УДК 681.513

С.Г. Удовенко, А.Р. Сорокин

Харьковский национальный университет радиоэлектроники, Харьков

НЕЧЕТКОЕ УПРАВЛЕНИЕ АВТОНОМНЫМ МОБИЛЬНЫМ РОБОТОМ С ПОДКРЕПЛЯЕМЫМ ОБУЧЕНИЕМ

Статья посвящена разработке метода автономной навигации колесных мобильных роботов в неизвестной среде с комбинированным применением нечеткой модели и RL-обучения, который позволяет последовательно улучшать набор нечетких правил системы, используя сигналы подкрепления. Реализуемые роботом действия для различных типов препятствий основаны на измерениях датчиков для определения положения цели и оценивания расстояния до препятствий. Вычислительная трудоемкость предлагаемого метода позволяет его реализовать в реальном масштабе времени.

Ключевые слова: *мобильный робот, нечеткая система, навигация, стратегии управления, нечеткий регулятор.*

Введение

Современные исследования показывают эффективность применения в интеллектуальных системах управления методов машинного обучения [1]. В последнее время получили распространение управляемые стохастические системы, в основе функционирования которых лежат метод обучения с подкреплением (reinforcement learning (RL)) [2]. В основе этого метода лежат принципы адаптивного поведения, которые позволяют мобильным объектам приспосабливаться к изменяющимся или неизвестным условиям окружающей среды. Отличительной чертой метода обучения с подкреплением является наличие скалярного сигнала подкрепления, который получает агент в процессе взаимодействия с внешней средой и который характеризует эффективность функционирования агента в данный момент времени. В исходном виде RL-метод использует конечное количество состояний внешней среды и возможных воздействий агента на внешнюю среду, а также взаимодействие агента с внешней средой в дискретные моменты времени. Обучение с подкреплением является методом, который позволяет находить оперативное решение, являющееся оптимальным в смысле получения максимального дохода в каждом из состояний. При этом он позволяет в процессе обучения допускать возможность кратковременных потерь, чтобы впоследствии максимизировать суммарный доход на длительном интервале. Вследствие этого, обучение с подкреплением является методом, концептуально приспособленным для эффективной работы в интеллектуальных системах, характеризующихся высоким уровнем изменения внешних и внутренних воздействий. Целью RL-алгоритмов является определение и реализация стратегии, основанной на текущем состоянии, и соответствующей максимальному значению длительной суммы сигналов подкрепления.

Модель такой задачи может быть описана процессом решений, алгоритм которого идентифицирует дискретный набор состояний окружающей среды S и выполняет одно из возможных действий из множества A [3]. В ответ на действие a_t в момент t при текущем состоянии среды s_t агент системы получает ответный сигнал подкрепления $r_t = r(s_t, a_t)$ от окружающей среды, после чего окружающая среда переходит в новое состояние $s_{t+1} = \delta(s_t, a_t)$. В алгоритме используются функции перехода $\delta(s_t, a_t)$. Функции перехода и подкрепления зависят только от текущих состояний и действий.

Рассмотрим возможность применения модифицированных методов машинного обучения с подкреплением для решения задачи нечеткого управления колесным мобильным роботом (МР) в непрерывной среде, характерной для систем интеллектуального управления стохастическими процессами. Перед МР ставится стандартная задача – добраться до цели, избежав столкновения с препятствиями. Критерием оценки эффективности функционирования МР служит среднее значение награды, полученной за время взаимодействия со средой. Рассмотрим модель робота, имеющего 3 колеса (2 задних и одно переднее), один двигатель, обеспечивающий управляемое перемещение, и один одометрический датчик для измерения положения и продольной скорости. Ориентация переднего колеса (угол его поворота) регулируется вторым двигателем. Это колесо, обеспечивающее устойчивость МР, оснащено датчиком ориентации, позволяющим измерять угол поворота шасси робота.

Пусть рассматриваемая система управления МР состоит из блоков, учитывающих 5 типов нечеткого поведения по модели Такаги-Сугено нулевого порядка: «движение к цели» (Goal Seeking Behavior -

GSB), «обход препятствий, расположенных прямо» (Front Obstacle Avoider - FOA), «обход препятствий, расположенных справо» (Right Obstacle Avoider - ROA), «обход препятствий, расположенных слева» (Left Obstacle Avoider - LOA) и «уменьшение скорости движения» (Velocity Reducing Behavior VRB).

Предположим, что МР имеет 7 ультразвуковых датчиков для обнаружения препятствий в трех направлениях (прямо, справа и слева).

Датчики сгруппированы в три модуля (по направлениям) и для каждого из модулей используются три датчика для выработки наилучшего управления движением. На рис. 1 показано расположение датчиков на МР и размещение их в модулях:

- модуль МД1 для обхода препятствий, расположенных прямо (FOA), использует расстояния d1, d2, d3;
- модуль МД2 для обхода препятствий, расположенных справа (ROA), использует расстояния d2, d4, d6;
- модуль МД3 для обхода препятствий, расположенных слева (LOA), использует расстояния d3, d5, d7.

Датчики Ci измеряют расстояния, наиболее близкие к di, i = 1,..,7. Схема размещения датчиков приведена на рис. 1.

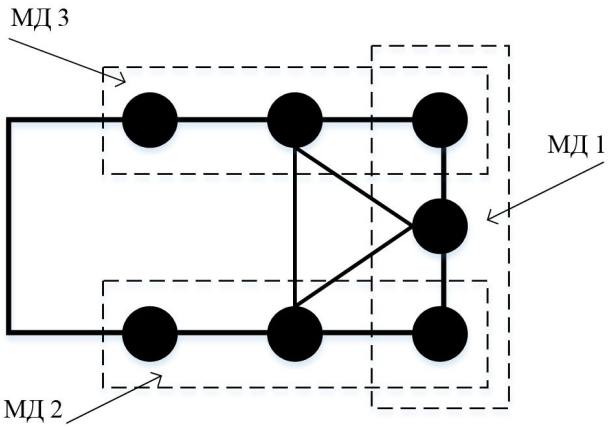

Рис. 1. Расположение датчиков на МР

Нечеткое управление рассматриваемым объектом может быть реализовано с помощью нечетких регуляторов (НР). В то же время синтез нечетких регуляторов (НР), позволяющих получить желаемое поведение МР, не всегда является тривиальным. Основным недостатком применения таких регуляторов является отсутствие обобщенной методологии при наличии значительного числа определяемых параметров (параметров функций принадлежности, правил нечеткого вывода и т.п.).

Улучшить качество нечеткого управления можно с применением RL-обучения, которое позволяет использовать априорные знания для адаптации нечетких правил управления на основе максимизации среднего значения получаемых подкреплений.

Целью статьи является разработка метода автономной навигации МР в неизвестной среде с комбинированным применением нечеткой модели и RL-методов, который позволяет улучшать набор нечетких правил, используя сигналы подкрепления.

Модифицированный алгоритм нечеткого управления с RL-обучением

Наиболее распространенным алгоритмом обучения с подкреплением (RL-обучения), который может быть использован в системах управления динамическими системами, является базовый алгоритм Q-обучения [4].

В этом алгоритме для определения оптимальной стратегии используется Q-функция, итеративную процедуру обновления которой можно представить в следующем виде:

$$Q_{t+1}(s,a) \leftarrow r + \gamma \cdot \max_{a \in A} Q(s',a), \qquad (1)$$

где α – действие, вызывающее переход среды из состояния s в состояние s'; α $(0 \le \alpha \le 1)$ – коэффициент нормирования значений Q-функции.

Q-функция должна быть определена для всех пар «состояние-действие». Для дискретных задач малой размерности можно наглядно показать применение соответствующих таблиц. Табличное представление Q-функций становится затруднительным для непрерывных пространств состояний и действий, а также для случая непрерывного входа с дискретными действиями. Для задания действий здесь могут быть использованы искусственные нейронные сети (ИНС) [4]. Нейросетевое представление Q-функций использует аппроксимирующие свойства ИНС, но не гарантирует сходимости Q к оптимальному значению Q*. Такая архитектура содержит N нейронных сетей типа «многослойный персептрон» (N – число действий α_i), каждая из которых используется для аппроксимации функции $Q(s,\alpha_i)$ для действия α_i. Главным недостатком здесь является применение двух достаточно медленных подходов: Q-обучения и обучения нейронной сети. После применения действия α_k в состоянии s, разность $Q_{t+1}(s,\alpha_k) - Q_t(s,\alpha_k)$ между эволюцией качества для шагов t и $t+1$ может рассматриваться как сигнал ошибки в нейронной имплементации. Для оптимизации ИНС обычно используется алгоритм обратного распространения, минимизирующий критерий следующего вида:

$$E_t(s,\alpha_k) = \frac{1}{2}\left[Q_{t+1}(s,\alpha_k) - Q_t(s,\alpha_k)\right]^2. \qquad (2)$$

Нечеткая версия такого представления для непрерывного пространства состояний и дискретных действий, именуемая Q-FUZ, предложена в [5]. Функция качества реализуется здесь нечеткой сис-

темой с N выходами. После выбора функций принадлежности задача обучения состоит в оптимизации правил вывода. Эти правила позволяют получить искомые значения $s \to Q(s,\alpha_j)$ для $j=1...N$.

Такая структура не использует свойства интерполяции нечетких систем и имеет дискретные выходы. Рассмотрим возможность расширения Q-FUZ представления для оптимизации нечетких правил вывода Такаги-Сугено (ТС) и его адаптации к задаче навигации МР.

Принцип работы предлагаемого модифицированного алгоритма нечеткого управления с RL-обучением (Q-FUZM) состоит в получении множества выводов для каждого нечеткого правила и ассоциации для каждого вывода функции качества, которая будет оцениваться с применением фиксированной функции принадлежности. При настройке по алгоритму Q-FUZM нечеткий регулятор (НР) мобильного робота должен корректировать выводы из правил ТС на основе сигналов подкрепления. Задача состоит в аппроксимации функции качества Q следующей нечеткой функцией SIF (System Inference Fuzzy):

$$s \to y = \hat{Q} = SIF(s). \qquad (3)$$

Если выбрать нечеткую ТС-систему нулевого порядка (ТС0), такая функция определится правилами следующего вида:

$$\begin{aligned} &\text{если } s = S_1, \text{ то } y = c_1; \\ &\text{если } s = S_2, \text{ то } y = c_2; \\ &\text{если } s = S_m, \text{ то } y = c_r, \end{aligned} \qquad (4)$$

где m – число правил, а прототипы i-го правила S_i определяются как: x_1 есть A_1^i и... и x_n есть A_n^i.

Для входного вектора s_t эволюция величины действия задается следующим уравнением:

$$\hat{Q} = SIF(s_t) = \sum_{i=1}^{m} w_i(s_t)c_i, \qquad (5)$$

где $w_i(s)$ - SIF-коэффициенты, определяемые по функциям принадлежности; c_i – выводы нечетких правил типа (4).

Выводы нечетких правил $(c_i)_{i=1}^m$ могут интерпретироваться как ограничение функции $s \to \hat{Q}(s,\alpha)$ прототипами S_i, т.е.:

$$q[S_i,\alpha] \equiv \hat{Q}(s,\alpha), \text{ если } s = S_i \text{ для } i = 1,...r,$$

где $q[S_i,\alpha]$ – функция величины действия α в состоянии S_i (для правила i).

Подставляя в (5) $c_i = q[S_i,\alpha]$, получаем:

$$\hat{Q} = \sum_i w_i(s_t)q[S_i,\alpha]. \qquad (6)$$

Процесс обучения по алгоритму Q-FUZM с применением процедуры обновления типа (1) позволяет определить набор правил, максимизирующих будущие подкрепления Начальная база правил состоит из m правил следующего вида:

$$\begin{aligned} &\text{если } s = S_i, \text{ то } y = \alpha[i,1]; \text{ при } q[i,1] = 0; \\ &\text{или } y = \alpha[i,2]; \text{ при } q[i,2] = 0; \\ &\text{или } y = \alpha[i,N]; \text{ при } q[i,N] = 0, \end{aligned} \qquad (7)$$

где $q[i,j]$ при $i=1,...m$ и $j=1,...N$ – потенциальные решения.

В процессе обучения вывод по каждому правилу выбирается по средним значениям сигналов подкрепления $C_r(i) \in \{1...N\}$. В этом случае результирующий выход определяется как:

$$A(s) = \sum_{i=1}^{N} w_i(s)q[i,C_r(i)]. \qquad (7)$$

Качество такого действия оценивается следующим образом:

$$\hat{Q}(s,A(s)) = \sum_{i=1}^{N} w_i(s)q[i,C_r(i)]. \qquad (8)$$

Алгоритм одношагового нечеткого управления состоит в реализации следующей последовательности действий:

1- Выбрать структуру системы нечеткого вывода (SIF);

2- Инициализировать $q[i,1] = 0$; $i = 1,...m$ (количество правил),

3- Повторить (для каждого эпизода):

3.1. Наблюдение состояния s

3.2. Повторить (для каждого эпизода)

3.2.1. Для каждого правила i вычислить $w_i(s)$

3.2.2. Для каждого правила i выбрать вывод с помощью C_r

3.2.3. Вычислить выход $A(s)$ и его качество, соответствующее $\hat{Q}(s,A(s))$

3.2.4. Применить действие $A(s)$, приводящее к новому состоянию s'

3.2.5. Получить подкрепление r

3.2.6. Для каждого правила i вычислить:

$$V^*(s') = \sum_{i=1}^{m} w_i(s')q[i,\max(i)]$$

3.2.7. Вычислить

$$\Delta Q = \beta\left[r + \gamma V^*(s') - \hat{Q}(s,A(s))\right]$$

3.2.8. Пересчитать для нового состояния элементы функции качества для выбираемых выводов

$$q[i,j] = q[i,j] + \Delta Q[i,j].$$

Действия по пунктам 3.1-3.2 повторяются до конечного S_t.

Рассматриваемый алгоритм одношагового нечеткого Q- обучения может быть использован для реализации разных типов поведения АМР (напри-

мер, для поиска цели, обхода препятствий, и движения вдоль стен). В наиболее простом варианте робот использует исходную базу данных типа TS0, определяющую возможные ситуации для желаемого поведения. При этом реализация алгоритма состоит в формировании множества выводов типа «импульс» для каждого правила и ассоциации с каждым выводом функции качества, которая оптимизируется во времени. Целью фазы обучения является определение набора выводов правил, максимизирующих среднее значение сигналов подкреплений. После реализации выбранного действия и получения сигнала подкрепления анализируется новая ситуацич.

Нечеткое RL-управление для различных типов поведения МР

Рассмотрим три типа стратегий управления поведением МР в зависимости от наличия или отсутствия препятствий в зоне прямого наблюдения: С1 – стратегия управления МР при непосредственном движении к цели; С2 – стратегия управления МР при обходе препятствий; С3 – стратегия управления МР при движении вдоль стен.

Стратегия С1 позволяет реализовать действие «движение к цели» МР на основе знания его положения относительно координат навигационной среды (т.е. сориентировать робот по направлению к цели).

Для достижения цели в среде, свободной от препятствий, МР должен продвигаться вперед, поворачивать направо или налево с различной скоростью в соответствии с переменными входа регулятора. В этом случае, как показано в предыдущих примерах для одного и того же типа поведения, МР должен измерять расстояние «робот-цель» (d_{rg}) и угловой ошибки (θ_{rg}). Используя эти две величины, НР, оптимизируемый по RL-алгоритму, будет генерировать два действия по передвижению к цели (угол изменения направления и скорость движения). Для этого поведение определяется с помощью нечеткой системы типа TS0. Для входных переменных используются функции принадлежности, представленные на рис. 2 и 3 со следующими лингвистическими переменными: N (Near), M (Middle), B(Big), NB (Negative Big), NM (Negative Middle), Z (Zero), PM (Positive Middle), PS (Positive Small), и PB (Positive Big). При этом образуются 15 правил. Для каждого нечеткого правила предлагаются 3 вывода для выходной переменной «угол поворота» следующего вида:

«Если s есть S_i, то $\alpha = \alpha_{i1}$ при качестве $q[i,1]$; $\alpha = \alpha_{i2}$ при качестве $q[i,2]$; или $\alpha = \alpha_{i3}$ при качестве $q[i,3]$; для $i = 1...15$; $\alpha_1 = -\pi/5$, $\alpha_2 = 0$, $\alpha_3 = \pi/5$». (9)

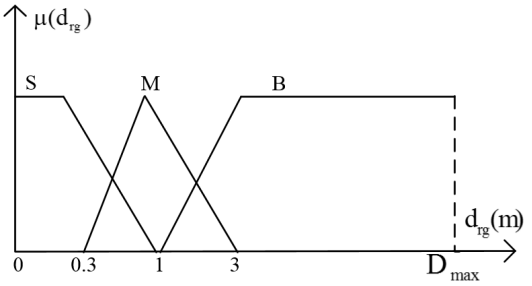

Рис. 2. Функции принадлежности для d_{rg}

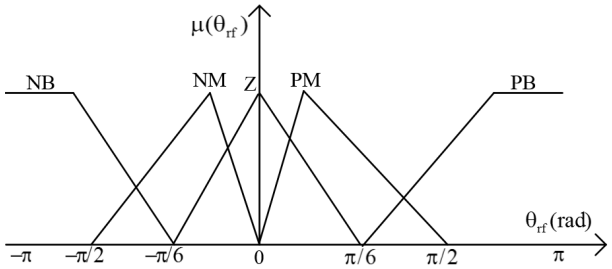

Рис. 3. Функции принадлежности для θ_{rg}

В нашем случае для обучения МР при типе поведения «движение к цели» используется алгоритм нечеткого Q- обучения дважды для двух управлений (для угла поворота и для скорости движения). В первом случае во время фазы настройки МР получает те же значения сигнала подкрепления, что и на предыдущем шаге. При этом во время перемещения МР формирует следующие сигналы подкрепления r (выигрыша или штрафа):

4, если робот достигает цели;

3, если d_{rg} уменьшается и $\theta_{rg} = 0$;

2, если d_{rg} и θ_{rg} уменьшаются;

-1, если d_{rg} уменьшается, а θ_{rg} увеличивается;

-2, если d_{rg} увеличивается;

-3, если МР наталкивается на стены среды.

После фазы обучения МР выбирает для каждого правила вывод с наилучшей функцией качества $q[i,j]_{j=1}^{N}$. Если МР может дойти до конечной цели с фиксированной скоростью, применяется второй алгоритм для оптимизации регулятора скорости, при этом предпосылки правил являются теми же, что используются при расчете угла поворота, но в части выводов имеются различия. База правил представляется здесь следующим уравнением:

«Если s есть S_i, то $V_r = V_{i1}$ при качестве $q[i,1]$; $V_r = V_{i2}$ при качестве $q[i,2]$; или $V_r = V_{i3}$ при качестве $q[i,3]$; для $i = 1...15$. (10)

В процессе обучения скорости МР получает следующие сигналы подкрепления r (выигрыша или штрафа):

-1, если d_{rg} и θ_{rg} уменьшаются, а скорость увеличивается;

-1, если θ_{rg} уменьшается, а скорость и d_{rg} увеличиваются;

1, если МР достигает цели при малой скорости;

0 – в остальных случаях.

В случае наличия препятствий (статических или динамических), которые препятствуют движению МР к цели, робот должен иметь эффективную возможность их обхода.

Рассмотрим стратегию С2, при которой генерируются адекватные действия для избегания столкновений с однотипными (простыми) препятствиями, если в окрестности МР обнаруживаются один или несколько объектов с помощью перцептуальных средств. Обход препятствий является основной задачей, решаемой всеми МР, так как она позволяет роботу перемещаться в неизвестном пространстве, избегая столкновений с окружающими объектами. Будем предполагать, что МР способен измерять расстояния до препятствий с трех сторон (прямо, справа и слева) на основе эффективной перцептивной системы. Система автономной навигации, основанная на нечетком Q-обучении, использует в качестве входов расстояния до препятствий в трех направлениях и должна генерировать (после фазы обучения) угол поворота переднего колеса (α) и скорость передвижения робота (V_r). Две лингвистические переменные (близко (P) и далеко (L)) используются для описания каждого из трех расстояний. Эти переменные представлены функциями принадлежности на рис. 4. База нечетких правил, полученная с использованием этих функций, содержит 8 основных ситуаций.

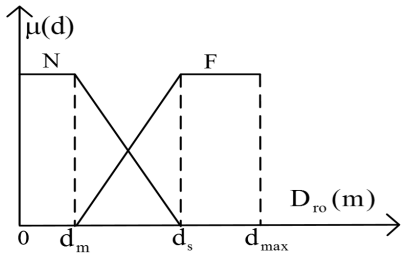

Рис. 4. Функции принадлежности для d_{rg}

Алгоритм RL-обучения используется для оптимизации нечеткого навигатора, генерирующего количество величин-кандидатов для каждого решения. Для каждого правила и каждого вывода ассоциируется функция качества $q[i,j]_{j=1}^{J}$ для интерпретации j правила i, с: $i=1...8$ и $j=1,2,3$. База правил имеет здесь следующий вид:

«Если s есть S_i, то $\alpha=\alpha_{i1}$ при качестве $q[i,1]$; $\alpha=\alpha_{i2}$ при качестве $q[i,2]$; или $\alpha=\alpha_{i3}$ при качестве $q[i,3]$; для $i=1...8$ и $j=1,2,3$; $\alpha_1=-\pi/5$, $\alpha_2=0$, $\alpha_3=\pi/5$.

Глобальное качество выхода нечеткого регулятора определяется следующим образом:

$$Q(s,\alpha)=\sum_{i=1}^{8} w_i(s)q\left[i,j^*(i)\right]. \qquad (10)$$

Используемый сигнал подкрепления r штрафует каждый вывод активируемого правила, когда МР имеет коллизию с препятствием. В процессе обучения скорости МР получает следующие сигналы подкрепления r (выигрыша или штрафа):

-4, если МР сталкивается с препятствием;

-1, если d_i уменьшается и $d_i < D_{max}/2$;

0 – в остальных случаях.

Значения подкрепления, используемые для обучения скорости:

-4, если МР сталкивается с препятствием;

-1, если $d_i > D_{max}$ и V_r низкая для $i=1...3$;

-1, если $d_i < D_{max}$ и V_r увеличивается для $i=1...3$;

0 – в остальных случаях.

Для этого применения значение подкрепления должно служить для определения лучшего вывода из трех выводов, предлагаемых для 2-х управлений.

Задачей навигации МР при движении вдоль стен является сохранение безопасной дистанции робота до стен, детектируемых датчиками справа или слева, осуществляя движение типа «вправо» (стратегия С3). Для этого МР должен собирать необходимую информацию. Будем использовать для этой задачи НР Такаги-Сугено 0-го порядка (TS0) с RL-обучением. Сигнал подкрепления для МР определяется следующим образом:

-2, если МР сталкивается с препятствием;

-1, если $d_i < d_m$ для $i=1...3$;

0 – в остальных случаях.

Этот сигнал используется для определения наилучшей числовой интерпретации используемых лингвистических термов. Алгоритм нечеткого Q-обучения используется для извлечения знаний, при этом предлагаются три интерпретации для каждого уровня выхода (изменение направления). Например, $\alpha=PG$ может быть интерпретировано с помощью $\alpha=45^\circ$, $\alpha=55^\circ$ или $\alpha=35^\circ$. Количественные значения подкреплений, задаваемых роботу для обучения скорости, задаются следующим образом:

-2, если МР наезжает на препятствие;

-1, если $d_i < d_s$ и V_r повышается (для $i=1...3$);

0 – в остальных случаях.

Результаты моделирования

Рассмотрим примеры моделирования навигации МР в различных средах для подтверждения эффективности предложенных схем управления. Используемая среда учитывает ограничения моделирования и движения МР в различных ситуациях, в том

числе, в свободном пространстве и в пространстве со статическими препятствиями.

Стратегия С1. Когда датчики робота не фиксируют никаких препятствий перед МР, то задача сводится к непосредственной ориентации к цели для ее последующего достижения (свободной навигации к цели). Для различных точек старта (S) и финиша (F) МР были получены траектории МР при различных начальных положениях шасси, одна из которых (после этапа RL-обучения) приведена на рис. 5, а.

Отметим, что поведение МР улучшается в процессе обучения (средние значения сигналов подкрепления полученные для каждого шага обучения, приближаются к константе 2, начиная с шага 180, что свидетельствует о сходимости сошелся к оптимальному решению для начальной позиции). Чтобы оптимизировать нечеткую систему навигации с RL-обучением, начальные положения МР должны быть выбраны случайным образом во время фазы обучения. В серии проведенных экспериментов (для НР с 15 правилами и 3 выводами) каждый эпизод начинался со случайного положения и заканчивался, когда МР достигал цели. При этом поведение робота улучшалось со временем, что подтверждает сходимость используемых Q-FUZM-алгоритмов.

Стратегия С2. Если среда МР содержит одно или несколько препятствий, то он должен иметь возможность обойти их без столкновений. Как было отмечено выше, система автономной навигации поддерживает два элементарных типа поведения: (движение к цели (тип 1) и обход препятствий (тип 2). МР осуществляет адекватное действие для достижения финального результата, избегая столкновения с препятствиями, выбирая один из этих двух типов поведения в соответствии с текущей ситуацией. При этом значение подкрепления должно служить для определения лучшего вывода из трех выводов, предлагаемых для двух управлений.

Предположим, что робот при старте находится в зоне, свободной от препятствий и ориентируется по направлению к цели для ее достижения, избегая коллизий. В процессе обучения МР должен вырабатывать возможные действия для каждой регистрируемой ситуации. МР перемещается к цели, когда одно из препятствий детектировано, с одной из трех сторон (анфас, направо или налево). При этом активизируется поведение «обход препятствия» для выработки соответствующих управлений и реализации действий, позволяющих избежать столкновения.

В серии проведенных экспериментов (для разных типов препятствий) каждый эпизод начинался со случайного положения и заканчивался, когда МР достигал цели. Примеры полученных траекторий (после этапа RL-обучения) приведены на рис. 5, б, в.

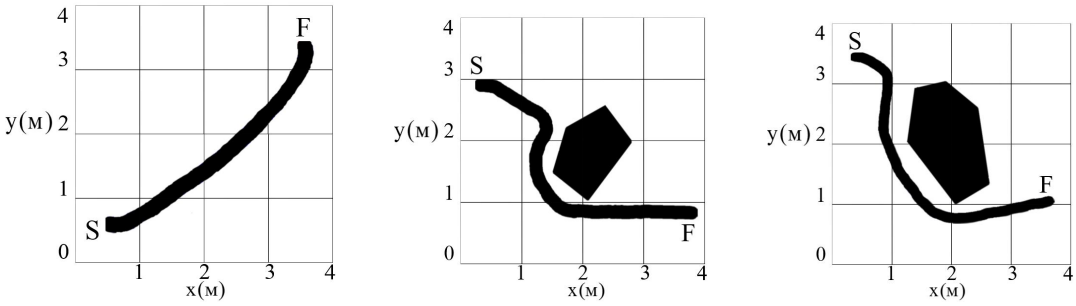

Рис. 5. Траектория МР для стратегии С1 (а – пример 1) и для стратегии С2: б – пример 1, в – пример 2

Стратегия С3. Эта стратегия применяется, если МР должен осуществить обход стен по контуру или достичь цели, огибая стены. Эксперименты с МР без этапа RL-обучения показали, что колесный робот может эффективно двигаться вдоль стен, но в случае, если препятствие содержит углы, МР не способен избегать коллизий. В серии проведенных экспериментов для МР, управляемых по Q-FUZM-алгоритмам (для разных типов стен) каждый эпизод начинался со случайного положения и заканчивался, когда МР полностью обходил контур или (в соответствии с заданием) достигал цели. Примеры полученных траекторий обхода роботом стен (после этапа RL-обучения) приведены на рис. 6, а, б. Пример траектории МР при движении вдоль стен для достижения цели в закрытом пространстве приведен на рис. 6, в.

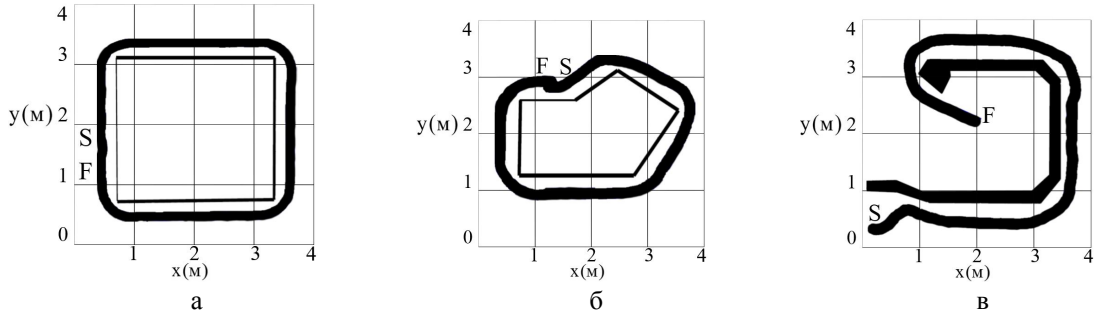

Рис. 6. Траектория МР для стратегии С3: а – пример 1, б – пример 2, с – пример 3

Результаты экспериментов показали, что Q-FUZM-алгоритм дает приемлемые результаты для реализации стратегии С3.

Выводы

В статье рассматривается задача автономной навигации колесного мобильного робота с применением Q-обучения с подкреплением (RL-обучения) в комбинации с нечеткими системами для управления МР. Обучение с подкреплением является эффективным инструментом для получения оптимальных решений. Применение нечеткого Q-обучения основано на введении в схему управления нечеткого регулятора, при этом исходная база правил улучшается в процессе обучения с использованием сигнала подкрепления. Это является перспективным направлением для решения задач навигации АМР, где алгоритм нечеткого RL-обучения позволяет оптимизировать параметры нечеткой системы и расширить возможности Q-обучения для непрерывного случая, основываясь на сигнале подкрепления. Этот сигнал обратной связи позволяет корректировать стратегию навигации для улучшения качества системы. Представление глобальной функции качества Q с помощью нечекой логики эквивалентно определению элементарной функции качества для каждого модального вектора (то есть для каждого правила). Задача обучения состоит в линейном улучшении базы нечетких правил навигации МР.

В соответствии с полученными результатами, RL-обучение является возможным решением для оптимизации нечетких регуляторов, предназначенных для решения задачи навигации МР. Алгоритм нечеткого управления с RL-обучением (Q-FUZM), позволяет построить НР, оптимизируемый с помощью подкреплений. К преимуществам такой техники следует отнести:

- возможность получения оценок качества, благодаря универсальным аппроксимирующим свойствам нечетких систем;
- возможность применения непрерывных управлений и пространства состояний;
- интеграция априорных знаний для получения приемлемого начального поведения МР и ускорения процедуры обучения.

Представленные результаты моделирования дают приемлемые решения для автономной навигации МР в сложных средах. Во всех рассмотренных случаях МР достигает цели, обходя препятствия.

Исследования могут быть продолжены в следующих перспективных направлениях:

- практическое внедрение предложенных методов для реальных МР;
- применение представленных структур для среды с динамическими препятствиями;
- модификация представленного метода путем применения эволюционных алгоритмов (например, ГА или алгоритмов муравьиных колоний).

Список литературы

1. Khriji L. Mobile Robot Navigation Based on Q-learning Technique / L. Khriji, Al Yahmedi // Int. journal of advanced Robotic System. – 2012. – Vol. 8, no. 1. – P. 45-51.

2. Удовенко С.Г. Гибридные методы машинного обучения в системах управления динамическими объектами / С.Г. Удовенко, А.А. Гришко, Л.Э. Чалая // Біоніка інтелекту. – 2012. – № 1 (78). – С. 78-84.

3. Sutton R.S. Reinforcement Learning with Replacing Eligibility Traces / R.S. Sutton // Machine Learning. – 1996. – Vol. 22. – P. 123-158,

4. Boskoski P. Neuro-Fuzzy Controllers and Application to Autonomous Robots / P. Boskoski, M. Stankovski // Mechanics, Automatic Control and Robotics. – 2008. – Vol. 7, no.1. – P.123-132.

5. Cherroun L. Designing of Goal Seeking and Obstacle Avoidance Behaviors for a Mobile Robot Using Fuzzy Techniques / L. Cherroun, M. Boumehraz // Journal of Automation and Systems Engineering (JASE). – 2012. – Vol. 6, no. 4. – P. 164-171.

Поступила в редколлегию 18.05.2016

Рецензент: д-р техн. наук, проф. Е.В. Бодянский, Харьковский национальный университет радиоэлектроники, Харьков.

НЕЧІТКЕ КЕРУВАННЯ АВТОНОМНИМ МОБІЛЬНИМ РОБОТОМ З ПІДКРІПЛЕНИМ НАВЧАННЯМ

С.Г. Удовенко, А.Р. Сорокін

У роботі розглянуто задачу автономної навігації колесного мобільного робота з використанням навчання з підкріпленням (RL-навчання) та нечітких регуляторів. База правил системи автономної навігації робота покращується в процесі навчання з використанням сигналу підкріплення. Розглянуто приклади моделювання навігації мобільного робота у різних середовищах. Застосування наведеного підходу дозволяє враховувати конфігурації перешкод та корегувати стратегію навігації для поліпшення якості системи.

Ключові слова: *мобільний робот, нечітка система, навігація, стратегії управління, нечіткий регулятор.*

FUZZY CONTROL FOR A AUTONOMOUS MOBILE ROBOT WITH REINFORCEMENT LEARNING

S.G. Udovenko, A.R. Sorokin

The task of autonomous navigation of the wheeled mobile robots with the use of reinforcement learning (RL- learning) and fuzzy controllers is examined in the article. The rule base of the system of autonomous navigation of robot gets better in the process of learning with the use of reinforcement signal. The examples of design of mobile robots navigation in different environments are considered. Application the presented approach over allows to take into account configurations of obstacles and correct strategy of navigation for the improvement of the system quality.

Keywords: *mobile robot, fuzzy system, navigation, control strategies, fuzzy controller.*