УДК 004.896, 681.5
DOI: 10.15827/0236-235X.130.195-203

Дата подачи статьи: 15.11.19
2020. Т. 33. № 2. С. 195–203

Проблема специализации в иерархических обучающих системах управления на примере задачи фуражировки

Е.Е. Овсянникова¹, инженер-исследователь, eeovsyana@gmail.com

М.А. Ровбо¹, инженер-исследователь, rouboba@gmail.com

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

В статье рассматривается проблема специализации в мультиагентной системе, состоящей из одинаковых, способных к обучению агентов, для которой в поставленной задаче можно выделить слабо связанные подцели.

Была предложена и рассмотрена иерархическая система управления с обучением на задаче фуражировки с различными типами ресурсов, основанная на ϵ -жадном Q-обучении. Перед группой изначально одинаковых агентов стоит задача сбора на базу нескольких типов ресурсов, расположенных в среде, при этом ресурсы требуется собирать в определенной пропорции и награда дается лишь за транспортировку на базу недостающего ресурса.

Предложенная система управления состоит из нескольких уровней. Верхний уровень отвечает за выбор типа ресурса, который агент собирается принести на базу, нижний – за выбор алгоритмов, каждый из которых способен к обучению и определяет навык сбора определенного типа ресурса. Таким образом, в подобной многоагентной системе иерархия в управлении позволяет изначально одинаковым агентам формировать в течение работы разные навыки по сбору ресурсов.

Рассматриваемая гипотеза состоит в том, что, несмотря на отсутствие в задаче предпосылок к оптимизации путем специализации (идеально работающие агенты могли бы собирать разные ресурсы попеременно), возникающая в ходе работы разница в накопленном опыте может сделать специализацию разных агентов на разных ресурсах более эффективной. Результаты были получены с помощью вычислительного эксперимента на компьютерной модели. Эффект специализации наблюдался при определенных условиях. Также была отмечена стабилизация эффективности работы многоагентной системы за счет специализации и введения иерархии в систему управления.

Ключевые слова: обучение с подкреплением, многоагентная система, робототехника, фуражировка, специализация.

Проблема построения группы обучающихся автоматических систем управления рассматривается исследователями с разных сторон [1], включая возможность совместного обучения в группе, передачи опыта обучения из схожих задач, совместного решения задачи без прямого взаимодействия, что свойственно роевым системам.

Алгоритмы управления на основе теории функциональных систем используют иерархические структуры, в которых различные уровни отвечают разной степени детализации ситуации или решаемой подзадачи [2, 3]. То же свойство можно обнаружить и у систем управления на основе семиотических моделей, в которых иерархия проявляется в планировании на уровне семиотической сети, в свою очередь, содержащей описания действий в отдельных узлах [4, 5], а также у некоторых видов алго-

ритмов обучения с подкреплением, показывающих результаты на сложных задачах [6]. Иерархическое построение алгоритмов управления давно известно и применяется в разных формах, однако сейчас можно выделить проблему использования обучения в иерархических системах управления, например, в BDI агентах [7], и интеграции иерархии в обучающиеся алгоритмы, как правило, для ускорения обучения, путем выделения значимых элементов задачи, примером чего являются опции в обобщении обучения с подкреплением [8]. Иерархические алгоритмы управления также могут использоваться для повышения интерпретируемости работы системы [9], для чего применяется автоматическое построение иерархического алгоритма по уже обученному другому алгоритму, при этом эффективность работы в целом сохраняется.

В работе [10] рассматривается задача фуражировки с одним ресурсом и весьма простым механизмом адаптации, в которой агенты учились находить самую эффективную пропорцию в группе между разными, заранее определенными ролями. В этой же работе исследуется более сложная задача – каким образом может возникнуть специализация с разными ролями в изначально гомогенной мультиагентной системе.

В данной статье на относительно простом примере предлагается рассмотреть, каким образом обучающиеся агенты без прямой коммуникации могут повысить эффективность выполнения групповой задачи путем выделения ролей и таким образом избежать затрат, связанных с временем обучения каждым агентом всем необходимым для решения целевой задачи подзадачам.

Перед группой одинаковых в начале работы системы агентов стоит задача сбора на базу нескольких типов ресурсов, расположенных в среде. При этом ресурсы требуется собирать в определенной пропорции (в простейшем случае поровну) и награда дается лишь за принос на базу недостающего ресурса. Особенность рассмотренной проблемы также заключается в том, что агентам необходимо научиться приносить каждый из ресурсов в отдельности, а также тому, что они требуются на базе не всегда.

Постановка задачи

Для исследования эффекта специализации необходимо выбрать задачу, в которой целевой критерий работы системы создает зависимость между действиями различных агентов и существуют разные подцели. В силу направленности исследования на приложения мобильной робототехники задача должна предусматривать перемещение агентов в пространстве и выполнение в нем каких-либо задач, связанных с их положением. Распространенной задачей, на которой можно исследовать группы мобильных агентов, является фуражировка. Ее наиболее известная постановка описывает дискретный мир, в котором агенты перемещаются по прямоугольной сетке и собирают расположенные в клетках ресурсы. При этом некоторые клетки проходимы, а некоторые содержат препятствия. Как правило, по краям мира расположены препятствия, ограничивающие поле. Для удовлетворения указанных выше требований к

постановке задачи необходимо сделать следующие уточнения:

- одна из свободных клеток помечена как гнездо или база – это место, куда агенты должны приносить ресурсы;
- агенты могут перемещаться на одну клетку вверх, вниз, влево и вправо, но не по диагонали;
- существуют фиксированные в пространстве источники нескольких типов ресурсов, и прохождение по ним агентов позволяет собрать этот ресурс (он возобновляется в источниках, то есть всегда там есть);
- агент может нести только один тип ресурса за раз и автоматически оставляет его в гнезде при передвижении на клетку гнезда;
- целью группы является сбор ресурсов в гнездо в равной пропорции, а соответствующая награда выдается агенту только за «нужный» ресурс, то есть такой, которого в гнезде в данный момент меньше (когда их поровну, награда дается за любой ресурс).

Такая постановка позволяет предположить, что эффективным способом сбора ресурсов может быть сбор одним агентом одного ресурса, а другим – другого (при прочих равных условиях, для чего ресурсы стоит сделать одинаково доступными), то есть специализация. Собирать разные ресурсы каждым из агентов может быть невыгодно в силу того, что необходимо обучиться их сбору, поэтому, предположительно, система будет работать эффективнее, когда каждый из агентов будет обучаться собирать один из ресурсов. Однако эффективность такого поведения агентов система должна обнаружить самостоятельно.

Агент принимает на вход локальную информацию: координаты клетки, в которой он находится, нахождение в ней гнезда, статус того, несет ли он ресурс определенного типа, а также награды за принесение ресурса определенного типа и за принесение требуемого типа ресурса в гнездо.

Система управления

Существует множество алгоритмов управления, способных к обучению, включая групповое обучение [11]. Для исследования в данном случае были выбраны более простые алгоритмы: ϵ -жадное Q-обучение и описываемый в данной статье иерархический алгоритм, состоящий из нескольких алгоритмов Q-обучения (один из них выбирает подцель, которую дол-

жен пытаться выполнить агент, а остальные отвечают за решение конкретной подцели). Это позволяет избежать влияния дополнительных эффектов, присущих алгоритмам группового обучения.

Алгоритм ϵ -жадного Q-обучения решает задачу оптимизации целевого критерия $G(S) \doteq \sum_{t=0}^{\infty} \gamma^t r_t$, где r_t – награды, получаемые агентом

на шаге t ; $\gamma \leq 1$ – коэффициент, влияющий на предпочтение ближайших по времени наград по отношению к удаленным в будущее; S – текущее состояние (в момент времени $t = 0$) агента. Алгоритм итеративно меняет стратегию выбора действий, то есть функцию выбора действия α , $\pi: S \rightarrow \alpha$, путем обновления оценки эффективности соответствующей пары состояние–действие, постепенно согласуя ее с реальными наградами в соответствии с формулой $Q(S, a) \leftarrow Q(S, a) + \alpha(r + \gamma \max_{a'} Q(S', a') - Q(S, a))$, где $Q(S, a)$ – функция оценки эффективности выбора действия a в состоянии S (штрих обозначает следующий момент времени); r – награда за последний шаг.

Выбор действия осуществляется путем нахождения максимально эффективного в текущей ситуации согласно Q . С некоторой вероятностью ϵ действие выбирается среди всех равновероятно доступных для накопления опыта и обнаружения более эффективных стратегий.

Однако в рассматриваемой задаче агенты не обладают полной информацией о меняющемся состоянии мира: не знают статуса другого агента и количества ресурсов в гнезде. Это усиливает зависимость результата действий одного агента от другого и, вообще говоря, нарушает обычные предположения о допустимости представления задачи, решаемой агентом, как марковского процесса принятия решений. Агент оказывается в среде, в которой одни и те же последовательности действий могут менять свою эффективность (в статистическом смысле получаемую награду) в течение времени из-за действий другого агента. Более того, классический алгоритм Q-обучения в таком представлении задачи не сможет различить ситуацию, в которой агент приносит один ресурс и получает награду, и ситуацию, когда тот же самый ресурс нет смысла нести на базу. Это вынуждает агентов постоянно переучиваться или специализироваться. По этой же причине параметр случайного выбора действий ϵ в данном случае выбран неубывающим, как обычно при-

нято для обеспечения теоретической сходимости.

Другой предлагаемый в данной статье алгоритм представляет собой иерархию из мета-алгоритма, выбирающего из действий «принести ресурс А» и «принести ресурс В». Мета-алгоритм принимает на вход только награду за предыдущую операцию и решение о новом действии, когда агент находится в гнезде, из подчиненных ему алгоритмов – навыков, которые работают так, как описано выше, но вместо награды за принесение требуемого ресурса в гнездо получают награду за принесение ассоциированного с ними ресурса. Например, один навык отвечает за доставку ресурса А, другой – за доставку ресурса В. Таким образом, нижний уровень учится выполнять подзадачи сбора конкретных типов ресурсов, а верхний отвечает за оптимизацию принятия решений о необходимом ресурсе. Все алгоритмы являются ϵ -жадным Q-обучением.

Численное моделирование

Рассматриваемая среда и ее начальное состояние изображены на рисунке 1, где препятствия по краям обозначены светло-серыми квадратами, в центре находится гнездо, обозначенное оранжевым, в противоположных углах два источника разных типов ресурсов (круги багрового и оливкового цветов), агенты обозначены кругами розового цвета различной яркости.

Среда представляет собой квадрат размером 7 на 7 с расположенным в центре гнездом и симметрично расставленными по краям ресурсами разных типов. Всего присутствуют два агента и два ресурса. Это обеспечивает равную сложность доставки обоих ресурсов агентами на базу, однако обучаться доставке каждого ресурса приходится отдельно. Агенты стартуют

Рис. 1. Рассматриваемая среда задачи (слева) и ее начальное состояние (справа)

Fig. 1. The problem environment under consideration (left) and its starting state (right)

из гнезда с количеством каждого ресурса в гнезде, равным 0. За успешную доставку ресурса агент получает награду 1 за соответствующий ресурс для навыка и целевую награду 1 за выполнение цели, если был доставлен недостающий ресурс.

Результаты

На рисунках 2 и 3 приведены результаты работы группы агентов с классическим алгоритмом Q-обучения, без иерархии с навыками, для

одного эксперимента. При этом использовались параметры скорости обучения $\alpha = 0.1$, важности будущих наград $\gamma = 0.1$, вероятности случайного действия $\epsilon = 0.1$. Здесь и далее график эффективности показывает суммарные награды за доставку требуемого ресурса в гнездо группой агентов по эпизодам (каждый эпизод состоит из 1 000 шагов). Эксперименты проводились по 200 000 шагов с двумя агентами. Некоторые результаты показаны для одного эксперимента, другие усреднены для каждого эпизода по десяти экспериментам.

Рис. 2. Эффективность выполнения задачи группой агентов с классическим Q-обучением для $\gamma = 0.1$

Fig. 2. The problem solution efficiency by a group of agents with classical Q-learning for $\gamma = 0.1$

Рис. 3. Динамика сбора ресурсов и навыков агентов с классическим Q-обучением для $\gamma = 0.1$

Fig. 3. The dynamics of resource collection and agent skills with classical Q-learning for $\gamma = 0.1$

Графики на рисунке 3 показывают (сверху вниз) разность между количеством ресурса А и ресурса В в гнезде, награды первого агента и награды второго агента. В тех случаях, когда рисунок динамики состоит из четырех графиков, первый из них показывает разность ресурсов в гнезде, а второй – коэффициент специализации, представляющий собой разность между количеством ресурсов А и В, которые соответствующий агент принес за эпизод.

Несмотря на улучшение в результате обучения эффективности работы группы, наблюдается нестабильное поведение системы, при котором агенты собирают сначала один ресурс с избытком, не получая за него награды, после чего компенсируют его интенсивным сбором другого ресурса, за который половину времени получают награду, после чего набирают уже его больше, чем необходимо, что вызывает большие амплитуды колебаний в эффективности.

Усредненные результаты эксперимента (<http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/15.jpg>, <http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/16.jpg>) с этой группой агентов позволяют увидеть, что явление сохраняется, хотя и с некоторым смещением по времени, а средняя эффективность сохраняется на уровне около 50.

В рассматриваемых случаях специализация не наблюдается: оба агента продолжают вести себя схожим образом на протяжении всего эксперимента. Однако стабилизация системы для данной задачи может наблюдаться при параметре $\gamma = 0.9$, что в какой-то момент приводит к прекращению колебаний и выходу работы системы в устойчивое состояние, в котором у агентов различная скорость сбора двух типов ресурсов, причем их сумма сохраняется (см. <http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/9.jpg>, <http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/10.jpg>). Таким образом, наблюдается специализация агентов (хотя и неполная, то есть они продолжают переключаться на общие подзадачи).

Введение иерархии в систему управления позволяет добиться быстрой стабилизации работы системы (рис. 4, 5), однако при этом специализация наблюдается лишь кратковременно, пока количество ресурсов на базе не выравнилось. Можно предположить, что основной причиной является то, что агенты продолжительное время учатся решать подзадачи, а к моменту, когда начинают успешно их выполнять, практически одинаково хорошо владеют

обоими навыками. Система остается гомогенной, и оптимальным режимом работы является эффективное выполнение подзадач обоими агентами по мере надобности.

Чтобы на данной задаче пронаблюдать эффект специализации, можно искусственно добавить период в начале работы системы, когда каждый агент может обучаться лишь одной из подзадач (далее это называется гандикапом). В данном случае первый агент мог обучаться некоторое время лишь сбору ресурса В, а второй – ресурса А. Тогда оптимальным режимом работы системы является сбор каждым агентом ресурса, который тот лучше всего научился собирать. От агентов требуется обнаружить этот режим работы и следовать ему. Более всего данный эффект проявляется (см. <http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/11.jpg>, <http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/12.jpg>) при низких значениях параметра γ у алгоритмов обучения навыкам, например 0.1. В приведенных экспериментах гандикап длился 50 000 тактов (50 эпизодов).

Стоит отметить, что разделение на роли происходит в такой постановке задачи в начале стабилизации количества ресурсов на базе и сохраняется длительное время после. Через какое-то время система снова становится гомогенной, поскольку в используемых алгоритмах коэффициент ϵ , отвечающий за случайное поведение, не уменьшается со временем и случайный выбор действий вопреки обнаруженной эффективной стратегии приводит к тому, что оба агента обучаются обоим навыкам в совершенстве и оптимальный режим работы группы агентов снова становится таким, когда оба выполняют обе подзадачи по мере необходимости. Указанное распределение на роли систематически возникает в ряде экспериментов (<http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/17.jpg>, <http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/18.jpg>), хотя и несколько смещено по моменту возникновения.

Отдельно можно отметить иногда возникающую специализацию в постановке задачи без гандикапа (на рисунке 6 это период с 50-го по 60-й эпизоды эксперимента, а также кратковременные меняющиеся периоды специализации в следующих эпизодах).

Обсуждение результатов

Значение γ в алгоритмах навыков для наблюдения специализации было выбрано низкое (0.05), что в некотором смысле делает за-

Рис. 4. Эффективность выполнения задачи группой агентов с иерархией алгоритмов Q -обучения для $\gamma = 0.9$, усредненная по экспериментам

Fig. 4. The problem solution efficiency by a group of agents with a Q -learning algorithm hierarchy for $\gamma = 0.9$, averaged over experiments

Рис. 5. Динамика сбора ресурсов и навыков агентов с иерархией алгоритмов Q -обучения для $\gamma = 0.9$, усредненная по экспериментам

Fig. 5. The dynamics of resource collection and agent skills with a Q -learning algorithm hierarchy for $\gamma = 0.9$, averaged over experiments

дачу приобретения навыка более сложной (то есть достижения стратегии, близкой к оптимальной). Это несколько компенсирует проблему, когда в выбранной для рассмотрения задаче агенты могут весьма быстро обучиться обоим навыкам и специализация перестанет иметь смысл с точки зрения оптимизации решаемой группой агентов задачи.

Работа была направлена на исследование гипотезы о возможности возникновения эффекта специализации в групповых системах из гомогенных агентов, способных к обучению,

а также о полезности данного эффекта для повышения эффективности работы системы (то есть достижения цели группы).

Несмотря на постановку задачи, выбранную таким образом, чтобы специализация агентов имела смысл и проявлялась в отрыве от других эффектов, но при этом сохраняла особенности работы реальных групп роботов (мобильность, работа в пространстве, неполное описание текущего состояния, доступное агенту, и т.п.), дальнейший анализ показал, что специализация агентов в данной задаче может наблю-

даться лишь временно при возникновении временного, но большого преимущества, позволяющего агентам обучиться разным навыкам, а асимптотическое и более эффективное поведение группы заключается в состоянии, когда оба агента способны одинаково хорошо (и оптимально) выполнять каждую подзадачу (связанную с навыком). Из этого можно сделать вывод, что эффект специализации может иметь большую важность, когда агентам тяжело (долго или по каким-то причинам нежелательно) обучаться нескольким или даже всем навыкам сразу, а также существуют предпосылки для того, чтобы агент, выбрав навык или их подмножество, значительно продвинулся в их изучении перед получением возможности переключиться на другой. Последнее может обеспечиваться либо напрямую алгоритмом обучения (использованные в данной статье, к примеру, для этого не подходят, поскольку склонны к относительно частому переключению действий), либо спецификой рассматриваемой задачи, к примеру, если выбранное агентом первое действие вынуждает его выполнить ассоциированную с ней подзадачу перед тем, как он вернется в исходное состояние.

В процессе исследования поставленной проблемы и соответствующих алгоритмов были выявлены некоторые другие эффекты, а также трудности и возможные подходы к их решению, в частности: нестабильная эффективность работы группы обучающихся агентов,

стабилизирующее влияние иерархического управления, слишком быстрое обучение всем навыкам в задаче, приводящее к гомогенизации группы.

Эффективность выполнения целевой задачи группой обучающихся агентов наблюдалась в экспериментах (рис. 2 и 3) из-за необходимости переобучаться в силу нестационарности задачи, обусловленной требованием сбора разных ресурсов, частичным знанием состояния среды и влиянием агентов на эффективность работы друг друга. Возникающие эффекты специализации и иерархический алгоритм стабилизируют систему. В случае с иерархической системой, в которой агенты имели возможность научиться решать подзадачи сбора обоих ресурсов, случайность принятия решений в итоге приводит к идеально решающим подзадачи агентам, что убирает выгодность специализации и приводит к эффективно решающей задачу гомогенной группе.

Заключение

В статье была предложена и рассмотрена постановка задачи фуражировки с несколькими ресурсами для группы агентов, требующая выполнения различных подзадач. Исследована иерархическая система управления, основанная на мета-алгоритме, выбирающем подцель для решения, и нескольких алгоритмах нижнего уровня, обучающихся решать конкретную под-

задачу. Наблюдались эффекты неустойчивой работы системы, в которой эффективность решения задачи имела сильные колебания. Эффект специализации, проявляющийся в системе для определенных параметров при классическом алгоритме Q-обучения и для иерархической системы управления из нескольких таких алгоритмов, приводил к стабилизации системы. Анализ результатов показывает также, что выбранная

задача не очень хорошо демонстрирует преимущества специализации по ряду причин, включая простоту обучения подзадачам агентов и наличие оптимального поведения группы, когда все агенты работают по одной и той же логике и фактически группа становится гомогенной. Эффект, возможно, имеет большее значение для более сложных задач, что является предметом дальнейших исследований.

Работа выполнена при частичной финансовой поддержке РФФИ, грант № 16-29-04412 офи_м.

Литература

1. Lopes Silva M.A., Ricardo de Souza S., Souza M.J.F., Felizardo de França Filho M. Hybrid metaheuristics and multi-agent systems for solving optimization problems: A review of frameworks and a comparative analysis. *Appl. Soft Comput. J.*, 2018, vol. 71, pp. 433–459.
2. Демин А.В., Витяев Е.Е. Логическая модель адаптивной системы управления // *Нейроинформатика*. 2008. Т. 3. № 1. С. 79–107.
3. Vorobev V., Rovbo M. Analysis of semantic probabilistic inference control method in multiagent foraging task. *Proc. Conf. OSTIS, Minsk*, 2019, pp. 237–242.
4. Ровбо М.А., Сорокоумов П.С. Архитектура системы управления интеллектуальным агентом на основе семиотической сети // *Открытое образование*. 2018. Т. 22. № 5. С. 84–93.
5. Kiselev G.A., Panov A.I. Sign-based approach to the task of role distribution in the coalition of cognitive agents. *Proc. SPIRAS*, 2018, vol. 2, no. 57, pp. 161–187.
6. Kulkarni T.D., Saeedi A., Narasimhan K.R., Tenenbaum J.B. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation. 2016. URL: <https://arxiv.org/abs/1604.06057> (дата обращения: 10.11.2019).
7. Singh D., Sardina S., Padgham L., James G. Integrating learning into a BDI agent for environments with changing dynamics. *Proc. IJCAI*, 2011, vol. 3, pp. 2525–2530.
8. Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. Cambridge. MIT Press, 2018, 552 p.
9. Brown A., Petrik M. Interpretable reinforcement learning with ensemble methods. 2018. URL: arxiv.org/pdf/1809.06995 (дата обращения: 10.11.2019).
10. Ровбо М.А. Распределение ролей в гетерогенном муравьино-подобном коллективе // *КИИ-2016: сб. матер. конф.* Смоленск, 2016. Т. 2. С. 363–371.
11. Guo H., Meng Y. Distributed reinforcement learning for coordinate multi-robot foraging. *J. Intell. Robot. Syst. Theory Appl.*, 2010, vol. 60, no. 3–4, pp. 531–551.

Software & Systems
DOI: 10.15827/0236-235X.130.195-203

Received 15.11.19
2020, vol. 33, no. 2, pp. 195–203

The specialization problem in hierarchical learning control systems for the foraging problem

*E.E. Ovsyannikova*¹, *Research Engineer, eeovsyan@gmail.com*
*M.A. Rovbo*¹, *Research Engineer, rovboma@gmail.com*

¹*National Research Centre “Kurchatov Insitute”, Moscow, 123182, Russian Federation*

Abstract. The paper considers the specialization problem in a multi-agent system that consists of identical agents capable of learning that also has weakly connected sub-goals in the problem.

The authors propose and consider a hierarchical control system, which is based on ϵ -greedy Q-learning, with learning on the foraging problem with different types of resources. A group of initially identical agents face the problem of collecting several types of environmental resources and bringing them to a base. The resources must be collected in a certain proportion and the reward is only for transporting the missing resource to the base.

The proposed control system has several levels. The upper level is responsible for choosing the resource type that the agent is going to bring to the base. The lower one selects the learning algorithms and determines the skill of collecting a certain type of resource. Thus, in such multi-agent system, the hierarchy in control allows initially identical agents to form different skills of collecting resources during their work.

The hypothesis under consideration is that, despite the problem does not have prerequisites for optimization through specialization (ideally working agents could collect different resources alternately), the difference in experience gained during operation can make more efficient specialization of the different agents on different resources. The results were obtained using a computational experiment on a computer model. The effect of specialization was observed under certain conditions. Also, specialization and the introduction of a hierarchy in the control system have notably stabilized the performance efficiency of the multi-agent system.

Keywords: reinforcement learning, multi-agent system, robotics, foraging, specialization.

Acknowledgements. The work was partially supported by RFBR, grant no. 16-29-04412 офу_м.

References

1. Lopes Silva M.A., Ricardo de Souza S., Souza M.J.F., Felizardo de França Filho M. Hybrid metaheuristics and multi-agent systems for solving optimization problems: A review of frameworks and a comparative analysis. *Appl. Soft Comput. J.*, 2018, vol. 71, pp. 433–459.
2. Demin A.V., Vityaev E.E. Logical model of the adaptive control system. *Neuroinformatics*, 2008, vol. 3, no. 1, pp. 79–107 (in Russ.).
3. Vorobev V., Rovbo M. Analysis of semantic probabilistic inference control method in multiagent foraging task. *Proc. Conf. OSTIS*, Minsk, 2019, pp. 237–242.
4. Rovbo M.A., Sorokoumov P.S. Control system architecture of an intelligent agent based on a semiotic network. *Open Education*, 2018, vol. 22, no. 5, pp. 84–93 (in Russ.).
5. Kiselev G.A., Panov A.I. Sign-based approach to the task of role distribution in the coalition of cognitive agents. *Proc. SPIIRAS*, 2018, vol. 2, no. 57, pp. 161–187.
6. Kulkarni T.D., Saeedi A., Narasimhan K.R., Tenenbaum J.B. *Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation*. 2016. Available at: <https://arxiv.org/abs/1604.06057> (accessed November 10, 2019).
7. Singh D., Sardina S., Padgham L., James G. Integrating learning into a BDI agent for environments with changing dynamics. *Proc. IJCAI*, 2011, vol. 3, pp. 2525–2530.
8. Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. Cambridge, MIT Press, 2018, 552 p.
9. Brown A., Petrik M. *Interpretable Reinforcement Learning with Ensemble Methods*. 2018. Available at: arxiv.org/pdf/1809.06995 (accessed November 10, 2019).
10. Rovbo M.A. Role assignment in a heterogeneous ant-like collective. *Proc. 15th CAI*, Smolensk, 2016, vol. 2, pp. 363–371 (in Russ.).
11. Guo H., Meng Y. Distributed reinforcement learning for coordinate multi-robot foraging. *J. Intell. Robot. Syst. Theory Appl.*, 2010, vol. 60, no. 3–4, pp. 531–551.

Для цитирования

Овсянникова Е.Е., Ровбо М.А. Проблема специализации в иерархических обучающихся системах управления на примере задачи фуражировки // Программные продукты и системы. 2020. Т. 33. № 2. С. 195–203. DOI: 10.15827/0236-235X.130.195-203.

For citation

Ovsyannikova E.E., Rovbo M.A. The specialization problem in hierarchical learning control systems for the foraging problem. *Software & Systems*, 2020, vol. 33, no. 2, pp. 195–203 (in Russ.). DOI: 10.15827/0236-235X.130.195-203.