A Network Selection Method Based on Q-Learning in Power Wireless Communication System

Gang Shao¹², Wenqin Wu¹², Lei Yin¹² and Chenyang Ding³*
¹State Grid Chongqing Electric Power Company Nan'yan Power Supply Branch
Chongqing, 404100
²State Grid Corporation of China, Beijing, 100031
³NARI Group Corporation (State Grid Electric Power Research Institute) Nanjing, 211000
Email: nanjingxnd88@126.com
*Corresponding Author

Abstract. In power wireless communication network, the traditional network selection algorithm can’t deal with the problem of dynamic selection according to network condition, so this paper proposes a network selection algorithm based on Q-learning. Q-learning algorithm is a model-free reinforcement learning algorithm, which is often applied in the field of wireless communication due to its flexibility and adaptability, and obtains the optimal strategy by the Q-value function estimation of state-action pair. In this paper, we establish the Q-learning network selection model, determine the network status in the analysis of network load and service type, then select the maximum cumulative rewards to the action, with which we can select the optimal network. Since the Q value of the Q-learning algorithm is updated iteratively, the algorithm can adapt to the dynamic network selection. The simulation results show that the network selection algorithm based on Q-learning can effectively reduce the blocking rate of voice traffic and data packet loss rate, and improve the average network throughput.

1. Introduction
With the development of the power industry, the scale of the power grid has gradually expanded, and the network topology has become increasingly complex. The current power communication backbone network based on optical fiber communication has been unable to meet the needs of various services. For long-distance access nodes, direct laying of optical fiber lines is costly and lacks practical value; in the event of sudden disasters, it is difficult to repair fiber-optic lines in time; for new substations, the number of optical fiber lines is large and the construction period is long, which often affects the substation is put into production. In areas where wired communications are not playing a key role, wireless communication network technology and its networking systems can provide superior communication security [1-3].

Wireless communication technology plays an increasingly important role in the smart grid. At present, power wireless communication uses a variety of standards, including 230MHz radio network, WiMax, GPRS public network, CDMA public network, LTE230 and LTE1800 and other different networks [4]. Faced with the current large number of users and complex business types, any single-standard network can’t fully meet the needs of users, thus showing the coexistence of multiple heterogeneous networks and complementary advantages. How to choose a highly reliable and highly adaptable wireless communication option has become a hot research topic in the field of power communication.
Commonly used network selection methods are mostly Multiple Attribute Decision Making (MADM), such as Analytic Hierarchy Process (AHP), Fuzzy Logic [6] (FL, Fuzzy Logic), based on Optimum Sorting Method [7] (TOPSIS, Technique for Order Preference by Similarity to an Ideal Solution), Gray Relational Method [8] (GRA, Gray Rational Analysis) and so on. Based on these algorithms, the Ref.[9] combines AHP and TOPSIS, and utilizes AHP to calculate network decision attribute weights, and then uses TOPSIS to select the network closest to the ideal network as the best network. While [10] improved the TOPSIS in [9] to reduce the probability of sorting anomalies. [11] and [12] introduces the fuzzy function into the network selection process, and blurs the attributes that cannot be quantified to improve the switching performance. [13] uses the gray-scale correlation method to select the network that best matches the reference network, which can effectively solve the situation that the decision attributes are not monotonous.

Although the multi-attribute decision-making method can comprehensively consider the network parameters, it needs to determine the network weight in advance according to experience or expert judgment, which will result in poor adaptability of the network selection algorithm to the network environment. The appropriate access network can’t be dynamically selected in the environment. Reinforcement Learning is a non-mentor online learning technology that learns from environmental state to action mapping, enabling users to adopt optimal strategies based on maximum reward values, often used in the field of automatic path finding and wireless communication. The Q-learning algorithm is a more commonly used method of reinforcement learning. The Q-value function is learned by evaluating the action-state pairs to find the optimal action strategy. [14-16] uses the Q-learning algorithm for heterogeneous network selection, which can adapt to the dynamic changes of the network environment and effectively improve the performance of the selection algorithm.

In this paper, the Q-learning algorithm is combined with the power wireless communication network selection problem. The Q value function is constructed step by step according to the current network load status and service type, and the Q-learning algorithm is used to find the optimal strategy and select the best access network. The advantage of this algorithm is that it can adapt to the dynamic network selection problem. In addition, the optimal strategy is set as a function of the network load state. Through the online learning ability of the Q-learning algorithm, the algorithm can effectively improve the network throughput, reduce the traffic blocking rate and improve the network selection performance.

2. System Model

With the development of the energy Internet, in order to achieve full coverage of power service terminals, it is urgent to construct a power wireless communication network combining wide-narrowband convergence, wide coverage and deep coverage. Therefore, the wireless communication network in this paper chooses a hybrid distribution model of 230MHz wireless private network and 1.8GHz wireless private network to achieve complementary network coverage areas, which is shown in Figure 1. Consider two common services in power communication systems: voice services and data services. Users within the coverage are connected to the optimal network according to their business needs according to the network selection algorithm.

![Figure 1. Network coverage model](image)
In the network scenario of this paper, the coverage of the 230MHz wireless private network is larger than the 1.8GHz wireless private network. In a network with only one network coverage area directly connected to the network, this paper only considers the network selection problem in the overlapping areas of the two networks. In the actual environment, the user's service arrives randomly, its duration and resource demand are a random variable, and the network status is also variable. If only the various attributes of the two networks are weighted, the multi-attribute decision making method is used to select the network with the best utility function or decision function. Since the weight of the method needs to be judged by experience or assessed by experts, it often cannot reflect the Dynamic changes of network status.

This paper introduces the Q-learning algorithm into the network selection, so that it has certain learning ability, and can adapt to the dynamically changing network, and find the optimal network selection strategy for each session with considering the network load and service type.

3. System Model

3.1 Network Selection Algorithm Based on Q-learning

Reinforcement Learning [17] is a collection of questions about the evolution of Agents through simple scalar signals from the external environment while analyzing the consequences of their actions. It has the characteristics of learning initiative and self-adaptation, and is widely used in the field of automatic control and wireless communication. In general, the basic enhanced learning model consists of the set of states \( S = \{ s_1, s_2, \ldots, s_M \} \) that reflect the current environment, the action set \( A = \{ a_1, a_2, \ldots, a_N \} \) that Agent can choose, return function \( r \) and search strategy \( \pi : S \rightarrow A \). Fig.2 shows the basic principle. Under the clear goal, the Agent selects the optimal action by perceiving the environment information and the search strategy, and causing the change of the environmental state and getting the immediate return, then updating the evaluation function, and finally performing a new round of learning after the completion of the learning until the satisfied condition Terminate learning.

The Q-learning algorithm [18] is a model-free enhanced learning algorithm proposed by Watkin in 1989. Assume that in step \( n \), the Agent records the status of the current environment \( s_n \in S \), and select an action \( a_n \in A \) based on the state at this time, then an instant reward function is generated after the action is completed, which is determined by the current state and the selection action.

Under the search strategy \( \pi \), the value of the state \( s \) is defined as

\[
V^\pi(s) = r_n(\pi(s)) + \gamma \sum_y P_{s_n,y_n}^n [\pi(y)] V^\pi(y)
\]

where \( \pi(s) \) is the action selected according to the policy under state \( s \). \( \gamma \in [0,1] \) discount factor reflecting the relative proportion of delayed rewards. \( P_{s_n,y_n}^n [a_n] = \text{Prob}[y_n = y | s_n, a_n] \) is the probability that the environment state will be transferred to \( y_n \) under the environmental state \( s_n \) and the selection action \( a_n \) in the \( n \)-th step operation. From this formula, the actual meaning of the value of the state \( s \) is
that the Agent immediately gets the reward \( r \) after performing the action selected via strategy \( \pi \), and transfers a "valid" state with a state value of \( V^\pi(y) \) with a probability \( P_{\pi(y)} \).

According to the DP (Dynamic programming) theorem [18], there is at least one optimal strategy \( \pi^* \) that takes the value of the state \( s \) to the maximum value, which is recorded as

\[
V^*(s) = V^\pi^*(s) = \max_a \left\{ r_s(a) + \gamma \sum_y P_{\pi(y)}(a)V^\pi(y) \right\}
\]

For a strategy \( \pi \), define a Q-valued function (also known as an action-valued function) as

\[
Q^\pi(s, a) = r_s(a) + \gamma \sum_y P_{\pi(y)}(\pi(s))V^\pi(y)
\]

in other words, the Q value is a predictive estimate of the return value after taking an action.

It can be seen from equations (2) and (3) that

\[
V^*(s) = \max_a Q(s, a)
\]

Assuming that the action that maximizes the Q value at this time is \( a^* \), then the best strategy can be expressed as

\[
\pi^*(s) = a^*
\]

Therefore, if the Agent can learn the Q-value function, it is easy to choose the best strategy, that is, perform the action that maximizes the Q value.

In the actual process, the learning of the Q value is done by iteration. Each iteration process updates a Q value \( Q(s, a) \). In order to update all Q values, the Agent needs to constantly interact with the environment. When the Q value does not change much after multiple iterations, it can be considered that the Q value converges and the Q-learning process ends. In each iteration, the Q value can be updated according to the method of equation (6).

\[
Q_n(s, a) = (1 - \alpha_n)Q_{n-1}(s, a) + \alpha_n(r_n + \gamma \max_{a'} Q_n(s', a'))
\]

where \( \alpha \in [0, 1) \) is the learning rate; \( s' \) and \( a' \) are the states obtained after the selection action is taken, that is, the selectable corresponding actions.

According to [18], when the reward function \( r \) is bounded, the learning rate \( 0 \leq \alpha_n < 1 \), and satisfies the equation (7), then

\[
\sum_{i=1}^\infty \alpha_i^n(x, a) = \infty \sum_{i=1}^\infty \left[ \alpha_i^n(x, a) \right]^2 < \infty, \forall x, a
\]

The Q-valued function \( Q(s, a) \) always converges with probability 1, that is

\[
Q(s, a) \rightarrow Q^*(s, a), n \rightarrow \infty, \forall x, a.
\]

3.2 Q-learning Element Design
In this section, the Q-learning algorithm introduced above is applied to the network selection problem of the scene of Figure 1, so that it can adapt to the dynamically changing network selection problem.

As can be seen from the above, Q-learning is mainly composed of a state set \( s \), an action set \( A \) that the agent can select, a reward function \( r \), and a search strategy \( \pi \). To apply Q-learning to the network selection problem of power wireless communication systems, the various elements of Q-learning should first be mapped into the network model.
1) State space $s$

The network status is mainly related to the network load status and the type of arrival service. According to the load size of the 230MHz wireless private network and the 1.8GHz wireless private network, the load status of each network is divided into three states: idle, busy and blocked, then considering the data service and the voice service, the state vector can be expressed as

$$ s = [BW_1, BW_2, k] $$

where $BW_1$ and $BW_2$ are the load states of the two networks, and the value is 0, 1 or 2, which means that the network is idle, busy, and blocked. $K$ is the reaching service type. As this paper only considers both voice and data services, the value of $k$ is 0 or 1. Therefore, the network has $3 \times 3 \times 2 = 18$ different states.

2) Action set $A$

In the network model shown in Figure 1, two networks are considered and the user needs to select a most suitable network access according to the network selection algorithm, so the user can take an action set to access the network, that is

$$ A = \{1, 2\} $$

where 1 indicates that the user accesses the 230MHz wireless private network; 2 indicates that the user accesses the 1.8GHz wireless private network. It can be obtained that the Q-valued function $Q(s, a)$ is a list of values of 18 rows and 2 columns, which is continuously updated during the iterative process.

3) Return function $r$

Immediate action is achieved when the action is performed in each state, and the immediate return is related to the network load status. If the user can successfully access the network and request the service to be executed, an immediate return is obtained

$$ r(a_1) = r(a_2) = 1 $$

If the network is busy at this time and cannot complete the corresponding service, the return is

$$ r(a_1) = r(a_2) = 0 $$

3.3 The Procedure of the Proposed Algorithm

Considering the network load status and the arrival service attribute, based on the Q learning method, the network selection method of the 230 MHz wireless network and the 1.8 GHz wireless network can be obtained. Fig. 3 is a flow chart based on the Q learning network selection algorithm, and the specific steps are as follows:

1) Initialize the Q value table and set the discount factor and learning rate;

2) Determine the type $k$ of the service arriving at a certain moment and the load rate $BW_i$ of the current two networks, and get the current status $s_i$.

3) Select the available action in action set $A$, and record the action and the next state $s$.

4) Calculate the immediate return function $r$, based on the network status after the selection action is performed;

5) According to the formula (6), update the Q value function $Q_n(s, a)$, and the learning rate $\alpha$ is gradually reduced to 0 according to the inverse proportional function rule;

6) Repeat steps (2)-(5) until the Q value converges, that is, the Q value difference before and after the update is less than the threshold value;

7) Return to step (3) to select the action and access the best network.
3.4 The Procedure of the Proposed Algorithm

The advantages of the network-based network selection algorithm based on Q-learning are summarized as follows:

1. iteratively updating the Q-value function, thus adapting to the dynamically changing network selection problem;
2. The optimal strategy considers the network load status, which can effectively reduce the voice blocking rate and data packet loss rate, and improve the network average throughput.

4. Simulation and Analysis

This paper considers the heterogeneous scenarios of the 230MHz wireless network and the 1.8GHz wireless network in the power wireless communication system. The network parameters are shown in Table 1. In the Q-learning algorithm, the discount factor is $\gamma = 0.8$; the learning factor is $\alpha = 0.5$. Assume that the voice service arrival interval obeys the Poisson distribution of 60s, and the duration obeys the 80s exponential distribution. If the network does not have enough resources during the service duration, the voice service is blocked. The data service arrival rate interval is the 20s exponential distribution. The duration is subject to the exponential distribution of 80s. If the network cannot provide the resources required for the data, the data service is discarded. The bandwidth required for voice services and data services is 12.2 kbps and 32 kbps, respectively.

| Bandwidth/MHz | Delay/ms | BER /10-6 | Shake /ms |
|---------------|----------|------------|-----------|
| 230MHz        | 1        | 200        | 40        | 10        |
| 1.8GHz        | 5        | 50         | 25        | 6         |

In the simulation, the Q-learning network selection algorithm and the random network selection method are used to compare the performance of the two network selection methods. Fig.4 shows the trend of the voice blocking rate as the traffic arrival rate increases. It can be seen that as the service arrival rate increases, the network becomes more and more busy, and the blocking rate of the voice service gradually increases. Because the Q-learning algorithm considers the network load condition, the action of maximizing the Q value function is selected in each iteration, which can better adapt to the dynamic change of the network, and thus can obtain a lower voice traffic blocking rate. Fig.5 shows the trend of data packet loss rate as the two services increase. Similar to the voice service, as the service arrival rate increases, the packet loss rate increases gradually, but the Q-learning algorithm can effectively reduce the packet loss rate of the data service and improve the network selection performance.
Fig. 4. Speech blocking rate under different business arrival rate

Fig. 5. Data packet loss rate under different business arrival rate

Fig. 6 shows the trend of the average network throughput as the business increases. Because the Q-learning algorithm can dynamically adapt to network changes and reduce the voice blocking rate and data packet loss rate, it can effectively increase the average network throughput and improve network resource utilization.

Fig. 6. Average throughput under different business arrival rate

5. Conclusion
This paper proposes a Q-learning-based network selection algorithm in power wireless communication networks, which can solve the problem that traditional network selection algorithms can’t consider
network conditions for online dynamic access control, and adapt to dynamically changing network access. By establishing a Q-learning network access model, the algorithm determines the network state after analyzing the network load and service type, and selects the action that can achieve the maximum cumulative return, so that the best access network can be selected. The simulation results show that the Q-learning based network selection algorithm can effectively reduce the voice service blocking rate and data service packet loss rate and improve the network average throughput.

6. Acknowledgement
This work is supported by Key project of state grid Chongqing electric power company(Research and demonstration on the application of new technology of wide and narrow band fusion Internet of things for power multi-service)

7. References
[1] Yu G, Liu J S, Cai S L. Research on LTE Wireless Network Planning in Electric Power System [J]. Electric Power Information and Communication Technology, 2016(10):7-11.
[2] Luo K B, Tan Y G, Zhou B. The analysis and research of wireless communication technology in power communication network [J]. Electronic Test, 2016(24):93-94.
[3] Xu K H. Application of wireless communication technology in power system [J]. Communication world, 2013(21):122-123.
[4] Zhai Z. Hotspots of wireless communication technique and application prospect in electric power communication network [J]. Telecommunication for Electric Power System, 2006, 27(5):1-2.
[5] Wang L, Kuo G S G S. Mathematical Modeling for Network Selection in Heterogeneous Wireless Networks — A Tutorial[J]. IEEE Communications Surveys & Tutorials, 2013, 15(1):271-292.
[6] Preethi J, Palaniswami S. A new hybrid approach for heterogeneous radio access technology selection[J]. Journal of Computer Science, 2012, 8(5).
[7] Sgora A, Vergados D D, Chatzimisios P. An access network selection algorithm for heterogeneous wireless environments[C]// The IEEE Symposium on Computers and Communications. IEEE Computer Society, 2010:890-892.
[8] Savitha K, Chandrasekar C. Grey Relation Analysis for Vertical Handover Decision Schemes in Heterogeneous Wireless Networks[J], 2011.
[9] Sun Q, Ge X B, Liu L, st. Multi-Attribute Network Process Comprehensive Evaluation Method for Smart Grid and Its Application [J]. Power System Technology, 2012, 36(10):49-54.
[10] Zhuo L, Nie J, Xiao J W, st. A network selection algorithm in heterogeneous multi-networks coexistence of electric power wireless communications [J]. Computer Engineering and Science, 2016, 38(1):78-83.
[11] Gu H, Xu M H, Yan H J. A Fuzzy AHP Networks Selection Algorithm in the Heterogeneous Networks [J]. Journal of Nanjing University of Posts and Telecommunications(Natural Science) , 2012, 32(6):50-57.
[12] Zhao H Q, Tang X W. Evaluation model of network service performance based on fuzzy analytic hierarchy process [J]. Journal of Computer Applications, 2013, 33(11):3035-3038.
[13] Bari F, Leung V C M. Use of Non-monotonic Utility in Multi-Attribute Network Selection[M]// Wireless Technology. Springer US, 2009:1-8.
[14] Kudo T, Ohsuki T. Q-Learning Based Cell Selection for UE Outage Reduction in Heterogeneous Networks[C]. IEEE Vehicular Technology Conference, 2014:1-5.
[15] Zhao Y Q, Zhou W F, Zhu Q. Q-Learning Based Heterogeneous Network Selection Algorithm[M]// Recent Advances in Computer Science and Information Engineering. Springer Berlin Heidelberg, 2012:1461-1464.
[16] Xu Y, Chen J, Ma L, et al. Q-Learning Based Network Selection for WCDMA/WLAN Heterogeneous Wireless Networks[C]// Vehicular Technology Conference, IEEE, 2015:1-5.
[17] Kaelbling L P, Littman M L. Moore Reinforcement Learning: A Survey[J]. Journal of Artificial Intelligence Research, 1996, 4(1):1 - 6.
[18] Christopher J.C.H. Watkins. Technical Note: Q-learning. Machine Learning,8,279-292(1992).