ПРЕДОТВРАЩЕНИЕ СТОЛКНОВЕНИЙ ПРИ БЕЗЭКИПАЖНОМ СУДОВОЖДЕНИИ НА ОСНОВЕ АЛГОРИТМА ГЛУБОКОГО ДЕТЕРМИНИРОВАННОГО ГРАДИЕНТА СТРАТЕГИИ

Л. А. Баракат, И.Ю. Квятковская (Астрахань)

Быстрое развитие цифровых технологий, искусственного интеллекта (ИИ), информационных систем интеграции и анализа данных, наблюдаемое в последние два десятилетия, не оставляет сомнения в их важности в инновационных системах дистанционного управления судами, но столь же быстрое их устаревание ставит под вопрос о возможности повышения уровни безопасности судовождения и снижения риска столкновений.

Ответом на этот вопрос стало применение различных методов машинного обучения и, в частности, методов глубокого обучения с подкреплением (ГОП) в задачах анализа ситуаций и интеллектуального принятия решений в режиме реального времени при возникновении риска столкновений безэкипажного судна с другими объектами.

ГОП является сочетанием обучения с подкреплением (Reinforcement Learning) и глубокого обучения (Deep Learning), при котором судно (агент) и окружающая среда взаимодействуют друг с другом, способствуя обучению [1].

Для каждого дискретного времени t судно наблюдает окружающую среду и получает информацию (состояние s_t), предоставляемую разрозненными датчиками и навигационными информационными системами, например, камеры, Лидар (англ. Light Detection and Ranging – Lidar), Радар, автоматическая идентификационная система (АИС) и др [2].

После этого судно выполняет действие a_t. В результате окружающая среда меняет свое предыдущее состояние s_t на s_{t+1}, и агент получает вознаграждение (награду) $r_t = r\,(s_t, a_t)$ (рис.1).

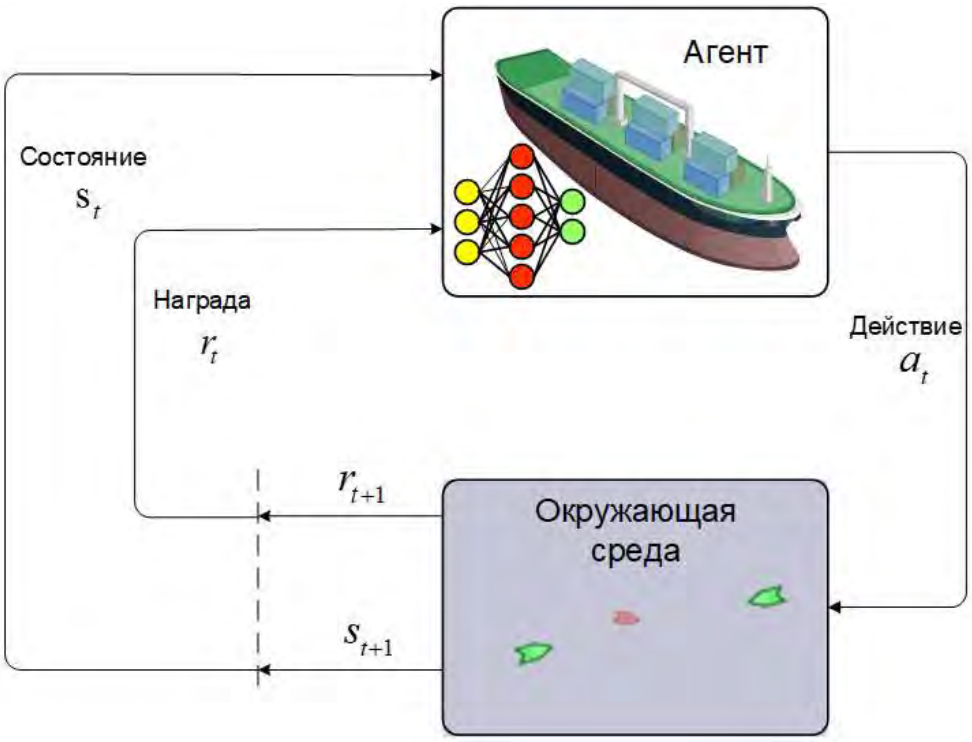

Рис. 1 – Общий принцип ГОП при БЭС

Процесс ГОП является конечным, длится T шагов, называемых эпизодами. Агент учиться, чтобы максимизировать совокупную награду R_t с учѐтом коэффициента дисконтирования $\gamma \in [0,1]$:

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i) \quad (1)$$

Выбор действий агента определяется стратегией $\pi(a=\pi(s))$, которая может быть, как детерминированной, так и стохастической.

Благодаря методу ГОП, судно может работать в соответствии с текущей ситуацией сближения с учетом международных нормативных правил предупреждения столкновений судов (МППСС-72).

При БЭС для выполнения задачи постоянного управления движением судна с непрерывными пространствами действий в рамках ГОП применяется алгоритм глубокого детерминированного градиента стратегии (DDPG).

Более того, цель детерминированности заключается в облегчении градиента стратегии во избежание случайного выбора и вывода конкретного значения действия [3].

DDPG основан на алгоритме типа «исполнитель–критик», в котором глубокие нейронные сети используются для обучения как исполнителя, так и критика. Структура алгоритма «исполнитель–критик» показана на рисунке 2.

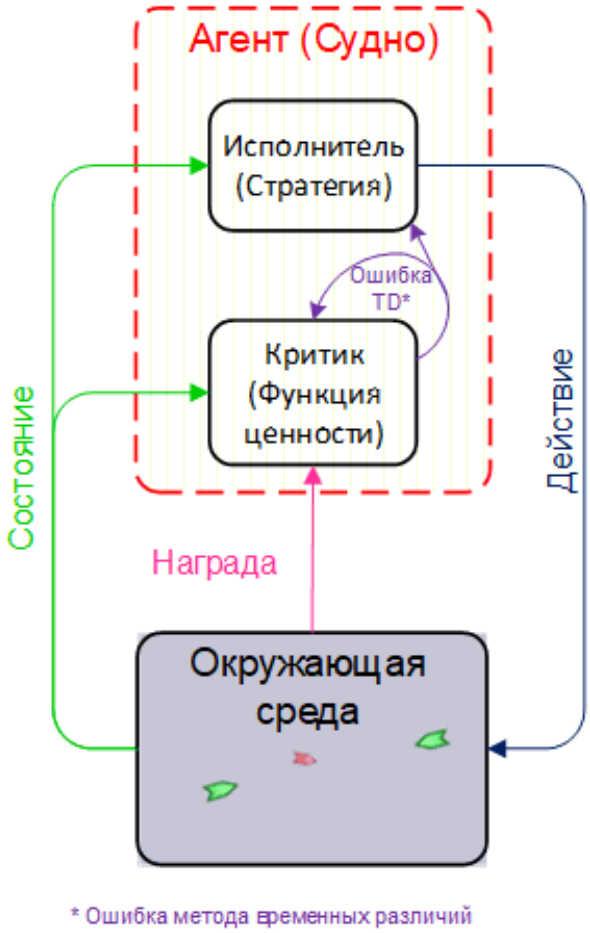

Рис. 2 – Структура алгоритма типа «исполнитель–критик»

Следует отметить, что стратегия известна как исполнитель используется для повышения ценности действий и их выбора, а функция ценности известна как критик оценивает новое состояние, давая исполнителю лучшую оценку градиента и, наконец, получая оптимальную стратегию. Эта оценка является ошибкой метода временных различий (TD), которая определяется по формуле:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{2}$$

где V – текущая функция значения, реализованная критиком.

DDPG имеет буфер воспроизведения для хранения информации обо всех переходах, который является кортежем (s_t,a_t,r_t,s_{t+1}), а затем выбирается мини-пакет для обучения исполнителя и критика (рис.3).

Кроме того, в DDPG используются четыре нейронных сети: основная сеть критика $Q(s,a|\theta^Q)$; основная сеть исполнителя $\mu(s|\theta^\mu)$; целевая сеть критика $Q'(s,a|\theta^{Q'})$; целевая сеть исполнителя $\mu'(s|\theta^{\mu'})$.

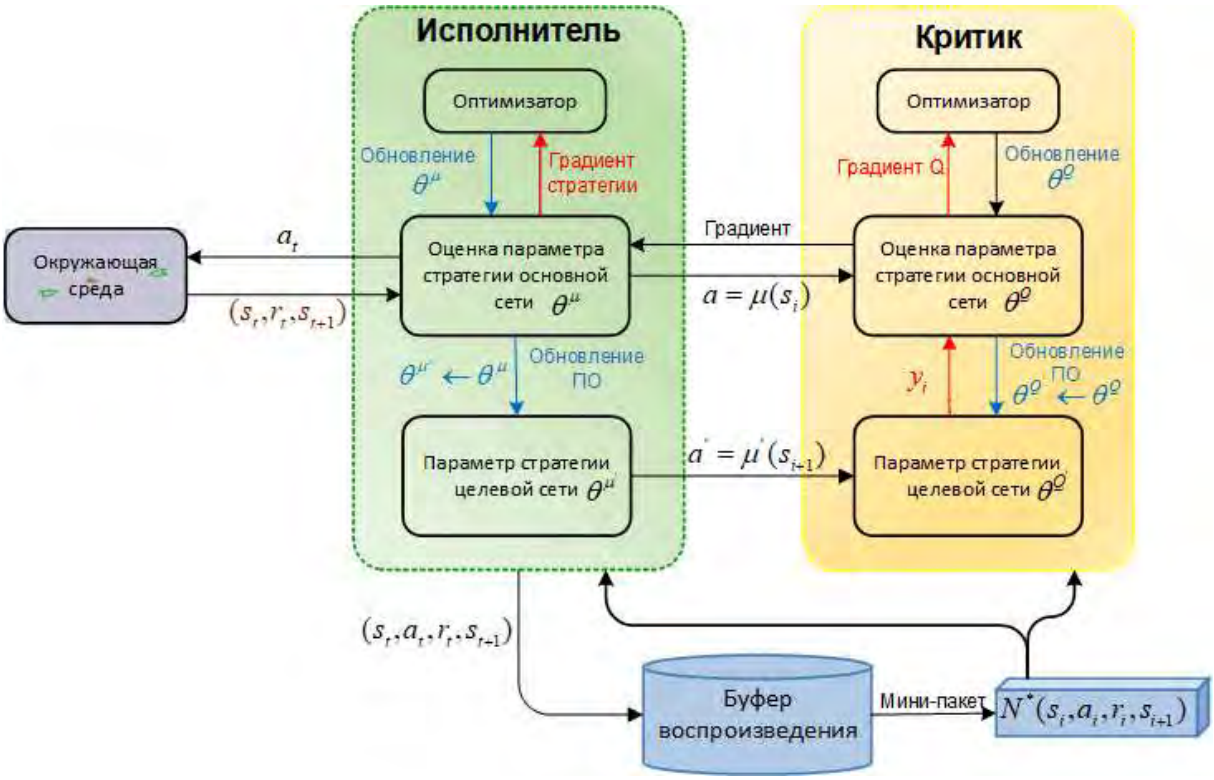

Рис. 3 – Структурная схема DDPG

Исполнитель в основной сети оптимизируется с помощью метода градиента стратегии:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s,a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i} \tag{3}$$

где s_i – переменные состояния текущего времени.

Критик в основной сети использует среднеквадратичную ошибку для определения потерь и оптимизации параметров сети методом градиентного спуска:

$$\begin{cases} L = \frac{1}{N}\sum_i (y_i - Q(\,(s_i, a_i|\theta^Q))^2 \\ y_i = r_i + \gamma Q'\big(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'}\big) \end{cases} \quad (4)$$

Целевая сеть обновляется следующим образом:

$$\begin{cases} \theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'} \\ \theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'} \end{cases} \quad (5)$$

где τ – параметр частоты обновления.

Таким образом, DDPG является одним из основных алгоритмов ГОП, который используется для решения задач с пространствами непрерывного действия, в частности, задач предотвращения столкновения и интеллектуального принятия решений при управлении безэкипажных судов.

Литература

1. **Баракат Л.** Предотвращение столкновений безэкипажных судов с использованием глубокого обучения с подкреплением / Л. Баракат // Пятая международная научно-практическая конференция ИКМ МТМТС-2019. Труды конференции. – 2019, С.102-105. (ISBN 978-5-00150-311-8).
2. **Баракат Л.** Интеллектуальная автономная система предупреждения столкновений безэкипажных судов на основе машинного обучения. [Электронный ресурс] / Л. Баракат, И. Ю. Квятковская // Информационные технологии и технологии коммуникации: современные достижения. (Астрахань, 1–5 октября 2019 года). – Астрахань: Изд-во АГТУ, год. – Режим доступа: 1 CD-диск.
3. **Guo S.** An autonomous path planning model for unmanned ships based on deep reinforcement learning/S. Guo, X. Zhang, Y. Zheng, Y. Du //Sensors. – 2020. – Т. 20. – № 2.
4. **Баракат Л.А.** Интеллектуальное принятие решений по автономному предотвращению столкновений безэкипажных судов на основе алгоритма глубокой Q-сети [Электронный ресурс] / Л.А. Баракат, И.Ю. Квятковская // 64-я Международная научная конференция Астраханского государственного технического университета, посвященная 90-летнему юбилею со дня образования Астраханского государственного технического университета. Материалы конференции. (Астрахань, 20-25 апреля 2020 года). – Астрахань: Изд-во АГТУ, год. – Режим доступа: 1 CD-диск. – № гос. регистрации 0322002778.
5. **Lillicrap T.P.** Continuous control with deep reinforcement learning / T. P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D.Wierstra // 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016.