

УДК 004.85

doi: 10.15622/rcai.2025.060

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ С ВНЕШНИМ УЧИТЕЛЕМ ДЛЯ УПРАВЛЕНИЯ ЭНЕРГООБЪЕКТАМИ

О.Ю. Марьясин (*maryasin2003@list.ru*)

А.Н. Плохотнюк (*admin@nixson.ru*)

Ярославский государственный технический университет,
Ярославль

Рассмотрена задача управления энергообъектом с применением алгоритмов обучения с подкреплением. Для решения проблемы длительного обучения агента авторы предложили подход, заключающийся в использовании внешнего учителя. В работе описаны различные способы применения внешнего учителя и рассмотрена постановка задачи управления энергообъектом, допускающая использование популярных алгоритмов обучения с подкреплением с внешним учителем. Результаты обучения агентов на виртуальной среде показали, что для всех алгоритмов обучения с подкреплением применение внешнего учителя позволяет достичь гораздо меньшего энергопотребления за фиксированное время чем для подобных алгоритмов без внешнего учителя. Таким образом, использование алгоритмов обучения с подкреплением с внешним учителем может как значительно ускорить обучение агентов, так и повысить эффективность управления на начальном этапе обучения.

Ключевые слова: управление энергообъектами, обучение с подкреплением, внешний учитель, оптимизация энергопотребления.

Введение

Одним из многообещающих подходов к управлению энергетическими объектами является обучение с подкреплением (Reinforcement Learning – RL). RL это область машинного обучения, в которой используется понятие агента [Graesser et al., 2019]. В RL агент взаимодействует со средой, наблюдает текущее состояние среды и использует данную информацию при выборе своего действия. Среда в результате действия агента переходит в следующее состояние и возвращает агенту свое новое состояние и вознаграждение, позволяющее агенту оценить успешность его действия.

В настоящее время алгоритмы RL широко применяются для решения сложных задач в таких областях как игры, робототехника, автономные мобильные транспортные средства и др. Растет интерес, связанный с использованием алгоритмов RL для управления сложными техническими объектами и системами, в том числе энергетическими. Например, согласно [Perera et al., 2021] число публикаций посвященных использованию методов RL в области энергетических систем с 2000 года неуклонно растет, а с 2016 года темпы роста числа публикаций значительно ускорились. В том числе растет число публикаций, описывающих применение методов RL для управления такими энергетическими объектами как здания и инженерные системы зданий [Sierla et al., 2022].

Первый опыт применения RL для управления энергообъектами выявил, как достоинства, так и некоторые проблемы, связанные с использованием данного подхода. Несомненным достоинством алгоритмов RL по сравнению с классическими методами оптимального управления, таким как Model Predictive Control [Afram et al., 2014], является то, что для их работы не требуется знания точных и сложных моделей энергообъектов [Yu et al., 2021]. Агенты RL могут методом проб и ошибок изучить оптимальную стратегию управления энергообъектом. Это также решает проблемы, связанные с изменчивостью и неопределенностью параметров энергообъекта.

Главной проблемой, с которой столкнулись исследователи при применении алгоритмов RL для управления энергообъектами является неприемлемо длительное время обучения агента. По этой причине ряд авторов пришел к выводу, что RL является проблематичным подходом к управлению энергообъектами [Sierla et al., 2022]. Кроме того, поскольку агент RL обучается методом проб и ошибок, то длительное время стратегия управления может быть неэффективной или вообще энергозатратной. Поэтому большинство авторов считают, что основное обучение агента RL не должно производиться на реальном объекте в реальном масштабе времени. Для этого необходимо использовать виртуальные среды и проводить обучение в сжатом модельном времени [Sierla et al., 2022]. Последующее применение и дообучение агента может происходить в реальных условиях.

К другим вопросам, которые необходимо решить при применении алгоритмов RL для управления энергообъектами относятся: выбор переменных среды, образующих множество состояний, выбор функции вознаграждения, адекватной поставленной цели управления, выбор алгоритмов RL и настройка их гиперпараметров [Sierla et al., 2022]. Анализ литературы показывает, что сначала наиболее популярным алгоритмом RL, который использовался для решения задач управления энергообъектами был Q-Learning [Wang et al., 2023]. Это связано с простотой реализации данного алгоритма. Затем, когда появились и стали доступными алгоритмы глубокого RL и алгоритмы RL с непрерывным множеством действий, то они также стали применяться для управления энергообъектами [Yu et al., 2021].

1. RL с внешним учителем

В научной литературе по RL нет изобилия публикаций, в которых описываются подходы, близкие к рассматриваемому в данной работе. В [Argerich et al., 2020] внешний репетитор (Tutor) применялся для улучшения качества обучения агента за счет использования внешних знаний для управления решениями агента. Авторы назвали данный метод обучением с подкреплением с использованием внешних знаний (External knowledge). Внешние знания, такие как экспертные или предметные знания, выражаются в виде программируемых функций, которые агент использует на этапе обучения. В [Argerich et al., 2020] рассматривается два вида программируемых функций – это функции ограничения и направляющие функции. Функция ограничения принимает текущее состояние среды и возвращает вектор, указывающий, можно ли выполнить соответствующее действие агента или нет. Направляющая функция принимает состояние и вознаграждение в качестве входных данных и выдает вектор, представляющий вес каждого действия агента.

Более известен другой подход к RL, получивший название система учитель-ученик (Teacher-student) [Zimmer et al., 2014]. В рамках этой системы агент-ученик учится выполнять задачу посредством RL, в то время как агент-учитель может оказывать помощь ученику учиться быстрее. В этой схеме предполагается что учитель также является агентом RL и уже усвоил оптимальную политику. Кроме того, учитель может дать лишь ограниченное количество советов, называемое бюджетом. В [Zimmer et al., 2014] представлено четыре эвристических метода для определения того, когда учитель может давать совет, включая раннее консультирование, консультирование по важности, исправление ошибок и прогнозное консультирование. При этом учитель может дать совет в форме действия, которое ученик должен выполнить (действие, лучшее, чем объявленное учеником).

В данной работе предлагается подход, использующий некоторые идеи из отмеченных ранее источников. В данном подходе в классическую схему RL, включающую агента и среду вводится внешний блок учитель (Teacher). Учитель принимает информацию о множестве состояний среды, действиях агента и некоторую внешнюю информацию (внешние сигналы), недоступную для агента. Когда наступает определенное событие, связанное со средой или агентом, учитель тем или иным способом оказывает воздействие на агента. При этом учитель может не только подсказывать агенту какое действие будет лучшим из сформированных агентом действий, но и непосредственно участвовать в формировании действия или даже выполнять определенные действия вместо агента. Этим учитель отличается от репетитора из [Argerich et al., 2020], способного лишь советовать агенту предпочтительные действия. Имеется два основных отличия пред-

лагаемой системы от системы учитель-ученик. Первым является то, что учитель не обязан быть агентом RL и может быть реализован любым другим способом. Например, он может быть построен на основе правил. Правила не должны охватывать все возможные ситуации. Они служат для отработки наиболее важных или критических событий. Второе отличие в том, что учителю доступна дополнительная внешняя информация, которая недоступна агенту. Важно отметить, что учитель не может заменить ученика. Учитель не обладает всеми способностями, которые получает агент в процессе обучения. Учитель только помогает агенту учиться быстрее, за счет того, что он позволяет агенту избегать неэффективных стратегий управления.

Далее рассматриваются различные способы использования внешнего учителя. Это стимулирование агента учителем, непосредственное изменение его действия и комбинированный способ. Блок-схема поясняющая способ стимулирования агента учителем показана на рис. 1. На рис. 1 использованы следующие обозначения: s_t – состояние среды в момент t , $s_t \in S$, r_t – вознаграждение (reward), a_t – действие (action) агента, $a_t \in A$, g_t – функция, выполняющая роль стимула для агента.

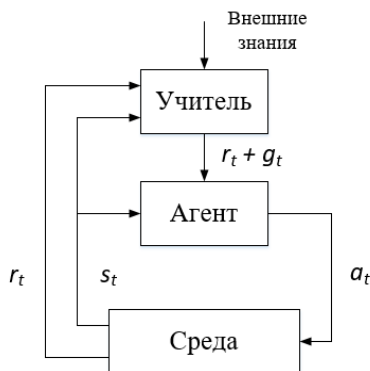


Рис. 1. Блок-схема для стимулирования агента учителем

Способ стимулирования заключается в том, что учитель изменяет вознаграждение r_t , получаемое агентом увеличивая (стимулируя) или уменьшая его. Функция стимула g_t выполняет роль функции штрафа в методах оптимизации. Например, когда необходимо заставить агента соблюдать ограничения накладываемые на состояния среды s Учитель для формирования функции g_t может использовать как информацию о состоянии s_t , передаваемую агенту, так и внешнюю информацию о среде, которая не доступна агенту.

На рис. 2 показана блок-схема, поясняющая способ непосредственного изменения действия агента учителем. На рис. 2 к обозначениям, показанным на рис. 1 добавлены: h_t – сигнал учителя, f_t – преобразованное действие. Блок трансформации F – в общем случае нелинейный блок, в котором на основании информации о действии агента a_t и сигнала учителя формируется новое действие f_t , передаваемое в среду.

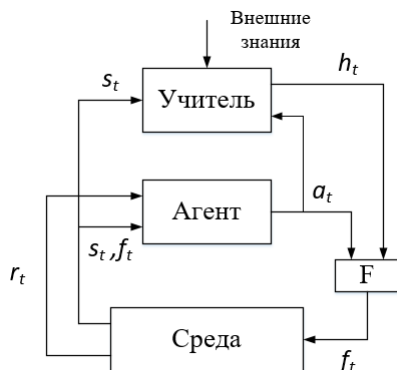


Рис. 2. Блок-схема для непосредственного изменения действия агента учителем

Способ непосредственного изменения действия агента учителем заключается в том, что при наступлении определенного события (внутреннего или внешнего), связанного со средой или агентом, учитель может изменить действие агента, усиливая или, наоборот, ослабляя его. При этом учитель изменяет действия агента только для конечного числа определенных ситуаций и только в том случае если действие, предложенное агентом, не соответствует требуемой реакции на данную ситуацию. Таким образом, учитель может корректировать действия агента, приводящие к неэффективным стратегиям управления, что особенно важно на начальных этапах обучения.

Учитель для формирования h_t может использовать как информацию о состоянии s_t , передаваемую агенту, так и внешнюю информацию, которая не доступна агенту. Информация о преобразованном действии должна быть добавлена к переменным состояния s_t и использоваться при обучении агента. По мере того, как агент накапливает больше опыта, он изучает свои действия и действия, скорректированные учителем, тем самым улучшая свои результаты. Дальнейшее развитие данного способа может привести к имитационному обучению (Imitation Learning) [Shenfeld et al., 2023], когда агент учится имитировать действия учителя.

Блок-схема, поясняющая комбинированный способ использования учителя показана на рис. 3. Комбинированный способ объединяет возможности двух предыдущих способов.

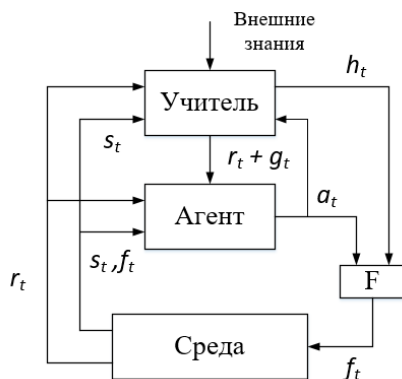


Рис. 3. Блок-схема для комбинированного способа использования учителя

2. Задача управления энергообъектом

Прежде чем решить задачу управления энергообъектом с использованием RL необходимо ответить на ряд вопросов: какая цель управления, что будет использоваться в качестве среды, какие переменные среды войдут в множество состояний s_t , какое вознаграждение будет получать агент и какие управляющие воздействия на среду составят набор действий агента a

В качестве энергообъектов в данной работе рассматриваются здания. На здания приходится около 30–40% всей энергии, потребляемой как в развитых, так и в развивающихся странах [Hossain et al., 2023]. Следовательно, необходимо разрабатывать эффективные методы управления энергопотреблением зданий, которые могут обеспечить оптимальный компромисс между потреблением энергии и обеспечением комфортного микроклимата помещений здания. Одной из наиболее энергоемких инженерных систем здания является система отопления, вентиляции и кондиционирования (ОВК). Параметры работы ОВК оборудования могут сильно влиять на энергопотребление здания [Марьясин и др., 2017]. Поэтому целью управления будет снижение энергозатрат на отопление и охлаждение здания с учетом сохранения комфортного микроклимата внутри помещений здания.

В настоящее время в качестве среды для обучения агентов RL при решении задачи управления энергопотреблением зданий принято использовать виртуальные среды на базе энергомоделей, построенных с применением систем энергомоделирования (Building Energy Modeling – BEM)

[Wang et al., 2023]. При этом, в большинстве случаев, в качестве ВЕМ системы применялась популярная программа EnergyPlus [Wang et al., 2023]. Поэтому в данной работе в качестве виртуальной среды также использовалась ВЕМ система EnergyPlus. Энергомоделирование производилось для того же 5-зонного одноэтажного здания, с той же системой ОВК и при тех же условиях, что и в [Maryasin et al., 2023].

В научной литературе нет единого мнения, какие переменные среды необходимо включать в состав состояния s_t при решении задачи управления энергопотреблением здания с помощью RL. Включение в состав состояния s_t большого числа переменных приводит к резкому росту трудоемкости обучения агентов RL (проклятие размерности) [Jia et al., 2019]. Поэтому в состав состояния s_t стараются включить только наиболее важные, по мнению авторов, переменные. В данной работе состояние s_t включает температуру зон здания, показатель, определяющий степень комфорта людей в помещении (Predicted Mean Vote [Dyvia et al., 2021]) в каждой из зон здания и наличие людей в помещениях зон здания. Кроме того, в зависимости от способа использования учителя, в состояние s_t могут включаться задания локальным регуляторам системы ОВК.

Для достижения поставленной цели функция вознаграждения должна поощрять агента к снижению энергопотребления при сохранении комфортного микроклимата внутри помещений здания. В соответствии с этим функция вознаграждения r_t агента будет иметь вид

$$r_t = -q_t - \lambda_1 p_t + \lambda_2 g_t. \quad (2.1)$$

где q_t – суммарное количество энергии, затраченное на отопление и охлаждение всех зон здания в момент времени t , p_t – функция штрафа за нарушение ограничений по температуре в каждой из зон здания, g_t – функция стимула учителя, λ_1, λ_2 – заданные коэффициенты. Функция штрафа p_t может быть постоянной, штрафую за любые нарушения ограничений по температуре или она может зависеть от величины отклонения от заданного температурного диапазона

$$T_{ztl} \leq T_{zt} \leq T_{ztl}, z = 1, \dots, Z, t = 0, \dots, H, \quad (2.2)$$

где T_{ztl}, T_{ztl} – минимальное и максимальное значения температуры в z -й зоне здания в момент времени t , H – горизонт управления, Z – число зон здания. Функция стимула также может быть постоянной, штрафую за любые неправильные действия агента или она может зависеть от степени различия действий агента и учителя. Для способа непосредственного изменения действия агента учителем $\lambda_2 = 0$.

Набор действий агента $a_t = \{T_{zhsb}, T_{zcsf}\}$ включает задания на отопление T_{zhsb} и охлаждение T_{zcsf} локальных регуляторов температуры зон здания. Таким образом, агент RL учится решать задачу определения значений заданий

по температуре T_{zhst} и T_{zcst} максимизирующих функцию вознаграждения агента (2.1), тем самым минимизируя количество энергии, затраченное на отопление и охлаждение всех зон здания с учетом ограничения (2.2).

Как было сказано в разделе 1 учитель может обладать некоторой внешней информацией, недоступной для агента. В качестве такой информации в задаче управления энергопотреблением здания с учетом наличия людей в помещениях выступают данные о графике работы персонала здания в рабочие и выходные дни. Тогда с учетом данной информации функция стимула g_t в (2.1) может быть реализована следующим образом. Эта функция равна нулю в рабочее время и равна

$$g_t = |T_{zhst} - T_{zhstl}| + |T_{zcst} - T_{zcstl}| \quad (2.3)$$

в нерабочее время, где T_{zhstl} , T_{zcstl} – минимальные значения задания для отопления и охлаждения в z -ой зоне здания в момент времени t . Аналогично может быть реализован блок F в схеме для непосредственного изменения действия агента учителем. В рабочее время выход блока f_t равен действию агента a_t , в нерабочее время f_t принимает значения

$$f_t = \min(a_t, h_t), \quad (2.4)$$

где $h_t = \{T_{zhstl}, T_{zcstl}\}$ – значения, сообщаемые учителем. При применении комбинированного способа использования учителя функции (2.3) и (2.4) реализуется одновременно.

3. Решение задачи управления энергообъектом

Для управления энергопотреблением здания использовались алгоритмы RL с непрерывным множеством действий Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), Deterministic Policy Gradient (DPG) и Deep Deterministic Policy Gradient (DDPG).

В табл. 1 приведены значения суточного энергопотребления для отопления и охлаждения помещений здания для двух вариантов, соответствующих одному и тому же зданию. Вариант 1 включает значения, полученные с помощью указанных алгоритмов RL для способа непосредственного изменения действия агента учителем. Вариант 2 – значения, полученные для способа стимулирования. Во всех вариантах используется график посещения людьми различных зон здания и график работы системы ОВК здания, предусматривающие снижение температуры в нерабочее время. Обучение агентов производилось в течение 10000 шагов.

Анализ данных из табл. 1 показывает, что способ непосредственного изменения действий агента учителем, обеспечивает получение лучших результатов, чем способ стимулирования. Это объясняется тем, что в первом способе учитель непосредственно принимает участие в управлении, выдавая правильные действия, в то время как для способа стимулирования агент не успевает обучиться достаточно хорошо за фиксирован-

ное время. Следовательно, применение способа непосредственного изменения действия агента учителем позволяет быстрее добиться лучших результатов.

Таблица 1

Алгоритм	Вариант 1, кВт/час	Вариант 2, кВт/час
PPO	76,310	109,561
A2C	75,408	107,257
DPG	69,010	102,635
DDPG	68,643	86,851

Кроме алгоритмов RL с непрерывным множеством действий для управления энергопотреблением здания были применены алгоритмы глубокого RL с дискретным множеством действий Deep Q-Network (DQN) и Double DQN (DDQN). Для данных алгоритмов была исследована зависимость суточного энергопотребления от степени дискретности действий агентов. При этом весь диапазон изменения действий агентов разбивался на заданное число интервалов, равное коэффициенту дискретности. Значение дискретности были взяты равными 4, 10, 20. Исследование проводилось для способа непосредственного изменения действия агента учителем. Результаты исследования приведены в табл. 2.

Таблица 2

Дискретность	DQN, кВт/час	DDQN, кВт/час
4	64,713	71,876
10	64,007	69,068
20	64,725	62,600

Анализ данных из табл. 2 показывает, что для алгоритма DQN изменение дискретности слабо влияет на суточное энергопотребление. Для алгоритма DDQN при увеличении дискретности суточное энергопотребление немного снижается. Для обоих алгоритмов значение энергопотребления, при дискретности равной 20, немного ниже, чем для алгоритмов с непрерывным множеством действий. Это говорит о небольшом преимуществе использования алгоритмов глубокого RL с дискретным множеством действий для управления энергопотреблением зданий.

Заключение

В работе рассмотрена задача управления энергообъектом (зданием) с применением алгоритмов RL. Для решения проблемы длительного обучения агента RL авторы предложили подход, заключающийся в использовании внешнего учителя. В работе описаны различные способы применения внешнего учителя и рассмотрена постановка задачи управления энергопо-

реблением зданий, допускающая использование популярных алгоритмов RL с внешним учителем. Результаты обучения агентов RL на виртуальной среде показали, что для всех алгоритмов RL применение внешнего учителя позволяет достичь гораздо меньшего энергопотребления за фиксированное время чем для подобных алгоритмов без внешнего учителя. Таким образом, использование алгоритмов RL с внешним учителем может как значительно ускорить обучение агентов RL, так и повысить эффективность управления на начальном этапе обучения.

В работе также произведено сравнение эффективности применения различных алгоритмов RL, в том числе алгоритмов RL с непрерывным и дискретным множеством действий. Результаты сравнения не выявили значительного преимущества использования алгоритмов глубокого RL с дискретным множеством действий (при дискретности равной четырем и более) перед алгоритмами RL с непрерывным множеством действий.

Список литературы

- [Марьясин и др., 2017] Марьясин О.Ю., Колодкина А.С. Управление тепловым режимом зданий с использованием прогнозирующих моделей // Вестник Сам-ГТУ. – 2017. – № 1 (53). – С. 122-132.
- [Afram et al., 2014] Afram A., Janabi-Sharifi F. Theory and applications of HVAC control systems – A review of model predictive control (MPC) // Building and Environment. – 2014. – Vol. 72. – P. 343-355.
- [Argerich et al., 2020] Argerich M.F., Furst J., Cheng B. Tutor4rl: Guiding reinforcement learning with external knowledge // Proc. AAAI Spring Symposium Combining Machine Learning with Knowledge Engineering. – 2020. – P. 1-4.
- [Dyvia et al., 2021] Dyvia H.A., Arif C. Analysis of thermal comfort with predicted mean vote (PMV) index using artificial neural network // IOP Conference Series: Earth and Environmental Science. – 2021. – Vol. 622. – P. 1-12.
- [Graesser et al., 2019] Graesser L., Keng W.L. Foundations of Deep Reinforcement Learning. – Addison-Wesley, 2019.
- [Hossain et al., 2023] Hossain J., Kadir A.F.A., Hanafi A.N., Shareef H., Khatib T., Baharin K.A., Sulaima M.F. A Review on Optimal Energy Management in Commercial Buildings // Energies. – 2023. – Vol. 16. – P. 1-40.
- [Jia et al., 2019] Jia R., Jin M., Sun K., Hong T., Spanos C. Advanced Building Control via Deep Reinforcement Learning // Energy Procedia. – 2019. – Vol. 158. – P. 6158-6163. – doi: <https://doi.org/10.1016/j.egypro.2019.01.494>.
- [Maryasin et al., 2023] Maryasin O.Yu., Plohotnyuk A. Reinforcement Learning-Based Approach to Optimization of Energy Consumption in a Building // Proc. 5th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency. – 2023. – P. 685-690.
- [Perera et al., 2021] Perera A.T.D., Kamalaruban P. Applications of reinforcement learning in energy systems // Renewable and Sustainable Energy Reviews. – 2021. – Vol. 137. – P. 1-22. – doi: <https://doi.org/10.1016/j.rser.2020.110618>.

- [Shenfeld et al., 2023]** Shenfeld I., Hong Z., Tamar A., Agrawal P. TGRL: An Algorithm for Teacher Guided Reinforcement Learning // Proc. 40th International Conference on Machine Learning. – 2023. – P. 1-18.
- [Sierla et al., 2022]** Sierla S., Ihasalo H., Vyatkin V. A Review of Reinforcement Learning Applications to Control of Heating, Ventilation and Air Conditioning Systems // Energies. – 2022. – Vol. 15. – P. 1-25.
- [Wang et al., 2023]** Wang M., Willes J., Jiralerspong T., Moezzi M. A Comparison of Classical and Deep Reinforcement Learning Methods for HVAC Control // arXiv preprint arXiv:2308.05711. – 2023. – P. 1-7.
- [Yu et al., 2021]** Yu L., Qin S., Zhang M., Shen C., Jiang T., Guan X. A Review of Deep Reinforcement Learning for Smart Building Energy Management // IEEE Internet of Things Journal. – 2021. – Vol. 8(15). – P. 12046-12063.
- [Zimmer et al., 2014]** Zimmer M., Viappiani P., Weng P. Teacher-Student Framework: A Reinforcement Learning Approach // Proc. AAMAS Workshop Autonomous Robots and Multirobot Systems. – 2014. – P. 1-17.