УДК 007:519.816
DOI: 10.15827/0236-235X.117.028-033
Дата подачи статьи: 17.11.16
2017. Т. 30. № 1. С. 28–33

РЕАЛИЗАЦИЯ МЕТОДОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ НА ОСНОВЕ ТЕМПОРАЛЬНЫХ РАЗЛИЧИЙ И МУЛЬТИАГЕНТНОГО ПОДХОДА ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ РЕАЛЬНОГО ВРЕМЕНИ

А.П. Еремеев, д.т.н., профессор, eremeev@appmat.ru;
А.А. Кожухов, аспирант, saaanchezzz@yandex.ru
(Национальный исследовательский университет «Московский энергетический институт», ул. Красноказарменная, 14, г. Москва, 111250, Россия)

В работе описана реализация методов обучения с подкреплением на основе временных (темпоральных) различий и мультиагентной технологии. Рассмотрены возможности комбинирования методов обучения со статистическими и экспертными методами прогнозирования с целью последующей интеграции в инструментальную программную среду для использования в современных перспективных интеллектуальных системах реального времени типа интеллектуальных систем поддержки принятия решений реального времени.

Даны анализ методов обучения с подкреплением (RL-обучения) в плане использования в интеллектуальных системах реального времени, их основные компоненты, преимущества и решаемые задачи. Основное внимание уделено методам RL-обучения на основе временных (темпоральных) различий (TD-методам), разработаны соответствующие алгоритмы. Рассмотрены возможности включения методов RL-обучения в мультиагентную среду и их комбинирования со статистическими и экспертными методами прогнозирования с целью последующей интеграции в инструментальную среду для использования в интеллектуальных системах реального времени типа интеллектуальных систем поддержки принятия решений реального времени для управления и диагностики сложных технических объектов.

Разработана архитектура прототипа подсистемы прогнозирования, включающая эмулятор, моделирующий состояние проблемной области (объекта и внешнего окружения), и модули прогнозирования, анализа и принятия решений, RL-обучения. Выполнена программная реализация прототипа подсистемы прогнозирования с применением мультиагентного подхода для решения задачи экспертного диагностирования сложного технического объекта.

Результаты тестирования и апробации разработанной системы показали ее достаточную эффективность и целесообразность включения в состав современных интеллектуальных систем поддержки принятия решений реального времени.

***Ключевые слова**: искусственный интеллект, интеллектуальная система, реальное время, обучение с подкреплением, прогнозирование, поддержка принятия решений, программное средство.*

Методы обучения с подкреплением (reinforcement learning, RL) [1], основанные на использовании большого количества информации для обучения в произвольной окружающей среде, являются одной из наиболее активно развиваемых областей искусственного интеллекта, связанных с разработкой перспективных *интеллектуальных систем реального времени* (ИС РВ), типичными примерами которых являются *интеллектуальные системы поддержки принятия решений реального времени* (ИСППР РВ) [2, 3].

Одним из наиболее перспективных в плане использования в ИС РВ, относящихся к классу динамических интеллектуальных систем [4, 5], является обучение на основе темпоральных различий (temporal-difference, TD) [1], когда процесс обучения основывается непосредственно на получаемом опыте без предварительных знаний о модели поведения окружающей среды. Ключевой особенностью TD-алгоритмов является обучение на основе различий во временных последовательных предсказаниях. TD-методы, предназначенные для многомерных временных рядов, способны обновлять расчетные оценки, основанные в том числе и на других полученных оценках, не дожидаясь окончательного результата, то есть являются самонастраиваемыми. Последнее свойство весьма важно для ИС семиотического типа, способных адаптироваться (подстраиваться) к изменениям в управляемом объекте и/или окружающей среде [3].

Использование мультиагентного подхода в динамических ИС, в том числе ИС РВ (ИСППР РВ), системах распределенного управления и системах интеллектуального анализа данных, способного улучшить эффективность и надежность таких систем, является быстроразвивающимся и перспективным подходом [6].

При разработке современных ИС РВ большое внимание должно быть уделено также средствам прогнозирования развития ситуации на объекте и последствий принимаемых решений, экспертным методам и средствам обучения [3, 7]. Эти средства необходимы для модификации и адаптации ИС РВ к изменениям на объекте и во внешней среде, а также для расширения области применения и улучшения эффективности функционирования систем.

Далее будет дан анализ ряда методов обучения с подкреплением, в частности TD-методов, в плане их последующей интеграции в инструментальную среду для ИС РВ типа ИСППР РВ с применением мультиагентного подхода.

Методы обучения с подкреплением

Будем предполагать, что неопределенность поступающей в БД ИС РВ информации о текущем состоянии проблемной области, объекта и окружающей среды связана в основном с ошибочной работой датчиков (сенсоров) или ошибками соответствующего оперативно-диспетчерского персонала (лиц, принимающих решения, ЛПР). В функции RL-обучения входит адаптация немарковской модели принятия решений к сложившейся ситуации за счет анализа предыстории процесса принятия решений, вследствие чего повышается качество принимаемых решений [1, 8, 9].

В RL-обучении модуль принятия решений, способный посредством взаимодействия с внешней средой и анализа оценочной функции (функции платежа) корректировать стратегию принятия решений, называется *агентом*. Задачей агента является нахождение в процессе обучения оптимальной (для марковского процесса) или приемлемой (если процесс не является марковским) стратегии принятия решений, называемой также *политикой*. Интеллектуальный агент должен уметь поддерживать несколько путей обучения и адаптировать накопленный опыт к изменениям в окружающей среде. В RL-обучении взаимодействие «агент–окружающая среда» моделируется посредством контроллера, связывающего ИС и среду. Процесс восприятия отображает состояния среды (проблемной области) во внутренние представления агента, а процесс воздействия отображает предлагаемые агентом воздействия в действия (преобразования) внешней среды. Обобщенная схема взаимодействия «агент–окружающая среда» приведена на рисунке 1.

Целью RL-обучения является максимизация ожидаемой выгоды R_t, которая определяется как некоторая функция, заданная на последовательности вознаграждений: $R_t = r_{t+1} + r_{t+1} + \ldots + r_{t+T}$, где T – завершающий временной шаг; r_t – вознаграждение на временном шаге t. Данный подход применим в прикладных задачах, когда завершающий шаг можно определить естественным образом исходя из природы решаемой задачи, то есть когда взаимодействие «агент–окружающая среда» можно разбить на последовательности, называемые *эпизодами*.

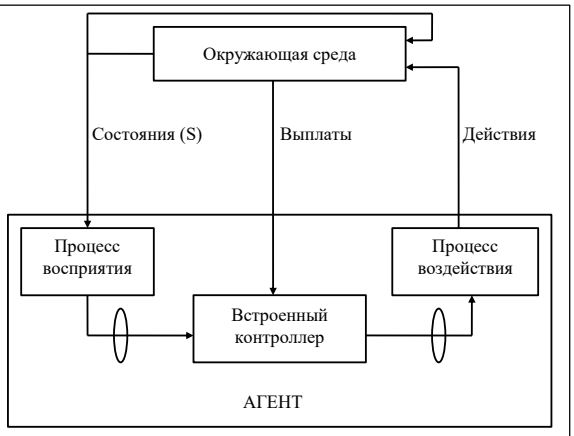

Рис. 1. Схема взаимодействия «агент–среда»

Рис. 1. "Agent–environment" interaction scheme

Основной проблемой RL-обучения является нахождение агентом компромисса между изучением и применением. Для получения большего вознаграждения агент должен предпочитать действия, ранее уже применявшиеся и показавшие свою эффективность с точки зрения получения поощрения. С другой стороны, чтобы обнаруживать такие действия, агенту необходимо пробовать выполнять новые действия. Таким образом, агент должен как применять уже известные действия, так и изучать новые для возможности иметь наилучший выбор в будущем. Важной характеристикой RL-обучения является получение отложенных вознаграждений, которые имеют место в сложных динамических системах. Это означает, что действие агента может повлиять не только на текущую награду, но и на все последующие.

В плане применения в ИС РВ TD-методы могут решать несколько задач: *задачу предсказания* значений некоторых переменных в течение нескольких временных шагов и *задачу управления*, основанную на RL-обучении агента тому, как влиять на окружающую среду. Таким образом, агент должен предсказывать последующие состояния окружающей среды и использовать эти значения для ее изменения с целью максимизации вознаграждений.

Для возможности обучения и адаптации к изменениям внешней среды агент должен обладать памятью для хранения предыстории. При этом возникает ряд проблем, связанных с объемом доступной агенту информации о прошлом, с запоминанием, хранением, использованием доступной информации и т.д. Для решения этих проблем агент может использовать скользящее окно для истории (наиболее простой метод) или строить зависящую от состояния прогнозную модель окружающей среды. Можно применить комбинацию этих подходов, когда агент анализирует чувствительную к предыстории политику непосредственно при обучении.

Несмотря на проблему поиска компромисса между изучением и применением, RL-обучение имеет ряд важных достоинств для применения в ИС РВ типа ИСППР РВ:

– использование простой обратной связи на основе скалярных платежей;

– поддержка режима оперативного реагирования, когда агенту необходимо быстро адаптироваться к изменениям внешней среды;

– интерактивность и возможность изменения (пополнения) анализируемых данных (предыстории);

– действенность в недетерминированных средах;

– эффективность в сочетании с темпоральными моделями для задач нахождения последовательных решений;

– открытость к модификации и сравнительная простота включения в интеллектуальные системы различного назначения (планирования, управления, обучения и т.д.).

Методы обучения с подкреплением на основе темпоральных различий

Рассмотрим RL-методы на основе темпоральных различий (TD-методы) в плане их применения в ИС РВ [1, 10]. TD-методы для решения задачи предсказания используют имеющийся опыт. При наличии некоторого опыта следования избранной стратегии TD-методы корректируют свои оценки, например, если имеет место посещение нетерминального состояния S_t в момент времени t, то корректируются оценки $V(S_t)$, основываясь на том, что случилось после этого посещения, то есть для корректировки оценки необходимо дождаться только следующего временного шага. Непосредственно в момент времени $t+1$ формируется целевое значение оценки и производится необходимая корректировка с учетом уже имеющегося вознаграждения r_{t+1} и оценки $V(S_{t+1})$.

Для наиболее простого TD-метода (метода TD(0)) справедливо $V_{St} \leftarrow V(S_t) + \alpha[r_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$, где γ – ценность терминального состояния. При корректировке целью будет величина $r_{t+1} + \gamma V(S_{t+1})$. Так как в своих корректировках TD-метод частично основывается на существующих оценках, он является *самонастраивающимся.* Одним из преимуществ TD-методов является то, что они не требуют знания модели окружающей среды с ее вознаграждениями и вероятностным распределением последующих состояний.

Следует отметить, что при обучении на основе вероятностного метода Монте-Карло каждый раз необходимо ждать завершения эпизода, так как только тогда становится известной оценка (выгода), в то время как при использовании TD-методов необходимо дождаться лишь следующего временного шага. Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ, так как в некоторых ситуациях эпизоды могут быть настолько продолжительными, что задержки процесса обучения, связанные с необходимостью завершения эпизодов, будут слишком велики. Возможны также ситуации, когда возникают непрерывные задачи, а эпизоды как таковые отсутствуют.

TD-методы обучаются на основе каждого перехода вне зависимости от осуществляемых в дальнейшем действий и, соответственно, не чувствительны к ситуациям, когда необходимо игнорировать эпизоды или снижать значимость эпизодов, в которых предпринимаются экспериментальные действия, что может сильно замедлить обучение. TD-методы в целом можно разделить на две основные категории – методы с интегрированной (on-policy) и методы с разделенной (off-policy) оценкой ценности стратегий. В методах с интегрированной оценкой стратегия, используемая для управления, аналогична оценочной стратегии, которая совершенствуется во время обучения. В методах с разделенной оценкой стратегия управления не имеет взаимосвязи с оценочной стратегией.

Метод SARSA – TD-метод с интегрированной оценкой ценности стратегий. Для данного метода необходимо оценить функцию $Q^{\pi}(s, a)$ для текущей стратегии π и для всех состояний s и действий a: $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1}+\gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$, где α – постоянная длина шага; γ – ценность терминального состояния. Данная корректировка имеет место после каждого перехода из нетерминального состояния s_t. Если состояние s_{t+1} является терминальным (заключительным), то значение $Q(s_{t+1}, a_{t+1})$ полагается равным нулю. Это правило использует каждый элемент из пятерки $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$, когда происходит переход от одной пары «состояние–действие» к другой. В методах с интегрированной оценкой ценности стратегий все время оценивается функция Q^{π} для стратегии поведения π и в то же время стратегия π делается более жадной по отношению к Q^{π}.

Свойство сходимости алгоритма SARSA непосредственно связано с зависимостью стратегии от функции Q. Например, можно использовать ε-жадную и ε-гибкую стратегии. Алгоритм сходится с вероятностью 1 к оптимальной стратегии и функции ценности действия при условии, что все пары «состояние–действие» посещались бесконечное число раз, и стратегия сходится в пределе к жадной стратегии (что может быть осуществлено, например, при помощи ε-жадных стратегий с $\varepsilon = 1/t$).

Метод Q-обучения – метод с разделенной оценкой ценности стратегий, который находит оптимальные значения функции Q для выбора последующих действий и одновременно определяет оптимальную стратегию. Аналогично методу TD(0) в каждой итерации есть только знание о двух состояниях: s и одного из его предшествующих. Таким образом, значения функции Q позволяют получить некоторое представление о будущем качестве действий в предшествующих состояниях и сделать задачу выбора действия проще.

Для данного метода нужно оценить функцию ценности действия $Q^{\pi}(s, a)$ для текущей стратегии π и для всех состояний s и действий a, где эпизод состоит из последовательности перемежающихся состояний и пар «состояние–действие». Одношаговое Q-обучение характеризуется зависимостью $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max\alpha Q(s_{t+1}, a) - Q(s_t, a_t)]$, где α – постоянная длина шага; γ – ценность терминального состояния. В этом случае искомая функция ценности действия Q непосред-

ственно аппроксимирует Q^{π} – оптимальную функцию ценности действия независимо от применяющейся стратегии. Стратегия определяет, какие пары «состояние–действие» посещаются и корректируются. В *Q*-обучении правило обновления всегда выполняется на основе жадной детерминированной стратегии, которая совершенствуется. И в то же время действия, выбранные для управления, основаны на другой стратегии (не зависящей от $Q(s, a)$). Например, для генерирования действий может быть использована стратегия с равномерным распределением по пространству действий.

Для обеспечения сходимости необходимо, чтобы все пары продолжали корректироваться. Это является минимальным требованием в том смысле, что каждый метод, гарантированно находящий оптимальную линию поведения, в общем случае должен удовлетворять данному условию. Установлено, что при таком условии и в случае стохастической аппроксимации для последовательности значений длины шага функция Q_t сходится к Q^{π} с вероятностью 1 [1].

Метод TD(λ) **–** метод, в котором временное различие имеет протяженность в *n* шагов. Вводится дополнительная переменная памяти, соответствующая каждому состоянию, – *след приемлемости*. Для состояния *s* в момент времени *t* след приемлемости обозначается $e^t(s)$. На каждом шаге следы приемлемости для всех состояний убывают с коэффициентом γλ, а след приемлемости для посещаемого на данном шаге состояния увеличивается на 1: $e^t(s) = \gamma\lambda e_{t-1}(s)$, если $s \neq s_t$, $e^t(s) = \gamma\lambda e_{t-1}(s) + 1$, если $s = s_t$, где γ – ценность терминального состояния; λ – коэффициент затухания. Такие следы показывают степень приемлемости каждого состояния при происходящих изменениях в обучении, если возникает подкрепляющее событие. Таким образом, для метода TD(λ) имеем: $V_{St} \leftarrow V(S_t) + e(s)\alpha[r_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$, $\forall s \in S : e(s) \neq 0$.

Используя следы приемлемости, все состояния должны быть обновлены на каждом шаге (выбор действия *a* в состоянии s_t и получение вознаграждения *r* в состоянии s_{t+1}). В то же время информация о текущих выплатах распространяется обратно к состояниям с более высокими значениями следов приемлемости. Можно показать, что при значении λ = 0 алгоритм становится аналогичным алгоритму TD(0), обновляя только состояния s_t на шаге $t+1$. При значении λ = 1 алгоритм поиска решения эквивалентен полному прогону метода, который имеет смысл только для эпизодических задач при оценке значений состояний после получения всех вознаграждений и подсчета полной выгоды.

Обучение с подкреплением в мультиагентной системе

Известно, что мультиагентная система – это группа автономных взаимодействующих между собой субъектов (агентов), имеющих общую интеграционную среду и способных получать, хранить, обрабатывать и передавать информацию в интересах решения как собственных, так и корпоративных (общих для группы агентов) задач анализа и синтеза информации [9]. Структура мультиагентной системы для RL-обучения аналогична рассмотренной ранее схеме взаимодействия «агент–окружающая среда» (см. рис. 1) с отличием в том, что на окружающую среду оказывают влияние несколько агентов одновременно и, соответственно, действия каждого агента могут зависеть от действий остальных агентов системы.

К преимуществам мультиагентных систем в RL-обучении можно отнести следующие:

– возможность параллельных вычислений, так как используется распределенный характер взаимодействия агентов, в рамках ускорения работы системы;

– обмен опытом между агентами, средствами обучения и имитации, позволяющий помочь RL-агентам с похожими задачами обучаться быстрее и достичь более высокой производительности;

– отказоустойчивость – при выводе из строя одного или несколько агентов система продолжает функционировать;

– масштабируемость – включение или исключение агента из системы не влияет на работу системы в целом.

Но при этом возникают определенные сложности:

– сложность задания цели обучения;

– нестационарность проблемы обучения, возникающая из-за того, что все агенты обучаются одновременно и каждый агент сталкивается с проблемой изменяющейся цели обучения, поэтому основная стратегия может меняться при изменении стратегий других агентов; таким образом, RL-агенту необходимо найти компромисс между использованием текущих знаний и исследованием среды для сбора информации и улучшения этих знаний;

– необходимость координации;

– экспоненциальный рост дискретного пространства состояний-действий, так как основной алгоритм Q-обучения оценивает значения всех возможных пар «состояния–действия», что ведет, соответственно, к экспоненциальному увеличению вычислительной сложности.

Реализация подсистемы прогнозирования для ИС РВ с включением методов обучения

На основе статистических и экспертных методов прогнозирования был предложен ***комбинированный (интегрированный) метод прогнозирования*** [10], который заключается в усреднении результатов, полученных на основе методов скользящей средней и Байеса [11] с учетом весовых коэффициентов. Затем полученный прогноз кор-

ректируется относительно значений ряда, полученного на основе метода экспоненциального сглаживания, а далее – с учетом экспертных методов ранжирования и непосредственной оценки. Вероятность каждого исхода, полученного статистическими методами, корректируется (увеличивается или уменьшается) в зависимости от значений экспертных оценок для указанных исходов.

Предложенная архитектура подсистемы прогнозирования (рис. 2) включает

– эмулятор, моделирующий состояние среды с использованием различных алгоритмов изменения параметров системы в оперативной БД;

– модуль прогнозирования на основе статистических методов (методов экстраполяции по скользящей средней, экспоненциального сглаживания и байесовского подхода) и прогнозирования на основе экспертных методов (методы ранжирования и непосредственной оценки);

– мультиагентный модуль RL-обучения, состоящий из группы независимых агентов, каждый из которых обучается на основе одного из разработанных TD-методов (TD(0), TD(λ), SARSA, Q-обучение), а также используемый для накопления знаний об окружающий среде и способный к адаптации, модификации и накоплению знаний;

– модуль принятия решений, предназначенный для анализа данных, поступающих от модулей прогнозирования, RL-обучения и принятия решений о последующих действиях, способе корректировки стратегий управления и т.д.

Выполнена программная реализация прототипа подсистемы прогнозирования с использованием статистического и экспертного модулей для решения задач экспертного диагностирования сложного технологического объекта – одной из подсистем АЭС (подсистема «1 контур» ВВЭР АЭС) с целью выполнения прогнозирования для оценки развития ситуации на объекте [10]. По результатам тестирования установлено, что необходимо привлечение дополнительных средств: методов RL-обучения на основе темпоральных различий, которые позволяют выявить имеющиеся закономерности посредством анализа предыстории процесса и таким образом уменьшить влияние случайных явлений.

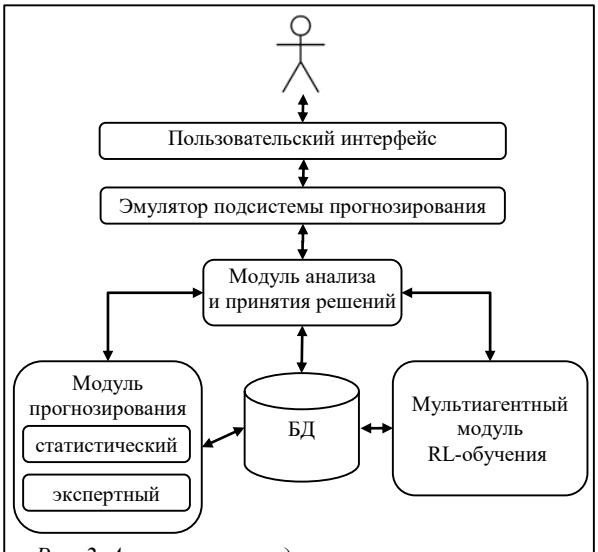

Рис. 2. Архитектура подсистемы прогнозирования

Рис. 2. Architecture of the forecasting subsystem

Были разработаны и исследованы различные алгоритмы TD-методов (TD(0), TD(λ), SARSA, Q-обучение) с целью анализа возможности их применимости в интегрированной среде, а также для сравнения результатов прогнозирования при использовании различных методов и их комбинаций [10]. Задача исследования – проектирование мультиагентной системы RL-обучения и ее интеграция в подсистему прогнозирования, а также нахождение наиболее предпочтительных для включения в состав ИС РВ типа ИСППР РВ методов RL-обучения и прогнозирования и оценка эффективности функционирования мультиагентных систем в рамках ИСППР РВ.

Заключение

В работе были проанализированы различные методы RL-обучения и реализованы соответствующие алгоритмы в плане их последующей интеграции в блок прогнозирования для ИС РВ типа ИСППР РВ. Особое внимание уделено методам на основе темпоральных различий (TD-методам). Предложен комбинированный метод прогнозирования, основанный на статистических и экспертных методах прогнозирования, и реализованы алгоритмы для комбинированного метода. Предложена архитектура подсистемы прогнозирования, включающая модуль прогнозирования, мультиагентный модуль RL-обучения и модуль анализа и принятия решений.

В настоящее время разрабатывается мультиагентный модуль RL-обучения для его включения в интегрированную среду, ориентированную на использование в ИСППР РВ семиотического типа, с целью расширения области применения, повышения производительности и эффективности функционирования современных ИСППР РВ.

Работа выполнена при финансовой поддержке РФФИ (проекты №№ 17-07-00553, 16-51-00058) и проекта по государственному заданию № 2.737.2014/К.

Литература

1. Саттон Р.С., Барто Э.Г. Обучение с подкреплением; [пер. с англ.]. М.: БИНОМ. Лаборатория знаний, 2011. 400 с.

2. Вагин В.Н., Еремеев А.П. Некоторые базовые принципы построения интеллектуальных систем поддержки принятия решений реального времени // Изв. РАН. Теория и система управления. 2001. № 6. С. 114–123.

3. Shani G., Brafman R.I., Shimony S.E. Model-based online learning of POMDPs. Proc. Europ. Conf. on Machine Learning, 2005, pp. 353–364.

4. Рыбина Г.В., Паронджанов С.С. Технология построения динамических интеллектуальных систем: учеб. пособие. М.: Изд-во НИЯУ МИФИ, 2011. 240 с.

5. Осипов Г.С. Методы искусственного интеллекта. М.: Физматлит, 2011. 296 с.

6. Busoniu L., Babuska R., and De Schutter B. Multi-agent reinforcement learning: An overview. Chapter 7 in Innovations in Multi-Agent Systems and Applications-1 (D. Srinivasan and L.C. Jain, eds.), vol. 310 of Studies in Computational Intelligence, Berlin, Germany: Springer, 2010, pp. 183–221.

7. Doshi-Velez F., Pineau J., Roy N. Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs, Artificial Intelligence, 2012, no. 187, pp. 115–132.

8. Еремеев А.П., Подогов И.Ю. Обобщенный метод иерархического подкрепленного обучения для интеллектуальных систем поддержки принятия решений // Программные продукты и системы. 2008. № 2. С. 35–39.

9. Sort J., Singh S., Lewis R.L. Variance-based rewards for approximate Bayesian reinforcement learning. Proc. Conf. Uncertainty in Artificial Intelligence, 2010, pp. 564–571.

10. Еремеев А.П., Кожухов А.А. Разработка интегрированной среды на основе методов прогнозирования и обучения с подкреплением для интеллектуальных систем реального времени // IS&IT'16: тр. Конгресса по интеллект. сист. и информ. технологиям. Науч. изд. в 3 т. Таганрог: Изд-во ЮФУ, 2016. Т. 1. С. 140–149.

11. Ross S., Chaib-draa B., Pineau J. Bayes-adaptive POMDPs. Proc. Conf. Advances in Neural Information Processing Systems 20, 2007, pp. 1225–1232.

Software & Systems
DOI: 10.15827/0236-235X.117.028-033

Received 17.11.16
2017, vol. 30, no. 1, pp. 28–33

IMPLEMENTATION OF REINFORCEMENT LEARNING METHODS BASED ON TEMPORAL DIFFERENCES AND A MULTI-AGENT APPROACH FOR REAL-TIME INTELLIGENT SYSTEMS

A.P. Eremeev [1]*, Dr.Sc. (Engineering), Professor, eremeev@appmat.ru*
A.A. Kozhukhov [1]*, Postgraduate Student, saaanchezzz@yandex.ru*

[1] *National Research University "MPEI", Krasnokazarmennaya St. 14, Moscow, 111250, Russian Federation*

Abstract. The paper describes implementation of reinforcement learning methods based on time (temporal) differences and a multi-agent technology. The authors examine the possibilities of combining learning methods with statistical and expert methods of forecasting for further integration into an instrumental software environment to use in modern and advanced real-time intelligent systems (RT IS), a type of real-time intelligent decision support systems (RT IDSS).

There is an analysis of reinforcement learning (RL-learning) methods in terms of using them in RT IS, main components, benefits and tasks. The paper focuses on the methods of RL-learning based on time (temporal) differences (TD-methods) and presents the developed corresponding algorithms. The authors consider the possibility of including RL-learning methods into a multi-agent environment and combining them with statistical and expert forecasting methods in terms of integration into the environment, which was developed for RT IDSS for complex technical object control and diagnosis.

The paper proposes the architecture of the forecasting subsystem prototype consisting of an emulator, which simulates the state of environment, forecasting module, analysis and decision-making module and a multi-agent RL-learning module. There is software implementation of the forecasting subsystem prototype using a multi-agent approach in order to solve the problem of the complex technological object expert diagnosis.

According to the results of testing and validation of the developed system, the paper considers the conclusions about the efficiency and expediency of including into the RT IDSS.

Keywords: artificial intelligence, intelligent system, real time, reinforcement learning, forecasting, decision support, program tools.

Acknowledgements. *The work has been financially supported by RFBR (projects no. 17-07-00553, 16-51-00058) and a project by the state order no. 2.737.2014/K.*

References

1. Sutton R.S., Barto A.G. *Reinforcement Learning*. London, MIT Press, 2012, 320 p. (Russ. ed.: Moscow, BINOM Publ., 2011, 400 p.).

2. Vagin V.N., Eremeev A.P. Some basic construction principles of real-time intelligent decision support systems. *Izv. RAN. Teoriya i sistemy upravleniya* [Journal of Computer and Systems Sciences International]. 2001, no. 6, pp. 114–123 (in Russ.).

3. Shani G., Brafman R.I., Shimony S.E. Model-based online learning of POMDPs. *Proc. European Conf. on Machine Learning*. 2005, pp. 353–364.

4. Rybina G.V., Parondzhanov S.S. *Tekhnologia postroeniya dinamicheskikh intellektualnykh sistem* [The technology of building dynamic intelligent systems]. Moscow, MEPHI Publ., 2011, 240 p.

5. Osipov G.S. *Metody iskusstvennogo intellekta* [Methods of artificial intelligence]. Moscow, Fizmatlit Publ., 2011, 296 p.

6. Busoniu L., Babuska R., De Schutter B. *Multi-agent reinforcement learning*. Berlin, Germany, Springer Publ., 2010, ch. 7., vol. 310, pp. 183–221.

7. Doshi-Velez F., Pineau J., Roy N. Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs. *Artificial Intelligence*. 2012, vol. 187–188, pp. 115–132.

8. Eremeev A.P., Podogov I.U. Generalized method of hierarchical reinforcement learning for intelligent decision support systems. *Programmnye produkty i sistemy* [Software and Systems]. 2008, no. 2, pp. 35–39 (in Russ.).

9. Sort J., Singh S., Lewis R.L. Variance-based rewards for approximate Bayesian reinforcement learning, *Proc. Uncertainty in Artificial Intelligence.* 2010, pp. 564-571.

10. Eremeev A.P., Kozhukhov A.A. Analysis and development of reinforcement learning methods based on temporal differences for real time intelligent systems. *Proc. 15th National Conf. on Artificial Intelligence with International Participation KII-2016*. Vol. 1, Smolensk, Universum Publ., 2016, pp. 323–330 (in Russ.).

11. Ross S., Chaib-draa B., Pineau J. Bayes-adaptive POMDPs. *Advances in Neural Information Processing Systems*. 2008, vol. 20, pp. 1225–1232.