МОДЕЛИРОВАНИЕ ЭВОЛЮЦИИ АВТОНОМНЫХ АДАПТИВНЫХ АГЕНТОВ*

О.П. Мосалов, В.Г. Редько, Д.В. Прохоров
Московский физико-технический институт
Институт оптико-нейронных технологий РАН, Москва
Toyota Technical Center in Ann Arbor, MI, USA

Исследуется модель эволюции популяций самообучающихся агентов и анализируется взаимодействие между обучением и эволюцией. Система управления агента основана на нейросетевых адаптивных критиках, обучаемых методом обучения с подкреплением. Модель исследуется на примере простого агента-брокера, который предсказывает изменения биржевого курса и использует эти предсказания для выбора действий. Проведено сравнение трех вариантов модели, в которых включены 1) либо обучение и эволюция одновременно, 2) либо отдельно эволюция, 3) либо отдельно обучение. Показано, что в данной модели может наблюдаться эффект Болдуина, т.е. первоначально приобретаемые навыки агентов в процессе эволюции становятся наследуемыми. Проведено сравнение поведения модельных агентов с поисковым поведением простейших животных.

SIMULATION OF EVOLUTION OF AUTONOMOUS ADAPTIVE AGENTS

Oleg P. Mosalov, Vladimir G. Red'ko, Danil V. Prokhorov
Moscow Institute of Physics and Technologies
Institute of Optical Neural Technologies, Russian Academy of Science
Toyota Technical Center in Ann Arbor, MI, USA

A model of evolving populations of self-learning agents is studied and the interaction between learning and evolution is analyzed. Each agent is equipped with a neural network adaptive critic design for behavioral adaptation. The model is investigated for the case of a simple agent-broker that predicts stock price changes and uses its predictions for selecting actions. Three cases are analyzed in which either evolution or learning, or both, are active in this model. It is shown that the Baldwin effect can be observed in this model, viz., originally acquired adaptive policy of agents becomes inherited over the course of the evolution. Also the behavioral tactics of our agents is compared to the searching behavior of simple animals.

1. ВВЕДЕНИЕ

Одно из актуальных направлений, которое активно развивается в последние годы в вычислительном интеллекте – исследование и применение автономных адаптивных агентов [1]. Такие агенты, подобно живым организмам, могут обладать собственными целями, собственными знаниями, формировать собственную политику поведения, выполнять те или иные действия, а также взаимодействовать с другими агентами.

В настоящей работе исследуется модель эволюции популяции автономных адаптивных агентов. Рассматриваемые агенты обучаются методом обучения с подкреплением [2], в котором обучение происходит без учителя, путем непосредственного взаимодействия агента с внешней средой. В зависимости от выполняемых действий агент получает подкрепления r: поощрения ($r > 0$) или наказания ($r < 0$). При этом агент стремится максимизировать суммарную награду U (сумму значений r), которую он может получить в будущем.

Система управления агента состоит из двух нейросетевых блоков, которые предназначены 1) для прогноза будущих ситуаций $\mathbf{S}$ во внешней среде и 2) для оценки качества $V(\mathbf{S})$ тех или иных ситуаций. Обучение происходит в результате модификации весов синапсов нейронных сетей агента. Часть действий агент выбирает случайно, тем самым он имеет возможность протестировать новые ситуации.

Помимо обучения происходит эволюционная адаптация агентов. А именно, предполагается, что имеется популяция агентов, начальные веса синапсов нейронных сетей, получаемые агентами при рождении, передаются от родителей к потомкам, испытывая при этом

* Работа выполнена при частичной поддержке программы Президиума РАН "Интеллектуальные компьютерные системы" и РФФИ (проект № 07-01-00180).

малые мутации. Причем, чем больше суммарная награда агента U, полученная им в течение жизни поколения, тем больше шансов имеет агент дать потомков в следующее поколение.

Таким образом, нейросетевая система управления агентов адаптируется как за счет обучения, происходящего в течение жизни каждого поколения, так и в процессе эволюции, происходящей в течение ряда поколений. При этом критерий, по которому проводится та и другая адаптация, один и тот же – суммарная награда агента U.

Исследование рассматриваемой модели проводится на примере агентов-брокеров, которые продают и покупают акции, стремясь максимизировать свой суммарный капитал.

2. ОПИСАНИЕ МОДЕЛИ

2.1. Обучение с подкреплением

Схема обучения исследуемых агентов основана на теории обучения с подкреплением (Reinforcement Learning), которая была развита в работах Р. Саттона и Э. Барто (Массачусетский университет). Общая схема обучения с подкреплением [2] показана на рис. 1. Рассматривается агент, взаимодействующий с внешней средой. Время предполагается дискретным: $t = 1,2,\ldots$ В текущей ситуации агент $\mathbf{S}(t)$ выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию $\mathbf{S}(t+1)$.

Цель агента – максимизировать суммарную награду U, которую можно получить в будущем в течение длительного периода времени. Оценка величины U выполняется с учетом коэффициента забывания:

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

где $U(t)$ – оценка суммарной награды, ожидаемой после момента времени t; γ – коэффициент забывания $(0 < \gamma < 1)$, с помощью которого учитывается, что чем дальше агент «заглядывает» в будущее, тем меньше у него уверенность в оценке награды.

2.2. Схема агента-брокера

Следуя [3], рассматриваем агента-брокера, который имеет ресурсы двух типов: деньги и акции; сумма этих ресурсов составляет капитал агента $C(t)$; доля акций в капитале равна $u(t)$. Внешняя среда определяется временным рядом $X(t)$, $t = 0,1,2,\ldots$, $X(t)$ – курс акций на бирже в момент времени t. Агент стремится увеличить свой капитал $C(t)$, изменяя значение $u(t)$. Динамика капитала определяется выражением [3]:

$$C(t+1) = C(t)\ \{1 + u(t+1)\ \Delta X(t+1) / X(t)\}\ [1 - J\,|u(t+1) - u(t)|], \tag{2}$$

где $\Delta X(t+1) = X(t+1) - X(t)$, J – параметр, учитывающий расходы агента на покупку/продажу акций. Следуя [4], используем логарифмическую шкалу для ресурса агента, $R(t) = \ln C(t)$. Текущее подкрепление агента $r(t) = R(t+1) - R(t)$ равно:

$$r(t) = \ln\{1 + u(t+1)\ \Delta X(t+1) / X(t)\} + \ln[1 - J\,|u(t+1) - u(t)|]. \tag{3}$$

Полагаем, что переменная u может принимать только два значения $u = 0$ (весь капитал в деньгах) или $u = 1$ (весь капитал в акциях). Соответственно, выбор действия производится следующим образом: в следующий такт времени весь капитал переводится в акции: $u(t+1) = 1$, либо в деньги: $u(t+1) = 0$.

2.3. Алгоритм обучения

Система управления агента представляет собой адаптивный критик (адаптивные критики – конструкции, обучаемые методом обучения с подкреплением, подробней см. [5,6]), состоящий из двух нейронных сетей (НС): Модель и Критик (рис. 2). Цель адаптивного критика – максимизировать величину $U(t)$, определяемую выражением (1).

Делая разумное предположение $\Delta X(t) << X(t)$, полагаем, что ситуация $\mathbf{S}(t)$, характеризующая состояние агента, зависит только от двух величин, $\Delta X(t)$ и $u(t)$: $\mathbf{S}(t) = \{\Delta X(t), u(t)\}$.

Модель предназначена для прогнозирования изменения курса временного ряда. На вход Модели подается m предыдущих значений изменения курса $\Delta X(t-m+1), \ldots, \Delta X(t)$, на выходе формируется прогноз изменения курса в следующий такт времени $\Delta X^{pr}(t+1)$. Модель представляет собой двухслойную НС, работа которой описывается формулами:

$\mathbf{x^M} = \{\Delta X(t-m+1), \ldots, \Delta X(t)\}$,

$$y_j^M = \mathrm{th}\left(\sum_i w_{ij}^M x_i^M\right),$$

$$\Delta X^{pr}(t+1) = \sum_j v_j^M y_j^M ,$$

где $\mathbf{x^M}$ – входной вектор, $\mathbf{y^M}$ – вектор выходов нейронов скрытого слоя, w_{ij}^M и v_j^M – веса синапсов НС.

Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$, а именно, оценки функции полезности $U(t)$ (см. формулу (1)) для агента, находящегося в рассматриваемой ситуации $\mathbf{S}$. Критик представляет собой двухслойную НС, работа которой описывается формулами:

$\mathbf{x^C} = \mathbf{S}(t) = \{\Delta X(t), u(t)\}$,

$$y_j^C = \mathrm{th}\left(\sum_i w_{ij}^C x_i^C\right),$$

$$V(t) = V(\mathbf{S}(t)) = \sum_j v_j^C y_j^C ,$$

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

Каждый момент времени t выполняются следующие операции:

1) Модель предсказывает следующее изменение временного ряда $\Delta X^{pr}(t+1)$.

2) Критик оценивает величину V для текущей ситуации $V(t) = V(\mathbf{S}(t))$ и для предсказываемых ситуаций для обоих возможных действий $V_u^{pr}(t+1) = V(\mathbf{S}_u^{pr}(t+1))$, где $\mathbf{S}_u^{pr}(t+1) = \{\Delta X^{pr}(t+1), u\}$, $u = 0$ либо $u = 1$.

3) Применяется ε-жадное правило [2]: действие, соответствующее максимальному значению $V_u^{pr}(t+1)$, выбирается с вероятностью 1 - ε, альтернативное действие выбирается с вероятностью ε ($0 < \varepsilon << 1$). Выбор действия есть выбор величины $u(t+1)$: $u(t+1) = 1$, либо $u(t+1) = 0$.

4) Выбранное действие $u(t+1)$ выполняется. Происходит переход к моменту времени $t+1$. Подсчитывается подкрепление $r(t)$ согласно (3). Наблюдаемое значение $\Delta X(t+1)$ сравнивается с предсказанием $\Delta X^{pr}(t+1)$. Веса НС Модели подстраиваются так, чтобы минимизировать ошибку предсказания методом обратного распространения ошибки [7]:

$$\Delta v_i^M(t+1) = -\alpha_M \left(\Delta X^{pr}(t+1) - \Delta X(t+1)\right) y_j^M ,$$

$$\Delta w_{ij}^{M}(t+1) = -\alpha_M \left(\Delta X^{pr}(t+1) - \Delta X(t+1)\right) v_j^M \left(1-(y_j^M)^2\right) x_i^M \ .$$

где α_M – скорость обучения Модели ($\alpha_M > 0$).

5) Критик подсчитывает $V(t+1) = V(\mathbf{S}(t+1))$; $\mathbf{S}(t+1) = \{\Delta X(t+1), u(t+1)\}$. Рассчитывается ошибка временной разности [2]:

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

Величина $\delta(t)$ характеризует ошибку в оценке $V(t) = V(\mathbf{S}(t))$ – суммарной награды, которую можно получить, исходя из состояния $\mathbf{S}(t)$. Ошибка $\delta(t)$ рассчитывается с учетом текущей награды $r(t)$ и оценки суммарной награды $V(\mathbf{S}(t+1))$, которую можно получить, исходя из следующего состояния $\mathbf{S}(t+1)$.

6) Веса НС Критика подстраиваются так, чтобы минимизировать величину $\delta(t)$, это обучение осуществляется градиентным методом, аналогично методу обратного распространения ошибки:

$$\Delta v_i^C(t+1) = \alpha_C \delta(t) y_j^C \ ,$$

$$\Delta w_{ij}^C(t+1) = \alpha_C \delta(t) v_j^C \left(1-(y_j^C)^2\right) x_i^C \ .$$

α_C – скорость обучения Критика ($\alpha_C > 0$).

Смысл обучения Модели – уточнение прогнозов будущих ситуаций. Смысл обучения Критика состоит в том, чтобы итеративно уточнять оценку качества ситуаций $V(\mathbf{S}(t))$ в соответствии с поступающими подкреплениями.

2.4. Схема эволюции

Эволюционирующая популяция состоит из n агентов. Каждый агент имеет ресурс $R(t)$, который изменяется в соответствии с подкреплениями агента: $R(t+1) = R(t) + r(t)$.

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

Начальные веса синапсов обоих НС (Модели и Критика) формируют геном агента $\mathbf{G}=\{\mathbf{W}_{M0},\mathbf{W}_{C0}\}$. Геном $\mathbf{G}$ задается в момент рождения агента и не меняется в течение его жизни. В противоположность этому текущие веса синапсов НС $\mathbf{W}_M$ и $\mathbf{W}_C$ подстраиваются в течение жизни агента путем обучения, описанного в п. 2.3.

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)-ое поколение. Геномы потомков $\mathbf{G}$ отличаются от генома родителя небольшими мутациями.

Более конкретно, в начале каждого нового (n_g+1)-го поколения для каждого агента полагается $G_i(n_g+1) = G_{best,\,i}(n_g) + \mathrm{rand}_i$, $\mathbf{W}_0(n_g+1) = \mathbf{G}(n_g+1)$, где $\mathbf{G}_{\mathbf{best}}(n_g)$ – геном лучшего агента предыдущего n_g-го поколения, rand_i – нормально распределенная случайная величина с нулевым средним и стандартным отклонением P_{mut} (интенсивность мутаций), которая добавляется к каждому весу.

Таким образом, геном $\mathbf{G}$ (начальные веса синапсов, получаемые при рождении) изменяется только посредством эволюции, в то время как текущие веса синапсов $\mathbf{W}$ дополнительно к этому подстраиваются посредством обучения. При этом в момент рождения агента $\mathbf{W} = \mathbf{W}_0 = \mathbf{G}$.

3. РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

3.1. Общие особенности адаптивного поиска

Изложенная модель исследовалась путем компьютерного моделирования. В вычислительных экспериментах использовались два варианта модельного временного ряда $X(t)$: 1) синусоида:

$$X(t) = 0{,}5(1 + \sin(2\pi t/20)) + 1, \qquad (5)$$

2) стохастический временной ряд, использованный в [3]:

$$X(t) = \exp(p(t)/1200), \qquad p(t) = p(t\text{-}1) + \beta(t\text{-}1) + k\,\lambda(t), \qquad \beta(t) = \alpha\beta(t\text{-}1) + \mu(t), \qquad (6)$$

где $\lambda(t)$ и $\mu(t)$ – два нормальных процесса с нулевым средним и единичной дисперсией, $\alpha = 0{,}9$, $k = 0{,}3$.

Некоторые параметры модели имели одно и то же значение для всех экспериментов: фактор забывания $\gamma = 0{,}9$; 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 $\alpha_M = 0{,}01$, $\alpha_C = 0{,}01$; параметр ε-жадного правила $\varepsilon = 0{,}05$; интенсивность мутаций $P_{mut} = 0{,}1$; расходы агента на покупку/продажу акций $J = 0$. Остальные параметры (продолжительность поколения T и численность популяции n) принимали разные значения в разных экспериментах, см. ниже.

Анализировались следующие варианты рассматриваемой модели:
- Случай L (чистое обучение); в этом случае рассматривался отдельный агент, который обучался методом, изложенным в п. 2.3;
- Случай E (чистая эволюция), т.е. рассматривалась эволюционирующая популяция без обучения;
- Случай LE (обучение + эволюция), т.е. анализировалась полная модель, изложенная выше.

Было проведено сравнение ресурса, приобретаемого агентами за 200 временных тактов для этих трех способов адаптации. Для случаев E и LE бралось $T = 200$ (T – 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 $T = 200$ 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 $R_{max}(n_g)$ для синусоиды (5) показаны на рис. 3. Чтобы исключить уменьшение значения $R_{max}(n_g)$ из-за случайного выбора действий при применении ε-жадного правила для случаев LE и L, полагалось $\varepsilon = 0$ после $n_g = 100$ для случая LE и после $n_g = 2000$ для случая L (на рис. 3 видно резкое увеличение $R_{max}(n_g)$ после $n_g = 100$ и $n_g = 2000$ для соответствующих случаев). Результаты усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

Рис. 3 показывает, что обучение, объединенное с эволюцией (случай LE), и чистая эволюция (случай E) дают одно и то же значение конечного ресурса $R_{max}(500) = 6{,}5$. Однако обучение и эволюция вместе обеспечивают нахождение больших значений R_{max} быстрее, чем эволюция отдельно – существует симбиотическое взаимодействие между обучением и эволюцией.

Из (2) следует, что существует оптимальная стратегия поведения агента (в настоящей работе пренебрегаем затратами на покупку/продажу акций, т.е. всюду полагаем $J = 0$): вкладывать весь капитал в акции ($u(t+1) = 1$) при росте курса ($\Delta X(t+1) > 0$), вкладывать весь капитал в деньги ($u(t+1) = 0$) при падении курса ($\Delta X(t+1) < 0$).

Анализ экспериментов, представленных на рис. 3, показывает, что в случаях LE (обучение + эволюция), и E (чистая эволюция) такая оптимальная стратегия находится. Это соответствует асимптотическому значению ресурса $R_{max}(500) = 6{,}5$.

В случае L (чистое обучение) асимптотическое значение ресурса ($R_{max}(2500) = 5{,}4$) существенно меньше. Анализ экспериментов для этого случая показывает, что одно обучение обеспечивает нахождение только следующей «субоптимальной» стратегии поведения: агент держит капитал в акциях при росте и при слабом падении курса и переводит капитал в деньги при сильном падении курса (рис.4). Та же тенденция к явному предпочтению вкладывать

капитал в акции при чистом обучении наблюдается и для экспериментов на стохастическом ряде (6).

Итак, результаты, представленные на рис. 3, демонстрируют, что хотя обучение в настоящей модели и несовершенно, оно способствует более быстрому нахождению оптимальной стратегии поведения по сравнению со случаем чистой эволюции (см. графики LE и E на рис. 3).

3.2. Взаимодействие между обучением и эволюцией. Эффект Болдуина

Как показано на рис. 3 для синусоидального временного ряда в случае E оптимальная стратегия находится во всех экспериментах. В случае стохастического временного ряда оптимальная стратегия также может быть найдена при помощи только эволюции, но лишь в некоторых экспериментах. Например, при $N_g = 300$ и $T = 200$ в случае E оптимальная стратегия была найдена в 8 из 10 экспериментов. Типичный пример оптимальной стратегии поведения представлен на рис. 5.

Рис. 3 также демонстрирует, что при $T = 200$ наличие обучения ускоряет поиск оптимальной стратегии по сравнению со случаем одной эволюции. Если длительность поколения T была достаточно большой (1000 и более тактов времени), то для случая LE часто наблюдалось и более четкое влияние обучения на эволюционный процесс. В первых поколениях эволюционного процесса существенный рост ресурса агентов наблюдался не с самого начала поколения, а спустя 200-300 тактов, т.е. агенты явно обучались в течение своей жизни находить более или менее приемлемую стратегию поведения, и только после смены ряда поколений рост ресурса начинался с самого начала поколения. Это можно интерпретировать как проявление известного эффекта Болдуина: исходно приобретаемый навык в течение ряда поколений становился наследуемым [8,9]. Этот эффект наблюдался в ряде экспериментов, один из которых представлен на рис. 6.

3.3. Особенности предсказания Модели

Система управления агента включает в себя нейронную сеть Модели, предназначенную для предсказания изменения значения $\Delta X\,(t+1)$ временного ряда в следующий такт времени $t+1$. Была проанализирована работа Модели и обнаружено, что нейронная сеть Модели может давать неверные предсказания, однако агент, тем не менее, может использовать эти предсказания для принятия верных решений. Например, рис. 7 показывает предсказываемые изменения $\Delta X^{pr}(t+1)$ и реальные изменения $\Delta X(t+1)$ стохастического временного ряда в случае чистой эволюции (случай E). Предсказания нейронной сети Модели достаточно хорошо совпадают по форме с кривой $\Delta X(t+1)$. Однако, предсказанные значения $\Delta X^{pr}(t+1)$ отличаются примерно в 25 раз от значений $\Delta X(t+1)$.

На рис. 8 приведен другой пример особенностей предсказания нейронной сети Модели в случае LE (обучение, объединенное с эволюцией). Этот пример показывает, что предсказания нейронной сети Модели $\Delta X^{pr}(t+1)$ могут отличаться от реальных данных $\Delta X(t+1)$ не только масштабом, но и знаком.

Хотя предсказания Модели могут быть неверными количественно, естественно полагать, что правильность предсказаний $\Delta X^{pr}(t+1)$ после линейных преобразований (например, изменения знака) приводит к тому, что Модель является полезной для адаптивного поведения. Эти предсказания эффективно используются системой управления агентов для нахождения оптимальной поведения: стратегия поведения агентов для обоих приведенных примеров работы Модели была подобна стратегии, представленной на рис. 5.

Естественно считать, что наблюдаемое увеличение значений ΔX^{pr} нейронной сетью Модели полезно для работы нейронной сети Критика, так как реальные значения $\Delta X(t+1)$ слишком малы (порядка 0,001). Таким образом, нейронная сеть Модели может не только предсказывать значения $\Delta X^{pr}(t+1)$, но также осуществлять полезные преобразования этих значений.

3.4. Сравнение с поведением простейших животных

Исследуемые агенты имеют две поведенческие тактики (продавать или покупать акции) и выбирают действия, переключаясь между этими тактиками. Можно сопоставить особенности

этого поведения с переключением между двумя тактиками при поисковом поведении простейших животных. Например, некоторые виды личинок ручейников используют аналогичные тактики [10,11]. Личинки живут на речном дне и носят на себе «домик» – трубку из песка и других частиц, которые они собирают на дне водоемов. Личинки строят свои домики из твердых частичек разной величины. Большие частицы распределены случайно, но обычно встречаются группами. Используя большие частицы, личинка может построить домик гораздо быстрее и эффективнее, чем используя маленькие, и, естественно, предпочитает использовать большие частицы. Личинка использует две тактики: 1) тестирование частиц вокруг себя и использование выбранных частиц, 2) поиск нового места для сбора частиц. Исследование поведения личинок обнаруживает инерцию в переключении с первой тактики на вторую [10,11]. Если личинка находит большую частицу, она продолжает тестировать частицы, пока не найдет несколько маленьких, и только после нескольких неудачных попыток найти новую большую частицу, переходит ко второй тактике. Во время поиска нового места личинка время от времени тестирует частицы, которые попадаются на ее пути. Она может переключиться со второй тактики на первую, если найдет большую частицу; при этом переключении также может проявляться инерция. Таким образом, переключение между тактиками имеет характер случайного поиска с явным эффектом инерции. Процесс инерционного переключения позволяет животному использовать только общие крупномасштабные свойства окружающего мира и игнорировать мелкие случайные детали.

В проведенных компьютерных экспериментах поведение агента-брокера, подобное поведению животных с инерционным переключением между двумя тактиками, наблюдалось, когда система управления агента оптимизировалась с помощью чистой эволюции при достаточно большой численности популяции. То есть фактически происходила оптимизация методом случайного поиска в достаточно большой области возможных решений. Рис. 9 показывает фрагмент стратегии поведения агента, найденной на ранней стадии эволюции в большой популяции, $n = 100$. Эта стратегия агента подобна описанному выше поведению животных с инерционным переключением между двумя тактиками. Стратегия переключения между $u = 0$ и $u = 1$ представляет собой реакцию только на общие изменения в окружающей среде (агент игнорирует мелкие флуктуации в изменении курса акций). Кроме того, переключение явно обладает свойством инерционности.

4. ЗАКЛЮЧЕНИЕ

Итак, построена модель эволюции автономных агентов, система управления которых основана на нейросетевых адаптивных критиках. Проведено исследование взаимодействия между обучением и эволюцией в популяциях самообучающихся агентов-брокеров. Проанализировано три варианта модели, в которых веса синапсов нейронных сетей настраивались 1) обучением, 2) эволюцией или 3) комбинацией обучения и эволюции.

Показано, что оптимальная стратегия обеспечивается в случаях эволюции или комбинации обучения и эволюции. Одно обучение не обеспечивает нахождения оптимальной стратегии, тем не менее, оно способствует более быстрому нахождению оптимальной стратегии поведения для случая комбинации обучения и эволюции по сравнению со случаем чистой эволюции.

При достаточно большой длительности жизни агентов наблюдалась эволюционная ассимиляция приобретенных навыков, что можно интерпретировать как проявление известного эффекта Болдуина.

Проведено сравнение поведение агента-брокера с поисковым поведением простых животных. Исследования поискового поведения животных показывают, что часто в процессе поиска проявляются инерционные эффекты при переключении между разными поведенческими тактиками. Такая инерция помогает животному адаптивно реагировать только на общие изменения в окружающей среде. В исследованной модели подобное инерционное поведение формируется в процессе эволюционного поиска на ранних стадиях эволюции при условии достаточного большого размера популяции.

Литература:

1. *Тарасов В.Б.* От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. М.: Эдиториал УРСС, 2002. 352 с.

2. *Sutton R., Barto A.* Reinforcement Learning: An Introduction. – Cambridge: MIT Press, 1998. See also: http://www.cs.ualberta.ca/~sutton/book/the-book.html

3. *Prokhorov D., Puskorius G., Feldkamp L.* Dynamical neural networks for control // In J. Kolen and S. Kremer (Eds.) A field guide to dynamical recurrent networks. NY: IEEE Press, 2001, pp. 257-289.

4. *Moody J., Wu L., Liao Y., Saffel M.* Performance function and reinforcement learning for trading systems and portfolios // Journal of Forecasting, 1998, vol.17, pp. 441-470.

5. *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.

6. *Prokhorov D.V., Wunsch D.C.* Adaptive critic designs // IEEE Transactions on Neural Networks, 1997, vol.8, pp.997-1007.

7. *Rumelhart D.E., Hinton G.E., Williams R.G.*: Learning representation by back-propagating error. Nature. 323 (1986) 533-536.

8. *Baldwin J.M.* A new factor in evolution // American Naturalist, 1896, vol. 30, pp. 441-451.

9. *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, V.4, N.3, 1996.

10. *Nepomnyashchikh V.A.* Selection behaviour in caddis fly larvae // In R. Pfeifer et al (Eds.) From Animals to Animats 5: Proceedings of the Fifth International Conference of the Society for Adaptive Behavior. Cambridge, MA: MIT Press, 1998, pp.155-160.

11. *Непомнящих В.А.* Как животные решают плохо формализуемые задачи поиска // Синергетика и психология. Тексты. Выпуск 3. Когнитивные процессы (под ред. Аршинова В.И., Трофимовой И.Н., Шендяпина В.М.) М.: Издательство Когнито-центр, 2004. С. 197-209.

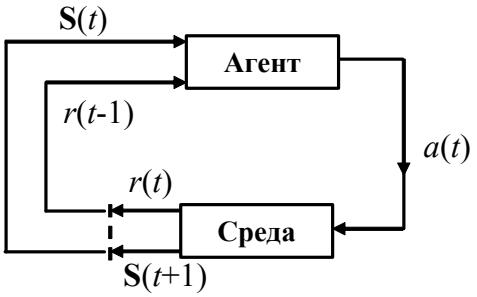

Рис. 1. Схема обучения с подкреплением.

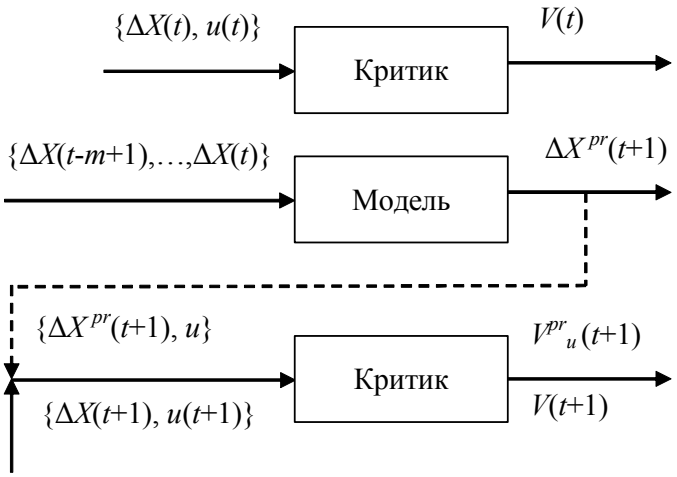

Рис. 2. Схема системы управления агента. НС Критика показана для двух последовательных тактов времени. Модель предназначена для прогнозирования изменения курса временного ряда. Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$ для текущей ситуации $\mathbf{S}(t) = \{\Delta X(t), u(t)\}$, для ситуации в следующий такт времени $\mathbf{S}(t+1)$ и для предсказываемых ситуаций для обоих возможных действий $\mathbf{S}^{\mathbf{pr}}_{u}(t+1) = \{\Delta X^{pr}(t+1), u\}$, $u = 0$ или $u = 1$.

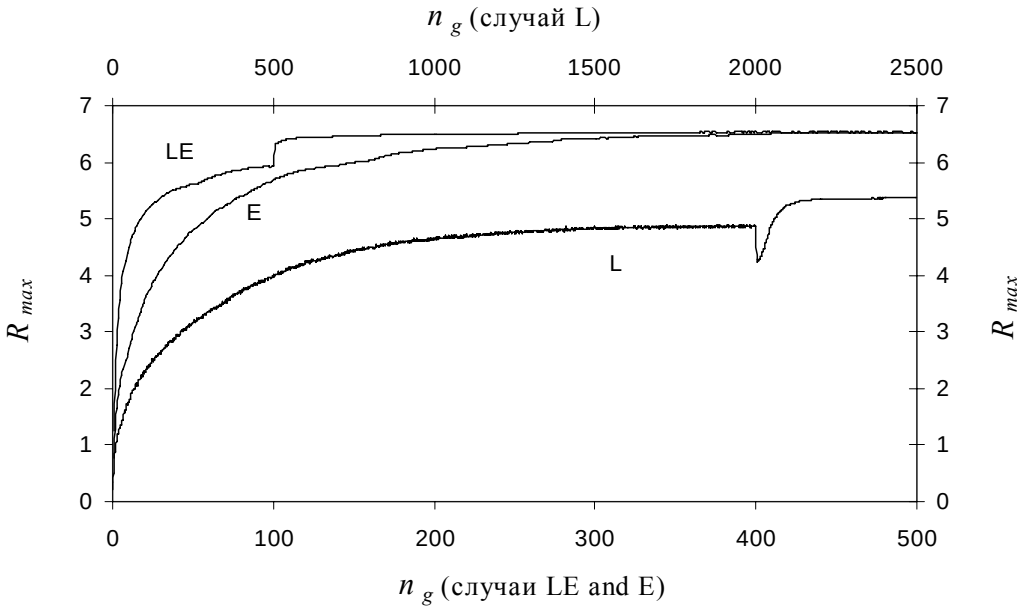

Рис. 3. Зависимости $R_{max}(n_g)$. Кривая LE соответствует случаю обучения, объединенного с эволюцией, кривая E – случаю чистой эволюции, кривая L – случаю чистого обучения. Временная шкала для случаев LE и E (номер поколения n_g) представлена снизу, для случая L (индекс n_g) – сверху. Моделирование проведено для синусоиды, кривые усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

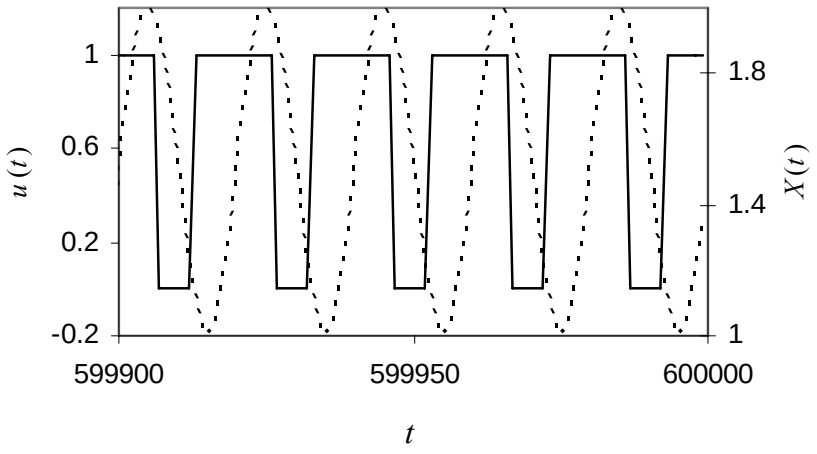

Рис. 4. Динамика поведения обучающегося агента для синусоиды (5). Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. Случай L.

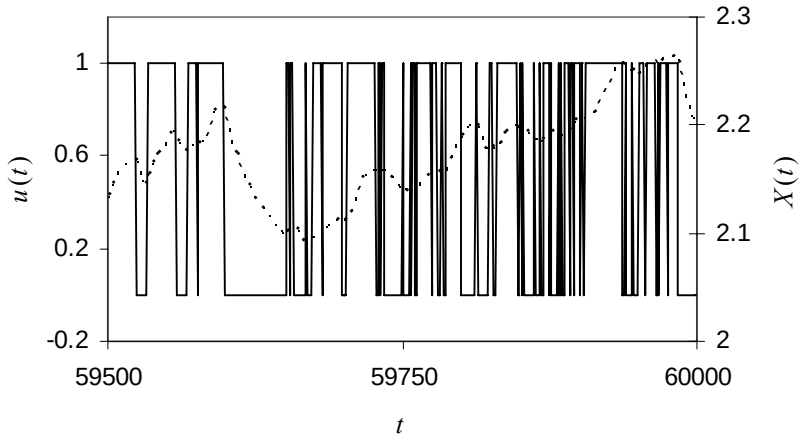

Рис. 5. Стратегия поведения агента. Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. Случай E. $n = 10$, $T = 200$. Стратегия поведения агента практически оптимальна: агент покупает/продает акции при росте/падении курса акций.

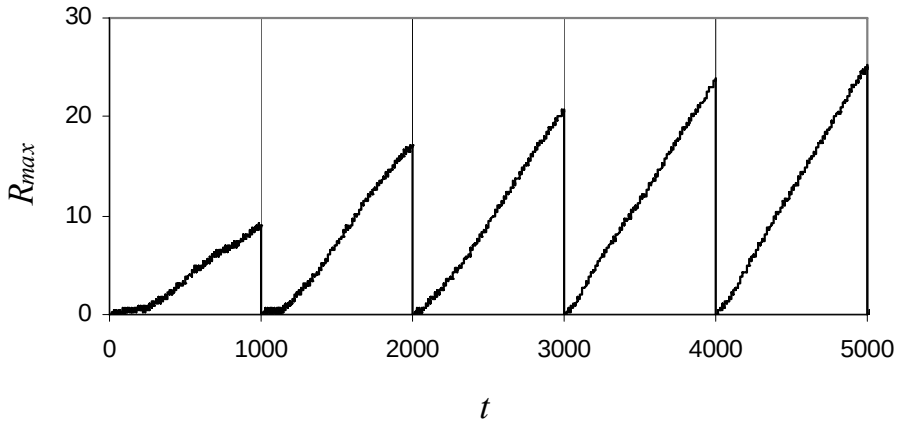

Рис. 6. Зависимость ресурса лучшего в популяции агента R_{max} от времени t для первых пяти поколений. Случай LE. $n = 10$, $T = 1000$. Моменты смены поколений показаны вертикальными линиями. Для первых двух поколений есть явная задержка в 100-300 тактов времени в росте ресурса агента. К пятому поколению агент обладает хорошей стратегией поведения с самого рождения, т.е. стратегия, изначально приобретаемая посредством обучения, становится наследуемой.

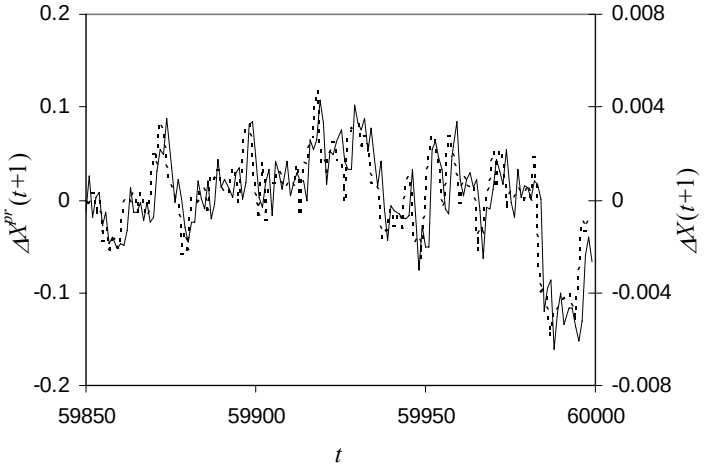

Рис. 7. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай E. $n = 10$, $T = 200$. Хотя обе кривые имеют сходную форму, по величине $\Delta X^{pr}(t+1)$ и $\Delta X(t+1)$ сильно различаются.

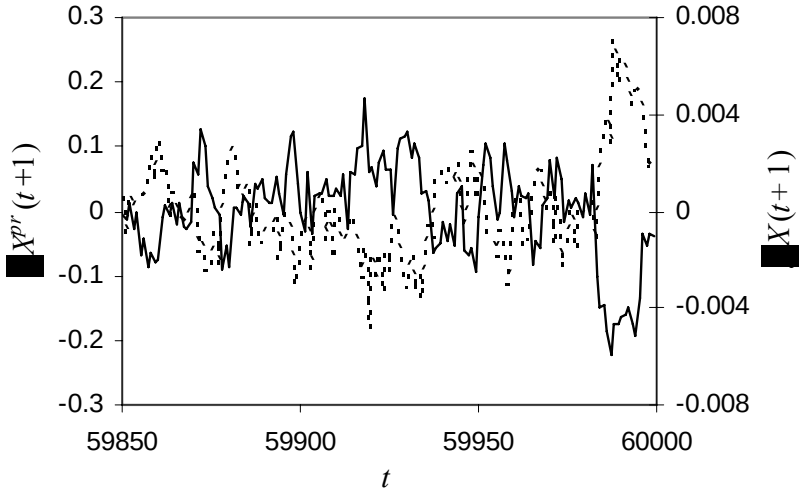

Рис. 8. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай LE. $n = 10$, $T = 200$. Кривые $\Delta X^{pr}(t+1)$ и $\Delta X(t+1)$ различаются как величиной, так и знаком.

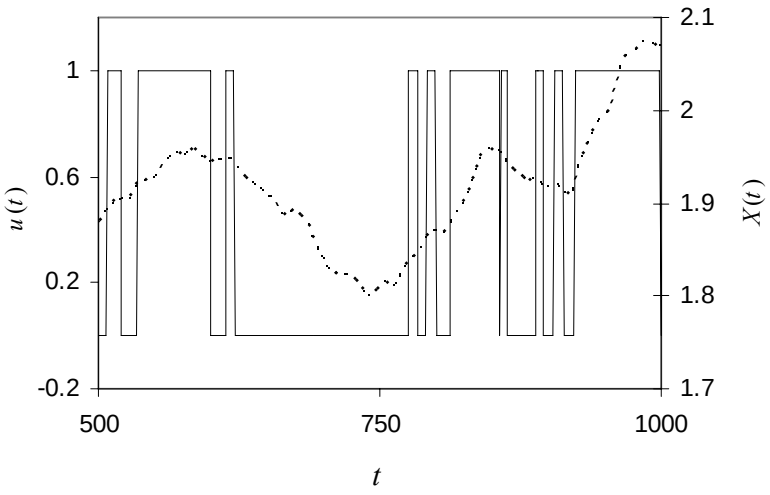

Рис. 9. Стратегия поведения агента в популяции. Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. $n = 100$, $T = 200$. Стратегия агента подобна поведению животных с инерционным переключением между двумя тактиками.