УДК 004.942, 519.876.5, 004.032.26

ИССЛЕДОВАНИЕ ВЛИЯНИЯ СПОСОБА ФОРМИРОВАНИЯ ФУНКЦИИ ВОЗНАГРАЖДЕНИЯ ПО МЕТОДУ «ДВОЙНИКА» ДЛЯ АЛГОРИТМА ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

Т. Ю. КИМ, канд. техн. наук, доц. Г. А. ПРОКОПОВИЧ
(Объединённый институт проблем информатики
Национальной академии наук Беларуси, г. Минск)

Аннотация. *В работе предложен новый метод управления мобильным роботом по лабиринту. Метод основан на повторении классического обучения с подкреплением в сочетании с правосторонним алгоритмом, который позволил обучить мобильного робота передвигаться по лабиринту. Предложенный метод основан на работе двух алгоритмов, взаимодействующих между собой – алгоритм правой руки и алгоритма обучения с подкреплением. Первый алгоритм является дискретным, который реализует детерминированный алгоритм движения по лабиринту. Скорость движения, которого зависит от второго алгоритма. Второй алгоритм предназначен для «копирования» действий первого алгоритма, имитирующего идеальное движение робота. Разработанная функция вознаграждения позволяет удерживать центр масс робота в центре коридора и при необходимости поворачивать, следуя алгоритму.*

Ключевые слова: *мобильные робот, кинематическая модель, обучение с подкреплением, алгоритм правой руки, лабиринт.*

Введение. Логистические роботы, роботы-уборщики [1], роботы-доставщики и т. д. [2, 3] в последнее время стали обычным явлением. Современная тенденция требует от роботов выполнения сложных задач и навигации в незнакомой среде. Одной из важных задач является навигация по «лабиринту» или «полосе препятствий». Цель одного из методов машинного обучения [4] – обучение с подкреплением – научить робота быстро передвигаться. Возможность алгоритма обучения с подкреплением позволяет роботу проводить исследования и находить наилучшее решение, а его преимущество в том, что он ищет оптимальную модель, и ему не нужно создавать случайную выборку или настраивать ее в автономном режиме. Обучение с подкреплением сочетает в себе две задачи: изучение новой ситуации и использование этого опыта для принятия лучших решений. Однако, как показала практика, существует ряд задач, для которых достаточно проблематичным или невозможным является процесс описания задачи в виде простой целевой функции в виде непрерывной и гладкой функции. Например, к таким задачам можно отнести ветвящиеся алгоритмы.

В данной работе предлагается метод, который совмещает в себе использование детерминированного управления для идеального (ИМР) и аналоговое – для реального (РМР) мобильного робота. Преимущество предложенного метода заключается в том, что функционирование ИМР можно описать алгоритмами любой сложности, причём поведение его может быть достаточно сложным. В то же время, РМР может копировать действия ИМР путём анализа только его внешних действий, сопоставляя текущие входные значения сенсорной системы с требуемыми значениями управляющими сигналами для исполнительной системы самого робота. Таким образом, обучение с подкреплением способен будет обучить РМР ориентироваться в неизвестной среде по правилу правой руки.

Формирование проблемы. Стоит задача с помощью известного алгоритма обучения с подкреплением научить мобильного робота передвигаться по лабиринту следуя одному из выбранных алгоритмов. Алгоритм обучения с подкреплением представим в качестве аналогового объекта управления, где основной целью является разработка вознаграждения в виде скаляра. Цель алгоритма обучения с подкреплением – максимизировать общее вознаграждение. Особое внимание следует обратить на отдельные типы и их взаимосвязь между вознаграждениями: положительное, то есть насколько в целом положительными являются значения вознаграждения при обучении; отрицательный, насколько быстро обучаемый агент избежит отрицательного наказания и поощрения в конце обучения, чтобы агент стремился успешно завершить обучение.

Разработка награды для решения задачи, возможно, является самым важным в обучении с подкреплением. Награда должна помочь агенту выполнить действие, которое принесет максимально долгосрочную награду. Дополнительным критерием является, когда агент получает награду за выполненное действие. Цель Агента – получить награду, которая будет соответствовать поставленной задаче.

Для того чтобы Агент успешно обучился, для него были сформированы следующие критерии:

1. Чем больше расстояние преодолел мобильный робот, тем больше он получил награду

$$dist = (dist_1 + dist_1') / (t \cdot v_{\max}), \quad (1)$$

где $dist_1$ – расстояние от начальной точки до текущей;

$dist_1'$ – расстояние на шаг назад;

t – время моделирования;

$v_{\max}$ – максимально допустимая скорость мобильного робота.

2. Определить повороты с помощью показаний данных с лидара ($lid_n = lid_5$)

$$lidar = \frac{\sum_{n=5}^{1}(lid_n)}{5} - \frac{width}{2}, \tag{2}$$

где $width$ – ширина коридора в лабиринте.

3. Уменьшить влияние угловой скорости (w_{rew}), чтобы мобильный робот не ходил кругами;

4. Дать высокую награду, когда мобильный робот достиг финиша:

$$reward = dist \cdot + lidar + v_{rew} \cdot 0.8 + w_{rew}{}^2 \cdot 0.0015 + finish \cdot 10. \tag{3}$$

5. Сократить расстояние между ИМР и РМР.

6. Побудить РМР двигаться в том же направлении, что и ИМР согласно алгоритму правой руки:

$$rot_{\gamma} = 1 - V\gamma_{real}V - \gamma_{stf}V, \tag{4}$$

где γ_{real} – направленность РМР;

γ_{stf} – направленность ИМР.

7. Дать наивысший штраф, если удовлетворяет условие критерии остановки $isdone = 0$

$$one = \begin{cases} \sum_{i=1}^{v=20} \frac{v(v+4)}{8} = 0; & (5.1) \\ lid_{\min} \le 0.02; & (5.2) \\ \begin{cases} x_{real} = x, \\ y_{real} = y; \end{cases} & (5.3) \\ C > C_{\max}, & (5.4) \end{cases}$$

где v – линейная скорость равна 0, в течение 20 последующих раз (5.1);

lidar – показания минимального луча с лидара меньше, чем 0,02 м (5.2);

(x, y) – координаты РМР (5.3);

C – причем награда нивелирует награду за выход из лабиринта; расстояние между роботами больше, чем C_{max} = 0,6 (5.4).

8. На рисунке 1 представлен дополнительный блок для преодоления тупика в лабиринте, где учитываются показания угловой и линейной скоростей с переключателем.

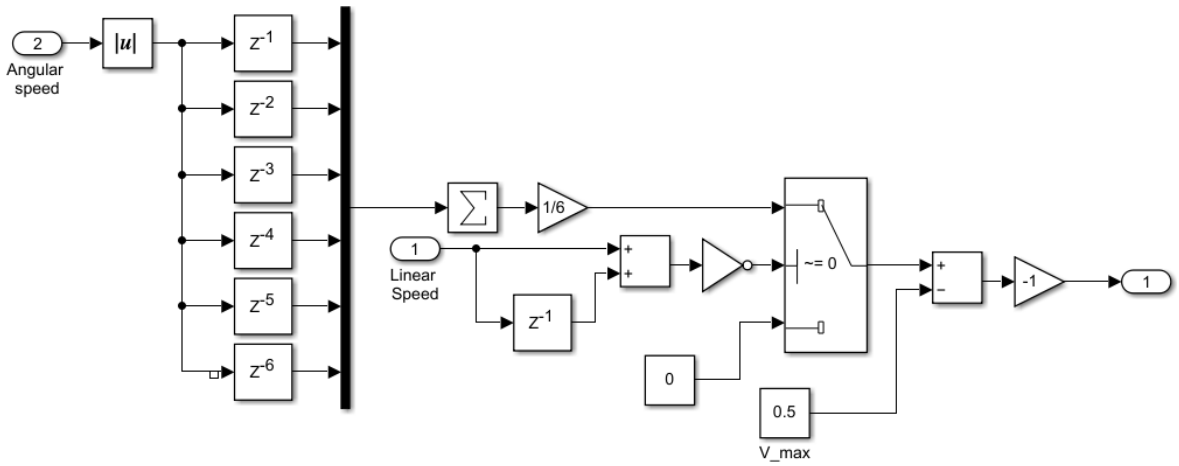


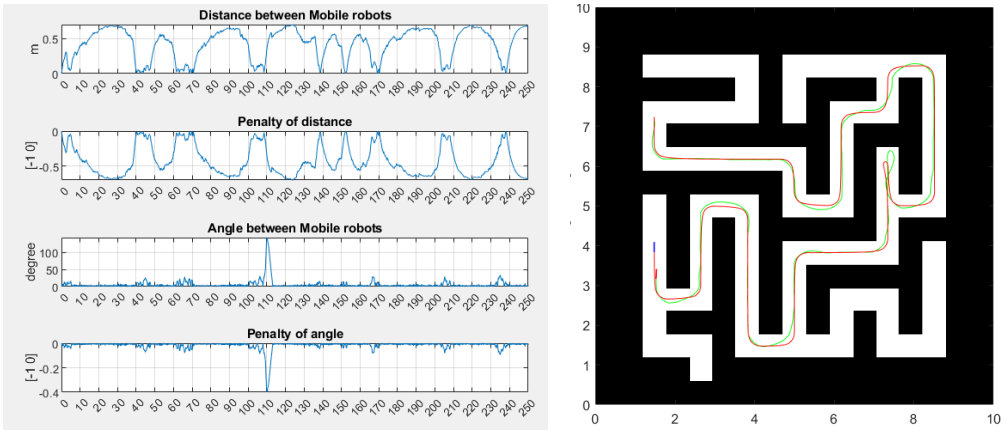
Рисунок 1. – Функция вознаграждения для выхода из тупика

$$penalty = -(v + C + rot_{\gamma} \cdot 0.8 + isdone \cdot 1.5 + impass \cdot 0.3).$$

Окончательная функция вознаграждения имеет вид:

$$reward_{agent} = reward + penalty.$$

Результаты. На рисунке 2 реализован один из моментов обучения, когда PMP следует за ИМР, на визуализации хорошо отслеживается, как разница между направленностью роботами и расстояние между ними влияет на общую награду.



а *б*

Рисунок 2. – Статистика (*а*) и визуализация (*б*) во время обучения на одном из эпизодов

Обученный агент передвигается в незнакомой среде, где больше вариантов выбора повернуть направо или налево, количество стен и тупиковых ситуаций, вариации ширины коридоров и стен. В этом разделе продемонстрированы сравнительные эксперименты и аналитические результаты.

Верификация обученного агента в различной среде. Последовательно усложняя функцию вознаграждения, удалось решит задачу решения поиска пути в лабиринте. Для верификации обученного Агента было решено проверить поведение РМР в незнакомой среде.

По результатам обучения, выяснили что совместное использование алгоритма обучения с подкреплением и алгоритма правой руки, а также разработанной функцией вознаграждения побудило РМР держаться центра коридора, совершать повороты и развороты по лабиринту, а также избегать столкновения.

Заключение. Предложенный метод предполагает, совместное использование детерминированного управления – для ИМР и непрерывного – для РМР. Преимущество предлагаемого метода в том, что несмотря на то, что алгоритмы управления ИМР могут быть достаточно сложными, и даже не известными наблюдателю, но анализируя его действия предложенный метод может обучить РМР сложному поведению, например, ориентироваться в неизвестной среде по правилу правой руки. Разработанный метод имеет большой потенциал для использования обучения с подкреплением в тех областях, где его было сложно реализовать. В результате обученные мобильный робот обобщает эти действия и выводит собственные правила. будет иметь множество применений.

ЛИТЕРАТУРА

1. LA P. Reinforcement learning with function approximation for traffic signal control / LA P., Bhatnagar S. // IEEE Transactions on Intelligent Transportation System – 2011. – Vol. 12, № 2. – P. 412–421.
2. Rezaee K. Application of reinforcement learning with continuous state space to ramp metering in real-world conditions / Rezaee K., Abdulhai B., Abdelgawad H. // In 2012 15th International IEEE Conference on Inteligent Transportation System. – 2012. – P. 1590–1595.
3. Mohammadi M. Semisupervised deep reinforcement learning in support of IoT and smart city services / Mohammadi M., Al-Fuqaha A., Guizani M., Oh J. // IEEE Internet of things journal. – 208. – Vol. 5, № 2. – P. 624–635.
4. Huang W. Learning to drive via apprenticeship learning and deep reinforcement learning / Huang W., Braghin F., Wang Z. // ArXiv:2001.03864. – 2020.