Research on Optimal Cooperative Attack Strategy of Multiple UAVs

Wenhua Wang¹,a, Xiaoping Wang¹,b*, Bo Liu¹,c, Wen Zhou¹,d

¹Aeronautics Engineering College Air Force Engineering University Xi’an, Shaanxi, China
a1411708905@qq.com, b*Corresponding author:81875303@qq.com, cliubowqn@163.com, d2531106025@qq.com

Abstract. Aiming at the problem of the optimal coordinated attack strategy of multi-UAV system against multiple moving targets in a known environment, a PMMAXQ layered reinforcement learning algorithm applied to multi-UAV cooperative attack is proposed. Firstly, the problem of multi-UAV coordinated strike, the successful strike conditions and the movement strategy of moving targets are described. Then the moving targets are allocated rationally through the 0-1 planning method. Finally, the MAXQ algorithm is improved by the Bayesian probability statistical formula. The simulation results show that, compared with the MAXQ algorithm, the PMMAXQ algorithm converges faster and the average cumulative reward value that the UAV can obtain is higher.

1. Introduction

Due to the complexity and dynamics of the combat environment and the limitations of the performance of a single UAV, how to coordinate and cooperate with multiple UAVs to achieve optimal strike decisions on moving targets is one of the hot issues that have been widely concerned in the field of UAVs in recent years [1-3].

The hierarchical reinforcement learning algorithm not only enables the drone to adjust its strategy online during repeated interactions with the environment and other drones, and better adapt to the dynamic environment, but also can overcome the "dimensionality curse" problem in the learning process. Therefore, reinforcement learning technology can be applied to complex domain problems. Typical HRL models are Option[4], HAM[5] and MAXQ[6]. They have studied HRL technology from different perspectives. Literature [7] proposes a combined hierarchical reinforcement learning method that integrates segmentation options with the traditional MAXQ algorithm, and automatically obtains all necessary parameters through learning for real-time collaboration of multiple robots in a completely unknown environment. Gunady [8] uses hierarchical reinforcement learning to solve the problem of minimizing repetitive work when multi-agents collaborate to complete tasks. Two aggregation levels are designed, topology aggregation and hidden aggregation, with higher convergence speed and up to 10 times the space reduction performance.

Aiming at the problem that the experience cannot be reused in the learning process, this article establishes an action state prediction table to make the UAV use the historical behavior state to predict the actions that other UAVs will take, so as to select better actions and improve teamwork ability.
2. Problem Description

2.1. Description of the Combat Problem

Assuming that the task area to be executed is a finite rectangle, some obstacles of different shapes, sizes, and positions are set in the environment. In this paper, the grid model representation method is used to divide the environmental area (including obstacles) into $L_x \times L_y$ cells, the set of all grids is $E = \{(i,j)|i=1,2,\ldots,L_x; j=1,2,\ldots,L_y\}$, and $(i,j)$ represents the grid located in the $i$-th row and $j$-th column.

After the action space is discretized, as shown in Fig. 1, the UAV and the moving target have 9 possible directions of movement at any time, but they can only perform one action, which are \{up, down, left, right, upper left, lower left, upper right, lower right, still\}.

![Figure 1](image.png)

In a known environment, we know the location of static obstacles, the location of UAVs and moving targets at the initial moment. $U = \{U_1, \ldots, U_n\}$ represents a collection of $n$ UAVs, and $G = \{G_1, \ldots, G_m\}$ represents a collection of $m$ moving targets. Both the UAVs and the moving targets are learning agents with self-aware decision-making capabilities $H$, $H = U \cup G$. At time $t \in T = \{1, 2, 3, \ldots\}$, the position of the UAV is $X_{U(i)} = (x_{U1(i)}, \ldots, x_{Un(i)})$ and the position of the moving target is $X_{G(i)} = (x_{G1(i)}, \ldots, x_{Gm(i)})$.

2.2. Conditions for Success

After a strike team is established for a target, the UAVs in the team cooperate with each other to strike the target. Assuming that the UAV and the moving target have the same speed during the strike, both have a global field of view, and there is ideal communication between the UAVs. The relative distance between the moving target and the UAV can be expressed as shown in (1):

$$d_i = \sqrt{(x_{g(i)} - x_u)^2 + (y_{g(i)} - y_u)^2}$$

Among them, $(x_{g(i)}, y_{g(i)})$ is the position of the moving target, and $(x_u, y_u)$ is the position of the $i$-th UAV in the team. Let $d_{min}$ be the minimum distance between UAVs to avoid collisions, and $d$ be the strike distance of UAVs. When the distance between each UAV and the moving target satisfies $d_{min} < d_i < d$, it is deemed to have completed the strike mission to this target, that is, satisfies $\prod \zeta_i = 1$. The calculation method of $\zeta_i$ can be expressed as shown in (2):

$$\zeta_i = \begin{cases} 1 & d_{min} < d_i < d \\ 0 & d_i > d \end{cases}$$

2.3. Sports Strategy

The movement target can adopt the intelligent movement strategy according to the current environmental situation. Fig. 2 shows a schematic diagram of the movement strategy of the movement goal. Assuming that the threat distance of the moving target is $L_v$, when there is no UAV within the threat distance of the moving target, the artificial potential field method [9] is used for reference, and the movement strategy of "vector repulsion superposition" is adopted to make the moving target and the UAV maximize the distance between each other. On the contrary, adopt a "maximum included angle" movement strategy,
take its position as the apex, form an angle with the UAVs in the team, and use the angle bisector of the maximum angle as the direction to move away from the UAV group.

![Diagram](image)

(a) When the UAV is far away from the moving target

(b) When the UAV is close to the moving target

Figure 2 Schematic diagram of sports target sports strategy

When there is no UAV within the threat range of the moving target, the moving direction of the moving target is obtained by superimposing the repulsive force vector of each UAV acting on the moving target. The magnitude of the repulsive force is related to the distance between the UAV and the moving target. The closer the distance, the greater the repulsion; the further the distance, the smaller the repulsion. It can be expressed as shown in (3):

\[
f_i = \frac{k}{2d_i^2}
\]

The repulsive force vector of each UAV acting on the moving target is superimposed, and the moving direction of the moving target is obtained as shown in (4):

\[
F_{\text{total}} = \sum_{i=1}^{n} f_i
\]

3. Target Allocation

This paper defines the process of multi-UAVs to strike each target as an independent stage, so task \( \Gamma \) can be decomposed into \( m \) strike subtasks \( \Gamma_1, \ldots, \Gamma_m \). If the \( U_i \) obtains a profit of \( c_{ij} \) when performing the subtask \( \Gamma_j \), the goal of the distribution plan is to maximize the total profit obtained.

Each UAV has one and only one moving target as the current strike mission at any time, and each moving target is executed by at least one UAV. Therefore, \( x_{ij} \) can be set as shown in (5):

\[
x_{ij} = \begin{cases} 
0 & \text{if } U_i \text{ does not perform task } \Gamma_j \\
1 & \text{if } U_i \text{ perform task } \Gamma_j
\end{cases}
\]

The 0-1 planning model is obtained as shown in (6):

\[
\begin{align*}
\max J &= \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} \\
\sum_{i=1}^{n} x_{ij} &\geq 1 \quad j = 1, \ldots, m \\
\sum_{j=1}^{m} x_{ij} &= 1 \quad i = 1, \ldots, n \\
x_{ij} &= 0, 1
\end{align*}
\]

The profit \( c_{ij} \) obtained when \( U_i \) performs subtask \( \Gamma_j \) is calculated based on four indicators: strike intention, mission completion rate, strike time, and mission income. The specific definition and calculation method are:
(1) Strike intention \( e_i \): Indicates the degree of intention of the UAV to the target. It can be expressed as shown in (7):

\[
e_i = \frac{1}{\sqrt{2\pi} \sigma} \int_{\theta_i} e^{-\frac{x^2}{2\sigma^2}} dx_i
\]  

(7)

The most ideal movement direction of the UAV is taken as the center, and each sub-set angle interval of the action is converted to the deflection angle space of the most ideal action. Among them, \( \sigma \) is a normal distribution parameter, \( \theta_i \) is the angle between the UAV and the apex of the obstacle and the horizontal line, \( z(\theta_i, \theta_2) = [\theta_i + \rho, \theta_2 + \rho] \) is the angle interval conversion function, and \( \rho \) is the angle between the UAV's movement direction and the target.

(2) Task completion rate \( e_s \) can be expressed as shown in (8):

\[
e_s = \frac{N_s}{N_x}
\]  

(8)

Among them, \( N_s \) represents the number of successfully completed strike missions, and \( N_x \) represents the total number of missions.

(3) Strike time \( e_t \) can be expressed as shown in (9):

\[
e_t = \frac{t - \tau}{\tau}
\]  

(9)

Among them, \( \tau \) is the maximum time allowed to strike this target, and \( \tau \) is the estimated time for the UAV to complete the mission.

(4) Task income \( e_j \): Determined by the type of sports goal.

In summary, the income \( c_j \) obtained by \( U \) when performing subtask \( \Gamma \), can be expressed as shown in (10):

\[
c_j = \lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3 + \lambda_4 e_4
\]  

(10)

Among them, \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) is the weight parameter, and \( \sum_{i=1}^{4} \lambda_i = 1 \).

For each moving target, calculate the profit that each UAV can obtain when performing a strike mission to the target, and then select the strike team in order according to the number of UAVs required by the target type. If a UAV is selected by multiple targets at the same time, join the team of the target with the most profit.

4. Cooperative Strike Method based on pmmaxq Layered Reinforcement Learning Algorithm

4.1. PMMAXQ Learning Algorithm

First, design the initial hierarchical structure based on prior knowledge, as shown in Fig. 3. The simple actions that the UAV can perform are: up, down, left, right, top left, bottom left, top right, bottom right, and still. The UAV coordinated strike mission consists of sub-tasks such as obstacle avoidance, reduction of the strike area and target orientation. The obstacle avoidance sub-task includes avoiding static obstacles and avoiding other UAVs. Then establish an action prediction table to record the state sequence \( \text{seq} \) experienced during the execution of the subtask. After a subtask is completed, input its complete state sequence \( \text{seq} \) into the action prediction table, and then clear \( \text{seq} \).
After extending the MAXQ algorithm to the field of multi-UAVs, this paper uses Bayesian formula and probability statistics to construct a PMMAXQ hierarchical reinforcement learning algorithm based on probability prediction to estimate the action probability of other UAVs. If $UAV_i$ and $UAV_j$ belong to the same strike team, the probability that $UAV_i$ believes that $UAV_j$ may take action $a_k$ in state $s$ can be expressed as $p'(s, j, a_k)$. Assuming that the number of times that $UAV_j$ performs different actions in state $s$ is $N'_j$, and the number of times that action $a_k$ is executed is $N'_j(k)$, the calculation formula of $p'(s, j, a_k)$ is shown in (11):

$$p'(s, j, a_k) = \frac{N'_j(k)}{N'_j}$$  \hspace{1cm} (11)

$UAV_i$ is in the current state $s$, according to the probability that other UAVs in the team will take actions in this state, the optimal action selection is made, so that the strike team can achieve the optimal equilibrium state after performing joint actions. As in (12), when all UAVs take their own actions, $UAV_i$ observes the new state after the joint action and the actions taken by other UAVs in the team, and updates the probability that $UAV_i$ may take their own actions against other UAVs in state $s$.

$$p(a' \mid a', s') = \frac{p(s' \mid a', a') \cdot p(a')} {p(s' \mid a')} = \frac{p(s' \mid a', a') \cdot p'(s, j, a_k)} {p(s' \mid a')}$$  \hspace{1cm} (12)

Among them, $p(a' \mid a', s')$ represents the probability that if $UAV_i$ takes action $a'$ in state $s'$, $UAV_j$ takes action $a'$ to maximize the interests of the team. $p(s' \mid a', a')$ represents the probability that if $UAV_i$ takes action $a'$ and $UAV_j$ takes action $a'$, the strike team is in state $s'$. $p(s' \mid a')$ represents the probability that $UAV_i$ will be in state $s'$ after taking action $a'$.

In the strike team, the actions of $UAV_i$ itself are recorded as $a_i$, and the joint actions of other UAVs in the team are recorded as $a'_j$. The probability formula for estimating the joint actions of other UAVs through $UAV_i$ is shown in (13):

$$p(a'_j \mid a_i, s') = \prod_{a'_j} p(a'_j \mid a'_j, s')$$  \hspace{1cm} (13)

4.2. Reward Function

In the reinforcement learning process, instantaneous rewards are generally divided into target rewards and action penalties. In order to obtain more reward signals in the learning process, this article chooses to use action penalty to represent the reward function, and give different punishments or rewards to different actions in the same state. The actions of UAVs are given instantaneous rewards in terms of moving towards the target, reducing the strike area, avoiding static obstacles and other UAVs. The instantaneous reward model is shown in (14):
Among them: \( q_r \) represents the reward signal when the UAV tends to the moving target. \( r_d \) represents the reward signal when avoiding static obstacles or other UAVs. \( r_s \) represents the reward signal when narrowing the strike area. And \( a_1, a_2, a_3 \) is the corresponding weight parameter.

When the drone tends to the moving target, it will give a positive reward signal; otherwise, it will give a negative reward signal. The specific definition is shown in (15):

\[
    r_q = \begin{cases} 
    1 & \Delta d < 0 \\
    -1 & \Delta d > 0 \\
    0 & \Delta d = 0 
\end{cases}
\]  

(15)

Let \( d(x_i(t), x_j(t)) \) be the distance between the UAV and the moving target at time \( t \), and \( \Delta d \) is expressed as shown in (16):

\[
    \Delta d = d(x_i(t), x_j(t)) - d(x_i(t-1), x_j(t-1))
\]  

(16)

In the process of hitting the target, the UAV can treat other UAVs as obstacles in the environment. Therefore, the definition of the reward signal for the UAV to avoid static obstacles or other UAVs is shown in (17):

\[
    r_d = \begin{cases} 
    1 & \zeta > d_a \\
    \frac{d_d - \zeta}{\zeta - d_b} & d_b < \zeta < d_a \\
    -1 & \zeta < d_b 
\end{cases}
\]  

(17)

Among them, \( \zeta \) is the distance between the UAV and obstacles or other UAVs, \( d_a \) and \( d_b \) are the distance thresholds.

When the UAVs are all outside the threat range of the moving target, the reward signal is determined according to the coordination between the trending target and the UAV mentioned above. When a UAV enters the threat range of a moving target, the reward signal is determined according to the extent to which the UAV narrows the strike area, as defined in (18):

\[
    r_s = \begin{cases} 
    1 & \Delta N < 0 \\
    -1 & \Delta N > 0 \\
    0 & \Delta N = 0 
\end{cases}
\]  

(18)

Let \( N(G_j(t)) \) denote the unoccupied number of adjacent grids of the moving target \( G_j \) at time \( t \), then \( \Delta N \) denote: \( \Delta N = N(G_j(t)) - N(G_j(t-1)) \)

The weights of the three reward signals are updated according to the position between the UAV and the moving target. The update rule can be expressed as shown in (19):

\[
    a_i = \begin{cases} 
    0.1 & u \in L_0 \\
    0.5 & u \in L_0 \\
    0.4 & u \in L_0 \\
    0.2 & u \in L_0 \\
    0.5 & u \in L_0 
\end{cases}
\]  

(19)

\( L_0 \) is the threat range of the moving target \( G_j \).

4.3. Action Selection Strategy

In order to enable the UAV to effectively explore and compare various actions when performing action selection, the Boltzmann distribution probability is introduced here for action selection, and the randomness of the selected action is increased. Then the probability of subtask \( M_i \) selecting action \( a_j \) in state \( s \) can be expressed as shown in (20):

\[
    p(a_j | M_i, s) = \frac{e^{\beta(M_i, a_j)T}}{\sum_{a_i \in A} e^{\beta(M_i, a_i)T}}
\]  

(20)

Among them, \( T \) is the temperature value, which represents the randomness. In the initial stage of learning, the \( T \) value is relatively large, and it has a relatively large exploratory movement ability. As the learning progresses, the \( T \) value should gradually decrease after each successful learning cycle.
ensure the learning convergence. Then the temperature of the $k$-th learning cycle can be expressed as shown in (21):

$$T_k = T_0 \cdot \varphi^k$$

Among them, $T_0$ is the initial temperature, $\varphi$ is the temperature drop coefficient and $\varphi \in [0, 1]$.

5. Simulation Experiment and Algorithm Analysis

Set up the multi-UAV cooperative attack simulation environment: Set the environment as a rectangular area of 300*300 and discretize it into a grid area of 30*30 according to the size of 10*10. The black solid area represents obstacles, the UAV is represented by a red solid circle, and the moving target is represented by a blue solid circle. The schematic diagram of the distribution of the UAV, the moving target and the obstacle is shown in Fig. 4. Randomly generate the initial positions of UAVs and moving targets in the environment, and move a grid each time. The strike range of the UAV is 1 grid, which is the area enclosed by the red frame in the figure. The threat distance of the moving target is 5 grids, that is the area enclosed by the blue frame in the picture.

![Figure 4: Schematic diagram of the distribution of UAVs, moving targets and obstacles](image)

After successfully completing the task of a certain sports goal, the goal disappears from the environment, and the reward of the third power of the target type is obtained, and it is evenly distributed to the team members. Then the team disbanded and randomly generated a target of the same type in the environment. The simulation parameters are shown in Table 1.

| Parameter | Value | Definition                                      | Parameter | Value | Definition          |
|-----------|-------|-------------------------------------------------|-----------|-------|---------------------|
| $d_{\text{min}}$ | 10    | Minimum collision avoidance distance between machines | $d_s$     | 60    | Distance threshold  |
| $d$       | 10    | UAV strike distance                              | $d_s$     | 20    | Distance threshold  |
| $\sigma$  | 1     | Normal distribution parameters                   | $\gamma$  | 0.9   | Discount factor     |
| $t$       | 800   | Maximum time allowed                             | $\alpha$  | 0.1   | Learning rate       |

In order to verify the effect of the multi-UAV cooperative attack on moving targets based on the PMMAXQ layered reinforcement learning algorithm proposed in this paper, assuming that there are 10 UAVs and 6 moving targets in the environment, 20 experiments are performed under the same initial conditions.
conditions, 3000 trainings are performed for each trial, and the maximum number of simulation cycles for each training is 1000. Analyze the experimental results from the following two aspects:

1) Comparison between different learning algorithms. Compare the PMMAXQ algorithm in this paper with the MAXQ algorithm in terms of convergence and the average cumulative reward of each UAV. Fig. 5(a) shows the comparison of the convergence between the PMMAXQ hierarchical reinforcement learning algorithm and the MAXQ learning algorithm. Due to the existence of obstacles, the convergence process is more volatile, but it can be seen that the PMMAXQ learning algorithm is in convergence speed and effect Both are better than MAXQ algorithm. Fig. 5(b) shows the comparison of the cumulative reward value of each UAV during the learning process of the two algorithms. In the early stage, the PMMAXQ algorithm needs to build an action state table, and the average reward value increases slower than the MAXQ algorithm. After about 500 trainings, the action status table is well constructed, the strike team can complete the task quickly, and the accumulated reward value increases rapidly.

2) Compare the impact of different target allocation mechanisms on the combat results. In the experiment, the number of UAVs was 10, the number of moving targets was changed to 8, and the types were randomly generated, and other test conditions were unchanged. Compare the random allocation strategy and the 0-1 planning allocation strategy on the total time to complete the task. Fig. 6 shows the time required for the two allocation strategies to hit the target. It can be seen that the performance of the 0-1 planning allocation strategy is significantly better than the random allocation strategy.
6. Conclusion
Through the research of reinforcement learning methods, a PMMAXQ layered reinforcement learning algorithm applied to multi-UAV cooperative strike is proposed. Based on the 0-1 planning distribution mechanism, the optimal strike team is formed by comprehensive consideration from four aspects of strike intention, mission completion rate, strike time and mission income indicators. Using probability statistics and Bayesian formulas to improve the MAXQ algorithm, by establishing a motion state prediction table, the UAV can predict the actions that other UAVs will take, and further improve the collaboration ability between teams. The simulation results show that the PMMAXQ algorithm proposed in this paper can improve the ability of UAVs to perform tasks cooperatively, and has faster convergence.

References
[1] I. K. Ha, Y.Z. Cho. “A Probabilistic Target Search Algorithm Based on Hierarchical Collaboration for Improving Rapidity of Drones,” in ISensors, 2018, 18(8).
[2] E. T. Alotaibi, S. S. Alqefari and A. Koubaa, "LSAR: Multi-UAV Collaboration for Search and Rescue Missions," in IEEE Access, vol. 7, 2019, pp. 55817-55832.
[3] X. W. Fu, G. W. Wei, and X. G. Gao, “Cooperative area search algorithm for multi-UAVs in uncertainty environment,” in Systems Engineering and Electronics, 2016, 38(04), pp. 821-827.
[4] R. S. Sutton, D. Precup, and S. R. Singhm, “Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning” in Artificial Intelligence, 1999, 112(1-2), pp. 181-211.
[5] R. Parr, Hierarchical Control and Learning for Markov Decision Processes. Berkeley, CA: University of California, 1998.
[6] T. G. Dietterich, “Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition” in Journal of Artificial Intelligence Research, vol. 13, 2000, pp. 227-303.
[7] Y. Cai, S. X. Yang, and X. Xu, “A combined hierarchical reinforcement learning based approach for multi-robot cooperative target searching in complex unknown environments” IEEE, 2013, pp. 218-230.
[8] M. K. Gunady, W. Gomaa, I. Takeuchi, “Aggregate Reinforcement Learning for multi-agent territory division: The Hide-and-Seek game” in Engineering Applications of Artificial Intelligence, 2014, 34(09), pp. 122-136.
[9] Y. B. Chen, G. C. Luo, and Y. S. Mei, “UAV path planning using artificial potential field method updated by optimal control theory” in International Journal of Systems ence, 2016, 47(06), pp. 1407-1420.