Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 1, 2023

Dynamic Allocation Mechanism of Cloud Computing Resources Driven by Neural Network

Yining Ou

Santa Clara University, Santa Clara, US

Abstract: With the popularization of cloud computing technology, the dynamic allocation mechanism of cloud computing resources has become an important research field to improve resource utilization and meet the needs of diversified workloads. The purpose of this study is to explore the dynamic allocation mechanism of cloud computing resources driven by neural network and introduce the powerful ability of deep learning into cloud computing environment. We put forward a comprehensive framework, which combines data collection, analysis, decision-making and implementation to realize intelligent resource allocation. These data will be used to train BP neural network (BPNN). In order to predict the bidding price, a BPNN is designed, which usually includes input layer, hidden layer and output layer. The number of nodes in the input layer is equal to the dimension of the input feature, and the number of nodes in the output layer is 1, which indicates the prediction of the bidding price. Through experiments and simulations, we verify the effectiveness of the dynamic resource allocation mechanism driven by neural network. The results show that this mechanism can better adapt to the changing workload requirements, improve resource utilization and reduce resource waste. In addition, it provides better performance and user experience, thus enhancing the competitiveness of cloud computing systems.

Keywords: Neural Network; Cloud Computing; Dynamic Allocation.

1. Introduction

With the rapid development and wide application of cloud computing technology, cloud computing has become one of the key driving forces in the field of modern information technology [1]. Cloud computing provides opportunities for flexibility, scalability and resource sharing, enabling various applications to be deployed and run on a global scale. However, the effective allocation of cloud computing resources has always been a challenging issue. Traditional resource allocation methods are usually based on static rules or simple heuristic algorithms, which are difficult to adapt to the changing workload requirements.

In recent years, the rapid development of deep learning and neural network technology has brought new opportunities and challenges to the dynamic allocation of cloud computing resources. The dynamic allocation mechanism of cloud computing resources driven by neural network combines the advantages of deep learning and cloud computing, which can better adapt to various workload requirements, improve resource utilization, and provide better performance and user experience [2-3].

The purpose of this paper is to discuss the principle, method and application of dynamic allocation mechanism of cloud computing resources driven by neural network. We will introduce the basic concepts and key components of this mechanism. Through this study, we hope to provide researchers and practitioners in the field of cloud computing with an in-depth understanding of the dynamic resource allocation mechanism driven by neural networks, and provide useful inspiration and guidance for future research and application. The progress in this field will help to improve the performance, efficiency and reliability of cloud computing system, thus promoting the wider application and development of cloud computing technology.

2. Dynamic Allocation Mechanism Framework of Cloud Computing Resources

The dynamic allocation mechanism of cloud computing resources refers to the methods and strategies for automatically adjusting and allocating computing, storage and network resources according to actual needs in the cloud computing environment [4]. This mechanism aims to optimize resource utilization, improve performance and meet user needs, while minimizing resource waste.

Virtual machine (VM) dynamic allocation: This is one of the most common resource allocation mechanisms in cloud computing. By monitoring the performance and load of virtual machines, the cloud platform can automatically adjust the number and configuration of virtual machines to meet different workload requirements. This includes vertical expansion (changing virtual machine configuration) and horizontal expansion (increasing or decreasing the number of virtual machines).

Dynamic allocation of containers: The resource allocation mechanism based on container technology is lighter and faster. Containers can be migrated between different hosts to adapt to load changes. Container orchestration tools such as Kubernetes can automatically manage the deployment and scheduling of containers.

Load balancing: The load balancer can distribute traffic to different servers to ensure the even use of resources. According to the load balancing strategy, it can automatically route the request to the most suitable server according to the load of the server.

Elastic storage: Cloud computing platforms usually provide elastic storage solutions, allowing users to expand or reduce storage capacity according to their needs. This allows users to adjust the storage capacity in real time according to data requirements.

Intelligent decision engine: Some advanced cloud computing resource dynamic allocation mechanisms use machine learning, deep learning and artificial intelligence technologies to make smarter resource allocation decisions to improve performance and resource utilization [5].

The goal of these mechanisms is to provide enough computing, storage and network resources according to actual needs without wasting resources. Through dynamic allocation, the cloud computing environment can achieve higher efficiency, scalability and availability, thus better meeting the needs of users [6-7].

The functions of the resource allocation mechanism proposed in this paper are mainly realized in the provider agent, the consumer agent and the auction intermediary. The provider agent and the consumer agent are responsible for providing decision support for the objects they serve, such as initializing the bid, deciding the bidding price, submitting the bid, scoring, etc. The auction intermediary is responsible for collecting the bid, running the winning bid determination algorithm, notifying the auction results, releasing the market operation data and managing the prestige system, etc. The system framework is shown in Figure 1.

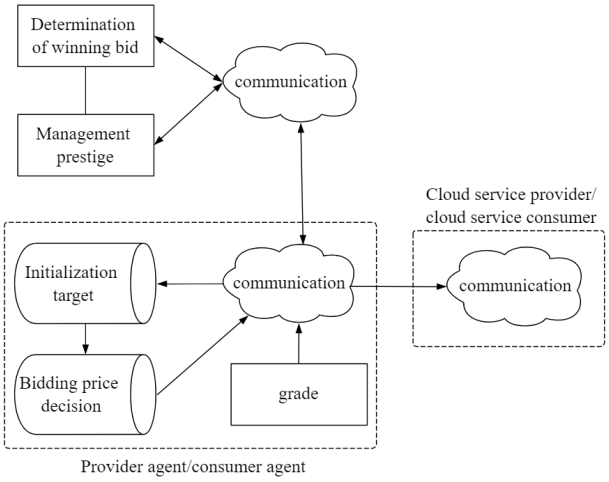

Figure 1. Dynamic allocation mechanism framework of cloud computing resources

3. BPNN Training Method for Bidding Price Decision Mechanism

BP neural network (BPNN) is a type of artificial neural network (ANN), which is used for machine learning and deep learning tasks. BPNN is a feedforward neural network, which includes input layer, hidden layer and output layer. The core idea of this network is to train the model through back propagation algorithm, so as to learn the feature representation of input data and make predictions.

The main components of BPNN include:

Input layer: accepts feature inputs from data sets, and each input corresponds to a neuron in the network.

Hidden layer: one or more layers of neurons located between the input layer and the output layer, used to learn and extract higher-order features of data. The depth and width of the hidden layer can be set according to the complexity of the task.

Output layer: used to generate the final prediction result of the model, and each output neuron corresponds to a category or value.

Weight and bias: The connection connecting each pair of adjacent neurons has an associated weight, which is used to adjust the transmission intensity of the signal. In addition, each neuron has a bias term to adjust the activation threshold of neurons.

The bidding price decision mechanism in the dynamic allocation of cloud computing resources is a way for cloud computing service providers and users to negotiate and decide the price of resource leasing. This mechanism is usually used to lease virtual machines or other computing resources in the cloud computing market to ensure the effective allocation and optimal utilization of resources. Cloud computing resources can be allocated by auction, and users can bid for resources in the auction. This mechanism is usually used for high competition for resources, but it may lead to fierce price fluctuations. Users can choose different resource types and price models to combine resources according to their needs. This mechanism allows users to meet their needs more flexibly, but costs need to be carefully managed. Different bidding price decision-making mechanisms are suitable for different use situations and needs. Users and cloud computing service providers need to choose appropriate mechanisms according to their specific conditions to balance cost, resource availability and performance. At the same time, intelligent algorithms and automation systems also play an important role in the dynamic allocation of resources to help optimize decision-making and manage resources.

The introduction of bidding price decision mechanism and the training method using BPNN can enhance the efficiency and performance of dynamic allocation of cloud computing resources. This method will consider the price of resources in order to better meet the needs of users and reduce costs [8]. Collect historical data, including resource requirements, workload characteristics, resource prices, etc. These data will be used to train BPNN. In order to predict the bidding price, a BPNN is designed, which usually includes input layer, hidden layer and output layer. The number of nodes in the input layer is equal to the dimension of the input feature, and the number of nodes in the output layer is 1, which indicates the prediction of the bidding price.

The characteristics of the input layer include: resource requirements, workload load, timestamp, historical bid price, etc. The output layer of neural network will give the predicted value of bidding price. The loss function is selected to measure the error between the predicted value and the actual bidding price. A common choice is Mean Squared Error(MSE):

$$MSE = \frac{1}{n} * \sum (y - y\_pred)^2 \tag{1}$$

Where n represents the number of samples, y represents the actual bidding price, and y_pred represents the predicted value of the neural network.

Back propagation algorithm is used to calculate the gradient of loss function relative to the weight of neural network. Then, according to the gradient descent method, the weights of the neural network are updated to reduce the value of the loss function.

$$\Delta W = -\eta * \nabla L \tag{2}$$

Where ΔW is the update of the weight, η is the learning rate, and ∇L is the gradient of the loss function with respect to the weight.

Repeat the process of back propagation and weight updating until the loss function converges or reaches a predetermined stop condition. Once the neural network training is completed, it can be used to predict the bidding

price of future resources. Pass the new input features to the network, and get the predicted bid price [9-10].

This training method allows the cloud computing system to predict the bidding price of future resources in order to make more informed resource allocation decisions. It needs to reasonably select input features, neural network structure and training parameters to improve the accuracy and reliability of prediction [11]. The advantage of using BPNN to predict the bidding price is that it can learn complex price patterns and trends, and constantly improve the prediction according to the actual data. This will help to improve the cost-effectiveness of resource allocation and better meet the needs of users.

4. Experimental Analysis

Based on the SimJava2.0 toolkit, this paper uses JDK1.6 for simulation. The service type is set with reference to Amazon cloud computing platform, the processing capacity of various resources is set with reference to the resource list of TeraGrid partners, and the price is set with reference to the resource price of Amazon cloud platform. The parameters of BP bidding decision-making mechanism are set according to reference [6]. In this paper, the benchmark mechanism for performance comparison is the stable continuous two-way auction mechanism proposed in reference [8], and the comparison data are obtained by averaging each mechanism for 20 times under specified circumstances.

The fixed task granularity is normal, and the relationship between supply and demand in the market is balanced. Figure 2 shows the comparison of the time expenditure of the two mechanisms under different market scales, in which Figure 2(a) counts the training time of the neural network in the proposed method into the time expenditure, and Figure 2(b) shows the comparison of the time expenditure when the neural network is trained offline, that is, the time expenditure of the proposed method does not include the training time of the neural network.

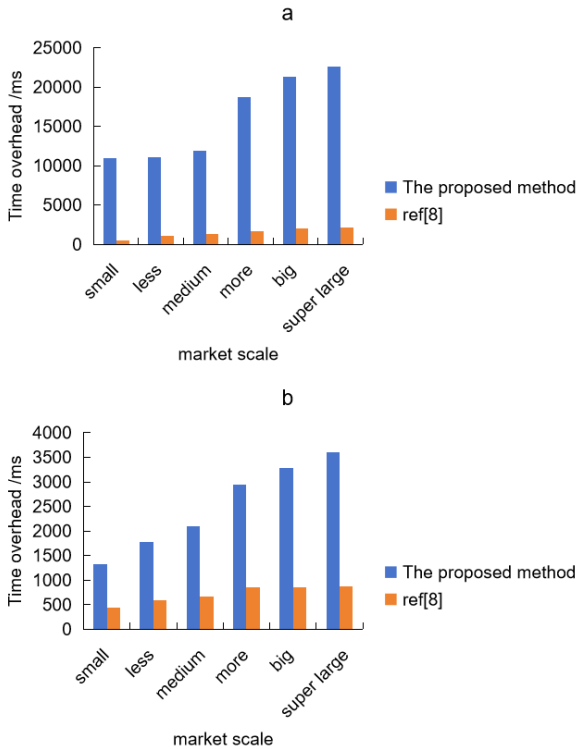

Figure 2. Comparison of time expenditure under different market sizes

It can be seen that the time cost of the proposed method is higher than that of ref[8]. This is because with the expansion of the market, the time for training the neural network and determining the winning bid of intelligent algorithm will increase obviously. However, the resource allocation in the cloud environment is usually not very strict, and when enough training samples are obtained, the system can train the neural network offline and directly use the neural network to bid online, which will greatly reduce the time cost of the proposed method.

The performance of the mechanism in this paper is better than that of the benchmark mechanism, but the time cost is large. If the neural network is trained offline, the time cost of this mechanism can be significantly reduced. Although the performance of the benchmark mechanism is obviously not as good as that of this mechanism, the time cost is low. The superiority of this mechanism in performance is mainly due to the adoption of the bidding decision mechanism using historical bidding data and the intelligent winning bid determination method, which is the biggest difference between this mechanism and related research work.

5. Conclusion

The dynamic resource allocation mechanism driven by neural network can predict the resource demand more accurately to meet the workload demand, thus improving the resource utilization rate. This helps to reduce the cost of cloud computing and reduce the waste of resources. This mechanism can adjust the resource allocation in real time according to the changing workload requirements, thus enhancing the adaptability of the cloud computing system. This enables the system to better adapt to different application scenarios and load fluctuations. Through more intelligent resource allocation decision, neural network-driven mechanism can improve the performance and user experience of cloud computing system. Users can get higher performance and better meet their service quality requirements. Through more intelligent, adaptive and efficient resource allocation, this mechanism is expected to promote the further development of cloud computing technology and provide users with better services and economic benefits. Future work will continue to explore new methods and technologies to further optimize the performance and usability in this field.

References

[1] Lin, W., Wang, J. Z. , Liang, C. , & Qi, D. (2011). A threshold-based dynamic resource allocation scheme for cloud computing. Procedia Engineering, 23(12), 695-703.

[2] Wei, W. , Fan, X. , Song, H. , Fan, X. , & Yang, J. (2018). Imperfect information dynamic stackelberg game based resource allocation using hidden markov for cloud computing. IEEE Transactions on Services Computing, 11(99), 78-89.

[3] Kavitha, J. , & Rao, K. T. (2022). Dynamic resource allocation in cloud infrastructure using ant lion-based auto-regression model. International journal of communication systems(6), 35.

[4] Bai, W. (2017). Virtual technology of cache and data real time allocation in cloud computing data center. Acta Technica CSAV (Ceskoslovensk Akademie Ved), 62(2), 187-197.

[5] Zhu, Q. , & Agrawal, G. (2012). Resource provisioning with budget constraints for adaptive applications in cloud environments. IEEE Transactions on Services Computing, 5(4), 497-511.

[6] Rochman, Y. , Levy, H. , & Brosh, E. (2017). Dynamic placement of resources in cloud computing and network applications. Performance Evaluation, 115(10), 1-37.

[7] Beloglazov, A. , Abawajy, J. , & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Generation Computer Systems, 28(5), 755-768.

[8] Zhang, Y. , Xu, K. , Shi, X. , Wang, H. , Liu, J. , & Wang, Y. (2018). Design, modeling, and analysis of online combinatorial double auction for mobile cloud computing markets. International journal of communication systems, 31(6), 3460.1-3460.17.

[9] Zhang, J. , Xie, N. , Zhang, X. , & Li, W. (2018). An online auction mechanism for cloud computing resource allocation and pricing based on user evaluation and cost. Future Generation Computer Systems, 89(9), 286-299.

[10] Kim, I. K. , Hwang, J. , Wang, W. , & Humphrey, M. (2020). Guaranteeing performance slas of cloud applications under resource storms. IEEE Transactions on Cloud Computing, (99), 1-1.

[11] Deng, Y. , Xu, Z. , & Tian, Z. (2019). Dynamic adaptive streaming over http-based 3d resource allocation algorithm. The Journal of Engineering, 2019(23), 8888-8890.