Редько В.Г., Редько О.В. Бионическая модель генетической ассимиляции приобретаемых навыков // Научная сессия НИЯУ МИФИ - 2010. XII Всероссийская научно-техническая конференция "Нейроинформатика-2010": Сборник научных трудов. В 2-х частях. Ч.1. М.: НИЯУ МИФИ, 2010. С. 191-198.

В. Г. РЕДЬКО, О. В. РЕДЬКО

Научно-исследовательский институт системных исследований РАН,
Москва
vgredko@gmail.com

БИОНИЧЕСКАЯ МОДЕЛЬ ГЕНЕТИЧЕСКОЙ АССИМИЛЯЦИИ ПРИОБРЕТАЕМЫХ НАВЫКОВ *

Построена модель агентов, которые подобны биологическим организмам, приспосабливающимся к изменению температуры T в окружающей среде. Система управления агента основана на нейросетевых адаптивных критиках и обеспечивает прогнозирование изменений T и принятие решения о перемещении агента в соответствии с изменениями температуры. Продемонстрировано, что приобретаемые в процессе индивидуального обучения агента свойства могут генетически ассимилироваться в его геноме в течение нескольких поколений дарвиновской эволюции.

1. Введение

Известен эффект Болдуина [1, 2] – генетическая ассимиляция приобретаемых путем индивидуального обучения навыков в процессе дарвиновской эволюции. Этот эффект работает в два этапа. На первом этапе эволюционирующие организмы, благодаря соответствующим мутациям, приобретают свойство обучиться некоторому полезному навыку. Приспособленность таких организмов увеличивается, следовательно, они распространяются по популяции. Но обучение имеет свои недостатки, так как оно требует энергии и времени. Поэтому возможен второй этап, который называют генетической ассимиляцией: приобретенный полезный навык может быть «повторно изобретен» эволюцией, в результате чего он записывается непосредственно в геном и становится наследуемым.

* Работа выполнена при финансовой поддержке РФФИ, проект № 07-01-00180

В [3] было продемонстрировано, что генетическая ассимиляция приобретаемых навыков может наблюдаться в модели эволюционирующей популяции агентов-брокеров, система управления которых основана на нейросетевых адаптивных критиках [4]. Система управления этих агентов оптимизируется как методом обучения с подкреплением [5], так и в результате дарвиновской эволюции популяции агентов. Причем оказалось, что генетическая ассимиляция навыков, приобретаемых путем обучения, может происходить быстро: в течение всего 3-5 поколений дарвиновской эволюции. Такая быстрота подразумевает, что определенные черты ламарковской эволюции характерны для дарвиновской эволюции. Но модель [3] рассматривала довольно далекий от биологии пример агента-брокера. Поэтому интересно проверить, будет ли работать подобный эффект для более близких к живым организмам моделей. В настоящей работе строится такая модель.

2. Описание модели

Модель основана на следующей аналогии. Рассматриваются модельные «ящерицы», которые адаптируются к изменениям температуры. Суть адаптации состоит в следующем. Есть два места, которые ящерицы могут выбирать: 1) место на камешке, 2) место в норке. Естественное поведение таково. При высокой температуре ящерица греется на камешке, при низкой температуре она забирается в норку и сохраняет накопленное тепло.

Ящерицы имеют систему управления, с помощью которой они выбирают место. Как в [3] предполагаем, что системы управления агентов-ящериц основаны на нейросетевых адаптивных критиках [4]. Система управления агента оптимизируется путем обучения с подкреплением [5] и посредством дарвиновской эволюции.

Температура внешней среды T_{ext} (температура на камешке) определяется зависимостью от времени $T_{ext}(t)$, t – дискретное время, $t = 0,1,2,...$

Ситуация $\mathbf{S}(t)$, в которой находится ящерица, определяется двумя величинами $T_{ext}(t)$ и $P(t)$, $\mathbf{S}(t) = \{T_{ext}(t), P(t)\}$, где $P(t)$ – параметр, определяющий положение ящерицы. Считаем, что $P(t) = 0$, если ящерица находится в норке, и $P(t) = 1$, если ящерица находится на камешке. Действия ящерицы состоят в выборе ее положения $P(t+1)$ в следующий такт времени.

Считаем, что есть некоторая оптимальная температура тела ящерицы T_0 , и когда ящерица находится в норке, то она обогревает норку своим телом; хотя температура внешней среды тоже немного сказывается на температуре в норке. В результате температура в норке $T_{int}(t)$ равна

$$T_{int}(t) = T_0 + k_1 [T_{ext}(t) - T_0] , \quad (1)$$

где k_1 – малый положительный параметр, $k_1 > 0$, $k_1 << 1$. Согласно (1) отсчитываемые от T_0 температуры $T_{int}(t)$ и $T_{ext}(t)$ пропорциональны друг другу.

Подкрепление, которое получает ящерица в момент времени t , пропорционально разности $T(t) - T_0$, где $T(t)$ – температура в том месте, где находится ящерица:

$$r(t) = k_2 [T(t) - T_0] , \quad (2)$$

где $k_2 > 0$. Для простоты считаем, что ящерица предсказывает $T_{ext}(t)$, а $T_{int}(t)$ может оцениваться ей согласно (1).

Система управления агента. Система управления агента-ящерицы предназначена для максимизации функцию полезности $U(t)$ [5]:

$$U(t) = \sum_{j=0}^{\infty} \gamma^j r(t + j), \qquad t = 1,2,..., \quad (3)$$

где $r(t)$ – текущее подкрепление, определяемое (2), γ – дисконтный фактор ($0 < \gamma < 1$, $1-\gamma << 1$). Величина $U(t)$ – это субъективная оценка ящерицей будущей суммарной награды.

Система управления агента состоит из двух нейронных сетей (НС): Модель и Критик. НС Модель предсказывает динамику температуры $T_{ext}(t)$. НС Критик оценивает функцию состояния $U(t)$ для текущей ситуации $\mathbf{S}(t)$, предсказываемых ситуаций для двух возможных положений агента в следующий такт времени и следующей ситуации $\mathbf{S}(t+1)$.

Работа и обучение системы управления. На вход Модели подается m предыдущих значений температуры $T_{ext}(t-m+1),\ldots,T_{ext}(t)$, на выходе формируется прогноз температуры T_{ext} в следующий такт времени $T^{pr}_{ext}(t+1)$. Модель представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x^M} = \{T_{ext}(t-m+1),\ldots,T_{ext}(t)\},\ y^M_j = \mathrm{th}\,(\textstyle\sum_i w^M_{ij} x^M_i),\ T^{pr}_{ext}(t+1) = \textstyle\sum_j v^M_j y^M_j ,$$

где $\mathbf{x^M}$ – входной вектор, $\mathbf{y^M}$ – вектор выходов нейронов скрытого слоя, w^M_{ij} и v^M_j – веса синапсов данной НС.

Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$, а именно, оценки функции полезности $U(t)$ для агента, находящегося в ситуации $\mathbf{S}$. Критик представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x}^{\mathbf{C}} = \mathbf{S}(t) = \{T_{ext}(t), P(t)\}, \quad y^C_j = \text{th}\,(\textstyle\sum_i w^C_{ij}\, x^C_i),\ V(t) = V(\mathbf{S}(t)) = \textstyle\sum_j v^C_j\, y^C_j\,,$$

где $\mathbf{x}^{\mathbf{C}}$ – входной вектор, $\mathbf{y}^{\mathbf{C}}$ – вектор выходов нейронов скрытого слоя, w^C_{ij} и v^C_j – веса синапсов НС.

При работе системы управления агента каждый момент времени t выполняются следующие операции:

1) Модель предсказывает внешнюю температуру в следующий такт времени $T^{pr}_{ext}(t+1)$.

2) Критик оценивает величину V для текущей ситуации $V(t) = V(\mathbf{S}(t))$ и для предсказываемых ситуаций для обоих возможных действий $V^{pr}_P(t+1) = V(\mathbf{S}^{\mathbf{pr}}_P(t+1))$, где $\mathbf{S}^{\mathbf{pr}}_P(t+1) = \{T^{pr}_{ext}(t+1), P(t+1)\}$, $P(t+1) = 0$ либо $P(t+1) = 1$.

3) Применяется ε-жадное правило [5]: с вероятностью $1 - \varepsilon$ выбирается действие, соответствующее максимальному значению $V^{pr}_P(t+1)$, в противном случае выбирается альтернативное действие ($0 < \varepsilon << 1$). Выбор действия есть выбор величины $P(t+1)$: поместиться в норку $P(t+1) = 0$, либо на камешек $P(t+1) = 1$.

4) Выбранное действие $P(t+1)$ выполняется. Происходит переход к моменту времени $t+1$. Подсчитывается подкрепление $r(t+1)$ согласно (2). Наблюдаемое значение $T_{ext}(t+1)$ сравнивается с предсказанием $T^{pr}_{ext}(t+1)$. Веса НС Модели подстраиваются так, чтобы минимизировать ошибку предсказания методом обратного распространения ошибки [6]. Скорость обучения Модели равна $\alpha_M > 0$.

5) Критик подсчитывает $V(t+1) = V(\mathbf{S}(t+1))$; $\mathbf{S}(t+1) = \{T_{ext}(t+1), P(t+1)\}$. Рассчитывается ошибка временной разности [5]:

$$\delta(t) = r(t+1) + \gamma\, V(t+1) - V(t)\,. \qquad (4)$$

6) Веса НС Критика подстраиваются так, чтобы минимизировать величину $\delta(t)$, это обучение осуществляется градиентным методом, аналогично методу обратного распространения ошибки. Скорость обучения Критика равна $\alpha_C > 0$.

Схема эволюции. Рассматривается эволюционирующая популяция, состоящая из n агентов. Каждый агент имеет ресурс $R(t)$, который изменя-

ется в соответствии с подкреплениями: $R(t+1) = R(t) + r(t+1)$, где $r(t+1)$ определяется выражением (2).

Эволюция происходит в течение ряда поколений, $n_g = 1,2,\ldots$ Продолжительность каждого поколения n_g равна T_g тактов времени (T_g – длительность жизни агента). В начале каждого поколения ресурс каждого агента равен нулю, т.е., $R(T_g(n_g-1)+1) = 0$.

Каждый агент имеет два набора весов синапсов нейронных сетей: **G** и **W**. Набор **G** представляет собой начальные веса синапсов НС, получаемые агентом в момент его рождения от агента-родителя. Этот набор **G** есть геном агента, который не меняется в течение его жизни. Набор **W** – текущие веса синапсов НС, которые подстраиваются в течение жизни агента путем обучения, описанного выше. В момент рождения агента **W = G**. Потомки агента наследуют геном **G** (с небольшими мутациями). Так как наследуется именно геном **G**, а не изменяемые в течение жизни веса **W**, то эволюция носит дарвиновский характер.

Процесс размножения происходит следующим образом. В конце каждого поколения определяется агент, имеющий максимальный ресурс $R_{max}(n_g)$ (лучший агент поколения n_g). Этот лучший агент порождает n потомков, которые составляют новое (n_g+1)-е поколение. Геномы потомков **G** отличаются от генома родителя небольшими мутациями. Более конкретно, предполагается, что в начале нового (n_g+1)-го поколения для каждого агента его геном формируется следующим образом $G_i(n_g+1) = G_{best,\,i}(n_g) + \mathrm{rand}_i$, где $G_{best,\,i}(n_g)$ – компоненты генома лучшего агента предыдущего n_g-го поколения, rand_i – нормально распределенные случайные величины с нулевым средним и стандартным отклонением P_{mut} (интенсивность мутаций), i – индекс веса синапса.

3. Результаты моделирования

Изложенная модель была реализована в виде компьютерной программы. Основные параметры модели имели следующее значение: дисконтный фактор $\gamma = 0.9$; количество входов НС Модели $m = 10$; количество нейронов в скрытых слоях НС Модели и Критика $N_{hM} = N_{hC} = 10$; скорость обучения Модели и Критика $\alpha_M = \alpha_C = 0.01$; параметр ε-жадного правила $\varepsilon = 0.05$; интенсивность мутаций $P_{mut} = 0.1$; продолжительность поколения $T_g = 1000$, численность популяции $n = 10$. Зависимость внешней температуры от времени задавалась в виде синусоиды с периодом 20 тактов времени:

$$T_{ext}(t) = 0.5\sin(2\pi/20) + T_0\,, \quad T_0 = 1.5.$$

Были проанализированы следующие варианты:

- Случай L (чистое обучение); в этом случае рассматривался отдельный агент, который обучался путем минимизации ошибки временной разности (4);
- Случай E (чистая эволюция), т.е. эволюционирующая популяция агентов без обучения;
- Случай LE (обучение + эволюция), т.е. полная модель, изложенная выше.

Было проведено сравнение ресурса, приобретаемого агентами за 1000 временных тактов для этих трех способов адаптации. Для случаев E и LE бралось T_g = 1000 (T_g – продолжительность поколения) и регистрировалось максимальное значение ресурса в популяции $R_{max}(n_g)$ в конце каждого поколения. В случае L (чистое обучение) рассматривался только один агент, ресурс которого для удобства сравнения со случаями E и LE обнулялся каждые T = 1000 тактов времени: $R(T(n_g\text{-}1)+1) = 0$. В этом случае индекс n_g увеличивался на единицу после каждых T временных тактов, и полагалось $R_{max}(n_g) = R(T\, n_g)$.

Графики $R_{max}(n_g)$ представлены на рис. 1, который показывает, что обучение, объединенное с эволюцией (случай LE), обеспечивает более эффективный рост R_{max} , чем обучение или эволюция отдельно (случаи L и E).

Для случая LE часто наблюдалось явное влияние обучения на эволюционный процесс. В первых поколениях эволюционного процесса существенный рост ресурса $R(t)$ агентов наблюдался не с самого начала поколения, а спустя 200-400 тактов, т.е. агенты явно обучались в течение своей жизни находить приемлемую стратегию поведения, и только после смены нескольких поколений рост ресурса начинался с самого начала поколения. Это можно интерпретировать как проявление известного эффекта Болдуина: исходно приобретаемый навык в течение ряда поколений дарвиновской эволюции становился наследуемым [1, 2]. Этот эффект наблюдался для ряда расчетов, один из которых представлен на рис. 2. Для этого примера было проанализировано, как изменяется значение ресурса наилучшего агента в популяции $R(t)$ в течение первых пяти поколений. Рис. 2 показывает, что в двух первых поколениях значительный рост ресурса лучшего в популяции агента начинается только после задержки 200-400 тактов времени; т.е., очевидно, что агент оптимизирует свою стратегию поведения при помощи обучения. От поколения к поколению агенты находят хорошую стратегию поведения все раньше и раньше. К пятому по-

колению лучший агент «знает» хорошую стратегию поведения с самого рождения, и обучение не приводит к существенному улучшению стратегии.

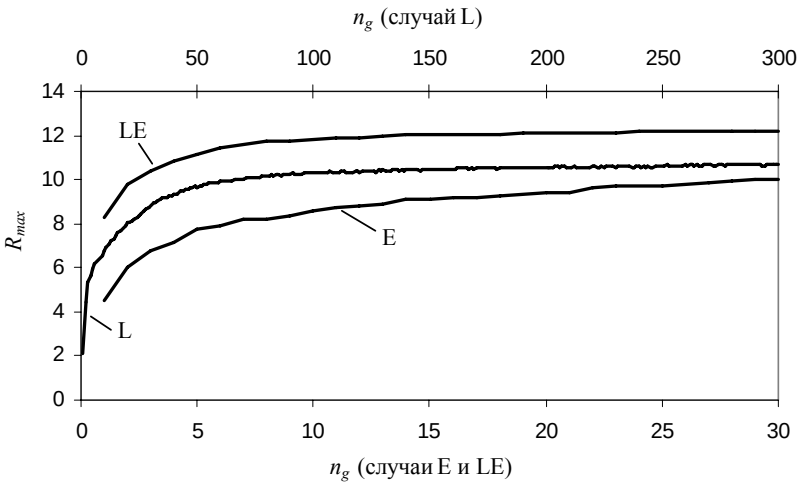

Рис. 1. Зависимости $R_{max}(n_g)$. Кривая LE соответствует случаю обучения, объединенного с эволюцией, кривая E – случаю чистой эволюции, кривая L – случаю чистого обучения. Временная шкала для случаев E и LE представлена снизу, для случая L – сверху. Кривые усреднены по 1000 расчетам.

На рис. 2 представлен пример расчета, в котором задержка в росте $R(t)$ четко видна. Был проведен ряд расчетов, различающихся датчиками случайных чисел. Во многих других расчетах эта задержка была примерно такая же, как на рис. 2. В некоторых расчетах задержка была менее длительной (примерно 100 тактов времени). Тем не менее, во всех проведенных расчетах задержка в росте $R(t)$, связанная с процессами обучения, наблюдалась.

Итак, моделирование показывает, что стратегия, изначально приобретаемая посредством обучения, становится наследуемой (эффект Болдуина), хотя эволюция имеет дарвиновский характер. Подчеркнем, что, как и в работе [3], в изложенной модели генетическая ассимиляция навыков,

приобретаемых путем обучения, может происходить быстро: в течение всего 3-5 поколений дарвиновской эволюции.

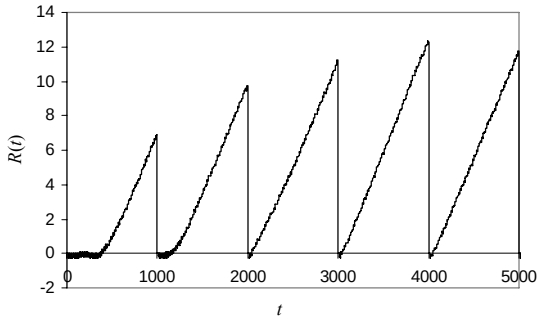

Рис. 2. Зависимость ресурса наилучшего в популяции агента $R(t)$ от времени t. Случай LE, $T_g = 1000$.

4. Выводы

Исследование биологически инспирированной модели продемонстрировало следующее.

1) Взаимодействие между обучением и эволюционной оптимизацией носит симбиотический характер: обучение, объединенное с эволюцией, обеспечивает более эффективную стратегию поведения агентов, чем обучение или эволюция отдельно.

2) В эволюционирующей популяции самообучающихся агентов наблюдается эффект Болдуина: навыки, изначально приобретаемые посредством обучения, генетически ассимилируются в процессе дарвиновской эволюции. Причем эта ассимиляция может происходить в течение всего 3-5 поколений. Тем самым продемонстрировано, что определенные черты ламарковской эволюции могут быть характерны для дарвиновской эволюции.

Список литературы

1. Baldwin J.M. A new factor in evolution // American Naturalist, 1896. Vol. 30. PP. 441-451.

2. Turney P., Whitley D., Anderson R. (Eds.). Evolution, Learning, and Instinct: 100 Years of the Baldwin Effect // Special Issue of Evolutionary Computation on the Baldwin Effect. Vol.4. No. 3. 1996.
3. Red'ko V.G., Mosalov O.P., Prokhorov D.V. A model of evolution and learning // Neural Networks, 2005. Vol. 18. No 5-6. PP. 738-745.
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. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. MIT Press, 1998.
6. Rumelhart D.E., Hinton G.E., Williams R.G. Learning representation by back-propagating error // Nature, 1986. Vol. 323. No. 6088. PP. 533-536