π-Economy. 2023, Том 16, № 5. С. 38–50.
π-Economy. 2023, Vol. 16, No. 5. Pp. 38–50.

Цифровые технологии и инновации в интеллектуальной экономике

Digital technologies and innovations in intelligent economy

Научная статья
УДК 519.857.3
DOI: https://doi.org/10.18721/JE.16503

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ КАК ТЕХНОЛОГИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ РЕШЕНИЯ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЗАДАЧ: ОЦЕНКА ПРОИЗВОДИТЕЛЬНОСТИ АЛГОРИТМОВ

Е.В. Орлова ✉

Уфимский университет науки и технологий,
г. Уфа, Российская Федерация

✉ ekorl@mail.ru

Аннотация. Обучение с подкреплением, с одной стороны, является классом методов машинного обучения и искусственного интеллекта, а с другой стороны – областью знания, в которой исследуется сама прикладная задача, а также методы ее решения. Одной из таких задач является задача управления социальными и экономическими системами, конструирование оптимального управления с учетом свойств самой системы, таких как разнообразие шкал характеристик изучаемых объектов, неоднородность выборок данных, неполнота и пропуски в данных, стохастичность данных, их мультиколлинеарность и гетероскедастичность. Методы обучения с подкреплением не чувствительны к этим особенностям и могут быть использованы с более высокой эффективностью в различных приложениях экономики, финансов и бизнеса. Обучение с подкреплением ближе всего к способам обучения людей, а решения возникающих проблем можно находить в области биологических самообучающихся систем на основе принципа проб и ошибок. Методы обучения с подкреплением представляет собой вычислительный подход к обучению, в ходе которого субъект управления (агент) обучается в процессе взаимодействия со сложным, динамическим, чаще стохастическим, объектом управления (средой) социально-экономической природы с целью максимизации общего вознаграждения. В процессе моделирования возникает проблема выбора таких алгоритмов обучения, которые адекватно отражают стохастическую динамику моделируемого объекта, и имеют высокую производительность. Бизнес-метрики и метрики качества, приемлемые для оценки качества методов обучения с учителем и без учителя в машинном обучении не вполне пригодны для оценки эффективности методов обучения с подкреплением, так как отсутствуют эмпирические данные для оценки. В работе предложены ряд показателей качества обучения для сгенерированных на основе методов обучения с подкреплением управленческих решений. На примере задачи управления человеческим капиталом предприятия произведено сравнение алгоритмов обучения – DQN, DDQN, SARSA, PRO для конструирования оптимальных траекторий профессионального развития работников предприятия. Осуществлена оценка предложенных показателей качества для всей группы методов обучения и выбран один из алгоритмов с наивысшей производительностью.

Ключевые слова: социально-экономические системы, индивидуальные траектории развития работников предприятия, искусственный интеллект, машинное обучение, обучение с подкреплением, качество алгоритмов обучения

Для цитирования: Орлова Е.В. (2023) Обучение с подкреплением как технология искусственного интеллекта для решения социально-экономических задач: оценка производительности алгоритмов. π-Economy, 16 (5), 38–50. DOI: https://doi.org/10.18721/JE.16503

Research article
DOI: https://doi.org/10.18721/JE.16503

REINFORCEMENT LEARNING AS AN ARTIFICIAL INTELLIGENCE TECHNOLOGY TO SOLVE SOCIO-ECONOMIC PROBLEMS: ALGORITHMS PERFORMANCE ASSESSMENT

E.V. Orlova ✉

Ufa University of Science and Technology, Ufa, Russian Federation

✉ ekorl@mail.ru

Abstract. Reinforcement learning is a class of machine learning and artificial intelligence methods, a field for the applied problem studied, as well as methods for solving it. One of these problems is management in social and economic systems, designing optimal control taking into account the systems' properties such as variety of characteristics scales, heterogeneity of data samples, incompleteness and gaps in the data, data stochasticity, their multicollinearity and heteroscedasticity. Reinforcement learning methods are not sensitive to these features and can be used with higher efficiency in various applications of economics, finance and business. Reinforcement learning is closest to the way humans learn, and solutions to emerging problems can be found in the field of biological self-learning systems based on the principle of trial and error. Reinforcement learning methods are a computational approach to learning, when the control subject (agent) learns under interaction with a complex, dynamic, often stochastic, control object (environment) like a socio-economic system in order to maximize the total reward. In the process of modeling, the problem of choosing such learning algorithms that adequately reflect the stochastic dynamics of the modeled object and have high performance is very important. Business and quality metrics that are appropriate for assessing the quality of supervised and unsupervised learning methods in machine learning are not entirely suitable for evaluating the effectiveness of reinforcement learning methods, since there is no empirical data for evaluation. The paper proposes a number of quality indicators of training for managerial decisions generated on the basis of training methods with reinforcement learning. We use an example for the corporate human resources management. A comparison for learning algorithms such as DQN, DDQN, SARSA, PRO for designing optimal trajectories for the proficiency training of the personnel is made. An assessment of the proposed quality indicators for the entire group of learning methods is carried out and one of the algorithms with the highest performance is selected.

Keywords: socio-economic systems, individual trajectories for employees' development, artificial intelligence, machine learning, reinforcement learning, quality of learning algorithms

Citation: Orlova E.V. (2023) Reinforcement Learning as an Artificial Intelligence Technology to Solve Socio-Economic Problems: Algorithms Performance Assessment. π-Economy, 16 (5), 38–50. DOI: https://doi.org/10.18721/JE.16503

Введение

Актуальность

Обучение с подкреплением (reinforcement learning, RL) является одним из самых активно изучаемых сфер искусственного интеллекта, представляет собой вычислительный подход к обучению в рамках методологии машинного обучения, в ходе которого субъект управления (агент) обучается, взаимодействуя со сложным, динамическим, часто стохастическим объектом управления (средой) с целью максимизации общего вознаграждения.

Для формализации задач последовательного принятия решений, когда последствия действий не детерминированы, используется марковский процесс принятия решений (markov decision process, MDP). Модель MDP определяет стохастическую динамику описываемой системы, а также полезность, связанную с эволюцией и со стратегий принятия решений. Этот класс задач разрешается с помощью алгоритмов RL, на основе которых агенты учатся, используя метод проб и

ошибок. RL как класс методов машинного обучения и искусственного интеллекта, использует теорию оптимального управления и понятие марковского процесса принятия решений. RL появился еще в 1950-х годах в контексте динамического программирования и квазилейных уравнений Беллмана. Задача RL может быть представлена в виде системы, состоящей из агента (управляемой подсистемы) и среды (подсистемы управления). Системы RL реализуют цикл управления с обратной связью, где агент и среда обмениваются сигналами, при этом агент стремится максимизировать целевую функцию. Обе стороны взаимодействуют непрерывно: агент выбирает действия, а вреда реагирует на эти действия и предлагает агенту новые ситуации. Среды генерируют вознаграждения – числовые значения, которые агент стремится со временем максимизировать посредством выбора действий.

Одной из проблем, возникающих при подборе необходимых алгоритмов обучения, является оценка их качества, производительности. Метрики, используемые в машинном обучении, не вполне пригодны для RL, так как отсутствуют обучающие выборки, то есть эмпирические данные [1, 2]. Сопоставление фактических и модельных результатов с помощью метрик машинного обучения, должно быть заменено иными показателями качества.

Цель исследования

Целью работы является проведение анализа показателей качества методов и алгоритмов RL и выбор приемлемых метрик при исследовании организационных, в том числе социально-экономических систем, в контуре которых присутствует человек как лицо, принимающее решение, и привносящее дополнительную неопределенность в систему. Данная цель декомпозирована на совокупность задач:

1) провести анализ решаемых задач и используемых алгоритмов RL в области экономики, финансов и бизнеса;

2) показать особенности наиболее часто используемых алгоритмов RL, учитывающих свойства исследуемого класса систем;

3) предложить показатели производительности алгоритмов RL применительно к исследованию социально-экономических систем;

4) на примере задачи управления человеческим капиталом предприятия провести сравнительную оценку производительности алгоритм RL.

Литературный обзор

Приложения систем обучения с подкреплением в организационных и социально-экономических системах разнообразны, связаны с задачами оптимизации (динамического программирования) процессов и систем. Новейшие исследования в области управления таким классом систем представлены ниже:

– в области промышленности, менеджмента RL используется по всему спектру задач управления ресурсами [3, 4], разработки принципов календарного планирования производства [5], разработке планов пополнения запасов, устанавливающих момент и объем пополнения запасов, разработке логистических маршрутов и цепочек поставок [6, 7];

– в робототехнике RL имеет множество приложений, включая улучшение движения, разработку автономных транспортных средств [8, 9].

– RL улучшают управление движением на дорогах и используется в алгоритмах управления умными городами [10];

– множество приложений RL в области здравоохранения используются для формирования схем расчета и дозирования лекарственных средств [11];

– при конструировании систем образования и электронного обучения, которые могут повысить свою эффективность за счет подбора учебных программ на базе RL [12].

В табл. 1 приведены приложения RL в области экономики, финансов и бизнеса, сгруппированные по общности используемых методов и алгоритмов обучения.

Таблица 1. Анализ решаемых задач и используемых алгоритмов RL в области экономики, финансов и бизнеса
Table 1. Analysis of the problems and the RL algorithms used in the field of economics, finance and business

Решаемая задача	Алгоритм	Ссылка на источник
Биржевая торговля Разработка стратегии принятия решений	DDPG (Deteministic Policy Gradirnt) Adaptive DDPG DQN (Deep Q-networks) RCNN (Recurrent Convolutional Neural Networks)	[13–16]
Управление инвестиционным портфелем (в том числе на рынках криптовалют) Задача алгоритмической торговли Оптимизация портфеля	DDPG Model-less CNN Model-free Model-based	[17–19]
Онлайн торговля и ритейл Разработка рекомендательных систем Разработка алгоритмов динамического ценообразования (в реальном времени)	Actor-critic method SS-RTB method (Sponsored Search Real-Time Bidding) (аукцион, построенный в реальном времени с привлечением спонсоров) DDPG DQN	[20–22]

В задачах управления человеческими ресурсами на уровне предприятия методы RL до сих пор не использовались. Предложенный автором методологический подход [23, 24] к управлению человеческим капиталом на основе индивидуализации управленческих решений, связанных с развитием потенциала работников, использует методы RL для выработки оптимальных стратегий по управлению человеческим капиталом предприятия.

Методы и материалы

Теоретико-методологической базой исследования служат труды зарубежных и отечественных исследователей, связанных с машинным обучением, обучением с подкреплением, изучением человеческого капитала и его влиянию на эффективность производственно-экономических систем. Используются общенаучные методы системного анализа, оптимального управления, теории принятия решений, стратегического управления, методы математического и компьютерного моделирования.

Описание алгоритмов обучения агента

Алгоритм обучения с подкреплением представляет собой последовательность адаптированных процедур, соответствующих динамическому изменению состояния системы. Таким образом, стратегия управления, разработанная на основе метода обучения с подкреплением, будет динамически меняться с течением времени по мере накопления наблюдений.

При построении алгоритмов RL важное значение имеет представление среды, то есть объекта управления. Различают алгоритмы, основанные на модели среды (model-based algorithm) и не использующие модели среды (model-free algorithm). Модель описывает поведение среды, предсказывает следующее ее состояние и вознаграждение для данных состояния и действия. Если модель известна, то для взаимодействия со средой в качестве выработки рекомендации будущих действий можно использовать алгоритмы планирования. Например, в средах с дискретными действиями потенциальные траектории можно смоделировать, применяя поиск по дереву методом Монте-Карло. Модель среды может быть либо задана заранее, либо обучена посредством взаимо-

действия с ней. Если среда сложная, динамичная, плохоформализуемая, то ее в процессе обучения можно аппроксимировать глубокой нейронной сетью.

Среда, представленная в виде марковского процесса принятия решения и гибкие алгоритмы RL реализуют способ последовательного принятия решений, когда выбранное действие влияет на следующие состояния объекта управления и результаты воздействия решений. Оптимальная стратегия достижения поставленной цели вырабатывается посредством взаимодействия объекта и субъекта управления.

В работе используются алгоритмы следующих классов – алгоритмы, основанные на полезности, алгоритмы, основанные на стратегии и комбинированные алгоритмы.

Алгоритмы, основанные на полезности (DQN-алгоритмы)

С использованием данных алгоритмов агент настраивает либо $V^{\pi}(s)$ (функцию ценности состояния s при стратегии π) либо $Q^{\pi}(s,a)$ (функция ценности действия a в состоянии s при стратегии π). Настроенная функция полезности используется для оценки пар (s,a) и порождения стратегии агента.

Алгоритм DQN (Deep Q-Networks) как алгоритм обучения глубоких нейронный сетей, основан на полезностях и методе временных различий, который аппроксимирует Q-функцию. Настроенная Q-функция используется агентом для выбора действий. Применяется для дискретного пространства действий.

Q-обучение основано на ценности действия, это алгоритм с разделенной стратегией. Для обновления текущей стратегии используется опыт, накопленный при реализации разных стратегий (не только текущей). В Q-обучении две стратегии: целевая (постоянно улучшается) и поведенческая ε – жадная, используемая для взаимодействия со средой. Агент на основе сведений о состоянии объекта управления s_t и полученном из среды вознаграждения r_t за действие a_t, переведшее состояние объекта в следующее состояние, вычисляет значение функции $Q(s,a)$ оценивающее ценность действия a_t в состоянии s_t. Настройка Q-функции осуществляется с помощью метода TD-обучения (метода временных различий), значение функции обновляется на накопленные дисконтированные будущие вознаграждения и определяет принцип оптимальности Беллмана:

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

где α – скорость обучения функции ценности (при $\alpha < 1$ осуществляется приближение старого состояния к новому, при $\alpha = 1$ осуществляется замена старого состояния новым); r_{t+1} – награды, полученные из среды, за действия a_t из состояния s_t; γ – коэффициент дисконтирования; $\max_a Q(s_{t+1},a)$ – максимальное ожидаемое значение из состояния s_{t+1} (новое значение); $Q(s_t,a_t)$ – предыдущая оценка Q-значения (старое значение).

Полученные Q-значения используются для обучения агента и для определения следующего действия. Для этого используется нейронная сеть (сеть полезности, value networks), которая оценивает Q-значения пар (s,a) и выбирает действия с максимальным Q-значением (максимальной полезностью):

$$Q^{\pi}_{target}(s,a) = r + \gamma \max_{a'_i} Q^{\pi_\theta}(s',a'). \quad (2)$$

У алгоритма SARSA идея такая же за исключением того, что алгоритм DQN рассчитывает функцию полезности за несколько временных шагов, то есть буферизует опыт. Алгоритм DQN реализует вычисления по множеству пакетов данных, это увеличивает вычислительную нагрузку на вычислительную систему, но при этом может значительно ускорить обучение. Если Q^{π_θ} со-

держит ошибки, то вычисленное максимальное значения $\max Q^{\pi_\theta}$ будет смещенным вправо и полученные итоговые Q-значения окажутся завышенными.

Для предотвращения ошибок в максимизации в Q-обучении и повышения устойчивости обучения используется алгоритм двойного Q-обучения. В алгоритме DQN для выбора действия и получения оценки Q-функции используется одна и та же нейронная сеть. В алгоритме двойной DQN (double DQN, DDQN) применяется две нейронные сети. Первая сеть обучаемая θ-сеть используется для выбора действия a, вторая сеть прогнозная φ-сеть используется для расчета Q-значения для пар (s, a), то есть для оценки этого действия a. Эти две сети обучаются на пересекающихся прецедентах. Применение прогнозной сети позволяет сделать обучение более устойчивым благодаря снижению скорости изменения целевого Q-значения Q^{π}_{target} :

$$Q^{\pi}_{target}(s, a) = r + \gamma Q^{\pi_\theta}\left(s_i', \max_{a_i'} Q^{\pi_\theta}(s_i', a_i')\right). \quad (3)$$

Применение двух сетей в данном алгоритме могут замедлять процесс обучения, если параметры θ и φ являются очень близкими значениями, в этом случае обучение может быть неустойчивым, но если φ меняется слишком медленно, процесс обучения может замедлится. Для поиска разумного соотношения между устойчивостью и скоростью обучения нужно настраивать гиперпараметр – частоту F, управляющий скоростью изменения φ.

Алгоритмы, основанные на стратегии (REINFORCE)

Данный класс алгоритмов предназначен для настраивания стратегии π. Хорошие состояния должны порождать действия, обеспечивающие траектории τ, которые максимизируют целевую функцию агента $J(\tau)$ как сумму дисконтированных вознаграждений, усредненную по нескольким эпизодам:

$$J(\tau) = E_\tau\left[R(\tau)\right] = E_\tau\left[\sum_{t=0}^{T}\gamma^t r_t\right], \quad (4)$$

где $J(\tau)$ – сумма дисконтированных вознаграждений за временные шаги $t = 0, ..., T$, целевая функция $J(\tau)$ – это отдача, усредненная по нескольким эпизодам (повторным прогонам).

Агенту нужно действовать в среде, а действия, который будут оптимальными в данный момент, зависят от состояния. Функция стратегии π принимает на входе состояние, а на выходе выдает действие $a \sim \pi(s)$. То есть агент может принимать эффективные решения в разных ситуациях.

Алгоритм строит параметризированную стратегию, которая получает вероятности действия по состояниям среды. Агент использует эту стратегию, чтобы действовать в среде. Представляют собой алгоритмы градиента стратегии, в которой для максимизации целевой функции ценности состояния используется ее градиент, который применяется для корректировки весов нейросети обратного распространения ошибки. Способ вычисления ошибки (потерь) основан на теореме о градиенте стратегии. В алгоритме не используется прошлый опыт: весь опыт, накопленный при следовании текущей стратегии, отбрасывается после перехода к другой стратегии. Оптимизация выполняется на основе пакета данных, сформированных на основании текущей реализованной и сохраненной траектории. Пакет данных включает все переходы. Процедура оптимизации заключается в обновлении весов нейросети в результате обратного распространения ошибки, полученной на обучающем пакете данных.

Комбинированные алгоритмы (PRO)

Алгоритм PRO (Proximal Policy Optimization, проксимальная оптимизация стратегии) представляет собой метод градиента стратегии с преобразованием целевой функции. Он комбинирует

алгоритм REINFORCE и алгоритм актора-критика. Существует два варианта выбора функции потерь: 1 – на основе расстояния Кульбака-Лейблера (с адаптивной штрафной функцией), 2 – на основе усеченной целевой функции. Применение преобразования целевой функции для стратегии может повысить устойчивость и эффективность выборок в процессе обучения за счет меньшей затратности вычислительных ресурсов и более высокой производительности. Однако есть и недостатки этого алгоритмы, например, низкая чувствительность к гиперпараметру ε, что дает близкие значения производительности при разных значениях этого параметра.

Результаты и обсуждение

Показатели производительности алгоритмов RL

Обучение с подкреплением представляет собой машинное обучение, в которой нет обучающей выборки, то метрики качества алгоритмов обучения отличаются от метрик качества алгоритмов классификации или регрессии. Кроме бизнес метрик (KPI и других) к метрикам качества классификации относят: Accuracy, Precision, Recall. В случае разбиения на два класса строят матрицы смежности (confusion matrix) с выделением различных исходов – True positives (TP), False positives (FP), True negatives (TN), False negatives (FN). ROC-кривая (Receiver operator characteristic) является часто используемой метрикой для представления результатов бинарной классификации [25]. В случае обработки данных с высокой степенью ассиметричности кривая PR (precision-recall) дает более информативную картину точности алгоритма.

К метрикам качества регрессии относят MAE/MAD (Mean Absolute Error, Mean Absolute Deviation) – средний модуль ошибки; MSE/MSD (Mean Squared Error / Deviation) – среднеквадратическая ошибка; RMSE (Root Mean Squared Error) – квадратный корень из метрики MSE, выражается в тех же единицах, что и изучаемый показатель; MAPE (Mean Absolute Percentage Error) – ошибка, выраженная в процентах от самой величины.

Агент обучения с подкреплением может обучаться, взаимодействуя с реальной системой, или с ее имитационной моделью (или ее частью), или с обоими источниками сразу. Имитационная модель реальной системы представляет собой среду, которую агент может исследовать без ограничений. В большинстве современных приложений RL обучение производится на имитированном опыте, поэтому можно сгенерировать неограниченное количество данных с меньшими затратами, чем получение реальной информации о системе. Проблема состоит в том, что имитация реальной системы ограничено достоверна. Особенно это касается организационных систем, то есть сред, динамика которых зависит от поведения людей – производственно-экономических систем, образовательных систем, систем здравоохранении, транспортных систем, систем государственного управления.

Эффективность агента и, соответственно, производительность алгоритма RL можно описать двумя абстрактными эффективности политики и эффективности обучения. Эффективность политики отражает, насколько хорошо алгоритм решает поставленную задачу. Эффективность обучения измеряет, насколько быстро можно обучить агента формированию оптимальной политики. В задаче с конечным горизонтом классическим способом измерения эффективности политики представляется суммарным вознаграждением. В задачах с бесконечным горизонтом используется дисконтированные вознаграждения.

Для задач управления в сложных дискретных средах с дискретным пространством состояний оценка производительности алгоритмов RL может быть основана на следующих показателях, первый относится к эффективности обучения, остальные – к эффективности политики:

1. Время обучения агента. Алгоритм, у которого время обучения минимально, является более эффективным. Однако оптимальность является асимптотическим результатом, поэтому скорость сходимости к оптимальности иногда более приемлемый показатель.

2. Значения средних вознаграждений. Они рассчитываются как скользящие средние по ряду контрольных точек по полным вознаграждениям, усредненных по результатам нескольких сес-

сий в испытаниях. Значения гиперпараметров в данном случае фиксируются. Алгоритм, имеющий максимальный показатель, является более эффективным.

3. Оценки чувствительности функции потерь (или средних вознаграждений) к изменению гиперпараметров. Определяются значения гиперпараметров, обеспечивающих максимальное среднее вознаграждение.

4. Оценка изменения политики при изменении начального состояния среды. Эффективность надежной политики должно постепенно ухудшаться при наличии неблагоприятных факторов. Политики, у которых есть необследованные области среды, близкие к оптимальной траектории, не являются надежными, так как небольшие отклонения могут привести к состоянию с неопределенной политикой.

5. В обучении онлайн применяется метрика – ошибка алгоритма предсказателя (regret). Измеряется как разница между вознаграждением, если агент вел себя оптимально в ретроспективе, и фактическим вознаграждением, полученным за все время обучения. Данная метрика используется как математический инструмент при формировании алгоритмов на основе политики, где политика оптимизируется, чтобы ограничить ошибку алгоритма предсказателя. Метрика важна концептуально, так как лицо, принимающее решение, опирается на имеющийся у него опыт для вывода причинно-следственных связей и ставит под сомнение полезность долгосрочных решений [26].

6. Статистические показатели [27] представляют собой способы количественной оценки устойчивости политики. Стандартное отклонение полученного вознаграждения по ряду испытаний является показателем устойчивости политики к изменению наблюдений. Для анализа различий между парными выборками, то есть тестируются гипотезы о статистически значимом различии между выборками на основе следующих тестов: параметрический t-тест Стьюдента (для независимых выборок), непараметрический теста Вилкоксона [28, 29].

Численные эксперименты

В рамках предложенного подхода [23, 24] разработана управленческая схема формирования индивидуальных траекторий развития работников на основе методов обучения с подкреплением с учетом текущего уровня человеческого капитала. Стратегия, которую вырабатывает алгоритм, определяет, как агент выбирает действие в данном состоянии, то есть какие методы управленческого воздействия будут наиболее приемлемыми для данного работника в данный момент времени. Выбирается такое решение (действие), которое максимизирует полное вознаграждение, которое может быть достигнуто из данного состояния, а не действие, которое приносит наибольший немедленный эффект (вознаграждение). Преследуется долгосрочная цель предприятия по улучшению качества человеческого капитала, росту производительности ресурсов и эффективности функционирования предприятия.

Работник представлен как среда, в которой заданы ограничения достижимости цели (конечного состояния среды), начальное состояние, функции переходов состояний, награды за эти переходы. Возможны переходы вправо и вниз, что соответствует перемещению работника на следующий уровень по одному из показателей при реализации управленческих решений в соответствующей группе. Действия дискретные и отражают одно из управленческих решений, предназначенных для данной категории работников. Под решением понимается реализация определенного мероприятия (например, проведение повышения квалификации), направленного на рост человеческого капитала и повышение производительности труда.

Проведена серия экспериментов для нескольких работников с разными значениями ЧК. Для каждого работника (среды) были реализованы различные алгоритмы обучения агента (предприятия), проводилась оценка сходимости алгоритма, и достижения наибольшего вознаграждения. Было проведено несколько имитационных экспериментов. А результате оценены средние суммы вознаграждения, полученные за каждый из 200 эпизодов моделирования. В каждом эпизоде

реализовывалось 50 испытаний. Результаты моделирования характеризуют быструю сходимость алгоритмов – для первого эксперимента использован алгоритм DDQN-обучения, для второго и третьего – алгоритм DQN-обучения, для четвертого эксперимента – алгоритм SARSA-обучения, для пятого – алгоритм PRO.

В табл. 2 представлены результаты оценки двух характеристик производительности алгоритмов – эффективности политики и эффективности обучения агента. Показано, что наилучший результат по этим двух критериям обеспечивает алгоритм DDQN, дающий сравнительно быстрое обучение и положительное вознаграждение. Значения средних вознаграждений, полученные за каждый из 200 эпизодов моделирования, усредненных по 50 испытаниям. Результаты получены со следующими гиперпараметрами: $\gamma = 0.99$, $\varepsilon = 0.04$.

Таблица 2. Результаты оценки эффективности политик и эффективности обучения
Table 2. Results of evaluation of policies and training efficiency

Алгоритм	Эффективность политики – среднее вознаграждение	Эффективность обучения – скорость сходимости (число эпизодов)
DQN	–0.15	0.27 (53)
DDQN	0.2	0.29 (58)
SARSA	–3.1	0.59 (117)
PRO	–52	–

Таким образом, разработка политик, то есть оптимальных программных мероприятий для работников предприятия в зависимости от рассчитанного значения его человеческого капитала целесообразна и более эффективна на основе агента, обученного с помощью DDQN алгоритма. Реализация предложенных политик позволит повысить качество человечного капитала предприятия, и обеспечит рост интегральных показателей производственно-экономической деятельности.

Заключение

Цель, поставленная в работе, достигнута. Задачи управления организационными, в том числе социально-экономическими системами могут быть представлены как задачи управления в сложных дискретных средах с дискретным пространством состояний. Их решение осложняется наличием дополнительной неопределенности, связанной с присутствием человека в контуре системы, действия которого не всегда возможно спрогнозировать. При построении систем поддержки принятия решений на базе алгоритмов RL возникает задача выбора приемлемых алгоритмов не только с точки зрения их адекватности содержанию социально-экономической проблемы, но и обладающих высокой производительностью.

В работе показано, что методы RL имеют доказанную эффективность, когда особенности решаемых задач управления следующие: во-первых, объект управления характеризуется стохастической динамикой своих показателей, а управленческие решения не детерминированы; во-вторых, задачи управления носят стратегический характер; в-третьих, решение задачи управления представляется в виде последовательного принятия решений.

Осуществлен подробный анализ показателей качества методов и алгоритмов RL и на примере задачи управления человеческим капиталом предприятия осуществлен выбор алгоритмов, обладающих максимальной эффективностью выработанной политики и эффективностью обучения.

На основе проведенных экспериментов было показано, что наилучшие результаты в смысле достижения максимальной полезности в кратчайшие сроки обеспечивает алгоритм DDQN на базе Q-обучения, который позволяет решать задачу оптимального управления социально-экономической системой.

Направления дальнейших исследований

Критически важными задачами при использовании методов и алгоритмов обучения в подкреплением является проектирование сигналов вознаграждения, так как именно они оценивают прогресс в достижении поставленной цели исследования. В последнее время значительно большее внимание исследователей стало уделяться построению функции вознаграждения, состоящей из двух компонент. Первая компонента формирует внутреннюю мотивацию агента, отражая его уровень социальной ответственности за принятые им решения. Вторая компонента связана с внешней мотивацией, она формируется как награда от объекта управления. Синтез таких комплексных наград может значительно улучшить процесс обучения агента за счет улучшения производительности используемых алгоритмов.

СПИСОК ИСТОЧНИКОВ

1. Sutton R.S., Barto A.G. (2020) *Reinforcement Learning. Am Introduction*. MIT Press, Cambridge, MA, 552 p.

2. LeCun Y., Bengio Y., Hinton G. (2015) Deep learning. *Nature*. 521 (7553), 436–444. DOI: https://doi.org/10.1038/nature14539

3. Ding Q., Jahanshahi H., Wang Y., Bekiros S., Alassafi M.O. (2022) Optimal Reinforcement Learning-Based Control Algorithm for a Class of Nonlinear Macroeconomic Systems. *Mathematics*, 10 (499). DOI: https://doi.org/10.3390/math10030499

4. Li Q., Lin T., Yu Q., Du H., Li J., Fu X. (2023) Review of Deep Reinforcement Learning and Its Application in Modern Renewable Power System Control. *Energies*, 16 (4143). DOI: https://doi.org/10.3390/en16104143

5. Wang R., Chen Z., Xing Q., Zhang Z., Zhang T. A. (2022) Modified Rainbow-Based Deep Reinforcement Learning Method for Optimal Scheduling of Charging Station. *Sustainability*, 14 (1884). DOI: https://doi.org/10.3390/su14031884

6. Abideen A.Z., Sundram V.P.K., Pyeman J., Othman A.K., Sorooshian S. (2021) Digital Twin Integrated Reinforced Learning in Supply Chain and Logistics. *Logistics*, 5 (84). DOI: https://doi.org/10.3390/logistics5040084

7. Yan Y., Chow A.H., Ho C.P., Kuo Y.H., Wu Q., Ying C. (2022). Reinforcement Learning for Logistics and Supply Chain Management: Methodologies, State of the Art, and Future Opportunities. *Transportation Research Part E: Logistics and Transportation Review*, 162 (102712).

8. Han D., Mulyana B., Stankovic V., Cheng S. A. (2023) Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation. *Sensors*, 23 (3762). DOI: https://doi.org/10.3390/s23073762

9. Orr J., Dutta A. (2023) Multi-Agent Deep Reinforcement Learning for Multi-Robot Applications: A Survey. Sensors, 23(3625). DOI: https://doi.org/10.3390/s23073625

10. Mohammadi M., Al-Fuqaha A., Guizani M., Oh J. (2018) Semi-supervised deep reinforcement learning in support of IoT and Smart City Services. *IEEE Internet of Things Joumal*, 5 (2), 624–635.

11. Yu C., Liu J., Nemati S. (2019) Reinforcement Learning in Healthcare: A Survey. *arXiv:*1908.08796. DOI: https://doi.org/10.48550/arXiv.1908.08796

12. Chi M., VanLehn K., Litman D. et al. (2011) Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Model User-Adapted Interaction*, 21, 137–180. DOI: https://doi.org/10.1007/s11257-010-9093-1

13. Azhikodan A.R., Bhat A.G., Jadhav M.V. (2019) Stock Trading Bot Using Deep Reinforcement Learning. Innovations in Computer Science and Engineering. Springer: Berlin/Heidelberg, Germany, 41–49.

14. Пономарев Е.С., Оселедец И.В., Чихоцкий А.С. (2019) Использование обучения с подкреплением в задаче алгоритмической торговли. *Информационные процессы*, 19 (2), 122–131.

15. Гурин А.С., Гурин Я.С., Горохова Р.И., Корчагин С.А., Никитин П.В. (2020) Повышение доходности торгового агента на основе метода Q-learning посредством использования производных финансовых показателей. *Современные информационные технологии и ИТ-образование*. 16 (3), 799–809. DOI: https://doi.org/10.25559/SITITO.16.202003.799-809

16. Гатауллин С.Т., Хасаншин И.Я., Никитин П.В., Семенов Д.Н., Круглов В.И., Мельникова А.И. (2021) Различные подходы применения технологии обучения с подкреплением в алгоритмической торговле. *Фундаментальные исследования*, 12, 86–91. DOI: https://doi.org/10.17513/fr.43158

17. Jiang Z., Liang J. (2017) Cryptocurrency portfolio management with deep reinforcement learning. In: *Proceedings of the 2017 Intelligent Systems Conference (IntelliSys), London, UK*, 905–913.

18. Yu P., Lee J.S., Kulyatin I., Shi Z., Dasgupta S. (2019) Model-based Deep Reinforcement Learning for Dynamic Portfolio Optimization. *arXiv:*1901.08740.

19. Amirzadeh R., Nazari A., Thiruvady D. (2022) Applying Artificial Intelligence in Cryptocurrency Markets: A Survey. *Algorithms*, 15 (428). DOI: https://doi.org/10.3390/a15110428

20. Feng L., Tang R., Li X., Zhang W., Ye Y., Chen H., Guo H., Zhang Y. (2018) Deep reinforcement learning based recommendation with explicit user-item interactions modeling. *arXiv:*1810.12027.

21. Liu J., Zhang Y., Wang X., Deng Y., Wu X. (2019) Dynamic Pricing on E-commerce Platform with Deep Reinforcement Learning. *arXiv:*1912.02572.

22. Zheng G., Zhang F., Zheng Z., Xiang Y., Yuan N.J., Xie X., Li Z. (2018) DRN: A deep reinforcement learning framework for news recommendation. In: *Proceedings of the 2018WorldWideWeb Conference, Lyon, France*, 167–176.

23. Orlova E.V. (2021) Design of Personal Trajectories for Employees' Professional Development in the Knowledge Society under Industry 5.0. *Social Sciences*, 10 (11), 427. DOI: https://doi.org/10.3390/socsci10110427

24. Orlova E. V. (2021) Assessment of the Human Capital of an Enterprise and its Management in the Context of the Digital Transformation of the Economy. *Journal of Applied Economic Research*, 20 (4), 666–700. DOI: https://doi.org/10.15826/vestnik.2021.20.4.026

25. Orlova E.V. (2021) Methodology and Models for Individuals' Creditworthiness Management Using Digital Footprint Data and Machine Learning Methods. *Mathematics*, 9 (15). DOI: https://doi.org/10.3390/math9151820

26. Hung C.C., Lillicrap T., Abramson J., Wu Y., Mirza M., Carnevale F., Ahuja A., Wayne G. (2019) Optimizing agent behavior over long time scales by transporting value. *Nature Communication*. 10 (1), 5223. DOI: https://doi.org/10.1038/s41467-019-13073-w

27. Colas C., Sigaud O., Oudeyer P.-Y. (2019) A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms. *arXiv:*1904.06979.

28. Orlova E.V. (2023) Inference of Factors for Labor Productivity Growth Used Randomized Experiment and Statistical Causality. *Mathematics*, 11 (4), 863. DOI: https://doi.org/10.3390/math11040863

29. Orlova E.V. (2022) Methodology and Statistical Modeling of Social Capital Influence on Employees' Individual Innovativeness in a Company. *Mathematics*, 10 (11),1809. DOI: https://doi.org/10.3390/math10111809

REFERENCES

1. Sutton R.S., Barto A.G. (2020) *Reinforcement Learning. Am Introduction*. MIT Press, Cambridge, MA, 552 p.

2. LeCun Y., Bengio Y., Hinton G. (2015) Deep learning. *Nature*. 521 (7553), 436–444. DOI: https://doi.org/10.1038/nature14539

3. Ding Q., Jahanshahi H., Wang Y., Bekiros S., Alassafi M.O. (2022) Optimal Reinforcement Learning-Based Control Algorithm for a Class of Nonlinear Macroeconomic Systems. *Mathematics*, 10 (499). DOI: https://doi.org/10.3390/math10030499

4. Li Q., Lin T., Yu Q., Du H., Li J., Fu X. (2023) Review of Deep Reinforcement Learning and Its Application in Modern Renewable Power System Control. *Energies*, 16 (4143). DOI: https://doi.org/10.3390/en16104143

5. Wang R., Chen Z., Xing Q., Zhang Z., Zhang T. A. (2022) Modified Rainbow-Based Deep Reinforcement Learning Method for Optimal Scheduling of Charging Station. *Sustainability*, 14 (1884). DOI: https://doi.org/10.3390/su14031884

6. Abideen A.Z., Sundram V.P.K., Pyeman J., Othman A.K., Sorooshian S. (2021) Digital Twin Integrated Reinforced Learning in Supply Chain and Logistics. *Logistics*, 5 (84). DOI: https://doi.org/10.3390/logistics5040084

7. Yan Y., Chow A.H., Ho C.P., Kuo Y.H., Wu Q., Ying C. (2022). Reinforcement Learning for Logistics and Supply Chain Management: Methodologies, State of the Art, and Future Opportunities. *Transportation Research Part E: Logistics and Transportation Review*, 162 (102712).

8. Han D., Mulyana B., Stankovic V., Cheng S. A. (2023) Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation. *Sensors*, 23 (3762). DOI: https://doi.org/10.3390/s23073762

9. Orr J., Dutta A. (2023) Multi-Agent Deep Reinforcement Learning for Multi-Robot Applications: A Survey. *Sensors*, 23 (3625). DOI: https://doi.org/10.3390/s23073625

10. Mohammadi M., Al-Fuqaha A., Guizani M., Oh J. (2018) Semi-supervised deep reinforcement learning in support of IoT and Smart City Services. *IEEE Internet of Things Journal*, 5 (2), 624–635.

11. Yu C., Liu J., Nemati S. (2019) Reinforcement Learning in Healthcare: A Survey. *arXiv:*1908.08796. DOI: https://doi.org/10.48550/arXiv.1908.08796

12. Chi M., VanLehn K., Litman D. et al. (2011) Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Model User-Adapted Interaction*, 21, 137–180. DOI: https://doi.org/10.1007/s11257-010-9093-1

13. Azhikodan A.R., Bhat A.G., Jadhav M.V. (2019) Stock Trading Bot Using Deep Reinforcement Learning. *Innovations in Computer Science and Engineering*. Springer: Berlin/Heidelberg, Germany, 41–49.

14. Ponomarev E.S., Oseledets I.V., Chikhotskii A.S. (2019) Ispol'zovanie obucheniya s podkrepleniem v zadache algoritmicheskoi torgovli. *Informatsionnye protsessy*, 19 (2), 122–131.

15. Gurin A.S., Gurin Ya.S., Gorokhova R.I., Korchagin S.A., Nikitin P.V. (2020) Povyshenie dokhodnosti torgovogo agenta na osnove metoda Q-learning posredstvom ispol'zovaniya proizvodnykh finansovykh pokazatelei. *Sovremennye informatsionnye tekhnologii i IT-obrazovanie*. 16 (3), 799–809. DOI: https://doi.org/10.25559/SITITO.16.202003.799-809

16. Gataullin S.T., Khasanshin I.Ya., Nikitin P.V., Semenov D.N., Kruglov V.I., Mel'nikova A.I. (2021) Razlichnye podkhody primeneniya tekhnologii obucheniya s podkrepleniem v algoritmicheskoi torgovle. *Fundamental'nye issledovaniya*, 12, 86–91. DOI: https://doi.org/10.17513/fr.43158

17. Jiang Z., Liang J. (2017) Cryptocurrency portfolio management with deep reinforcement learning. In: *Proceedings of the 2017 Intelligent Systems Conference (IntelliSys), London, UK*, 905–913.

18. Yu P., Lee J.S., Kulyatin I., Shi Z., Dasgupta S. (2019) Model-based Deep Reinforcement Learning for Dynamic Portfolio Optimization. *arXiv:*1901.08740.

19. Amirzadeh R., Nazari A., Thiruvady D. (2022) Applying Artificial Intelligence in Cryptocurrency Markets: A Survey. *Algorithms*, 15 (428). DOI: https://doi.org/10.3390/a15110428

20. Feng L., Tang R., Li X., Zhang W., Ye Y., Chen H., Guo H., Zhang Y. (2018) Deep reinforcement learning based recommendation with explicit user-item interactions modeling. *arXiv:*1810.12027.

21. Liu J., Zhang Y., Wang X., Deng Y., Wu X. (2019) Dynamic Pricing on E-commerce Platform with Deep Reinforcement Learning. *arXiv:*1912.02572.

22. Zheng G., Zhang F., Zheng Z., Xiang Y., Yuan N.J., Xie X., Li Z. (2018) DRN: A deep reinforcement learning framework for news recommendation. In: *Proceedings of the 2018WorldWideWeb Conference, Lyon, France*, 167–176.

23. Orlova E.V. (2021) Design of Personal Trajectories for Employees' Professional Development in the Knowledge Society under Industry 5.0. *Social Sciences*, 10 (11), 427. DOI: https://doi.org/10.3390/socsci10110427

24. Orlova E.V. (2021) Assessment of the Human Capital of an Enterprise and its Management in the Context of the Digital Transformation of the Economy. *Journal of Applied Economic Research*, 20 (4), 666–700. DOI: https://doi.org/10.15826/vestnik.2021.20.4.026

25. Orlova E.V. (2021) Methodology and Models for Individuals' Creditworthiness Management Using Digital Footprint Data and Machine Learning Methods. *Mathematics*, 9 (15). DOI: https://doi.org/10.3390/math9151820

26. Hung C.C., Lillicrap T., Abramson J., Wu Y., Mirza M., Carnevale F., Ahuja A., Wayne G. (2019) Optimizing agent behavior over long time scales by transporting value. *Nature Communication*. 10 (1), 5223. DOI: https://doi.org/10.1038/s41467-019-13073-w

27. Colas C., Sigaud O., Oudeyer P.-Y. (2019) A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms. *arXiv:*1904.06979.

28. Orlova E.V. (2023) Inference of Factors for Labor Productivity Growth Used Randomized Experiment and Statistical Causality. *Mathematics*, 11 (4), 863. DOI: https://doi.org/10.3390/math11040863

29. Orlova E.V. (2022) Methodology and Statistical Modeling of Social Capital Influence on Employees' Individual Innovativeness in a Company. Mathematics, 10 (11), 1809. DOI: https://doi.org/10.3390/math10111809

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

ОРЛОВА Екатерина Владимировна
E-mail: ekorl@mail.ru
Ekaterina V. ORLOVA
E-mail: ekorl@mail.ru
ORCID: https://orcid.org/0000-0001-6535-6727

Поступила: 07.08.2023; Одобрена: 27.09.2023; Принята: 02.10.2023.
Submitted: 07.08.2023; Approved: 27.09.2023; Accepted: 02.10.2023.