В.Г. РЕДЬКО, Г.А. БЕСХЛЕБНОВА
Научно-исследовательский институт системных исследований РАН,
Москва
E-mail: vgredko@gmail.com , gab19@list.ru

МОДЕЛЬ АДАПТИВНОГО ПОВЕДЕНИЯ АВТОНОМНЫХ АГЕНТОВ В ДВУМЕРНОЙ КЛЕТОЧНОЙ СРЕДЕ

Ставится задача моделирования адаптивного поведения автономных агентов с несколькими потребностями (питание, размножение, безопасность). Рассматривается поведение агентов в двумерной клеточной среде. Каждый агент способен выполнять следующие действия: делиться, съедать пищу, перемещаться на одну клетку вперед, поворачиваться направо или налево, наносить удар по соседнему агенту, отдыхать. Проведено моделирование для упрощенной задачи, для которой деление, удары по соседним агентам и отдых отсутствуют. Показана возможность формирования цепочек целенаправленных действий.

Введение

В настоящей работе ставится задача построения биологически инспирированных моделей поведения автономных агентов с несколькими потребностями: питания, размножения, безопасности. Хотя в настоящее время имеется множество многоагентых моделей, тем не менее, многие вопросы, относящиеся к процессам формирования адаптивного поведения агентов с указанными выше основными потребностями, к накоплению и использованию необходимых знаний, остаются неисследованными. Например, хорошо известна «сахарная модель» [1], в которой анализируются процессы питания и поиска пищи (сахара). Но в этой модели рассматривается только одна потребность питания, при этом не учитываются переходы между типами поведения с разными потребностями и познавательные процессы, необходимые при формировании адаптивного поведения.

Основной целью данной работы являлось выполнение первого этапа разработки моделей автономного поведения, включающего в себя создание общей схемы построения моделей эволюционирующей популяции агентов, а также анализ методов формирования адаптивного поведения и познавательных способностей агентов. Проведено моделирование для упрощенной задачи, в которой рассматривается один самообучающийся агент с ограниченным числом действий, при этом поведение агента фор-

мируется с помощью обучения с подкреплением. Для такого агента показана возможность формирования цепочек действий, приводящих к пополнению ресурса агента. На основе проведенного моделирования намечены пути дальнейшей работы по построению моделей автономных адаптивных агентов.

1. Общая схема моделирования

В этом разделе излагается общая схема моделирования эволюции популяции автономных агентов. Для определенности рассматриваются агенты, система управления которых основана на классифицирующих системах [2], представляющих собой набор логических правил, формируемых как в процессе эволюции популяции, так и путем самообучения агентов. Этот набор правил составляет геном агента. Другие варианты систем управления агентов анализируются в разделе 3.

Предполагается, что имеется двумерная клеточная среда, в которой эволюционирует популяция, состоящая из n агентов. В любой клетке может находиться только один агент. Каждый агент имеет свое направление «вперед». В некоторых клетках, число которых фиксировано, имеется пища агентов, величина порции пищи в каждой из этих клеток тоже фиксирована. Агент обладает ресурсом $E(t)$. Ресурс агента увеличивается при съедании им пищи и уменьшается при выполнении им действий. Агенты функционируют в дискретном времени, $t = 0,1,\ldots$

В течение каждого такта времени агент выполняет одно из следующих действий: деление, питание, перемещение на одну клетку вперед, поворот направо или налево на 90°, нанесение удара по агенту, находящемуся впереди данного, отдых.

Действие «деление» происходит следующим образом: рождается потомок данного агента в одной из соседних клеток, случайно выбираемой; если все соседние клетки данного агента заняты, то потомок не рождается; геном рождающегося потомка отличается от генома родителя случайными мутациями.

При выполнении действия «питание» агент съедает всю порцию пищи в той клетке, в которой он находится.

Если агент ударяет находящегося впереди него другого агента, то нападающий агент отнимает у ударяемого определенный ресурс. Если оба агента нападают друг на друга, то ресурс обоих уменьшается на величину, расходуемую на действие «ударить».

Размер двумерного мира равен $N_x N_y$ клеток ($x = 1,\ldots,N_x$; $y = 1,\ldots,N_y$).

Клеточный мир замкнут: если агент, находящийся в клетке с координатой $x = N_x$, движется вправо (т.е. пересекает «границу мира»), то он перемещается в клетку с координатой $x = 1$, аналогично происходит движение агента при пересечении других границ мира.

Выбор действий агента обеспечивается имеющейся у него системой управления. Система управления агента представляет собой набор правил вида:

$$R_k = \mathbf{S}_k(t), A_k(t) \rightarrow \mathbf{S}_k(t+1), \qquad (1)$$

где $\mathbf{S}_k(t)$ – текущая ситуация, $A_k(t)$ – действие, соответствующее этому правилу, $\mathbf{S}_k(t+1)$ – ситуация, ожидаемая в следующий такт времени, k – номер правила. Набор правил вида (1) представляет собой классифицирующую систему [2]. Каждое правило имеет свой вес W_k , веса правил модифицируются при обучении агента. $\mathbf{S}_k(t)$ есть вектор, компоненты которого принимают значения 0, 1, либо #. Значения 0 и 1 характеризуют внешнюю среду агента, например, они соответствуют отсутствию или наличию порции пищи в определенной клетке. Символ # означает, что соответствующая компонента не имеет значения при применении правила.

Каждый такт времени агент осуществляет выбор действия и обучается. Выбор действия осуществляется следующим образом. Определяется текущая ситуация $\mathbf{S}(t)$ и формируется набор выделенных правил $\{\mathbf{R}\}$, в этот набор включатся правила, для которых все существенные (не равные #) компоненты вектора $\mathbf{S}_k(t)$ совпадают с компонентами вектора $\mathbf{S}(t)$. Из правил, входящих в $\{\mathbf{R}\}$, выбирается правило, определяющее действие агента в данный такт времени. При этом используется ε-жадный метод [3]: с вероятностью $1-\varepsilon$ выбирается то правило из числа выделенных $\{\mathbf{R}\}$, для которого вес правила W_k максимален, с вероятностью ε выбирается произвольное правило из $\{\mathbf{R}\}$ ($1 >> \varepsilon > 0$). Далее выполняется действие $A_k(t)$, соответствующее номеру выбранного правила.

При обучении веса правил W_k модифицируются методом обучения с подкреплением [3], который состоит в следующем. Меняется вес того правила, которое использовал агент в предыдущий такт времени t-1, этот вес изменяется в соответствии с изменением ресурса агента при переходе к такту t и весом правила, применяемого в такт t. Пусть вес правила, примененного в такт t-1, равен $W(t\text{-}1)$, вес правила, применяемого в такт t, равен $W(t)$, ресурс агента в эти такты времени равен $E(t\text{-}1)$ и $E(t)$ соответственно. Тогда изменение веса $W(t\text{-}1)$ равно

$$\Delta W(t-1) = \alpha\left[E(t) - E(t-1) + \gamma W(t) - W(t-1)\right], \qquad (2)$$

где α – параметр скорости обучения, γ – дисконтный фактор; $0 < \alpha << 1$, $0 < \gamma < 1$, $1-\gamma << 1$.

Процесс эволюции популяции агентов предполагает, что при делении ресурс родителя делится пополам между родителем и потомком. Если ресурс агента $E(t)$ в результате его действий становится меньше определенного порога E_{min} ($E(t) < E_{min}$), то данный агент погибает.

Изложенная схема, хотя и конкретна, тем не менее, она предлагает общую структуру моделей формирования адаптивного поведения автономных агентов. Отдельные алгоритмы в рамках этой структуры могут быть заменены другими. Например, набор правил классифицирующей системы может быть заменен нейросетевым адаптивным критиком, обеспечивающим оценку качества ситуаций и прогноз будущих ситуаций [4], метод обучения с подкреплением может быть заменен методом семантического вывода [5]. Таким образом, в рамках данной структуры возможны варианты конкретных систем управления агентов. Подробнее направления развития изложенной схемы обсуждаются в разделе 3. Но перед этим обсуждением приведем пример компьютерного моделирования, в котором анализируется возможность формирования цепочек действий рассматриваемых автономных агентов.

2. Пример компьютерного моделирования. Формирование цепочек действий

Предложенная схема модели формирования адаптивного поведения автономных агентов была реализована в виде компьютерной программы, на основе которой выполнен ряд расчетов. В данном разделе приводятся результаты расчетов по упрощенной версии программы, с помощью которой анализировалась возможность формирования цепочек действий одним самообучающимся агентом. Для простоты мы ограничивались только одной потребностью – потребностью питания. Действия деления, удары по другим агентам, а также отдых не рассматривались. Агент мог выполнять только 4 действия: питаться, двигаться вперед и поворачиваться на 90° направо либо налево. Также были упрощены векторы ситуаций: рассматривалось поле зрения агента, состоящее из 4-х клеток: той клетки, в которой находится агент, клетки впереди, слева и справа агента. Для одного отдельного агента вектор ситуации характеризовал наличие пищи в этих 4-х клетках, т.е. было всего $2^4 = 16$ возможных ситуаций. Прогноз

будущих ситуаций в правилах (1) в упрощенной модели не учитывался.

Итак, для одного отдельного агента можно определить 16 возможных ситуаций и 4 действия. Классифицирующая система, управляющая агентом, включала 64 всевозможных правила. Веса правил исходно задавались случайно, а в последующем модифицировались с помощью метода обучения с подкреплением. В соответствии с весами выбирались действия агента. Порог минимального ресурса агента считался равным нулю: $E_{min} = 0$.

Перед агентом ставились следующие простые задачи.

А) Агент помещался в клетку, в которой имеется порция пищи. В остальных клетках пища отсутствовала. Если агент съедал пищу, то порция пищи вновь помещалась в эту же клетку. Агент должен был научиться каждый такт времени выполнять действие «питание».

Б) Агент помещался в центр мира из 9 клеток, координаты которых равны $x = 1, 2, 3$; $y = 1, 2, 3$. Исходные координаты агента были равны $x = 2$, $y = 2$. Порция пищи была одна, она помещалась в клетку с координатами $x = 2$, $y = 3$. Исходное направление «вперед» у агента соответствовало направлению «вверх», т.е. в ту точку, в которой находится пища. После того, как пища была съедена, и агент, и новая порция пищи помещались в исходные клетки. Направление «вперед» у агента после съедания пищи не менялось. Агенту необходимо было сформировать цепочку действий: 1) «перемещение вперед», 2) «питание».

В) Модификация задачи Б). Исходные координаты агента равны $x = 2$, $y = 2$. Направление «вперед» у агента соответствовало направлению «вверх». Пища помещалась справа от агента, в клетку с координатами $x = 3$, $y = 2$. После того, как пища была съедена, агент и новая порция пищи помещались в исходные клетки. Направление «вперед» у агента после приема пищи не менялось. Агенту необходимо было сформировать цепочку действий: 1) «поворот направо» (т.е. повернуться в сторону клетки с пищей), 2) «перемещение вперед», 3) «питание».

Для каждой задачи расчет проводился для нескольких случаев, которые различались исходными случайными весами правил агента W_k. Получены следующие результаты моделирования.

Задача А легко решалась. Агент достаточно быстро находил нужное действие. Пример роста ресурса агента для этой задачи представлен на рис. 1.

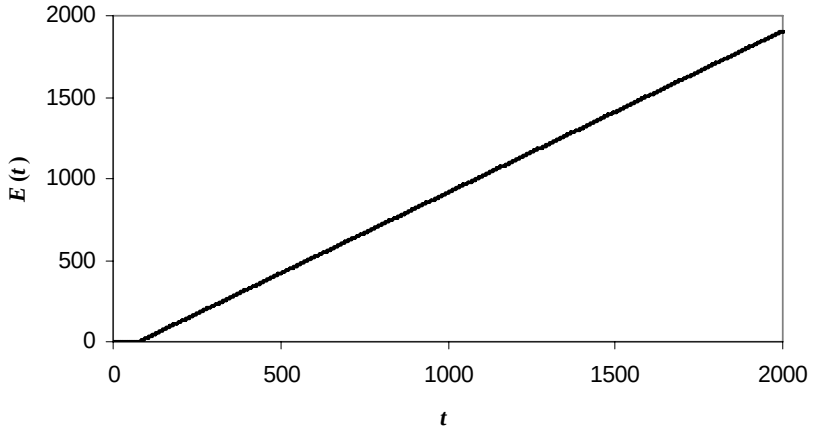

Рис. 1. Зависимость ресурса агента $E(t)$ от номера такта времени t. Агент находит решение задачи А после $t = 75$ тактов времени. Параметры расчета: $\varepsilon = 0$, $\alpha = 0.1$, $\gamma = 0.9$. Расход ресурса на каждое действие равен 0.01. Увеличение ресурса при съедании порции пищи равно 1.0. Если агент при выполнении действий расходовал весь свой ресурс, то агент заменялся новым. Начальный ресурс нового агента равен 1.0.

Задача Б также решалась (правда, несколько дольше, чем задача А). Однако при этом для немного модифицированной задачи, когда направление «вперед» после съедания пищи не задавалось, было найдено и неожиданное решение, при котором агент сначала перемещался вниз ($x = 2$, $y = 1$), затем через границу замкнутого мира попадал в клетку с порцией пищи ($x = 2$, $y = 3$). Неожиданное решение обусловлено тем, что агент сначала случайным образом менял направление «вперед», переворачивался, при этом направление «вперед» соответствовало направлению «вниз», и только после этого находил цепочку из 3-х звеньев: 1) «перемещение вперед», 2) «перемещение вперед», 3) «питание». Примеры изменения ресурса агента для обычного и неожиданного решений приведены на рис. 2.

Также наблюдалось и правильное решение задачи В.

Итак, проведены компьютерные расчеты для простых задач формирования цепочек действий при поиске пищи.

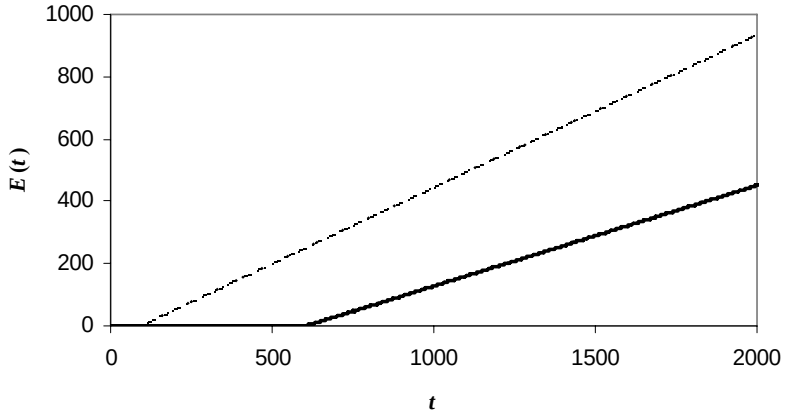

Рис. 2. Зависимость ресурса агента $E(t)$ от номера такта времени t. Пунктирная линия – зависимость ресурса от времени для обычного решения, сплошная жирная линия – для неожиданного решения (после изменения ориентации «вперед» на противоположную). В первом случае агент находит решение задачи Б после $t = 99$ тактов времени, во втором случае – после $t = 611$ тактов времени. Параметры расчета такие же, как на рис. 1.

3. Направления дальнейшей работы

Реализованная версия программы выглядит несколько упрощенной, так как расчеты проведены для довольно простых задач. Поэтому необходимо развитие методов формирования адаптивного поведения автономных агентов. Для промоделированных задач нетрудно представить такое развитие. Например, можно рассмотреть прогнозы всех 4-х действий для всех 16-ти ситуаций. Используя эти прогнозы, можно путем перебора составить цепочку действий, приводящих к увеличению ресурса агента.

Изложенный подход близок к постановке работ по проекту «Животное», который разрабатывался М.М. Бонгардом с сотрудниками в 1970 годах [6]. В этом проекте была рассмотрена общая схема поведения модельных организмов, использующих запоминание фактов, имеющих вид, близкий к правилам по формуле (1). Хотя проект «Животное» по существу не был разработан, он, как и предложенная выше структура, представляет биологически инспирированную общую схему автономного поведения с несколькими потребностями.

Рассмотрим подробнее методы формирования автономного поведения

в рамках изложенного подхода.

Выше изложен метод обучения с подкреплением для процедуры обучения агентов. Проанализируем возможности других вариантов обучения. Во-первых, этот метод можно модифицировать, применив его не к оценке правил выбора действий, а к оценке прогнозируемых ситуаций. Во-вторых, интересно проанализировать возможность замены метода обучения с подкреплением методами семантического вывода [5]. Последнее весьма важно, так как дает связь с когнитивными методами познания закономерностей взаимодействия с внешней средой. Например, интересно проанализировать соотношение между методом семантического вывода и известным ДСМ-методом индуктивного вывода [7], используемым в направлении исследований «Искусственный интеллект». Сокращение ДСМ происходит от инициалов Джона Стюарта Милля, подход которого к правдоподобным рассуждениям был положен в основу ДСМ-метода. Также имеет смысл рассмотреть биологически инспирированные схемы обучения с прогнозом будущих ситуаций, например, такие, какие были использованы в работе [8].

Другим важным направлением будущих исследований является анализ поведения агентов с несколькими потребностями, например, с потребностями питания и безопасности. При высокой потребности питания агенты должны искать скопления пищи, при высокой потребности безопасности агенты должны избегать области скопления других агентов. Для моделирования потребности питания имеет смысл ввести мотивацию к поиску пищи $M(t)$ и анализировать динамику мотивации аналогично работе [9]. При наличии нескольких потребностей возможно формирование иерархических систем управления агентом. В частности, переключение от одного блока управления к другому возможно на основе регулирования одной или нескольких мотиваций $M(t)$.

В заключение отметим, что анализируемые методы формирования адаптивного поведения автономных агентов составляют определенный этап работ по проекту «Мозг анимата», направленному на изучение адаптивного поведения модельного организма (анимата) с несколькими естественными потребностями [10,11].

Итак, предложена общая схема моделирования процессов, на основе которых формируется адаптивное поведение автономных агентов с несколькими потребностями. Проведено компьютерное моделирование обучения цепочкам действий. Намечены пути развития моделей путем усовершенствования методов обучения и моделирования когнитивных процессов познания закономерностей взаимодействия с внешней средой и

использования этих закономерностей при адаптивном поведении.

Список литературы

1. Макаров В.Л. Искусственные общества // Искусственные общества. 2006. Т.1. № 1. С.10-24. См. также: http://www.artsoc.ru/docs/Journal/1.pdf
2. Holland J.H., Holyoak K.J., Nisbett R.E., Thagard P. Induction: Processes of Inference, Learning, and Discovery. Cambridge, MA: MIT Press, 1986.
3. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. MIT Press, 1998.
4. Редько В.Г., Прохоров Д.В. Нейросетевые адаптивные критики // Научная сессия МИФИ-2004. VI Всероссийская научно-техническая конференция «Нейроинформатика-2004». Сборник научных трудов. Часть 2. М.: МИФИ, 2004. С.77-84.
5. Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. Новосибирск: НГУ, 2006. 293с.
6. Бонгард М.М., Лосев И.С., Смирнов М.С. Проект модели организации поведения – «Животное» // Моделирование обучения и поведения. М.: Наука, 1975. С.152-171.
7. Финн В.К. О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф.Бэкона - Д.С. Милля // Семиотика и информатика. 1983. М.: ВИНИТИ. Вып. 20. С.35-101.
8. Butz M.V., Hoffmann J. Anticipations control behavior: animal behavior in an anticipatory learning classifier system // Adaptive Behavior, Vol. 10, No. 2, 75-96 (2002).
9. Непомнящих В.А., Попов Е.Е., Редько В.Г. Бионическая модель адаптивного поискового поведения // Известия РАН. Теория и системы управления. 2008. № 1. С.85-93.
10. Анохин К.В., Бурцев М.С., Зарайская И.Ю., Лукашев А.О., Редько В.Г. Проект «Мозг анимата»: разработка модели адаптивного поведения на основе теории функциональных систем // Восьмая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. М.: Физматлит, 2002. Т.2. С.781-789.
11. Red'ko V.G., Anokhin K.V. et al. Project “Animat Brain”: Designing the animat control system on the basis of the functional systems theory // In Butz, M.V., Sigaud, O., et al (Eds.), Anticipatory Behavior in Adaptive Learning Systems: From Brains to Individual and Social Behavior. LNAI 4520, Berlin, Heidelberg: Springer Verlag. 2007. PP. 94-107.