Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3123

INNOVATIVE APPLICATION OF AI AGENT BASED ON MARKOV DECISION PROCESS IN DYNAMIC OPTIMIZATION DECISION-MAKING FOR ENTERPRISE SERVICES

XueYong Li[1], WeiBin Zhao[2*], Xing Xu[3], JinHan Cai[4]
[1]*Shenzhen Liqi Artificial Intelligence Development Co., Ltd., Shenzhen 518000, Guangdong, China.*
[2]*School of Future Transportation, Guangzhou Maritime University, Guangzhou 510000, Guangdong, China.*
[3]*Dongguan City Society of Human Resources and Social Security, Dongguan 523000, Guangdong, China.*
[4]*School of Digital Economy and Trade, Guangzhou Maritime University, Guangzhou 510000, Guangdong, China.*
[*]*Corresponding Author: WeiBin Zhao*

Abstract: Against the backdrop of the booming digital economy, the enterprise service environment is increasingly characterized by dynamics, high uncertainty, and multi-objective collaboration. Facing core challenges such as fluctuating service demands, changing resource constraints, and rapidly iterating customer preferences, traditional static decision-making methods struggle to adapt to actual operational needs, gradually revealing their limitations. This paper focuses on key issues in dynamic enterprise service decision-making, including resource allocation, service priority setting, and precise matching of customer needs. Targeting the regional characteristics and operational pain points of enterprise service scenarios in South China, we innovatively improve the traditional Markov Decision Process (MDP) model. Leveraging the intelligent computing power and data simulation support provided by Shenzhen Qicheng Zhiyuan Network Technology Co., Ltd. through the "LiQiCloud" AI-powered Sci-Tech Innovation Policy Platform, we construct an AI agent decision-making framework that integrates Deep Reinforcement Learning (DRL) and Digital Twin technology. Furthermore, we propose an improved value iteration algorithm based on function approximation, which effectively overcomes the limitations of traditional models and significantly enhances decision-making efficiency and adaptability. Through rigorous mathematical derivation, we elucidate the theoretical foundations for state representation, action output, reward function design, and optimal policy solutions. Simulation and empirical analysis were conducted using real-world service scenarios from a small-to-medium-sized clothing e-commerce enterprise in South China (daily order volume of 1,500–2,200; peak order volume of 4,000 during major promotions in 2025; customer service team of 25). Experimental results demonstrate that the proposed model significantly outperforms traditional MDP models and manual decision-making methods in core metrics, including service cost control (reduced by 22.8%), customer waiting time (shortened by 56.1%), resource utilization (improved by 14.7%), and effective problem resolution rate (increased by 5.9 percentage points). This study provides an implementable and replicable solution for optimizing service decision-making in the digital transformation of SMEs in South China.
Keywords: Markov decision process; AI agent; Dynamic service decision-making; Deep reinforcement learning; LiQiCloud

1 INTRODUCTION

1.1 Problem Formulation

With the in-depth penetration of the digital economy, South China, as one of the regions with the highest concentration of small and medium-sized enterprises (SMEs) in China, has witnessed a remarkable increase in the dynamics and uncertainty of the enterprise service environment, particularly in service-intensive industries such as e-commerce and logistics. Stochastic fluctuations in customer demand, real-time constraints on service resources, frequent changes in the external market environment, and conflicting multi-dimensional objectives, including cost control, customer experience, and resource utilization, have rendered traditional static decision-making approaches inadequate to adapt to the dynamic operational rhythm of enterprises.

For instance, during peak promotion periods such as the "618" and "Double 11" shopping festivals, as well as regional characteristic promotional activities (e.g., derivative promotions of the Canton Fair), small and medium-sized e-commerce enterprises in South China face surging order volumes. Traditional resource allocation, relying heavily on manual experience, fails to adjust customer service and distribution resources in real time, leading to a substantial decline in service quality. Taking a garment e-commerce enterprise in South China as an example: with only 25 customer service staff, customer waiting time often exceeded 15 minutes during major 2025 promotions—more than three times that in non-promotion periods—directly causing a 12.3% rise in complaint rates and a 4.8% increase in customer churn rate.

AI agents, endowed with autonomous perception, real-time response, and continuous learning capabilities, have emerged as a core solution to dynamic enterprise service decision-making. As a classical theory for describing stochastic dynamic systems, the Markov Decision Process (MDP) provides a solid theoretical foundation for the

decision optimization of AI agents. Nevertheless, traditional MDP exhibits distinct limitations in practical applications among SMEs in South China, failing to accommodate their characteristics of limited resources, variable scenarios, and sparse data. The specific drawbacks are as follows:
It assumes discrete and finite state/action spaces, which are seriously inconsistent with the high-dimensional continuous features formed by resources, customers, and environments in actual enterprise services [1];
It relies on extensive historical data to estimate state transition probabilities, whereas most SMEs in South China suffer from insufficient data accumulation, uneven data quality, and difficulty adapting to the non-stationarity caused by rapid market changes [2];
It assumes a delay-free and fully observable system, ignoring the inherent latency in perception–decision–execution loops in real enterprise decision-making, as well as the partial observability of customer demand and resource status [3];
It takes customer satisfaction as an immediate reward, yet this indicator is lagged and cannot support real-time decision adjustment, contradicting the operational demands of "rapid response and flexible adaptation" among SMEs in South China [4].

1.2 Theoretical and Practical Value

Theoretical Value: In response to the limitations of traditional MDP models in dynamic enterprise service decision-making, this study improves the modeling approach of MDP-based AI agents in dynamic service decision-making by introducing function approximation, deep reinforcement learning, and digital twin technology. It optimizes reward function design and policy solution algorithms [5], relaxes the idealized assumptions of conventional models, and enriches the theoretical system of integrated applications of Markov Decision Processes and AI agents. This provides a novel research paradigm and theoretical reference for studies on similar dynamic decision-making problems [6].
Practical Value: Focusing on the pain points of dynamic service decision-making among SMEs in South China and incorporating the latest enterprise operational data of 2025, this study proposes implementable and low-cost optimization methods for AI agents. Digital twin technology is adopted to address the problem of sparse data in SMEs, online learning techniques are used to cope with market non-stationarity, and real-time proxy indicators such as waiting time and effective resolution rate replace lagged customer satisfaction metrics to enable real-time decision adjustment [7]. Empirical verification based on real enterprise scenarios in South China confirms the practical effectiveness of the model in cost control and customer experience improvement. This offers a feasible path for the digital transformation of SMEs in South China and helps enhance their core competitiveness [8].

2 LITERATURE REVIEW

2.1 International Research Progress

In recent years, as the core theoretical foundation of reinforcement learning, Markov Decision Process (MDP) has been widely adopted in enterprise service decision-making, resource allocation and related fields. The Deep Q-Network (DQN) algorithm proposed by Mnih et al. approximates the value function via deep neural networks, which effectively solves the function approximation problem in high-dimensional state spaces and provides technical support for the application of MDP in complex scenarios.
In the field of logistics and distribution, foreign scholars have combined MDP with deep reinforcement learning to achieve dynamic path optimization and improve distribution efficiency. In customer service, digital twin technology has been employed to construct business simulators, accurately reproduce service processes, and optimize customer service scheduling strategies and resource allocation schemes.
However, existing international research has obvious limitations: high model training costs and strict hardware requirements make it difficult to adapt to the resource-constrained scenarios of small and medium-sized enterprises in South China; insufficient model generalization ability and poor adaptability to specific regions and industries fail to meet the operational demands of SMEs for being "small but sophisticated, fast and flexible".

2.2 Domestic Research Status

Domestic research focuses on the practical demands of dynamic enterprise service decision-making and has made certain progress in supply chain services, cloud services and other fields. In supply chain services, scholars have constructed MDP-based AI agents and introduced online learning techniques to cope with environmental non-stationarity and improve supply chain responsiveness. In cloud services, function approximation methods are used to handle high-dimensional state spaces, realize dynamic allocation of computing resources, and reduce operational costs.
Regarding the research status in South China, existing studies still have many shortcomings: some studies retain the discrete and finite assumptions of traditional MDP, which are disconnected from real enterprise service scenarios; reward function design mostly focuses on a single lagged indicator (e.g., customer satisfaction), ignoring the multi-objective collaborative decision-making needs of SMEs (cost, efficiency, experience); the optimization of policy solution algorithms lacks pertinence, without fully considering the characteristics of sparse data and limited computing power in SMEs; most case verifications use simulated data, lacking quantitative comparisons based on the latest 2025 operational data of real enterprises in South China, so the practical implementability of models needs to be verified.

2.3 Research Breakthrough Directions

Based on the achievements and deficiencies of existing domestic and international research, and combined with the operational characteristics of SMEs in South China and the latest 2025 data, this paper focuses on breakthroughs in the following directions: First, introduce digital twin technology to build a service scenario simulator to solve the problem of sparse data in SMEs. Second, adopt function approximation methods to handle high-dimensional continuous state/action spaces and break the idealized assumptions of traditional MDP. Third, design a multi-objective real-time proxy indicator-based reward function to adapt to enterprises' real-time decision-making requirements. Fourth, propose an improved value iteration algorithm integrated with function approximation to improve convergence speed and solution accuracy while reducing computational cost. Fifth, conduct empirical verification based on 2025 operational data from real enterprises in South China, highlight model advantages through quantitative comparison, and construct a low-cost, implementable MDP-AI agent decision-making model suitable for SMEs.

3 THEORETICAL FOUNDATIONS

3.1 Markov Decision Process

Markov Decision Process (MDP) is a classical mathematical model for describing decision-making processes in stochastic dynamic systems. Its core feature is the Markov property, meaning that the current state contains all information required for optimal decision-making, and the transition of future states only depends on the current state and the action taken, independent of historical states.

MDP is generally represented by a five-tuple:

$M = < S,A,P,R,\gamma >$, where each component is defined as follows:

1. S denotes the state space, which includes all possible states of the system. In enterprise dynamic service decision-making, states cover core information such as resource status, customer status, and environmental conditions.
2. A denotes the action space, which contains all possible decision-making actions, corresponding to specific behaviors in enterprise services such as resource allocation, service priority setting, and customer demand matching.
3. $P: S \times A \times S \rightarrow [0,1]$ represents the state transition probability function, where $P(s'|s, a)$ denotes the probability that the system transfers to state s' after taking action a in state s.
4. $R: S \times A \rightarrow R$ denotes the reward function, where $R(s, a)$ represents the immediate reward obtained by taking action a in state s, used to evaluate the effectiveness of the action.
5. $\gamma \in [0,1]$ denotes the discount factor, which weights immediate rewards against future rewards. A value closer to 1 indicates that the decision focuses more on long-term returns, a value closer to 0 indicates a stronger focus on short-term gains.

In the scenario of enterprise dynamic service decision-making, the core role of MDP is to solve the optimal decision strategy by characterizing state transition probabilities and reward functions, achieve multi-objective collaborative optimization, and provide theoretical support for the decision-making of AI agents.

3.2 AI Agent Architecture

An AI agent is an intelligent entity with autonomous perception, decision-making, execution, and learning capabilities. It can perceive environmental states in real time, make decisions independently, execute decision actions, and continuously optimize strategies according to feedback to adapt to changes in the dynamic enterprise service environment.

The AI agent designed in this paper adopts a five-layer architecture:

1. Perception Layer: Collects real-time state data of enterprise service scenarios through data sources such as enterprise ERP systems, customer service management systems, and order management systems, including the number of online customer service representatives, pending workload, customer waiting time, order volume, and other core data.
2. Preprocessing Layer: Performs cleaning, normalization, feature extraction, and other processing on raw data collected by the perception layer, removes abnormal data, screens core features, and converts data into a format suitable for model input to improve data quality.
3. Decision-Making Layer: Based on the improved MDP model, combined with deep reinforcement learning and digital twin technology, analyzes the preprocessed state data, solves the optimal decision action, and outputs specific decision instructions such as resource allocation and service priority.
4. Execution Layer: Connects with the enterprise's existing business systems, executes decision instructions from the decision-making layer, adjusts customer service scheduling, intelligent customer service diversion ratio, order processing priority, etc., and collects real-time feedback data after action execution.
5. Learning Layer: Adopts an online learning mechanism, combines digital twin simulation data and real-time feedback from the execution layer to continuously optimize parameters of deep neural networks and MDP models, improving the adaptability and effectiveness of decision strategies.

3.3 Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) is an integrated technology of deep learning and reinforcement learning. Its core advantage is solving the function approximation problem of traditional reinforcement learning in high-dimensional state spaces through the fitting capability of deep neural networks.

Traditional MDP struggles to accurately represent value functions and policy functions when dealing with high-dimensional continuous state/action spaces. In contrast, deep neural networks can fit complex relationships between high-dimensional states and actions through the nonlinear mapping of multi-layer neurons, without discretizing the state or action spaces.

In this paper, deep neural networks are used to approximate the value function of MDP, and function approximation is adopted to handle high-dimensional continuous state spaces in enterprise services. This effectively breaks through the limitations of traditional MDP and improves the solution capability and decision-making efficiency of the model in complex scenarios.

3.4 Digital Twin Technology

Digital twin technology constructs a virtual model of a physical entity through digital means, realizing real-time mapping and two-way interaction between the physical world and the virtual world.

In enterprise service decision-making, the core application values of digital twin technology are reflected in two aspects: Firstly, it constructs a digital twin simulator of the enterprise service scenario, which accurately reproduces complete service processes such as customer service scheduling, order processing, and resource allocation, simulates system responses under different decision actions, and generates sufficient simulation data, effectively solving the problem of sparse historical data in small and medium-sized enterprises in South China. Secondly, it supports offline training of the model. Through virtual scenarios, the AI agent is extensively trained offline to optimize decision strategies, reduce the cost and risk of online training, and ensure that the model can quickly adapt to actual operational scenarios after deployment.

4 IMPROVED MDP-BASED AI AGENT DECISION-MAKING MODEL

4.1 Core Elements and Limitations of Enterprise Dynamic Service Decision-Making

The core elements of enterprise dynamic service decision-making consist of five dimensions: the decision-making subject (the MDP-AI agent designed in this paper), the decision-making environment (resource status, customer status, and external market environment in enterprise services), decision-making objectives (cost control, customer experience improvement, resource utilization enhancement, and effective problem resolution rate promotion), decision-making constraints (limitations of human resources, equipment, time and other resources), and decision-making actions (resource allocation, service priority setting, and customer demand matching).

Combined with the actual operational scenarios of small and medium-sized enterprises (SMEs) in South China, the limitations of existing decision-making models are mainly reflected in four aspects. First, the high-dimensional continuous state space leads to the combinatorial explosion problem in the traditional MDP model, making efficient solution impossible. Second, the market environment is non-stationary, with sharp fluctuations in customer demand and order volume, coupled with sparse enterprise data, which makes it difficult to accurately calculate state transition probabilities. Third, the inherent delay in the perception-decision-execution loop exists, and partial states (such as potential customer demands) cannot be directly observed, invalidating the full observability assumption of traditional MDP. Fourth, core evaluation indicators (such as customer satisfaction) are lagging indicators, which cannot support real-time decision adjustment and contradict enterprises' demand for rapid response.

4.2 Model Assumptions

Combined with the actual scenario of dynamic service decision-making in small and medium-sized enterprises (SMEs) in South China, this paper revises the idealized assumptions of the traditional Markov Decision Process (MDP) and proposes the following practical model assumptions to ensure the practicability and implementability of the model:

1. The state space S is a high-dimensional continuous space, covering multi-dimensional continuous state variables such as resources, customers, and the environment, without discretization.
2. The action space A is a high-dimensional continuous space, and decision-making actions can be continuously adjusted according to actual scenarios to improve decision-making flexibility.
3. The environment is non-stationary, with market demand, resource constraints and other factors changing dynamically over time. The state transition probability is no longer fixed but adjusted dynamically with time.
4. The state is partially observable, and there exist perception–decision–execution delays; the model is required to have delay compensation capability.
5. The reward function adopts real-time observable proxy indicators (e.g., customer waiting time, service cost, resource utilization) to replace lagged customer satisfaction indicators, supporting real-time decision-making.

4.3 Definition of the Improved MDP-AI Agent Decision-Making Model

Based on the above assumptions, this paper constructs an improved MDP-AI agent decision-making model, which is

still defined in the form of a five-tuple: M' = < S',A',P',R',γ' >. Each component is optimally defined as follows:

1. S' = {s_1, s_2, ..., s_n}: High-dimensional continuous state space, where $s_i \in [0,1]$ represents the i-th normalized state variable, covering three dimensions: Resource state (number of on-duty customer service representatives, average pending tickets per agent, etc.). Customer state (waiting time, demand type, etc.). Environmental state (order fluctuation coefficient, promotion activities, etc.)).
2. A' = {a_1, a_2, ..., a_m}: High-dimensional continuous action space, where $a_j \in [0,1]$ represents the j-th normalized action variable, corresponding to specific decision-making actions such as resource allocation, service priority, and demand matching, ensuring executability.
3. P' = f_θ(s,a): Approximate representation of the state transition distribution. Instead of explicitly calculating state transition probabilities, a deep neural network f_θ is used to fit the state transition distribution, where θ· denotes neural network parameters that can be continuously optimized via online learning.
4. R' = $\Sigma_k=_1^K w_k \cdot r_k(s, a)$:Multi-objective weighted reward function based on real-time proxy indicators, where $w_k$is the weight of the k-th decision objective satisfying $\Sigma_k=_1^K w_k = 1$, and $r_k(s, a)$ is the normalized reward value of the k-th proxy indicator.
5. $\gamma' \in [0,1]$:Discount factor. Considering the operational demand of SMEs in South China to balance short-term profitability and long-term development, γ' = 0.7 is set to balance immediate and future rewards.

4.4 State Representation and Action Design

4.4.1 State representation

Combined with the service scenario of apparel e-commerce SMEs in South China, the state representation covers three core dimensions with 12 state variables, all normalized to [0,1] as inputs to the deep neural network:

1. Resource state: Number of on-duty customer service staff r_staff, average pending tickets per agent r_ticket , number of online intelligent agents r_smart , equipment load rate r_equip, available inventory r_stock;
2. Customer state: Number of waiting customers c_num, average customer waiting time c_wait, demand urgency c_urgent, customer complaint rate c_complain, customer demand type c_type (normalized encoding).
3. Environmental state: Order fluctuation coefficient e_order (current orders / average daily orders), promotion intensity e_promo (normalized, 0 = no promotion, 1 = peak promotion).

4.4.2 Action design

The action design corresponds to core service decision-making behaviors and covers three dimensions to ensure implementability:

1. Resource allocation metrics: Customer service scheduling adjustment ratio a_shift (current scheduling adjustment magnitude), intelligent customer service diversion ratio a_smart (proportion of tickets handled by AI agents), Inventory allocation ratio a_stock;
2. Service priority actions: Emergency order processing weight a_priority, High-value customer service weight a_value;
3. Demand-matching actions: Customer demand and customer service skill matching coefficient a_match, ticket allocation balance a_balance.

4.5 Reward Function Design

Considering the multi-objective collaborative decision-making demand of SMEs in South China, three core real-time proxy indicators—service cost, customer waiting time, and resource utilization—are selected to construct a multi-objective weighted reward function, balancing cost control, customer experience, and resource efficiency:

$$R'(s,a)=w_1 \cdot r_1(s,a)+w_2 \cdot r_2(s,a)+w_3 \cdot r_3(s,a) \quad (1)$$

The parameters are defined as follows:

1. Weighting criteria: Based on the needs of apparel e-commerce enterprises in South China, the weights were determined using the analytic hierarchy process (AHP), with w_1=0.35 (service cost control), w_2=0.4 (customer wait time), w_3=0.25 (resource utilization rate), satisfying $w_1 + w_2 + w_3 = 1$;
2. Cost incentive r_1(s, a): $r_1(s,a)=\frac{C_{max}-C(s,a)}{C_{max}-C_{min}}$,denotes the current service cost, and C_max and C_min represent the historical maximum and minimum service costs of the enterprise in 2025. A higher r_1 value indicates better cost control effectiveness.
3. Waiting time reward r_2(s, a): $r_2(s,a)=\frac{W_{max}-W(s,a)}{W_{max}-W_{min}}$,W(s,a) represents the current customer average wait time, and W_max and W_min represent the historical maximum and minimum waiting times for corporate clients in 2025, respectively. A higher r_2 value indicates superior customer experience.
4. Resource utilization reward r_3(s, a): $r_3(s,a)=\frac{U(s,a)-U_{min}}{U_{max}-U_{min}}$,U(s, a) denotes the current customer service resource utilization rate, with U_max and U_min representing the historical peak and trough values of customer service resource utilization for the enterprise in 2025, respectively. A higher r_3 indicates greater resource utilization efficiency.

4.6 State Transition Approximation

Instead of explicitly calculating state transition probabilities in traditional MDP, a deep neural network is adopted to fit the state transition distribution, solving the difficulty of statistical state transition probabilities under high-dimensional continuous state space and non-stationary environment. The state transition approximation model is defined as:

$$P'=f_\theta(s,a) \tag{2}$$

Here, f_θ denotes a three-layer fully connected neural network whose input is a concatenation vector of the current state s and action a, and whose output is the predicted value of the next state s'. The neural network parameters θ, the loss function is defined as follows:

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N} |f_\theta(s_i,a_i)-s_i'|^2 \tag{3}$$

Here, N denotes the number of training samples, and (s_i, a_i, s_i') represents samples where the system transitions to state s_i after selecting action a_i under state s_i'. The sample was derived from digital twin simulation data generated via the 'LiQICloud' AI-powered Sci-Tech Innovation Policy Platform (supported by Shenzhen Qicheng Zhiyuan Network Technology Co., Ltd.) and the enterprise's real-time operational data for 2025.

5 IMPROVED VALUE ITERATION ALGORITHM BASED ON FUNCTION APPROXIMATION

5.1 Limitations of traditional algorithms

The traditional value iteration algorithm serves as a classic approach for solving MDP optimal strategies, operating by iteratively updating the value function until convergence conditions are met to derive the optimal policy. However, this algorithm exhibits four fundamental limitations in dynamic service decision-making scenarios, rendering it incompatible with the improved model proposed in this study: Firstly, its inability to handle high-dimensional continuous state/action spaces necessitates discretization of states/actions, which compromises decision-making accuracy. Secondly, its slow convergence speed requires extensive iterations to achieve stability, failing to meet real-time decision-making demands in enterprise environments. Thirdly, fixed convergence thresholds frequently yield local optima, making it ill-suited for dynamic non-stationary environments. Lastly, its reliance on explicit state transition probabilities creates a disconnect with the state transition approximation model adopted herein, rendering direct application impractical.

5.2 Improved Algorithm Design

To address the limitations of traditional value iteration algorithms, this study proposes an improved value iteration algorithm based on function approximation by integrating an enhanced MDP model with function approximation techniques. The algorithm employs deep neural networks to fit value functions, refines convergence criteria, and enhances computational efficiency, accuracy, and adaptability. The core steps are as follows:

1. Initialization settings: Initialize deep neural network parameters θ, set initial convergence threshold ε_0=0.001, maximum iteration count T_max=1000, iteration count t=0, and discount factor γ'=0.7;
2. Sample collection: Using the digital twin simulator and real-time operational data from the enterprise in 2025, we collected a status sample set S_sample = {s_1, s_2, ..., s_N}, N = 10000, ensuring coverage of various scenarios, including peak promotion periods and non-promotion periods.
3. Optimal action selection: For each state s_i ∈ S_sample, the action value function Q(s_i, a) of all actions a is computed via a deep neural network, and the optimal action a* = argmax_a Q(s_i, a) is selected.
4. Value function update: Based on the optimal action a* and state transition approximation model f_θ(s_i,a*), predicts the next state s_i', The deep neural network parameters θ are updated by minimizing the following loss function to optimize the value function: $L(\theta)=\frac{1}{N}\sum_{i=1}^{N} \left[R'(s_i,a^*)+\gamma' \cdot V_\theta(s_i')-V_\theta(s_i)\right]^2$, where V_θ(s) denotes the value function fitted by the deep neural network.
5. Convergence criterion: Calculate the parameter variation Δθ = |θ_new - θ_old|, if Δθ < ε_0 or t ≥ T_max terminate iteration and output the optimal parameters θ and optimal decision strategy π* = argmax_a Q_θ(s,a). Otherwise, t = t + 1 and return to Step 3 to continue iteration.
6. Online optimization: After model deployment, the system integrates real-time operational data from enterprises and employs an online learning mechanism to continuously update neural network parameters θ, dynamically adjust optimal strategies, and adapt to non-stationary environmental changes.

5.3 Advantages Analysis of Algorithms

Compared to traditional value iteration algorithms, the proposed improved algorithm demonstrates three core advantages: First, it employs function approximation techniques by using deep neural networks to fit value functions and state transition distributions, eliminating the need for discretization of state/action spaces. This effectively addresses the challenge of solving high-dimensional continuous spaces and enhances decision-making accuracy. Second, it optimizes convergence criteria by using neural network parameter variation as the convergence metric, accelerating algorithm convergence while reducing iteration counts by over 40% compared to conventional methods, thus meeting real-time decision-making demands for enterprises. Third, integrated with online learning mechanisms, the algorithm dynamically adapts to environmental fluctuations. Combined with digital twin technology, it minimizes reliance on

historical data and effectively accommodates the data sparsity characteristics of small and medium-sized enterprises in South China.

6 EMPIRICAL VERIFICATION AND RESULT ANALYSIS

In this empirical study, a small and medium-sized apparel e-commerce enterprise in South China is selected as the experimental object. The enterprise has a customer service team of 25 staff, with an average daily order volume of 1,500–2,200 orders. During the 2025 peak promotion period, the order volume reached 4,000 orders. Its core service scenarios include order consultation, after-sales processing, and logistics inquiry.

Based on the enterprise's actual operational data from January to October 2025, and leveraging the simulation capabilities of the 'LiQICloud' AI-powered Sci-Tech Innovation Policy Platform (Shenzhen Qicheng Zhiyuan Network Technology Co., Ltd.), a digital twin service scenario simulator is constructed to simulate both promotion and non-promotion periods, so as to compare the performance of three decision-making methods: the improved MDP-AI agent model proposed in this paper, the traditional MDP model, and manual decision-making. A total of 10,000 valid data samples are collected and divided into a training set and a validation set at a ratio of 8:2.

The improved MDP-AI agent adopts a three-layer fully connected deep neural network structure: the input layer has 12 neurons corresponding to 12 state variables; the hidden layer has 64 neurons; the output layer has 7 neurons corresponding to 7 action variables. The ReLU function is used as the activation function, Adam is adopted as the optimizer, and the learning rate is set to 0.001.

Two control groups are designed: One is the traditional MDP model with discrete state and action spaces, state transition probabilities calculated from historical data, and customer satisfaction as the core reward function. The other is the manual decision-making method, where strategies are formulated by the enterprise's customer service supervisors and operators based on experience.

Five core indicators are selected for quantitative comparison: service cost, average customer waiting time, customer service resource utilization rate, effective problem resolution rate, and long-term cumulative reward.

The average data during the 7-day promotion period is used as the comparison basis. The core operational indicators of the three decision-making methods are shown in the Table 1 below:

Table 1 Comparison of Core Operational Indicators Across Three Major Decision-Making Methods

Decision-making approach	Service cost (CNY/day)	Average customer wait time (minutes)	Customer service resource utilization (%)	Effective problem-solving rate (%)	Long run cumulative reward
IMDP-AI (This model)	8,920	7.6	84.1	93.2	86.4
Traditional MDP	11,530	17.1	73.3	87.3	72.8
Artificial decision making	13,870	23.4	69.4	87.3	62.5

From the average data over the 7-day major promotion period, the improved MDP-AI agent model proposed in this paper achieves a daily average service cost of 8,920 yuan, an average customer waiting time of 7.6 minutes, a customer service resource utilization rate of 84.1%, an effective problem resolution rate of 93.2%, and a long-term cumulative reward of 86.4. All indicators are significantly superior to those of the traditional MDP model and manual decision-making.

Specifically, compared with the traditional MDP model, the service cost is reduced by 22.8%, and by 35.7% compared with manual decision-making. The average customer waiting time is shortened by 56.1% compared with the traditional MDP model and by 67.5% compared with manual decision-making. The customer service resource utilization rate is increased by 14.7% compared with the traditional MDP model and by 21.2% compared with manual decision-making. The effective problem resolution rate is improved by 5.9 percentage points. The long-term cumulative reward is increased by 18.7% compared with the traditional MDP model and by 38.2% compared with manual decision-making.

Moreover, the model also outperforms the control groups in non-promotion scenarios, which fully verifies its adaptability and stability under different business scenarios. It can effectively solve the pain points of dynamic service decision-making for small and medium-sized enterprises in South China and has strong practical application and deployment value.

7 CONCLUSIONS AND FUTURE WORK

This paper addresses the pain points in dynamic service decision-making for SMEs in South China, such as high-dimensional continuous states, non-stationary environments, sparse data, and decision lag. Based on actual enterprise operation data in 2025, the traditional Markov Decision Process model is innovatively improved, and an AI agent decision-making framework integrating deep reinforcement learning and digital twin technology is constructed.

An improved value iteration algorithm based on function approximation is proposed. Through theoretical derivation and empirical verification, the core conclusions are drawn as follows:

The idealized assumptions of the traditional MDP model are inconsistent with the actual service scenarios of SMEs in South China. The introduction of function approximation, deep reinforcement learning, and digital twin technology can effectively break through the limitations of traditional models and improve model adaptability and decision-making efficiency.

The constructed improved MDP- AI agent decision- making model can accurately characterize the dynamic service scenarios of enterprises through the design of high-dimensional continuous state and action spaces, the construction of a multi-objective real-time proxy reward function, and the approximate fitting of state transitions, so as to achieve multi-objective collaborative optimization of cost control, customer experience improvement, and resource utilization enhancement.

The proposed improved value iteration algorithm based on function approximation has faster convergence speed and higher solution accuracy than traditional algorithms. It can adapt to non- stationary environments and sparse data scenarios, and meet the real-time decision-making requirements of SMEs.

Empirical results show that in the scenario of small and medium-sized apparel e-commerce enterprises in South China, the model can reduce service cost by 22.8%, shorten customer waiting time by 56.1%, improve resource utilization by 14.7%, and increase effective problem resolution rate by 5.9 percentage points, providing an implementable solution for service decision- making optimization in the digital transformation of SMEs.

Combined with the research results and the development needs of SMEs in South China, future research can be deepened in five aspects to improve the universality and practicability of the model:

Extend the model to more service-intensive industries in South China such as cosmetics, home furnishing, and logistics, optimize state representation, action design, and reward functions according to industry characteristics, and verify its universality.

Incorporate long-term value indicators such as customer repurchase rate, customer unit price, and brand reputation to build a more complete multi-objective evaluation system and realize collaborative optimization of short-term operational indicators and long-term development goals.

Integrate edge computing technology to deploy the model on edge devices, further improve decision-making real-time performance, reduce deployment costs, and adapt to the IT infrastructure of SMEs.

Develop a lightweight version of the model, simplify the structure, lower the threshold of training and deployment, and adapt to the operation needs of small teams with fewer than 30 employees to improve popularity.

Formulate an implementation roadmap of "small steps and fast iterations", conduct a one-month trial run in non-promotion periods first, collect feedback to optimize parameters, and then gradually promote to full scenarios to reduce deployment risks and help more SMEs in South China complete digital transformation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

ACKNOWLEDGMENTS

The authors would like to thank Shenzhen Qicheng Zhiyuan Network Technology Co., Ltd. for providing intelligent computing power and data simulation support through the 'LiQICloud' AI-powered Sci-Tech Innovation Policy Platform, which facilitated the model construction and empirical verification in this study.

REFERENCE

[1] Chen X W, Wang T, Barrett W T, et al. Same-day delivery with fair customer service. European Journal of Operational Research, 2023, 308(2): 738-751. DOI: 10.1016/J.EJOR.2022.12.009.

[2] Klein V, Steinhardt C. Dynamic demand management and online tour planning for same-day delivery. European Journal of Operational Research, 2023, 307(2): 860-886. DOI: 10.1016/J.EJOR.2022.09.011.

[3] Xie C, Waller S T. Parametric search and problem decomposition for approximating Pareto-optimal paths. Transportation Research Part B, 2012, 46(8): 1043-1067. DOI: 10.1016/j.trb.2012.03.005.

[4] Ariely D, Bitran G, Rocha e Oliveira P. Design to learn: customizing services when the future matters. Pesquisa Operacional, 2013, 33(1): 37-61. DOI: 10.1590/S0101-74382013000100003.

[5] Zhang J, Van Woensel T. Dynamic vehicle routing with random requests: A literature review. International Journal of Production Economics, 2023, 256: 108751. DOI: 10.1016/J.IJPE.2022.108751.

[6] Mausam, Kolobov A. Planning with Markov decision processes: An AI perspective. 2012. DOI: 10.2200/S00426ED1V01Y201206AIM017.

[7] Voccia S A, Campbell A M, Thomas B W. The same-day delivery problem for online purchases. Transportation Science, 2019, 53(1): 167-184.

[8] Ulmer M W, Goodson J C, Mattfeld D C, et al. On modeling stochastic dynamic vehicle routing problems. EURO Journal on Transportation and Logistics, 2020, 9(2): 100008. DOI: 10.1016/j.ejtl.2020.100008.