УДК 004.056.57:032.26

КОЛЛЕКТИВНОЕ ПОВЕДЕНИЕ В МНОГОАГЕНТНЫХ СИСТЕМАХ НА ОСНОВЕ ПОДКРЕПЛЯЮЩЕГО ОБУЧЕНИЯ

Кабыш А.С.
УО «Брестский государственный технический университет», г. Брест

Введение. Многоагентное обучение – новая область исследований в машинном обучении (machine learning). Также подкрепляющее обучение – это последнее, самостоятельно выделившееся, направление машинного обучения, делающее только первые шаги. На пересечении этих двух областей находится множество нерешенных проблем. Одна из них – это коллективное обучение, или задача адаптации одиночного обучения на многоагентный подход. Агент – сущность, находящаяся в какой-то среде и способная действовать в ней. Многоагентная система (МАС) – это группа взаимодействующих агентов. Более формально[1], многоагентный подход – это целостная парадигма к построению сложных систем, состоящих из взаимодействующих агентов, которые, оперируя локальными знаниями и ограниченными возможностями (ресурсами), способны, благодаря своему взаимодействию, достигать ожидаемого глобального поведения системы. Так, муравьи образуют муравейники – превосходно организованные общества с разделением труда; взаимодействие множества людей порождает мировую экономику, а взаимодействие живых клеток создает функционирующий организм. Обучение многоагентных систем имеет свои базовые принципы:

1. Один агент не может обучаться отдельно от других агентов.
2. Структуры взаимодействий внутри многоагентной системы должны быть включены в алгоритм обучения.
3. Многоагентное обучение – это во многом принципиально новый вид обучения, не сводящийся к раздельному обучению отдельных агентов[1].

Цель данной работы – адаптация подкрепляющего обучения для многоагентных систем. Как результат – мы получаем новый класс алгоритмов обучения для широкого круга задач. Рассмотрим более подробно подкрепляющее обучение.

Подкрепляющее обучение

Подкрепляющее обучение (Reinforcement Learning, RL)– это область искусственного интеллекта и способ машинного обучения, предназначенная для обучения автономных агентов путем их взаимодействия с внешней средой для достижения в ней оптимального поведения [2]. Подкрепляющее обучение возникло на пересечении таких областей наук, как динамическое программирование, машинное обучение, исследование рефлексов, когнитивные процессы и биология живых организмов. Подкрепляющее обучение применяется в робототехнике, теории адаптивного управления, теории адаптивного обучения и в теории игр.

Стандартная модель подкрепляющего обучения имеет вид [2,3]:

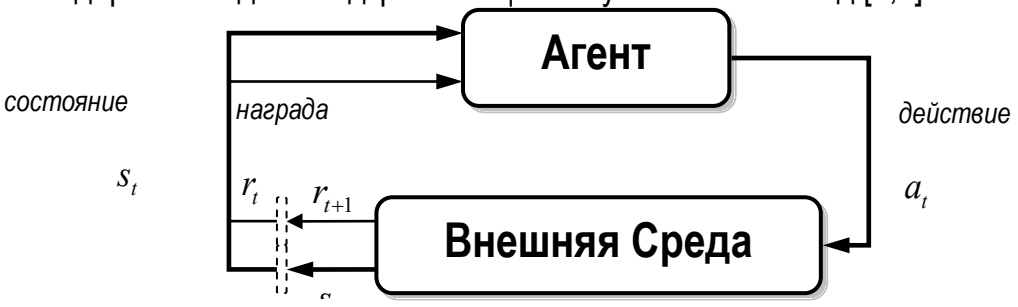

Рисунок 1 – Стандартная модель подкрепляющего обучения

Агент взаимодействует с внешней средой в дискретные моменты времени $t = 0,1,2,3,....$. В каждый момент времени t агент получает некоторое представление внешней среды – внутреннее состояние $s_t \in S$, где S - это множество всех возможных состояний. На основании текущего состояния выбирается действие $a_t \in A(s_t)$, где $A(s_t)$ множество возможных действий в состоянии s_t. В следующий момент времени, после выполнения действия, агент и переходит в новое состояние s_{t+1}, и получает численную **награду** $r_{t+1} = \Re(s_t, a_t, s_{t+1})$, где $\Re(s_t, a_t, s_{t+1})$ - функция награды. Имея значение награды, агент может оценить качество или полезность перехода $s_t \xrightarrow{a_t} s_{t+1}$ и произвести обучение. В каждый момент времени агент осуществляет отображение из состояния в вероятности выбора каждого возможного действия. Это отображение называется *политикой* π агента и обозначается $\pi(s,a)$, где при условии, что $s = s_t$, выбирается некоторое $a = a_t$.

Награда играет роль обратной связи, говорящей о том, какое действие было «плохим» или «хорошим». По сути, подкрепляющее обучение – это метод проб и ошибок. Агент исследует среду и, если находит положительное действие, получает положительную награду, и наоборот – отрицательную; алгоритм подкрепляющего обучения устроен таким образом, что вероятность в будущем выбрать действие с положительной наградой возрастет.

Математическое обоснование подкрепляющего обучения получено из динамического программирования, а также статистических методов моделирования (метод Монте-Карло), и основано на итеративном варианте формулы Беллмана. Подкрепляющее обучение работает на Марковских процессах принятия решений (МППР). Другими словами, среда, в которой действует агент, должна описываться МППР; как следствие, подкрепляющее обучение хорошо подходит к теории игр.

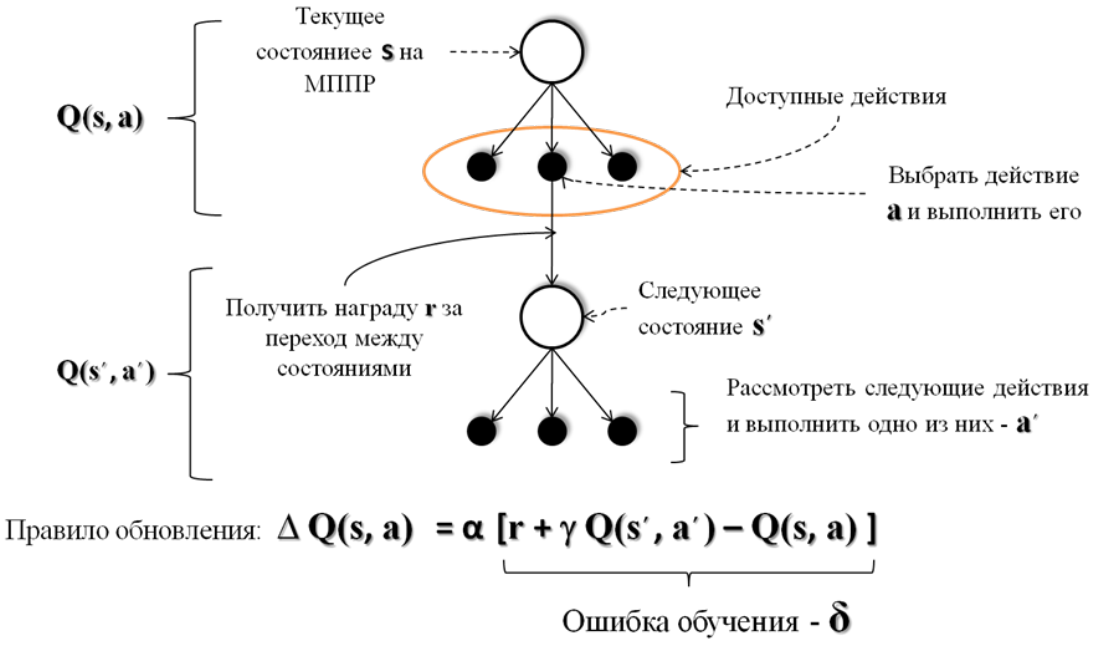

Рисунок 2 - Итеративное подкрепляющее обучение по методу временной разности

Где $Q(s,a)$ – это значение Q-функции, функции полезности некоторого действия *a* в некотором состоянии *s*. Разность ценности двух пар состояние-действие с учетом награды составляет ошибку обучения δ, которая может быть использована для обновления значений ценности:

$$Q(s,a) = Q(s,a) + \alpha[r_t + \gamma Q(s',a') - Q(s,a)] \quad (1)$$

Другой, наиболее известный алгоритм подкрепляющего обучения называется Q-*Learning* и выполняет обновление по следующей формуле:

$$Q(s,a) = Q(s,a) + \alpha[r_t + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (2)$$

Обучение законченно, когда ошибка обучения δ становится равной 0, или меньше заданного порога, тогда оптимальная политика выбирается при помощи следующего выражения:

$$\pi^* = \max Q(s,a) \quad (3)$$

Для хранения Q-функции используется либо таблица (для небольших пространств), либо аппроксиматор. Основное преимущество аппроксиматора – способность проводить обобщение в пространстве состояний-действий. Аппроксиматором может быть любая параметрическая функция, например, деревья решений, нейронные сети, RBF-функция и другие.

Основным преимуществом подкрепляющего обучения является способность адаптироваться к обучаемой среде; она может быть полностью неизвестна, но и даже в этом случае агент может достичь оптимального поведения.

Многоагентное подкрепляющее обучение

В первых работах по коллективному подкрепляющему обучению [5-6] был использован подход, при котором вся Многоагентная система рассматривалась как один агент – объединенное обучение. Такой подход позволял успешное обучение, но оно ничем не отличалось от стандартной RL модели. Сходимость алгоритма была медленной, а пространство состояний-действий характеризовалось большой размерностью. Возникла потребность в многоагентном обучении, которое решало бы вышеназванные проблемы подкрепляющего обучения. Эта модель была разработана и представлена в [7]. Основные её положения сводятся к следующим принципам:

1. Любую сложную задачу всегда можно разложить на подзадачи и поручить её решение отдельным агентам. Награда задает некоторую цель и корректирует процесс её достижения. Следовательно, *выделив награды, можно выделить цели*. Но способ начисления наград зависит от структуры многоагентной системы; один агент может начислять награду другому и т.д. Данный принцип получил название принципа декомпозиции и лежит в основе предложенного проекционного подхода к подкрепляющему обучению. Он позволяет декомпозировать задачу обучения на подзадачи, чем во многом решает проблему проклятья размерности. Для полноценного обучения декомпозиции недостаточно.

2. Тесная взаимосвязь агентов, выраженная в направленности на решение общей задачи, делает их обучение взаимосвязанным. Аналогичное решение было применено и к подкрепляющему обучению. В многоагентной системе действия агентов направлены на других агентов, не только во внешнюю среду, как показано на рисунке 1. Если предположить, что значение $Q(s',a')$ в формулах (1) и (2) может быть получено от другого агента, то становится возможным обучение агентов относительно друг друга. Такое разделение отражает тот факт, что текущее состояние одного агента может зависеть от действий другого. Таким образом, можно определить ошибку действий одного агента, относительно другого.

Данные принципы были опробованы на задаче управления многозвенным роботом. Многозвенный робот - это робот с N степенями свободы, где N - количество узлов. Каждый узел робота представляет собой одного агента. Каждый узел робота, кроме последнего, может изменять положение, ориентацию всех последующих узлов относительно своей позиции на 360 градусов. Другими словами, произвольный узел может вращать всю последующую структуру робота по всей окружности с центром в точке с данным узлом (радиус вращения зависит от количества сегментов робота и размерности среды). Центральный (корневой) узел робота не изменяет своего местоположения. Последний, терминальный узел, не изменяет своего местоположения; его положение зависит от со-

гласованных действий предыдущих агентов. *При обучении необходимо научиться согласовывать движение частей робота таким образом, чтобы терминальный элемент попадал в целевую точку*.

В задаче управления пятизвенным роботом, благодаря декомпозиции, пятимерное пространство состояний действий было преобразовано в пять одномерных, обучение в которых происходит значительно быстрее. Относительное обучение позволило обучаться с учетом структуры многоагентной системы.

Моделирование многоагентной системы, построенной на вышерассмотренных принципах, показало следующие результаты. Во первых, сходимость обучения была на порядок быстрее, чем при объединенном обучении (60-100 эпизодов против 1000). Во вторых, была выявлена способность многоагентной системы к синхронизации действий. В третьих, был наглядно продемонстрирован принцип, что «обучение – это обобщение», т.к. робот легко перестраивался с цели на цель. В четвертых, стратегия поведения робота значительно изменялась в зависимости от выбранного алгоритма обучения. В пятых, моделирование показало наиболее оптимальное поведение для данного робота, которое значительно отличалось от ожидаемого.

Литература

1. Hose M. Vidal. *Fundamentals of Multiagent Systems with Net Logo Examples.* (www.multiagent.com)

2. Richard S. Sutton, Andrew G. Barto. *Reinforcement Learning: An Introduction* Cambridge : MIT Press., 1998

3. Tesauro, G. J. (1994). TD--gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215--219. (http://www.research.ibm.com/massive/tdl.html)

4. Dr. Florentin Woergoetter, Dr. Bernd Porr. Статья *Reinforcement Learning* на ресурсе *http://www.scholarpedia.org. (*http://www.scholarpedia.org/article/Reinforcement_learning).

5. Кабыш, А.С. *Коллективное поведение агентов на основе подкрепляющего обучения.* Нейроинформатика / А.С. Кабыш, В.А. Головко. – 2009. – Часть 1. – С. 191-200.

6. Kabysh, A.S., Golovko V.A., *Collective Behavior in Multiagent Systems Based on Reinforcement Learning*, PRIP-2009: Proceedings of the Tenth International Conference (19-21 May, Minsk, Republic of Belarus) / Kabysh, A.S., Golovko V.A. – Minsk, 2009. – С. 260-264.

7. Kabysh, A.S. *Collective Behavior in Multi-Agent Systems,* OWD 2009 Ph.D. workshop, Eastern Europe Summer School, 12-24 October, Selesian University of Technology. – Poland. – 2009. – P. 92-97.

УДК 004.89

ИСПОЛЬЗОВАНИЕ ИСКУССТВЕННЫХ ИММУННЫХ СИСТЕМ И НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБНАРУЖЕНИЯ КОМПЬЮТЕРНЫХ АТАК

Комар М.П.

Тернопольский национальный экономический университет, г. Тернополь, Украина

В настоящее время обеспечение безопасности информации является одной из ключевых задач. Развитие компьютерных сетей и их объединение в глобальную сеть Интернет привело к росту числа преступлений, связанных с нарушением основополагающих принципов информационной безопасности: доступности, целостности и конфиденциальности информации. Несмотря на развитие средств защиты, таких как брандмауэры, количество