A Brief Review on Prediction Methods for Cloud Resource Management

Chenxing Kuang, Yunyun Qiu, Weipeng Cao, Zhijiao Xiao and Zhong Ming

November 22, 2023

A Brief Review on Prediction Methods for Cloud Resource Management

Chenxing Kuang[1], Yunyun Qiu[1], Weipeng Cao[2(✉)], Zhijiao Xiao[1(✉)], and Zhong Ming[1,2]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China 518060
cindyxzj@szu.edu.cn
[2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen), Shenzhen, China 518107
caoweipeng@gml.ac.cn

Abstract. Nowadays, managing cloud resources appropriately is undoubtedly crucial to fully utilize computing resources in the cloud environment, and improve service quality. An effective resource management strategy is to forecast potential future cloud resource demands or abnormal events in order to schedule resources in advance. This study provides a brief overview of these resource prediction methods from the view of different prediction objects.

Keywords: Cloud computing · Prediction model · Resource management · Cloud brain.

1 INTRODUCTION

Cloud data centers suffer from a high level of resource waste, with server utilization rate typically below 30%[1]. In addition, abnormal events such as hard disk failures and software failures occur in cloud computing environments, leading to downtime from various sources of failure[2]. These issues can seriously impact the quality of service (QoS), causing unnecessary waste of resource. Therefore, accurate prediction of resource utilization and abnormal events in cloud data centers is essential for capacity planning[3], resource management[4, 5], and energy efficiency[6].

While several surveys have been conducted on prediction methods in cloud environments, they have mostly examined models that can only forecast one particular type of object. For instance, Amiri et al. [7] provided a thorough overview of the literature on application prediction models, but did not classify the predicted items in great depth. Aldossary et al. [8] focused on predictive models related to workload, energy consumption, and cost of cloud services, with a specific emphasis on energy-related cost issues in cloud computing. Similarly, Vashistha et al. [9] reviewed only prediction techniques for workloads in cloud environments, while Ramoliya et al. [10] reviewed only failure and fault prediction techniques. In contrast, this paper provides a comprehensive overview of the research on prediction models for forecasting workloads, resource requirements, QoS metrics, and abnormal events.

Figure 1 illustrates different categories objects in prediction research for cloud resource management. The paper includes a brief review of recent prediction models for different objects in Section II and Section III identifies common challenges and suggests future research directions.

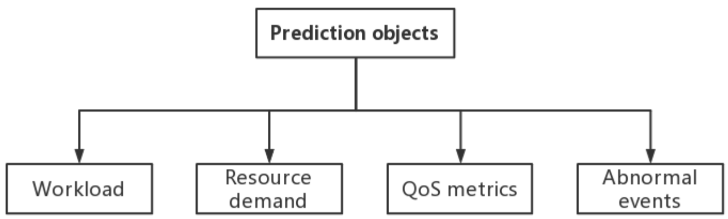

Fig. 1. Main objects for the prediction research in cloud resource management.

2 PREDICTION METHODS FOR DIFFERENT PREDICTION OBJECTS

This paper classifies the prediction objects related to cloud resource management into four categories: workload (e.g., number of user requests, task arrival rate), resource demand (e.g., CPU, memory, disk, and network utilization), QoS metrics (e.g., response time and throughput), and abnormal events (e.g., job and task failures). Prediction methods for workload, resource requirements, and abnormal events fall into four categories: statistical, machine learning, deep learning, and hybrid methods, while collaborative filtering is used for QoS metrics prediction.

2.1 Prediction for workload

Workload in the context of cloud resource management refers to all input requests from end-users interacting with cloud services or batch jobs[11]. Among studies that predict workload, the number of user requests and task arrival rate are the most commonly considered metrics. Current models for workload prediction mainly use deep learning techniques. Dang et al.[12] proposed a Bi-LSTM-based model for web load prediction that improved prediction accuracy by approximately 50% compared to the conventional LSTM model. Saxena et al.[13] developed an improved 3D adaptive differential evolution algorithm AADE to train a feedforward neural network that adaptively optimizes neuron connections and counter-intuitively learns workload patterns. Arbat et al.[14] introduced a prediction model named WGANgp Transformer, which accurately captures dynamic patterns in cloud workloads.

The methods mentioned above are all used for point value prediction, which is preferred in existing prediction methods over trend prediction. While point values can predict future workload values, they do not consider changes in demand trends, such as peaks and troughs, nor do they reflect the workload characteristics in future periods [15]. To address this issue, Xia et al. [3] and Li et al. [15] utilized Piecewise Linear Regression (PLR) algorithms to divide and label datasets. Xia et al.[3] approached cloud capacity planning as a classification problem and used weighted SVM to fit statistical information and labels for each cycle, predicting the next cycle's trend. Li et al.[15] designed a multitasking cloud workload turning point prediction algorithm, Cloudtrend, that uses feature-enhanced improved LSTM to capture implicit information.

The above-mentioned prediction methods are all based on a single predictor. However, methods that rely on a single predictor are often insufficient to cope with dynamic changes and have poor performance for unknown workload patterns. Therefore, an integrated model that incorporates multiple predictors is a promising direction for future research. Kim et al. [16] proposed CloudInsight, an online integrated model based on multiple predictors, which uses multiple regression to estimate the relative accuracy of predictor variables and dynamically assigns weights to each predictor variable.

2.2 Prediction for future resource demand

In studies on resource demand prediction, the most commonly considered metrics are CPU and memory usage. The most frequently used techniques are LSTM and its improved models. Bi et al. [17] proposed a BG-LSTM model that integrates the BiLSTM model and GridLSTM, which can extract complex features from the time series of task arrival rates, CPU, and RAM usage. Nguyen et al. [18] proposed the LSTM-ED, which improves the ability of the LSTM to learn long-term dependencies by constructing an internal representation of the host load data.

The classical ARIMA model is mostly used in conjunction with neural networks to construct hybrid forecasting models, as it specializes in capturing the linear components of the time series. Xie et al. [19] used ARIMA to mine the linear relationships of the time series and used triple exponential smoothing to mine the nonlinear relationships. Devi et al. [20] used an ARIMA-ANN model to predict future CPU and memory utilization in multiple steps. In this hybrid model, ANN was used to predict the nonlinear components of the residuals obtained from the original data and ARIMA. The hybrid model was shown to provide more accurate multi-step predictions than a single model.

The presence of attention mechanisms, encoder-decoder models, and new deep neural networks has also provided new ideas for cloud resource demand prediction research. Many studies have begun to explore using these techniques to improve prediction accuracy. Al-Sayed et al. [21] proposed an attention-seq2seq based prediction technique for CPU and memory usage prediction. They addressed the problem of unstable cloud workloads and user demand variability by dividing the prediction sequence into multiple subintervals and constructing a specific model for each subinterval. Singh et al. [22] proposed an evolutionary quantum neural network (EQNN) to predict future resource (CPU and memory) utilization and workload (job arrival demand) in cloud data centers.

The methods mentioned above are all based on single predictors for prediction, and multi-predictor prediction methods have better generality than single-predictor models. Ding et al. [23] proposed COIN, a container workload prediction method that builds both source and target prediction methods. The source prediction method uses migration learning to learn common variations of workloads, while the target prediction method uses online learning to learn individual salient variations. Furthermore, the method library dynamically selects the appropriate method based on the historical accuracy of each method.

2.3 Prediction for QoS metrics

QoS refers to a set of non-functional properties (e.g., response time, reliability, cost) that can impact the overall quality of service delivery [24]. The primary approach for predicting QoS metrics is collaborative filtering. Syu et al. [25] conducted a comprehensive study on dynamic QoS attribute modeling and prediction, which demonstrated that machine learning methods and several proposed hybrid methods outperformed most statistical methods. While Ghafouri et al. [26] presented a comprehensive discussion of QoS prediction methods for web services, their focus was primarily on papers published before 2019. This paper concentrates on prediction methods for QoS metrics in cloud environments proposed after 2019. The current research trend is towards neural network models, including LSTM and hybrid models.

Gao et al. [27] extended the QoS concept by introducing additional value and cost calculations for service invocation. They considered properties such as response time, throughput, and signal strength, and used LSTM for QoS prediction. Li et al. [28] proposed a topology-aware neural network (TAN) based model for predicting QoS (response time, throughput, and reliability). The TAN model constructs end-to-end and path features and synthetically models service requests and responses. Liu et al. [29] proposed a hybrid model, HAP, that integrates two QoS prediction methods (response time and throughput). HAP incorporates a local prediction method using similarity-enhanced CF (L-CF) and a global QoS prediction method based on case inference (G-CBR).

2.4 Prediction for abnormal events

Abnormal events in cloud computing systems can include failures of jobs and tasks, cloud service system node failures, CPU overload and memory bottleneck failures, as well as other causes. These anomalies can be attributed to a variety of factors, such as software and hardware glitches, service failures, power outages, natural disasters, and more [30]. A single abnormal event can trigger a series of cascading failures, leading to significant resource loss for the cloud data center. Therefore, accurately predicting abnormal events is both critical and extremely challenging. By anticipating abnormal situations, resource waste can be reduced, and corrective actions can be taken in a timely manner.

The prediction of job and task failures is a topic of increasing research interest. The current research trend is to use deep learning models or develop generic models for anomalous event prediction. For instance, Gao et al. [31] proposed a multilayer Bi-LSTM-based fault prediction model to predict the likelihood of task and job failure. The model's multilayer structure can better handle multiple input features for higher accuracy. Similarly, PMarahatta et al. [32] used a deep neural network (DNN) to predict the failure rate of each incoming task and classified them into "failure-prone tasks" and "non-failure-prone tasks" based on the prediction results.

There are also numerous studies that employ multiple machine learning and deep learning algorithms to develop hybrid models with higher prediction accuracy and generalizability. For instance, Jassas et al. [33] utilized various machine learning classification algorithms (e.g., decision tree (DT), random forest (RF), etc.) to build a new generic model for predicting unsuccessful tasks in advance. Li et al. [34] created a series of prediction models using three machine learning algorithms (LSTM, MING, and random forest) and two different data sampling techniques (interval and oversampling).

3 CHALLENGES

Resource prediction is one of the barriers to the development of cloud computing. The most critical challenges in forecasting are the following.

3.1 High variability of cloud workloads

Cloud resources are constantly in flux, and workloads are highly non-stationary[15, 21]. For instance, the autocorrelation and periodicity of workloads in DUX-based clusters can exhibit different characteristics across various time scales.[35]. Moreover, a detailed analysis of Aliyun data center revealed that its average CPU utilization can vary significantly, ranging from 5% to 80%, during a highly volatile day [36]. The volatility in resource usage poses a major challenge for accurate resource forecasting in data centers.

3.2 Non-universality of predictive models

Single static prediction models are often inadequate to deal with complex and variable real-world cloud application workloads, which may exhibit short-term interleaved patterns with different characteristics[16]. It is essential to develop generic models that can integrate multiple prediction models with multiple predictors and include intelligent selection strategies to make predictions at a lower cost. Multi-objective prediction models have better generality. However, developing such models is even more challenging due to the need for considering multiple objectives simultaneously.

3.3 Real-time prediction

While prediction accuracy is crucial, the efficiency of prediction is equally important, particularly in real-time scenarios. Moreover, when dealing with real-time prediction requirements, it is challenging to reduce the time cost of training models while maintaining accuracy, even if the training samples are limited, and the feature dimensions are small. Therefore, there is a need to develop models that can balance accuracy and efficiency, and quickly adapt to dynamic changes in workloads.[3].

3.4 Fine-grained prediction

Prediction research at the container level is a future trend. However, predicting at the container level faces many challenges. For example, since each container in the cloud starts and shuts down in a relatively short time, it is difficult, if not impossible, to collect sufficient historical workload data in advance for the predicted containers[19]. Furthermore, container workloads are streaming, and the changing relationships in the new workload data may never be encountered and learned by the prediction model[23].

4 CONCLUSIONS

This paper provides a brief survey of prediction models used in cloud resource management. It discusses the reasons and necessities of prediction and provides a brief review of the latest and most prominent prediction models based on the prediction objects. The paper also highlights the issues and difficulties facing the research of predictive models in cloud resource management, and briefly describes future research trends in accordance with the dynamic nature of the cloud and deficiencies of current models.

References

1. Javad Dogani, Farshad Khunjush, and Mehdi Seydali. Host load prediction in cloud computing with discrete wavelet transformation (dwt) and bidirectional gated recurrent unit (bigru) network. *Computer Communications*, 198:157–174, 2023.
2. Malte Schwarzkopf, Derek Gordon Murray, and Steven Hand. The seven deadly sins of cloud computing research. In *USENIX Workshop on Hot Topics in Cloud Computing*.
3. Bin Xia, Tao Li, Qifeng Zhou, Qianmu Li, and Hong Zhang. An effective classification-based framework for predicting cloud capacity demand in cloud services. *IEEE Transactions on Services Computing*, 14(4):944–956, 2021.

4. Ali Asghar Rahmanian, Mostafa Ghobaei-Arani, and Sajjad Tofighy. A learning automata-based ensemble resource usage prediction algorithm for cloud computing environment. *Future Generation Computer Systems*, 79:54–71, 2018.
5. Célia Ghedini Ralha, Aldo H. D. Mendes, Luiz A. Laranjeira, Aletéia P. F. Araújo, and Alba C. M. A. Melo. Multiagent system for dynamic resource provisioning in cloud computing platforms. *Future Generation Computer Systems*, 94:80–96, 2019.
6. Shuja-ur-Rehman Baig, Waheed Iqbal, Josep Lluis Berral, and David Carrera. Adaptive sliding windows for improved estimation of data center resource utilization. *Future Generation Computer Systems*, 104:212–224, 2020.
7. Maryam Amiri and Leili Mohammad Survey on prediction models of applications for resources provisioning in cloud. 82:93–113, 2017.
8. Mohammad Aldossary. A review of energy-related cost issues and prediction models in cloud computing environments. 36(2):353–368, 2021.
9. Avneesh Vashistha and Pushpneel Verma. A literature review and taxonomy on workload prediction in cloud data center. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 415–420, 2020.
10. Dipak Ramoliya, Akash Patel, Khushi Patel, Gaurang Patel, Priyalba Vaghela, and Aishwariya Budhrani. Advanced techniques to predict and detect cloud system failure: A survey. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 788–793, 2022.
11. Mohammad Masdari and Afsane Khoshnevis. A survey and classification of the workload forecasting methods in cloud computing. *Cluster Computing*, 23(4):2399–2424, 2020.
12. Minh Dang and Myungsik Yoo. A web application load prediction model using recurrent neural network in cloud. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 510–514, 2020.
13. Deepika Saxena and Ashutosh Kumar Singh. Auto-adaptive learning-based workload forecasting in dynamic cloud environment. *International Journal of Computers and Applications*, 44(6):541–551, 2022.
14. Shivani Arbat, Vinodh Kumaran Jayakumar, Jaewoo Lee, Wei Wang, and In Kee Kim. Wasserstein adversarial transformer for cloud workload prediction. In *AAAI Conference on Artificial Intelligence*.
15. Li Ruan, Yu Bai, Shaoning Li, Jiaxun Lv, Tianyuan Zhang, Limin Xiao, Haiguang Fang, Chunhao Wang, and Yunzhi Xue. Cloud workload turning points prediction via cloud feature-enhanced deep learning. *IEEE Transactions on Cloud Computing*, PP:1–1, 2022.
16. In Kee Kim, Wei Wang, Yanjun Qi, and Marty Humphrey. Forecasting cloud application workloads with cloudinsight for predictive resource management. *IEEE Transactions on Cloud Computing*, 10(3):1848–1863, 2022.
17. Jing Bi, Shuang Li, Haitao Yuan, and Mengchu Integrated deep learning method for workload and resource prediction in cloud systems. 424:35–48, 2020.
18. Hoang Minh Nguyen, Gaurav Kalra, and Daeyoung Kim. Host load prediction in cloud computing using long short-term memory encoder–decoder. *The Journal of Supercomputing*, 75(11):7592–7605, 2019.
19. Yulai Xie, Minpeng Jin, Zhuping Zou, Gongming Xu, Dan Feng, Wenmao Liu, and Darrell Long. Real-time prediction of docker container resource load based on a hybrid model of arima and triple exponential smoothing. *IEEE Transactions on Cloud Computing*, PP:1–1, 2020.
20. K. Lalitha Devi and Sushmitha Time series-based workload prediction using the statistical hybrid model for the cloud environment. pages 1 – 22, 2022.

21. Mustafa M. Workload time series cumulative prediction mechanism for cloud resources using neural machine translation technique. 20, 2022.
22. Ashutosh Singh, Deepika Saxena, Jitendra Kumar, and Vrinda Gupta. A quantum approach towards the adaptive prediction of cloud workloads. *IEEE Transactions on Parallel and Distributed Systems*, PP:1–1, 2021.
23. Zhijun Ding, Binbin Feng, and Changjun Jiang. Coin: A container workload prediction model focusing on common and individual changes in workloads. *IEEE Transactions on Parallel and Distributed Systems*, 33(12):4738–4751, 2022.
24. Qiang He, Jun Yan, Hai Jin, and Yun Yang. Servicetrust: Supporting reputation-oriented service selection. In Luciano Baresi, Chi-Hung Chi, and Jun Suzuki, editors, *Service-Oriented Computing*, pages 269–284. Springer Berlin Heidelberg.
25. Yang Syu, Chien-Min Wang, and Yong-Yi Fanjiang. Modeling and forecasting of time-aware dynamic qos attributes for cloud services. *IEEE Transactions on Network and Service Management*, 16(1):56–71, 2019.
26. Seyyed Hamid Ghafouri, Seyyed Mohsen Hashemi, and Patrick C. K. Hung. A survey on web service qos prediction methods. *IEEE Transactions on Services Computing*, 15(4):2439–2454, 2022.
27. Honghao Gao, Wanqiu Huang, and Yucong Duan. The cloud-edge-based dynamic reconfiguration to service workflow for mobile ecommerce environments: A qos prediction perspective. *ACM Transactions on Internet Technology*, 21:1–23, 2021.
28. Jiahui Li, Hao Wu, Jiapei Chen, Qiang He, and Ching-Hsien Hsu. Topology-aware neural model for highly accurate qos prediction. *IEEE Transactions on Parallel and Distributed Systems*, 33(7):1538–1552, 2022.
29. Jian Liu and Youling Chen. Hap: A hybrid qos prediction approach in cloud manufacturing combining local collaborative filtering and global case-based reasoning. *IEEE Transactions on Services Computing*, 14(6):1796–1808, 2021.
30. Yogesh Sharma, Bahman Javadi, Weisheng Si, and Daniel W. Comput. Appl. Sun. Reliability and energy efficiency in cloud computing systems: Survey and taxonomy. 74:66–85, 2016.
31. Jiechao Gao, Haoyu Wang, and Haiying Shen. Task failure prediction in cloud data centers using deep learning. *IEEE Transactions on Services Computing*, 15(3):1411–1422, 2022.
32. Avinab Marahatta, Qin Xin, Ce Chi, Fa Zhang, and Zhiyong Liu. Pefs: Ai-driven prediction based energy-aware fault-tolerant scheduling scheme for cloud data center. *IEEE Transactions on Sustainable Computing*, 6(4):655–666, 2021.
33. Mohammad S. Jassas and Qusay H. Mahmoud. Analysis of job failure and prediction model for cloud computing using machine learning. 22(5):2035, 2022.
34. Yangguang Li, Zhen Ming Jiang, Heng Li, Ahmed E. Hassan, Cheng He, Ruirui Huang, Zhengda Zeng, Mian Wang, Pinan Software Engineering Chen, and Methodology. Predicting node failures in an ultra-large-scale cloud computing platform. 29:1 – 24, 2020.
35. Peter A. Dinda. The statistical properties of host load. In David R. O'Hallaron, editor, *Languages, Compilers, and Run-Time Systems for Scalable Computers*, pages 319–334. Springer Berlin Heidelberg.
36. Jing Guo, Zihao Chang, Sa Wang, Haiyang Ding, Yihui Feng, Liang Mao, and Yungang Bao. Who limits the resource efficiency of my datacenter: An analysis of alibaba datacenter traces. In *Proceedings of the International Symposium on Quality of Service*, IWQoS '19, New York, NY, USA, 2019. Association for Computing Machinery.