A Method of UAV Formation Transformation Based on Reinforcement Learning Multi-agent

Kunfu Wang, Ruolin Xing, Wei Feng, and Baiqiao Huang(✉)

System Engineering Research Institute of China State Shipbuilding Corporation, BeiJing, China
bq_huang@126.com

Abstract. In the face of increasingly complex combat tasks and unpredictable combat environment, a single UAV can not meet the operational requirements, and UAVs perform tasks in a cooperative way. In this paper, an improved heuristic reinforcement learning algorithm is proposed to solve the formation transformation problem of multiple UAVs by using multi-agent reinforcement learning algorithm and heuristic function. With the help of heuristic back-propagation algorithm for formation transformation, the convergence efficiency of reinforcement learning is improved. Through the above reinforcement learning algorithm, the problem of low efficiency of formation transformation of multiple UAVs in confrontation environment is solved.

Keywords: Multi UAV formation · Formation transformation · Agent · Reinforcement learning

1 Introduction

With the development of computer, artificial intelligence, big data, blockchain and other technologies, people have higher and higher requirements for UAV, and the application environment of UAV is more and more complex. The shortcomings and limitations of single UAV are more and more prominent. From the functional point of view, a single UAV has only part of the combat capability and can not undertake comprehensive tasks; From the safety point of view, a single UAV has weak anti-jamming ability, limited flight range and scene, and failure or damage means mission failure. Therefore, more and more research has turned to the field of UAV cluster operation. UAV cluster operation is also called multi UAV cooperative operation, which means that multiple UAVs form a cluster to complete some complex tasks together [1]. In such a multi UAV cluster, different UAVs often have different functions and play different roles. Through the cooperation among multiple UAVs, some effects that can not be achieved by a single UAV can be achieved. Based on the reinforcement learning algorithm of multi-agent agent, this paper introduces the heuristic function, and uses the heuristic reinforcement learning of multi-agent agent to solve the formation transformation problem of multi UAV formation in unknown or partially unknown complex environment, so as to improve the solution speed of reinforcement learning.

Z. Qian et al. (Eds.): WCNA 2021, LNEE 942, pp. 187–195, 2022.
https://doi.org/10.1007/978-981-19-2456-9_20

2 Research Status of UAV Formation

With the limited function of UAV, facing the increasingly complex combat tasks and unpredictable combat environment, the performance of a single UAV can not meet the operational requirements gradually. UAV more in the way of multi aircraft cooperative operation to perform comprehensive tasks. Multi UAV formation is an important part of multi UAV system, and it is the premise of task assignment and path planning. But it has also been challenged in the dynamic environment of high confrontation, including: (1) the multi UAV formation constructed by the existing formation method can not be satisfied both in formation stability and formation transformation autonomy (2) When formation is affected, it is necessary to adjust, the formation transformation speed is not fast enough, the flight path overlaps and the flight distance is too long.

The process of multi UAV system to perform combat tasks includes: analysis and modeling, formation formation, task allocation, path allocation, and task execution. When encountering emergency threat or task change, there are formation transformation steps. Among them, the formation method of UAV is always used as the foundation to support the whole task. The formation control strategy of UAV is divided into centralized control strategy and distributed control strategy [2]. The centralized control strategy requires at least one UAV in the UAV formation to know the flight status information of all UAVs. According to these information, the flight strategies of all UAVs are planned to complete the combat task. Distributed control strategy does not require UAVs in formation to know all flight status information, and formation control can be completed only by knowing the status information of adjacent UAVs (Table 1).

Table 1. Parison of advantages and disadvantages between centralized control and distributed control

Name	Advantage	Disadvantage
Centralized Control Strategy	Simple and complete theory	Lack of flexibility, fault tolerance, communication pressure
Distributed Control Strategy	High flexibility and low communication requirements	It is difficult to realize and is likely to be disturbed

The advantages of centralized control strategy are simple implementation and complete theory; The disadvantages are lack of flexibility and fault tolerance, and the communication pressure in formation is high [3]. The advantage of distributed control strategy is that it reduces the requirement of UAV Communication capability and improves the flexibility of formation; The disadvantage is that it is difficult to realize and the formation may be greatly disturbed [4].

Ru Changjian et al. designed a distributed predictive control algorithm based on Nash negotiation for UAVs carrying different loads in the mission environment, combined with the multi-objective and multi person game theory and the Nash negotiation theory of China. Zhou shaolei et al. established the UAV virtual pilot formation model

and introduced the neighbor set, adopted distributed model predictive control to construct the reconfiguration cost function of multi UAV formation at the same time, and proposed an improved quantum particle swarm optimization algorithm to complete the autonomous reconfiguration of multi UAV formation. Hua siliang et al. studied the communication topology, task topology and control architecture of UAV formation, analyzed the characteristics of task coupling, collision avoidance and dynamic topology of UAV formation reconfiguration, and proposed a model predictive control method to solve the UAV formation reconfiguration problem. Wang Jianhong transformed the nonlinear multi-objective optimization model based on autonomous reconfiguration of multi UAV formation into a standard nonlinear single objective optimization model, and solved the optimal solution through the interior point algorithm in operational research. Mao Qiong et al. proposed a rule-based formation control method aiming at the shortcomings of existing methods in UAV formation control and the characteristics of limited range perception of UAV system [5–8].

3 Agent and Reinforcement Learning

3.1 Agent

The concept of agent has different meanings in different disciplines, and so far there has been no unified definition. In the field of computer, agent refers to the computer entity that can play an independent role in the distributed system. It has the following characteristics:

1) Autonomy: it determines its own processing behavior according to its own state and perceived external environment;
2) Sociality: it can interact with other agents and work with other agents;
3) Reactivity: agent can perceive the external environment and make corresponding response;
4) Initiative: be able to take the initiative and show goal oriented behavior;
5) Time continuity: the process of agent is continuous and circular;

A single agent can perceive the external environment, interact with the environment and other agents, and modify its own behavior rules according to experience, so as to control its own behavior and internal state. In the multi-agent system, there are agents who play different roles. Through the dynamic interaction, they make use of their own resources to cooperate and make decisions, so as to achieve the characteristics that a single agent does not have, namely, emergence behavior. Each agent can coordinate, cooperate and negotiate with each other. In the multi-agent system, each agent can arrange their own goals, resources and commands reasonably, so as to coordinate their own behaviors and achieve their own goals to the greatest extent. Then, through coordination and cooperation, multiple agents can achieve common goals and realize multi-agent cooperation. In the agent model, the agent has belief, desire and intention. According to the target information and belief, the agent can generate the corresponding desire and make the corresponding behavior to complete the final task (Fig. 1).

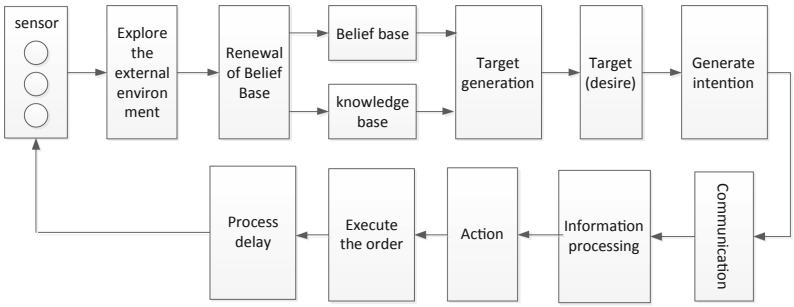

Fig. 1. Agent behavior model

When there are multiple agents in a system that can perform tasks independently, the system is called multi-agent system. In the scenario of applying multi-agent system to deal with problems, the focus of problem solving is to give full play to the initiative and autonomy of the whole system, not to emphasize the intelligence of a single agent. In some scenarios, it is often impossible to simply use the reinforcement learning algorithm of single agent to solve the problem of multi-agent (Fig. 2).

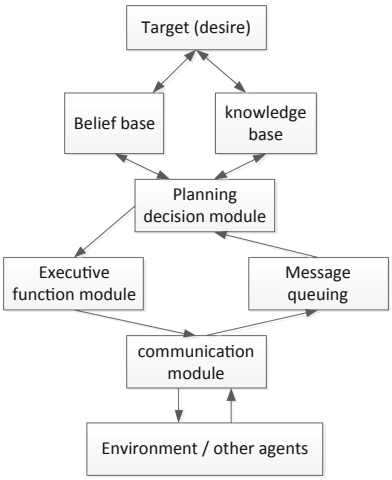

Fig. 2. The structure of agent in combat simulation architecture

According to the classification of Multi-Agent Reinforcement learning algorithm, it can be divided into the following categories according to the types of processing tasks

(1) Multi agent reinforcement learning algorithm in the case of complete cooperation. All the participants in the system have the same optimization goal. Each agent makes its own action by assuming that the other agents choose the optimal action in the current state, or makes some combination action through the cooperation mechanism to obtain the optimal goal.

(2) Multi agent reinforcement learning algorithm under complete competition. The goals of all participants in the system are contrary to each other. Each agent assumes that the other agents make the actions to minimize their own benefits in the current state, and make the actions to maximize their own benefits at this time.
(3) Reinforcement learning algorithm of multi-agent agent under mixed tasks. It is the most complex and practical part in the current research field.

3.2 Reinforcement Learning

The standard reinforcement learning algorithm mainly includes four elements: environment, state, action and value function. The problem can be solved by constructing mathematical model, such as Markov decision process (Fig. 3).

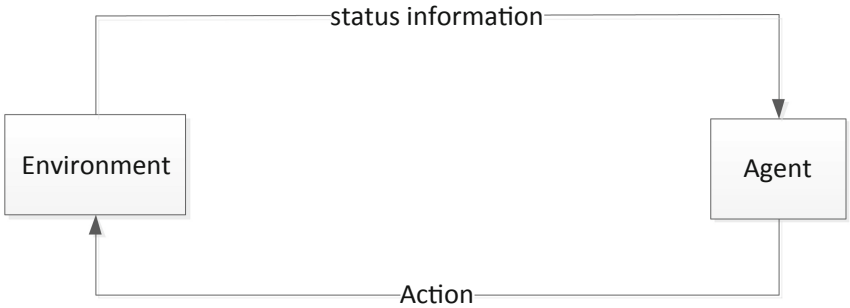

Fig. 3. Basic concept map of reinforcement learning

At present, the research on agent reinforcement learning algorithm has built a perfect system and achieved fruitful results. However, the processing ability and efficiency of a single agent are always limited. It is an effective way to solve the problems in complex environment by using the Multi-Agent Reinforcement learning algorithm. When there are multiple agents in a system that can perform tasks independently, the system is called multi-agent system. In the scenario of multi-agent system, the key point of problem solving is to give full play to the initiative and autonomy of the whole system, not the intelligence of single agent. In some scenarios, it is difficult to use the reinforcement learning algorithm of single agent to solve the problem of multi-agent. Therefore, the research and attention of experts and scholars on the reinforcement learning algorithm of multi-agent is improving.

4 A Method of UAV Formation Transformation Based on Reinforcement Learning Multi-agent

4.1 Description of UAV Formation Transformation Model

The core model of reinforcement learning: Markov decision-making process is usually composed of a quadruple: M = (S, A, P_{sa}, R). S represents the states in finite space; A

represents the actions in finite space; P_sa represents the probability set of state transfer, that is, in the current s ∈ S state, the probability that action a ∈ A will be transferred to other states after action a ∈ A is selected; R represents the return function, which is usually a function related to state and action, which can be expressed as r(s, a). The agent takes action a under state s, and performs the following actions. The expected return can be obtained as follows:

$$R_{sa} = E\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} | S = s, A = a\right] \tag{1}$$

γ is a discount factor with a value between 0 and 1, which makes the effect of the later return on the return function smaller. It simulates the uncertainty of the future return and makes the return function bounded.

In this paper, four tuples (S, A, P, R) are used to represent the Markov decision process model for formation transformation of multiple UAVs. Where S is the state space set of UAV, A is the action space set of UAV, P is the state transition probability of UAV, and R is the action return function of UAV.

Let the UAV move in the constructed two-dimensional grid, and use $Z(Z > 0)$ to represent a positive integer, then the two-dimensional grid space is Z^2, and the UAV coordinate in the two-dimensional grid space is (x_{ti}, y_{ti}), indicating the state s of UAV $s_{ti} \in Z^2$, and toward the corresponding target point $G_i(i = 1, 2, 3, .. N)$ motion, the target point of each UAV will be given in advance according to the conditions. During the flight of UAV I, action set $A_i(s) = \{up, down, left, right, stop\}$.

4.2 A Method of UAV Formation Transformation Based on Reinforcement Learning Multi Agent Agent

The fundamental goal of reinforcement learning is to find a strategy set $\langle S, A\rangle$ so that the expected return of agent in any state is the largest. The agent can only get the immediate return of the current step each time. We choose the classical Q-learning algorithm state action value function Q(s, a) instead of R_{sa}. According to a certain action selection strategy, the agent makes an action in a certain state and gets immediate feedback from the environment. The Q value increases when it receives positive feedback, and decreases when it receives negative feedback. Finally, the agent will select the action according to the Q value. The action selection function of traditional Q-learning algorithm is as follows:

$$\pi(s) = \begin{cases} \arg\max\left[Q(s, a)\right], & if\ q < 1 - \varepsilon \\ a_{randon} & otherwise \end{cases} \tag{2}$$

ε is a parameter of $\varepsilon -$ greedy, When the random number q is less than $1 - \varepsilon$ Choose the behavior a that makes the Q value maximum, otherwise choose the random behavior a. In the practical algorithm design, the iterative approximation method is usually used to solve the problem:

$$Q^*(s, a) = Q(s, a) + \alpha\left[r(s, a) + \gamma maxQ\left(s', a\right) - Q(s, a)\right] \tag{3}$$

where a is the learning factor, the larger the value of a is, the less the results of previous training are retained; $maxQ\left(s', a\right)$ is the prediction of Q value, as shown in algorithm 1:

Algorithm 1 Q-learning algorithm

Input: iteration times T, state set S, learning rate a, exploration rate ϵ, Discount factor γ
Output: state action value function Q (S, A)
1.Initialize the Q values of all States and actions
2.For i = 1 to T do:
3. Initialize state s as the first state
4. While the final state is not reached:
5. use $\varepsilon -$ greedy selects action A according to the current state S
6. Perform action A in current state S, get new status S′and reward r (S, A)
7. Update Q value:$Q(S,A) = Q(S,A) + \alpha[r(S,A) + \gamma \max Q(S',A) - Q(S,A)]$
8 $S = S'$
9. End Wbile
10. End For
11. Return Q (S, A)

In this paper, the multi UAV formation problem based on reinforcement learning can be described as: UAV interacts with the environment, learning action strategy, so that the whole UAV group can reach their respective target points with the minimum consumption steps without collision. In the process of learning the optimal action strategy, when all UAVs arrive at the target point, the group will get a positive feedback r_+, otherwise it will get a negative feedback r_.

The reinforcement learning algorithm of multi-agent needs to change the action of each agent in each state to $a_{si}(i = 1, 2, \ldots n)$ is regarded as a joint action $\rightarrow_{a_{si}}$ can be considered. The learning process of the algorithm is complex, consumes more resources and is difficult to converge. Therefore, we introduce heuristic function H to influence the action selection of each agent. Formula 1.2 can be changed as follows:

$$\pi^H(s) = \begin{cases} argmax\big[Q(s,a) + \beta H(s,a)\big], & if\ q < 1 - \varepsilon \\ a_{randon'} & otherwise \end{cases} \tag{4}$$

where β is the real number that controls the effect of the heuristic function on the algorithm. The heuristic function H needs to be large enough to affect the agent's action selection, and it should not be too large to prevent the error that affects the result. when β is 1, the mathematical expression of heuristic function H can be defined as:

$$\pi^H(s) = \begin{cases} argmax\big[Q(s,a) + \beta H(s,a)\big], & if\ q < 1 - \varepsilon \\ a_{randon'} & otherwise \end{cases} \tag{5}$$

where δ is a relatively small real number, which makes the heuristic function H larger than the difference between Q values and does not affect the learning process of reinforcement learning. The whole process of improved heuristic reinforcement learning is as follows (Fig. 4):

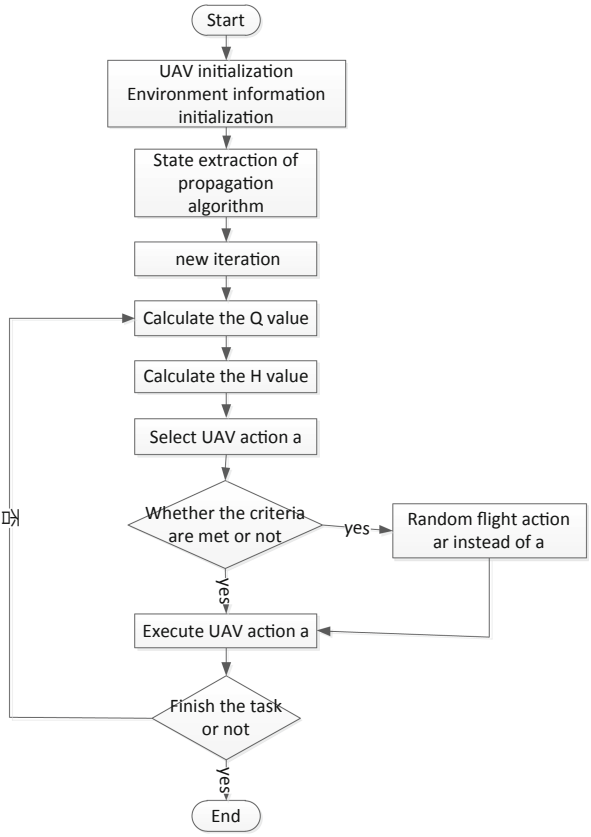

Fig. 4. The whole flow chart of improved heuristic reinforcement learning

5 Summary

In this paper, a reinforcement learning based multi-agent UAV formation transformation method is proposed. The heuristic algorithm is used to improve the traditional reinforcement learning algorithm, and the optimal path without collision is planned for the multi UAV system in the formation transformation stage, which solves the problem that the reinforcement learning algorithm consumes a lot of computing resources when facing the multi-agent problem.

References

1. Jia, Y., Tian, S., Li, Q.: Recent development of unmanned aerial vehicle swarms. Acta Aeronautica ET Astronautica Sinica 1–12 [2020–02–19]
2. Li, L., Xu, Y., Jiang, Q., Wang, T.: New development trends of military UAV equipment and technology in the world in 2018. Tactical Missile Technol. **02**, 1–11 (2019)
3. Wang, Q.-Z., Cheng, J.-Y., Li, X.-L.: Method research on cooperative task planning for multiple UCAVs. Fire Cont. Comm. Cont. **43**(03), 86–89+94 (2018)

4. Chen, X., Serrani, A., Ozbay, H.: Control of leader-follower formations of terrestrial UAVs. IEEE Conf. Deci. Cont. **1**(1), 498–503 (2004)
5. Jie, Y., et al.: UAV Form. Cont. Based Impr. APF. **3160**, 358–364 (2014)
6. Ili, P., Wang, H., Li, X.: Improved ant colony algorithm for global path planning. Advances in Materials, Machinery, Electronics I (2017)
7. Marsella, S., Gratch, J.: Evaluating a computational model of emotion. Autonomous Agents and Multi-Agent Systems (S1387–2532) **11**(1), 23–43 (2006)
8. Martins, M.F., Bianchi Reinaldo, A.C.: Heuristically-accelerated reinforcement learning: a comparative analysis of performance. In: 14th Annual Conference on Towards Autonomous Robotic Systems (TAROS) (2013)