Intelligent Countermeasure Design of Radar Working-modes Unknown

Xing Qiang, Zhu Wei-gang, Jia Xin
Equipment Academy
Beijing, China
e-mail:xingyanqiang1989@126.com, yi_yun_hou@sina.com, jiaxinpro@sina.com

***Abstract*—With the progress of radar technology, the development of radar tends to multifunction and intellectualization. The anti-jamming capability of radar is enhanced, and the combat effectiveness of radar countermeasure method for conventional radar is decreasing. So it is urgent to study the intelligent radar countermeasure (IRC) method, especially for number of working modes that is unknown. Based on this, this paper expounds the intelligent radar countermeasure method, and compares the difference between intelligent radar countermeasure and traditional radar countermeasure (TRC). The basic principle of reinforcement learning (RL) is introduced. For the situation that number of working modes is unknown, intelligent radar countermeasure based on reinforcement learning is proposed, and the algorithm steps are given. The simulation results show that the intelligent radar jamming method can independently learn and make decision intelligently according to the jamming effect, which improves the adaptability of the radar countermeasure system, and can resist the multi-working mode radar simultaneously.**

***Keywords*—intellectualization; radar countermeasure; reinforcement learning; Q-learning.**

I. INTRODUCTION

With the development of radar technology, radar has developed from fixed working mode to multi-working mode, and the working parameters can be changed in real time according to its working environment and task demand[1]. The main problems of the traditional radar jamming measures are: the radar intelligence reconnaissance, acquisition and analysis is becoming more and more difficult because of the complex electromagnetic environment; with the radar tending to intellectualization, the jamming mode needs to be improved urgently; for unknown threat radar target, real-time performance of traditional jamming mode is poor[2]. In this context, cognitive electronic warfare (CEW) emerged and developed rapidly[3], which is the combination of cognition and EW technology. The key technologies include cognitive reconnaissance technology, intelligent jamming method synthesis technology and real-time jamming effect evaluation technology[4]. Which, cognitive reconnaissance technology and real-time jamming effect evaluation technique provide the prior knowledge and the basis of the updating of jamming measures for intelligent jamming synthetic technology. It can be seen that the intelligent jamming synthetic technology is the core of the key technology in the CEW.

As early as in the 1980s, artificial intelligence (AI) technology is proposed to apply to electronic warfare to improve the agility and intelligence [5-7]. And then it may not be reported in the public literature for secrecy reasons. Until 2010, DARPA successively released BLADE, CommEx, ARC and other project announcements[8 - 11], application of AI in the EW develops rapidly. Among them, reinforcement learning, called the hope of real AI, is one of the most active research fields in AI[12-13]. It focus on how the Agent changes through the environment state, deciding what action to take to obtain the maximum expected return. Compared with the process of CEW, the jamming system determines the jamming countermeasure to obtain the best results based on the change of the target state, and the working process of the two is almost identical, so this paper proposes that the reinforcement learning is applied to the IRC process design. This will enable the radar jamming system to synthesize jamming measures intelligently in a short time, and improve the real-time performance and intelligent. The threat target of multi-working mode can be resisted, and the problem of single function of jammer in traditional radar countermeasure is solved with better adaptability.

II. OVERVIEW OF INTELLIGENT RADAR COUNTERMEASURE

Intellectualization refers to the application of modern communication and information technology, computer network technology, industry technology, intelligent control technology to a certain aspect[14]. The process from perception to memory, judgment and imagination called "wisdom", the result of which needs to be expressed through behavior or language, which is called "ability", and this whole process is called intellectualization. Characteristics of intellectualization are: (1) can perceive the external world and obtain external information; (2) the ability to store the perceived information and think using knowledge; (3) ability to learn, interact with external environments and adapt to external environmental changes; (4) can react independently to external stimuli.

This paper studies precisely this kind of intelligent radar countermeasure for radar working-modes unknown. The

978-1-5386-3142-3/17/$31.00 ©2017 IEEE

jamming receiver recognizes the state and behavior of the radar target and synthesize the jamming style independently according to the target state. The jamming system continue to detect the radar target state, evaluating the jamming effect, and feedback the jamming effect to the jamming decision module.

The United States first began the IRC research, and launched the" ARC" project, aiming to quickly perceive new, unknown radar threats, automatic synthesis radar countermeasures, and accurately assessing its countermeasure effect. November 4, 2016, BAE Systems announced a contract that has been awarded by DARPA to continue to develop the "ARC" system for DARPA to help airborne EW systems resist new unknown adaptive radars[15]. Architecture of ARC is shown in Fig.1.

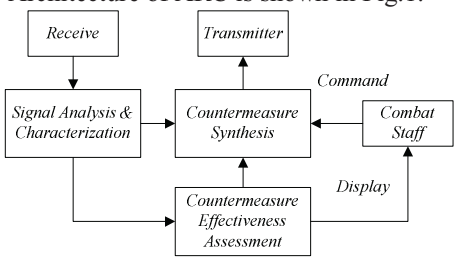

Fig.1 Architecture of IRC

Difference between intelligent radar countermeasure and traditional radar countermeasure is shown in Tab.1.

Table 1 Difference of IRC and TRC

TRC	IRC
Single function	Simultaneously against multi-working mode radars
Limited to single-node or dedicated platforms	Distributed, networked, or single platform
Poor feedback	Closed-loop feedback
Poor environment perception	Real-time surroundings perception
Implemented by prior programming	Intelligent decision
Frequency band limitations	Adjust the working band according to the surroundings

It can see from Tab. 1 that the IRC is distinct from the TRC, mainly including the following aspects:

(1) IRC with real-time and accurate situational awareness, as well as the memory and learning ability to target;

(2) Jamming measures can be synthesized and optimized by analyzing the surroundings;

(3) The jamming effect can be evaluated based on the detection and analysis of external environment;

(4) A closed-loop feedback is formed, and the jamming style and parameters are automatically adjusted according to the feedback result.

III. INTELLIGENT RADAR COUNTERMEASURE METHOD RESEARCH

A. Principle of Reinforcement Learning

Reinforcement learning[12] is based on the principles of animal physiology and psychology, using human and animal "trial and error" mechanism, and learning from the interaction with the environment. Learning process only needs to obtain the evaluation of feedback signal, and take the max reward as the learning goal. The characteristic is the initiative temptation to the environment; the environment generates feedback to the temptation action which is evaluated, the Agent adjusts the future behavior according to the feedback. Therefore, the advantages of RL are also generated: self-learning, on-line learning and updating. The interaction process is shown in Fig. 2.

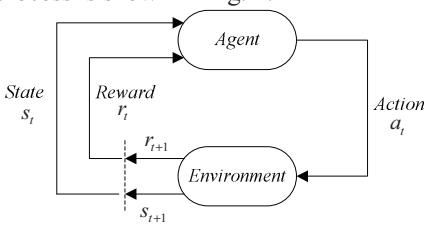

Fig. 2 Principle of RL

The interaction process is (1) The Agent perceives the current state of the environment; (2) According to the current state and reward value, the Agent selects an action, and performs the action; (3) The action chosen by the Agent acts on the environment, causes the environment to move to the new state, and gives the new reward; (4) The Agent calculates the return value according to the feedback reward value, and takes the return value as the basis of updating strategy. $s_t \in S$ represents the state of the Agent at a time *t*, and S is a collection of all States. $a_t \in A(s_t)$ represents the action that the Agent performs at time *t*, and $A(s_t)$ represents all possible sets of actions under State s_t. When in state s_t, the Agent chooses and executes the action a_t, receiving the reward $r_{t+1} \in R$, and transfers to the new state s_{t+1}.

In the above process, the Agent did not tell which action to take, but by itself according to the environment feedback information. The principle is to maximize the probability that the Agent acquire the positive reinforcement signal from the environment in the process of subsequent learning.

The purpose of RL is to form a strategy that enables agent to take action with optimal performance. Therefore, it is need to define a target function to determine what is the optimal action in the long run. Policy defines the behavior of the agent and is the mapping from State to action: $\pi : S \to A$. The policy defines the actions $a_t = \pi(s_t)$ that can be performed in any state: s_t. The value $V^{\pi}(s_t)$ of the strategy π is the desired cumulative reward of the agent who follows the policy from state s_t.

In a finite stage or fragment model, the agent attempts to maximize the expected rewards for the next *N* steps:

$$V^{\pi}(s_t) = E[r_{t+1} + r_{t+2} + \cdots + r_{t+N}] = E[\sum_{i=1}^{N} r_{t+i}] \quad (1)$$

In infinite order model, there is no sequence length limit, but the future rewards will be discounted:

$$V^{\pi}(s_t) = E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots] = E[\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}] \quad (2)$$

Where, $\gamma \in [0,1]$ represents a discount rate to ensure that the reward returned is limited. For each strategy, there is a value $V^{\pi}(s_t)$. If the best strategy π^*, there is

$$V^*(s_T) = \max_{\pi} V^{\pi}(s_t), \ \forall s_t \quad (3)$$

In some applications, the state-action value $Q(s_t, a_t)$ is more meaningful than the strategy value $V(s_t)$, $Q(s_t, a_t)$ represents the value obtained by action a_t under state s_t. The $Q^*(s_t, a_t)$ indicates the desired cumulative reward obtained by the optimal strategy under state s_t by taking action a_t.

$$\begin{aligned} V^*(s_t) &= \max_{a_t} Q^*(s_t, a_t) \\ &= \max_{a_t} E[\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}] \\ &= \max_{a_t} E[r_{t+1} + \gamma \sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i+1}] \\ &= \max_{a_t} E[r_{t+1} + \gamma V^*(s_{t+1})] \end{aligned} \quad (4)$$

$$V^*(s_t) = \max_{a_t}(E[r_{t+1}] + \gamma \sum_{s_{t+1}} p(s_{t+1} \mid s_t, a_{t+1})) \quad (5)$$

The probability of moving to each possible next state s_{t+1} is $p(s_{t+1} \mid s_t, a_{t+1})$, and following the optimal strategy. The cumulative expected reward is $V(s_{t+1})$, for the state-action value:

$$Q^*(s_t, a_t) = E[r_{t+1}] + \gamma \sum_{s_{t+1}} p(s_{t+1} \mid s_t, a_t) \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}) \quad (6)$$

If the state-action value $Q^*(s_t, a_t)$ is obtained, it is possible to define the execution action a_t of the policy π, which has the maximum value $\pi^*(s_t)$ in all $Q^*(s_t, a_t)$: selecting a_t^*, where $Q^*(s_t, a_t^*) = \max_{a_t} Q^*(s_t, a_t)$. Then the action sequence of the optimal strategy is obtained.

Commonly used RL algorithms include dynamic programming, Monte Carlo method, temporal-difference learning, Q-Learning and so on. Dynamic programming based on a complete model for strategy optimization, and the computation increases exponentially with the state, so there is a "dimensional catastrophe" problem[16]; Monte Carlo method is a model-independent RL method. The requirement for Markov properties is not strict, but this method can only update the state value function at the end of each learning task, therefore the algorithm is slow; temporal-difference learning is combined with the merits of dynamic programming and Monte Carlo, the value function is updated step by step, and the model is not required[17]; Based on the temporal-difference learning, according to the different behavior decision modes in interacting with the environment, Watkins and Rummery respectively propose Q-Learning and Sarsa algorithm. The difference is that the iteration of behavioral decision and value function in Q Learning is independent for each other. It is an off-line algorithm that uses the maximum value function to iterate, the update of R value depends on all kinds of assumption decisions; the iteration of behavior decision is consistent with the value function in Sarsa learning, and it is an on-line R learning using actual Q value to iterate. According to the above characteristics, This paper chooses Q for IRC research.

Basic form of Q-learning is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, a') - Q(s_t, a_t)] \quad (7)$$

Where, $Q(s_t, a_t)$ represents sum of discount rewards using action a_t under state s_t. $\alpha \in (0,1)$ is the learning rate, $\gamma \in [0,1]$ is the discount rate.

B. Intelligent Radar Countermeasure Based on Q-learning

Principle of intelligent radar countermeasure is shown in Fig. 3.

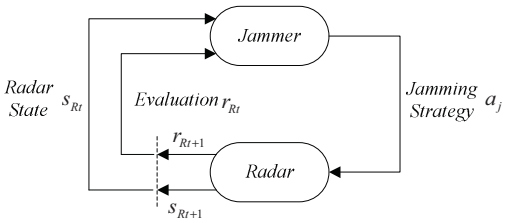

Fig. 3 Principle of IRC

$s_{Rt} \in S$ indicates the working mode of the multifunction radar in the moment *t*, and *S* is a set of all radar working states. $a_j \in A(s_{Rt})$ indicates the jamming style used by the jammer at the moment *t*. $A(s_{Rt})$ means the set of jamming styles that the radar is under working mode s_{Rt}. When the radar is in working state s_{Rt}, the jammer selects and executes the jamming mode a_j, receives the jamming performance evaluation $r_{Rt+1} \in R$, and transfers to the new working status s_{Rt+1}.

Updating formula of intelligent radar countermeasure is:

$$Q(s_{Rt}, a_j) \leftarrow Q(s_{Rt}, a_j) + \alpha[r_{Rt+1} + \gamma \max Q(s_{Rt+1}, a') - Q(s_{Rt}, a_j)] \quad (8)$$

Where, $Q(s_{Rt}, a_j)$ is radar state-jamming style value, r_{Rt+1} represents evaluation of jamming effect.

Each learning process of a jammer can be seen as starting with a random state of the radar, using a strategy, such as a ε greedy strategy or Boltzamann distribution strategy, to select the jamming style. After performing the selected jamming style, the jammer observes the new state and jamming rewards of the radar, then updates the last working state-jamming style Q value based on the maximum Q value and the jamming reward of the new radar state. The jammer will continue to select the jamming style according to the new radar state until it reaches a termination state.

Radar has a variety of operating modes, and jamming style set A contains a variety of interference styles. Define an immediate return value:

$$r = \begin{cases} 100 & TR \to \min \\ 0 & TR \leftrightarrow TR \quad (TR \not\to \min) \\ -1 & TR \not\leftrightarrow TR \end{cases} \tag{9}$$

Wherein: $TR \to \min$ represents the conversion to the current minimum threat level operating mode of the radar; $TR \leftrightarrow TR\,(TR \not\to \min)$ means switching between working modes except the lowest threat level; $TR \not\leftrightarrow TR$ indicates no conversion between working modes.

For the choice of jamming style, the jammer needs to solve the contradiction between exploratory and utilization. This paper uses the ε greedy strategy to select the action.

The jamming style selection index of intelligent jamming is: finding a strategy (jamming style sequence) that can maximize the future work mode-jamming style value.

Here is a description of the algorithm:

Step1. Initialize $Q(s_{Rt}, a_j)$ to the first order zero matrix, give the parameter α 、 γ initial value;
Step2. Loop
(1) Reconnaissance of current environment and identification of radar working mode;
(2) Select the jamming style a_j and execute according to ε greedy strategy;
(3) Continue reconnaissance radar working mode, Judge new working mode s_{Rt+1}, calculate efficiency evaluation value r_{Rt};
(4) update $Q(s_{Rt}, a_j)$ matrix according to $Q(s_{Rt}, a_j) \leftarrow r_{Rt+1} + \gamma \max_{a'} Q(s_{Rt+1}, a')$, then add corresponding ranks if the new mode;
(5) $s_{Rt} \leftarrow s_{Rt+1}$;
Until the number of working modes stabilizes, the cycle stops.
Step3. Outputs the final jamming style sequence and the target State (the minimum threat level of working mode)

Based on the above algorithm, the intelligent countermeasure process is shown in Fig. 4 in the condition that the number of radar working modes is unknown.

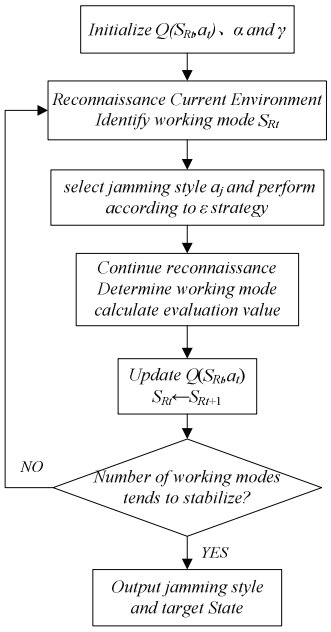

Fig.4 IRC process with unknown radar working modes

IV. SIMULATION EXPERIMENT

Radar has a variety of working modes, such as search, tracking, guidance, recognition and so on. Similarly, there are different jamming styles for various working modes, such as noise amplitude modulation, noise phase modulation, noise frequency modulation, speed deception and so on. This paper assumes that there are 5 working states $S_1, S_2, ...S_5$ from high to the end according to the threat equivalence, and the state S_5 is the goal state. Conversion between different working states is shown in Fig. 5.

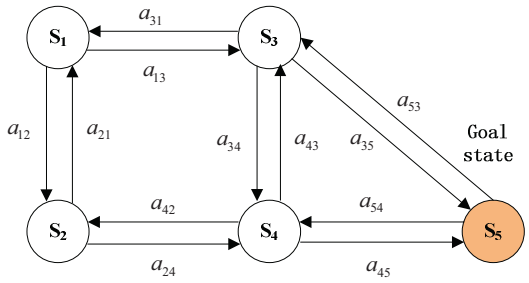

Fig. 5 Diagram of radar state conversion

Fig. 5 shows that the radar is converted from State S_i to state S_j under jamming style a_{ij}. But the Jammer has no prior knowledge, and the number of working modes and the conversion between different working modes need to be determined by RL.

The Q matrix is initialized to 0, and the discount constant is 0.5. The Q matrix value is obtained according to the algorithm steps:

$$Q = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 25 & 50 & 0 & 0 \\ 25 & 0 & 0 & 50 & 0 \\ 25 & 0 & 0 & 50 & 100 \\ 0 & 25 & 50 & 0 & 100 \\ 0 & 0 & 50 & 50 & 100 \end{bmatrix} \end{array}$$

Radar state-jamming styles value is shown in Q matrix. If the initial state of the radar is S_1, the optimal path $S_1 \xrightarrow{a_{13}} S_3 \xrightarrow{a_{35}} S_5$ is quickly found based on the Q-Learning convergence result. The process is shown in Fig.6.

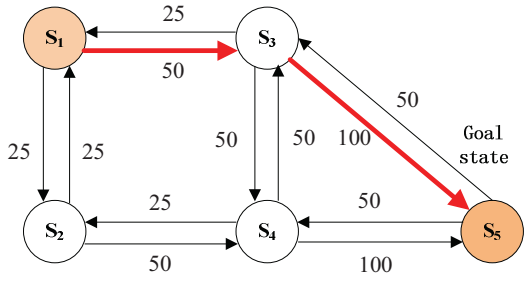

Fig. 6 Optimal path selection diagram

The convergence time of Q matrix is 4.118s in the simulation experiment. The jamming style is selected according to the convergent matrix. So we can draw conclusions as follows: the method greatly shortens the synthetic time of the jamming measures compared with the traditional radar countermeasure process, and improve the real-time of interference; providing ideas for designing jamming systems that integrate multiple jamming styles and real-time synthesis and design of new jamming patterns; resisting the multifunctional radar with various working modes.

V. Conclusion

The intelligent radar countermeasure is introduced in this paper, the difference between traditional radar countermeasure and intelligent radar countermeasure are compared. Aiming at the main problems encountered in the conventional radar countermeasure process, the intelligent radar countermeasure based on reinforcement learning is researched and simulation experiments are carried out. The simulation results show that the intelligent radar countermeasure system can learn and select jamming measures independently. It improves the adaptability of the radar countermeasure system, and can resist the multi-working state radar simultaneously.

With the development of artificial intelligence and correlation algorithm and the application in radar field, the new radar technology and radar countermeasure (such as deep reinforcement learning) will also develop in a spiral way.

References

[1] Haykin S, "Cognitive Radar: a Way of the Future," IEEE Signal Procesing Magazine, vol.23, no.1, 2006, pp.30-40.

[2] K Zhang, X Zhang, and J C Jin, "Elementary Study on Cognitive Electronic Warfare," Aerospace Electronic Warfare, vol.29, no.1, 2013, pp.53-56.

[3] N Shen, L Xiao, W Xie, and A G Shi, "Development of Cognitive Electronic Warfare System," Electronic Information Warfare Technology, vol.26, no.6, 2011, pp.22-26.

[4] H Y Dai, B Zhou, H Lei, and X J Shen, "Development and analysis of key technologies of cognitive EW," Aerodynamic Missile Journal, no.9, 2014, pp.57-60.

[5] H M Gong, "New field of electronic warfare-AI," Aerospace Shanghai, vol.2, 1986, pp.36-42.

[6] Z C Li, "Application of AI technology in EW," Electronic Warfare Technology, vol.2, 1988, pp.27-39.

[7] J W Zhao, and H C Chen, "Application of AI technology in EW," Electronic Warfare Technology, vol.5, 1988, pp16-26.

[8] DARPA, "Behavior Learning for Adaptive Electronic Warfare," https://www.fbo.gov, 2010-10-06.

[9] DARPA, "Adaptive Radar Countermeasures," https://www. fbo.gov, 2012-8-27.

[10] Air Force, "Cognitive Jammer," https://www.fbo. gov, 2010-1-20.

[11] DARPA, "Communications Under Extreme RF Spectrum Conditions," https://www.fbo.gov, 2010-9-10.

[12] X S Wang, M Q Zhu, and Y H Cheng, The principle of reinforcement learning and its application, Beijing: Science Press, 2014, pp.56-57.

[13] Sutton R, and Barto A, Reinforcement Learning: An Introduction, MIT Press, 2005, pp.9-21.

[14] Autovalmet, Intellectualization. http://baike.baidu.com/item/Intellec-tualization, 2017-5-5.

[15] C L Zhang, "2016 EW Dynamic Research," https://mp.weixin.qq.com/s/ndnXamPyTNh-peNrVik0Y9A##, 2017-01-03.

[16] Y H Luan, and P Zhang, "Comparative analysis of reinforcement learning method," Computer Era, no.12, 2015, pp.93-97.

[17] M L Xu, "Research on the reinforcement learning and its application," Wuxi: Jiangnan University , 2010, pp.20-22.