В. Г. Редько, Г.А. Бесхлебнова. Моделирование адаптивного поведения автономных агентов. Нейрокомпьютеры: разработка, применение. № 3. 2010. С. 33-38.

УДК 32.81

Моделирование адаптивного поведения автономных агентов*

В. Г. Редько, Г.А. Бесхлебнова

Аннотация. Построена и исследована компьютерная модель адаптивного поведения автономных агентов, имеющих несколько естественных потребностей: питание, размножение, безопасность. Система управления агента основана на правилах вида «Если имеет место ситуация **S**, то необходимо выполнить действие *A*». Каждое правило имеет свой вес. Веса правил модифицируются как путем обучения с подкреплением, так и в процессе эволюционной оптимизации. Моделирование продемонстрировало формирование естественного поведения агентов.

Ключевые слова: формирование адаптивного поведения, автономные агенты, обучение с подкреплением, эволюционная оптимизация

1. Введение

Современные исследования адаптивного поведения и интеллектуальных систем включают активные работы по изучению автономных адаптивных агентов и их интеллектуальных и когнитивных свойств (см., например, [1-3]). Тем не менее, фактически отсутствуют работы по моделированию процессов формирования адаптивного поведения агентов с естественными потребностями. В настоящей работе исследуется модель адаптивного поведения автономных агентов, обладающих потребностями, аналогичными основным потребностям биологических организмов: питание, размножение, безопасность.

Система управления агентов формируется путем обучения агентов и эволюционной оптимизации. Каждый агент имеет ресурс, который пополняется при питании агента и расходуется при выполнении им действий. Обучение осуществляется методом обучения с подкреплением [4], т.е. путем самообучения агента на основе изменения его ресурса. Эво-

* Работа выполнена при финансовой поддержке Президиума РАН (Программа «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация», проект № 2.15) и РФФИ, проект № 07-01-180.

люционная оптимизация происходит в результате дарвиновской эволюции популяции агентов.

В разработанной модели агенты обладают элементарными когнитивными способностями: они запоминают закономерности взаимодействия с внешней средой в системе логических правил. Проведенное исследование может рассматриваться как начальный этап моделирования когнитивной эволюции [5, 6].

2. Описание модели

В модели предполагается, что имеется эволюционирующая популяция, численность популяции не превышает N_{max} агентов. Время t дискретно, t = 1,2,… Ресурс агента равен $R(t)$. $R(t)$ увеличивается при питании агента и уменьшается при выполнении им действий, а также в опасных ситуациях. Если ресурс агента $R(t)$ становится меньше определенного порога R_{min} , то данный агент умирает.

Мир, в котором находятся агенты, состоит из двух клеток: одна клетка является опасной для агентов, вторая – безопасной. С периодом T_D тактов времени статус клеток меняется: опасная клетка становится безопасной, и, наоборот, клетка, бывшая безопасной, становится опасной. Агент, находящийся в опасной клетке, каждый такт времени теряет ресурс r_D. В мире имеется пища агентов, которая с определенной вероятностью пополняется каждый такт времени t.

Каждый такт времени агент выполняет одно из следующих 4-х действий: деление, питание, перемещение в другую (альтернативную из двух) клетку, отдых.

Сенсорная система агента определяет ситуацию $\mathbf{S}(t)$, характеризующую внешнюю и внутреннюю среду агента. Вектор $\mathbf{S}(t)$ имеет 3 компоненты, принимающие значения 0 либо 1 и определяющие следующее: 1) есть ли в мире достаточное количество пищи, превышающее определенный порог f_{th} , 2) превышает ли собственный ресурс агента $R(t)$ заданный порог r_{th} , 3) опасна ли клетка, в которой находится агент. Таким образом, имеется 8 различных ситуаций $\mathbf{S}(t)$.

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

$$\mathbf{S}_k \rightarrow A_k , \qquad (1)$$

где $\mathbf{S}_k$ и A_k – ситуация и действие, соответствующие этому правилу, k – номер правила. Каждое правило имеет свой вес W_k , веса правил модифицируются при обучении агента. Так как общее число различных ситуаций равно 8, а число действий равно 4, то всего име-

ется 32 различных правила. Начальный набор весов этих правил $\{W_{0k}\}$, получаемый агентом от родителя (с небольшими мутациями), представляет собой геном агента. В противоположность геному текущие веса правил $\{W_k\}$, которые использует агент при выборе действия, модифицируются при обучении агента. Таким образом, каждый агент имеет два набора весов правил: начальные веса $\{W_{0k}\}$, составляющие геном агента и не меняющиеся в течение его жизни, и текущие используемые веса $\{W_k\}$, модифицируемые при жизни агента путем обучения. В момент рождения агента текущие веса равны начальным: $\{W_k\} = \{W_{0k}\}$.

Отметим, что описанные правила аналогичны логическим правилам в известных классифицирующих системах [7].

Чтобы учесть ограничение на возраст агентов, считается, что агент с определенной вероятностью P_d ($P_d << 1$) может погибнуть каждый такт времени (от случайных факторов), это соответствует средней продолжительности жизни агентов порядка $1/P_d$.

Опишем действия агента. Действие «деление» происходит следующим образом: рождается потомок данного агента, ресурс родителя делится пополам между родителем и потомком; геном $\{W_{0k}\}$ рождающегося потомка отличается от генома родителя случайными мутациями.

При выполнении действия «питание» агент съедает определенную часть r_{eat} пищи, если такое количество пищи имеется в данный такт времени в мире. Ресурс агента увеличивается на величину r_{eat}.

При выполнении одного из действий «деление», «питание», «перемещение» и «отдых» ресурс агента уменьшается на величину r_d, r_e, r_t, r_r, соответственно. Действия «деление» и «питание» соответствуют потребностям размножения и питания. Действие «перемещение» соответствует потребности безопасности, так как оно может обеспечить движение агента из опасной клетки в безопасную; моделирование показывает, что такое обеспечение действительно происходит.

Каждый такт времени агент осуществляет выбор действия и обучается. При выборе действия агента определяется текущая ситуация $\mathbf{S}(t)$ и выделяются 4 правила, для которых $\mathbf{S}_k = \mathbf{S}(t)$. Далее используется ε-жадный метод: с вероятностью $1-\varepsilon$ из этих выделенных правил выбирается то, для которого вес W_k максимален, а с вероятностью ε – произвольное из этих правил ($1 >> \varepsilon > 0$). Действие A_k, соответствующее выбранному правилу, выполняется.

При обучении веса правил W_k модифицируются методом обучения с подкреплением [4]. Сигналами поощрения или наказания служат 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 [4]:

$$\Delta W(t\text{-}1) = \alpha\,[R(t) - R(t\text{-}1) + \gamma W(t) - W(t\text{-}1)]\,, \qquad (2)$$

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

3. Результаты моделирования

При компьютерном моделировании изучалось поведение агентов и осуществлялся грубый подбор параметров, при котором формировалось достаточно естественное адаптивное поведение. Основные параметры расчета были следующими. Максимальная численность популяции составляла $N_{max} = 100$ (если численность популяции достигала величины N_{max} , то новые агенты не рождались). Расход ресурса на каждое из действий (r_d , r_e , r_t , r_r) был равен 0.01. Период смены статуса клеток (опасная ↔ неопасная) составлял $T_D = 100$ тактов времени. Уменьшение ресурса агента за один такт времени при нахождении его в опасной клетке было равно $r_D = 10$. Увеличение ресурса агента при питании составляло $r_{eat} = 10$. Вероятность гибели агента от случайных факторов составляла $P_d = 0.001$. Параметры обучения с подкреплением были равны: $\alpha = 0.1$, $\gamma = 0.9$. Параметр ε-жадного метода при случайном выборе правила составлял $\varepsilon = 0.1$. Изменение порогов R_{min} , f_{th} , r_{th} не сильно влияло на поведение агентов; в типичных расчетах эти величины составляли: минимальный ресурс агента $R_{min} = 0$, порог значимого количества пищи в мире $f_{th} = 10$, порог значимого собственного ресурса агента $r_{th} = 1$. Система управления каждого агента состояла из всех 32-х возможных правил, в начале расчета веса правил W_{0k} , составляющие геном агента, были случайными и малыми по сравнению с последующими значениями W_{0k}. При мутациях к начальным весам правил рождающихся агентов W_{0k} добавлялась случайная величина, равномерно распределенная в интервале $[-0.5P_m, 0.5P_m]$; $P_m = 0.1$ – ин-

тенсивность мутаций. В каждый такт времени при выборе действия каждого агента с вероятностью 0.5 в мир добавлялась порция пищи, равная 10.

Специальным выбором параметров задавались следующие случаи:

– Случай L (чистое обучение); в этом случае интенсивность мутаций полагалась нулевой: $P_m = 0$.

– Случай E (чистая эволюция), в этом случае интенсивность обучения была нулевой, вероятность выбора случайного правила также обнулялась: $\alpha = 0$ и $\varepsilon = 0$.

– Случай LE (обучение + эволюция), т.е. полная модель, с приведенными выше параметрами.

На рис. 1 представлены зависимости среднего по популяции ресурса агента <*R*> от номера такта времени *t* для случаев чистого обучения L и чистой эволюции E. Зависимости усреднены по 100 различным расчетам, выполненным для разных используемых последовательностей случайных чисел.

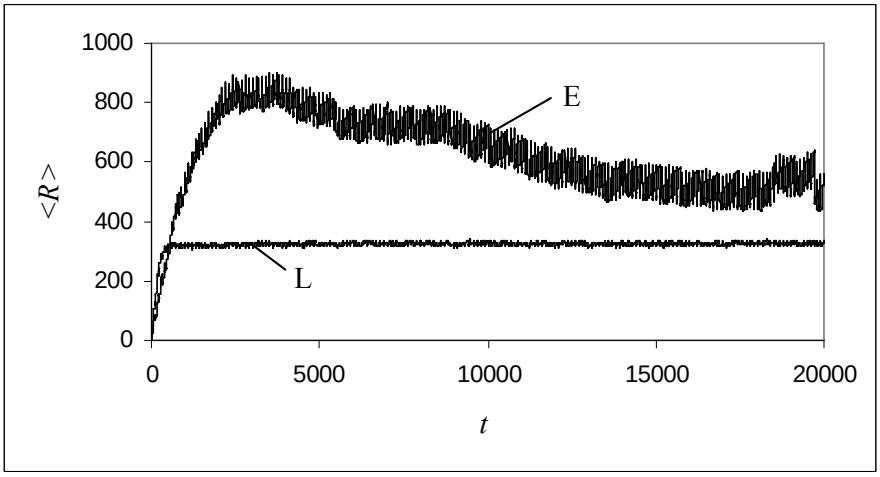

Рис. 1. Зависимости <*R*>(*t*) для случаев чистого обучения L и чистой эволюции E. Усреднено по 100 различным расчетам.

Согласно рис. 1 в случае эволюционной оптимизации при больших значениях *t* средний ресурс постепенно уменьшается, а в случае обучения <*R*> при больших *t* становится постоянным. Расчет в случае E показывает, что при больших *t* средний ресурс <*R*> и далее уменьшается (рис. 2) и только при $t > 100000$ наблюдается выход на асимптотическое значение <*R*>, меньшее, чем асимптотическое значение для случая L. Анализ кривых <*R*>(*t*) при $t < 1000$ показывает, что в случае L рост величины <*R*> происходит примерно в два раза быстрее, чем в случае E.

В случае полной модели LE зависимость $<R>(t)$ близка к таковой в случае чистого обучения L.

Таким образом, анализ зависимостей $<R>(t)$ показывает, что за исключением времени локального увеличения величины $<R>$ при $100000 > t > 1000$ чистое обучение имеет преимущество перед эволюционной оптимизацией.

Анализ случая E показывает, что локальное увеличение и максимум зависимости $<R>(t)$ обусловлен тем, что при чистой эволюции важную роль играет размножение агентов (именно при размножении происходят отбор и мутации агентов, обеспечивающие оптимизацию их поведения), которое в расчетах становится достаточно частым не с самого начала эволюции популяции, а спустя некоторое время, при $t > 100000$. При этих временах агенты выполняют действие «деление» примерно в 40% тактов времени. При размножении агенты-родители отдают половину своего ресурса потомкам, поэтому средний ресурс агентов популяции $<R>$ уменьшается. Если в компьютерной программе искусственно исключить деление ресурса $R(t)$ пополам между родителями и потомками при рождении новых агентов, то в случае E максимум в кривых $<R>(t)$ исчезает. При обучении роль размножения невелика (агенты выполняют это действие всего в 3% тактов времени) и соответствующего уменьшения ресурса не происходит, поэтому в случаях L и LE установившийся средний ресурс агентов выше, чем в случае E.

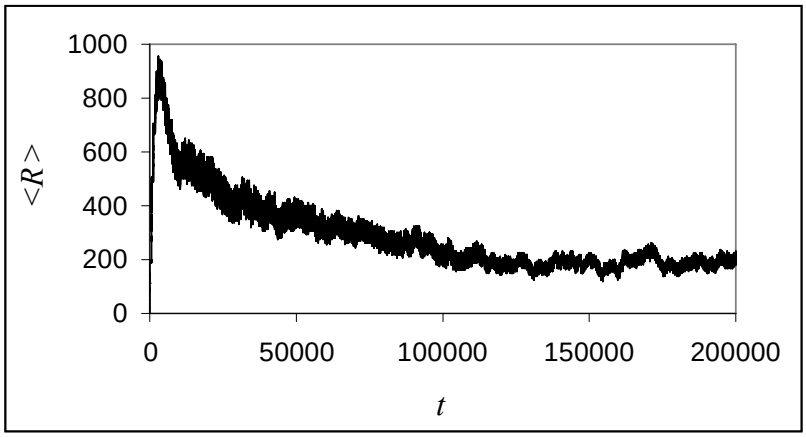

Рис. 2. Зависимость $<R>(t)$ для случая чистой эволюции E. Усреднено по 100 различным расчетам.

При эволюционной оптимизации (случай E) при $t \approx 200000$ (когда поведение агентов уже практически не менялось) действия агентов распределялись следующим образом. Действие «отдых» осуществляло небольшое число агентов (примерно 5% агентов популя-

ции), действие «питание» – 55% агентов, действие «деление» – 40% агентов. При этом в моменты, непосредственно следующие за сменой статуса опасности клеток (5-10 тактов времени), доля агентов, выполнявших действие «деление», резко уменьшалась до 5% , доля агентов, выполнявших действия «отдых» и «питание», также уменьшалась, но всего на величину 2-5% от общего числа агентов популяции. В эти же моменты средняя по популяции частота действия «перемещение» возрастала от почти нулевого значения до 40%.

При обучении (случай L) поведение агентов выходило на стационарный режим уже при $t \approx 5000$. В этом случае действия агентов распределялись следующим образом. Действие «отдых» осуществляли примерно 25% агентов популяции, действие «питание» – 70% агентов, действие «деление» – 3% агентов. При этом в моменты, непосредственно следующие за сменой статуса опасности клеток, частота действия «деление» практически не менялась, а частота выполнения агентами действий «отдых» и «питание» уменьшалась до 5% и 30%, соответственно. Частота действия «перемещение» сразу после смены статуса опасности клеток возрастала от 5% до 60%. Таким образом, динамика действий агентов в случаях E и L была сходна между собой. Отличие в основном состояло в том, что при эволюционной оптимизации существенно возрастала частота действия «деление» за счет других действий.

В случае полной модели LE (обучение + эволюция) динамика частоты действий агентов только немного отличалась от таковой в случае чистого обучения L.

4. Обсуждение. Выводы

Итак, моделирование продемонстрировало формирование достаточно естественного поведения агентов. Существенно, что при эволюционной оптимизации важную роль играет размножение. При эволюции оптимизация происходит медленней, чем при обучении. При совмещении обучения с эволюционным поиском именно обучение играет основную роль, и результаты моделирования в случае полной модели близки к таковым в случае одного обучения.

Тот факт, что в настоящей модели обучение является более эффективным, чем эволюционная оптимизация, отличает данную модель от модели [8], в которой система управления автономных агентов основана на нейросетевых адаптивных критиках [9]. Более того, в настоящей модели не наблюдался эффект Болдуина (генетическая ассимиляция приобретаемых навыков), который был продемонстрирован в [8]. Эти отличия связаны с тем, что в данной модели при рождении потомка ресурс агента-родителя уменьшался, а в

модели работы [8] рождение новых агентов не было связано с передачей ресурса от родителя к потомку. Можно ожидать, эффект Болдуина должен наблюдаться или не наблюдаться в зависимости от величины ресурса, передаваемой от родителей к потомкам.

Настоящая модель построена при ряде упрощающих допущений: простая сенсорика агентов, элементарные действия, достаточно простые правила принятия решений. Тем не менее, модель описывает систему управления с основными потребностями, подобными потребностям биологического организма, и демонстрирует вполне естественное поведение агентов.

Как развивать изложенную модель? Интересное направление развития – введение мотиваций, количественно характеризующих потребности. Несложно ввести конкуренцию между мотивациями, что эквивалентно конкуренции между потребностями. «Выигравшая» конкуренцию потребность формирует посредством обучения или эволюции свой блок системы управления агента, и этот блок функционирует при принятии решений. Использование блочной системы управления должно позволить построить более структурированную систему управления, в которой для каждой потребности формируется и оптимизируется свой блок – набор правил поведения. Моделирование динамики отдельной мотивации можно провести аналогично работе [10], в которой исследовалась биологически инспирированная модель мотивации.

Приведем схему возможной модели с мотивациями. В работе [11] предложена модель автономных агентов с несколькими потребностями. Популяция агентов эволюционировала в двумерной клеточной среде. Система управления агентов была основана на правилах вида (1). Детальное моделирование проводилось для одной потребности – потребности питания. Продемонстрировано формирование 2-3-звенных цепочек действий агентов. Дополнительно к потребности питания несложно ввести потребности размножения и безопасности. При размножении рождается потомок рассматриваемого агента в соседней с ним клетке. Способность к обеспечению безопасности целесообразно промоделировать путем введения хищников, которые представляют угрозу для жизни агентов. Для обеспечения своей безопасности автономные агенты должны удаляться от хищников. В такой модели система управления автономных агентов по сравнению с моделью данной работы существенно усложняется и набор правил поведения более естественно формировать для каждой потребности отдельно. Причем процесс формирования правил для одной потребности уже продемонстрирован в [11]. Блоки системы управления агентов (каждый блок соответствует своей потребности) в полной модели вполне могут получиться автоматически, в процессе самообучения или эволюционной оптимизации поведения агентов.

Приведем еще некоторые возможности развития изложенной модели. Набор правил вида (1) может быть заменен нейросетевым адаптивным критиком, обеспечивающим оценку качества ситуаций и прогноз будущих ситуаций [8, 9]. Отметим также, что в [11] наблюдалось обобщение ситуаций и фактическое возникновение элементарной семантики, т.е. автоматическое появление некоторых множеств ситуаций, единообразно используемых агентом, которые можно охарактеризовать так: «в клетке агента находится пища», «пища находится в клетке впереди агента», «пища находится в клетке, расположенной справа/слева от агента». При наличии предсказания и обобщения ситуаций возможны дальнейшие шаги к использованию логических выводов при оптимизации и планировании поведения. Кроме того, возможно использование методов семантического вывода [12]. Таким образом, имеются подходы к развитию моделей интеллектуальных и когнитивных свойств агентов, к моделированию процессов возникновения простейших логических выводов.

Список литературы

1. Witkowski M. An action-selection calculus // Adaptive Behavior, 2007. V. 15. No. 1. PP. 73-97.
2. Butz M.V., Sigaud O., Pezzulo G., Baldassarre G. (Eds.). Anticipatory Behavior in Adaptive Learning Systems: From Brains to Individual and Social Behavior. LNAI 4520, Berlin, Heidelberg: Springer Verlag, 2007.
3. Vernon D., Metta G., Sandini G. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents // IEEE Transactions on Evolutionary Computation, special issue on Autonomous Mental Development, 2007. V. 11. No. 2. PP. 151-180.
4. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. MIT Press, 1998.
5. Редько В.Г. Перспективы моделирования когнитивной эволюции // Третья международная конференция по когнитивной науке. Тезисы докладов в 2-х томах. Т. 2. М.: Художественно-издательский центр, 2008. С. 576-577.
6. Red'ko V.G. Evolution of cognition: Towards the theory of origin of human logic // Foundations of Science, 2000. V. 5. No. 3. PP. 323-338.
7. Holland J.H., Holyoak K.J., Nisbett R.E., Thagard P. Induction: Processes of Inference, Learning, and Discovery. Cambridge, MA: MIT Press, 1986.

8. Red'ko V.G., Mosalov O.P., Prokhorov D.V. A model of evolution and learning // Neural Networks, 2005. V. 18. No. 5-6. PP. 738-745.
9. 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.
10. 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.
11. Редько В.Г., Бесхлебнова Г.А. Модель формирования адаптивного поведения автономных агентов // «Интегрированные модели и мягкие вычисления в искусственном интеллекте». Сборник трудов V-й Международной научно-практической конференции. Т. 1. М.: Физматлит, 2009. С. 70-79.
12. Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. Новосибирск: НГУ, 2006.