基于大数据的云数据中心智能运维系统

张 颖

（华东交通大学 网络信息中心，江西 南昌 330013）

摘 要： 针对云计算中心容量趋势无法预测和异常无法监测的问题，基于ARIMA模型的时间序列算法和Isolation Forest的异常检测算法，提出一种以大数据分析为基础对云计算中心进行智能运维的方法。实验表明，基于Isolation Forest的服务器异常检测方法准确率约为88%，基于ARIMA模型的磁盘容量预测平均绝对误差为0.158 6，证明了该系统相较于传统运维平台具有更好的稳定性、健壮性及服务能力。

关键词： 云计算中心；智慧运维；孤立森林；ARIMA模型；趋势预测；异常检测

DOI： 10. 11907/rjdk. 231188 **开放科学（资源服务）标识码（OSID）：**

中图分类号： TP311.13 **文献标识码：** A **文章编号：** 1672-7800(2024)011-0153-05

Intelligent Operation and Maintenance System of Cloud Data Center Based on Big Data

ZHANG Ying

（*Network & Information Center*, *East China Jiaotong University*, *Nanchang* 330013, *China*）

Abstract: Aiming at the problem of unpredictable capacity trends and inability to monitor anomalies in cloud computing centers, a method for intelligent operation and maintenance of cloud computing centers based on big data analysis is proposed using ARIMA time series algorithm and Isolation Forest anomaly detection algorithm. The experiment shows that the accuracy of the server anomaly detection method based on Isolation Forest is about 88%, and the average absolute error of disk capacity prediction based on ARIMA model is 0.158 6, proving that this system has better stability, robustness, and service capabilities compared to traditional operation and maintenance platforms.

Key Words: cloud computing center; intelligent operation and maintenance; isolation forest; ARIMA model; trend prediction; anomaly detection

0 引言

随着智慧校园建设不断深入，高校对计算、存储等资源的需求呈增长趋势，而云计算作为当前数据中心的主要服务模式，可提供计算、存储等资源[1]。随着数据中心规模扩大、设备数量增加，传统被动等待式运维方式已无法满足当前运维需求，因此如何保障云数据中心稳定运行、健壮性及其服务能力已成为当前需要迫切解决的问题。

为了解决这一问题，本文引入大数据相关技术，通过建模、分析云计算中心服务器的性能数据，实现预见式和主动式的运维管理生态，并将其应用到高校云数据中心的运维管理中，对实际运维具有重要意义。

1 相关研究

随着互联网快速发展，大数据在云数据中心智能运维方面的应用成为了研究热点。国内外学者和企业进行了深入研究，主要关注云数据中心的性能监控、故障诊断、容量规划、任务调度和能耗管理等方面。Malik等[2]结合混合遗传算法和粒子群优化的功能链接神经网络预测云数据中心多资源，并在Google集群中得到应用。胡小宁[3]构建了基于GRU循环神经网络的应用故障预测模型，分析处理监控数据预测应用故障。卢洪明等[4]采用随机森林算法和网格搜索方法构建模型预测服务器能耗。孙湛冬等[5]提出一种基于随机Petri网的任务调度模型。来风刚

收稿日期： 2023-03-01
扫描二维码阅读全文：
基金项目： 江西省教育厅科技研究项目（GJJ191660）

作者简介： 张颖（1987-），女，硕士，华东交通大学网络信息中心工程师，研究方向为数据中心建设、大数据。

等[6]设计了基于门控循环单元的深度学习框架预测机房设备故障。廖恩红等[7]提出一种基于改进时序卷积网络模型预测云服务器性能。清华大学研究团队开发了一种基于机器学习的云数据中心故障预测系统,实现了对故障发生的预测和预警[8]。

此外,阿里巴巴推出一款基于大数据和人工智能的云数据中心运维管理系统,极大提升了运维效率和管理水平[9]。由于ARIMA或Isolation Forest在智能运维方面应用不多,本文结合ARIMA与Isolation Forest算法构建了一个智能运维系统分析、预测云数据中心的运行状况,并对异常情况进行识别和处理。

2 智能运维系统整体框架

智能运维系统主要包括服务器性能数据采集、数据预处理、基于时间序列算法与异常检测算法的大数据分析、大数据预警4个阶段,系统框架如图1所示。由此可见,服务器性能数据采集是智能运维系统的首要任务,只有具备了海量服务器性能数据,才能分析、挖掘和探索数据之间潜在的关系和规律;数据预处理主要对采集的服务器性能数据进行去重、属性重构、过滤、替换、关联等,实现数据标准化;大数据分析主要将数据按照不同维度,采用对应算法(异常检测、关联分析、时间序列等)挖掘主题,从而实现精准预测、异常预警等;在应用层主要通过预警程序提前发现、干预和处理问题与故障。

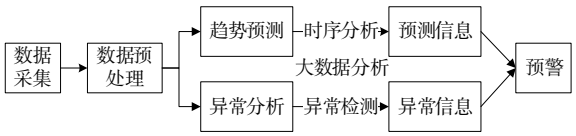

Fig. 1 Intelligent operation and maintenance system framework based on big data

图1 基于大数据的智能运维系统框架

3 相关模型

3.1 ARIMA模型

ARIMA模型是ARMA模型的延伸和补充,实质是差分运算+ARMA模型,即将原始时序数据进行d次一阶差分后拟合ARMA模型,可表示为ARIMA(p,d,q)=AR(p)+MA(q)+Difference(d)[10-12]。其中,AR(p)为p阶自回归过程;MA(q)为q阶移动平均过程;Difference(d)为差分模型。

ARIMA(p,d,q)相关概念和数学表示为:

(1)自回归过程AR(p)。

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + u_t \tag{1}$$

式中:φ_i为自回归参数;u_t为白噪声过程。

(2)移动平均过程MA(q)。

$$x_t = \sum_{i=1}^{q} \theta_i u_{t-i} + u_t \tag{2}$$

式中:θ_i为自回归参数。

(3)自回归移动平均过程ARMA(p,q)。

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + u_t + \sum_{i=1}^{q} \theta_i u_{t-i} \tag{3}$$

(4)一阶差分和二次一阶差分。

$$\nabla x_t = x_t - x_{t-1}, \nabla^2 x_t = \nabla x_t - \nabla x_{t-1} \tag{4}$$

由此计算d次一阶差分,即$\nabla^d x_t$。

3.2 孤立森林模型

孤立森林是一种适用于连续数据的高效异常检测方法,通过递归地随机分割大规模数据集的策略得到的异常点路径通常较短[13-15]。相关概念和数学表示如下:①孤立树(iTree)节点要么是叶子节点,要么是只有两个子节点(T_l,T_r)的内部节点,每次分割都包含特征q和分割值p,将$q<p$数据分到左节点T_l,将$q\geqslant p$的数据分到右节点T_r;②二叉树森林(iForest)由t棵孤立树iTree组成二叉树森林;③样本x的路径长度$h(x)$为从根节点到x所在节点经历的总边数;④所有数据样本期望值$E(h(x))$;⑤n个样本的二叉树,树的平均路径长度$C(n)$;⑥数据点x的异常指数$s(x,n)$。

$$E\left(h(x)\right) = \sum_{i=0}^{t} h(x) \tag{5}$$

$$C(n) = 2\mathrm{H}(n-1) - \frac{2(n-1)}{n} \tag{6}$$

$$s(x,n) = 2^{\frac{-E(h(x))}{C(n)}} \tag{7}$$

式中:H(i)为调和数,约为ln(i)+0.5 772 156 649(欧拉常数)。

4 智能运维系统主要功能实现

4.1 容量预测

容量预测主要通过建立资源预测模型,依据现有资源时序数据预测未来一段时间内的数据序列。本文以云数据中心磁盘容量预测为例,通过现有磁盘使用数据预测未来一段时间内磁盘的使用情况,以便管理员从整体预判磁盘运行情况,并根据预测结果制定相应的应急方案,例如删除僵尸虚拟机或制定相关扩容方案等,以有效避免因磁盘容量不足而导致业务系统不稳定、不连续等问题。

本文采用ARIMA模型预测磁盘使用情况趋势的基本思路为:首先,将原始磁盘使用时序数据进行平稳性检验,若处于非平稳状态则进行差分运算直至平稳为止;其次,评估模型参数选择最优模型,以预测未来一段时间内的磁盘使用数据序列[16]。图2给出了基于ARIMA模型的的磁盘容量趋势预测流程图。具体的趋势预测流程主要分析步骤如下:

步骤1:平稳性检测。采用单位根检验方法对标准化后的磁盘容量时序数据进行平稳性检测,如果序列为非平稳则进行差分运算直至平稳。

步骤2:白噪声检验。若未通过该检验则说明有用信

息没有被提取完毕,将进行下一步;否则流程结束。

步骤3:参数估计。采用BIC信息准则估计模型参数,确定最优p、q参数。

步骤4:模型评价。依托模型计算实际值与预测值的平均绝对误差、均方根误差和平均绝对百分误差判定模型是否有效,如果超出预先设定值则重新进行步骤3。

步骤5:模型应用。根据现有磁盘容量序预测未来一周磁盘的使用情况。

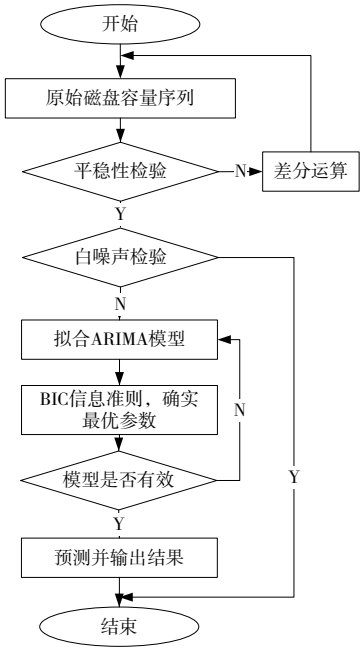

Fig. 2　Trend prediction process

图2　趋势预测流程

4.2　异常检测

异常分析主要通过异常检测算法从海量服务器性能时序数据中发现异常数据,运维人员根据异常预警信息事前排查和处理故障,然后主动防御,从而有效降低故障率。常用的异常检测方法包括基于距离、聚类和预测模型[17]。本文采用基于距离的孤立森林异常检测算法,基本思路是反复随机分割大规模服务器性能数据集,直到所有服务器性能数据孤立,从而孤立出所有异常数据[18]。图3为基于Isolation Forest的服务器异常检测流程图。主要分析步骤如下:

步骤1:采集子样本。对于大规模服务器性能数据集$\boldsymbol{D}$,随机采样一定数量的服务器性能数据得到n个子样本集。

步骤2:构造二叉树iTree。随机选取子样本集中某一属性q,随机选择一个分割点p,其中$p \in (q_{\min}, q_{\max})$,若$q < p$:则将样本$x$归为左子树,反之归为右子树,以此递归直至样本集的最后一条数据或达到树的最大高度。

步骤3:构造孤立森林。重复步骤2构建$t-1$棵树,将t棵iTree合并为孤立森林。

步骤4:计算高度平均值。对每一样本数据x遍历每一颗iTree,求x到该树根节点的距离$h(x)$,并根据式(5)、式(6)计算高度平均值。

步骤5:计算异常指数。依据式(5)—式(7)计算异常指数以判断x是否异常,若$s(x,n) \to 1$表示x为异常的可能性越大;若$s(x,n) \to 0$则x为正常的可能性越大。

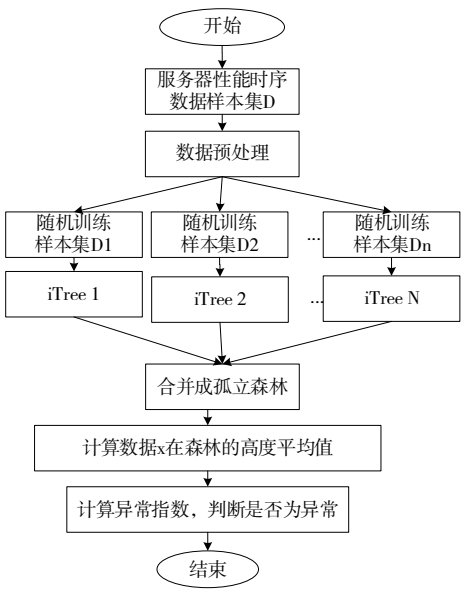

Fig. 3　Abnormal analysis process

图3　异常分析流程

5　实验结果与分析

本文实验操作系统为Window 10,硬件设备为Intel(R) Core(TM) i7-6700 CPU、3.4 GHz,8 GB内存,算法编程语言为Python 3.6。通过采集服务器性能时序数据进行评价,包括CPU使用率、磁盘使用率、内存使用率、CPU已使用容量、磁盘已使用容量、内存已使用容量。

5.1　异常检测算法有效性验证

为了验证Isolation Forest服务器异常检测算法的有效性,选取2020/01/01-2021/02/12期间的服务器性能样本数9 265个。磁盘容量异常检测图如图4所示。对本文算法并与Statistics算法[19]、Frequency算法[20]进行比较,主要对磁盘使用率、CPU使用率和内存使用率3个字段进行检测,采用异常点个数、准确率和误报率作为检验算法的标准,实验结果如表1所示。实验前,已知实验期间服务器共发生了7次异常,由此可知基于Isolation Forest算法检测到的异常点数量最接近实际数量,准确率较高。

5.2　预测算法有效性验证

为了验证ARIMA算法的可行性,采集2021/02/01-2021/03/20时间段的服务器磁盘数据,预测未来一周的利用率。图5为算法的磁盘使用率预测效果,真实值、预测值和相对误差如表2所示。

由图5可见,基于ARIMA方法预测的磁盘使用率与实

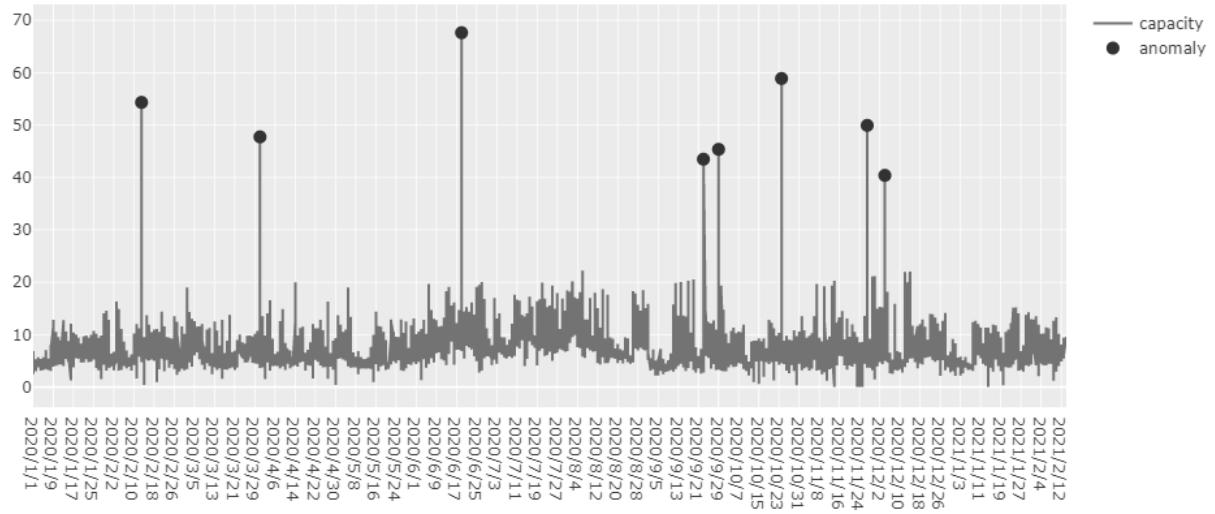

Fig. 4 Disk capacity anomaly detection

图 4 磁盘容量异常检测

Table 1 Experimental comparison of anomaly detection algorithms

表 1 异常检测算法比较

算法	异常点数量	准确率/%
Isolation Forest	8	88
Statistics	11	64
Frequency	13	54
BPNN	10	70
RF	9	77

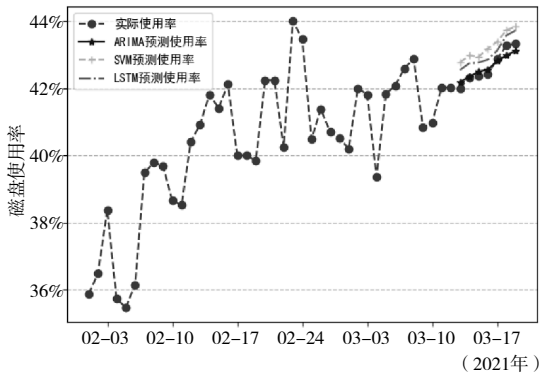

Fig. 5 Comparison of disk utilization prediction

图 5 磁盘使用率预测比较

Table 2 Comparison of disk capacity

表 2 磁盘容量比较

日期	实际值	预测值	相对误差
2021/3/13	42.00	42.18	−0.18
2021/3/14	42.32	42.34	−0.02
2021/3/15	42.37	42.49	−0.13
2021/3/16	42.43	42.55	−0.12
2021/3/17	42.86	42.81	0.05
2021/3/18	43.28	42.97	0.31
2021/3/19	43.33	43.12	0.21

际使用率曲线拟合效果最好。由表2可知，基于ARIMA方法预测的磁盘容量真实值与磁盘容量预测值之间的差值较小，平均绝对误差为0.158 6，均方根误差为0.184 5，平均绝对百分误差为0.370 7。综上，本文所提方法可较准确预测未来趋势，对容量趋势预测具有一定的指导意义。

6 结语

本文提出基于ARIMA模型和Isolation Forest算法的智能运维系统，不仅适用于数据中心，还能应用于电力设备、供应链管理、交通运输和智能建筑等领域。该系统可帮助企业实现监测分析、预测、检测各种设备的状态与性能，提高了生产效率。本文系统虽然具有较高的准确性和效率，但由于数据样本过少、不平衡和分布不一致等因素也存在一定的局限性。

未来，将通过增加数据样本、优化模型参数和算法扩展性等措施，进一步提升系统准确性和泛化能力，使模型适应更多应用场景；还可引入实时机器学习技术（集成学习、深度学习等）提供更精确、高效的技术支撑，从而推进智能运维系统的实践应用和发展。

参考文献：

[1] WANG B F, SU J S, CHEN L. Review of the design of data center network for cloud computing[J]. Journal of Computer Research and Development, 2016, 53(9): 2085-2106.
王斌锋，苏金树，陈琳. 云计算数据中心网络设计综述. 计算机研究与发展，2016，53(9)：2085-2106.

[2] MALIK S, TAHIR M, SARDARAZ M, et al. A resource utilization prediction model for cloud data centers using evolutionary algorithms and machine learning techniques[J]. Applied Sciences, 2022, 12(4): 2160.

[3] HU X N. Application failure prediction method of cloud data center based on GRU recurrent neural network [J]. Railway Computer Application, 2022, 31(2): 7-11.
胡小宁. 基于GRU循环神经网络的云数据中心应用故障预测方法[J]. 铁路计算机应用，2022，31(2)：7-11.

[4] LU H M, LIU X F, ZHOU Z, et al. Research on energy consumption model of cloud data center based on a machine learing method[J]. Journal of Chinese Computer Systems, 2023, 44(9): 1-10.

卢洪明,刘先锋,周舟,等. 机器学习方法的云数据中心能耗模型研究[J]. 小型微型计算机系统,2023,44(9):1-10.

[5] SUN Z D, JIAO J, LI W, et al. A task scheduling strategy for a power cloud data center based on an improved ant colony algorithm[J]. Power System Protection and Control,2022,50(2):95-101.
孙湛冬,焦娇,李伟,等. 基于改进蚁群算法的电力云数据中心任务调度策略研究[J]. 电力系统保护与控制,2022,50(2):95-101.

[6] LAI F G, LIU J, LI J W, et al. Cloud data center fault detection based on Pytorch and neural network[J]. Computer Systems & Applications, 2020, 29(11):40-46.
来风刚,刘军,李济伟,等. 基于Pytorch和神经网络的云数据中心故障检测[J]. 计算机系统应用,2020,29(11):40-46.

[7] LIAO E H, SHU N, LI J W, et al. A prediction model of cloud server performance based on temporal convolutional network[J]. Journal of South China Normal University (Natural Science Edition), 2020, 52(4): 107-113.
廖恩红,舒娜,李加伟,等. 基于时序卷积网络的云服务器性能预测模型[J]. 华南师范大学学报(自然科学版),2020,52(4):107-113.

[8] YU R, JIN H, YANG X, et al. Machine learning based cloud data center fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2018, 14(9): 4003-4011.

[9] WANG Y, HUANG K, TAN K, et al. A cloud data center operations and maintenance management system based on big data and artificial intelligence[J]. IEEE Network, 2019,33(6): 32-39.

[10] WANG E, ZHANG T. Application of time series in GDP forecasting of human province —— based on ARIMA model[J]. Journal of Qingdao University(Natural Science Edition), 2019,32(3):136-140.
王鄂,张霆. 时间序列在湖南省GDP预测中的应用——基于ARIMA模型[J]. 青岛大学学报(自然科学版),2019,32(3):136-140.

[11] YUAN Y, GUO T T. Research on apparel sales forecast based on ARIMA-RF combination model[J]. Software Guide,2021,20(9):33-38.
袁远,郭天添. ARIMA-RF组合模型的销售预测研究[J]. 软件导刊,2021,20(9):33-38.

[12] CAO H, QIN J T. Research on freight volume prediction based on ARIMA-BP combination model[J]. Software Guide,2022,21(2):32-36.
曹慧,秦江涛. 基于ARIMA-BP组合模型的货运量预测研究[J]. 软件导刊,2022,21(2):32-36.

[13] ARYAL S, TING K M, WELLS J R, et al. Improving iForest with relative mass[C]// Proceedings of the 18th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining,2014: 510-521.

[14] XIONG K, DING Q, ZHU H M. Fault detection for chiller based on isolated forest and KDE-LOF[J]. Computer Applications and Software, 2023,40(1):84-89.
熊坤,丁强,祝红梅. 基于孤立森林与KDE-LOF的冷水机组故障检测[J]. 计算机应用与软件,2023,40(1):84-89.

[15] SUN R S, HAN S H. Ultra limit incident prediction of flight approach based on isolation forest[J]. Journal of Safety and Environment, 2022, 22(4):2010-2016.
孙瑞山,韩韶华. 基于孤立森林的航班进近超限事件预测[J]. 安全与环境学报,2022,22(4):2010-2016.

[16] WANG Y W, MA S C. Time series forecasting based on ARIMA_DLSTM hybrid model[J]. Computer Applications and Software, 2021, 38(2):291-298.
王英伟,马树才. 基于ARIMA和LSTM混合模型的时间序列预测[J]. 计算机应用与软件,2021,38(2):291-298.

[17] GAO Y F, WANG J P, LI L F. Outlier detection method of hydrological time series based on Cauchy distribution[J]. Journal of Hohai University(Natural Sciences),2020,48(4):307-313.
高熠飞,王建平,李林峰. 基于柯西分布的水文序列异常值检测方法[J]. 河海大学学报(自然科学版),2020,48(4):307-313.

[18] XIAO W Y. Anomaly detection and analysis of air quality data based on isolation forest algorithm[J]. China Computer & Communication, 2019, 31(17):38-40.
肖伟洋. 基于孤立森林算法的空气质量数据异常检测分析[J]. 信息与电脑(理论版),2019,31(17):38-40.

[19] CAO C X, TIAN Y L, ZHANG Y K, et al. Application of statistical methods in outlier detection for time series data[J]. Journal of Hefei University of Technology(Natural Science),2018,41(9):1284-1288.
曹晨曦,田友琳,张昱堃,等. 基于统计方法的异常点检测在时间序列数据上的应用[J]. 合肥工业大学学报(自然科学版),2018,41(9):1284-1288.

[20] LI H L, WU X L. Time series anomaly detection method based on frequent pattern discovery[J]. Journal of Computer Applications, 2018, 38(11):3204-3210.
李海林,邬先利. 基于频繁模式发现的时间序列异常检测方法[J]. 计算机应用,2018,38(11):3204-3210.

(责任编辑:刘嘉文)