DOI: 10.25728/mlsd.2023.1308

РЕШЕНИЕ ЗАДАЧИ ОБНАРУЖЕНИЯ И РАЗРЕШЕНИЯ КОНФЛИКТОВ В ВОЗДУШНОМ ПРОСТРАНСТВЕ С ПОМОЩЬЮ ГЛУБОКОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

Кулида Е.Л., Лебедев В.Г.
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
elena-kulida@yandex.ru, lebedev-valentin@yandex.ru

Аннотация. Рассматриваются перспективные методы обнаружения и разрешения конфликтов воздушных судов. Рассматриваются алгоритмы решения задачи на основе иерархического глубокого обучения с подкреплением при полетах по заданным маршрутам и алгоритм исполнитель-критик для свободных полетов.

Ключевые слова: обнаружение и разрешение конфликтов, обучение с подкреплением, функция ценности состояния.

Введение

В последние годы происходит бурное развитие технологий ИИ, связанных с методами машинного обучения и на их основе получены важные практические результаты. Методы и модели машинного обучения все шире используются в различных областях человеческой деятельности, в том числе в управлении воздушным движением (УВД). Задачи, возникающие в УВД, в значительной степени связаны с неопределенностью вследствие метеоусловий, человеческого фактора, состояния техники и т.д. Перспективное направление адаптивного управления связано с обучением с подкреплением, которое осуществляется во взаимодействии со средой и не предполагает наличия заранее подготовленных наборов обучающих данных. При этом возможна адаптация моделей к изменяющимся условиям окружающей среды, поскольку исследование на основе отклика среды продолжается в процессе эксплуатации.

Одной из важнейших задач УВД является задача обнаружения и разрешения конфликтов между воздушными судами (ВС). Для разрешения конфликтов используются различные стратегии управления высотой полета, траекторией полета (курсом и боковыми отклонениями) и скоростью полета. Эта задача решается диспетчерами, но поскольку ее сложность постоянно возрастает в связи с ростом интенсивности воздушного движения, все более актуальной становится проблема помощи диспетчеру на основе автоматической генерации вариантов решений.

Конфликт между ВС возникает, если между ними одновременно нарушены минимальные необходимые расстояния разделения по горизонтали и по вертикали. Разрабатывались различные подходы к решению задачи разрешения конфликтов между ВС: геометрический метод, методы оптимального управления, смешанно-целочисленное линейное или нелинейное программирование, генетические алгоритмы, оптимизация роя частиц [1]. Главным недостатком этих методов, затрудняющим их практическое применение, является большое время, необходимое для решения. В последние годы в связи с успехами в применении методов обучения с подкреплением появились исследовательские работы по решению задачи обнаружения и разрешения конфликтов между ВС при помощи обучения с подкреплением [2]. Проведенные исследования свидетельствуют о перспективности предлагаемых методов и алгоритмов, поскольку обученные агенты могут получить варианты решений очень быстро, за доли секунды.

1. Задача обучения с подкреплением [3]

В основе обучения с подкреплением лежит идея создания системы, которая будет самонастраиваться с целью максимизации положительного отклика окружающей среды. Обучение с подкреплением реализует простую модель взаимодействия агента со средой, в основе которой лежат понятия состояний среды, действий агента и получаемых агентом вознаграждений при переходе среды в новое состояние. Суть такого обучения заключается не в обучении конкретным действиям, а в обратной связи в виде оценки действий агента при взаимодействии со средой. Успешность обучения зависит от того, насколько хорошо система обобщает накопленный в процессе обучения опыт. Агент должен изменять свое поведение, наблюдая за результатами своих действий (получая численное вознаграждение), при этом агент сможет реагировать на происходящие в окружающей среде изменения.

Центральным звеном метода обучения с подкреплением является алгоритм расчета функции ценности, при помощи которой определяется долгосрочная ценность состояний и действий, что

позволяет получить наибольшее суммарное вознаграждение с учетом дальнейших состояний и вознаграждений.

Целью методов обучения с подкреплением является создание систем, адаптирующихся к изменяющимся условиям реального мира. Для этого, в отличие от методов обучения с учителем, обучение должно продолжаться в процессе использования.

1.1. Постановка задачи

В основе задачи обучения с подкреплением лежит модель взаимодействия агента и среды, модель среды агенту не известна. В момент времени t среда находится в состоянии s_t из множества возможных состояний $S = \{s\}$. В состоянии s_t агент осуществляет одно из возможных в этом состоянии действий $a_t \in \{A(s_t)\}$, при этом агент получает численное вознаграждение r_t, а состояние среды изменяется на s_{t+1}.

Целью агента является максимизация ожидаемой выгоды – суммарного вознаграждения за длительный период взаимодействия со средой:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \quad (1)$$

где γ – коэффициент дисконтирования, $0 \le \gamma \le 1$, чем ближе этот коэффициент к 1, тем больше для агента значимость будущих вознаграждений. В процессе обучения формируется стратегия выбора действия в зависимости от состояния среды: $\pi(s, a)$ – вероятность выбора действия a в состоянии s.

В этой модели у агента нет другого способа обучения, кроме многократных взаимодействий со средой и накопления опыта методом проб и ошибок в процессе этих взаимодействий. Опыт агента аккумулируется в виде функций ценности.

Функция ценности состояния для стратегии π представляет собой математическое ожидание ожидаемой выгоды при следовании агента стратегии π:

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}. \quad (2)$$

Функция ценности действия для стратегии π представляет собой математическое ожидание ожидаемой выгоды при выборе агентом действия a в состоянии s и следовании стратегии π:

$$Q^{\pi}(s, a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\}. \quad (3)$$

1.2. Итерация по стратегиям

В теории обучения с подкреплением рассматриваются среды, обладающие марковским свойством. Это значит, состояние среды s_{t+1} в момент времени $t + 1$ зависит только от s_t, a_t в момент времени t и не зависит от состояний и действий в более ранние моменты времени. Теоретической основой обучения с подкреплением является задаваемая уравнениями Беллмана связь между функциями ценности состояния s и функциями ценности последующих состояний s'.

Для вычисления функции ценности состояния для стратегии π начальное приближение V_0 выбирается произвольно, последующие приближения получаются в процессе итеративного оценивания стратегии по уравнению Беллмана, которое связывает ценность состояния с ценностями состояний, следующих за ним:

$$V_{k+1}(s) = E_{\pi}\{r_{t+1} + \gamma V_k(s_{t+1}) | s_t = s\} \quad (4)$$

Зная функцию ценности $V^{\pi}(s)$ для стратегии π, можно улучшить эту стратегию, выбирая для каждого состояния наилучшее действие. Улучшенная стратегия π' называется жадной по отношению к функции ценности $V^{\pi}(s)$:

$$\pi'(s) = arg\max_a Q^{\pi}(s, a) = arg\max_a E\{r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s, a_t = a\}. \quad (5)$$

Для новой стратегии π' оценивается функция ценности $V^{\pi'}(s)$, в соответствии с новой функцией ценности формируется жадная стратегия π'' и т.д.

Этот процесс называется итерацией по стратегиям и в пределе он сходится к оптимальной функции ценности

$$V^*(s) = \max_{\pi} V^{\pi}(s) \; \forall s \in S, \quad (6)$$

которой соответствует оптимальная стратегия π^*.

2. Разные постановки задачи обнаружения и разрешения конфликтов в воздушном пространстве

2.1. Стратегическое и тактическое разрешение конфликтов

Задача обнаружения и разрешения конфликтов в воздушном пространстве решается на двух уровнях: стратегическом и тактическом. При перспективном стратегическом планировании предполагается заблаговременно анализировать совокупность четырехмерных траекторий всех полетов в масштабах целой страны или континента и ставится задача такой организации потоков ВС, при которой минимизируются потенциальные конфликты между ВС. Особенностью стратегического планирования является необходимость учета неопределенности, поскольку в реальном полете по различным причинам могут возникнуть отклонения от первоначально запланированной четырехмерной траектории. С целью учета неопределенности предлагается ввести в модель неопределенность положения ВС в каждый момент времени, т.е. предполагается, что ВС может находиться не точно в точке запланированной четырехмерной траектории, а в некоторой области вокруг этой точки с фиксированными отклонениями по каждому из 4 измерений (широта, долгота, высота, время). Потенциальный конфликт в этом случае возникает при пересечении в четырехмерном пространстве таких областей для разных ВС. При выборе параметров неопределенности необходим компромисс, поскольку чем больше будут допуски модели, связанные с неопределённостью положения ВС, тем меньше вероятность появление конфликтов в реальных полетах, однако по мере увеличения допусков будет снижаться эффективность использования воздушного пространства.

Каким эффективным ни было бы стратегическое планирование, в реальном полете могут возникнуть обстоятельства, связанные с неопределенностью метеоусловий, человеческого фактора, технического состояния ВС, которые могут привести к потенциальному конфликту. Поэтому всегда будет оставаться актуальной задача тактического разрешения конфликтов в процессе полета. Задача тактического обнаружения и разрешения конфликтов подразделяется на две: в крейсерском полете и при предпосадочном маневрировании в районе крупных аэропортов. Районы крупных аэропортов являются известным узким местом в системе управления воздушным движением, поскольку плотность ВС в них очень высокая. В этой связи организация потоков воздушных судов, особенно находящихся в воздухе в районе крупного аэропорта, должна быть продумана особенно тщательно. В основе оптимизации использования воздушного пространства лежит оптимизация очереди захода ВС на посадку с учетом различных критериев оптимизация, рациональная организация зон ожидания для ВС, ожидающих разрешения на посадку, и организация слияния нескольких потоков ВС в точке перед заходом на взлетно-посадочную полосу.

2.2. Построение модели

В литературе исследуются различные постановки задачи обнаружения и разрешения конфликтов в воздушном пространстве с помощью глубокого обучения с подкреплением.

Первым шагом при решении задачи на основе обучения с подкреплением является построение модели марковского процесса принятия решений, включающего пространство действий, пространство состояний и функцию вознаграждения.

Пространство состояний и режим управления могут быть непрерывными или дискретными.

Информация о ВС может передаваться в виде структур данных, в виде необработанных пикселей (изображений), или в виде их комбинации.

Определяется, в каком режиме реализуются полеты. В настоящее время большинство гражданских ВС осуществляют полеты по заранее определенным маршрутам. Однако из-за перегрузки воздушных трасс в качестве перспективной технологии обсуждается реализация свободных полетов, при которой ВС сможет выбирать траекторию движения от начальной точки к конечной. В литературе исследуются подходы к решению задачи обнаружения и разрешения конфликтов для обоих режимов полетов.

Определяется, является количество ВС, участвующих в конфликте, фиксированным или переменным. Решение задач с фиксированным количеством ВС проще, но при этом для сценариев с разным количеством ВС может потребоваться обучать разных агентов.

В зависимости от целей задача может решаться в двумерном или трехмерном пространстве. В двумерном пространстве движение ВС регулируется скоростью и курсом, в трехмерном – высотой, скоростью и курсом.

Большое влияние на скорость обучения, сходимость и производительность агентов оказывает функция вознаграждения. Главное ее назначение – стимулировать обеспечение безопасности и предотвращения конфликтов между ВС, поддерживая безопасную дистанцию. Важно при этом

обеспечить эффективность и рациональность принятия решений, назначая штрафы за недопустимые действия, отклонения от заданного маршрута, задержки, излишние действия по изменению высоты, курса, скорости.

Для решения задачи обнаружения и разрешения конфликтов в воздушном пространстве с помощью глубокого обучения с подкреплением разработано много различных алгоритмов [4].

Для решения задачи может быть реализован один агент, который будет корректировать полеты всех ВС, либо для каждого ВС будет управляться отдельным агентом. Каждый из подходов имеет достоинства и недостатки.

Для дискретного пространства действий и полетов по заданным маршрутам предназначен алгоритм, описанный в разделе 3. Алгоритм для непрерывного пространства действий и свободного полета представлен в разделе 4.

3. Обнаружение и разрешение конфликтов между воздушными судами при полетах по заданным маршрутам

В работе [5] был предложен метод решения задачи обнаружения и разрешения конфликтов между ВС в структурированном воздушном пространстве, названный иерархическим глубоким агентом. При этом подходе используются два агента – родительский агент выполняет действия по изменению маршрутов ВС, дочерний агент контролирует действия по изменению скорости ВС. Разделение функций по управлению маршрутами и скоростями между разными агентами после однократного принятия решения по изменению маршрутов позволяет продолжать принимать решения по корректировке скоростей.

Родительский агент для получения исходной информации использует методы компьютерного зрения, получая в качестве входных данных представление экрана, на котором отображено текущее состояние, в виде набора пикселей. На рис. 1 представлен пример экрана для двух ВС. Каждое ВС может достичь целевой точки MOD по двум различным маршрутам, т.е. возможно четыре различных комбинации маршрутов. С помощью обучения с подкреплением с двойным *Q*-обучением родительский агент анализирует ситуацию и выполняет действия по выбору маршрутов ВС.

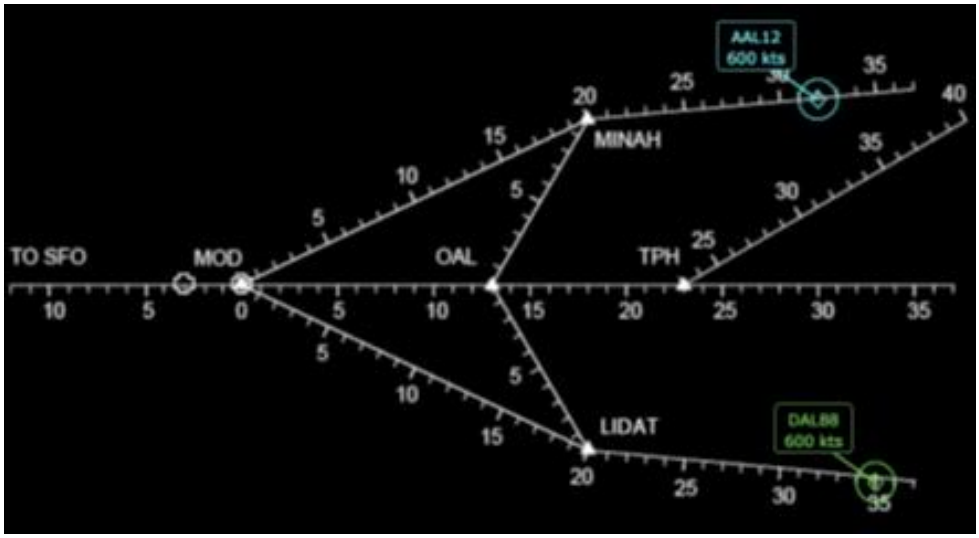

Рис. 1. Пример маршрутов двух конфликтующих ВС

В 1989 году Уоткинсом был разработан фундаментальный алгоритм обучения с подкреплением – алгоритм *Q*-обучения [6]. При *Q*-обучении на каждом временном шаге значения функции ценности действия обновляются по формуле:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right], \quad (7)$$

где α – скорость обучения. После обновления функции ценности действия обновляется стратегия действий агента. При небольшом числе состояний стратегия может быть представлена в виде таблицы. Однако при большом пространстве состояний опыта взаимодействия со средой становится недостаточно, чтобы формировать функцию Q в каждом возможном состоянии и ее необходимо аппроксимировать. Для аппроксимации функции Q используется глубокая нейронная сеть и

рассматривается функция $Q(s,a,\theta)$, где θ – параметры нейронной сети. Цель глубокого обучения определяется следующим образом:

$$Y_t' = r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_t'). \quad (8)$$

Для определения наилучшего действия используется вторая нейронная сеть, цель которой определяется следующим образом:

$$Y_t'' = r_{t+1} + \gamma Q(s_{t+1}, \operatorname{argmax}_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_t'); Q_t''). \quad (9)$$

Для дочернего агента, представляющего i-е ВС, входными данными являются позиция и скорость ВС и информация о маршруте, полученная от родительского агента. Дочерний агент принимает решения об изменении скорости ВС.

Проведенные численные эксперименты позволили сделать вывод о перспективности предлагаемого подхода.

4. Обнаружение и разрешение конфликтов между воздушными судами при свободном полете

В работе [7] рассматривается использование метода глубокого обучения с подкреплением для обнаружения и разрешения конфликтов при свободном полете. Разработана среда для обучения агентов, разработан агент на основе алгоритма исполнитель-критик, предложен алгоритм с K контрольными моментами в каждом эпизоде.

Рассматриваемый сценарий воздушного движения для N ВС представлен на рис. 2. $N-1$ ВС находятся в секторе радиуса L, в сектор входит еще одно ВС. У каждого ВС определена текущая и целевая точки. Задача каждого ВС – перелететь из текущей точки в целевую за минимальное время без конфликтов с другими ВС. Положение n-го ВС в момент времени t: $(x_n(t), y_n(t), \varphi_n(t))$.

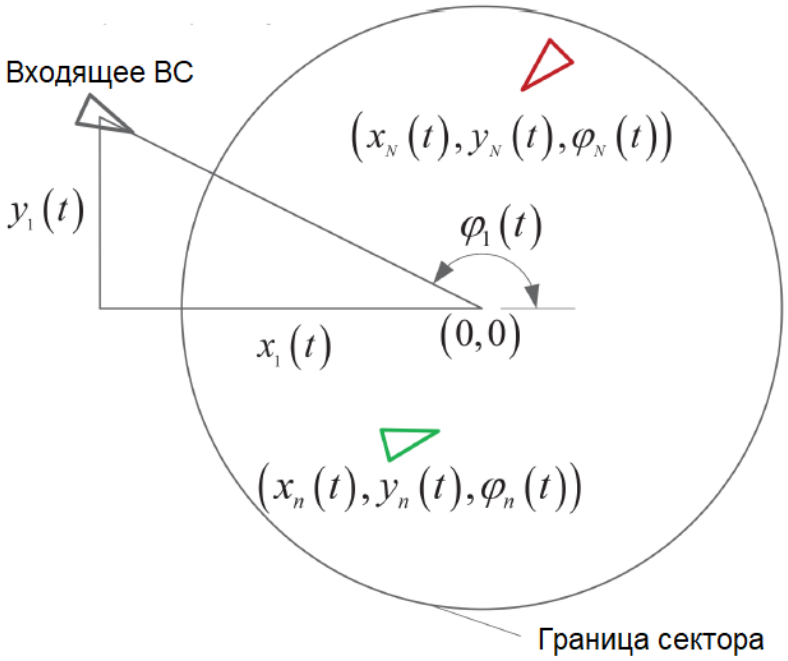

Рис. 2. Сценарий воздушного движения

Состояние среды представляет собой N-мерный вектор, $N = N_n \times N_p \times N_d$, N_n – количество ВС, N_p – количество точек пути одного ВС от текущего положения до пункта назначения, $N_d = 3$ – координаты ВС. Высота фиксирована, две другие координаты изменяемы.

Действием является выбор на каждом временном шаге новой точки, определяемой полярной координатой, в которую ВС летит из текущего положения:

$$A = \{\rho, \varphi | \rho \in [0, L], \varphi \in [-\pi, \pi]\},$$

где L – радиус сектора, ρ, φ – полярный радиус и угол.

Функция вознаграждения

$$R_t = \begin{cases} -1, \text{если есть конфликт} \\ -1/K \times |\Delta\varphi_t/\pi|, \text{иначе,} \end{cases} \quad (10)$$

где $K -$ число контрольных моментов (моментов выбора действия агентом).

Процесс обучения состоит из эпизодов. В начале каждого эпизода инициируется состояние S_0. На каждом временном шаге t агент получает состояние среды S_t, выполняет действие A_t, получает вознаграждение R_t, среда переходит в состояние S_{t+1}. Кортеж $[S_t, A_t, R_t, S_{t+1}]$ сохраняется в памяти. Процесс повторяется до конечного состояния, после чего агент выбирает данные из памяти и обучается.

Обучаются нейронные сети исполнителя и критика.

Параметры сети критика δ определяются по формуле:

$$\delta_t = R_t + \gamma\hat{V}(S_{t+1}, w) - \hat{V}(S_t, w), \quad (11)$$

где $R_t -$ немедленное вознаграждение, $\hat{V}(S_{t+1}, w) -$ значение функции ценности в следующем состоянии S_{t+1}, $\hat{V}(S_t, w) -$ значение функции ценности в текущем состоянии S_t. Функция ценности действия аппроксимируется при помощи нейронной сети $V(S, w) \approx V_\pi(S)$, где w – веса нейронов. Параметры w обновляются по методу наименьших квадратов:

$$w \leftarrow w + \alpha^w \nabla\delta^2. \quad (12)$$

Для сети исполнителя используется метод градиента политики:

$$ln\,\pi(\rho_t, \varphi_t|S_t, \theta) = ln\,\pi(\rho_t|S_t, \theta) + ln\,\pi(\varphi_t|S_t, \theta), \quad (13)$$

где $\pi(\rho_t, \varphi_t|S_t, \theta) -$ вероятность выбора ρ_t, φ_t в состоянии S_t и параметрах θ, $\pi(\rho_t|S_t, \theta) -$ вероятность выбора ρ_t в состоянии S_t и параметрах θ, $\pi(\varphi_t|S_t, \theta) -$ вероятность выбора φ_t в состоянии S_t и параметрах θ. Параметры θ обновляются по формуле:

$$\theta \leftarrow \theta + \alpha^\theta \delta_t \nabla\, ln\,\pi(\rho_t, \varphi_t|S_t, \theta). \quad (14)$$

Предложенный метод предполагается доработать для использования в будущих системах УВД. Основное преимущество предлагаемого подхода заключается в том, что хорошо обученный агент может сгенерировать решение очень быстро, за доли секунды, в то время как предыдущие методы требовали десятков и даже сотен секунд для расчетов. Эффективность предложенного подхода была продемонстрирована в разработанной среде обучения при помощи численного моделирования при фиксированной высоте и скорости полетов. Необходимо исследование предлагаемого подхода на исторических данных реальных полетов.

5. Заключение

В связи с феноменальными результатами, полученными в последние годы в области машинного обучения, было разработано и исследовано много различных моделей и алгоритмов решения задачи обнаружения и разрешения конфликтов в воздушном пространстве с помощью глубокого обучения с подкреплением. Однако при управлении воздушным движением требуется, чтобы были успешно разрешены все конфликты, но предлагаемые модели и алгоритмы являются приближенными и стопроцентный результат не гарантируют даже при моделировании. Поэтому предлагаемые методы и алгоритмы пока можно использовать только в качестве вспомогательного средства при принятии решений диспетчером. Для автономного решения задач обнаружения и разрешения конфликтов в воздушном пространстве с помощью глубокого обучения с подкреплением в реальных условиях требуются дополнительные исследования.

Обучение с подкреплением – одно из быстро развивающихся направлений машинного обучения, на основе которого уже получены значимые практические результаты в некоторых важных областях человеческой деятельности. Однако при применении обучения с подкреплением в области авиации еще остается множество неисследованных проблем, над которыми работают ученые во многих странах мира. Возможность применения и эффективность алгоритмов обучения с подкреплением при управлении воздушным движением в реальном воздушном пространстве, их поведение и сходимость с учетом всех факторов неопределенности, в том числе с учетом метеорологических условий, еще только предстоит изучить.

Литература

1. *Кулида Е.Л., Лебедев В.Г.* Методы решения задач планирования и регулирования потоков воздушного движения. Ч. 1. Стратегическое планирование четырехмерных траекторий // Проблемы управления. 2023. № 1. – С. 3–14.
2. *Кулида Е.Л., Лебедев В.Г.* Методы решения задач планирования и регулирования потоков воздушного движения Ч. 2. Применение методов глубокого обучения с подкреплением // Проблемы управления. 2023. № 2. – С. 3–18.
3. *Sutton, R.S., Barto, A.G.* Reinforcement learning: an introduction – London, UK: MIT Press, 2017.
4. *Wang, Z.; Pan,W.; Li, H.; Wang, X.; Zuo, Q.* Review of Deep Reinforcement Learning Approaches for Conflict Resolution in Air Traffic Control // Aerospace. – 2022, 9 (6), 294. – DOI: 10.3390/aerospace9060294.
5. *Brittain,M., Wei, P.* Autonomous aircraft sequencing and separation with hierarchical deep reinforcement learning // In Proceedings of the 8th International Conference on Research in Air Transportation – Barcelona, Spain, 26–29 June 2018.
6. *Watkins, C. J. C. H.* Learning from delayed rewards // Ph.D. thesis, King's College, Cambridge. – 1989.
7. *Wang, Z.; Li, H.; Wang, J.; Shen, F.* Deep reinforcement learning based conflict detection and resolution in air traffic control // IET Intelligent Transport System. – 2019. – Vol. 13. – P. 1041–1047.