

УДК 004.89

doi: 10.15622/rcai.2025.062

ИНТЕЛЛЕКТУАЛЬНАЯ ПОДДЕРЖКА СИТУАЦИОННЫХ РЕШЕНИЙ НА ОСНОВЕ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ¹

А.П. Шапкин (*shapkinap1999@gmail.com*)

В.В. Борисов (*vbor67@mail.ru*)

Филиал Национального исследовательского университета «МЭИ»
в г. Смоленске, Смоленск

В статье решается задача интеллектуальной поддержки ситуационных решений на основе предложенной разновидности адаптивных нечетких ситуационных моделей (АНСМ) и их обучения с подкреплением. Выполнено обоснование подхода и предложен способ оценки управляющих решений в АНСМ на основе нейросети *Deep Recurrent Q-Network*, позволяющей учитывать ретроспективу принятых ранее управляющих решений, а также результаты структурно-параметрической настройки АНСМ. Приведен пример интеллектуальной поддержки ситуационных решений при управлении энергетической мини-сетью на основе предложенной разновидности АНСМ и обучения с подкреплением.

Ключевые слова: адаптивные нечеткие ситуационные сети, интеллектуальная поддержка ситуационных решений, обучение с подкреплением.

Введение

Нечеткий ситуационный подход эффективно применим для поддержки принятия решений в условиях неполноты и неопределенности данных, с учетом выполнения требований и ограничений при переходе через промежуточные ситуации для достижения целевой ситуации [Мелихов и др., 1990]. Одной из основных особенностей этого подхода является формирование модели рассматриваемой проблемы, совмещенной с моделью управления ею в виде совокупности управляющих решений для достижения целевой ситуации [Борисов и др., 2021].

¹ Работа выполнена в рамках государственного задания (проект № FSWF-2023-0012).

В работе [Мелихов и др., 1990] предложены нечеткие ситуационные сети, состоящие из нечетких ситуаций и дуг, соответствующих переходам между ситуациями при воздействии управляющих решений на соответствующие ситуационные признаки.

В статье [Борисов и др., 2009] предложены нечеткие ситуационно-событийные модели, учитывающие неопределенность воздействия управляющих решений на ситуации за счет использования лингвистических лотерей, а также продолжительность выполнения решений.

Развитием этих моделей являются адаптивные нечеткие ситуационные модели (АНСМ), предложенные способы адаптации которых позволяют изменять их структуру и параметры в зависимости от изменения системных и внешних факторов, стратегий принятия решений и ограничений [Denisenkov et al., 2018.].

Однако определение наилучшей последовательности управляющих решений для перехода из текущей в целевую ситуацию и выбор следующей нечеткой ситуации в соответствии с заданной стратегией для АНСМ представляет собой нетривиальную задачу в условиях необходимости ее структурно-параметрической настройки.

Подход к обучению с подкреплением создает хорошие предпосылки для решения указанной выше проблемы оценки управляющих решений АНСМ для поиска наилучших последовательностей управляющих решений с максимизацией заданной функции полезности с учетом выбранной стратегии и постоянно накапливаемого опыта.

В статье предлагается способ оценки управляющих решений АНСМ на основе обучения с подкреплением, а также рассматривается пример поддержки ситуационных решений при управлении энергетической минисетью на основе предложенной разновидности АНСМ.

1. Адаптивные нечеткие ситуационные модели

Предложенная в работе [Борисов и др., 2025] разновидность АНСМ и дополненная предлагаемым способом оценки управляющих решений представляется в виде:

$$AFSM = \langle P, S, R \cup A, E \rangle,$$

где P, S, U, R – множества нечетких ситуационных признаков, ситуаций, переходов, управляющих решений, соответственно; A – способ адаптации; E – способ оценки управляющих решений.

Нечеткие ситуации представляются в виде нечетких множеств 2-го уровня на совокупности ситуационных признаков:

$$S = \{s_l\}, \quad l = 1, \dots, L,$$

$$s_l = \left\{ \left(\mu(p_i) / p_i \right) \right\}, \quad p_i \in P,$$

$$\mu(p_i) = \left\{ \left(\mu_{\mu(p_i)}(T_j^i) / T_j^i \right) \right\}, \quad i \in \{1, \dots, I\}, \quad j \in \{1, \dots, J_i\},$$

$$p_i = \langle T_i D_i \rangle, \quad i \in \{1, \dots, I\},$$

$T_i = \{T_1^i, \dots, T_{m_i}^i\}$ – терм-множество ситуационного признака p_i , m_i – число p_i ; D_i – базовое множество p_i ; термы T_j^i , $j \in \{1, \dots, J_i\}$ задаются нечеткими переменными $\langle T_j^i, D_i C_j^i \rangle$, то есть нечеткими множествами C_j^i на D_i :

$$C_j^i = \left\{ \left(\mu_{C_j^i}(d) / d \right) \right\}, \quad d \in D_i.$$

Управляющие решения представляются в виде:

$$R_k = \langle Lr_k, Tr_k, Dr_k \rangle, \quad k \in \{1, \dots, K\},$$

где $Lr_k = \{Lr_1^k, Lr_2^k, Lr_3^k\}$ – терм-множество направленности воздействия управляющего решения r_k ; $Tr_k = \{Tr_1^k, \dots, Tr_{L_k}^k\}$ – терм-множество силы его воздействия; Dr_k – шкала силы воздействия r_k , $[-1, 1]$.

Множество переходов $U = \{u_1, u_2, \dots, u_y\}$ между ситуациями сопоставляется с управляющими решениями с учетом их значимости.

Способ адаптации A предназначен для структурно-параметрической настройки АНСМ на основе обучения с подкреплением.

Способ оценки управляющих решений E предназначен для определения управляющих решений, наиболее подходящих для достижения целевой нечеткой ситуации, в соответствии с выбранной стратегией управления.

2. Обоснование подхода к оценке управляющих решений в АНСМ на основе обучения с подкреплением

Политика обучения с подкреплением для ситуационного подхода определяется как отображение ситуаций в управляющие решения в зависимости от контекста и системы предпочтений.

При решении задачи поиска наилучших управляющих решений в АНСМ с неограниченным количеством ситуаций и управляющих решений традиционные алгоритмы обучения с подкреплением, такие как *Q-learning* или *Deep Q-Learning (DQN)*, сталкиваются с принципиальными ограничениями. Так, алгоритм *Q-learning* не масштабируется на многомерные признаковые пространства из-за необходимости хранения *Q*-таблиц, а алгоритм DQN, хоть и использует нейросетевую аппроксимацию, не учитывает зависимости между смежными ситуациями [Шапкин и др., 2025].

Deep Recurrent Q-Network (DRQN) сочетает преимущества глубокого обучения с механизмом рекуррентных нейросетей, что позволяет эффективно обрабатывать частично наблюдаемые признаковое пространство и учитывать ретроспективу взаимодействия системных и внешних факторов [Zhu et al., 2018], [Moreno-Vera, 2019]. Достоинства применения *DRQN* для оценки управляющих решений в АНСМ заключаются в его способности учитывать историю переходов между нечеткими ситуациями. Помимо этого, *DRQN* позволяет учитывать новые ситуации при адаптации АНСМ, за счет кодирования особенностей АНСМ в скрытых состояниях [Hausknecht et al., 2017].

3. Оценка управляющих решений в АНСМ на основе Deep Recurrent Q-Network

3.1. Способ оценки управляющих решений для АНСМ

Необходимость в оценке управляющих решений для АНСМ возникает при определении наилучшей последовательности управляющих решений для достижения целевой нечеткой ситуации. Поэтому, перед оценкой управляющих решений определяется целевая нечеткая ситуация s_{targ} , например, в виде значений функций принадлежности по каждому признаку нечеткой ситуации.

Для получения оценки управляющих решений применяется глубокая рекуррентная нейросеть E , основанная на *DRQN*.

Предлагаемый способ оценки управляющих решений для АНСМ включает в себя перечисленные этапы.

Этап 1. Идентификация текущей нечеткой ситуации s_{cur} .

Этап 2. Определение возможных управляющих решений для нечеткой ситуации s_{cur} . Результатом этого этапа является маска m_{cur} , представляющая собой массив бинарных чисел по числу нечетких ситуаций в АНСМ. Каждой ситуации, в которую возможен переход соответствует «1», а недостижимым ситуациям – «0» [Shengyi et al., 2020].

Этап 3. Формирование массива входных данных для сети E , который включает в себя информацию о текущей нечеткой ситуации s_{cur} и целевой нечеткой ситуации s_{targ} .

Этап 4. Входной массив подается на вход нейросети E .

Этап 5. На выходе нейросети E получается множество оценок управляющих решений, размер которого зависит от количества нечетких ситуаций.

Этап 6. Для исключения оценок несуществующих переходов из текущей нечеткой ситуации s_{cur} , ко множеству оценок управляющих решений применяется маска m_{cur} .

Результатом работы способа является множество оценок управляющих решений, в котором нечеткая ситуация, в которую необходимо совершить переход с учетом выбранной стратегии имеет наивысшее значение оценки.

При необходимости оценки управляющих решений в АНСМ, моделирующих разные сложные системы одного класса, применяется нормализация данных. Информация о весах нечетких переходов не подается на вход явным образом, а кодируется через функцию вознаграждения.

3.2. Обучение нейросети, основанной на Deep Recurrent Q-Network

Ключевое преимущество применения *DRQN* для оценки управляющих решений заключается в его способности интегрировать особенности заданной стратегии управления непосредственно в процесс обучения. Это достигается через соответствующую параметризацию функции вознаграждения $R(s, a, s')$, которая кодирует критерии оптимальности пути (минимизацию количества переходов, избегание нежелательных ситуаций). В результате обученная *Q*-функция $Q(s, a; \theta)$ позволяет получить оценку действий, которые отражают долгосрочные последствия решений в контексте выбранной стратегии управления [AlMahamid et al., 2022]. Благодаря этому свойству, в каждой ситуации s , выбирается наилучшее управляющее действие a , позволяющее достичь целевую ситуацию s_T .

Дополнительно для обеспечения устойчивости обучения используется целевая сеть – нейросеть с идентичной структурой основной нейросети *DRQN*, но отличающейся частотой обновления параметров.

Периодическое обновление параметров целевой нейросети θ^- позволяет стабилизировать обучение, уменьшая корреляцию между текущими *Q*-значениями и целевыми. Интервал обновления подбирается эмпирически: слишком частые обновления приводят к нестабильности, а слишком редкие – к замедлению обучения.

В контексте ситуационного подхода политика, полученная в результате обучения, представляет собой систему предпочтений, отражающую стратегию управления. Формально политика $\pi(a|s)$ определяет вероятность выбора действия a в ситуации s . В случае ϵ -жадной стратегии с вероятностью $(1 - \epsilon)$ выбирается действие с максимальным *Q*-значением, а с вероятностью ϵ – исследует альтернативные варианты.

Процесс обучения начинается с инициализации весов *DRQN*. Входом сети является текущая ситуация s_t , представленная в виде множества признаков, а выходом – *Q*-значения для всех возможных управляющих решений $a \in A(s_t)$.

При последовательном переходе из одной нечеткой ситуации в другую формируется награда r_t за каждый такой переход. Награда определяется в соответствии со стратегией управления: например, если стратегия требует минимизации суммарного веса переходов, то награда может быть определена как отрицательный вес нечеткого перехода. В случае, если переход в определенную ситуацию нежелателен при выбранной стратегии, то может быть добавлен значительный штраф при переходе в эту нечеткую ситуацию.

Рекуррентный слой нейросети *DRQN* позволяет учитывать не только текущую ситуацию, но и контекст предыдущих переходов, что особенно важно для работы АНСМ. На каждом шаге скрытое состояние h_t обновляется в соответствии с выбранным действием и наблюдением новой ситуации, формируя представление, которое влияет на принятие решений.

Обучение происходит путем минимизации функции потерь:

$$L(\theta) = E_{(s,a,r,s') \sim D} \left(\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right),$$

где θ – параметры обучаемой нейросети, θ^- – параметры целевой нейросети, обновляемые периодически, γ – коэффициент дисконтирования, а D – база воспроизведения опыта.

Во время обучения из D случайным образом выбирается последовательность управляющих решений, затем для каждого из них, используя целевую нейросеть, вычисляется целевое Q -значение. Используя градиент ошибки для выбранной последовательности управляющих решений обновляются параметры основной нейросети.

4. Пример поддержки ситуационных решений с использованием АНСМ на основе обучения с подкреплением

Рассмотрим пример оценивания управляющих решений на примере АНСМ, предназначенной для моделирования процесса управления энергетической мини-сетью [Борисов и др., 2025]. Определим множество признаков $P = \{p_1, p_2, p_3, p_4\}$, где p_1 – уровень заряда аккумуляторов, p_2 – производство энергии солнечными панелями, p_3 – потребление энергии, p_4 – количество топлива резервного генератора.

Определим терм-множества ситуационных признаков и соответствующие им базовые множества:

$$\begin{aligned} T_{p1} &= \{T_1^{p1} = \text{«Низкий»}, \bar{T}^{p1} = \text{«Средний»}, \bar{T}^{p1} = \text{«Высокий»}\}, D_1 = [0, 40]. \\ T_{p2} &= \{T_1^{p2} = \text{«Низкое»}, \bar{T}^{p2} = \text{«Среднее»}, \bar{T}^{p2} = \text{«Высокое»}\}, D_2 = [0, 100]. \\ T_{p3} &= \{T_1^{p3} = \text{«Низкое»}, \bar{T}^{p3} = \text{«Среднее»}, \bar{T}^{p3} = \text{«Высокое»}\}, D_3 = [0, 100]. \\ T_{p4} &= \{T_1^{p4} = \text{«Низкое»}, \bar{T}^{p4} = \text{«Среднее»}, \bar{T}^{p4} = \text{«Высокое»}\}, D_4 = [0, 700]. \end{aligned}$$

Нечеткие ситуации задаются с учетом соответствия степеней принадлежности значениям признаков. Структура АНСМ для рассматриваемого примера представлена на рис. 1.

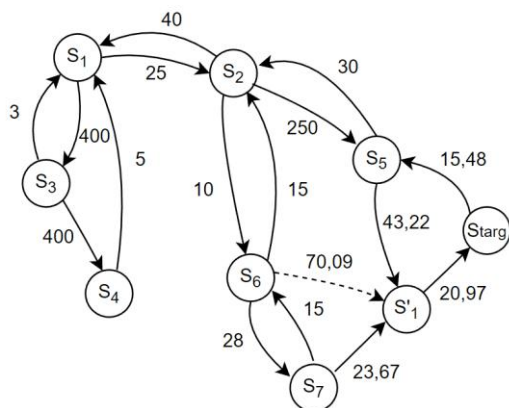


Рис. 1. АНСМ для моделирования процесса управления энергетической мини-сетью

Пусть энергетическая мини-сеть находится в нечеткой ситуации s_6 а до этого – в нечеткой ситуации s_2 . Стоит задача ее перехода в нечеткую ситуацию s_{targ} . Выбрана стратегия минимизации стоимости соответствующих управляющих решений. Однако переход через некоторые ситуации является нежелательным. Например, непосредственный переход в нечеткую ситуацию s'_1 является более «дорогим», чем переход в эту ситуацию через нечеткую ситуацию s_7 , однако, ситуация s_7 характеризуется повышенными эксплуатационными рисками.

Массив исходных данных для нейросети *DRQN* формируется из описания целевой нечеткой ситуации, которое формируется по значениям функций принадлежности признаков нечеткой ситуации:

$$s_{targ} = (0,9; 0, 2; 0, 4; \dots; 0,2)$$

и описания текущей нечеткой ситуации:

$$s_6 = (0, 2; 0, 4; 0, 7; \dots; 0,9)$$

После обработки массива исходных данных, нейросеть формирует значения оценок для всех управляющих решений АНСМ, к которому применяется маска доступных нечетких переходов из нечеткой ситуации s_6

$$M_{s_6} = (0; 1; 0; 0; 0; 0; 1; 0)$$

Массив итоговых значений оценки выглядит следующим образом:

$$Q(s_6) = (0; 3, 4; 0; 0; 0; 0; 12, 7; 15)8; 0$$

Таким образом, оценка управляющего решения, приводящего к переходу в нечеткое состояние s'_1 выше, чем у управляющего решения, приводящего к переходу в нечеткое состояние s_7 . Такая оценка соответствует выбранной стратегии управления.

Заключение

Применение обучения с подкреплением для интеллектуальной поддержки ситуационных решений является перспективным направлением развития теории и практики принятия решений.

В статье поставлена и решается задача интеллектуальной поддержки ситуационных решений на основе предложенной разновидности АНСМ и их обучения с подкреплением.

Выполнено обоснование подхода и предложен способ оценки управляющих решений в АНСМ на основе нейросети *DRQN*, позволяющего учитывать ретроспективу принятых ранее управляющих решений и взаимодействия системных и внешних факторов, а также результаты динамичной структурно-параметрической настройки АНСМ, обеспечивать выбор наилучших управляющих решений в соответствии с заданной стратегией.

Приведен пример интеллектуальной поддержки ситуационных решений при управлении энергетической мини-сетью на основе предложенной разновидности АНСМ и обучения с подкреплением.

Дальнейшие исследования направлены на расширение класса решаемых задач с использованием предложенной разновидности АНСМ, включая многокритериальную оптимизацию и интеллектуальную поддержку ситуационных решений в мультиагентных системах.

Список литературы

- [Борисов и др., 2009] Борисов В.В., Зернов М.М. Реализация ситуационного подхода на основе нечеткой иерархической ситуационно-событийной сети // Искусственный интеллект и принятие решений. – 2009. – № 1. – С. 17-30.
- [Борисов и др., 2021] Борисов В.В., Авраменко Д.Ю. Нечеткое ситуационное управление сложными системами на основе их композиционного гибридного моделирования // Системы управления, связи и безопасности. – 2021. – № 3. – С. 207-237.
- [Борисов и др., 2025] Борисов В.В., Шапкин А.П. Нечеткие ситуационные сети с адаптацией на основе обучения с подкреплением // Вестник Московского энергетического института. – 2025. – № 3. – С. 135-143.
- [Мелихов и др., 1990] Мелихов А.Н., Берштейн Л.С., Коровин С.Я. Ситуационные советующие системы с нечеткой логикой. – М.: Наука, 1990. – 272 с.
- [Шапкин и др., 2025] Шапкин А.П., Борисов В.В. Способы внедрения обучения с подкреплением в адаптивные нечеткие ситуационные сети // XXII Международная научно-техническая конференция студентов и аспирантов (Смоленск, 23–24 апреля 2025 г.): Труды конференции. В 2-х т. Т. 1. – Смоленск: Универсум, 2025. – С. 470-474.

- [AlMahamid et al., 2022] AlMahamid F., Grolinger K. Reinforcement Learning Algorithms: An Overview and Classification [Электронный ресурс] // arXiv.org. – 2022. URL: <https://arxiv.org/abs/2209.14940> (дата обращения: 16.03.2024).
- [Denisenkov et al., 2018] Denisenkov M.A., Borisov V.V. Modeling the behavior of intelligent agents based on adaptive fuzzy situational networks // Proc. 3th Russian-Pacific Conference on Computer Technology and Application (RPC). Vladivostok, Russia, 18-25 Aug. 2018. – DOI: 10.1109/RPC.2018.8482217.
- [Hausknecht et al., 2017] Hausknecht M., Stone P. Deep Recurrent Q-Learning for Partially Observable MDPs // arXiv.org. – 2017. – URL: <https://arxiv.org/abs/1507.06527> (дата обращения: 16.03.2024).
- [Moreno-Vera, 2019] Moreno-Vera F. Performing Deep Recurrent Double Q-Learning for Atari Games [Электронный ресурс] // arXiv.org. – 2019. – URL: <https://arxiv.org/abs/1908.06040> (дата обращения: 16.03.2024).
- [Shengyi et al., 2020] Shengyi H., Onta ñón S. A Closer Look at Invalid Action Masking in Policy Gradient Algorithms [Электронный ресурс] // arXiv.org. – 2020. – URL: <https://arxiv.org/abs/2006.14171> (дата обращения: 16.03.2024).
- [Zhu et al., 2018] Zhu P., Li X., Poupart P., Miao G. On Improving Deep Reinforcement Learning for POMDPs [Электронный ресурс] // arXiv.org. – 2018. – URL: <https://arxiv.org/abs/1804.06309> (дата обращения: 16.03.2024).