Journal of Computers Vol. 36 No. 5, October 2025, pp. 1-20
https://doi.org/10.63367/19911599202510360500 1

Network Access Selection Algorithm Based on Simulated Annealing and Q-Learning in Space-Air-Ground Integrated Network

Chen-Wei Feng, Ya-Wei Sun*, and Xiao-Xin Wu

School of Opto-electronic and Communication Engineering, Xiamen University of Technology, Xiamen, 361000, China

cwfeng@xmut.edu.cn, sunyw22@163.com, 1289819501@qq.com

Received 1 September 2024; Revised 14 January 2025; Accepted 25 January 2025

Abstract. With the explosive growth in user numbers and business demands, the limitations of ground network coverage have become increasingly prominent, and its limited resources are not enough to meet the overloaded access requests in hotspot areas. Considering the characteristics of satellites with wide-area coverage, the rapid deployment of UAV lift-off base stations can effectively cope with the sudden changes in user demand and alleviate the load. Therefore, based on the Space-Air-Ground integration scenario, this paper improves and optimizes the Q-learning algorithm in network access decision-making with the help of the Metropolis criteria in the simulated annealing algorithm, improves the convergence speed of the algorithm, and realizes the access selection of heterogeneous wireless network.

Keywords: network access selection, Q-learning, simulated annealing, space-air-ground integrated network

1 Introduction

In recent years, the rapid development of wireless network technology has given rise to diversified application scenarios. Considering the differentiated service requirements and the heterogeneity of networks, multiple networks are required to integrate with each other so as to collaborate to guarantee the quality of user communications. In addition, the terrestrial mobile communication system, limited by network capacity and coverage, has been difficult to adapt and meet the ubiquitous demand for seamless full-coverage connectivity services in modern society [1]. In some hot spots, such as large stadiums, concerts and other crowded places, the limited communication resources of the network will not be able to cope with the overloaded access requests, which will easily cause network congestion, communication paralysis and other serious consequences. At the same time, sudden natural disasters may destroy the normal communication of the infrastructure, making it difficult for users to carry out timely communication services.

In view of the challenges faced by terrestrial mobile communications, in order to break through the limitations of traditional networks and optimize the overall communication performance, the Space-Air-Ground Integrated Network (SAGIN) is formed based on the terrestrial network and the introduction of satellites and un-manned aerial vehicles (UAVs) to improve the adaptability in a variety of complex communication scenarios, which is regarded as a key trend in the future development of communication [2, 3]. The Space-Air-Ground integrated network mainly consists of three parts: space-based, air-based and ground-based. Its structure is shown in Fig. 1. Satellite networks can provide ubiquitous communication services to a large number of terminals. In aerial networks, high-altitude platforms and drones can serve as relays to improve long-distance communications and flexibly meet business needs. Terrestrial networks can provide users with real-time and high-speed communication services within limited coverage. The integration of space and space can provide wide area coverage, flexible mobility, and safe and reliable services, which is an inevitable trend in the development of heterogeneous wireless networks (HWNs).

Different types of wireless networks have their own performance characteristics, such as significant differences in latency, mobility support and reliability. The uniqueness and complementarity of each RAT is utilized to provide multiple access networks and multiple network services for a large number of user devices [4]. Although heterogeneous fusion networks are capable of supporting large-scale data transmission requirements, the surge in data requests can cause network congestion, and the distribution of different resources in the network is usually

* Corresponding Author

managed in a dynamic manner. Therefore, the design of efficient and intelligent access selection strategy not only has important theoretical value, but also has far-reaching significance for improving the performance of wireless communication system and optimizing the user experience. How to choose the right network for users to access to ensure the optimization of communication quality and user experience, and realize the load balance of the network at the same time, has also become a hot spot of current research.

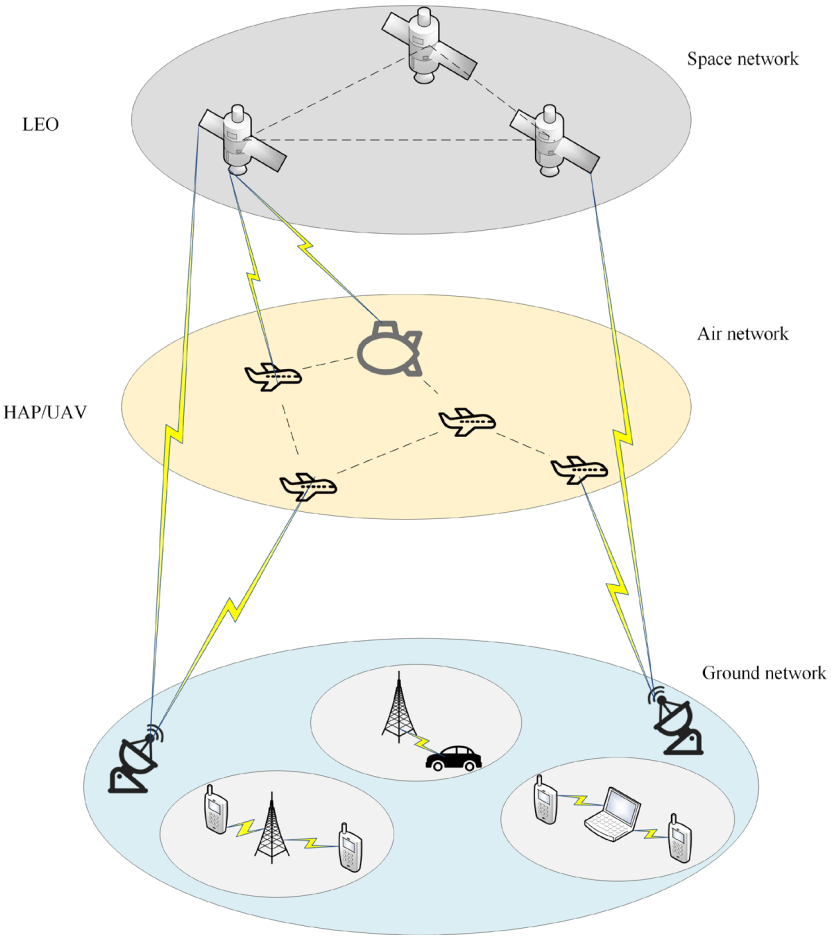

Fig. 1. Space-Air-Ground integrated structure diagram

Intelligence is an important feature of the 5G network, and the application of AI techniques in HWN can provide a new way of thinking for network access algorithms, making network access algorithms more efficient and intelligent. Different from the necessity of manual collection and annotation of samples in supervised learning, Reinforcement Learning (RL), as an important element in the field of AI and Machine Learning, has the ability of agents to make trial and error in their environment, so as to gradually master and achieve the action strategies required for the set goals [5]. Therefore, RL, with its excellent self-learning and adaptive capabilities, has been widely used in the selection of Vertical Handoff (VHO) strategies in heterogeneous wireless network environments [6].

Q-learning uses the interaction between the agent and the environment to iteratively learn the optimization strategy, so as to achieve the optimization goal. It can effectively solve the network selection problem in the heterogeneous network environment, and has been widely used. References [7] and [8] are based on traditional RL algorithms for access selection. In [7], the researcher proposed an intelligent handover scheme driven by Quality of Experience (QoE) in view of the characteristic of rapid changes in satellite networks. This scheme solves the optimization problem involving multiple metrics through RL to reach the best satellite network switching choice. In [8], based on the multi-armed bandit model, different reward functions are designed by distinguishing between

real-time and non-real-time user services, which helps to maximize the long-term benefit of the network and directs users to connect to the most appropriate network. In [9], a dynamic network selection algorithm based on QoE is proposed for QoE and network energy consumption by using the online dynamic learning characteristics of Q-learning method and considering user QoE, network energy consumption and handover time comprehensively. The algorithm can dynamically select the network, obtain the maximum user QoE, and optimize the network energy consumption and handover time.

Q-learning algorithm, as a classical algorithm in RL, has achieved remarkable results in practical applications, but it also has some limitations, such as easy to fall into local optimal solution and low learning efficiency. In order to improve these shortcomings, researchers have proposed a variety of improved methods combined with Q-learning algorithm or the distributed Q-learning algorithm. In [10], a handover decision algorithm combining TOPSIS and Q-learning is proposed. TOPSIS is used to reduce the size of the action space in Q-learning. Simulations show that the proposed algorithm enhances the handover performance and has scalability. In [11], an online multi-agent Q-learning algorithm is proposed for performing real-time load balance, user association and handover in dense cellular networks. The algorithm has the characteristics of low complexity and fast convergence, which can adapt to the dynamics of the network and reduce the handover frequency in real time. In [12], a multiagent Q-learning network selection algorithm based on Nash multiagent Q-learning is proposed, which can learn the joint optimal selection policy to improve system throughput and reduce user blocking while ensuring the requirements of IoT services. In [13], a joint relay selection and access control scheme based on distributed Q-learning is proposed for IoT-oriented satellite terrestrial relay networks with massive IoT devices and multiple relays. In [14], a novel RL method is investigated to reduce the handover failure problem by learning the current policy action to determine whether a fast moving vehicle needs to be handed over or not. In addition, it combines fuzzy logic and graph hierarchical structure to select the correct network, so as to select an efficient and suitable network on the vehicular communication, eliminating the difficulty of network selection.

For network selection in different application scenarios, the traditional RL algorithm has been improved and extended, and deep reinforcement learning (DRL) has been widely adopted. In [15], a DRL scheme based on UE-driven is proposed, in which a centralized agent deployed on the backhaul side of the NT-BS is responsible for training the parameters of the deep Q-network (DQN), and each UE independently makes its own access decision based on the parameters of the trained DQN. With the proposed scheme, each UE is able to intelligently access the appropriate non-terrestrial base station (NT-BS) to improve the long-term sys-tem throughput and avoid frequent handovers among NT-BSs. In [16], a new DRL model is proposed for solving the network selection problem, aiming to optimize the delivery of healthcare data in heterogeneous healthcare systems in order to minimize the transmission energy consumption and latency while meeting the QoS requirements of various applications. In [17], aiming to improve the QoS for cognitive users, an online network selection method of CRNs with multiple primary networks based on DRL is proposed based on the RL method and the dynamic spectrum access results of cognitive radio in a single primary network environment. In [18], a network selection system model in SAGIN is designed and a network selection algorithm based on evolutionary game is proposed to study the autonomous decision-making process of network selection. In addition, a network selection algorithm based on deep deterministic policy gradient is proposed to deal with continuous and high-dimensional action spaces. In a comprehensive comparison, the algorithm is superior to the repeated game and PPO algorithm.

Although the existing network access selection algorithms are effective in improving user experience to a certain extent, most of them only focus on optimizing the specific performance of the network system in a single service type scenario, or simplify the characteristics of different services, ignoring the differentiated service requests. And they fail to fully consider the state of the network at the next moment and the long-term return, and easily fall into local optimal solution. And most of them are for static environment, do not have dynamic adaptability, can not adapt to the development of intelligence and difficult to realize the full use of network resources. Some of the algorithms are more complex and require a large amount of data and computational resources during the training process, especially for complex tasks where the training time and cost increase significantly.

SAGIN has the advantages of wide coverage, high throughput and strong elasticity, which can meet a variety of services and applications with different quality of service requirements in different scenarios. The emergence of the SAGIN structure effectively addresses the shortcomings existing in the current field of wireless communication. Therefore, in this paper, an access selection algorithm based on simulated annealing and Q-learning is proposed for the access selection problem in the SAGIN heterogeneous network scenario. The contributions of this paper are summarized as follows:

(1) The Heterogeneous Networks (HetNet) system model and end-user mobility model are established, and a variety of 5G new service requirements are considered. The access selection is modeled as MDP, and the network state is described as different service types and the status of each network resource, where the network resources

are not completely evenly divided to make a more detailed judgment when the network resources tend to be saturated.

(2) In order to balance the QoS requirements of users and the network load, the reward function is constructed by using the Analytic Hierarchy Process (AHP) and Gray Relation Analysis (GRA).

(3) The Metropolis criterion in the simulated annealing algorithm is used to dynamically adjust the action decision of the agent under multi-network overlap, and the network access selection mechanism of Q-learning is optimized, in order to avoid falling into the local optimum. Thus, the intelligence of access selection is realized.

The rest of the paper is organized as follows: Section II describes the HetNet system model and the channel model. Section III describes in detail the access selection algorithm flow for the heterogeneous wireless network combining simulated annealing and Q-learning. Section IV evaluates and analyzes the performance of the proposed algorithm through simulation experiments. Finally, the paper is summarized.

2 System Model

2.1 Heterogeneous Network Model

As shown in Fig. 2, LEO, UAV, LTE-A (MBS) and 5G micro base station (SBS) constitute a scenario of space-to-earth heterogeneous wireless network coexistence, in which the gray area is a hot spot area. Considering the large coverage of the satellite, it is assumed that its position remains unchanged at a fixed altitude in the center of the network area; The drone is equipped with a mobile base station and moves back and forth along a specific trajectory at a fixed height in the hotspot area; Assuming that the user terminals are all multimodal mobile terminals, with an initial random distribution and randomly changing the direction of motion at regular intervals; The ground base station consists of LTE-A and 5G micro base stations. In addition, four new 5G services were considered, namely 4K/8K high-definition video service, telemedicine service, big data collection service and office automation service. Different business types not only have different requirements for network attributes such as latency and packet loss rate, but also have varying degrees of requirements for security and support for mobility. Fig. 3 depict the specific requirements of the four new 5G services mentioned in this paper in terms of latency, reliability, security, and mobility. Among them, 0-5 represents the degree of requirement for the attribute, and the larger the value, the stricter the requirement for the attribute. Considering the large number of users in hot spots and the most complex network selection, this paper mainly analyzes the user access selection in hot spots.

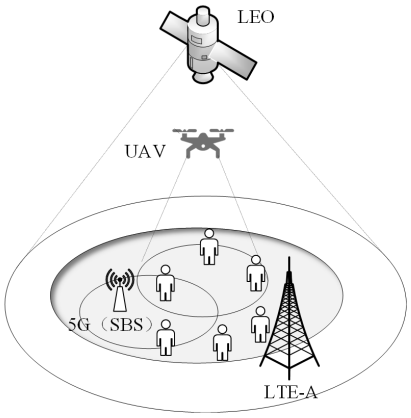

Fig. 2. Heterogeneous network model

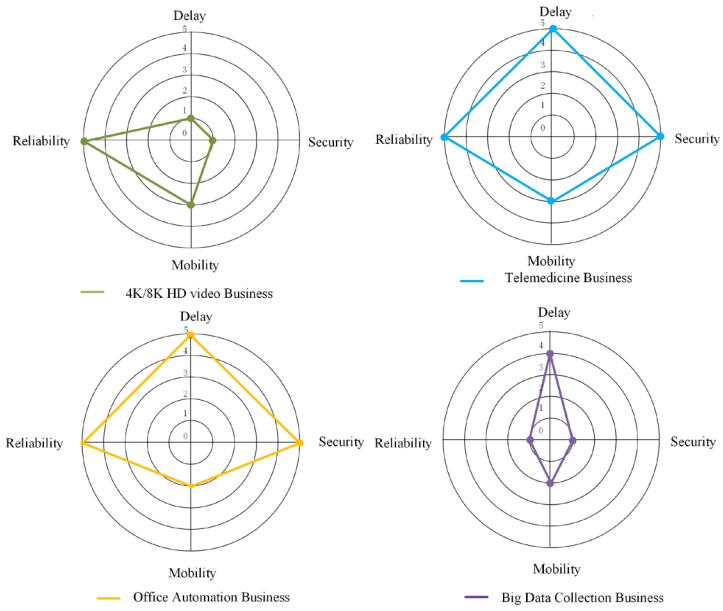

Fig. 3. Attribute requirements for four new 5G services

2.2 Channel Model

In the network discovery phase, users move within the network coverage area and periodically collect network reception signals. When the signal strength received by the network is higher than the specified threshold, it means that the network can provide normal business services to users, and vice versa. The user received power P_r is calculated as follows

$$P_r = P_t - PL. \tag{1}$$

Among them, P_t is the transmit power of the base station, and PL is the path loss value.

Assume that the simulation scene in this paper is located in a suburban hotspot area, and different signal propagation models are used to consider the performance of various communication equipment. Since ground base stations are arranged in suburban areas and the communication distance between satellites and ground equipment is usually long, the line-of-sight communication environment can usually be described by the free space loss model [19], so consider using this model to evaluate the coverage and signal quality of satellite platforms and ground networks; Considering that UAVs usually fly in the air and are greatly affected by factors such as terrain, vegetation, weather, etc., their path loss model is composed of two parts: free space loss and additional loss. Due to the complexity of real-world scenarios, it is difficult to distinguish whether the channel belongs to a line of sight link or a non line of sight link. Therefore, considering the average path loss, the calculation formula is as follows

$$PL = P_{LoS} \times PL_{LoS} + P_{NLoS} \times PL_{NLoS}. \tag{2}$$

$$P_{LoS} + P_{NLoS} = 1. \tag{3}$$

Among them, P_{LoS} is the probability of line-of-sight transmission, and P_{NLoS} is the probability of non-line-of-sight transmission. PL_{LoS} and PL_{NLoS} are the path losses corresponding to the two channels respectively. Their calculation formula is as follows

$$PL_{LoS} = 20log(\frac{4\pi f_c d}{c}) + \eta_{LoS}. \tag{4}$$

$$PL_{NLoS} = 20log(\frac{4\pi f_c d}{c}) + \eta_{NLoS}. \tag{5}$$

Among them, f_c is the carrier frequency, c is the speed of light, η_{LoS} and η_{NLoS} are the additional losses corresponding to the two channels.

According to the International Telecommunications Union definition, P_{LoS} can be represented by the following probability function

$$P_{LoS} = \frac{1}{1 + a \cdot \exp\left(-b \cdot \arctan\left(\theta\right) - a\right)}. \tag{6}$$

$$\theta = \arctan(\frac{h}{r}). \tag{7}$$

Among them, parameters a and b are determined by the environment, θ is the elevation angle, h is the height of the drone, and r is the horizontal distance from the drone to the terminal.

Table 1. Comparison table of different environmental parameters

	Suburbs	City
a	4.88	9.61
b	0.43	0.28
η_{LoS}	0.1	1
η_{NLoS}	21	20

The final average path loss PL can be expressed by the following formula, where the specific parameter values are shown in Table 1.

$$PL = \frac{\eta_{LoS} - \eta_{NLoS}}{1 + a\exp\left(-b\left(\theta - a\right)\right)} + 10\log\left(h^2 + r^2\right) + 20\log f_c + 20\log(\frac{4\pi}{c}) + \eta_{NLoS}. \tag{8}$$

3 An Access Selection Algorithm for the Heterogeneous Wireless Network Combining Simulated Annealing and Q-learning

3.1 Design of the Network Selection Scheme Based on Q-learning

Q-learning is a model-free reinforcement learning algorithm that does not require complete knowledge of the environment, and constantly learns how to effectively match states with actions through trial and error to maximize returns, with greater versatility and a wide range of applications. In Q-learning, $Q(s, a)$ represents the sum of the reward obtained by performing action a in the current state s and the expected discounted reward from subsequent decisions. In the case of unknown reward function or state transition probability, the agent only considers each available action a in the current state s, uses the reward value obtained after executing the action to calculate the Q value of the action and updates the Q matrix. Afterwards, the intelligent agent continuously tries and

learns, iteratively optimizing the Q-matrix until all elements converge. Finally, regardless of the starting state, the agent can successfully achieve a specific goal based on the converged Q matrix. According to the Markov state transition model, the Q-value iteration formula is as follows

$$\begin{aligned} Q^{\pi}(s,a) &= E_{\pi}[\sum_{i=0}^{\infty}\gamma^{i} r_{t+i} \mid s_t = s, a_t = a] \\ &= E_{\pi}[r_t \mid s_t = s, a_t = a] + E_{\pi}[\sum_{i=1}^{\infty}\gamma^{i} r_{t+i} \mid s_t = s, a_t = a] \\ &= R(s,a) + \gamma E\left[\sum_{i=0}^{\infty}\gamma^{i} r_{t+1+i} \mid s_t = s, a_t = a\right] \\ &= R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') Q(s',a'). \end{aligned} \tag{9}$$

Among them, $R(s, a)$ is the mathematical expectation for immediate return r_t, and $P(s, a, s')$ is the probability of transitioning to the next state s' after executing action a in state s. The optimal value action function $Q^*(s, a)$ for the Q function is shown below.

$$\begin{aligned} Q^{*}(s,a) &= \max_{\pi} Q^{\pi}(s,a) \\ &= R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') \max_{a' \in A} Q(s',a'). \end{aligned} \tag{10}$$

According to the principle of Bellman optimality, the state value under the optimal strategy should be equal to the expected benefit of taking the optimal action in that state. In other words, once $Q^*(s, a)$ is obtained, the optimal strategy for a given arbitrary state $s \in S$ is to choose the action that maximizes $Q^*(s, a)$.

$$\pi^{*} = \arg\max_{a \in A} Q^{*}(s,a), \forall s \in S. \tag{11}$$

Q-learning updates the action value function, and its update rules are as follows

$$\begin{aligned} Q(s,a) &= (1-\alpha) Q(s,a) + \alpha(r(s,a) + \gamma \max_{a' \in A} Q(s',a')) \\ &= Q(s,a) + \alpha(r(s,a) + \gamma \max_{a' \in A} Q(s',a') - Q(s,a)). \end{aligned} \tag{12}$$

Among them, α is the learning rate, with a value ranging from $0 \le \alpha \le 1$, which determines the proportional relationship between newly acquired experience and current experience. If α is 1, it means that the agent only pays attention to the latest information and ignores previous experience or knowledge; When α is 0, it means that the agent will not learn. a' is the set of feasible strategies in the new state.

In heterogeneous wireless networks, different access technologies have their own characteristics, and they jointly provide users with more extensive and flexible access options. Multi-attribute decision making aims at comprehensively evaluating multiple unrelated attributes of the same object, so as to realize a comprehensive consideration of the performance of the object. Therefore, in this paper, after using the multi-objective decision-making method based on AHP to obtain the network attribute weights, GRA is introduced to enhance the objectivity of AHP decision. GRA is a technology specially proposed to deal with the problems of incomplete information and uncertainty in the decision-making process [20]. It uses the "gray correlation degree" index to evaluate the degree of similarity and difference between factors in the dynamic process. In this theory, the interdependence between factors is quantified by correlation, meaning that the greater the similarity, the more related factors.

(1) Construct a multi-attribute decision matrix

By collecting the delay, packet loss rate, mobility support and security of each candidate network, an $m \times n$ decision matrix D is constructed. The structure is as follows

$$D_{m\times n}=\begin{bmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{m1} & \cdots & d_{mn} \end{bmatrix}. \tag{13}$$

In the above formula, d_{mn} represents the value of attribute n in network m.

(2) Network parameter normalization

By transposing the multi-attribute decision matrix D, the comparison matrix D' is obtained. Since the units of each attribute of the network are different, in order to eliminate the dimensional effects of different attributes, the matrix D' is normalized. According to the expected value of network attributes, these attributes are divided into two categories: benefit type and cost type. Benefit indicators tend to be better for larger values, while cost attribute indicators are based on the criterion of smaller values. Here, the minimum-maximum method is used to normalize these two types of parameters.

$$a_{ij}=\frac{d_{ij}-\min\limits_{i} d_{ij}}{\max\limits_{i} d_{ij}-\min\limits_{i} d_{ij}}, i=1,2,...,m; j=1,2,...,n. \tag{14}$$

$$a_{ij}=\frac{\max\limits_{i} d_{ij}-d_{ij}}{\max\limits_{i} d_{ij}-\min\limits_{i} d_{ij}}, i=1,2,...,m; j=1,2,...,n. \tag{15}$$

Among them, a_{ij} represents the normalized value of the corresponding attribute element.

(3) Calculate the correlation coefficient to determine user preferences for different business types

By taking the maximum value of the benefit attribute index and the minimum value of the cost attribute index, the optimal value of each attribute index in each candidate network is determined, and the reference sequence element value is obtained.

$$a_{0j}=\max_{i}\left\{a_{ij}\right\}, i=1,2,...,m; j=1,2,...,n. \tag{16}$$

$$a_{0j}=\min_{i}\left\{a_{ij}\right\}, i=1,2,...,m; j=1,2,...,n. \tag{17}$$

The gray correlation coefficient is calculated through the absolute value, minimum value and maximum value of the difference between each attribute element and the reference sequence element.

$$R=\frac{\min\limits_{i}\min\limits_{j}\left|a_{0j}-a_{ij}\right|+\lambda\max\limits_{i}\max\limits_{j}\left|a_{0j}-a_{ij}\right|}{\left|a_{0j}-a_{ij}\right|+\lambda\max\limits_{i}\max\limits_{j}\left|a_{0j}-a_{ij}\right|}, i=1,2,...,m; j=1,2,...,n. \tag{18}$$

In the formula, λ is the discrimination coefficient. The smaller its value, the greater the difference between correlation coefficients. Its value is usually set to 0.5.

Use AHP to calculate the weight of network attributes under different services, then use GRA to calculate the correlation coefficient between candidate network attributes and ideal network attributes, and weight the network preferences of users of different business types to improve the objectivity of AHP decision-making to a certain extent.

Q-learning is a value based reinforcement learning algorithm that focuses on learning optimal decisions through the interaction between agents and the environment. In heterogeneous network scenarios, the Q-learning algorithm is introduced into the problem of network access selection, allowing the agent to dynamically perceive the current terminal device status and network environmental conditions, including signal quality, network con-

gestion, battery life and other factors. Through learning and analysis, the agent can adaptively learn and decide the optimal network access or handover strategy. Integrating Q-learning into heterogeneous network access selection can greatly simplify the user's network access process and effectively improve the management efficiency of the entire network system and the user experience. The core elements in reinforcement learning algorithms include state, action and reward function. Therefore, when designing an access selection algorithm based on reinforcement learning, the key step is to clearly define the state space S, action space A, and reward mechanism R.

(1) State space *S*

In addition to selecting the appropriate network for users and optimizing load balancing, the network access selection algorithm also needs to consider the differences in user business needs. Therefore, when expressing the environmental status, in addition to using the load occupancy rate of the network, the user terminal business status is also added. The system state can be defined as

$$S = \{B, L\}. \tag{19}$$

Among them, *B* represents the type of service currently requested by the user terminal, including 4K/8K high-definition video service, telemedicine service, big data collection service and office automation service. *L* represents the quantified value of the current usage of each network resource. Users will make corresponding network selections under different network loads based on the service type of the call. This paper considers non-completely uniform division of network load, with the purpose of achieving more refined judgment when network resource occupation is close to saturation, so as to make reasonable network choices. As shown in Table 2.

Table 2. Network load quantification level

Network load	$\left[0,\frac{1}{3}\right)$	$\left[\frac{1}{3},\frac{1}{2}\right)$	$\left[\frac{1}{2},\frac{2}{3}\right)$	$\left[\frac{2}{3},\frac{7}{9}\right)$	$\left[\frac{7}{9},\frac{8}{9}\right)$	$\left[\frac{8}{9},1\right]$
Quantitative level	1	2	3	4	5	6

(2) Action space *A*

At each decision-making moment, the user terminal will make network access choices based on its own needs and network conditions. Define *A* as the set of all selectable actions.

$$A = \{a_1, a_2 \ldots a_i\}. \tag{20}$$

Where a_i represents the current user terminal's choice to access network *i*.

(3) Reward function *R*

In the Q-learning algorithm, the reward value is a signal that evaluates the quality of network selection actions in a given state. Therefore, the design of the reward function is a core part of the reinforcement learning algorithm. In order to characterize network QoS more reasonably and comprehensively, this paper selects four parameters: network delay, packet loss rate, mobility support and security. In addition to considering the preferences of different services, the load balancing of the network is also taken into consideration. If at the network access selection moment, the available resources of the network meet the business requirements, the user terminal successfully accesses or switches to the network; Otherwise, it indicates that the access or handover fails, and the user terminal maintains the original connection. If the user terminal has left the coverage of the original connection at this time, the access will fail and the access selection will be made again until the next decision time. Therefore, the reward function is defined as follows

$$r(s, a_i) = \begin{cases} l_i \times \sum_{j=1}^{4} w_j^k \rho_{ij}, & \text{Access network } i \\ 0, & \text{Access failed.} \end{cases} \tag{21}$$

Among them, $r(s, a_i)$ represents the reward value for choosing to access network i. li is the ratio of the available resources of network i to the total resources; $\sum_{j=1}^{4} w_j^k \rho_{ij}$ is the preference value of business k for network i, w_j is the weight of attribute j obtained by using AHP, ρ is the correlation obtained by gray correlation analysis coefficient.

In order to evaluate the long-term return value of the overall network access selection or switching of the system, the cumulative utility u_a is defined as

$$u_a = \sum_t \sum_n r_n (s, a_i). \tag{22}$$

Among them, t represents the various decision moments when the system performs network access selection or switching, and n represents the business calls that exist at the decision moment.

3.2 Strategy Update Mechanism Based on Simulated Annealing

The simulated annealing algorithm is a heuristic stochastic optimization algorithm derived from Monte-Carlo iteration. Its idea was first proposed by the American physicist Metropolis in 1953. During the search process, the local optimum is overcome with a certain attenuated probability, thereby gradually tending to the global optimal solution. The principle comes from the process of cooling an object after it heats up.

The algorithm concept of simulated annealing is abstracted from the annealing phenomenon in physics. When an object is at high temperature, its internal molecules are in a more active disordered state. As the temperature gradually decreases, the molecules in the system begin to transform into an ordered structure, until a stable state is finally formed where the internal energy reaches the lowest level [21]. Based on such characteristics, this physical process is simulated and applied to the problem of finding the optimal solution to the mathematical objective function in a specific solution space. The internal energy of the object is simulated into the objective function value. The simulation process connects the high internal energy state of the material at high temperature with the disordered molecules. At this time, the algorithm has a higher probability of exploring the space to skip the local optimal point; As the temperature gradually decreases, the global random search ability weakens accordingly, while the ability to explore local optimal solutions is strengthened. The molecules in the system eventually turn to an orderly arrangement. This process corresponds to the algorithm finding the global optimal solution to the objective function.

The Metropolis acceptance criterion is the core mechanism for decision-making in the implementation of the simulated annealing algorithm. It receives the new state of the system according to a specific probability and is used to calculate the equilibrium properties of the physical system. The Metropolis criterion defines the internal energy probability of an object from state i to i' at a certain temperature T. The expression is as follows

$$P = \begin{cases} 1 \ , E(i) \le E(i') \\ exp\left(\dfrac{E(i') - E(i)}{kT} \right) \ , E(i) > E(i'). \end{cases} \tag{23}$$

Where $E(i)$ and $E(i')$ are the internal energies in states i to i' respectively, k is Boltzmann's constant, and T is the annealing temperature. It can be seen from the above formula that $E(i) \le E(i')$, the system accepts the new solution i with a probability of 1. On the contrary, the system will accept i with a probability of $exp\left(\dfrac{E(i) - E(i')}{kT} \right)$.

The Metropolis acceptance criterion follows: At higher simulation temperatures, the system tends to accept new states as current values. As the temperature strategy is executed during simulated annealing, that is, the temperature gradually decreases, Metropolis will reduce the probability of accepting new states.

The selection of intelligent agent actions is the key to the learning process and also the core of heterogeneous network access selection. Q-learning algorithms usually adopt the ε-greedy strategy, which aims to balance the exploration and utilization of the agent in the action selection process. Generally speaking, as the number of iterations increases, the Q value obtained by the agent becomes closer and closer to the optimal, and it is no longer reasonable to conduct large-scale exploration at this time. The parameters of the greedy strategy often cannot meet the needs of agents in different exploration stages. Therefore, this paper uses the idea of simulated annealing algorithm to dynamically adjust the action selection mechanism in heterogeneous network access selection.

(1) In the initial learning stage, actions are selected based on a random action strategy, and then actions are selected according to the Metropolis acceptance criteria in simulated annealing. First, the agent randomly selects the action a_r, updates the exploration probability W according to the following formula and generates a random number δ. If $\delta < W$ is satisfied, the random action a_r is selected. Otherwise, the optimal action a_o is selected.

$$W = exp\left(\frac{Q(s,a_r)-Q(s,a_o)}{T}\right). \tag{24}$$

$$\delta = rand\left(0,1\right). \tag{25}$$

In the formula, $Q(s, a_r)$ represents the Q value obtained by the agent selecting action a_r according to the random action strategy in state s; $Q(s, a_o)$ represents the Q value obtained by the agent selecting the optimal action a_o according to the Q table; T is the annealing temperature. The agent's action selection probability vector is represented by $p(i)(i = 1, 2, ..., n)$, then the probability of selecting the i-th action is as follows.

If the i-th action selected is not the best action corresponding to the largest Q value, the probability of being randomly selected is $1/n$. Action a_i is only selected by the agent when the condition $rand(0,1) < W$ is satisfied, so $p(i) = W/n$.

If the i-th action selected is the action with the largest Q value in the current state, it is the optimal action. At this time, the probability value $p(i)$ consists of two parts: first, when the i-th action is randomly selected with probability $1/n$, that is, the Q value of the randomly selected action is equal to the Q value of the optimal action, then $W = 1$, $p(i) = 1/n$. Second, when the i-th action is not randomly selected, the corresponding W value is calculated according to (24). Afterwards, when the condition $rand(0,1) > W$ is met, the action selected by the agent is still the optimal action. At this time, the calculation formula of $p(i)$ is as follows

$$\begin{aligned} p(i) &= \frac{1}{n}\exp(\frac{Q(s,a_i)-Q(s,a_i)}{T}) + \frac{1}{n}\sum_{j=1,j\neq i}^{n}\{1-\exp(\frac{Q(s,a_j)-Q(s,a_i)}{T})\} \\ &= \frac{1}{n} + \frac{n-1}{n} - \frac{1}{n}\sum_{j=1,j\neq i}^{n}\exp(\frac{Q(s,a_j)-Q(s,a_i)}{T}) \\ &= 1 - \frac{1}{n}\sum_{j=1,j\neq i}^{n}\exp(\frac{Q(s,a_j)-Q(s,a_i)}{T}). \end{aligned} \tag{26}$$

The expression for the probability $p(i)$ of the intelligent agent choosing the i-th action can be derived as follows

$$p(i) = \begin{cases} 1 - \frac{1}{n}\sum_{j=1,j\neq i}^{n}\exp(\frac{Q\left(s,a_j\right)-Q\left(s,a_i\right)}{T}), a_i = a_o \\ \frac{1}{n}\exp(\frac{Q\left(s,a_i\right)-Q(s,a_o)}{T}), a_i \neq a_o. \end{cases} \tag{27}$$

(2) Observe the reward value R, obtain the next state s', and update the learning parameters. Calculate the immediate reward value R obtained after the agent performs the action, and update the annealing temperature T and learning rate.

In the process of exploration and utilization by the agent, the annealing function determines the efficiency of the agent in exploring the optimal solution. The three classic annealing functions that affect temperature changes in the simulated annealing algorithm are algebraic, logarithmic, and exponential, and there is a slight gap in their annealing speeds. Among them, the logarithmic expression has a slower convergence speed, while the algebraic expression has a faster convergence speed. This paper uses the exponential annealing function.

$$T = \frac{\lambda}{1+k} T_0. \tag{28}$$

$$T = \frac{\lambda}{log\left(1+k\right)} T_0. \tag{29}$$

$$T = \lambda^k T_0. \tag{30}$$

Among them, T_0 is the initial temperature, T is the k-th annealing temperature, and λ is the temperature decay factor.

The learning rate represents the learning speed of the model and determines the difference between the current Q value and the newly obtained Q value. Its setting is crucial. If the setting is smaller, it means that the descent speed along the gradient is slower, which is helpful to ensure that no minimum value point is missed, but it also means that it takes longer for the model to reach a stable state; The higher the value is set, the more attention is paid to the newly obtained return value, and the gradient may fluctuate near the minimum value, even causing the model to fail to converge. Therefore, an adaptive learning strategy is obtained by improving the learning rate, which is a variable learning rate that gradually decreases as the number of learning iterations increases. α is large in the early stage of learning and small in the later stage of learning.

$$\alpha = \frac{1}{k \times i}. \tag{31}$$

Where k is a constant and i is the number of iterations. As the learning process continues, gradually reducing the learning rate can ensure that the state space is fully traversed in the early stage and accelerate the convergence of the algorithm in the later stage.

(3) Update Q value

According to the immediate reward obtained after the execution of the network selected action, the Q value corresponding to the combination of the current state and the selected action is updated. The update formula is as follows

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)). \tag{32}$$

3.3 Algorithm Flow

The flow chart of the Simulated Annealing and Q-learning-based Heterogeneous Wireless Network Access Decision Algorithm (SQAD) in this paper is shown in Fig. 4. The specific process is roughly as follows: First, perform parameter initialization operations, including setting the Q value table to zero, setting the initial temperature T, initial learning rate α, discount factor γ, etc. Secondly, obtain the current status s, including network status and service characteristics of the user terminal. The first learning adopts a random strategy to select actions, and the subsequent learning process selects actions according to the Metropolis criterion, that is: in state s, select the action a_r according to the random action strategy, and obtain the corresponding $Q(s, a_r)$; Select the optimal action a_o according to the function Q and obtain the corresponding $Q(s, a_o)$. After that, a random number is generated

$\delta(\delta(0,1))$ and the exploration probability $W = exp\left(\frac{Q(s,a_r)-Q(s,a_o)}{T}\right)$ is calculated, and the corresponding action is selected according to the size of δ and W at this time. If $\delta < W$, the exploration operation is performed and the agent selects action a_r. Otherwise, the optimal action is performed and the agent selects action a_o. After the action is executed, an immediate return r is obtained, and at the same time, the movement change of the user terminal and the network, the generation and end of the service call and the change of the network load are updated. Transfer to the next state s', update the annealing temperature T and learning rate, and then update the value function according to the optimal value of the next state $V^*(s) = \underset{a\in A}{max} Q^*(s,a)$ until the iteration to the final state.

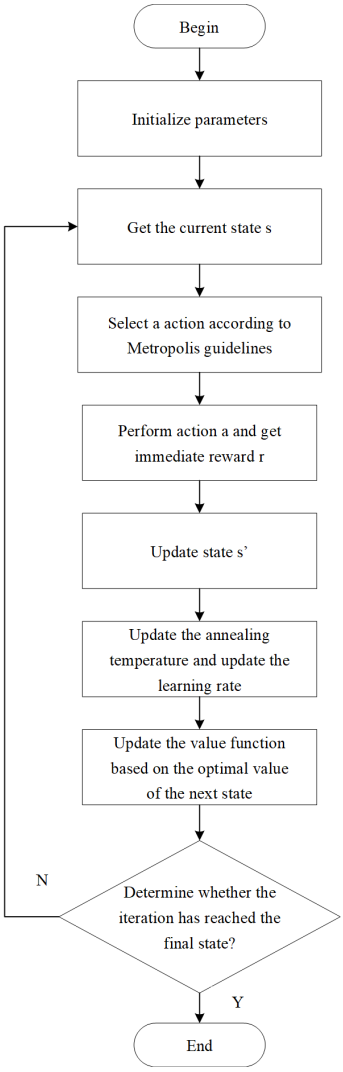

Fig. 4. The flow chart of the SQAD algorithm

4 Simulation and Analysis

This paper studies the network selection and switching of users in hotspot areas. Assume that the user terminal can run multiple services, the arrival of user calls obeys the Poisson distribution with parameter $\lambda = 0.05$, the service time obeys the exponential distribution with a mean value of 60, and the user call service type obeys the

uniform distribution. Considering that ground base stations are susceptible to interference from terrain, buildings and other wireless signals; Satellites have a wide coverage area and can be used to provide global communication services. They are basically not affected by geographical environment, natural disasters, etc., and can ensure the stability and reliability of communications, especially in response to emergencies. In addition, it has high encryption technology, but due to transmission distance, the delay is relatively large; Although UAVs have flexible mobility, their coverage is small and susceptible to factors such as mobility, flight environment, and signal interference. Therefore, the network attribute parameter settings in this paper are as shown in Table 3. The larger the attribute parameter value, the more obvious the corresponding attribute characteristics. The specific value can be adjusted according to the actual network environment. Other simulation parameters are shown in Table 4.

Table 3. Network attribute parameters

	LEO	UAV	5G micro base station (SBS)	LTE-A (MBS)
Security level	9	5	8	7
Delay (ms)	50	20	1	10
Packet loss rate level	1	7	2	4
Support mobility level	10	4	5	8

Table 4. Simulation parameter settings

Parameters	Value
Maximum number of users	100
Capacity ratio of each network	1:1:1:1
Maximum number of iterations	100
Initial temperature T_0	100K
Initial learning rate α_0	0.5
Discount factor γ	0.8
Temperature attenuation factor λ	0.98

(1) Assume that the number of multi-mode mobile users in the hotspot area is 100, and user calls are made every T=1s during the simulation process. The following is the convergence of the cumulative reward for 100 iterations of the Q-learning access selection algorithm (QL) based on the ε greedy strategy and the SQAD algorithm in this paper.

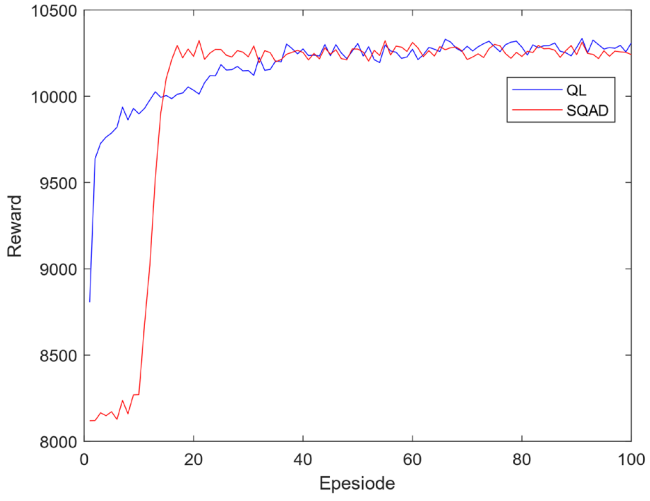

Fig. 5. Cumulative reward chart

It can be seen from Fig. 5 that the convergence trends of both algorithms gradually stabilize within a certain range as the number of iterative learning increases. The SQAD algorithm proposed in this paper improves the action selection mechanism and conducts full exploration in the early stage of iterative learning to avoid falling into the local optimal solution, so the initial reward value is small. After completing the exploration of the environment, it quickly converged to an optimal strategy based on the cooling strategy. The reward value of the SQAD algorithm reached stability around 17 iterations. However, the QL algorithm tends to stabilize after about 40 iterations, indicating that the SQAD algorithm has improved the convergence rate and reached the convergence value faster. That is, faster convergence is achieved by considering the overall long-term benefits. Therefore, the following focuses on comparing the proposed algorithm with other common access selection algorithms.

(2) Compare the SQAD algorithm mentioned in this paper with the AHP & GRA algorithm based on business preferences (AGP), Multi-Attribute Decision-Making Algorithm Considering Business Preference and Network Load (PLDM) and Random Access Decision Algorithm (RAD). The performance of the algorithm is evaluated from four aspects: the access failure rate, the cumulative utility, the proportion of each network access under different services and the average load occupancy rate of the network.

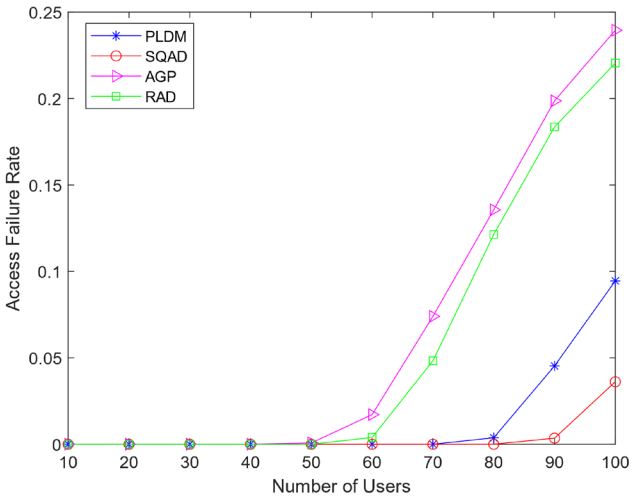

Fig. 6. Changes in access failure rate with different user numbers

Fig. 6 compares the changes in service access failure rate as the number of users changes among the four algorithms. Due to limited available network resources, the access failure rates of the four algorithms increase as the number of users increases. Due to the fact that the AGP algorithm only considers business characteristics and ignores network load conditions, and the same network may be the optimal choice for multiple services, it is easy for a large number of user terminals to simultaneously request the same network, resulting in access failure. Therefore, the AGP algorithm is the first to encounter access failure situations. The RAD algorithm has a randomized access method and does not need to consider the requirements of business or network load. Therefore, its access failure rate is slightly lower than that of the AGP algorithm. The PLDM algorithm takes into account the business characteristics and network load, so its access failure rate is lower than that of both AGP and RAD algorithms. The SQAD algorithm starts to experience access failure only when the number of users reaches 80. This is because it not only takes into account the requirements of business characteristics and network load, but also learns a better action selection strategy through Q-learning in a dynamic environment to avoid access conflicts. Therefore, the access failure rate is smaller than the other three algorithms.

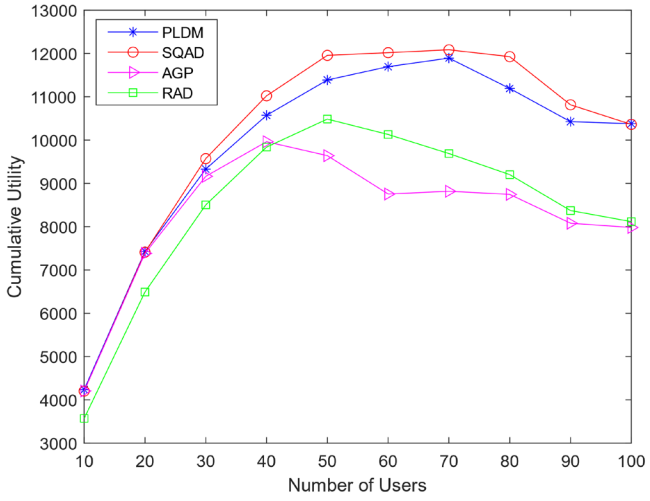

Fig. 7. Changes in cumulative utility with different number of users

Fig. 7 shows the changes in the cumulative utility of the four algorithms as the number of users increases. When the number of users is small, the available network resources are sufficient, and the cumulative utility value is mainly determined by business preference requirements. Therefore, the cumulative utility values of PLDM, SQAD and AGP algorithms considering the business characteristics are not very different, but they are all higher than the RAD algorithm. However, as the number of users increases, the available network resources change. Both PLDM and SQAD algorithms comprehensively consider user QoS requirements and network load, while the AGP algorithm only considers a single factor, resulting in the cumulative utility of the AGP algorithm gradually becoming the lowest. After the number of users reaches a certain scale (around 60), the cumulative utility of each algorithm shows a downward trend. The reason is that the larger the user scale, the more network resources are occupied, which leads to a smaller proportional factor of available resources in the reward function, further reducing the overall cumulative utility value. The SQAD algorithm can dynamically adjust the access strategy according to the current network load, while taking into account the business characteristics, and can adaptively access to a more satisfactory network. Therefore, its cumulative utility values are higher than other algorithms.

Considering that both PLDM and SQAD algorithms take into account business characteristics and network load conditions, and as service calls arrive and leave, each network load changes dynamically. Therefore, the performance of the two algorithms is analyzed by counting the average occupancy of each network load within the complete time period under different numbers of users.

Fig. 8 and Fig. 9 show the changing trend of the average load occupancy of each network with different number of users in the PLDM and SQAD algorithms respectively. When the number of users is 10 and 20, the network resources are relatively sufficient. Since LEO and SBS networks have higher weights for business characteristics and can provide better services, users under both algorithms tend to choose LEO and SBS networks. It can also be seen from Fig. 9 that starting from the number of users of 30, the reduction of available network resources affects the access decision of the SQAD algorithm. The average load occupancy rate of the MBS network with larger coverage increased significantly, and a small number of services began to access the UAV network. When the number of users reaches 50, the MBS network with large coverage capacity begins to accommodate more service requests than the SBS network, while the PLDM algorithm only has services access the UAV network after the number of users reaches 40. Moreover, on the whole, the SQAD algorithm has stronger load balancing among all networks. This is especially obvious for the access situation of MBS and SBS networks. Therefore, compared with the PLDM algorithm, the SQAD algorithm can better improve the load of the network and fully use the resources of each network by balancing the number of network access users. To evaluate whether the SQAD algorithm still meets the QoS requirements of different services while balancing network loads and reducing access failure rates. Under the maximum number of users, the system dynamics is the most complex, the network available resources are the most poor, the statistics of different business types access to each network to analyze the performance of the algorithm.

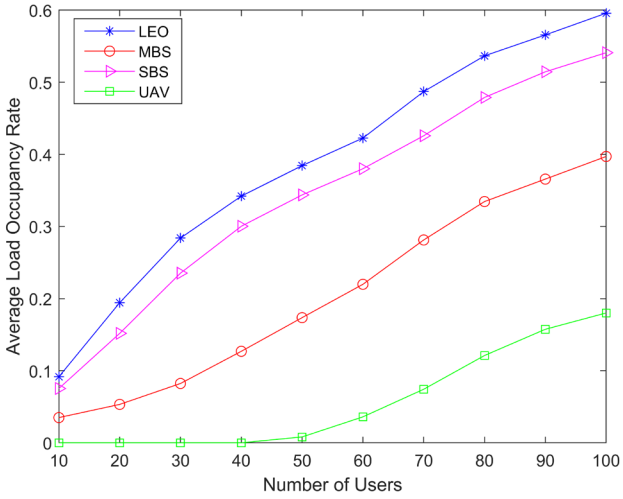

Fig. 8. Change trend of average load occupancy rate of each network in PLDM algorithm

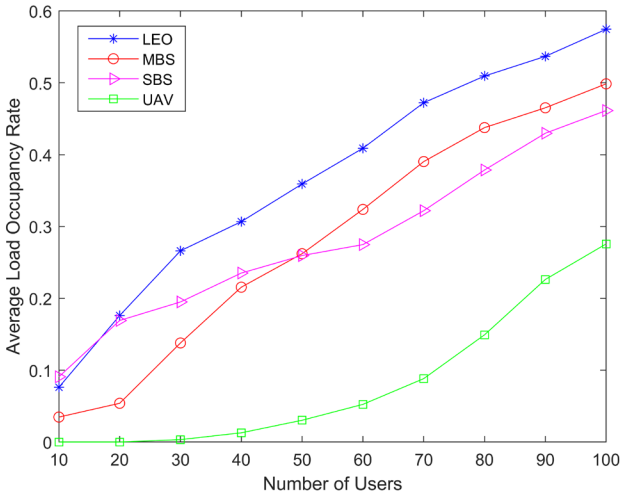

Fig. 9. Change trend of average load occupancy rate of each network in SQAD algorithm

Fig. 10 shows the preferences of different service types for each network calculated by combining the AHP and GRA algorithms in this article. For 4K/8K high-definition video services, it is more sensitive to reliability and supported mobility, and the LEO network has the lowest packet loss rate and the highest mobility support level, so the LEO network is preferred, which is in line with reality. Telemedicine services and office automation services have high requirements on latency, reliability and security, and have low requirements on supporting mobility. The security and reliability of the LEO network are the best among the four types of networks, followed by the SBS network. However, the delay of the LEO network is relatively large among the four types of networks, so the preferred networks for these two types of services are SBS, LEO, and MBS. LEO can support high mobility and SBS has the smallest latency, while the MBS network ranks second in both performance. The big data collection business is more concerned about latency, followed by its ability to support mobility, so the more preferred network ranking is SBS, MBS, LEO and UAV networks.

Fig. 11 shows the situation in which the four services in the SQAD algorithm access each network when the number of users is the largest. For 4K/8K high-definition video services, the LEO network access ratio is 0.68, which is much higher than the other three networks, thus meeting the network preference of 4K/8K high-defi-

nition video services. The network load access ratio rankings of telemedicine services are SBS, MBS, LEO and UAV networks, and the network load access ratio rankings of office automation services are SBS, LEO, MBS and UAV networks respectively. This is because under the maximum number of users, there are many business calls, and the network resources are limited, resulting in greater resource competition. Based on the arrival of business calls, more access is made to the SBS network, which is the most preferred by the business, while the rest of the network chooses MBS and LEO networks with a larger coverage range. At the same time, in order to improve the access success rate under limited network capacity, the access rate of the UAV network has reached about 0.15%. Overall, the proportion of SBS network access in telemedicine services is close to 40%, which is in line with business preference requirements; For the office automation business, SBS and LEO network access proportions are larger, which is in line with the network preference ranking of this business. The big data collection business prefers SBS and MBS networks, and the algorithm with a higher proportion of access to this business is the MBS network with strong coverage ability, followed by SBS network, which is in line with reality and meets the QoS requirements of the business.

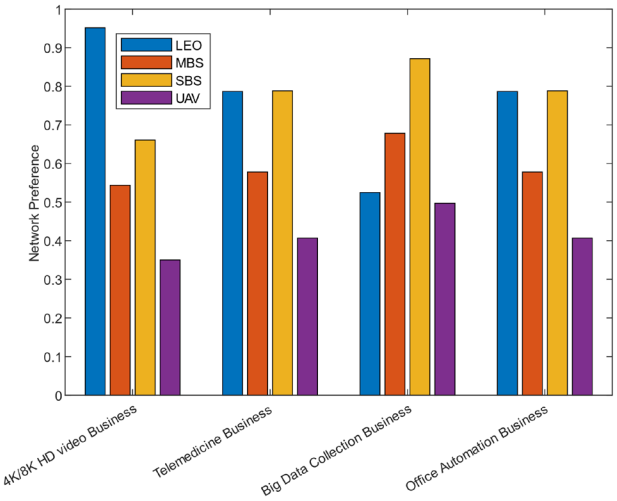

Fig. 10. Preference of different business types for each network

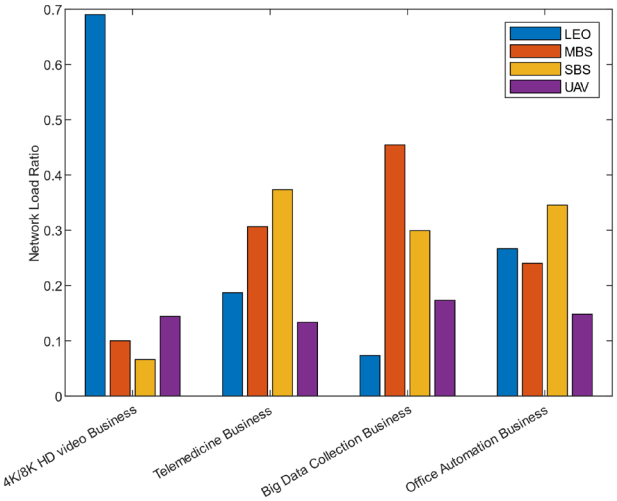

Fig. 11. The situation of four types of services accessing each network

In summary, the proposed algorithm considers the long-term benefits and can select the appropriate network for users with different business requirements. In addition, it effectively reduces the network access failure rate and improves the network resource utilization to a certain extent.

5 Conclusion

This article proposes a network access selection algorithm based on simulated annealing and Q-learning in SAGIN. Firstly, the user mobility model and heterogeneous network system model are established. Considering the network preferences of various network QoS parameters and different business types, the MDP model is constructed, and the return function is constructed by hierarchical analysis and gray correlation analysis. Secondly, SA is combined with the Q-learning algorithm, and the exploration-exploitation probability value is dynamically adjusted through the Metropolis criterion in the simulated annealing algorithm to optimize the action selection mechanism of the agent in the heterogeneous network, avoiding the algorithm from falling into the local optimal solution, improving the convergence speed and stability. Finally, the performance of the algorithm is evaluated from four aspects: access failure rate, cumulative utility, network access proportion under different services, and average network load occupancy. The simulation results show that the algorithm proposed in this paper can meet business preference requirements to a certain extent, reduce the network access failure rate, and improve network resource utilization. However, due to the complexity of the current SAGIN architecture, the differentiation of service guarantees for network users, and the dynamics of network resources, the development of the SAGIN is still restricted by many aspects.

6 Acknowledgement

This work was supported by Natural Science Foundation of Fujian Province (Grant No. 2022J011276, and No. 2023I0044), and High-level Talent Project of Xiamen University of Technology (Grant No. YKJ22030R).

References

[1] P. Qin, X. Wu, Z. Cai, X. Zhao, Y. Fu, M. Wang, S. Geng, Joint trajectory plan and resource allocation for UAV-enabled C-NOMA in air-ground integrated 6G heterogeneous network, IEEE Transactions on Network Science and Engineering 10(6)(2023) 3421-3434.
https://doi.org/10.1109/TNSE.2023.3261278

[2] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, X.S. Shen, Software defined space-air-ground integrated vehicular networks: Challenges and solutions, IEEE Communications Magazine 55(7)(2017) 101-109.
https://doi.org/10.1109/MCOM.2017.1601156

[3] P. Qin, S. Wang, Z. Lu, Y. Xie, X. Zhao, Deep reinforcement learning-based energy minimization task offloading and resource allocation for air ground integrated heterogeneous networks, IEEE Systems Journal 17(3)(2023) 4958-4968.
https://doi.org/10.1109/JSYST.2023.3266769

[4] A. Roy, P. Chaporkar, A. Karandikar, Optimal radio access technology selection algorithm for LTE-WiFi network, IEEE Transactions on Vehicular Technology 67(7)(2018) 6446-6460.
https://doi.org/10.1109/TVT.2018.2805190

[5] J. Liu, F. Gao, X. Luo, Survey of deep reinforcement learning based on value function and policy gradient, Chinese Journal of Computers 42(6)(2019) 1406-1438.
https://doi.org/10.11897/SP.J.1016.2019.01406

[6] J. Sun, Z. Qian, X. Wang, X. Wang, ES-DQN-based vertical handoff algorithm for heterogeneous wireless networks, IEEE Wireless Communications Letters 9(8)(2020) 1327-1330.
https://doi.org/10.1109/LWC.2020.2990713

[7] H. Xu, D. Li, M. Liu, G. Han, W. Huang, C. Xu, QoE-driven intelligent handover for user-centric mobile satellite networks, IEEE Transactions on Vehicular Technology 69(9)(2020) 10127-10139.
https://doi.org/10.1109/TVT.2020.3000908

[8] B. Yang, X. Wang, Z. Qian, A multi-armed bandit model-based vertical handoff algorithm for heterogeneous wireless networks, IEEE Communications Letters 22(10)(2018) 2116-2119.
https://doi.org/10.1109/LCOMM.2018.2861731

[9] L. He, D. Jiang, C. Wei, A QoE-based dynamic energy-efficient network selection algorithm, Wireless Networks 27(5) (2021) 3585-3595.
https://doi.org/10.1007/s11276-019-02231-z
[10] J. Zhong, L. Zhang, M. Alhabo, J. Serugunda, S.N. Mugala, A hybrid scheme using TOPSIS and Q-learning for handover decision making in UAV assisted heterogeneous network, IEEE Access 12(2024) 31422-31430.
https://doi.org/10.1109/ACCESS.2024.3368916
[11] A. Alizadeh, B. Lim, M. Vu, Multi-agent Q-learning for real-time load balancing user association and handover in mobile networks, IEEE Transactions on Wireless Communications 23(8)(2024) 9001-9015.
https://doi.org/10.1109/TWC.2024.3357702
[12] M. Ma, A. Zhu, S. Guo, Y. Yang, Intelligent network selection algorithm for multiservice users in 5G heterogeneous network system: Nash Q-learning method, IEEE Internet of Things Journal 8(15)(2021) 11877-11890.
https://doi.org/10.1109/JIOT.2021.3073027
[13] B. Zhao, G. Ren, X. Dong, H. Zhang, Distributed Q-learning based joint relay selection and access control scheme for IoT-oriented satellite terrestrial relay networks, IEEE Communications Letters 25(6)(2021) 1901-1905.
https://doi.org/10.1109/LCOMM.2021.3061717
[14] P.P. Kumar, K. Sagar, Reinforcement learning and neuro-fuzzy GNN-based vertical handover decision on internet of vehicles, Concurrency and Computation: Practice and Experience 35(12)(2023) e7688.
https://doi.org/10.1002/cpe.7688
[15] Y. Cao, S.-Y. Lien, Y. -C. Liang, Deep reinforcement learning for multi-user access control in non-terrestrial networks, IEEE Transactions on Communications 69(3)(2021) 1605-1619.
https://doi.org/10.1109/TCOMM.2020.3041347
[16] Z. Chkirbene, A.A. Abdellatif, A. Mohamed, A. Erbad, M. Guizani, Deep reinforcement learning for network selection over heterogeneous health systems, IEEE Transactions on Network Science and Engineering 9(1)(2022) 258-270.
https://doi.org/10.1109/TNSE.2021.3058037
[17] Y. Yang, Y. Wang, K. Liu, N. Zhang, S. Gu, Q. Zhang, Deep reinforcement learning based online network selection in CRNs with multiple primary networks, IEEE Transactions on Industrial Informatics 16(12)(2020) 7691-7699.
https://doi.org/10.1109/TII.2020.2971735
[18] K. Fan, B. Feng, X. Zhang, Q. Zhang, Network selection based on evolutionary game and deep reinforcement learning in space-air-ground integrated network, IEEE Transactions on Network Science and Engineering 9(3)(2022) 1802-1812.
https://doi.org/10.1109/TNSE.2022.3153480
[19] E. De Santis, A. Giuseppi, A. Pietrabissa, M. Capponi, F.D. Priscoli, Satellite integration into 5G: Deep reinforcement learning for network selection, Machine Intelligence Research 19(2)(2022) 127-137.
https://doi.org/10.1007/s11633-022-1326-3
[20] S.A. Javed, A. Mahmoudi, S. Liu, Grey absolute decision analysis (GADA) method for multiple criteria group decision-making under uncertainty, International Journal of Fuzzy Systems 22(4)(2020) 1073-1090.
https://doi.org/10.1007/s40815-020-00827-8
[21] H. Wang, K. Li, W. Pedrycz, A routing algorithm based on simulated annealing algorithm for maximising wireless sensor networks lifetime with a sink node, International Journal of Bio-Inspired Computation 15(4)(2020) 264-275.
https://doi.org/10.1504/ijbic.2020.108596