GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Economics and Management Innovation

Article

Application of Machine Learning in Cloud Service Cost Prediction and Resource Optimization

Fangyuan Li [1,*]

[1] Amazon Web Services, Inc., AWS Global Sales, Washington, 98121, USA
* Correspondence: Fangyuan Li, Amazon Web Services, Inc., AWS Global Sales, Washington, 98121, USA

Abstract: With the continuous development of cloud computing, the demand for intelligent cost management and resource optimization is increasing. This paper focuses on the application of machine learning in the whole process of cloud services, including time series modeling, regression analysis and transfer learning in cost prediction, load modeling in resource optimization, reinforcement learning scheduling and multi-tenant allocation mechanism. The research results also verify that machine learning helps cloud service platforms improve prediction accuracy and resource utilization effect and is conducive to building efficient self-adaptive service architecture.

Keywords: cloud computing; machine learning; cost forecasting; resource optimization

1. Introduction

With the development of cloud computing, enterprises have an increasing demand for cost reduction and reasonable allocation of resources. Traditional management methods are unable to respond accurately to changing environments, which limits efficiency and flexibility. Therefore, the data-driven nature of machine learning provides a new solution, namely cost prediction and resource optimization, to solve these problems. This paper mainly studies the specific practices and advantages in model selection, job scheduling, and system adaptability in the aspects of cost prediction and resource optimization.

2. Basic Concepts of Machine Learning

Machine learning refers to a class of algorithms that learn from data to automatically find patterns and make predictions and decisions based on them. The core idea of machine learning is to learn patterns from past data through algorithms and to exhibit better generalization ability when faced with new data. In cloud computing, in addition to analyzing resource usage trends, machine learning is also widely used in automatic capacity expansion, load balancing, quality of service monitoring and other fields [1]. Depending on the learning styles applied, machine learning can be divided into three categories: supervised learning, unsupervised learning and reinforcement learning. Supervised learning is a common processing method for labeled data sets, which is often used for cost prediction and classification tasks. In unsupervised learning it is used for data clustering and pattern discovery, such as different types of workloads; Reinforcement learning, on the other hand, is well-suited for dynamic decision-making problems and has been widely used in resource scheduling and policy optimization [2].

3. Application of Machine Learning in Cloud Service Cost Forecasting

3.1. Time Series Prediction Model

In the cloud computing environment, resource consumption and cost data often have a strong time correlation, which is mainly reflected in the two aspects of cycle change and trend change. We can use existing historical data to capture these characteristics in order

Received: 22 April 2025
Revised: 24 April 2025
Accepted: 14 May 2025
Published: 18 May 2025

to make accurate cost predictions [3]. Traditional data models, such as the autoregressive moving average model, are used in situations where environmental changes are relatively stable, and their structure is clear, easy to interpret, and easy to install and maintain in resource-limited systems. Its prediction function can be expressed as:

$$\hat{C}_{t+1} = f(C_t,\ C_{t-1}, \ldots, C_{t-n}) \quad (1)$$

Where C_t represents the cost at moment t, $\hat{C}_{t+1}$ is the projected cost at the next moment, and n is the window length. However, with the increase of service load, it is difficult for the model of simple linear structure to accurately establish the non-stationary sequence. Especially in the context of frequent events or rapid changes in resource use, linear models become very constrained. In recent years, neural network-based methods have been widely used in this field, especially in recurrent neural networks and their variants such as long-term memory networks, which have superior generalization ability to deal with complex patterns and abrupt changes. LSTM can capture the temporal correlation well and has a good effect on constructing the periodic resource usage trend model. Compared with traditional models, deep learning methods require a wealth of high-quality historical data as a basis, and its training and application costs are higher. Therefore, the selection of the model should comprehensively consider the actual demand, data volume, and system operation efficiency, in order to achieve the desired prediction accuracy and maintain optimal system expenditure [4].

3.2. Regression Algorithm Modeling

In addition to time, other resource usage metrics also affect cloud service costs, such as CPU utilization, memory allocation, and storage access times. Such complex structures are well suited to the construction of regression models. Regression algorithms estimate future costs by creating a mapping relationship between input attributes and predicted costs. Commonly used methods include linear regression, support vector regression, random forest, gradient lifting decision tree and so on. Although linear regression is simple to construct, it has limited ability to express nonlinear relationships or complex connections between features. Ensemble learning methods offer better stability and accuracy when dealing with high-dimensional feature sets. At the same time, training efficiency and generalization performance should also be considered in model selection, so as to avoid over-adaptation to small samples [5]. The applicability of common regression algorithms is summarized and compared in Table 1 below:

Table 1. Comparison of Common Regression Models.

Model	Peculiarity	Application scenario	Restrict
Linear regression	Simple and easy to explain	The linear correlation is obvious	The ability to fit complex relationships is weak
SVR	Support nonlinear modeling	Small scale complex data	Parameter tuning is difficult and the calculation cost is high
XGBoost	High precision and robustness	Large-scale feature space	The structure of the model is complex and the training time is long

Table 1 summarizes the cost estimation characteristics and applicability of common regression models. Linear regression modeling is simple and suitable for linear relationships. SVR can build nonlinear models, but it requires a lot of computation. XGBoost has strong accuracy and anti-interference ability, which is more suitable for handling large amounts of data. The choice of the model should be judged according to the actual application requirements and data characteristics [6].

3.3. Feature Selection and Preprocessing

Ensuring the quality of input features is essential for building an effective cost prediction model. Since features are generally obtained from multi-dimensional monitoring, features usually contain noise, missing and redundant variables, etc. Without proper feature selection, irrelevant or redundant variables may introduce errors; similarly, lack of data pre-processing can significantly reduce model performance. Therefore, feature selection and pre-processing of data have become a very important step in modeling.

The function of feature selection is to discard the elements that are irrelevant to the cost change or have little correlation, reduce the complexity and improve the model performance. Typical selection methods include filtering methods based on statistics (such as variance screening and correlation filtering), wrapping methods based on model performance (such as recursive elimination), and embedded selection (such as L1 regularization, tree model feature importance ranking, etc.). Table 2 below is a comparison of common feature selection methods:

Table 2. Comparison of Common Feature Selection Methods.

Method type	Common technique	Advantage	Restrict
Filtration method	Variance threshold, Pearson correlation coefficient	Simple and fast for high-dimensional data	The combination effect between features is not considered
Wrapping	Recursive feature elimination	High precision, based on model feedback optimization	The calculation cost is high and the training time is long
Embedding method	Lasso, tree model weights	Can be integrated with the training process, automatic selection	It relies on the model structure, and its generality is weak

Table 2 compares the common feature selection methods. The filter method can be used for efficient initial screening. The wrapping method is based on the enhancement of model performance, which has better accuracy but a large amount of calculation. The embedding method integrates feature selection directly into the model training process, making it the most flexible among the three approaches. The selection method should be comprehensively considered according to data scale, model type and computing resources.

In the process of data analysis, the purpose is to make the data more accurate and standardized, mainly including the processing of missing data, mining of outliers, standardization and time window construction. Missing data can be dealt with by means of averaging, interpolation, or building model predictions; The processing of outliers is usually Z-score, box graph or isolated forest algorithm. Standardization can reduce differences to avoid certain features being too significant. The main purpose of the sliding window is to transform raw historical data into structured sample data that can be used to infer cost trends over time. Figure 1 shows the feature processing flow chart:

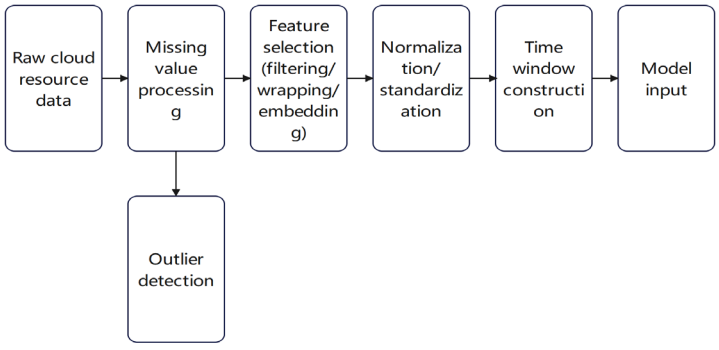

Figure 1. Feature Processing Flows Chart.

After the processing of the above steps, the data is more concentrated, and the model can be more stable and generalized. It should be noted that each type of model has a different sensitivity to the shape and distribution of data. For example, neural networks benefit more from data standardization, while decision trees and similar models are less sensitive to outliers. Therefore, it is necessary to select different preprocessing methods according to different algorithms.

3.4. Model Migration and Generalization

For predictive models in multi-tier or hybrid cloud environments, they need to be run repeatedly on different machines. However, due to the different pricing mechanisms, load characteristics, and operating environments on each machine, the learned patterns of the model on the training machine are difficult to transfer to other machines. Therefore, improving the portability and generalization of the model has become a key challenge in the practical application of cost estimation. One effective solution is transfer learning, which essentially takes general knowledge from the original environment and fine-tunes the model to suit the new environment with less data from the target environment. The commonly used methods include feature transfer, model parameter movement, expression conversion, etc. The choice depends on the similarity between the source domain and the target domain. If there is a large distribution difference between the two, it may be necessary to use domain adaptation to reduce the impact of inter-domain differences on the model effect. Figure 2 shows the framework diagram of model migration:

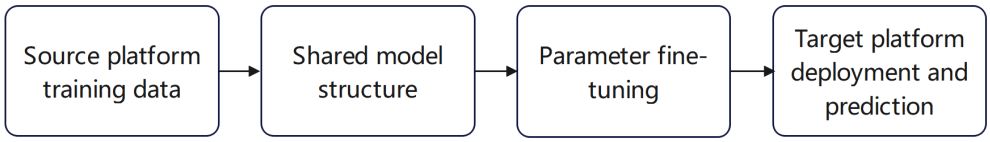

Figure 2. Model Migration Diagram.

Furthermore, good generalization ability also requires proper normalization planning, data extension techniques, and control of the learning process. For example, considering multi-device universality when establishing features, or establishing learning architectures that formally input multiple tasks, can make the model scalable for a variety of cloud environments. Designing multi-task learning architectures not only allows for the sharing of underlying network structures but also enhances the generalization ability of the model through the integration of multiple tasks, which is suitable for cloud environment applications with limited resources or complicated tasks. In addition, different types of sub-models can also be jointly trained to train in different data partitions or regions with different characteristics and then combine the results during inference to improve the overall prediction reliability.

4. Application of Machine Learning in Cloud Service Resource Optimization

4.1. Workload Modeling

In the cloud service model, the effect of resource scheduling policy depends on the accurate identification and modeling of cloud system workload characteristics. Workloads refer to various computing tasks on the cloud platform, which have different resource usage patterns, task deadlines, access modes and other attributes. For different types of tasks, they have different requirements for computing, storage, and network resources. Therefore, establishing an appropriate classification and modeling method can improve the overall resource utilization and scheduling intelligence of the system. Machine learning technology is used to develop a data-based job type recognition method to extract the relevant characteristics of various job types through historical resource usage data. This process usually includes three steps: feature extraction, feature reduction and

classification. The resource utilization vector is used to map the task behavior into multi-dimensional data feature vector, and the input clustering or classification algorithm is used for training and decomposition. For unlabeled flow data, unsupervised learning methods can be automatically divided into multiple categories based on similarity, and these categories can be further used to guide future scheduling or resource allocation strategies.

It is important to note that building a workload model should not only focus on the current resource usage, but also on its change over time and periodic behavior patterns. Using time series analysis or sliding window method can improve the accuracy of the description of workload change process. Figure 3 shows the workload modeling flowchart:

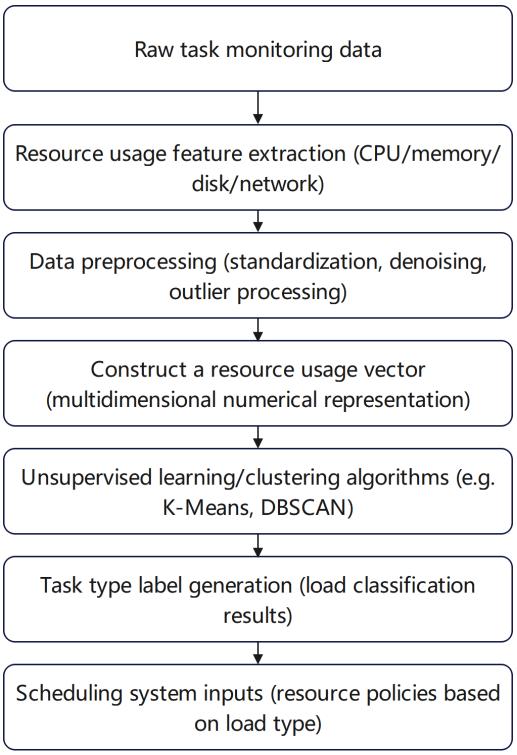

Figure 3. Workload Modeling Flowchart.

4.2. Reinforcement Learning Expansion and Contraction

Elastic resource allocation and scheduling is a major feature of cloud service platform, mainly reflected in the ability to automatically adjust or return resources according to real needs, which is of great significance to achieve service stability and optimize operating costs. Traditional capacity expansion technologies typically rely on pre-set thresholds, which may perform poorly in dynamic or unpredictable workload scenarios due to their lack of adaptability. Therefore, an automatic reinforcement learning optimization based on online feedback can be established to realize resource control. In the modeling process, system state S_t represents the current resource usage, such as CPU, memory load; Action a_t indicates expansion, contraction, or unchanged. Environment feedback an instant reward r_t, used to evaluate the effectiveness of the current operation. The learning goal is to maximize cumulative discounted rewards:

$$Q(s_t, a_t) = r_t + \gamma \cdot \max_{a'} Q(s_{t+1}, a') \quad (2)$$

Where γ is the discount factor for future rewards, and the Q -value function guides policy updates. Reinforcement learning is a strategic learning method to optimize environmental feedback, and it is also suitable for solving such complex problems that require timely response to frequent changes in resource status. In the reinforcement learning

framework, resource management is modeled as an interactive process between agents and the environment. The agent decides whether to expand or reduce capacity based on the system state (such as CPU usage and memory utilization) and uses the environmental reward feedback to adjust its strategy, aiming to optimize resource efficiency and system performance. Compared with fixed strategies, reinforcement learning is better at dealing with nonlinear and multi-objective optimization scenarios. Attention should be paid to the design of state space, the applicability of reward function and the effect of model deployment when formulating strategies, so as to ensure the robustness of strategies and platform compatibility.

4.3. Priority Scheduling Optimization

In a cloud environment, dealing with thousands of jobs running concurrently is a key issue for the cloud. Work priority typically governs task scheduling and execution order throughout the processing lifecycle. Algorithms in machine learning are more in line with a more appropriate ordering of actual work in the cloud than rule-based and pre-set rules. The core of scheduling optimization is to assign a priority score to each task T_i to be scheduled:

$$P_i = f(\text{Deadline}_i, \text{Age}_i, \text{Resource}_i, \text{SLA}_i) \quad (3)$$

Where, Deadline_i is the cut-off time, Age_i is the waiting time, Resource_i is the resource requirement, and SLA_i is the service level indicator. When the model calculates the output timing, it can prioritize and allocate resources. Through task feature modeling, each task will have a task scheduling level to evaluate the auxiliary scheduling decision process. In this training process, the model can learn a set of task sorting models according to the historical scheduling data and running results, to update the scheduling queue order in real time and maximize scheduling efficiency.

In general, machine learning scheduling optimization algorithms are adaptive and can automatically adjust strategies to changes in business volume. Combined with reinforcement learning technology, the long-term benefit optimization goal can be introduced into the strategy, so that the system will not only consider the current load, but also predict the possibility of future resource competition, so that the task can be managed in time.

4.4. Multi-Tenant Resource Allocation

Multi-tenant resource allocation requires a dynamic, controllable and predictable sharing mechanism in the resource sharing pool. The resource usage and dynamic service requirements of different tenants vary greatly. Therefore, static and fixed allocation cannot balance resource usage and improve resource efficiency in the shared pool. With the help of machine learning methods, tenant behavior model can be established, which becomes an important part of resource allocation prediction and optimization. The tenant's resource requirements can be modeled as a time series vector:

$$D_i(t) = [CPU_t, Mem_t, I/O_t]^{(i)} \quad (4)$$

Predict the resource application trend in a future period based on historical data. $\hat{D}_i(t+1)$ The system dynamically adjusts resource allocation to improve the overall resource utilization. At the same time, in order to measure resource fairness among tenants, the resource utilization ratio index is introduced:

$$U_i = \frac{A_i}{D_i} \quad (5)$$

Where A_i indicates the resource obtained by tenant, D_i indicates the resource requested by tenant i, and U_i indicates the higher the resource satisfaction. Through machine learning, a new multi-tenant resource management scheme can be proposed. The solution models and forecasts each tenant's future resource needs based on historical usage patterns, enabling proactive and optimized allocation decisions. Predictive models often employ regression or series-building techniques and are constantly updated with a picture of each tenant's usage through real-time monitoring data. In addition, a dynamic

quota adjustment strategy based on reinforcement learning has been gradually introduced, and the current contribution and real usage status of each tenant are constantly evaluated during the implementation process to achieve refined resource management. In multi-tenant environments, fairness must be balanced with resource optimization to ensure equitable access and efficient resource use. Therefore, a measure of fairness can be introduced as an optimization goal, such as the number of resources received by a user in a specific time is equal to the number of actual needs. In addition, at the whole system level, the differentiated needs of services should be considered to ensure that customers with higher priorities can have better services and more resources during peak hours, so as to avoid performance degradation or disservice.

5. Conclusion

The increasing complexity of cloud computing infrastructures means that traditional processing methods can no longer effectively handle cost and resource management. This paper analyzes and completely expounds the process of cloud service cost prediction and resource optimization based on machine learning technology, from model construction to application landing. The research shows that the solution based on machine learning can effectively enhance the accuracy of prediction, optimize resource scheduling and strengthen the adaptive ability of the system. In the future, the intelligent strategy combining multi-task learning, online optimization and cross-platform migration is the main means to further improve the level of cloud services and provide the basis for the intelligent development of cloud computing.

References

1. I. Pintye, J. Kovács, and R. Lovas, "Enhancing machine learning-based autoscaling for cloud resource orchestration," *J. Grid Comput.*, vol. 22, no. 4, pp. 1–31, 2024, doi: 10.1007/s10723-024-09783-1.
2. B. Guindani, D. Ardagna, A. Guglielmi, R. Rocco, and G. Palermo, "Integrating Bayesian optimization and machine learning for the optimal configuration of cloud systems," *IEEE Trans. Cloud Comput.*, vol. 12, no. 1, pp. 277–294, 2024, doi: 10.1109/TCC.2024.3361070.
3. P. Nawrocki and M. Smendowski, "Optimization of the use of cloud computing resources using exploratory data analysis and machine learning," *J. Artif. Intell. Soft Comput. Res.*, vol. 14, no. 4, pp. 287–308, 2024, doi: 10.2478/jaiscr-2024-0016.
4. S. Petale and S. Subramaniam, "CLARA+: dual machine learning optimized resource assignment for translucent SDM-EONs," *J. Opt. Commun. Netw.*, vol. 16, no. 10, pp. F1–F12, 2024, doi: 10.1364/JOCN.527846.
5. P. Nawrocki and M. Smendowski, "FinOps-driven optimization of cloud resource usage for high-performance computing using machine learning," *J. Comput. Sci.*, vol. 79, p. 102292, 2024, doi: 10.1016/j.jocs.2024.102292.
6. O. C. Agomuo, O. W. B. Jnr, and J. H. Muzamal, "Energy-aware AI-based optimal cloud infra allocation for provisioning of resources," in *Proc. 2024 IEEE/ACIS 27th Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distributed Comput. (SNPD)*, pp. 269–274, Jul. 2024, doi: 10.1109/SNPD61259.2024.10673918.