0.1. *Федулов В.А., Товарнов М.С.* **Имитационное моделирование движения двух антагонистических дронов, управляемых акторами – моделями обучения с подкреплением**

Беспилотные летательные аппараты (дроны) имеют большие перспективы использования, в частности, в пространстве «умных городов» [1]. Однако, являясь частью городской среды и частью «интернета вещей», дроны несут в себе множество потенциальных угроз [2]. Одним из способов ликвидации (опасного) *целевого* дрона (ЦД) является его перехват *противодействующим* дроном (ПД) [3].

В работе решались задачи навигации ЦД и ПД, управляемых обученными акторами. Задача ЦД – из точки A прилететь в точку B при заданных начальных и граничных условиях полёта. При этом свободному полёту ЦД мог препятствовать ПД, задача которого не допустить достижение конечной точки первым. Моделирование проведено в среде разработанной имитационной модели (ИМ) для трёх соответствующих задач навигации:

1. В отсутствии ПД.
2. При наличии ПД под управлением «идеального» актора, реализующего известный закон наведения.
3. При наличии ПД под управлением актора, обученного по тому же алгоритму, что и актор ЦД.

Акторы дронов тренировались следующими тремя алгоритмами обучения с подкреплением [4]:

1. Мягкий актор — критик (Soft Actor — Critic, SAC).
2. Градиентный спуск по сильно детерминированным стратегиям (Deep Deterministic Policy Gradient, DDPG).
3. Дважды отсроченный градиентный спуск по сильно детерминированным стратегиям (Twin Delayed Deep Deterministic Policy Gradient, TD3).

Функция вознаграждения акторов:

$$r_i = \Delta t_{i-1}^{\min} - \Delta t_i^{\min}, \qquad (1)$$

где r_i – величина награды актора при его переходе из состояния s_{i-1} в состояние s_i; $t_{i-1}^{\min}$ и $t_i^{\min}$ – величины наименьшего времени перемещения дрона из начальной точки в конечную соответственно для предыдущей $(i-1)$-й и текущей i-й прогонок ИМ. Наградой для актора являлось положительное значение r_i, что свидетельствовало об уменьшении времени полёта до точки назначения (до точки B для ЦД и до центра масс ЦД для ПД).

В результате работы показано, что функция (1) может использоваться в алгоритмах обучения с подкреплением при решении рассмотренных в работе задач навигации. Все три модели обучения продемонстрировали высокую эффективность управляющих действий акторов ЦД и ПД, причём наибольшая эффективность в среднем достигнута при использовании моделей DDPG и TD3.

Исследование выполнено при финансовой поддержке РФФИ (проект № 19-29-06090 мк).

Научный руководитель — к.т.н. Быков Н.В.

Список литературы

[1] Qi F., Zhu X., Mang G., Kadoch M., Li W. UAV Network and IoT in the Sky for Future Smart Cities // IEEE Network. 2019. Vol. 33. P. 96–101.

[2] Бойко А. Проблемы и опасности, связанные с беспилотниками. Инциденты. [Электронный ресурс]. URL: http://robotrends.ru/robopedia/problemy-i-opasnosti-svyazannye-s-bespilotnikami.-incidenty (дата обращения 18.08.2021).

[3] Sang Y., Cai Z., Lin Q., Wang Y. Planning Algorithm Based on Airborne Sensor for UAV to Track and Intercept Moving Target in Dynamic Environment // Proc. Intern. Conf. «Chinese Guid. Navig. Control Conf. CGNCC»>. China: IEEE, 2014. P. 1972–1977.

[4] Sutton R. S., Barto A. G. Reinforcment Learning: An Introduction / Massachusetts: The MIT Press, 2014. 352 p.