

基于ARIMA和LSTM的高性能计算平台资源使用的预测研究

李思琪, 俞琨, 陈宇皓

引用本文

李思琪, 俞琨, 陈宇皓. 基于ARIMA和LSTM的高性能计算平台资源使用的预测研究[J]. 计算机科学, 2025, 52(9): 178-185.

LI Siqi, YU Kun, CHEN Yuhao. Prediction of Resource Usage on High-performance Computing Platforms Based on ARIMA and LSTM[J]. Computer Science, 2025, 52(9): 178-185.

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)

向量DSP的数组计算高效代码生成技术研究

Research on Efficient Code Generation Techniques for Array Computation for Vector DSPs
计算机科学, 2025, 52(6A): 240300156-7. <https://doi.org/10.11896/jsjcx.240300156>

一种基于产品复用模型的高效遥感共性产品生产算法

Efficient Remote Sensing Common Product Production Algorithm Based on Product Reuse Model
计算机科学, 2025, 52(6): 316-323. <https://doi.org/10.11896/jsjcx.240300019>

并行计时偏差评测指标及工具

Metrics and Tools for Evaluating the Deviation in Parallel Timing
计算机科学, 2025, 52(5): 41-49. <https://doi.org/10.11896/jsjcx.241200053>

面向国产超算的操作系统评测与优化

Performance Evaluation and Optimization of Operating System for Domestic Supercomputer
计算机科学, 2025, 52(5): 11-24. <https://doi.org/10.11896/jsjcx.240500103>

AI+HPC:“智能+”驱动下的超算系统软件及应用技术发展综述

AI+HPC: An Overview of Supercomputing System Software and Application Technology Development Driven by “AI+”
计算机科学, 2025, 52(5): 1-10. <https://doi.org/10.11896/jsjcx.241100177>

基于 ARIMA 和 LSTM 的高性能计算平台资源使用的预测研究

李思琪 俞琨 陈宇皓

华东师范大学数据科学与工程学院 上海 200062

(51275903034@stu.ecnu.edu.cn)

摘要 随着科学研究和工程模拟中数据规模和实验复杂度的不断提升,对高性能计算资源的需求日益增长。然而由于资源有限,如何高效利用现有计算资源成为亟待解决的问题。基于 2022 年 1 月至 2023 年 11 月期间华东师范大学高性能计算中心集群收集的 40 万条作业数据,整理得到集群每日运行作业数和 CPU 资源利用率,以表征计算资源的使用情况。采用 ARIMA 模型、由 LSTM 改进的 2DLSTM 模型和 ARIMA-2DLSTM 组合模型对历史数据进行拟合,实现了对集群计算资源使用情况的长短期预测。通过平均绝对误差(MAE)和均方误差(MSE)指标评估模型预测效果,实验结果表明,ARIMA-2DLSTM 组合模型在预测准确性上优于单独使用 ARIMA 模型和 2DLSTM 模型,且能够精确预测趋势变化以及波峰与低谷出现的时间,为高性能计算中心的资源分配提供了有效支持。

关键词: 高性能计算;时间序列分析;ARIMA 模型;LSTM 模型;ARIMA-LSTM 组合模型

中图分类号 TP183

Prediction of Resource Usage on High-performance Computing Platforms Based on ARIMA and LSTM

LI Siqi, YU Kun and CHEN Yuhao

School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

Abstract The rapid increase in data scale and experimental complexity in scientific research and engineering simulations has significantly heightened the demand for high-performance computing (HPC) resources. However, limited availability of such resources necessitates efficient utilization strategies. This paper analyzes over 400 000 job records collected from the HPC cluster at East China Normal University between January 2022 and November 2023. After preprocessing the data, daily job counts and CPU utilization rates are extracted to represent the cluster's resource usage patterns. Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory with Dual Dropouts (2DLSTM), and a hybrid ARIMA-2DLSTM are applied to fit the historical data and forecast short-term and long-term resource usage. Model performance is evaluated by mean absolute error (MAE) and mean squared error (MSE) metrics. Results indicate that the ARIMA-2DLSTM hybrid model achieves superior predictive accuracy compared to standalone ARIMA and 2DLSTM models. The hybrid model effectively captures trend changes of the cluster's resource usage patterns and accurately predicts peak and trough timings, providing critical insights for optimizing resource allocation in HPC environments.

Keywords High-performance computing, Time series analysis, ARIMA model, LSTM model, ARIMA-LSTM hybrid model

1 引言

近年来,随着高性能计算技术的飞速发展,高性能计算已进入 E 级计算时代^[1]。据《数字中国发展报告(2024 年)》显示,截至 2024 年底,我国算力总规模已达到 280 EFLOPS,年增长率近 22%。越来越多的高校、企业和科研机构将应用部署至高性能计算中心。据统计,2022 年全球高性能计算市场规模已达 2136.98 亿元,预计到 2028 年,将以 8.19% 的增速达到 3427.54 亿元^[2]。然而,高性能计算的广泛应用导致

HPC(High Performance Computing)集群的高能耗问题日益严重。过去十年间,我国高性能计算中心的年用电量增长超过 10%,2020 年已达 2000 亿千瓦时,占全国总用电量的 2.71%^[3]。高性能计算中心庞大的能源需求不仅带来了高昂的运营成本,也对环境造成严重污染和破坏^[4]。

在此背景下,如何优化资源调度策略以提升计算资源利用率、平衡集群能耗,避免高峰期资源拥挤和低谷期资源浪费,成为高性能计算资源管理者亟待解决的问题。

HPC 平台算力资源的使用性能可以通过程序运行过程

到稿日期:2024-11-28 返修日期:2025-01-26

基金项目:国家自然科学基金(62377013)

This work was supported by the National Natural Science Foundation of China(62377013).

通信作者:俞琨(kyu@cc.ecnu.edu.cn)

中的 CPU 资源使用情况、GPU 资源使用情况、用户提交作业数等因素进行定量描述^[5-6]。因此,本文基于华东师范大学高性能计算集群 40 万条作业数据,提取作业申请的资源数量和执行时长等特征,以每日作业数和计算资源利用率表征集群计算资源的使用情况,构建 ARIMA、2DLSTM 和 ARIMA-2DLSTM 组合模型,对集群每日运行作业数和资源利用率进行长短期预测。实验表明,ARIMA-2DLSTM 组合模型能精准捕捉集群每日运行作业数和资源利用率的变化趋势,尤其对峰值和低谷的时间预测更为准确,从而有助于高性能计算中心在短期内优化日常运营,在长期内做好战略规划,全面提升运营效率和服务质量,促进高性能计算技术的可持续发展。

2 相关工作

由于 HPC 平台资源使用情况的各项指标在集群内按照时间顺序收集,因此许多研究者采用统计学方法对时间序列数据进行研究。Kumar 等^[7]构建了多种自回归滑动平均模型(Autoregressive Moving Average Model, ARMA),包括自回归积分移动平均模型(Autoregressive Integrated Moving Average Model, ARIMA)、季节性差分自回归滑动平均模型(Seasonal Autoregressive Integrated Moving Average, SARIMA)、分整自回归移动平均模型(Autoregressive Fractionally Integrated Moving Average, ARFIMA)和奇异谱分析(Singular Spectrum Analysis, SSA)等,用于预测集群中 CPU 和 RAM 的占用情况,结果表明,ARIMA 模型的预测效果显著优于其他复杂的 ARMA 类模型。Bi 等^[8]结合小波分解与 ARIMA 模型,先用 Savitzky-Golay 滤波器平滑时间序列数据,再将其分解为多个分量并分别建立 ARIMA 模型,以预测下一个时间间隔提交到 HPC 集群的任务数量,结果表明,该混合方法的预测结果优于单一 ARIMA 模型。

近年来,随着机器学习技术的迅速发展,研究者逐渐将机器学习方法应用于集群计算资源使用情况的分析和预测工作中。Ipek 等^[9]率先利用复杂神经网络模型研究 HPC 平台运行 SMG2000 并行程序的输入规模与运行时间的关系。Borghesi 等^[10]尝试使用线性回归、支持向量机、决策树等方法,结合 Bagging、AdaBoost 等集成学习算法预测 HPC 系统作业的能耗。

在时间序列预测方面,有研究比较了使用单一的统计学方法或机器学习方法,以及构建混合模型的预测性能。Kon-topoulou 等^[11]的研究表明,混合统计-机器学习模型通常能结合两者优势,在处理复杂数据时表现出更好性能。Shariff^[12]对比了 ARIMA 模型和长短期记忆网络(Long Short-Term Model, LSTM)在电能消耗预测中的表现,发现 LSTM 擅长捕捉数据的长短期依赖性,ARIMA 则更适用于处理线性趋势。Sirisha 等^[13]基于历史销售数据,构建 ARIMA、SARIMA 和 LSTM 模型预测未来销售利润,结果显示,LSTM 模型的预测精度最高(97.01%)。Fan 等^[14]提出 ARIMA-LSTM 混合模型,通过 ARIMA 提取线性趋势并将残差输入 LSTM 捕捉非线性波动,实现对油井产量的精准预测。

基于上述研究现状,本文分别构建 ARIMA 模型、由 LSTM 改进的 2DLSTM 模型和 ARIMA-2DLSTM 组合模型,对集群未来每日运行作业数和资源利用率进行长期(3 个月)

和短期(1 个月)预测。实验结果表明,ARIMA-2DLSTM 模型能够有效捕捉时间序列中的长短期依赖及线性趋势,其预测效果显著优于单独使用 ARIMA 或 2DLSTM 模型的结果,尤其在资源利用的波峰与低谷时间预测上表现优异,展示了其在复杂时间序列预测中的精确度和稳健性,能够帮助高性能计算中心的管理者更有效地分配计算资源,提高资源利用效率。

3 模型构建方法

3.1 滑动窗口构造 ARIMA 和 2DLSTM 模型

3.1.1 滑动窗口构建模型输入输出序列

为提升对时间序列短期依赖与局部模式的捕捉能力,本文采用滑动窗口法构建预测模型的输入和输出序列,即某时刻模型预测值仅取决于前 k 个时间点的历史数据(k 为滑动窗口大小),而非全部历史数据。由于集群资源使用情况随时间推移持续变化,且易受突发事件影响而骤变。因此,采用滑动窗口更新模型的输入序列和输出目标能够使模型持续更新,从而得到更优的预测效果。

如图 1 所示,利用大小为 k 的滑动窗口从时间序列中提取长度为 k 的子序列作为模型的输入特征,将第 $k+1$ 个时间点的值作为目标输出。整个数据集按一个时间步长(本文设为 1 天)滑动窗口,得到多个输入-输出对,构成该预测模型的数据集。

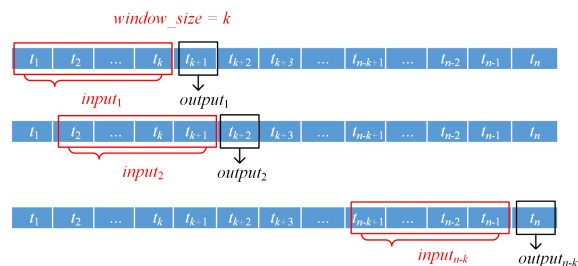


图 1 滑动窗口逻辑图

Fig. 1 Sliding window logic diagram

依据不同预测任务需求,滑动窗口大小 k 的选取存在差异。本文设定 7,14,30,45 天为短期预测可选的滑动窗口大小取值,以捕捉短期时间依赖性;设定 60,90,120,150 天为长期预测可选的滑动窗口大小取值,以捕捉长期变化趋势。基于不同滑动窗口大小构建模型输入输出序列,生成模型数据集,随后按 4:1 的比例将数据集划分为训练集与测试集。通过比较模型在测试集上的平均绝对误差(Mean Squared Error, MSE),确定最优滑动窗口大小^[15]。

针对集群每日运行作业数与资源利用率的长期和短期预测实验,分别确定 ARIMA 与 2DLSTM 模型滑动窗口大小的最优取值,综合两个模型表现,选择整体最优的滑动窗口大小。最终,在短期预测集群每日运行作业数与资源利用率的两组实验中,选择 14 天为滑动窗口大小;在长期预测的两组实验中,选择 90 天为滑动窗口大小,开展后续预测任务。

3.1.2 ARIMA 模型

ARIMA 模型是统计学中常用的时间序列分析方法,由自回归(Autoregressive, AR)、差分(Integrate, I)和移动平均(Moving Average, MA)3 部分组成。其核心是通过分析时间

序列的统计特征,建立模型捕捉线性变化趋势,预测未来发展趋势情况^[16]。

在 ARIMA 模型中,自回归部分揭示当前观测值与历史观测值之间的动态联系;差分部分通过平稳化处理原始时间序列,消除其非平稳性;移动平均部分刻画当前观测值与历史预测误差的关系,利用滞后残差调整预测值^[17]。通过这 3 部分的整合,ARIMA 构建了以自回归阶数 p 、差分次数 d 、移动平均阶数 q 为参数的模型框架。

ARIMA 模型的流程图如图 2 所示。对集群每日运行作业总数和资源利用率构建 ARIMA(p,d,q)模型,首先对时间序列进行 d 阶差分,保证数据平稳性。而后,检验平稳序列是否为非白噪声序列,若为非白噪声序列,则通过自相关函数(Autocorrelation Function, ACF)图和部分自相关函数(Partial Autocorrelation Function, PACF)图确定参数 p 和 q 的取值范围,并结合赤池信息准则(Akaike information criterion, AIC)和贝叶斯信息准则(Bayesian Information Criterion, BIC)热力图进一步优化选择。最终,若模型通过残差检验,即可确定 ARIMA(p,d,q)^[18]。

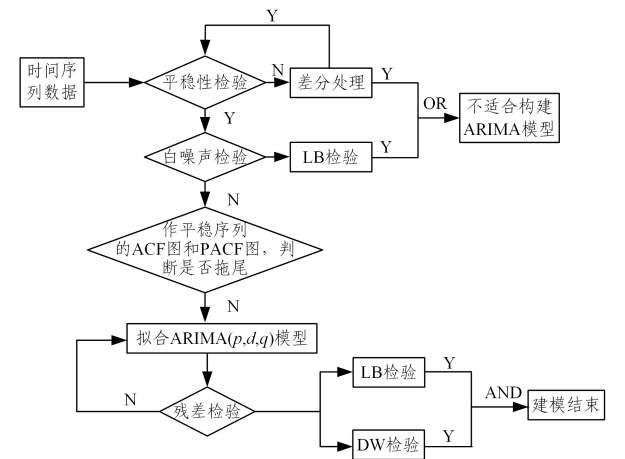


图 2 构建 ARIMA 模型的流程图

Fig. 2 Flowchart of building ARIMA model

3.1.3 LSTM 模型

LSTM 是一种改进的循环神经网络(Recurrent Neural Network, RNN),旨在解决传统 RNN 存在的梯度消失和梯度爆炸问题^[19]。如图 3 所示, LSTM 通过输入门(σ_f)、遗忘门(σ_f)和输出门(σ_o)3 种门控机制实现信息的存储与更新。输入门控制当前输入数据 x_i 进入记忆单元 c_i 的比例;遗忘门决定上一时刻信息的传递或遗忘;输出门调节记忆单元 c_i 中的信息对当前时刻输出的影响。这种门控机制使得 LSTM 能够有效捕获序列数据中的长时间依赖关系,挖掘复杂模式和规律,从而实现更精准的预测和生成任务^[20]。

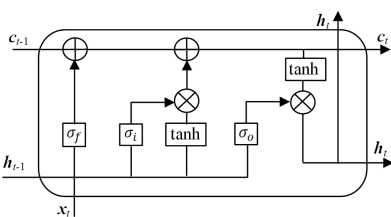


图 3 LSTM 单元结构

Fig. 3 Structure of LSTM unit

3.2 对 LSTM 网络的改进:2DLSTM 结构

LSTM 模型通常由输入层、LSTM 层和输出层 3 部分组成。为提高模型复杂度以更准确地预测集群每日运行作业数 and 资源利用率,同时避免过拟合,本文设计了“输入层-LSTM 层-dropout 层-LSTM 层-dropout 层-输出层”的模型结构,如图 4 所示。在训练阶段,以最小化损失函数为目标优化模型参数;在验证阶段,通过网格搜索优化模型的超参数组合,包括学习率(Learning Rate)、LSTM 单元数(Units)和 dropout 丢弃率(Rate),以平均绝对误差(Mean Absolute Error, MAE)为指标评估各超参数组合的 2DLSTM 模型效果,从中选取最优参数组合。最终,在测试集上验证模型的预测性能与泛化能力。

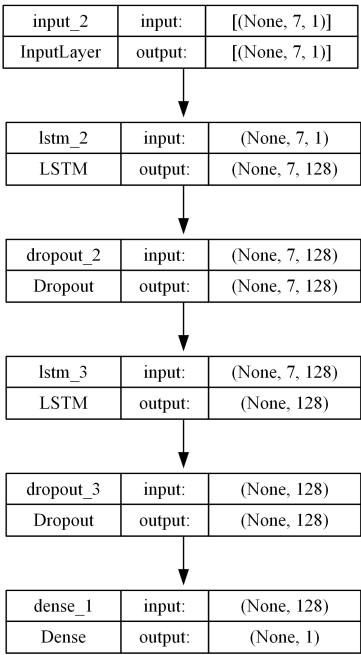


图 4 拟合集群每日运行作业数的 LSTM 模型结构

Fig. 4 LSTM model structure for short-term forecast of future job count

3.3 ARIMA-2DLSTM 组合模型

为充分融合统计模型与深度学习模型在时间序列预测中的优势,本文提出了 ARIMA-2DLSTM 组合模型。该组合模型利用 ARIMA 和 2DLSTM 分别对输入数据建模,通过加权求和整合两者的预测结果,以提升整体预测性能。加权系数依据验证集上两种模型的预测精度确定,从而实现组合模型在性能上的优化与平衡。

4 数据来源与预处理

4.1 HPC 集群数据来源

本文以华东师范大学高性能计算中心集群的真实数据为研究对象,该集群由 224 个计算节点组成。数据集涵盖 2022 年 1 月至 2023 年 11 月期间 75 名用户提交的超过 40 万条 HPC 作业任务的详细信息,如图 5 所示,包括作业 ID(Job_ID)、用户 ID(Log_ID)、申请资源类别(Queue_ID)和数量(CPU_Num)、作业提交时间(Submit_Time)、开始运行时间(Begin_Time)及结束时间(Finish_Time)等。

log_id	job_id	queue_id	queue_name	cpu_num	submit_time	begin_time	finish_time
1641005329140650	140654	29	gpu2	1	2022/1/1 10:48:49	2022/1/1 10:48:49	2022/1/1 10:48:51
1641005468140650	140655	12	normal	1	2022/1/1 10:51:08	2022/1/1 10:51:10	2022/1/1 10:51:54
1641005571140650	140656	29	gpu2	1	2022/1/1 10:52:51	2022/1/1 10:52:52	2022/1/1 10:53:06
1641005738140650	140657	29	gpu2	1	2022/1/1 10:55:38	2022/1/1 10:55:40	2022/1/1 10:55:51
1641007136140650	140658	12	normal	1	2022/1/1 11:18:56	2022/1/1 11:18:58	2022/1/1 11:19:14
1641007246140650	140659	29	gpu2	1	2022/1/1 11:20:46	2022/1/1 11:20:46	2022/1/1 11:21:18
1641007438140660	140660	29	gpu2	1	2022/1/1 11:23:58	2022/1/1 11:23:59	2022/1/1 11:24:11
1641007599140660	140661		special	1	2022/1/1 11:26:39	2022/1/1 11:26:41	2022/1/1 13:43:16
1641007658140660	140662	29	gpu2	1	2022/1/1 11:27:38	2022/1/1 11:27:39	2022/1/1 11:27:51
1641010207140660	140663	12	normal	1	2022/1/1 12:10:07	2022/1/1 12:10:07	2022/1/1 12:10:29
1641010414140660	140664	12	normal	1	2022/1/1 12:13:34	2022/1/1 12:13:36	2022/1/1 12:13:42
1641010983140660	140665	12	normal	1	2022/1/1 12:23:03	2022/1/1 12:23:04	2022/1/1 12:23:09
1641011125140660	140666	12	normal	1	2022/1/1 12:25:25	2022/1/1 12:25:25	2022/1/1 12:25:29
1641011215140660	140667	12	normal	1	2022/1/1 12:26:55	2022/1/1 12:26:55	2022/1/1 12:27:00
1641011302140660	140668	12	normal	1	2022/1/1 12:28:22	2022/1/1 12:28:23	2022/1/1 12:28:28
1641011373140660	140669	13	gpu	1	2022/1/1 12:29:33	2022/1/1 12:29:57	2022/1/1 12:45:36
1641011391140670	140670	12	normal	1	2022/1/1 12:29:51	2022/1/1 12:29:53	2022/1/1 12:29:57
1641011526140670	140671	12	normal	1	2022/1/1 12:32:06	2022/1/1 12:32:06	2022/1/1 12:32:11

图 5 HPC 集群用户提交作业信息

Fig. 5 Job details of the HPC cluster dataset

本文按日期统计集群每日运行作业总数和 CPU 资源利用率,生成 690 条时间序列数据,以表征集群计算资源的使用情况,从执行作业数和 CPU 利用率两个维度对集群的资源使用情况进行分析与预测。

4.2 数据预处理

由于每日提交作业总数和集群资源利用率的时间序列中存在离群值,为提升模型训练与预测的稳健性,本文采用标准分数(Z-score)衡量数据点与均值的偏离程度,将偏离超过 3 倍标准差的数据点视为离群值,并用对应序列的均值替代离群值。如图 6 和图 7 所示,实线表示集群每日运行作业数和计算资源利用率的实际值,虚线表示处理离群值后的结果。

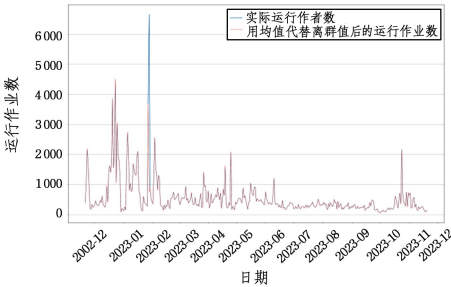


图 6 集群每日运行作业数进行数据预处理后的结果

Fig. 6 Daily job counts in the HPC cluster after data preprocessing

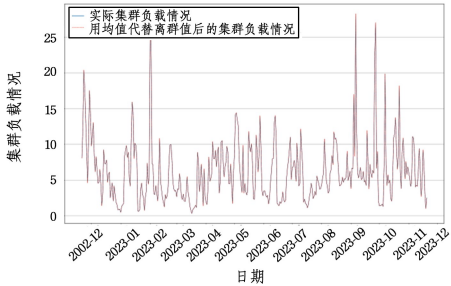


图 7 集群资源利用率进行数据预处理后的结果

Fig. 7 CPU utilization rate in the HPC cluster after data preprocessing

在构建 2DLSTM 模型时,对数据进行归一化处理,以消除特征量纲差异对预测性能的影响,降低梯度消失和梯度爆炸的风险,提升模型的稳定性和泛化能力。此外,通过滑动窗

口生成输入序列及对应输出标签,将原始时间序列转换为适合 2DLSTM 模型输入的格式。

4.3 划分数据集

将数据集按时间顺序划分为训练集、验证集和测试集。训练集用于模型训练,验证集用于超参数调优与模型结构选择,测试集用于评估模型预测值与真实值的差异,以检验其预测性能和准确性。

长期预测任务旨在预测未来一个季度(3 个月)的集群计算资源使用情况。为评估模型的泛化能力与预测效果,保留最后 90 条数据作为预测目标。

为避免验证集的预测目标与测试集的输入序列重叠对预测结果的影响,将预测目标前 90 条数据(对应长期预测的滑动窗口大小)作为测试集的第一个输入序列,而不纳入训练集、验证集或测试集划分。其余 510 条数据按照 7:3 的比例划分为训练集和验证集。

短期预测任务旨在预测未来一个月的集群计算资源使用情况。在 690 条数据中,前 646 条按 4:1 的比例划分为训练集与验证集,最后 30 条作为预测目标。同理,这 30 条数据中的前 14 条数据(对应短期预测的滑动窗口大小)仅用于构成测试集的第一个输入序列,不参与训练、验证或测试集划分。

5 日作业总量的长短期预测

5.1 日作业总量的短期预测

5.1.1 基于 ARIMA 模型的预测

通过 ADF 检验和 KPSS 检验对原始序列及其一阶差分序列的平稳性进行评估,结果如表 1 所列。一阶差分序列的 ADF 检验 p 值小于 0.05, KPSS 检验 p 值大于 0.05,表明其具有平稳性,而原始序列则不平稳。因此,构建 ARIMA 模型时需对原始序列进行一阶差分,模型设为 $ARIMA(p,1,q)$ 。

表 1 每日运行作业数序列的平稳性检验结果		
Table 1 Stationarity test for daily job count time series data		
序列	ADF p -value	KPSS p -value
原始序列	9.939×10^{-12}	0.013428
一阶差分得到的序列	3.517×10^{-16}	0.1

通过 Ljung-Box 检验对集群每日运行作业数的一阶差分序列进行分析,10 阶滞后项的 p 值均小于 0.05,表明该序列

并非白噪声。因此,可基于一阶差分序列构建 ARIMA 模型,以捕捉时间序列中的规律,从而更准确地预测集群每日运行作业数的未来变化趋势。

每日运行作业数的一阶差分序列的 ACF 图和 PACF 图如图 8 所示。两图中均未出现拖尾现象,表明该序列适用于参数 p 和 q 均不为 0 的 ARIMA 模型,且参数 p 取值接近 2,参数 q 取值接近 1。分别设置 p 和 q 的取值范围为 $[0,4]$,对所有可能的参数组合计算 AIC 和 BIC 值,并绘制热力图,结果如图 9 和图 10 所示。

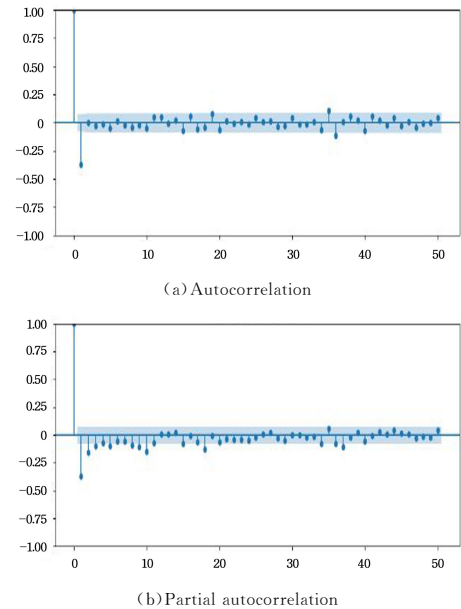


图 8 每日运行作业数序列的 ACF 图和 PACF 图
Fig. 8 ACF and PACF plots of daily job count time series

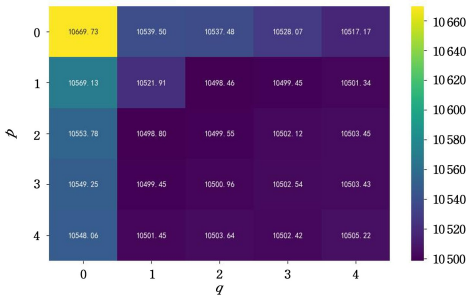


图 9 不同 p, q 组合的 AIC 值热力图
Fig. 9 Heatmap of AIC for different p and q

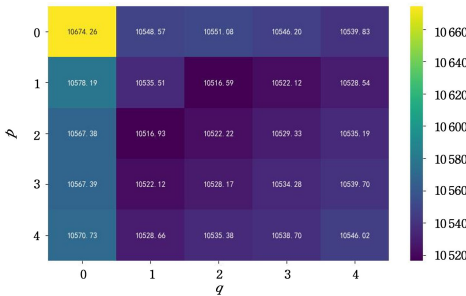


图 10 不同 p, q 组合的 BIC 值热力图
Fig. 10 Heatmap of BIC for different p and q
结果表明,当 (p, q) 分别取值 $(1, 2)$ 和 $(2, 1)$ 时,AIC 和

BIC 值均较小且接近,进一步构建 ARIMA(1,1,2)和 ARIMA(2,1,1)模型,比较其在验证集上的 MAE,分别为 509.433和 499.451,因此选择 ARIMA(2,1,1)构建每日运行作业数的短期预测模型。

基于 ARIMA(2,1,1)模型对数据进行拟合,并通过 Ljung-Box 检验验证残差序列是否为白噪声序列。检验结果如表 2 所列,10 阶滞后项的 p 值均大于 0.05,且残差序列的 Durbin-Watson 检验结果为 2.02,表明残差序列无自相关性,符合白噪声特征,说明 ARIMA(2,1,1)模型有效提取了每日运行作业数的趋势变化信息。

表 2 对 ARIMA(2,1,1)模型残差进行 Ljung-Box 检验结果
Table 2 Ljung-Box test results on residuals for ARIMA(2,1,1) model

lag	lb_stat	lb_pvalue
1	0.056 037	0.812 872
2	0.404 605	0.816 848
3	1.346 498	0.718 122
4	1.592 336	0.810 169
5	1.668 151	0.892 887
6	2.229 109	0.897 468
7	2.284 113	0.942 458
8	3.508 553	0.898 525
9	4.537 496	0.872 621
10	5.187 689	0.878 293

基于处理后的集群每日运行作业数历史数据,拟合 ARIMA(2,1,1)模型,并利用该模型对集群未来每日运行作业数进行短期预测,结果如图 11 所示,其中蓝线表示真实值,红线表示预测值。测试集上的平均绝对误差(MAE)为 170.01,模型摘要如表 3 所列。

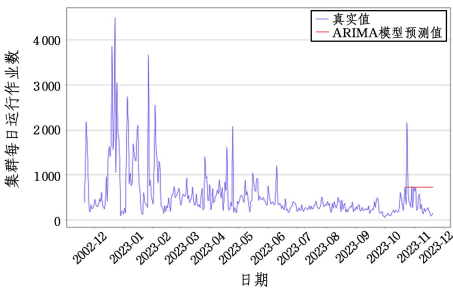


图 11 ARIMA(2,1,1)模型对未来运行作业数进行短期预测
(电子版为彩图)

Fig. 11 Short-term forecast of future job count by ARIMA(2,1,1) model

表 3 ARIMA(2,1,1)模型拟合效果摘要
Table 3 ARIMA(2,1,1) fitting effect summary

	coef	std err	z	$P> z $	$[0.025$	$0.975]$
ar. L1	0.465 8	0.026	18.096	0.00	0.415	0.516
ar. L2	0.203 2	0.027	7.514	0.00	0.150	0.256
ma. L1	-0.979 2	0.010	-95.12	0.00	-0.999	-0.959
sigma2	2.7×10^{-5}	8796.405	30.593	0.00	2.5×10^{-5}	2.8×10^{-5}

分析模型摘要,自回归和移动平均部分的拟合系数均显著。然而,由 ARIMA(2,1,1)模型的短期预测折线图可见,预测结果在某一固定值上下波动。这表明模型可能过于简单,未能有效捕捉数据中的复杂结构与波动特性。因此,进一步尝试采用改进的 LSTM 模型对数据进行拟合。

5.1.2 基于 2DLSTM 模型的预测

如 3.2 节所述,构建结构为“输入层-LSTM 层-dropout

层-LSTM层-dropout层-输出层”的 2DLSTM 模型。基于归一化数据,在训练阶段,以最小化损失函数为目标优化模型参数;通过网格搜索法确定超参数组合,在验证集上以 MAE 为评价指标选择最优配置。最终,利用最优参数组合对未来每日运行作业数进行短期预测,模型在测试集上的 MAE 为 0.1804。

将测试集的预测值和真实值反归一化后,绘制未来 30 天集群每日运行作业数的预测结果,如图 12 所示,其中蓝线表示真实值,红线表示预测值,对应的 MAE 为 134.41。

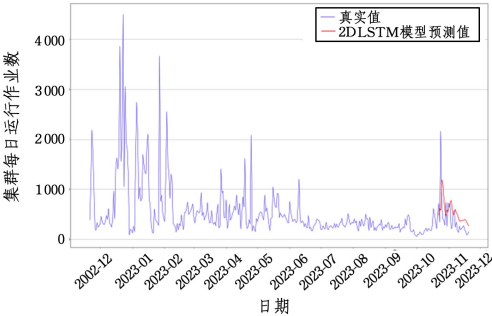


图 12 2DLSTM 模型对未来运行作业数进行短期预测
(电子版为彩图)

Fig. 12 Short-term forecast of future job count by 2DLSTM model

ARIMA 模型和 2DLSTM 模型在测试集上的 MAE 值分别为 170.01 和 134.41,表明 2DLSTM 模型拟合效果更优。通过预测值与真实值的折线图可观察到,2DLSTM 模型的拟合值及其拟合的变化趋势更接近真实情况,但对数据中突增或骤降值的拟合效果欠佳。

5.1.3 ARIMA-2DLSTM 组合模型与单独模型比较

本文利用加权求和构建的 ARIMA-2DLSTM 组合模型在测试集上的 MAE 值为 125.87,预测效果优于单独使用 ARIMA 或 2DLSTM 模型的结果,如图 13 所示。

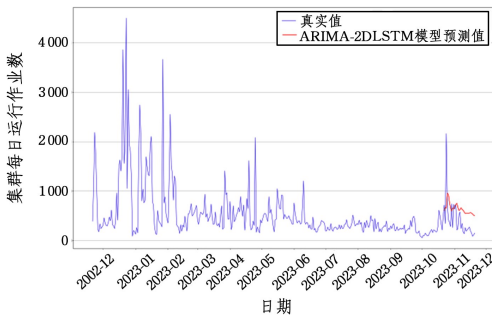


图 13 ARIMA-2DLSTM 模型对未来运行作业数进行短期预测
Fig. 13 Short-term forecast of future job count by ARIMA-2DLSTM model

图 14 整合了 3 种模型在测试集上的预测效果,其中蓝线表示真实值,红线、黄线和绿线分别对应 ARIMA,2DLSTM 和 ARIMA-2DLSTM 组合模型的预测结果。结合表 4 中各模型的 MAE 和 MSE 值可知,ARIMA-2DLSTM 模型的预测效果最佳,且能够较准确地预测集群运行作业数的波峰和低谷时间。

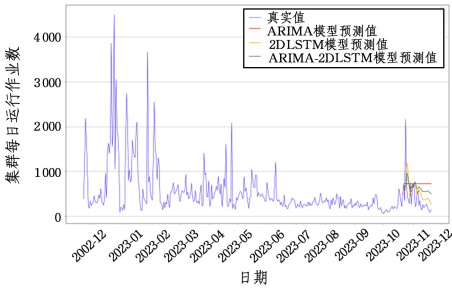


图 14 3 组模型对未来运行作业数进行短期预测对比
(电子版为彩图)

Fig. 14 Short-term forecasting comparison of future job counts by three models

表 4 不同模型短期预测未来运行作业数的效果对比

Table 4 Comparison of short-term forecasting for future job counts by three models

模型	MAE	MSE
ARIMA	170.0137768	279.7471904
2DLSTM	134.4070072	267.9426154
ARIMA-2DLSTM	125.8718333	255.2750526

利用该模型预测未来一个月集群每日运行作业数的变化趋势,为集群管理者优化资源分配和调度策略提供支持,避免资源浪费或短缺。此外,预测的短期趋势变化可能提示近期潜在的系统故障或性能瓶颈,有助于管理者提前制定应对方案。

5.2 日作业总量的长期预测

按照上述建模过程,重新划分数据集对集群每日运行作业数进行长期预测。表 5 列出了 ARIMA,2DLSTM 和 ARIMA-2DLSTM 组合模型的 MAE 和 MSE 值,3 组模型的拟合趋势如图 15 所示。

表 5 3 种模型长期预测未来运行作业数的效果对比

Table 5 Comparison of long-term forecasting for future job counts by three models

模型	MAE	MSE
ARIMA	225.3205351	274.4449281
2DLSTM	110.4949256	210.2895892
ARIMA-2DLSTM	155.0116779	219.9974892

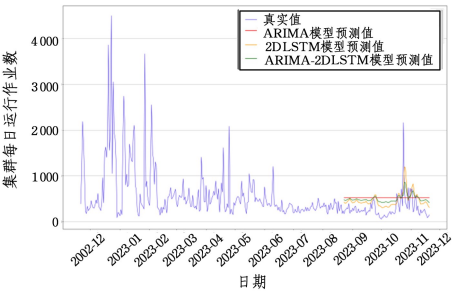


图 15 3 种模型对未来运行作业数进行长期预测对比
Fig. 15 Long-term forecasting comparison of future job counts by three models

结果显示,2DLSTM 模型的预测效果显著优于 ARIMA 模型和组合模型,说明其能够有效捕捉复杂的非线性趋势,而 ARIMA 模型在长期预测中表现较差,难以准确拟合和预测数据变化。

预测 HPC 集群未来 3 个月运行作业数的变化趋势,可为高性能计算中心制定季度资源扩展或调整计划提供支持,确保其资源总量满足科研和商业需求。同时,该预测可为设备升级与扩容等长期投资决策提供数据依据,助力资源配置的合理优化。

5.3 小结

通过对比上述长短期预测模型的 MAE 值可知,2DLSTM 模型在长期预测中表现优异,为集群季度规划提供了有效参考;而 ARIMA 模型虽未精准捕捉变化趋势,但在短期预测中部分抵消了 2DLSTM 模型的误差。结合每日运行作业数的短期和长期预测(见图 14 和图 15)中最后一个月的预测效果可知,短期预测更易受临近数据影响,而长期预测在趋势捕捉上更准确,尤其在峰值和低谷的时间预测方面。总体而言,短期预测注重即时运营效率,长期预测关注未来发展规划,两者结合可显著提升 HPC 集群的管理水平与资源利用率。

6 集群资源利用率的长短期预测

6.1 集群利用率的短期预测

利用集群当天的 CPU 利用率表示资源利用率,分别构建 ARIMA,2DLSTM 和 ARIMA-2DLSTM 模型对其进行短期预测。3 组模型的拟合趋势如图 16 所示,拟合效果的 MAE 和 MSE 值如表 6 所列。结果表明,ARIMA-2DLSTM 组合模型在测试集上的 MAE 和 MSE 均表现最优,说明加权组合 ARIMA 和 2DLSTM 模型的预测效果有效提升了短期预测的准确性。

表 6 3 种模型短期预测未来集群资源利用率的效果对比
Table 6 Comparison of short-term forecasting for future CPU utilization rate by three models

模型	MAE	MSE
ARIMA	3.116610231	24.1820589
2DLSTM	3.284649505	24.9540936
ARIMA-2DLSTM	3.085268496	23.1830871

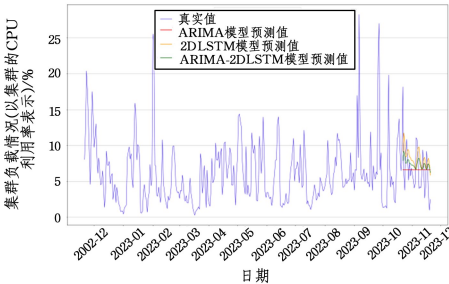


图 16 3 种模型对未来集群资源利用率进行短期预测对比
Fig. 16 Short-term forecasting comparison of future CPU utilization rate by three models

对未来一个月的集群资源利用率进行预测,有助于集群管理者实时优化资源分配与调度,从而提升系统性能与响应速度,避免资源浪费。此外,该预测还可帮助识别潜在的负载高峰期,便于管理者进行有针对性的维护,减少集群过载导致的故障和宕机。

6.2 集群利用率的长期预测

利用 ARIMA,2DLSTM 和组合模型对集群资源利用率进行长期预测,得到的 MAE 和 MSE 值如表 7 所列,预测结果的折线图如图 17 所示。在拟合效果上,ARIMA-2DLSTM 组合模型表现最佳,其次为 2DLSTM 模型,而 ARIMA 模型的拟合效果较差。

表 7 不同模型长期预测未来集群资源利用率的效果对比
Table 7 Comparison of long-term forecasting for future CPU utilization rate by three models

模型	MAE	MSE
ARIMA	3.074284912	18.15715092
2DLSTM	3.584188695	20.88338542
ARIMA-2DLSTM	3.207889831	17.97611990

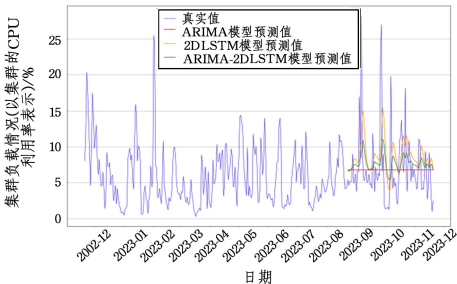


图 17 3 种模型对未来集群资源利用率进行长期预测对比
Fig. 17 Long-term forecasting comparison of future CPU utilization rate by three models

预测未来 3 个月集群资源利用率变化趋势的研究,为集群管理者合理规划集群容量提供了依据,确保在未来 3 个月内能够应对预期负载。同时,高性能计算中心可据此制定能效优化计划,减少能源消耗,提升集群的整体效率。

6.3 小结

根据表 6 和表 7 的 MAE 和 MSE 值,ARIMA 模型在对集群资源利用率的长短期预测中误差较低,但其预测趋势稳定于某一固定值附近,未能有效捕捉集群资源利用率的变化趋势(见图 16 和图 17)。相比之下,2DLSTM 模型的误差略高,但能够更好地适应非线性波动。而 ARIMA-2DLSTM 组合模型则在短期和长期预测中均表现出较为平衡的性能,结合了 ARIMA 模型的稳定性和 2DLSTM 模型对复杂波动的适应性,提升了预测精度。

结束语 本文基于华东师范大学高性能计算中心的真实数据,构建集群每日运行作业数和资源利用率的时间序列,表征集群每日计算资源使用情况。分别利用统计学中的 ARIMA 模型与机器学习领域的 LSTM 模型改进而来的 2DLSTM 模型拟合历史数据,预测未来集群的资源使用情况。鉴于集群资源使用模式的复杂性,本文引入 ARIMA-2DLSTM 组合模型,显著提高了预测准确性与稳健性,能够有效预测计算资源使用的峰值和低谷时间,从而帮助高性能计算中心预估未来集群资源的使用情况,合理分配计算资源。

致谢 感谢华东师范大学公共创新服务平台高性能计算中心为本研究提供数据与算力支持。

参 考 文 献

[1] LU P J,XIONG Z Y,LAI M C. Survey on High-performance Computing Technology and Standards[J]. Computer Science, 2023,50(11):1-7.

[2] Editorial Department of this Journal. Top 10 Frontiers of Science and Technology in 2023 [J]. Science and Technology Think Tank, 2023(1):1-4.

[3] WANG J Y,ZHOU B Y,ZHANG F,et al. Overview of Energy Consumption Models and Energy Efficiency Algorithms in Data Centers [J]. Journal of Computer Research and Development, 2019,56(8):1587-1603.

[4] CHEN Y J. Energy Consumption Analysis and Power Consumption Prediction of High-Performance Computing Clusters [D]. Jinan:Qilu University of Technology, 2023.

[5] SUN J W. Data-Driven Research on Execution Time Prediction and Optimization for High-Performance Computing Programs [D]. Hefei: University of Science and Technology of China, 2020.

[6] ZHANG J P. Research on Active Learning Methods for Empirical Performance Modeling of High-Performance Computing Programs [D]. Hefei:University of Science and Technology of China, 2020.

[7] KUMAR A S,MAZUMDAR S. Forecasting HPC workload using ARMA models and SSA[C]// 2016 International Conference on Information Technology(ICIT). IEEE, 2016:294-297.

[8] BI J,ZHANG L,YUAN H,et al. Hybrid task prediction based on wavelet decomposition and ARIMA model in cloud data center[C]// 2018 IEEE 15th International Conference on Networking, Sensing and Control(ICNSC). IEEE, 2018:1-6.

[9] IPEK E,DE SUPINSKI B R,SCHULZ M,et al. An approach to performance prediction for parallel applications[C]// Euro-Par 2005 Parallel Processing: 11th International Euro-Par Conference, Lisbon, Portugal, August 30-September 2, 2005. Proceedings 11. Berlin:Springer, 2005:196-205.

[10] BORGHESI A,BARTOLINI A,LOMBARDI M,et al. Predictive modeling for job power consumption in HPC systems[C]// High Performance Computing: 31st International Conference, ISC High Performance 2016, Frankfurt, Germany, June 19-23, 2016, Proceedings. Springer, 2016:181-199.

[11] KONTOPOULOU V I,PANAGOPOULOS A D,KAKKOS I,et al. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks[J]. Future Internet, 2023,15(8):255.

[12] SHARIFF S M. Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) network models for forecasting energy consumptions[J]. European Journal of Electrical Engineering and Computer Science, 2022,6(3): 7-10.

[13] SIRISHA U M,BELAVAGI M C,ATTIGERI G. Profit prediction using ARIMA, SARIMA and LSTM models in time series forecasting: A comparison[J]. IEEE Access, 2022, 10: 124715-124727.

[14] FAN D,SUN H,YAO J,et al. Well production forecasting based on ARIMA-LSTM model considering manual operations [J]. Energy, 2021,220:119708.

[15] SELVIN S,VINAYAKUMAR R,GOPALAKRISHNAN E A,et al. Stock price prediction using LSTM,RNN and CNN-sliding window model[C]// 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2017:1643-1647.

[16] SCHAFFER A L,DOBBINS T A,PEARSON S A. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions[J]. BMC Medical Research Methodology, 2021, 21:1-12.

[17] BOX G E P,JENKINS G M,REINSEL G C,et al. Time series analysis: forecasting and control[M]. John Wiley & Sons, 2015: 88-126.

[18] ARUNKUMAR K E,KALAGA D V,KUMAR C M S,et al. Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA) [J]. Applied Soft Computing, 2021,103:107161.

[19] HOCHREITER S,SCHMIDHUBER J. Long Short-term Memory[J]. Neural Computation, 1997,9(8):1735-1780.

[20] CHOUDAKKANAVAR G,MANGAI J A. A hybrid 1D-CNN-Bi-LSTM based model with spatial dropout for multiple fault diagnosis of roller bearing[J]. International Journal of Advanced Computer Science and Applications, 2022,13(8):0130873.

LI Siqi, born in 2002, postgraduate, is a member of CCF (No. V2924G). Her main research interests include high-performance computing and data analysis.

YU Kun, born in 1979, master, engineer, is a member of CCF (No. I8213M). Her main research interests include high-performance computing, computational resource management and optimization.

(责任编辑:柯颖)