口的方法进行处理,以使模型能够学习到更多的时间序列特征。

2.2 模型设计

本文提出的 Att-BiLSTM 负载预测模型主要由负载数据输入、负载数据处理、BiLSTM 网络层、注意力层和负载预测输出等 5 部分组成,如图 4 所示。

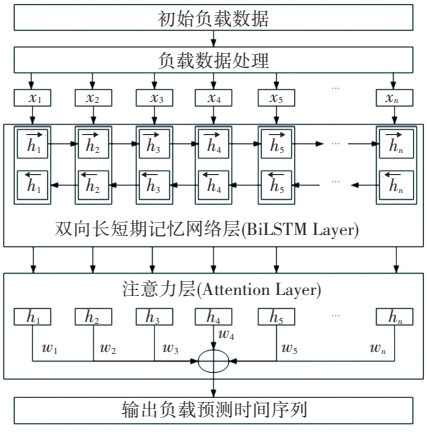

图 4 Att-BiLSTM 模型结构

Fig. 4 Structure of Att-BiLSTM model

首先，输入的负载数据经过处理后，传入 BiLSTM 网络层进行训练，其中隐藏层在前向和后向传播信息的过程中，综合处理两个方向的输入得到输出；然后，经过注意力机制模块，模型会根据所配置的时间步长，对先前某个时间点的负载数据重点关注，从而提升整个训练效果；最后，模型将负载预测结果输出。

2.3　评价指标

本文选取均方根误差（*RMSE*）、平均绝对误差（*MAE*）和平均绝对百分比误差（*MAPE*）3 个评价指标，对融合注意力机制的 BiLSTM 集群负载预测模型进行实验评估。其中，*RMSE* 衡量真实值和预测值之间偏离的大小情况；*MAE* 和 *RMSE* 类似，也是衡量偏离情况的指标；而 *MAPE* 则描述预测值误差的百分比。各评价指标的值越小，表示预测精度越高，其计算公式如下：

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

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

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

式中：y_i 代表负载真实值，$\hat{y}_i$ 代表负载预测值。

3　实验与分析

3.1　数据集

为了验证算法的有效性，本文采用阿里巴巴官方公开的集群跟踪数据作为实验数据集。该数据集记录了 4 000 个节点在 8 天内的跟踪数据，其中包含了以 10 s 为间隔测量的负载数据。为了处理数据集，首先从中提取了 4 种原始数据，即 CPU、磁盘、内存和网络负载数据；然后使用改进的负载模型中的公式获得汇总的负载时间序列数据；最后将数据以 7∶3 的比例划分为训练集和测试集。

3.2　实验参数

本文选择了均方误差（*MSE*）作为损失函数，使用自适应矩估计（Adam）优化器对参数进行优化，并引入早停（Early-Stopping）机制来加快训练速度。对于影响实验结果的精度、效率的主要模型参数（其中包括：每层的神经元数量、BiLSTM 的层数、网络训练输入的时间步长等），经过多次实验和不断调整，最优的参数设置为：神经元个数 60、层数 2、步长 10。

3.3　实验结果

经过对训练集数据的训练，使用该模型对测试集进行预测，预测结果如图 5 所示。图中展示了 Att-BiLSTM 模型的部分预测结果，其中实线表示真实值，虚线表示预测值。通过观察图形可以看出，本文提出的预测模型能够有效地预测负载值以及拟合负载变化的趋势。

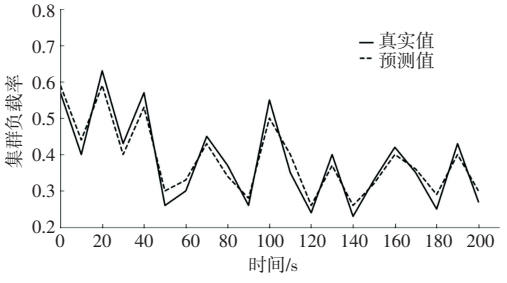

图 5　Att-BiLSTM 预测结果

Fig. 5　Att-BiLSTM prediction result

为了进一步验证本文预测模型的有效性，使用同一数据集对 ARIMA 预测模型、BiLSTM 预测模型、CNN-LSTM 预测模型进行对比实验，利用 *RMSE*、*MAE* 和 *MAPE* 指标进行评估，结果见表 2。

表 2　对比实验结果

Table 2　Comparison of experimental results

模型名称	*RMSE*	*MAE*	*MAPE*/ %
ARIMA	0.142	0.116	9.734
BiLSTM	0.063	0.057	6.395
CNN-LSTM	0.054	0.049	5.542
Att-BiLSTM	0.043	0.041	3.237

实验结果表明，本文提出的 Att-BiLSTM 负载预测模型的误差值比 ARIMA、BiLSTM 和 CNN-LSTM 预测模型都要小，验证了 Att-BiLSTM 预测模型的有效性与性能的优越性。

4　结束语

本文旨在研究和改进负载预测模型，建立了服务器集群的负载模型，针对现有负载预测模型存在预测精度低的问题，提出了一种融合注意力机制的 BiLSTM 负载预测模型。该模型将各种影响集群负载的因素充分考虑进去，利用 BiLSTM 神经网络双向传递信息的特性，并融合注意力机制对部分信息重点关注，从而将模型的预测性能提高，为服务器集群的资源分配提供了有效的决策依据。

参考文献

[1] 巩俊辉. 企业高可用服务架构研究与应用[J]. 电子技术与软件

工程,2022(13):238-241.
[2] 史爱武,张义欣,韩超,等. 基于 CEEMDAN-SE-TCN 的集群资源预测研究[J]. 软件导刊,2023,22(4):43-47.
[3] GANAPATHI A, CHEN Y, FOX A, et al. Statistics-driven workload modeling for the cloud[C]//Proceedings of 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010). IEEE, 2010: 87-92.
[4] 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.
[5] KUO P H, HUANG C J. A high precision artificial neural networks model for short-term energy load predictioning[J]. Energies, 2018, 11(1): 213.
[6] NGUYEN H M, WOO S, IM J, et al. A workload prediction approach using models stacking based on recurrent neural network and autoencoder[C]//Proceedings of 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2016: 929-936.
[7] ZHENG H, LIN F, FENG X, et al. A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(11): 6910-6920.
[8] 郭杨虎. 微服务环境下 docker 容器调度策略的研究与实现[D]. 北京:北京邮电大学, 2018.
[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.