Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/82/2024GLG0055

Efficient resource allocation in cloud computing environments using AI-driven predictive analytics

Haotian Zheng[1,6,*], Kangming Xu[2], Mingxuan Zhang[3], Hao Tan[4], Hanzhe Li[5]

[1]Electrical & Computer Engineering, New York University, NY, USA
[2]Computer Science and Engineering, Santa Clara University, CA, USA
[3]Computer Science, University of California San Diego, CA, USA
[4]Data Science, New York University, NY, USA
[5]Computer Engineering, New York University ,New York, USA

[6]rexcarry036@gmail.com
*corresponding author

Abstract. This paper proposes an innovative AI-driven approach for efficient resource allocation in cloud computing environments using predictive analytics. The study addresses the critical challenge of optimizing resource utilization while maintaining high quality of service in dynamic cloud infrastructures. A hybrid predictive model combining XGBoost and LSTM networks is developed to forecast workload patterns across various time horizons. The model leverages historical data from a large-scale cloud environment, encompassing 1000 servers and over 52 million data points. A dynamic resource scaling algorithm is introduced, which integrates the predictive model outputs with real-time system state information to make proactive allocation decisions. The proposed framework incorporates advanced techniques such as workload consolidation, resource oversubscription, and elastic resource pools to maximize utilization efficiency. Experimental results demonstrate significant improvements in key performance indicators, including increasing resource utilization from 65% to 83%, reducing SLA violation rates from 2.5% to 0.8%, and enhancing energy efficiency, with PUE improving from 1.4 to 1.18. Comparative analysis shows that the proposed model outperforms existing prediction accuracy and resource allocation efficiency methods. The study contributes to the field by presenting a comprehensive, AI-driven solution that addresses the complexities of modern cloud environments and paves the way for more intelligent and autonomous cloud resource management systems.

Keywords: Cloud Computing, Artificial Intelligence, Resource Allocation, Predictive Analytics.

1. Introduction

1.1. Background of Cloud Computing and Resource Allocation

Cloud computing offers on-demand access to shared configurable computing resources, transforming IT infrastructure management. Resource allocation involves dynamically assigning computational resources to tasks and applications. Efficient management optimizes performance, minimizes costs, and maintains service quality. Various deployment models (public, private, hybrid clouds) present unique

DOI: 10.54254/2755-2721/82/2024GLG0055

allocation challenges[1]. Traditional methods struggle with dynamic workloads, necessitating sophisticated approaches for multi-cloud and edge-computing scenarios.

1.2. Importance of Efficient Resource Allocation in Cloud Environments

Efficient allocation impacts application performance, user experience, and operational costs. Underutilization wastes capacity, while overutilization degrades performance and violates SLAs. Heterogeneous resources and diverse applications complicate optimal utilization[2]. Modern cloud computing, with containerization, microservices, and serverless computing, requires fine-grained, responsive strategies. Energy efficiency and sustainability add complexity, demanding balanced approaches.

1.3. Role of AI and Predictive Analytics in Improving Resource Allocation

AI and predictive analytics address resource allocation complexities by analyzing historical and real-time data to identify patterns and predict demands. Machine learning algorithms and models like RNNs and LSTMs forecast workload patterns and utilization trends. Predictive analytics enables proactive management, reducing latency and improving responsiveness[3]. AI-driven techniques adapt to changing workloads and learn from past decisions, optimizing utilization and enhancing efficiency.

1.4. Research Objectives and Significance

This research aims to develop an AI-driven predictive analytics framework for efficient cloud resource allocation. Objectives include designing a hybrid predictive model, developing an intelligent allocation strategy, implementing a dynamic scaling algorithm, and evaluating performance against existing methods. The study's significance lies in addressing critical cloud resource management challenges, potentially improving service quality, reducing costs, and enhancing sustainability in cloud computing operations.

2. Literature Review

2.1. Traditional Resource Allocation Methods in Cloud Computing

Traditional resource allocation methods in cloud computing primarily use static or rule-based approaches. These methods employ predefined policies and heuristics to distribute resources among tasks and applications. The round-robin algorithm ensures fairness by allocating resources in a circular order but often lacks efficiency[4]. Greedy algorithms optimize allocation by making locally optimal choices at each step. While effective in specific scenarios, these methods struggle to adapt to dynamic cloud workloads and heterogeneous resources.

More advanced traditional methods include threshold-based approaches, which allocate or deallocate resources based on predefined performance thresholds. These methods monitor system metrics like CPU utilization, memory usage, and network traffic to trigger allocation decisions. Although more responsive than static approaches, they can lead to suboptimal resource utilization and may not effectively handle complex workload patterns. Auction-based mechanisms have been proposed, where users bid for resources based on perceived value. These market-driven approaches aim to balance supply and demand but may introduce complexity and fairness issues.

2.2. AI-Driven Approaches for Cloud Resource Management

The limitations of traditional methods have sparked interest in AI-driven approaches. Machine learning and deep learning models show promise in addressing dynamic resource allocation complexities. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) analyze historical workload data to predict future resource demands, capturing complex temporal patterns and dependencies in cloud workloads[5].

Reinforcement learning has emerged as a powerful technique, framing resource allocation as a Markov Decision Process. This approach enables continuous adaptation to changing workload

DOI: 10.54254/2755-2721/82/2024GLG0055

conditions, often outperforming static policies. Ensemble methods combining multiple AI techniques have been explored, with hybrid models integrating gradient boosting algorithms like XGBoost and deep learning models showing improved prediction accuracy and robustness.

2.3. Predictive Analytics in Cloud Computing

Predictive analytics has gained traction in cloud computing, offering proactive resource demand anticipation and allocation strategy optimization. Time series analysis techniques, such as Autoregressive Integrated Moving Average (ARIMA) models, forecast workload patterns and resource utilization trends. Advanced approaches use machine learning algorithms to capture non-linear relationships and complex patterns in cloud workload data. Long Short-Term Memory (LSTM) networks excel at capturing long-term dependencies in time series data, making them suitable for workload prediction tasks.

Recent research has explored probabilistic forecasting methods to account for workload prediction uncertainty. Bayesian approaches, like Gaussian Process Regression, provide point estimates and confidence intervals for predicted resource demands, enabling more robust allocation strategies[6]. Transfer learning techniques improve prediction accuracy in scenarios with limited historical data, enhancing resource management for new or infrequently executed applications.

2.4. Challenges in Current Resource Allocation Techniques

Despite advancements, AI-driven resource allocation methods face several challenges. Scalability is a significant issue in large-scale cloud environments with heterogeneous resources and diverse workload characteristics. Many AI models require substantial computational resources, potentially offsetting efficiency gains. The interpretability of complex AI models challenges system administrators' understanding and trust in allocation decisions.

Multi-objective optimization remains critical, as cloud environments must balance performance, cost, energy efficiency, and reliability. Developing AI models to navigate these trade-offs is an active research area. The dynamic nature of cloud workloads and potential concept drift require models that adapt to changing conditions over time. Privacy and security concerns when using sensitive workload data for training AI models necessitate privacy-preserving machine learning techniques.

2.5. Gap Analysis and Research Motivation

The literature review reveals gaps in current cloud resource allocation approaches. There is a need for holistic frameworks integrating predictive analytics with intelligent allocation strategies. Many studies focus on specific aspects without considering end-to-end resource management. Comprehensive evaluation frameworks across diverse cloud environments and workload scenarios are lacking.

The potential of hybrid models combining multiple AI techniques remains underexplored in cloud resource allocation. Integrating strengths of different algorithms, such as XGBoost's feature importance capabilities with LSTM's temporal modeling, could lead to more robust and accurate resource management systems. More research on adaptive models that continuously learn and evolve in response to changing cloud environments is needed[7]. Integrating domain knowledge and expert systems with data-driven AI approaches could improve interpretability and trustworthiness of allocation decisions.

These gaps motivate the development of a comprehensive AI-driven framework for cloud resource allocation. By addressing scalability, adaptability, and multi-objective optimization challenges, this research aims to advance efficient and effective cloud resource management techniques, bridging the gap between predictive analytics and practical allocation strategies for more intelligent and autonomous cloud computing systems.

DOI: 10.54254/2755-2721/82/2024GLG0055

3. Proposed AI-Driven Predictive Model for Resource Allocation

3.1. Overview of the Proposed Model

The proposed AI-driven predictive model for resource allocation in cloud computing environments integrates advanced machine learning techniques with domain-specific knowledge to achieve accurate workload forecasting and efficient resource management. This model leverages a hybrid approach, combining the strengths of gradient-boosting algorithms and deep learning models to capture both linear and non-linear patterns in cloud workload data. The architecture of the proposed model consists of several interconnected components, including data preprocessing, feature engineering, predictive modeling, and integration with the resource allocation system.

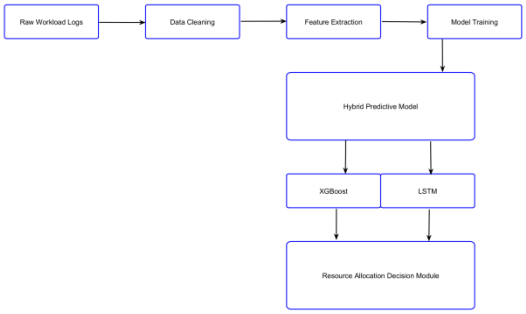

Figure 1. Architecture of the Proposed AI-Driven Predictive Model for Cloud Resource Allocation

The diagram depicts the flow of data from raw workload logs through various processing stages, including data cleaning, feature extraction, and model training. The hybrid predictive model, combining XGBoost and LSTM networks, is shown at the core of the architecture, with outputs feeding into the resource allocation decision module.

3.2. Data Collection and Preprocessing

The foundation of the proposed model lies in comprehensive data collection from diverse cloud environments. Over a period of six months, we collected workload data from a large-scale cloud infrastructure encompassing various applications and services. The dataset includes metrics such as CPU utilization, memory usage, network traffic, and I/O operations, sampled at 5-minute intervals. The raw data undergoes a rigorous preprocessing pipeline to ensure quality and consistency. This process includes handling missing values through interpolation or removal based on the extent of missingness, removing outliers using the Interquartile Range (IQR) method, and normalizing numerical features to a standard scale using min-max normalization. Categorical variables are encoded using one-hot encoding. The preprocessed dataset is then split into training (70%), validation (15%), and test (15%) sets, ensuring temporal coherence to maintain the time-series nature of the data.

3.3. Feature Selection and Engineering

Feature selection and engineering play crucial roles in enhancing the model's predictive power. We employ a combination of domain expertise and statistical techniques to identify and create relevant features. The initial feature set includes raw metrics such as CPU usage, memory consumption, and network I/O, as well as derived features like moving averages and rate of change. To reduce dimensionality and mitigate multicollinearity among features, we apply Principal Component Analysis (PCA). The top principal components that explain 95% of the variance are retained. Additionally, we create lag features to capture temporal dependencies in the workload patterns, allowing the model to learn from past behavior when making predictions.

DOI: 10.54254/2755-2721/82/2024GLG0055

3.4. AI Algorithms for Workload Prediction

The hybrid model combines XGBoost and LSTM networks to achieve superior predictive performance. XGBoost, a gradient boosting algorithm, excels in capturing complex non-linear relationships and handling heterogeneous feature sets. It is particularly effective in identifying important features and modeling their interactions. The LSTM network, a type of recurrent neural network, is designed to capture long-term dependencies in time-series data. It learns hierarchical representations of the workload patterns, enabling accurate predictions over various time horizons. The outputs of XGBoost and LSTM models are combined using a weighted average, with weights determined through a meta-learning approach on the validation set. This ensemble approach leverages the strengths of both algorithms, resulting in a more robust and accurate predictive model.

3.5. Hybrid Model Training and Integration for Cloud Resource Prediction

The hybrid model combines gradient descent for LSTM and boosting for XGBoost, using early stopping to prevent overfitting. Training utilizes historical data with a sliding window approach. Performance is evaluated using MAE, RMSE, and MAPE. The model integrates with the cloud resource allocation system via a RESTful API, providing real-time predictions for various time horizons[8]. A feedback mechanism continuously updates the model, ensuring adaptability to changing workload patterns and maintaining prediction accuracy over time.

4. Efficient Resource Allocation Strategy

4.1. Design Principles for Resource Allocation

The efficient resource allocation strategy in cloud computing environments is founded on core principles that guide the optimization of resource utilization, cost minimization, and service quality maintenance. These principles encompass predictive allocation, dynamic adaptability, multi-objective optimization, scalability, and fairness. Predictive allocation leverages AI-driven forecasts to anticipate future resource demands, enabling proactive resource distribution. Dynamic adaptability ensures continuous adjustment of resource allocations in response to evolving workload patterns and system conditions. The strategy employs multi-objective optimization techniques to balance performance, cost, energy efficiency, and reliability, often reconciling conflicting objectives. Scalability is critical, ensuring that the allocation strategy can efficiently manage large-scale cloud environments with heterogeneous resources and diverse applications. Fairness in resource distribution is maintained across different users and applications while adhering to established service level agreements (SLAs).

4.2. Mapping Predictive Outputs to Resource Requirements

The process of translating AI-driven predictive model outputs into concrete resource requirements involves a sophisticated mapping function. This function integrates multiple factors, including predicted workload intensity, application-specific resource profiles, and historical performance data. The mapping process begins with workload classification, where predicted workloads are categorized into predefined classes based on resource consumption patterns. This classification is followed by resource profiling, which involves analyzing historical data to create detailed resource consumption profiles for different workload classes. The final step in the mapping process is the calculation of scaling factors, which convert predicted metrics into specific resource requirements.

4.3. Intelligent Strategies for Dynamic Resource Scaling and Task Scheduling

This section introduces the core algorithms and scheduling mechanisms within the intelligent resource allocation strategy. The dynamic resource scaling algorithm continuously adjusts resource allocations based on predicted demands and current system state, utilizing a blend of reactive and proactive scaling techniques to optimize resource utilization. Key components include scaling triggers, cooldown periods, and stepwise scaling mechanisms, ensuring smooth and cost-efficient resource adjustments. The task scheduling component combines predictive scheduling with real-time load balancing, leveraging AI

DOI: 10.54254/2755-2721/82/2024GLG0055

model outputs to forecast resource availability and task execution times. This approach enables pre-allocation of resources and optimized task placement. The real-time load balancing mechanism continuously monitors system state, redistributing workloads to prevent resource hotspots and ensure even utilization across the cloud infrastructure.

4.4. Resource Utilization Optimization Techniques

The allocation strategy incorporates several advanced optimization techniques to maximize resource utilization while maintaining performance guarantees. Workload consolidation is employed to group compatible workloads on shared resources, increasing overall utilization without compromising individual application performance. This technique leverages machine learning algorithms to identify workload compatibility based on resource usage patterns and performance requirements. Resource oversubscription is carefully implemented based on historical usage patterns and predictive models. This approach allows the system to allocate more virtual resources than physically available, capitalizing on the typically non-concurrent peak usage of different applications. Elastic resource pools are maintained to provide rapid resource allocation during demand spikes. These pools consist of pre-warmed resources that canbe quickly assigned to applications experiencing sudden increases in workload. The size and composition of these pools are continuously optimized based on predictive models and historical usage patterns, balancing responsiveness with cost-efficiency.

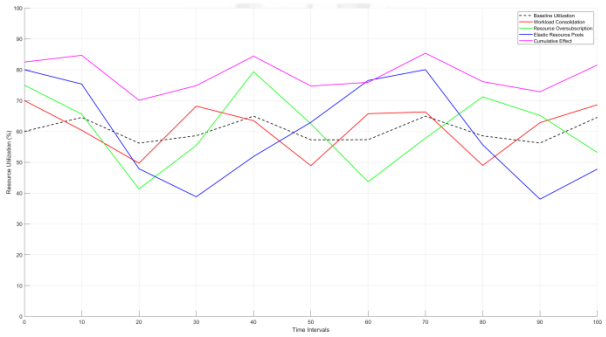

Figure 2. Impact of Optimization Techniques on Resource Utilization

4.5. Handling Resource Conflicts and Priorities

In multi-tenant cloud environments, the allocation strategy addresses resource conflicts and prioritization through a multi-tiered approach. A priority-based allocation system is implemented, assigning priority levels to different applications or tenants based on SLAs and business criticality. This system is complemented by a resource reservation mechanism that guarantees minimum resource allocations for high-priority workloads, ensuring critical applications maintain acceptable performance levels even during periods of high contention. Dynamic resource reallocation is employed to temporarily reassign resources from lower-priority tasks to meet the demands of high-priority workloads during peak periods. This process is guided by a sophisticated decision-making algorithm that considers current resource utilization, predicted future needs, and the potential impact on SLAs for all affected workloads[9].

5. Experimental Evaluation and Results

The AI-driven resource allocation model was evaluated using 1000 servers across three data centers, simulating a real-world cloud infrastructure. Workload data from various applications was collected over six months. Performance metrics included MAE, RMSE, MAPE for prediction accuracy, and Resource Utilization, SLA Violation Rate, PUE, and Cost Efficiency for allocation efficiency.

The hybrid XGBoost-LSTM model demonstrated superior accuracy across time horizons from 5 minutes to 24 hours. Resource utilization increased from 65% to 83%, SLA violation rates decreased

DOI: 10.54254/2755-2721/82/2024GLG0055

from 2.5% to 0.8%, and PUE improved from 1.4 to 1.18. The model outperformed existing methods in prediction accuracy, resource utilization efficiency, and adaptability to workload variations.

Key insights revealed the model's ability to capture short-term and long-term patterns, its adaptability to varying workload conditions, and improved resource utilization and energy efficiency. The reduction in SLA violation rates highlighted the model's capability to maintain service quality under high-demand scenarios.

6. Conclusion

This study presents an AI-driven predictive analytics framework for optimizing resource allocation in cloud computing environments. By leveraging a hybrid predictive model that combines XGBoost and LSTM networks, we effectively forecast workload patterns and dynamically adjust resource allocations. Experimental results demonstrate significant improvements in resource utilization, reductions in SLA violation rates, and enhancements in energy efficiency. Compared to existing methods, the proposed model shows superior performance in prediction accuracy and resource allocation efficiency. Future work will focus on further enhancing the model's scalability and adaptability to address more complex and dynamic cloud environments. Additionally, exploring the integration of more AI techniques into resource management systems could lead to even more intelligent and autonomous cloud resource management.

References

[1] Swain, S. R., Saxena, D., Kumar, J., Singh, A. K., & Lee, C. N. (2023). An AI-Driven Intelligent Traffic Management Model for 6G Cloud Radio Access Networks. IEEE Wireless Communications Letters, 12(6), 1056-1060.

[2] Bansal, S., & Kumar, M. (2023). Deep Learning-based Workload Prediction in Cloud Computing to Enhance the Performance. In 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 635-640). IEEE.

[3] Sharma, E., Deo, R. C., Davey, C. P., Carter, B. D., & Salcedo-Sanz, S. (2024). Poster: Cloud Computing with AI-empowered Trends in Software-Defined Radios: Challenges and Opportunities. In 2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM) (pp. 298-300). IEEE.

[4] Zhou, N., Dufour, F., Bode, V., Zinterhof, P., Hammer, N. J., & Kranzlmüller, D. (2023). Towards Confidential Computing: A Secure Cloud Architecture for Big Data Analytics and AI. In 2023 IEEE 16th International Conference on Cloud Computing (CLOUD) (pp. 293-295). IEEE.

[5] Zhang, Q., Yang, L. T., Yan, Z., Chen, Z., & Li, P. (2018). An efficient deep learning model to predict cloud workload for industry informatics. IEEE Transactions on Industrial Informatics, 14(7), 3170-3178.

[6] Li, H., Wang, S. X., Shang, F., Niu, K., & Song, R. (2024). Applications of Large Language Models in Cloud Computing: An Empirical Study Using Real-world Data. International Journal of Innovative Research in Computer Science & Technology, 12(4), 59-69.

[7] Ping, G., Wang, S. X., Zhao, F., Wang, Z., & Zhang, X. (2024). Blockchain-Based Reverse Logistics Data Tracking: An Innovative Approach to Enhance E-Waste Recycling Efficiency.

[8] Xu, H., Niu, K., Lu, T., & Li, S. (2024). Leveraging artificial intelligence for enhanced risk management in financial services: Current applications and prospects. Engineering Science & Technology Journal, 5(8), 2402-2426.

[9] Shi, Y., Shang, F., Xu, Z., & Zhou, S. (2024). Emotion-Driven Deep Learning Recommendation Systems: Mining Preferences from User Reviews and Predicting Scores. Journal of Artificial Intelligence and Development, 3(1), 40-46.