Int. J. Sci. R. Tech., 2025 2(1)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

A Review on the Integration of Machine Learning in Cloud Computing Resource Management

Thafzy V. M.*

Department of Computer Science and Engineering, Government Engineering College Wayanad

ABSTRACT

Efficiently managing resources in cloud computing poses a critical challenge. Over-provisioning inflates costs for both providers and customers, while under-provisioning spikes application latency and risks breaching service level agreements, leading providers to lose customers and revenue. Consequently, researchers are actively pursuing optimal resource management approaches in cloud environments, exploring container place- ment, job scheduling, and multi-resource scheduling. Machine learning plays a pivotal role in these endeavors. This paper offers an extensive survey of machine learning-based solutions for resource management in cloud computing projects, concluding with a comparative analysis of these initiatives. Additionally, it outlines future directions to steer researchers towards further advancements in this domain.

Keywords: computing, resource manage- ment, supervised learning, unsupervised learning, semisupervised learning, machine learning.

INTRODUCTION

Efficient resource management in cloud computing is a critical challenge. Over-provisioning of resources leads to increased costs for both cloud providers and customers, while under-provisioning results in application latency and potential service level agreement violations. To address these issues, researchers have explored various approaches, including con- tainer placement, job scheduling, and multi-resource schedul- ing. Machine learning techniques have emerged as powerful tools in this domain. In this survey, we delve into projects that leverage machine learning for resource management solutions within the cloud computing environment.

Presently, both industry and academia are migrating their applications to the cloud. Cloud computing offers developers a platform to run their applications without the burden of man- aging server setups and configurations. Simultaneously, cloud providers are persistently seeking avenues to enhance services for developers, prioritizing efficiency in their offerings. Re- source management is the allocation of resources such as CPU, Memory, Storage and Network Bandwidth to the virtualization unit, such as virtual machines or containers, in the cloud. There is no finite outline for proper resource management. Good resource management can vary between cloud providers based on their primary aspirations. The main objectives usually are minimizing job completion time, makespan time, increasing resource efficiency, cost reduction, and energy optimization.

Cloud computing has transformed how consumers access software and IT infrastructure, likening computing to a fifth essential utility. Despite this, effective resource management within data centers remains a challenging aspect of cloud com-

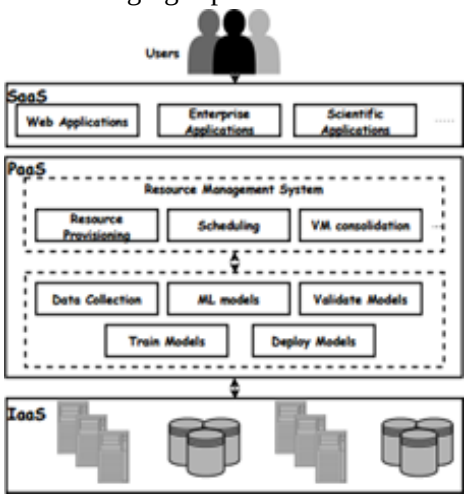

Fig. 1. Components of cloud computing paradigm using machine learning.

puting, heavily reliant on application workloads. Traditional cloud environments tethered applications to specific physical servers, often leading to

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

overprovisioning to tackle peak workloads. This approach resulted in wastage of resources and floor space, rendering data centers costly to manage in terms of resources. However, virtualization technology has emerged as a game-changer, simplifying data center operations by enabling server consolidation and increasing server utilization. This technological shift offers various advantages. Notably, industry behemoths like Google, Microsoft, and Amazon op- erate vast, intricately managed data centers, leveraging these advancements in resource management. Machine learning plays a pivotal role in optimizing re- source allocation. By analyzing historical data and patterns, ML models can make informed decisions about resource provisioning, scaling, and load balancing. These techniques enhance efficiency, reduce costs, and improve overall system performance. In our comprehensive review, we explore how ML algorithms are applied to tackle resource management challenges.Fig. 1. shows the components of cloud computing paradigm using machine learning. A wide range of projects that utilize machine learning techniques are encompassed here. We investigate container orchestration, workload prediction, auto-scaling, and dynamic resource allocation. By examining the strengths and limitations of each approach, we provide insights into the state of the field. Whether it's reinforcement learning, supervised learn- ing or unsupervised learning we explore how these methods contribute to effective resource management. We evaluate the strengths, weaknesses, and trade-offs of different ML- based resource management solutions. By understanding the nuances of each project, practitioners and researchers can make informed choices based on their specific requirements. Our goal is to facilitate decision-making and encourage the adoption of best practices.

II. RELATED WORK

In the realm of resource management research, three promi- nent types of machine learning models come into play: supervised learning, semi supervised, unsupervised learning, and reinforcement learning (RL). In supervised learning, the training dataset comprises both features and corresponding labels. The model learns to predict labels based on the pro-vided features. Essentially, it grasps the relationship between input features and output labels. This approach is widely used when labeled data is available for training. Unlike supervised learning, unsupervised learning operates without explicit labels in the dataset. Instead, the model explores the inherent patterns and relationships within the training data. By uncovering these hidden structures, it identifies clusters or associations. Unsupervised learning is particularly useful when labeled data is scarce or unavailable. In RL, an agent interacts with an environment. It observes the environment's state, takes actions that influence the next state, and receives a reward based on its actions. The reward serves as feedback, indicating how well the agent's actions align with the RL objective function. By exploring state-action-reward spaces, RL algorithms learn op- timal action sequences to maximize cumulative rewards. Semi-supervised learning stands as a hybrid method harnessing both labeled and unlabeled data for training machine learning models. While supervised learning banks on labeled examples exclusively, and unsupervised learning operates devoid of labels, semi-supervised learning takes a middle ground. It combines a limited set of labeled data with a more extensive pool of unlabeled data to train models effectively.

A. Supervised Learning

Supervised learning involves data samples comprising mul- tiple input features alongside corresponding labels. The learn- ing process aims to approximate a mapping function that connects these features to the label. Once established, this mapping function enables predictions of labels for new data based on their input features. This method [1] stands as the most extensively applied machine learning approach, finding use across diverse domains. For instance, classifying objects based on their attributes, such as categorizing mobile devices by brand and specifications, exemplifies supervised learning's classification task. When the objective of supervised learning is to predict a continuous variable, like stock pricing, it be- comes a regression task. Fig. 2 depicts the layout of supervised learning initiating the rehoming routine through an external trigger. It begins by listing feasible action combinations for the VNFs requiring relocation (such as those on a server slated for removal, update, or restart). For each combination, feature vectors are constructed, and models created during the offline stage come into play, predicting the costs associated with relocation. The system selects the combination with the

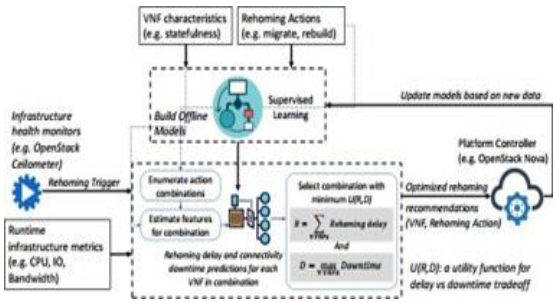

Fig. 2. Cost efficient supervised learning mechanism in cloud resource management.

lowest cost, providing rehoming suggestions to the platform controller for implementation.

B. Semi supervised learning

Owing to inadequate planning of virtual machine (VM) resources and the complexity of fluctuating loads, many en- terprises grapple with substantial wastage of VM resources. While existing solutions can identify idle VMs, most are tailored for private or public cloud settings, lacking effec- tiveness in managed cloud environments. These environments pose challenges like limited labels, data quality issues, and managing large-scale VMs in production.

To address this gap, analyzing resource usage data from thousands of VMs in an actual managed cloud is commenced. Based on this analysis, an innovative method to detect idle VMs, leveraging meticulous data processing, feature engi- neering, and model training is devised. The method in this paper [2] demonstrated excellent performance through exten- sive experimentation using real data from Sangfor's managed cloud, proving its effectiveness and practicality in production environments.

C. Unsupervised Learning

Dynamic resource provisioning in Web applications aims to maintain service-level objectives (SLOs) while minimizing operational costs. Yet, the intricate nature of multitier Web apps poses challenges in autonomously allocating resources for each tier without human oversight. This paper [3] in- troduced unsupervised machine learning techniques for the dynamic provisioning of multitier Web applications, ensuring adherence to user-defined performance objectives. This model operates in real time, leveraging learning al- gorithms to recognize workload patterns from access logs. It identifies bottlenecks corresponding to distinct workload patterns and dynamically formulates resource allocation poli- cies for each pattern. Through experiments using synthetic workloads on Amazon Elastic Compute Cloud (EC2), it val- idated the effectiveness of this approach and compared it against conventional rule-based autoscale strategies.The results demonstrate that this proposed techniques empower cloud infrastructure providers or application owners to autonomously manage multitier Web applications, meeting SLOs without prior knowledge of the applications' resource utilization or workload patterns.

D. Reinforcement Learning

As software systems handle larger and more intricate data volumes, there's a heightened demand in large-scale systems for high-performance distributed computing. The rise of Web 2.0 and the subsequent acceleration of the Internet have pro- pelled Cloud computing as a dynamic and adaptable paradigm, offering advantages in meeting computing demands efficiently. However, extensive Cloud use without proper scheduling ap- proaches can result in increased energy consumption, elevated costs, and significant carbon dioxide emissions. Inadequate scheduling may further diminish the lifespan of physical devices and lead to longer response times for user requests. Consequently, the efficient scheduling of resources or optimal allocation of requests, a typically NP-hard problem, stands out as a critical challenge in the evolving landscape of Cloud computing. Over the years, researchers have extensively explored re- source scheduling problems in Cloud computing, emphasizing improvements in quality of service (QoS), cost reduction, and environmental impact mitigation. However, the growing complexity of Cloud systems, functioning as super-massive distributed entities, has constrained the applicability of con- ventional scheduling approaches. To address this challenge, researchers have turned to machine learning, particularly deep reinforcement learning (DRL), which combines deep learning (DL) and reinforcement learning (RL). DRL presents a promis- ing avenue for resource scheduling in Cloud computing and has emerged as an innovative concept in recent years. This paper [4] conducts a survey on resource scheduling methodologies in Cloud computing, focusing specifically on DRL-based scheduling approaches. It reviews the application of DRL and discusses the challenges and future trajectories of DRL in the realm of Cloud computing scheduling.

III. CONCLUSION

The majority of studies utilizing supervised learning have focused on using recent workloads to forecast present or future workloads. Their primary goals revolve around reducing server or VM numbers to conserve energy and cut costs. When scheduling a job on a VM or server, the target is to en- sure sufficient resources without disrupting other concurrently running jobs on that specific VM/server. The selection process typically adopts a greedy approach, aiming to identify the best VM/server that meets these conditions. However, this method doesn't consistently guarantee the most optimal choice.

Following supervised learning, reinforcement learning emerges as the next frequently investigated method. In most reinforcement learning approaches, the state-space comprises jobs' directed acyclic graphs (DAGs), while the action space involves scheduling jobs and defining job parallelism levels. The primary aim here is often to reduce job completion time. However, a drawback of these solutions lies in their lack of online adaptability. When workload alterations occur, there's a delay in retraining the model and effectively responding to these changes. Hence, in scenarios demanding prompt adap-tation to dynamic changes, an alternative method alongside reinforcement learning becomes essential. Unsupervised learning isn't particularly well-suited for re- source management due to its tendency to group workloads into clusters, often leading multiple workloads to fall within the same cluster. Consequently, the system allocates identical resources to all members of a cluster. However, an exception arises where unsupervised learning proves advantageous: when there's a shift or change in the workload. In the realm of future research, an intriguing avenue in- volves delving into online learning coupled with reinforcement learning (RL) while integrating meta-learning techniques. Ex- ploring scenarios involving more than two conflicting ob- jectives with multiple agents presents an exciting direction for investigation. Additionally, researchers should delve into preemptible jobs, considering the utilization of a designated agent for managing job preemption decisions. An enhancement for RL models would entail reducing reliance on job resource usage profiles, given their occasional inaccuracies. Instead, a preferable approach involves employ- ing supervised learning to estimate the resource profiles for each job.

REFERENCE

1. M. Wajahat, "Cost-efficient dynamic management of cloud resources through supervised learning," Stony Brook University Journal, 2019.
2. Z. C. J. Y. X. C. B. L. . C. X. Xian Yu, Kejiang Ye, "A semi-supervised learning based method for identifying idle virtual machines in managed cloud: Application and practice," Springer Journal, 2023.
3. M. N. D. Waheed Iqbal and D. Carrera, "Unsupervised learning of dynamic resource provisioning policies for cloud-hosted multitier web applications," IEEE SYSTEMS JOURNAL, 2015.
4. R. B. Guangyao Zhou, Wenhong Tian, "Deep reinforcement learning- based methods for resource scheduling in cloud computing: A review and future directions," arXiv:2105.04086v1 [cs.DC], 2021.
5. R. B. Tahseen Khana, Wenhong Tiana, "Machine learning (ml)-centric re- source management in cloud computing: A review and future directions," arXiv:2021.04086v1 [cs.DC], 2021.
6. R. H. E. K. E. R. Sepideh Goodarzy, Maziyar Nazari, "Resource manage- ment in cloud computing using machine learning: A survey," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020

HOW TO CITE: Thafzy V. M., A Review on the Integration of Machine Learning in Cloud Computing Resource Management, Int. J. Sci. R. Tech., 2025, 2 (1), 18-21. https://doi.org/10.5281/zenodo.14592545