УДК 004.896, 681.5
DOI: 10.15827/0236-235X.131.413-419
Дата подачи статьи: 26.03.20
2020. Т. 33. № 3. С. 413–419

Применение передачи обучения в семиотических моделях к проблеме фуражирования с реальными роботами

В.В. Воробьев [1], *зам. начальника лаборатории робототехники, Vorobev_VV@nrcki.ru*
М.А. Ровбо [1], *инженер-исследователь, rovboma@gmail.com*

[1] *Национальный исследовательский центр «Курчатовский институт», г. Москва, 123182, Россия*

В статье рассматриваются особенности применения алгоритма передачи обучения для агентов с семиотическими моделями мира к задаче фуражировки с реальными роботами. Роботу необходимо собирать случайно размещаемую еду, которая при подборе появляется на новом случайном месте в пределах полигона. Мобильный робот управляется агентом с моделью мира, описывающей показания с датчиков как предикаты. Агент принимает решения на основе таблицы оценок состояний-действий с Q-обучением. Он предобучен на упрощенной модельной среде с дискретными состояниями, в которой действия выполняются гарантированно с детерминированным исходом.

В реальной среде и в ее модели с учетом физики действия могут быть выполнены некорректно в силу ошибки планировщика, погрешностей локализации и других проблем, а состояния среды определяются путем анализа данных с датчиков, работающих с непрерывным миром.

Показаны возможность реализации соответствующих интерфейсов и переносимость концепции с упрощенной модельной среды как на ее более полную модель с учетом физики, так и на реального робота. Перенос обучения также происходит успешно, однако итоговые показатели работы снижаются (вероятно, из-за неверности предположения о детерминированности мира в реальной среде) и роботу требуется дообучение. В качестве симулятора реальной среды с учетом физики использовался Gazebo, а реальный полигон был оборудован специальными маркерами и камерами для локализации. Использовались также элементы дополненной реальности в виде модуля виртуальной еды.

Ключевые слова: *обучение с подкреплением, робототехника, фуражировка, передача обучения, семиотические модели.*

Роботизированные системы с одним или несколькими агентами должны обладать адаптивными свойствами для надежного достижения различных целей в реальных неконтролируемых средах, что часто имеет место для мобильных роботов. Одной из выдающихся областей исследований, которые могут достичь этого, является обучение с подкреплением. Оно рассматривает роботов как агентов с предопределенным пространством ввода и дискретными или непрерывными действиями, функционирующих в среде и получающих награды за полезную работу. Это позволяет роботам изучать и оптимизировать целенаправленное поведение. С другой стороны, семиотические системы управления позволяют роботам иметь структурированные и более понятные модели мира, основанные на правилах и предикатах. Для использования методов обучения с подкреплением требуется некоторая адаптация соответствующих моделей и алгоритмов, поскольку семиотические системы управления используют знаки и связи между ними в качестве основы для описания мира, тогда как для обучения с подкреплением в основном используется векторизованное описание.

Знак состоит из четырех частей – имя, восприятие, функциональное значение и личное значение [1]. Восприятие представляет собой набор предикатов, используемых для описания концепции, закодированной знаком. Он может применяться для подключения датчиков агента и его логической части управления системой управления путем реализации некоторых предикатов в качестве алгоритмических функций данных датчика. Разные агенты могут иметь разные предикаты. Чтобы использовать обучение с подкреплением в многоагентной системе для таких агентов, требуется применение различных описаний предикатов при передаче опыта между агентами.

Передача опыта в обучении с подкреплением связана с применением знаний о решении задачи агентом к другой, несколько схожей проблеме, которая также может быть использована для передачи опыта между различными агентами. Важно учитывать различие между описаниями задачи в пространстве действий и

в пространстве задач, поскольку это может привести к более эффективным алгоритмам обучения [2]. Описания проблемного пространства обычно являются полными описаниями состояния среды и, следовательно, нецелесообразны для мобильных роботов, поэтому в данной работе основное внимание уделяется описаниям пространства агентов. Изучение переноса может быть выполнено различными способами, которые также зависят от специфики проблемы, в частности, является ли входное представление одним и тем же пространством, существуют ли общие полезные последовательности действий между задачами и т.д.

Идея метода, описанного в работе [3], состоит в том, чтобы использовать описания мира, свойственные семиотической системе управления, для облегчения передачи опыта между двумя обучающими агентами с различными описаниями состояния в пространстве агентов. С точки зрения таксономии, описанной в [4], рассматриваемая проблема заключается в переносе между задачами с разными доменами (с точки зрения пространства агента) или переносе между задачами с фиксированной областью (с точки зрения пространства задач представления), в то время как алгоритм использует особый случай передачи параметров, который учитывает разницу в представлениях состояния агентов.

Проблема поиска пищи рассматривается как тестовая среда для алгоритмов, поскольку она является общей для многоагентных роботизированных исследований, использовалась для оценки обучения с подкреплением и может служить моделью для некоторых приложений, таких как сбор ресурсов или сбор энергии [5, 6].

Алгоритмы и методы

В работе используются алгоритмы управления, ранее описанные авторами в [3]. Метод состоит из двух этапов: обучение на упрощенной модельной задаче и перенос обучения на более сложную среду (приближенную модель с учетом физики, реализованную в соответствующем симуляторе или на реальном полигоне с роботом). При этом представление робота в сложной среде необходимо также свести к аналогичному в упрощенной модели для возможности применения алгоритма на основе таблицы состояний-действий.

Это представление состоит из вектора, элементы которого описывают содержимое соответствующей клетки (области около робота) и принимают значения «еда», «пусто», «агент», «препятствие». Агенту доступна информация лишь о части всей среды в локальной области рядом с ним, в данном случае в виде содержимого клеток спереди, сзади, слева, справа и по центру агента на сетке из квадратных клеток (рис. 1).

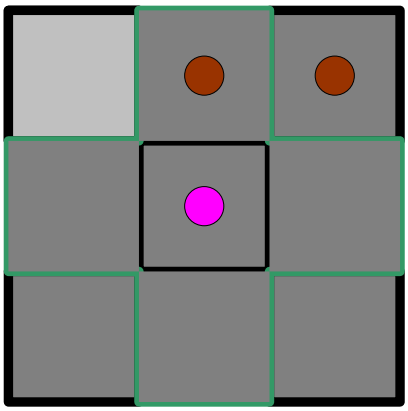


Рис. 1. Элемент упрощенной модельной среды, формирующий текущее состояние на входе агента

Fig. 1. Simplified model environment element that generates the current state at the agent input

Рассматриваемый алгоритм передачи обучения основан на идее группировки значений состояния-действия с использованием общих частей описаний предикатов или имен соответствующих описаний семиотического состояния. В данном случае они позволяют сопоставлять различные описания состояний агентов, определяя, какое подмножество состояний одного агента соответствует конкретному описанию состояния другого агента. Знак агента соединяет агента с именем *agent 1* с его сенсорными объектами в образе *p*, описывающем знак, выражает правила, которые определяют, когда конкретное действие может быть выполнено в значении *m*, и может сохранять его опыт в качестве правила перехода в личностном смысле *a* (рис. 2).

Агент использует ε-жадное Q-обучение для предобучения и для последующей работы после переноса обучения, выполняемого с помощью специальной функции переноса. Таким образом, обновление оценки состояния-действия происходит с помощью формулы $Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma Q(s', a') - Q(s, a))$, где s – прошлое состояние; a – выполненное действие; s' – новое состояние; a' – оптимальное

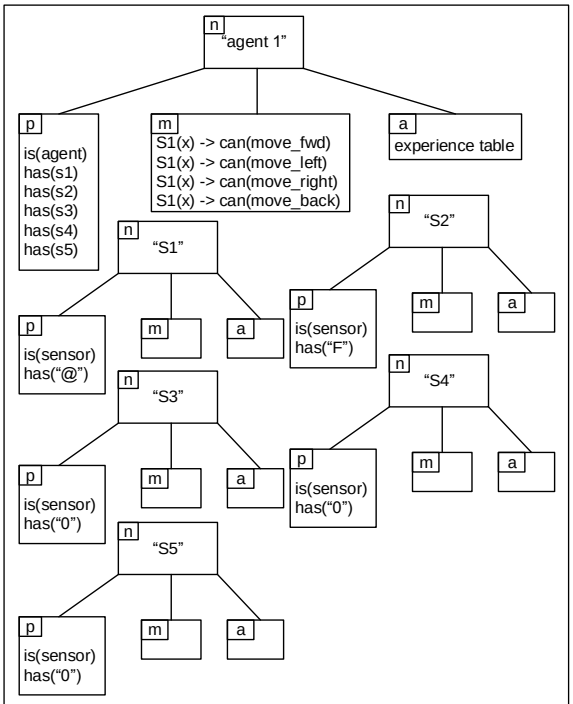


Рис. 2. Описание модели мира агента с пятью сенсорами на основе знаков (символы в значениях сенсоров обозначают пустое место (0), агента (@), пищу (F), препятствие (#))

Fig. 2. Description of the agent world models with five sensors based on signs (symbols in the sensor values indicate an empty space (0), an agent (@), food (F), an obstacle (#))

действие в новом состоянии (максимизирующее Q); α – скорость обучения; γ – коэффициент уменьшения важности будущих наград.

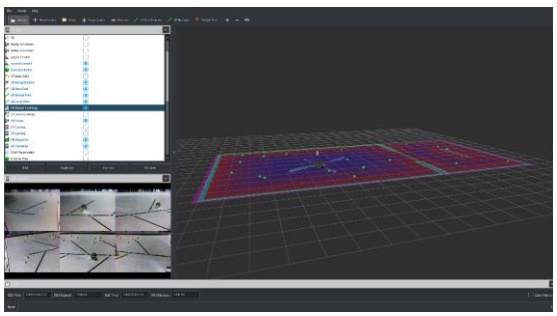
Выбор действия осуществляется с помощью функции вероятности выбора действия: $P(a) = \frac{e^{Q(s,a)}}{\sum_{i=0}^{n} e^{Q(s,a_i)}}$, где n – количество всех возможных действий в состоянии (в данной задаче постоянно для всех состояний); s – текущее состояние.

Перенос обучения между агентами с различными представлениями осуществляется с помощью функции, усредняющей значение оценки состояния-действия целевого агента по всем соответствующим ей значениям оценок исходного агента. Таким образом, оценка целевого агента для состояния $Q_t(s_t, a)$ получается из подмножества оценок исходного агента $Q_s(s_s, a)$ следующим образом: $Q_t(s_t, a) = \frac{1}{n_t} \sum_{s_s \subseteq s_t} Q_s(s_s, a)$, где n_t – число значений оценок $Q_s(s_s, a)$ из подмножества состояния целевого агента.

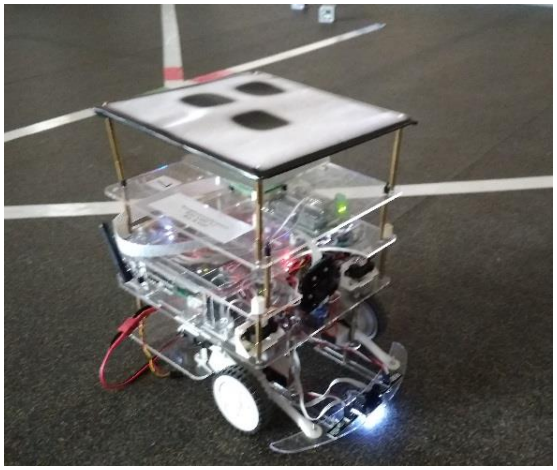
Отдельно необходимо отметить, что подобное представление фактов и правил может быть использовано не только при передаче опыта от одного робота к другому, но и в задаче логического вывода в группе роботов [7]. При этом особенно важно, что в таком случае может существенно снизиться скорость передачи данных между роботами. Это связано как с каналами локальной связи, отличающимися достаточно низкой пропускной способностью, так и со структурой знака, не подходящей для передачи по таким каналам. В указанной в [7] схеме логический вывод осуществляет единственный робот, в то время как остальные только ретранслируют ему необходимые данные. В связи с этим для уменьшения времени логического вывода имеет смысл хранить наиболее часто используемые факты и правила в базах знаний ближайших от него роботов, что сократит общее число ретрансляций. Иными словами, в момент ретрансляции роботы обмениваются фактами или правилами из своих баз знаний. Ближайший к лидеру сохраняет передаваемое лидеру сообщение, передавая соседу, от которого он его получил, один из фактов из своей базы знаний. Таким образом, наиболее часто используемые факты и правила будут постепенно перемещаться ближе к роботу, осуществляющему логический вывод, что сократит общее время этого процесса.

Экспериментальная система

Робот находится на полигоне, оборудованном ArUco-маркерами, по которым он с помощью бортовой камеры определяет свое местоположение, и внешними камерами, также служащими для локализации робота центральной системой контроля эксперимента. Помимо этого, для локализации используется одометрия по энкодерам на колесах робота. Вместе с реальным полигоном в систему интегрирован модуль виртуальной еды (рис. 3, 4), который отвечает за случайное помещение еды на полигон, симуляцию датчика обнаружения этой еды роботом и ее сбор при соответствующей команде робота, находящегося вблизи еды (на той же клетке). На рисунке 3 в системе визуализируется робот в виде модели, зеленым цветом отмечается положение виртуальной еды, в углу показан вид с 6 внешних камер, служащих для локализации робота. На рисунке 4 кубики с маркерами являются моделями еды, которые



а)



б)

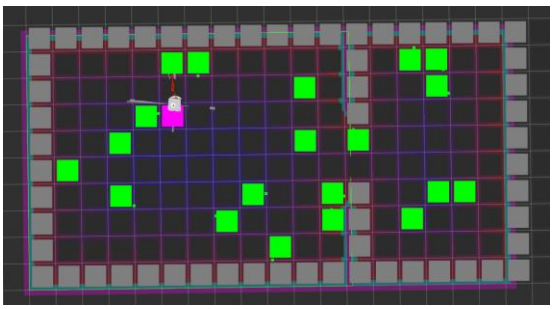
Рис. 3. Вид полигона с точки зрения центральной системы контроля эксперимента (а) и используемый робот модели YARP с маркером для локализации на крыше (б)

Fig. 3. View of the polygon from the point of view of the Central control system of the experiment (a) and the YARP robot used with a marker for localization on the roof (б)

робот может обнаружить собственной камерой, однако для эксперимента используются не они, а модуль, сообщающий роботу положение виртуальной возобновляемой еды.

Система управления робота является двухуровневой. Нижний уровень управления осуществляет сбор сенсорных данных и управление мотор-редукторами. Также он является интерфейсом между верхним уровнем управления и дополнительными контроллерами локальной связи и схвата, связь с которыми осуществляется через протокол I2C. Верхний уровень управления – это микрокомпьютер Raspberry Pi с установленной на нем системой ROS. С помощью предоставляемого ROS инструментария разработана система управления, включающая в себя указанные алгоритмы. Интерфейс между нижним и верхним уров-

а)



б)

Рис. 4. Полигон для фуражирования с ArUco-маркерами и внешними камерами для локализации (а) и то, как видит мир робот (б)

Fig. 4. Foraging polygon with ArUco-markers and external cameras for localization (a) and how the robot sees the world (б)

нями управления осуществляется с помощью UART.

Совместимые интерфейсы управления были реализованы также для симулятора Gazebo [8], поддерживающего моделирование роботов и среды с учетом динамики (трения, моментов силы и т.п.), что необходимо для длительных экспериментов, так как время автономной работы мобильного робота, использованного на реальном полигоне, без подзарядки составляет около часа. При этом еда также моделировалась на уровне виртуальных сенсоров робота, не отображая еду как физический объект, для максимального подобия симуляции и реального полигона (рис. 5). Робот и симуляция показали схожие результаты при качественной оценке работы системы. Виртуальная еда в симуляторе не отображается и присутствует лишь в виде виртуального сенсора робота, отображаемого в визуализации центральной системы контроля эксперимента.

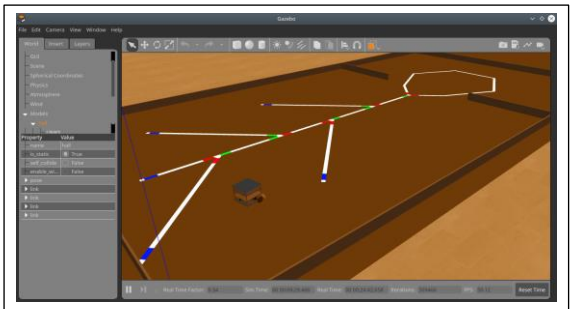


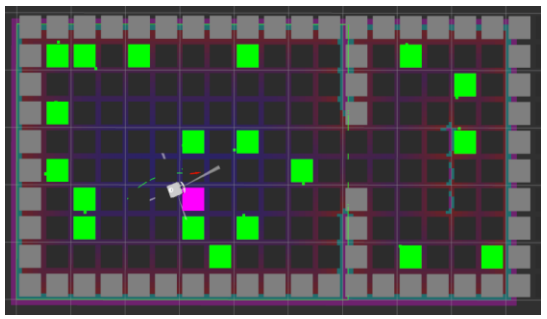
Рис. 5. Модель мира в симуляторе Gazebo с движком физического моделирования

Fig. 5. World model in the Gazebo-simulator with the physical modeling engine

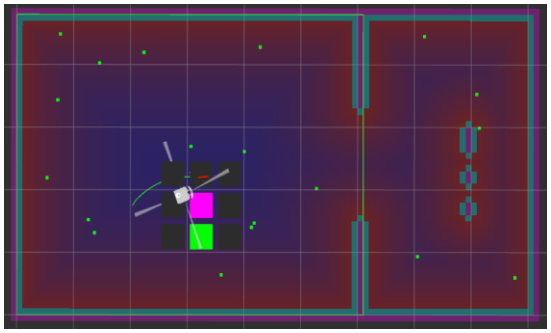
Важно отметить различия модели фуражировки, на которой обучался агент, и фактических интерфейсов агента, на которого переносится обучение. Исходный агент воспринимает мир в виде набора клеток, в которых может располагаться еда или препятствие. В реальной среде координаты агента и еды непрерывны, а препятствия могут представлять собой объекты сложной формы, пересекающие различные клетки среды. В данном случае в качестве сопоставления реальной среды и описанного представления используется процесс дискретизации пространства, в котором клетка помечается препятствием, если в ней содержится хотя бы часть известных препятствий (представленных, в свою очередь, растровым изображением поля с повышенным разрешением по сравнению с размером клеток среды фуражировки), а клетки без препятствий и с едой помечаются как еда. Аналогичным образом отмечается текущее положение робота на поле, однако в силу погрешностей локализации это может приводить к ошибочным действиям: робот может находиться в соседней клетке, но на момент завершения передвижения считать, что уже достиг своей цели. Это обусловливает появление вероятностных эффектов в среде, которые отсутствовали во время предобучения, а значит, агент хуже реагирует на такие ситуации.

Действия агента (движение в одну из соседних клеток) реализуются с помощью навигационного стека ROS с использованием глобального планировщика и локального планировщика на основе DWA (Dynamic Window Approach) [9]. Однако навигационный стек не всегда способен успешно привести робота в требуемую клетку, поскольку, помимо уже упомянутых ошибок локализации, дальномеры могут обнаруживать новые препятствия или шум, которые способны помешать достижению цели после принятия решения о движении в клетку. Это также добавляет недетерминированности в модель мира, используемую роботом. В данном случае настоящими препятствиями являются стенки полигона, нанесенные на статическую карту, а динамические препятствия, детектируемые дальномерами, были отключены для уменьшения недетерминированности среды робота, которая сильно затрудняла принятие решений роботом в соответствии с предобученными оценками ситуаций.

Отладка, сбор данных и общий контроль за ходом эксперимента осуществлялись с помощью централизованной системы, имеющей доступ к данным как робота, так и внешних датчиков (рис. 4), при этом система контроля эксперимента и модуль виртуальной еды формировали полное состояние клеточной среды фуражировки (рис. 6а), в то время как робот для принятия решений получал информацию лишь о локальных объектах (рис. 6б), которые переводились в клеточное представление.



а)



б)

Рис. 6. Полное состояние среды (а) и частичное (б), поступающее на вход роботу и формируемое из его ближайшего окружения

Fig. 6. The full state of the environment (a) and the partial state (b) that is received to the robot's input and formed from its immediate environment

Заключение

Был рассмотрен алгоритм переноса обучения для агента с семиотической моделью мира на задаче фуражирования с реальным роботом и с его приближенной к реальности моделью, предложена реализация на основе сопоставления определенных интерфейсов мобильного робота абстрактным действиям в упрощенной модели, на которой происходило предварительное обучение агента. Агент был основан на алгоритме Q-learning, использующем таблицу оценок состояний-действий, при этом обучение проходило на детерминированной модели в том смысле, что действия всегда приводят к одному и тому же результату в полном описании состояния мира, в то время как для целевого агента переноса обучения данное предположение не выполняется в силу особенностей работы локализации и навигации, а также погрешностей датчиков. Перенесенный опыт показывает возможность работы с предварительным обучением, однако после переноса агент работает хуже из-за отличия в динамике среды. Таким образом, была продемонстрирована возможность применения предложенного ранее алгоритма переноса обучения для семиотических агентов на реальных роботах.

Работа выполнена при частичной финансовой поддержке гранта РФФИ 18-37-00498 мол а.

Литература

1. Осипов Г.С., Чудова Н.В., Панов А.И., Кузнецова Ю.М. Знаковая картина мира субъекта поведения. М.: Физматлит, 2017. 264 с.
2. Konidaris G., Barto A. Building portable options: Skill transfer in reinforcement learning. IJCAI, 2007, pp. 895–900.
3. Rovbo M. Agent-space transfer learning for sign-based world model. Proc. OSTIS-2020, Minsk, 2020, pp. 327–332.
4. Lazaric A. Transfer in reinforcement learning: A framework and a survey. In: Reinforcement Learning. ALO (Wiering M., van Otterlo M., Eds.), 2012, vol. 12, pp. 143–173. DOI: 10.1007/978-3-642-27645-3_5.
5. Kernbach S., Nepomnyashchikh V.A., Kancheva T., Kernbach O. Specialization and generalization of robot behaviour in swarm energy foraging. Mathematical and Computer Modelling of Dynamical Systems, 2012, vol. 18, no. 1, pp. 131–152. DOI: 10.1080/13873954.2011.601421.
6. Yogeswaran M., Ponnambalam S.G. Reinforcement learning: Exploration-exploitation dilemma in multi-agent foraging task. Opsearch, 2012, vol. 49, no. 3, pp. 223–236. DOI: 10.1007/s12597-012-0077-2.
7. Воробьев В.В. Логический вывод и элементы планирования действий в группах роботов // КИИ-2018: 16 конф. по искусственному интеллекту. М., 2018. С. 88–96.
8. Gazebo. URL: http://gazebosim.org (дата обращения: 20.03.2020).
9. Fox D., Burgard W., Thrun S. The dynamic window approach to collision avoidance. IEEE Robotics & Automation Magazine, 1997, vol. 4, no. 1, pp. 23–33. DOI: 10.1109/100.580977.

Software & Systems
DOI: 10.15827/0236-235X.131.413-419

Received 26.03.20
2020, vol. 33, no. 3, pp. 413–419

Application of transfer learning for semiotic models to the foraging problem with real robots

V.V. Vorobev [1], *Deputy Head of Laboratory, eeovsyan@gmail.com*
M.A. Rovbo [1], *Research Engineer, rovboma@gmail.com*

[1] *National Research Centre "Kurchatov Institute", Moscow, 123182, Russian Federation*

Abstract. The paper considers the problem of applying a transfer learning algorithm for agents with semiotic models of the world to the foraging task with real robots. The robot needs to collect randomly placed food items, which when collected appear in a new random place within the polygon. The mobile robot is controlled by an agent with a model of the world that describes sensor readings as predicates.

The agent makes decisions based on a state-action value estimation table for Q-learning. The agent is pre-trained on a simplified model environment with discrete states in which actions are performed with a guaranteed deterministic outcome.

In a real environment and its model, taking into account physics, actions can be performed incorrectly due to a scheduler error, localization errors, and other problems, and the data analysis of sensor information gathered from the continuous world determines the environmental state.

The authors show the implementability of the corresponding interfaces and portability of the concept from a simplified model environment both to its more complete model that takes into account physics and a real robot. The transfer learning application is successful, but the final performance of the agent is reduced (probably due to the incorrect assumption of the determinism of the world in a real environment) and the robot needs additional learning after the transfer. Gazebo was used as a simulator that takes physics into account while the real polygon was equipped with special markers and cameras for localization. The authors also used elements of augmented reality in the form of a virtual food module.

Keywords: reinforcement learning, robotics, foraging, transfer learning, semiotic models.

Acknowledgements. *The work was partially supported by RFBR 18-37-00498 мол_а.*

References

1. Osipov G.S., Shudova N.V., Panov A.I., Kuznetsova Yu.M. *Sign-Based World Model for a Behavior Subject.* Moscow, 2017, 264 p.
2. Konidaris G., Barto A. Building portable options: Skill transfer in reinforcement learning. *IJCAI*, 2007, pp. 895–900.
3. Rovbo M. Agent-space transfer learning for sign-based world model. *Proc. OSTIS-2020*, Minsk, 2020, pp. 327–332.
4. Lazaric A. Transfer in reinforcement learning: a framework and a survey. In*: Reinforcement Learning. ALO* (Wiering M., van Otterlo M., Eds.), 2012, vol. 12, pp. 143–173. DOI: 10.1007/978-3-642-27645-3_5.
5. Kernbach S., Nepomnyashchikh V.A., Kancheva T., Kernbach O. Specialization and generalization of robot behaviour in swarm energy foraging. *Mathematical and Computer Modelling of Dynamical Systems*, 2012, vol. 1, no. 18, pp. 131–152. DOI: 10.1080/13873954.2011.601421.
6. Yogeswaran M., Ponnambalam S.G. Reinforcement learning: Exploration-exploitation dilemma in multi-agent foraging task. *Opsearch*, 2012, vol. 49, no. 3, pp. 223–236. DOI: 10.1007/s12597-012-0077-2.
7. Vorobev V.V. Inference and action planning elements in robot groups. *Proc. 16th. Conf. on Artificial Intelligence RCAI-2018*. Moscow, 2018, pp. 88–96. Available at: http://2018.rncai.ru/program/ (accessed March 20, 2020).
8. *Gazebo*. Available at: http://gazebosim.org (accessed March 20, 2020).
9. Fox D., Burgard W., Thrun S. The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, 1997, vol. 4, no. 1, pp. 23–33. DOI: 10.1109/100.580977.

Для цитирования

Воробьев В.В., Ровбо М.А. Применение передачи обучения в семиотических моделях к проблеме фуражирования с реальными роботами // Программные продукты и системы. 2020. Т. 33. № 3. С. 413–419. DOI: 10.15827/0236-235X.131.413-419.

For citation

Vorobev V.V., Rovbo M.A. Application of transfer learning for semiotic models to the foraging problem with real robots. *Software & Systems*, 2020, vol. 33, no. 3, pp. 413–419 (in Russ.). DOI: 10.15827/0236-235X.131.413-419.