МОДЕЛЬ АВТОНОМНЫХ АГЕНТОВ В ДВУМЕРНОЙ КЛЕТОЧНОЙ СРЕДЕ

Г.А. Бесхлебнова, В.Г. Редько
Научно-исследовательский институт системных исследований РАН, Москва
E-mail: gab19@list.ru

Аннотация. Построена и исследована модель поведения автономных агентов в двумерной клеточной среде. Формирование поведения происходит как путем самообучения агентов, так и в результате эволюции популяции агентов. Показана возможность формирования естественного поведения агентов в популяции. Изолированные агенты в результате самообучения формировали цепочки действий, приводящие к нахождению пищи.

Описание модели

В настоящей работе изучается биологически инспирированная компьютерная модель поведения автономных агентов с несколькими естественными потребностями: питания, размножения, безопасности.

Предполагается, что имеется двумерная клеточная среда, в которой эволюционирует популяция агентов. В любой клетке может находиться только один агент. Каждый агент имеет свое направление «вперед». В фиксированном числе клеток имеется пища агентов, величина порции пищи также фиксирована. Когда агент выполняет действие «питание», то он съедает всю порцию пищи в той клетке, в которой он находится; при этом новая порция помещается в случайно выбранную клетку. Агент обладает ресурсом, ресурс агента увеличивается при съедании им пищи и уменьшается при выполнении им действий.

Время дискретно. Каждый такт времени t агент выполняет одно из следующих действий: деление, питание, перемещение на одну клетку вперед, поворот направо или налево, нанесение удара по агенту, находящемуся впереди данного, отдых. Если один агент ударяет другого, то он отнимает ресурс у ударяемого. Система управления агента основана на наборе логических правил вида «если ситуация $\mathbf{S}_k(t)$, то действие $A_k(t)$». Ситуация $\mathbf{S}_k(t)$ определяется наличием пищи и других агентов в ближайшем окружении данного. Каждое правило имеет свой вес W_k , веса правил модифицируются при обучении агента.

Каждый такт времени агент осуществляет выбор действия и обучается. При выборе действия часто использовался «метод отжига»: на начальных тактах моделирования, когда логические правила еще не сформированы, действия агентов выбирались преимущественно случайно, а затем вероятность случайного выбора постепенно уменьшалась, и выбор действия осуществлялся в соответствии с правилами и их весами (чем больше вес правила, тем больше приоритет его использования). При обучении веса правил W_k модифицировались методом обучения с подкреплением [2]. Сигналами поощрения или наказания служили изменения ресурса агента. Изменение весов W_k при обучении происходит следующим образом. 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 $R(t\text{-}1)$ и $R(t)$, соответственно. Тогда изменение веса $W(t\text{-}1)$ равно [2]:

$$\Delta W(t\text{-}1)=\alpha[R(t)\text{-}R(t\text{-}1)+\gamma W(t)\text{-}W(t\text{-}1)], \quad (1)$$

где α – параметр скорости обучения, γ – дисконтный фактор; $0 < \alpha << 1$, $0 < \gamma < 1$, $1\text{-}\gamma << 1$. В результате обучения увеличивались веса тех правил, применение которых приводило к росту ресурса агента.

Процесс эволюции популяции агентов предполагает, что при делении агента ре-

сурс родителя делится пополам между родителем и потомком. Логические правила потомка отличаются от правил родителя малыми мутациями.

Результаты моделирования

Моделирование проводилось в рамках полной описанной модели и в рамках упрощенной версии. В последнем случае изучалось обучение одного агента.

В случае полной версии модели рассматривалась популяция, состоящая из 50 агентов, помещенная в мир из 100 клеток, в котором в половине клеток была случайно распределена пища. Было продемонстрировано, что в процессе эволюции и обучения агентов формировалось естественное их поведение: агенты преимущественно питались и часто отнимали ресурс друг у друга, изредка они выполняли и другие действия.

Пример расчета иллюстрируется рис. 1, 2. Параметры расчета были следующими. Порция пищи была в 50 клетках, прирост ресурса агента при съедании пищи был равен 1 (считаем ресурс безразмерным). Расход ресурса на каждое из действий, кроме удара, был равен 0.01, расход на удар составлял 0.02, при ударе ударяющий агент отнимал у ударяемого ресурс, равный 0.05. Параметры обучения с подкреплением равны: $\alpha = 0.1$, $\gamma = 0.9$. Применялся метод отжига: исходная вероятность случайного выбора действия была равна 1, характерное время уменьшения до нуля вероятности случайного выбора действия составляло 1000 тактов времени. Интенсивность мутаций была равна 0.

На рис. 1 представлена зависимость среднего по популяции ресурса агента $<R>$ от номера такта времени t. Видно, что в начале процесса ($t < 2000$) скорость роста $<R>$ мала, так как агенты еще плохо обучены. При $t > 5000$ скорость роста $<R>$ практически постоянна; стохастичность на этом участке обусловлена случайным перемещением агентов по клеткам мира, а также случайным размещением новых порций пищи в ячейках мира при съедании части пищи агентами.

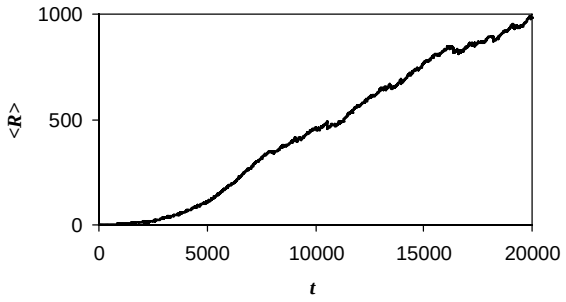

Рис. 1. Зависимость среднего по популяции ресурса агента $<R>$ от номера такта времени t.

На рис. 2 представлена зависимость числа агентов Ne, выполняющих действие «питание», от номера такта времени t. Видно, что при больших t примерно 30% агентов (из общего числа агентов популяции, равного 50) выполняют это действие. Наблюдаются сильные стохастические колебания числа Ne во времени. Примерно такая же зависимость от времени наблюдается и для числа агентов, выполняющих действие «нанесение удара»; число таких агентов при больших t равно примерно 15-20. Число агентов, выполняющих действие «деление», мало и составляет около 1. Число агентов, выполняющих каждое из остальных действий, равно примерно 3-5.

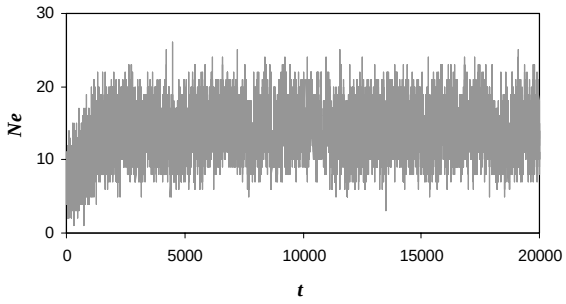

Рис. 2. Зависимость числа агентов Ne, выполняющих действие «питание», от номера такта времени t.

Таким образом, агенты обучаются выполнять преимущественно действия «питание» и «нанесение удара», приводящие к увеличению ресурса агента, и избегать действия «деление», которое приводит к уменьшению ресурса. Все остальные действия выполнялись с небольшой частотой.

В упрощенных версиях модели анализировалась возможность формирования цепочек действий одним самообучающимся

агентом. Рассматривалось два варианта формирования цепочек. В первом варианте агент мог выполнять только 4 действия: питаться, двигаться вперед и поворачиваться направо либо налево. Считалось, что есть только одна расположенная в определенной клетке порция пищи. Агенту надо было сформировать цепочку из одного, двух или трех заданных действий. Например, трехзвенная цепочка включала следующие действия: 1) «поворот направо», 2) «перемещение вперед», 3) «питание»; порция пищи исходно располагалась в клетке справа от той клетки, в которую исходно помещался агент. Основные параметры расчета были такими же, как и для полной модели. Метод отжига в этом варианте не использовался. Расчеты показали, что простые одно-, двух- и трехзвенные цепочки действий достаточно легко формировались в процессе самообучения агента.

Во втором варианте упрощенной версии к указанным 4-м действиям добавлялось еще действие «отдых». В этом случае было возможным формирование цепочек действий в мире, в котором, как и в полной модели, в половине клеток была случайно распределена пища. Применялся метод отжига, приводящий к случайному выбору действий на начальных тактах моделирования. Как и в первом варианте, основные параметры расчета были такими же, как в полной модели. Расчеты показали, что в этом варианте формировались заранее неизвестные цепочки действий из нескольких звеньев, приводящие к нахождению пищи. Зависимость ресурса R агента от времени для данного случая показана на рис. 3.

Анализ результатов для второго варианта показывает, что в конце расчета агент применяет только малое число логических правил, имеющих большой вес. Этот набор правил можно рассматривать как обобщающие эвристики, формируемые агентом в процессе самообучения. Эти эвристики сводятся к следующему. 1. Если есть порция пищи в той клетке, в которой находится агент, то нужно выполнить действие «питание». 2. Если нет порции пищи в той клетке, в которой находится агент, и есть пища в клетке впереди или справа/слева агента, то нужно выполнить действие «перемещение вперед» или «поворот направо/налево», соответственно. Следовательно, в процессе обучения агент самостоятельно формирует вполне естественную стратегию поведения.

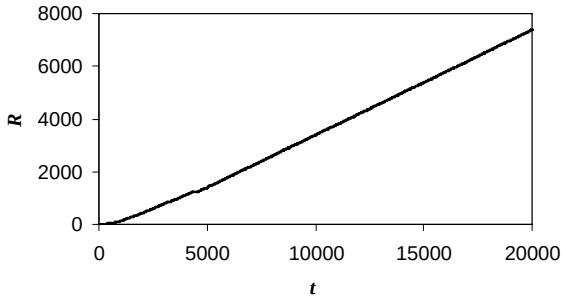

Рис. 3. Зависимость ресурса R самообучающегося агента от номера такта времени t.

Итак, построена модель автономных агентов, которые путем самообучения формируют свое поведение. Проведено компьютерное моделирование обучения цепочкам действий. Агенты обладают простыми когнитивными способностями: они запоминают закономерности взаимодействия с внешней средой в системе логических правил. Исследование поведения таких автономных агентов может рассматриваться как начальный этап моделирования когнитивной эволюции [1].

Работа выполнена при финансовой поддержке РФФИ (проект № 07-01-00180).

Список литературы

1. Редько В.Г. Перспективы моделирования когнитивной эволюции // Третья международная конференция по когнитивной науке: Тезисы докладов: в 2 т. М.: Художественно-издательский центр, 2008. Т. 2. С. 576-577.
2. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 1998.