Research on UAV Cluster's Operation Strategy Based on Reinforcement Learning Approach

Yi Mao(✉) and Yuxin Hu

State Key Laboratory of Air Traffic Management System and Technology, Nanjing 210007, China
mao_y@nuaa.edu.cn

Abstract. It is of necessity to formulate an overall UAV regulation scheme that covers each UAV's path and task implementation in the operation of UAV cluster. However, failure to fully realize consistency with pre-planned scheme in actual task implementation may occur considering changes of task, damage, addition or reduction of UAVs, fuel loss, unknown circumstance and other uncertainties, which thus entail a simultaneous online regulation scheme. By predicting UAV's 4D track, posture, task and demand of resources in regulation and clarifying data set in UAV cluster operation, quick planning of UAV's flight path and operation can be realized, thus reducing probability of scheme adjustment and improving operation efficiency.

Keywords: UAV · Operation strategy · Reinforcement learning approach

1 Introduction

UAV cluster operation is a complex multitask system that is affected and restricted by UAV's functional performance, load, information context, support equipment and other factors. Dynamic regulation of UAV cluster operation involves strategic training and optimization at the stage of gathering in takeoff, formation in flight, transformation of flight form, dispersion and collection of multiple UAVs of different types and structures [1–5].

Traditional training approaches, in general, include linear planning, dynamic planning, branch-and-bound method, elimination method and other conventional training optimization approaches frequently adopted in operational researches. In the application of optimal approach, the problem is often simplified for the sake of mathematical description and modeling so as to formulate an optimized regulation scheme [6–9]. UAV's specification, load, ammunition, coverage of support resources and battlefield context are all critical factors that impose influences in UAV cluster operation. Coupled with randomness and dynamicity in operation and a number of re-regulation tasks, UAV cluster operation is a NP-hard problem, characterized with great difficulties in solution, long time spent on solution and inability to realize simultaneous online UAV strategic cluster training [10–13].

Q. Liang et al. (Eds.): *Artificial Intelligence in China*, LNEE 572, pp. 28–35, 2020.
https://doi.org/10.1007/978-981-15-0187-6_4

The paper focuses on how to realize interactive learning, obtain knowledge and improve operation strategies in cluster operation assessment system so as to adapt to the environment and reach the ideal purpose [14–17]. UAV-borne computer is not informed of subsequent act, and it can only make judgment by trying every movement. Remarkably featured by trial-and-error-based search and delayed reward, reinforcement learning realizes enhancement of optimal actions and optimal movement strategies by making judgment on the environment's feedback about actions and instructing subsequent actions based on assessment. In this way, UAV cluster operation can better adapt to the environment [18].

2 UAV Cluster's Task Allocation and Modeling

Task allocation in UAV cluster operation aims at confirming UAV cluster's target and attack task, designing the path and realizing maximal overall performances and least cost in cluster attack. This is a multi-target optimization problem with numerous restrictions. Optimization indicators include: maximal value return of target, maximal coverage of target, minimal flight distance and least energy consumption. Restrictions include a certain definition ratio of target surveillance, necessity of target's location within UAV platform's attack diameter, limitations of each prohibition/flight avoidance zone and a certain number of UAV platforms that surveil the same target (no exceeding the limit), etc. [19].

Considering the features of multiple goals and restrictions, based on multi-target optimization theory, the paper sets up an overall multi-target integral planning model concerning UAV cluster's automatic task allocation.

Set the number of UAV platforms and targets as N_{v} and N_{T}, respectively. Decision-making variable in design is: $x_{i,j} \in \{0,1\}, i = \{1,2,\ldots,N_{\mathrm{T}}\}$. 1 means that UAV platform i targets at j, and it is the opposite in case of 0. Therefore, mathematical model of task allocation in UAV cluster attack is established.

Target function

(1) Cluster's least flight time f_1

One important indicator in UAV cluster's attack task allocation is "least time of UAV cluster's task implementation as much as possible," i.e., realizing shortest path of UAV platform allocated with a task.

$$P_{\mathrm{Fix}M_i} = \begin{cases} \mathrm{e}^{\frac{-R}{R_h}}, & R \le R_h \\ 0, & R > R_h \end{cases} \tag{2.1}$$

Thereinto, M_i is the total number of targets allocated to UAV platform i.

(2) Cluster's total flight time f_2

Another important indicator in UAV platform's allocation of collaborated attack is "least cost as much as possible," in which energy consumption is the key factor. Energy consumed in flight is related to flight distance and time. The less the time is, the less the

energy is consumed. Thus, indicator function is about the shortest flight time of UAV cluster.

$$\min f_2 = \sum_{i=1}^{N_v} \sum_{k=1}^{M_i} T_k \tag{2.2}$$

In order to ensure accuracy and feasibility of planned result when calculating the quantity of fuel burning/battery energy, initial path planning is indispensable, so is the calculation of energy consumption along initially planned path.

(3) Maximization of target's value f_3

Maximization of target's value is also an important indicator in task allocation on UAV platform. Under the condition of adequate attacking force, it is critical to realize maximization of target's values. The indicator function is as below:

$$\max f_3 = \sum_{i=1}^{N_v} \sum_{k=1}^{M_i} V_{\mathrm{k}} \tag{2.3}$$

Thereinto, V_{k} is the value of target K.

(4) Maximal target coverage f_4

Concerning task allocation on UAV platform, in addition to maximization of target's value, maximization of target coverage is also of great importance. Indicator function is as below:

$$\max f_4 = \frac{\sum_{i=1}^{N_v} M_i}{N_t} \tag{2.4}$$

In fact, UAV cluster's attack task allocation is about multi-target integral planning. With regard to multi-target planning, each target's relative weight can be confirmed according to decision-making intention, and multi-target integral planning can be transited to single-target planning:

$$\min f = \gamma_1(\alpha_1 f_1 + \alpha_2 f_2) + \gamma_2(\beta_1(1 - f_3) + \beta_2(1 - f_4)) \tag{2.5}$$

Thereinto, $0 \le \gamma_1, \gamma_2 \le 1, \gamma_1 + \gamma_2 = 1$

$$0 \le \alpha_1, \alpha_2 \le 1, \alpha_1 + \alpha_2 = 1$$
$$0 \le \beta_1, \beta_2 \le 1, \beta_1 + \beta_2 = 1$$

f_1, f_2 and f_3 are all normalized. Different valuations of α_1, α_2, β_1, β_2, γ_1 and γ_2 reflect preference of decision-making. Adjustment can be conducted automatically in accordance with the algorithm. For example, when attacking force is adequate, in general, $\beta_1 = 0$, $\beta_2 = 1$; while when attacking force is inadequate, $\beta_1 = 0.5$, $\beta_2 = 0.5$.

Restrictions

(1) Number of UAV platforms

The restriction imposes limitations on the number of attacking UAVs on UAV platform. The number should not exceed the total number of UAVs to be allocated with tasks on the platform.

$$\sum_{j=1}^{N_1} x_{i,j} \leq N_v \tag{2.6}$$

(2) Attack diameter

When cluster on UAV platform executes attack tasks, diameter should be the primary restriction to be taken into consideration, i.e., flight diameter on UAV platform must be within the attack diameter.

$$x_{i,j} l_{i,j_1} + \sum_{k=2}^{M_i} x_{i,j_k} l_{j_{k-1} j_k}, j_k \leq D_i \tag{2.7}$$

$l_{j_{k-1}j_k}$ refers to the distance between UAV platform and the first allocated target. $l_{j_{k-1}j_k}$, j_k refers to the distance between successive two targets allocated by UAV platform. Di refers to UAV platform's attack diameter $j = \{1, \ldots, M_i\}$.

Attack height

When cluster on UAV platform executes attack height, height should be the primary restriction to be taken into consideration, i.e., flight height on UAV platform must be within the attack height.

$$x_{i,j} H_j \leq H_i \tag{2.8}$$

H_j refers to the height of target j; H_i refers to the ascending limit of UAV platform i.

(4) Attack force

Target allocated to each UAV platform shall not exceed its attack force.

$$\sum_{k=1}^{M_i} x_{i,j_k} \leq \text{Attack}_i \tag{2.9}$$

Attack_i is each UAV platform's attack load, i.e., the maximal number of tasks to be executed.

3 Optimization of Operation Strategic Training Modeling

One key assumption is that UAV cluster's operation strategy is based on the interaction between UAV and environment. Such an interaction can be regarded as MDP (Markov decision process). Optimization of operation strategy based on reinforcement learning can adopt solutions to Markov problems. Supposing that the system is observed at time $t = t_1, t_2, \ldots, t_n$, one operation strategy's decision process is composed of five elements:

$$\left\langle S, A(S), P_{ss}^{a}, R_{ss}^{a}, V | S, S' \in S, \alpha \in A(S) \right\rangle \tag{3.1}$$

Each element's meaning is as below:

① S is the non-empty set composed of all possible states of the system. Sometimes, it is also called "system's state space," which can be a definite, denumerable or arbitrary non-empty set. S, S' are elements of S, implying the states:

② $s \in S$, $A(s)$ is the set of all possible movements at states.

③ When system is at the states at decision-making time t, after executing decision a, the probability of system's S' state at the next decision-making point $t + 1$ is P_{ss}^{a}. The transition probability of all movements constitutes one transition matrix cluster:

$$P_{ss}^{a} = \Pr\{S_{t+1} = s' | S_t = S, a_t = a\} \tag{3.2}$$

④ When system is at the states at decision-making time t, after executing decision a, the immediate return obtained by the system is R. This is usually called "return function":

$$R_{ss}^{a\prime} = E\{r_{t+1} | S_t = S, a_t = a, s_{t+1} = s'\} \tag{3.3}$$

⑤ V is criterion function (or objective function). Common criterion functions include: expected total return during a limited period, expected discounted total return and average return. It can be a state value function or state-movement function.

Based on the characteristics of UAV cluster's operation, UAV, battlefield's environment, load platform, information resources and other supporting resources are segmented in the establishment of Markov state transition model (MDP). Thereinto, MDP model's state variables include usability of platform and load, position of UAV, priority of multi-tasks and residual quantity of fuel. Usability refers to whether UAV and load are usable or not, and residual usability, whether takeoff/recycling equipment is usable or not. Variables of position state include current position, position of takeoff/recycling equipment and position of 4D track of task execution in sky. Priority is used to regulate UAVs' takeoff and landing sequence as well as priority of multiple tasks. Residual quantity of fuel can be classified into multiple levels to judge the order of UAVs' landing and priority of tasks. Finally, MDP model's state space can be obtained:

$$S = \prod_{i \epsilon A} B_i \times \sum_{j \epsilon \text{pos}} \text{PR}_m \prod_{i \epsilon A} B_i \times \sum_{n \epsilon \text{level}} \text{LE}_n \tag{3.4}$$

Thereinto, B refers to equipment's usability; A is the set of equipment; PS is UAV's position state; pos is the set of position states; PR is the priority of tasks; pri is the set of priority tasks; LE is the level of residual fuel; level is the set of levels of residual fuel.

By setting up a return function, as convergence conditions in learning, an optimized regulatory strategy can be generated by Q learning approach. The change of states corresponding to regulatory strategies is UAV regulation plan.

4 Research on UAV Cluster's Operation Strategy Algorithm

The purpose of reinforcement algorithm is to find out one strategy π so that the value of every state $V^{\pi}(S)$ or $(Q^{\pi}(S))$ can be maximized, that is:

Find out one strategy $\pi{:}S \rightarrow A$ so as to maximize every state's value:

$$V^{\pi}(S) = E\{r_1 + \gamma r_2 + \ldots + \gamma^{i-1} r_i + \cdots | s_0 = s\} \tag{4.1}$$

$$Q^{\pi}(s,a) = E\{r_1 + \gamma r_2 + \ldots + \gamma^{i-1} r_i + \cdots | s_0 = s, a_0 = a\} \tag{4.2}$$

$$v^*(s) = \max_{\pi}(v^{\pi}(s)) \tag{4.3}$$

$$Q^*(s,a) = \max_{\pi}(Q^{\pi}(s,a)) \tag{4.4}$$

Thereinto, means immediate reward at time t. $\gamma(\gamma \epsilon [0.1])$ is attenuation coefficient. $V^{\pi}(S)$ and $Q^{\pi}(s,a)$ are optimal value functions. Corresponding optimal strategy is:

In reinforcement learning, if optimal value function $Q^*(V^*)$ has been estimated, there are three movement patterns for option: greedy strategy, ε-greedy strategy and softmax strategy. In greedy strategy, movements with highest Q value are always selected, i.e., $\pi^* = \arg\max Q^*(s,a)$. In ε-greedy strategy, movements with highest Q value are selected under a majority of conditions, and random selections of movements occur from time to time so as to search for the optimal value. In softmax strategy, movements are selected according to weight of each movement's Q value, which is usually realized by Boltzmann machine. In the approach, the movements with higher Q value correspondingly have higher weight. Thus, it is more likely to be selected.

The mechanism of all reinforcement learning algorithms is based on the interaction between value function and strategy. Value function can be made use of to improve strategy; value function learning can be realized and value function can be improved by making use of assessment of strategy. In such an interaction in reinforcement learning, UAV cluster's operation strategy gradually concludes optimal value function and optimal strategy.

5 Conclusion

At present, there are mainly two types of reinforcement learning algorithms to solve UAV cluster's independent intelligent operation problems: first, value function estimation, which is adopted in reinforcement learning researches mostly extensively with quickest development; second, direct strategy space search approach, such as genetic algorithm, genetic program design, simulated annealing and other evolution approaches.

Direct strategy space search approach and value function estimation approach can be both used to solve reinforcement learning problems. Both improve Agent strategy by means of training learning. However, direct strategy space search approach does not use strategy as a mapping function from state to movement. Value function is not taken into consideration. In another word, environment status is not taken into consideration.

Value function estimation approach concentrates on value function; that is, environment status is regarded as a core element. Value function estimation approach can realize learning based on the interaction between Agent and environment, regardless of quality of strategy—either good or bad. On the contrary, direct strategy space search approach fails to realize such a segmented gradual learning. It is effective for reinforcement learning with enough small strategy space or sound structure that facilitates easiness to find out an optimal strategy. In addition, when Agent fails to precisely perceive environment status, direct strategy space search approach displays great advantages. By modeling UAV cluster's independent intelligent operation, we can find that value function estimation is better for large-scale complex problems because it can make better use of effective computing resources and reach the goal of solving UAV cluster's independent intelligent operation.

References

1. Raivio T (2001) Capture set computation of an optimally guided missile [J]. J Guid Control Dyn 24(6):1167–1175
2. Fei A (2017) Analysis on related issues about resilient command and control system design. Command Inf Syst Technol 8(2):1–4
3. Smart WD (2004) Explicit manifold representations for value-function approximation in reinforcement learning. In: AMAI
4. Keller PW, Mannor S, Precup D (2006) Automatic basis function construction for approximate dynamic programming and reinforcement learning. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp 449–456
5. Dearden R, Friedman N, Russell S (1998) Bayesian Q-learning. In: AAAI/IAAI, pp 761–768
6. Chase HW, Kumar P, Eickhoff SB et al (2015) Reinforcement learning models and their neural correlates: an activation likelihood estimation meta-analysis. Cogn Affect Behav Neurosci 15(2):435–459
7. Botvinick M (2012) Hierarchical reinforcement learning and decision making. Curr Opin Neurobiol 22(6):956–962
8. Smith RE, Dike BA, Mehra RK et al (2000) Classifier systems in combat: two-sided learning of maneuvers for advanced fighter aircraft. Comput Methods Appl Mech Eng 186(2):421–437

9. Salvadori F, Gehrke CS, Oliveira AC, Campos M, Sausen PS (2013) Smart grid infrastructure using a hybrid network architecture. IEEE Trans Smart Grid 3(4):1630–1639
10. Peng CH, Qian K, Wang CY (2015) Design and application of a VOC-monitoring system based on a Zigbee wireless sensor network. IEEE Sens J 15(4):2255–2268
11. Sara GS, Sridharan D (2014) Routing in mobile wireless sensor network: a survey. Telecommun Syst 57(1):51–79
12. Peters M, Zelewski S (2008) Pitfalls in the application of analytic hierarchy process to performance measurement. Manage Decis 46(7):1039–1051
13. Potts AW, Kelton FW (2011) The need for dynamic airspace management in coalition operations. Int C2 J 5(3):1–9
14. Yunru LI (2017) Joint tactical information system and technology development. Command Inf Syst Technol 8(1):9–14
15. Hoffman R, Jakobovits R, Lewis T et al (2005) Resource allocation principles for airspace flow control. In: Proceedings of AIAA guidance, navigation & control conference
16. Mukherjee A, Grabbe S, Sridhar B (2009) Arrival flight scheduling through departure delays and reroutes. Air Traffic Control Q 17(3):223–244
17. McLain TW, Beard RW (2005) Coordination variables, coordination functions, and cooperative timing missions. J Guid Control Dyn 28(1):150–161
18. Lau SY, Naeem W (2015) Cooperative tensegrity based formation control algorithm for a multi-aircraft system. In: American control conference, pp 750–756
19. Shi L, Liu W (2017) UAV management and control technology based on satellite navigation spoofing jamming. Command Inf Syst Technol 8(1):22–26