В.Г. Редько

Модели адаптивного поведения – биологически инспирированный подход к искусственному интеллекту[1]

Аннотация. Характеризуется направление исследований «Адаптивное поведение», в котором развивается биологически инспирированный подход к искусственному интеллекту. Излагаются примеры моделей адаптивного поведения, при этом особый акцент делается на модели самообучающихся автономных аниматов (модельных организмов). Характеризуется один из ключевых методов самообучения – метод обучения с подкреплением. Представлены исследования по оригинальному проекту «Мозг анимата».

Введение

В настоящей работе характеризуется направление исследований «Адаптивное поведение», которое сложилось сравнительно недавно, в начале 1990-х годов. В этом направлении изучаются архитектуры и принципы функционирования систем управления биологических или модельных организмов (например, роботов), которые обеспечивают приспособление организмов к внешней среде. Дальняя цель этого направления связана с исследованием когнитивных способностей биологических организмов, с интеллектом человека. Приложения моделей адаптивного поведения – искусственный интеллект, робототехника, модели адаптивного поведения в социально-экономических системах.

Раздел 1 характеризует это направление исследований в целом. В разделе 2 излагаются типичные примеры биологически инспирированных моделей адаптивного поведения. Раздел 3 описывает часто используемый в моделях адаптивного поведения метод самообучения – метод обучения с подкреплением. В разделе 4 излагается оригинальный проект «Мозг анимата», нацеленный на формирование общей платформы для систематического построения моделей адаптивного поведения.

1. Направление исследований «адаптивное поведение»

С начала 1990-х годов активно развивается направление «Адаптивное поведение» [1-3]. Основной подход этого направления – конструирование и исследование искусственных (в виде компьютерной программы или робота) «организмов», способных приспосабливаться к внешней среде. Эти организмы называются «*аниматами*» (от англ. animal + robot = animat).

Поведение аниматов имитирует поведение животных. Исследователи направления «Адаптивное поведение» стараются строить именно такие модели, которые применимы к описанию поведения *как реального животного, так и искусственного анимата* [4].

Программа-минимум направления «Адаптивное поведение» – исследовать архитектуры и принципы функционирования, которые позволяют животным или роботам жить и действовать в переменной внешней среде.

Программа-максимум этого направления – попытаться проанализировать эволюцию когни-

[1] Работа выполнена при финансовой поддержке программы № 14 Президиума РАН «Фундаментальные проблемы информатики и информационных технологий» (проект 2-45).

тивных способностей животных и эволюционное происхождение человеческого интеллекта [5].

Данное направление исследований рассматривается как бионический подход к разработке систем искусственного интеллекта [6].

Хотя официально «Адаптивное поведение» было провозглашено в 1990 г., существовали явные провозвестники этого направления, например, в нашей стране в 1960-70-х гг. подобные исследования вели М.Л. Цетлин, М.М. Бонгард, Д.А. Поспелов [7-9].

Это направление исследований использует ряд нетривиальных компьютерных методов:

– нейронные сети,

– генетический алгоритм [10] и другие методы эволюционной оптимизации,

– классифицирующие системы (Classifier Systems) [11],

– обучение с подкреплением (Reinforcement Learning) [12].

Нейросетевые и эволюционные методы достаточно хорошо известны, и здесь мы не будем на них останавливаться.

Классифицирующая система представляет собой набор правил вида «Если имеет место ситуация $\mathbf{S}(t)$, то нужно выполнить действие $\mathbf{A}(t)$, результатом действия будет следующая ситуация $\mathbf{S}(t+1)$» (t – дискретное время). Набор правил оптимизируется в результате обучения (путем модификации силы правил) и эволюционным путем (путем селекции правил и генерации новых правил). Отметим, что набор правил классифицирующей системы сходен с таковым в проекте «Животное» М.М. Бонгарда и сотрудников [8].

В обучении с подкреплением рассматривается анимат, взаимодействующий с внешней средой. В текущей ситуации $\mathbf{S}(t)$ анимат выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию $\mathbf{S}(t+1)$. Подкрепление может быть положительным (награда) или отрицательным (наказание). Цель анимата – максимизировать суммарную награду, которую можно получить в будущем в течение длительного периода времени. Подробнее метод обучения с подкреплением изложен в разделе 3.

Подчеркнем, что «Адаптивное поведение» – активно развивающееся направление исследований. Есть международное общество “International Society for Adaptive Behavior” (http://www.isab.org). Регулярно проводятся международные конференции “Simulation of Adaptive Behavior (From Animals to Animats)”. Издается журнал “Adaptive Behavior”.

В настоящее время исследования адаптивного поведения включают в себя работы по следующим темам [3]:

– сенсорные системы и управление,

– обучение и адаптация,

– выбор действий, навигация и внутренние модели мира,

– антиципаторное адаптивное поведение,

– нейроэволюция (настройка нейронных сетей аниматов эволюционными методами),

– возникновение языка и коммуникаций при адаптивном поведении,

– коллективное и социальное поведение,

– адаптивное поведение роботов,

– поведение и мышление как сложные адаптивные системы.

В следующем разделе характеризуются типичные примеры моделей адаптивного поведения.

2. Примеры моделей адаптивного поведения

2.1. Модели мозга и поведения в Институте нейронаук Дж. Эдельмана

В Институте нейронаук Дж. Эдельмана (http://www.nsi.edu) уже более 25 лет ведутся разработки поколений моделей работы мозга (Darwin I, Darwin II) и – в последние годы – исследования адаптивного поведения искусственного организма-устройства NOMAD (Neurally Organized Mobile Adaptive Device), построенного на базе этих моделей.

Принципы моделирования NOMAD’а (авторы называют его также Brain-based device) состоят в следующем:

1. устройство помещается в реальную физическую среду;

2. имеется некоторая поведенческая задача, которую должно решать устройство;

3. поведение устройства контролируется модельной нервной системой, которая отражает архитектуру мозга и динамику процессов в мозге;

4. поведение устройства и процессы в модельной нервной системе должны допускать сравнение с экспериментальными биологическими данными.

В одной из последних работ по NOMAD'у промоделировано поведение мыши в лабиринте Морриса [13].

Исследования поведения мыши или крысы в лабиринте Морриса – один из канонических биологических экспериментов, который состоит в следующем. Имеется бассейн с непрозрачной жидкостью (например, это может быть вода, подкрашенная молоком), на бортах бассейна есть рисунки, которые мышь видит и может использовать для ориентировки. В определенном месте бассейна есть скрытая платформа, которую мышь может найти и тем самым спастись – не утонуть. Мышь бросают в бассейн, она плавает некоторое время и либо находит платформу и спасается, либо начинает тонуть (тогда ее спасает экспериментатор). После ряда экспериментов мышь начинает использовать ориентиры на бортах бассейна и находить платформу за достаточно короткое время.

Поведение NOMAD'а в лабиринте Морриса моделировалась следующим образом. NOMAD представлял собой подвижное устройство на колесах, которое управлялось нейронной сетью, состоящей из 90000 нейронов (1.4·106 синапсов). В ней было выделено 50 различных нейронных областей, в частности, несколько областей гиппокампа. Программно нейронная сеть была реализована на основе компьютерного кластера.

Сенсорная система NOMAD'а включала зрение, обонятельную систему, позволяющую отслеживать свои собственные следы, систему инфракрасных приемников-излучателей, обеспечивающую избегание столкновений, и специальный детектор скрытой от зрения платформы, позволяющий обнаруживать эту платформу только тогда, когда NOMAD находится непосредственно над ней.

NOMAD помещался в комнату, в которой была скрытая платформа; на стенах комнаты были разноцветные полосы – ориентиры. В начале каждого из компьютерных экспериментов NOMAD помещался в разные участки комнаты, задача NOMAD'а была найти платформу. Обучение нейронных сетей NOMAD'а осуществлялось по модифицированному правилу Хебба на основе подкреплений (получаемых при нахождении скрытой платформы) и наказаний (получаемых при приближении к стенам комнаты).

Было продемонстрировано, что

1. NOMAD обучается находить платформу достаточно быстро (за 10-20 попыток),

2. в модельном гиппокампе формируются *нейроны места*, активные только тогда, когда NOMAD находится в определенных участках комнаты,

3. в модельном гиппокампе формируются связи между отдельными нейронными областями, отражающие причинно-следственные зависимости.

Изложенная модель представляет собой эмпирическое компьютерное исследование, хорошо продуманное с биологической точки зрения. Поведение NOMAD'а нетривиально и, вместе с тем, было бы полезно более формализованное исследование, дополнительное к этим эмпирическим работам.

2.2. Модель эволюционного возникновения коммуникаций в коллективе роботов

В данной модели (Д. Марокко, С. Нолфи, Институт когнитивных наук и технологий, Рим) [14] исследовались вопросы: Как могут эволюционно возникнуть коммуникации между модельными организмами? Как в эволюционном процессе может сформироваться сигнальная обработка информации?

Рассматривалась следующая проблема. Есть четыре двухколесных робота, каждый из которых управляется рекуррентной нейронной сетью, состоящей из 5-ти нейронов. На входы нейронов поступают сигналы от 8-ми инфракрасных датчиков и от 4-х датчиков, воспринимающих звуковые сигналы с разных сторон. Нейронная сеть имела 3 выходных нейрона, два из которых определяли скорость движения двух колес, а третий нейрон – интенсивность силы звука, издаваемого в данный момент роботом. В ограниченной области пространства имелось две кормушки, и роботам нужно было, используя свои нейронные сети и звуковые сигналы разной интенсивности, как можно быстрее распределиться по кормушкам: по 2 робота на каждую из кормушек.

Нейронные сети роботов оптимизировались эволюционным путем. В результате в течение 2000 поколений у роботов сформировались сигналы 5 различных видов (т.е. разной интенсивности). Используя эти сигналы, роботы достаточно устойчиво находили требуемое распределение по кормушкам.

В [14] продемонстрировано, что в эволюционирующей популяции роботов, управляемых рекуррентными нейронными сетями, может формироваться система коммуникаций, позволяющая решать достаточно нетривиальную задачу распределения роботов по кормушкам.

2.3. Бионическая модель поискового адаптивного поведения

Одно из актуальных направлений исследований в рамках моделирования адаптивного поведения – имитация поискового поведения животных. В нашей работе [15] исследовано поисковое поведение на примере личинок ручейников Chaetopteryx villosa – насекомых, обитающих на дне водоемов. Личинки носят на себе «домик» – трубку из песка и других частиц. Строительство требует меньше времени, усилий и белка, если личинки используют относительно крупные и плоские частицы: общая протяженность швов между составляющими его немногими крупными частицами оказывается меньше, чем в том случае, когда домик сооружается из большого числа мелких частиц. Однако поиск крупных частиц на дне водоема требует затрат времени и энергии, не известных ручейнику заранее. Задача осложняется еще и тем, что личинки при поиске частиц не пользуются зрением и могут обнаружить частицу и определить её размер только наощупь, что требует дополнительных затрат времени.

В [15] построена компьютерная модель поискового поведения личинок ручейников, строящих чехол-домик из частиц разного размера и ведущих поиск скоплений подходящих частиц. Модель использует понятие мотивации $M(t)$ к прикреплению частиц к домику. Продемонстрирована адекватность построенной модели биологическим экспериментальным данным. Модель характеризуется как своей спецификой, обусловленной памятью о размерах последних обработанных частиц, так и общими свойствами инерционного переключения, позволяющими животным выявлять и использовать при адаптивном поведении наиболее общие закономерности взаимодействия с внешней средой.

Хотя модель была биологически инспирирована, она вводит важное понятие мотивации M(t), причем схема регулирования мотивации проста и эффективна и может быть применена в ряде приложений. Эта схема включает в себя инерционность, случайные колебания и направленное воздействие (для личинок ручейников это воздействие увеличивает мотивацию при тестировании крупной частицы после мелкой и уменьшает мотивацию в противном случае). Как продемонстрировано в [15], такая схема регулирования мотивации может быть применена при поиске максимума функции f(x,y). Кратко изложим эту схему.

Рассматривается анимат, который может двигаться в двумерном пространстве(X,Y). Задача анимата – поиск максимума функции $f(x,y)$. Анимат может оценивать изменение текущего значения функции $f(x,y)$ по сравнению с предыдущим тактом времени $\Delta f(t) = f(t) - f(t\text{-}1)$; $t = 1,2,\ldots$ Каждый такт времени анимат совершает движение, при этом его координаты x, y изменяются на величины $\Delta x(t)$, $\Delta y(t)$ соответственно. Анимат имеет две тактики поведения: А - двигаться в выбранном направлении, Б - изменить направление движения случайным образом. Смещение анимата в следующий такт времени $\Delta x(t+1)$, $\Delta y(t+1)$ для этих тактик определяется различным образом. Переключение между тактиками регулируется величиной $M(t)$, зависимость от времени которой определяется выражением

$$M(t) = k_1 M(t\text{-}1) + \xi(t) + I(t), \qquad (1)$$

где k_1 – параметр, характеризующий инерционность переключения между тактиками ($0 < k_1 < 1$), $\xi(t)$ – нормально распределенная случайная величина со средним, равным 0, и средним квадратическим отклонением σ, $I(t)$ – интенсивность раздражителя. В простейшем случае интенсивность раздражителя $I(t)$ равна:

$$I(t) = k_2 \Delta f(t), \qquad (2)$$

где $k_2 > 0$. Предполагаем, что при $M(t) > 0$ анимат придерживается тактики А, при $M(t) < 0$ – тактики Б. Величину $M(t)$ можно рассматривать как мотивацию к выбору тактики А.

Тактика А. При движении в выбранном направлении анимат смещается на величину R_0: $\Delta x(t+1) = R_0 \cos\varphi_0$, $\Delta y(t+1) = R_0 \sin\varphi_0$, где угол φ_0 характеризует сохраняющееся направление движения анимата.

Тактика Б. При случайном повороте анимат также смещается на некоторую величину r_0, а направление его движения случайно варьируется: $\Delta x(t+1) = r_0 \cos\varphi$, $\Delta y(t+1) = r_0 \sin\varphi$, где $\varphi = \varphi_0 + w$,

φ_0 – угол, задающий направление движения в такт времени t, величина w нормально распределена (среднее значение w равно нулю, среднее квадратическое отклонение равно w_0), φ – угол, характеризующий направление движения в такт времени t+1.

Отметим, что данная схема легко обобщается на случай поиска оптимума функции многих переменных. Необходимо подчеркнуть, что схемы поиска экстремума функций, близкие к изложенной, разрабатывались рядом авторов (Цыпкин Я.З., Растригин Л.А., Неймарк Ю.И., например, [16]). Однако данная схема в наиболее явном и четком виде использует свойства инерционности, стохастичности и влияние изменения оптимизируемой функции (формулы (1)-(2)). Результаты моделирования по этой схеме и ее обсуждение в [15].

Представленные в данном разделе модели довольно разнородны. В разделе 4 излагается проект «Мозг анимата» [17,18], который нацелен на формирование общей платформы для систематического построения моделей адаптивного поведения. Однако предварительно в разделе 3 будет представлен метод обучения с подкреплением, который используется в первом варианте проекта «Мозг анимата».

3. Обучение с подкреплением

Метод обучения с подкреплением (Reinforcement Learning) был развит в работах Р. Саттона и Э. Барто [12]. В этом методе рассматривается анимат, взаимодействующий с внешней средой (Рис. 1). В текущей ситуации $\mathbf{S}(t)$ анимат выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию $\mathbf{S}(t+1)$; t = 1,2,… Подкрепление $r(t)$ может быть положительным (награда) или отрицательным (наказание).

Цель анимата – максимизировать суммарную награду, которую можно получить в будущем в течение длительного периода времени. Анимат имеет свою внутреннюю «субъективную» оценку суммарной награды и в процессе обучения постоянно совершенствует эту оценку. Эта оценка определяется с учетом дисконтного фактора γ:

$$U(t) = \sum_{k=0}^{\infty} \gamma^k r(t+k), \qquad (3)$$

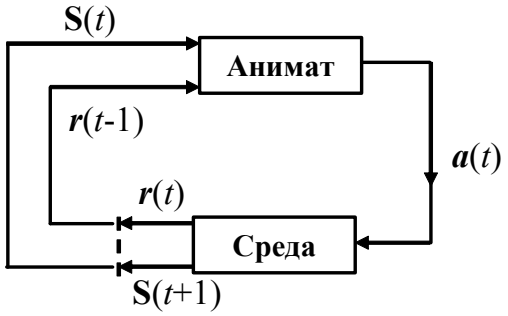

Рис. 1. Схема обучения с подкреплением

где $U(t)$ – оценка ожидаемой суммарной награды, $0 < \gamma < 1$. Дисконтный фактор γ учитывает, что чем дальше анимат «заглядывает» в будущее, тем меньше у него уверенность в оценке награды.

Если множество возможных ситуаций $\{\mathbf{S}_i\}$ и действий $\{a_j\}$ конечно, то существует простой метод обучения SARSA, каждый шаг которого соответствует цепочке событий $\mathbf{S}(t) \rightarrow a(t) \rightarrow r(t) \rightarrow \mathbf{S}(t+1) \rightarrow a(t+1)$.

Кратко опишем метод SARSA. В этом методе итеративно формируются оценки величины суммарной награды $Q(\mathbf{S}(t), a(t))$, которую получит анимат, если в ситуации $\mathbf{S}(t)$ он выполнит действие $a(t)$. Математическое ожидание награды равно:

$$Q(\mathbf{S}(t), a(t)) = \mathrm{E}\,[r(t) + \gamma\, r(t+1) + \gamma^2\, r(t+2) + \ldots] \mid \mathbf{S} = \mathbf{S}(t),\ a = a(t). \qquad (4)$$

Из (3), (4) следует $Q(\mathbf{S}(t), a(t)) = E\,[r(t) + \gamma\, Q(\mathbf{S}(t+1), a(t+1))]$. Ошибку естественно определить так [12]:

$$\delta(t) = r(t) + \gamma\, Q(\mathbf{S}(t+1), a(t+1)) - Q(\mathbf{S}(t), a(t)). \qquad (5)$$

Величина $\delta(t)$ называется ошибкой временной разности и равна разности между той оценкой суммарной величины награды, которая формируется у анимата для момента времени t после выбора действия $a(t+1)$ в следующей ситуации $\mathbf{S}(t+1)$, и предыдущей оценкой этой же величины. Предыдущая оценка равна $Q(\mathbf{S}(t), a(t))$, новая оценка равна $r(t) + \gamma\, Q(\mathbf{S}(t+1), a(t+1))$, что и отражает формула (5).

Каждый такт времени происходит как выбор действия, так и обучение анимата. Выбор действия происходит так:

- в момент t с вероятностью $1-\varepsilon$ выбирается действие с максимальным значением $Q(\mathbf{S}(t), a_j)$: $a(t) = \arg\max_j \{Q(\mathbf{S}(t), a_j)\}$

- с вероятностью ε выбирается произвольное действие, $0 < \varepsilon << 1$.

Такую схему выбора действия называют «ε-жадным правилом».

Обучение, т.е. переоценка величин $Q(\mathbf{S}, a)$ происходит в соответствии с оценкой ошибки $\delta(t)$:

$$\Delta Q(\mathbf{S}(t), a(t)) = \alpha\, \delta(t) = \\ = \alpha\, [r(t) + \gamma\, Q(\mathbf{S}(t+1), a(t+1)) - Q(\mathbf{S}(t), a(t))], \quad (6)$$

где α – параметр скорости обучения.

Метод обучения с подкреплением идейно связан с методом динамического программирования. И в том, и в другом случае общая оптимизация многошагового процесса принятия решения происходит путем упорядоченной процедуры одношаговых оптимизирующих итераций, причем оценки эффективности тех или иных решений, соответствующие предыдущим шагам процесса, переоцениваются с учетом знаний о возможных будущих шагах. Например, при решении задачи поиска оптимального маршрута в лабиринте от стартовой точки к определенной целевой точке сначала находится конечный участок маршрута, непосредственно приводящий к цели, а затем ищутся пути, приводящие к конечному участку, и т.д. В результате постепенно прокладывается трасса маршрута от его конца к началу. Обучение с подкреплением, нейросетевые адаптивные критики (см. ниже) и подобные методы часто называют приближенным динамическим программированием [19].

Важное достоинство метода обучения с подкреплением – его простота. Анимат получает от учителя или из внешней среды только сигналы подкрепления r(t). Это радикально отличает этот метод (фактически метод самообучения) от таких традиционных в теории нейронных сетей методах обучения, как метод обратного распространения ошибок, для которого учитель точно определяет, что должно быть на выходе нейронной сети при заданном входе.

Метод обучения с подкреплением был исследован рядом авторов [12] и был использован в многочисленных приложениях. В частности, применения этого метода включают в себя:

- оптимизацию игры в триктрак (достигнут уровень мирового чемпиона);
- оптимизацию системы управления работой лифтов;
- формирование динамического распределения каналов для мобильных телефонов;
- оптимизацию расписания работ на производстве.

Подчеркнем, что метод обучения с подкреплением может рассматриваться как развитие автоматной теории адаптивного поведения, разработанной в работах М.Л. Цетлина и его последователей [7,9]. В свою очередь, метод обучения с подкреплением получил свое развитие в работах по нейросетевым адаптивным критикам [20], в которых применяются нейросетевые аппроксиматоры функций оценки качества функционирования анимата. Простейшая схема нейросетевых адаптивных критиков рассматривается в следующем разделе.

4. Проект «Мозг анимата»

4.1. Общая архитектура «Мозга анимата»

Предполагается, что система управления аниматом имеет иерархическую архитектуру. Базовым элементом системы управления является отдельная функциональная система (ФС).

Верхний уровень соответствует основным потребностям организма: питания, размножения, безопасности, накопления знаний. Более низкие уровни системы управления соответствуют тактическим целям поведения. Все эти уровни реализуются с помощью ФС. Управление с верхних уровней может передаваться на нижние уровни и возвращаться назад. Предполагается, что система управления аниматом функционирует в дискретном времени, каждый такт времени активна только одна ФС.

В пп. 4.2, 4.3 рассматривается простая формализация ФС на основе нейросетевых адаптивных критиков. ФС моделирует: 1) прогноз результата действия, 2) сравнение прогноза и результата, 3) коррекцию прогноза путем обучения в соответствующих нейронных сетях, 4) формирование оценок качества ситуаций, уточнение этих оценок, 5) принятие решения.

Схема адаптивного критика состоит из двух блоков: Модель и Критик. Предполагается, что Модель и Критик – нейронные сети (НС) и что производные по весам НС этих блоков могут быть вычислены обычным методом обратного распространения ошибки [21]. Цель адаптивного критика – максимизировать функцию суммарной награды $U(t)$ (формула (3)).

Для развития проекта важно проверить, как функционируют простые схемы адаптивных

критиков в конкретных моделях. Ниже излагаются результаты исследования такой модели.

4.2. Модель эволюции популяции самообучающихся агентов на базе нейросетевых адаптивных критиков [22]

Далее исследуется модель эволюции популяции самообучающихся автономных агентов и анализируется взаимодействие между обучением и эволюцией. Модель отрабатывается на примере агента-брокера. Этот пример используется только для определенности, аналогично можно рассматривать функционирование модельного организма, более подобного биологическим прототипам, например, организма, помещенного во внешнюю среду, которая определяется зависимостью температуры от времени, аналогичной временной зависимости курса акций для агента-брокера.

Схема агента. Рассматривается модель 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* = 1,2,...; *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 [23]:

$$C(t+1) = C(t)\,[1 + u(t+1)\,\Delta X(t+1) / X(t)], \quad (7)$$

где $\Delta X(t+1) = X(t+1) - X(t)$ – текущее изменение курса акций. Используется логарифмическая шкала для ресурса агента, $R(t) = \log C(t)$ [24]. Текущее подкрепление агента $r(t) = R(t+1) - R(t)$ равно:

$$r(t) = \log [1 + u(t+1)\,\Delta X(t+1) / X(t)]. \quad (8)$$

Для простоты предполагается, что переменная u может принимать только два значения $u = 0$ (весь капитал в деньгах) или $u = 1$ (весь капитал в акциях).

Алгоритм обучения. Система управления агента – это простой адаптивный критик, состоящий из двух НС: Модель и Критик (Рис. 2). В предположении $\Delta X(t) << X(t)$ считаем, 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 *ΔX(t-m+1), … , ΔX(t),* на выходе

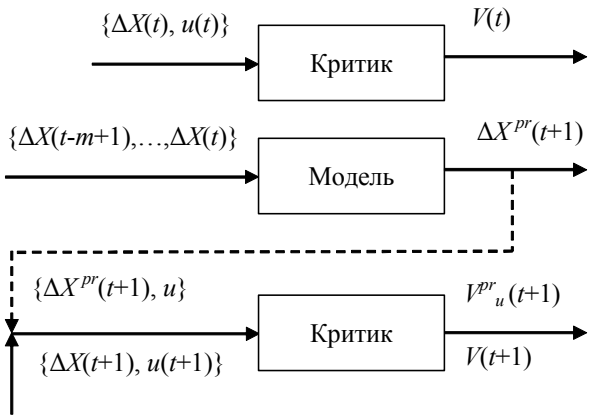

Рис. 2. Схема системы управления агента

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

формируется прогноз изменения курса в следующий такт времени $\Delta Xpr(t+1)$. Модель представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x}^{\mathbf{M}} = \{\Delta X(t-m+1),\ldots,\Delta X(t)\},$$
$$y^{M}{}_{j} = \text{th}\,(\textstyle\sum_i w^{M}{}_{ij}\, x^{M}{}_{i}),\ \Delta X^{pr}(t+1) = \textstyle\sum_j v^{M}{}_{j}\, y^{M}{}_{j},$$

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

Критик предназначен для оценки качества ситуаций *V(S),* а именно, оценки ожидаемой суммарной награды *U(t)* (см. (3)) для агента, находящегося в рассматриваемой ситуации S. Критик представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x}^{\mathbf{C}} = \mathbf{S}(t) = \{\Delta X(t), u(t)\},\ y^{C}{}_{j} = \text{th}\,(\textstyle\sum_i w^{C}{}_{ij}\, x^{C}{}_{i}),$$
$$V(t) = V(\mathbf{S}(t)) = \textstyle\sum_j v^{C}{}_{j}\, y^{C}{}_{j},$$

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.

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 $V^{pr}{}_u(t+1) = V(\mathbf{S}^{\mathbf{pr}}{}_u(t+1))$, где $\mathbf{S}^{\mathbf{pr}}{}_u(t+1)=\{\Delta X^{pr}(t+1), u\}$, $u = 0$ либо $u = 1$.

3) Применяется ε-жадное правило [12]: действие, соответствующее максимальному значению $V^{pr}{}_u(t+1)$ выбирается с вероятностью $1 - \varepsilon$, и альтернативное действие выбирается с вероятностью ε ($0 < \varepsilon << 1$). Выбор действия есть выбор величины $u(t+1)$: перевести весь капитал в деньги, $u(t+1) = 0$; либо в акции, $u(t+1) = 1$.

4) Выбранное действие $u(t+1)$ выполняется. Происходит переход к моменту времени $t+1$. 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 [21]. Скорость обучения Модели равна $\alpha_M > 0$.

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:

$$\delta(t) = r(t) + \gamma V(t+1) - V(t) \qquad (9)$$

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 изменяется в соответствии с подкреплениями агента: $R(t+1) = R(t) + r(t)$, где $r(t)$ определено в (8).

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

Начальные веса синапсов обеих 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.

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 $\mathbf{G}$ отличаются от генома родителя небольшими мутациями. Более конкретно, предполагается, что в начале каждого нового (n_g+1)-го поколения для каждого агента его геном формируется следующим образом $G_i(n_g+1) = G_{best,\ i}(n_g) + \text{rand}_i$, $\mathbf{W}_0(n_g+1) = \mathbf{G}(n_g+1)$, где $\mathbf{G}_{\mathbf{best}}(n_g)$ – геном лучшего агента предыдущего n_g-го поколения и rand_i – это 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 агента) 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 $\mathbf{W} = \mathbf{W}_0 = \mathbf{G}$.

4.3. Результаты моделирования [22]

Общие особенности адаптивного поиска. Изложенная модель была реализована в виде компьютерной программы. В компьютерных экспериментах использовалось два варианта временного ряда:

1) синусоида:

$$X(t) = 0.5(1 + \sin(2\pi t/20)) + 1, \qquad (10)$$

2) стохастический временной ряд [23]:

$$X(t) = \exp(p(t)/1200),\ p(t) = p(t\text{-}1) + \beta(t\text{-}1) + k_1\lambda(t),$$
$$\beta(t) = k_2\beta(t\text{-}1) + \mu(t), \qquad (11)$$

где $\lambda(t)$ и $\mu(t)$ – два нормальных процесса с нулевым средним и единичной дисперсией, $k_1 = 0.3$; $k_2 = 0.9$.

Некоторые параметры модели имели одно и то же значение для всех экспериментов: 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 $P_{mut} = 0.1$. Остальные параметры (продолжительность поколения T и численность популяции n) принимали разные значения в разных экспериментах.

Были проанализированы следующие варианты рассматриваемой модели:

– Случай L (чистое обучение); в этом случае рассматривался отдельный агент, который только обучался.

– Случай E (чистая эволюция), т.е. рассматривается эволюционирующая популяция без обучения.

– Случай LE (эволюция + обучение), т.е. 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 (10) 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. 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 Rmax(500) = 6.5. Однако эволюция и обучение вместе обеспечивают нахождение больших значений Rmax быстрее, чем эволюция отдельно – существует симбиотическое взаимодействие между обучением и эволюцией.

Из (8) следует, что существует оптимальная стратегия поведения агента: вкладывать весь капитал в акции (u(t+1) = 1) при прогнозе роста курса (ΔXpr(t+1) > 0), вкладывать весь капитал в деньги (u(t+1) = 0) при прогнозе падения курса (ΔXpr(t+1) < 0). Анализ экспериментов, представленных на Рис. 3, показал, что в случаях LE (обучение + эволюция) и E (чистая эволюция) такая оптимальная стратегия находится. Это соответствует асимптотическому значению ресурса Rmax(500) = 6.5.

В случае L (чистое обучение) асимптотическое значение ресурса (Rmax(2500) = 5.4) существенно меньше. Анализ экспериментов для этого случая показал, 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. Та же тенденция к явному предпочтению вкладывать капитал в акции при чистом обучении наблюдалась и для экспериментов на стохастическом ряде (11).

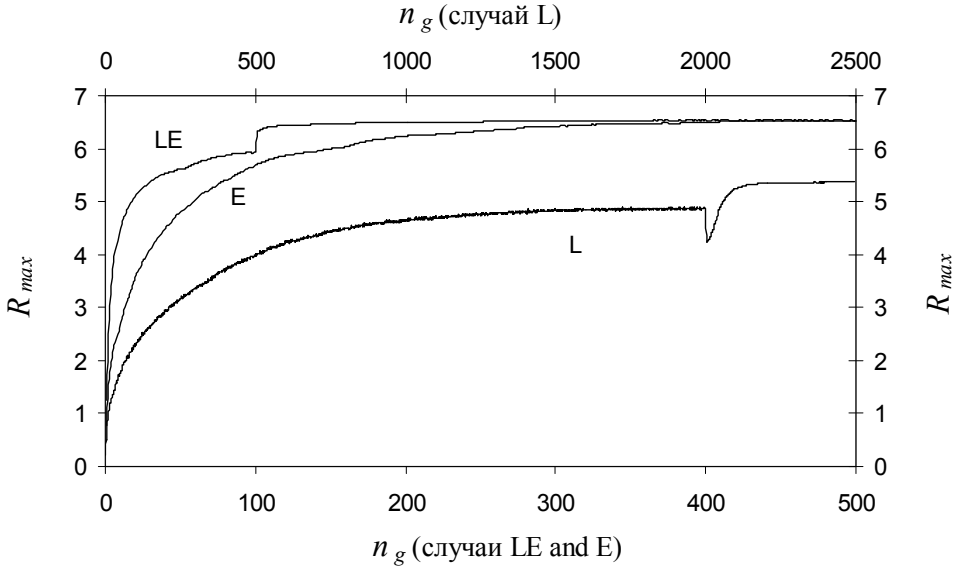

Рис. 3. Зависимости Rmax(ng)

Кривая LE соответствует случаю эволюции, объединенной с обучением, кривая E – случаю чистой эволюции, кривая L – случаю чистого обучения. Временная шкала для случаев LE и E (номер поколения ng) представлена снизу, для случая L (индекс ng) – сверху. Моделирование проведено для синусоиды (10), кривые усреднены по 1000 экспериментам; n = 10, T = 200.

Представленные результаты демонстрируют, что хотя обучение в настоящей модели и несовершенно, оно способствует более быстрому нахождению оптимальной стратегии поведения по сравнению со случаем чистой эволюции.

Взаимодействие между обучением и эволюцией. Эффект Болдуина. Если длительность поколения T была достаточно большой (1000 и более тактов времени), то для случая LE часто наблюдалось сильное влияние обучения на эволюционный процесс. В первых поколениях 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 (Рис. 4). Это можно интерпретировать как проявление известного эффекта Болдуина: исходно приобретаемый навык в течение ряда поколений становился наследуемым [25, 26].

Эволюция самообучающихся агентов и проект «Мозг анимата». Хотя исследованная модель эволюции популяции агентов на основе нейросетевых адаптивных критиков интересна сама по себе, однако как показывают компьютерные расчеты, НС Модель этих агентов в определенных условиях может делать неправильные предсказания: изменения курса акций предсказываются с точностью до знака и предсказанные изменения могут сильно отличаться от реальных по величине. И эти, строго говоря, неверные предсказания могут вполне разумно использоваться. Это показывает, что необходима определенная осторожность в выборе базовой модели функциональной системы для проекта «Мозг анимата». Поэтому имеет смысл рассмотреть и другие возможности для базовой модели ФС. В п. 4.4 излагается версия «Мозга анимата» на основе нейронных сетей, которые запоминают отображения между ситуациями и действиями, а также прогнозы будущих ситуаций.

4.4. Модель управления анимата на базе отображающих нейронных сетей [28]

Как и выше, предполагается, что есть эволюционирующая популяция модельных организмов (аниматов), системы управления которых оптимизируются как путем обучения, так и в процессе эволюции.

Система управления анимата основана на простых нейронных сетях (НС) и обеспечивает самообучение на основе поощрений или наказаний. В систему управления входят функциональные системы (ФС). Каждая ФС содержит две НС: Контроллер и Модель. На вход активной ФС поступает информация о текущей ситуации $\mathbf{S}(t)$. $\mathbf{S}(t)$ – набор параметров, характеризующий внешнюю и внутреннюю среду анимата. Контроллер по известной ситуации $\mathbf{S}(t)$ формирует действие анимата $\mathbf{A}(t)$, т.е. осу-

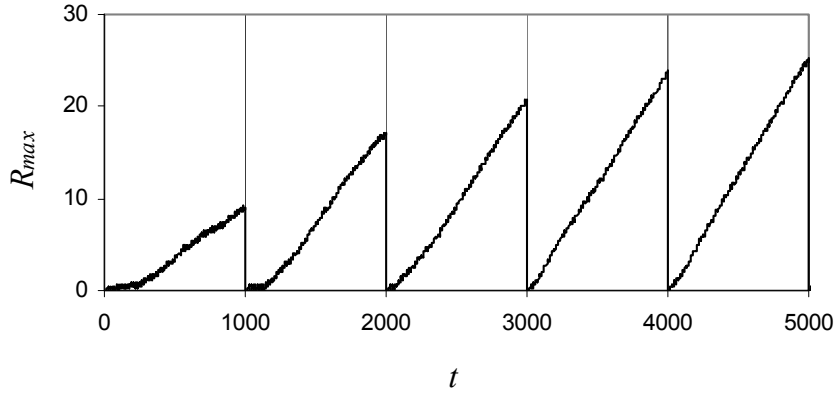

Рис. 4. Зависимость ресурса лучшего в популяции агента Rmax от времени t для первых пяти поколений

Случай LE (эволюция, объединенная с обучением); размер популяции n = 10, длительность поколения T = 1000. Моменты смены поколений показаны вертикальными линиями. Для первых двух поколений есть явная задержка в 200-300 тактов времени в росте ресурса агента. К пятому поколению лучший агент «знает» хорошую стратегию поведения с самого рождения, т.е. стратегия, изначально приобретаемая посредством обучения, становится наследуемой.

ществляет отображение $\mathbf{S}(t) \rightarrow \mathbf{A}(t)$. Часть действий $\mathbf{A}(t)$ соответствует формированию командных сигналов на эффекторы (собственно действия), часть действий соответствует передаче управления другим ФС. Модель по известным векторам $\mathbf{S}(t)$ и $\mathbf{A}(t)$ прогнозирует ситуацию в следующий такт времени $\mathbf{S}(t+1)$, т.е. формирует отображение $\{\mathbf{S}(t), \mathbf{A}(t)\} \rightarrow \mathbf{S}^{\mathbf{pr}}(t+1)$. Отображения $\mathbf{S}(t) \rightarrow \mathbf{A}(t)$ и $\{\mathbf{S}(t), \mathbf{A}(t)\} \rightarrow \mathbf{S}^{\mathbf{pr}}(t+1)$ запоминаются в весах синапсов НС Контроллера и Модели.

Каждый такт времени активна только одна ФС. Активная ФС может передать управление другой ФС. Есть матрица связей между ФС $||C_{ij}||$ ($C_{ik} > 0$). Величины C_{ij} определяют вероятности передачи управления между ФС. Ниже излагаются принципы работы ФС и схемы обучения и эволюции для этого варианта.

Нейронные сети отдельной ФС. На вход Контроллера подается вектор $\mathbf{S}(t)$. На выходе Контроллера формируется вектор действий $\mathbf{A}(t)$. Выбор действия производится по максимальной компоненте вектора $\mathbf{A}(t)$. Если выбрано действие «передать управление другой ФС», то та ФС, которой передается управление, выбирается случайным образом, с вероятностями, пропорциональными C_{ij} .

Работа НС Контроллера описывается формулами (для простоты рассматриваем двухслойные НС):

$$\mathbf{x}^{\mathbf{C}} = \mathbf{S}(t), \qquad y^C_j = \text{th}\,(\textstyle\sum_i w^C_{ij}\, x^C_i), \qquad A_k(t) = \textstyle\sum_j v^C_{jk}\, y^C_j\,, \tag{12}$$

где $\mathbf{x}^{\mathbf{C}}$ – входной вектор, $\mathbf{y}^{\mathbf{C}}$ – вектор выходов нейронов скрытого слоя, w^C_{ij} и v^C_{jk} – веса синапсов НС, $A_k(t)$ – компоненты вектора действий $\mathbf{A}(t)$.

На вход НС Модели подается составной вектор $\mathbf{x}^{\mathbf{M}} = \{\mathbf{S}(t), \mathbf{A}(t)\}$. На выходе Модели формируется вектор $\mathbf{S}^{\mathbf{pr}}(t+1)$, характеризующий прогнозируемые результаты действия. Работа НС Модели описывается формулами:

$$\mathbf{x}^{\mathbf{M}} = \{\mathbf{S}(t), \mathbf{A}(t)\}, y^M_j = \text{th}(\textstyle\sum_i w^M_{ij}\, x^M_i), \qquad S^{pr}_k(t+1) = \text{F}(\textstyle\sum_j v^M_{jk}\, y^M_j), \tag{13}$$

где $\mathbf{x}^{\mathbf{M}}$ – входной вектор, $\mathbf{y}^{\mathbf{M}}$ – вектор выходов нейронов скрытого слоя, w^M_{ij} и v^M_{jk} – веса синапсов НС, $S^{pr}_k(t+1)$ – компоненты вектора прогнозируемой ситуации $\mathbf{S}^{\mathbf{pr}}(t+1)$, F(.) – активационная функция выходных нейронов.

Обучение нейронных сетей. Имеется два режима обучения: 1) грубого поиска, 2) тонкой доводки.

Грубый поиск происходит, если есть сильное рассогласование между ожидаемым и полученным результатом: прогноз ситуации $\mathbf{S}^{\mathbf{pr}}(t+1)$ в активной ФС существенно отличается от реальной ситуации $\mathbf{S}(t+1)$. Под сильным рассогласованием будем подразумевать качественное различие в существенных компонентах векторов: например, ожидалось увеличение ресурса, а произошло уменьшение ресурса. Каждая ФС имеет свой вектор-маску $\mathbf{M}$, этот вектор имеет компоненты, равные 0 либо 1, единичные компоненты вектора-маски определяют существенные компоненты вектора ситуации $\mathbf{S}(t+1)$. Существенные компоненты определяют, какая именно закономерная связь проверяется данной Моделью.

Предполагаем, что при грубом поиске: а) происходит возврат управления к той ФС, от которой было передано управление текущей ФС, б) меняются связи между ФС. Изменение связей происходит следующим образом. Рассматривается связь C_{ij} между текущей ФС j и ФС i , от которой было передано управление текущей ФС, сработавшей в такт времени t. Будем считать, что такая передача управления была сделана неправильно (j-я ФС не справилась с «порученной ей задачей»), следовательно, нужно уменьшить коэффициент связи C_{ij} .

Это осуществляется следующим образом: во-первых, эта связь сильно уменьшается на короткое время, в течение которого повторяется попытка передать управление от i-й ФС какой-либо другой ФС, во-вторых, эта связь уменьшается на небольшую величину долговременно.

В режиме тонкой доводки обучение происходит путем настройки весов синапсов в НС. Обучение в Контроллерах происходит в соответствии с поступающими подкреплениями. При этом модифицируются веса синапсов в ФС, активной в текущий момент времени, и в ФС, бывших активными несколько предыдущих тактов времени. В данный такт времени t все веса синапсов в нейронных сетях Контроллеров модифицируются следующим образом:

$$\Delta W_{ij} = \alpha_C\, \gamma^k\, X_i(t\text{-}k)\; Y_j(t\text{-}k)\; [R(t) - R(t\text{-}1)]\,, \tag{14}$$

где W_{ij} – вес рассматриваемого синапса; $X_i(t\text{-}k)$ – сигнал на входе этого синапса; $Y_j(t\text{-}k)$ – сигнал на выходе нейрона, соответствующего данному синапсу; α_C – скорость обучения Контроллера; γ – коэффициент забывания ($0 < \gamma < 1$); k – разность

между текущим моментом времени и временем срабатывания рассматриваемой ФС; [$R(t)$ - $R(t\text{-}1)$] – величина текущего изменения ресурса анимата. Так как обучение происходит и в тех Контроллерах, которые были активны в предыдущие такты времени, то эта форма обучения позволяет формировать цепочки действий.

Обучение в Модели происходит, если есть существенное рассогласование между прогнозом $\mathbf{S}^{pr}(t+1)$ и результатом $\mathbf{S}(t+1)$. Обучение осуществляется методом обратного распространения ошибки [21]. Смысл обучения Модели – уточнение прогнозов $\mathbf{S}^{pr}(t+1)$.

Схема эволюции. Эволюционирующая популяция состоит из n аниматов. Каждый анимат имеет ресурс $R(t)$, который изменяется в соответствии получаемыми подкреплениями и расходами на действия.

Эволюция происходит в течение ряда поколений, n_g = 1,2,… В начале поколения ресурс каждого анимата равен нулю.

Начальная архитектура системы управления анимата (ФС и связи между ними) и начальные веса НС формируют геном анимата **G,** который в течение его жизни не меняется. Текущая архитектура и веса синапсов НС подстраиваются в течение жизни анимата путем обучения.

В конце поколения определяется анимат, имеющий максимальный ресурс $R_{max}(n_g)$. Этот лучший анимат порождает n потомков, которые составляют новое поколение. Геномы потомков **G** отличаются от генома родителя небольшими мутациями. В процессе мутаций могут генерироваться новые ФС и удаляться старые ФС.

Выводы по проекту «Мозг анимата». Итак, предложены два варианта архитектуры системы управления аниматом. Вариант на основе отображающих нейронных сетей интересен тем, что в нем предусмотрена простая возможность изменения архитектуры: связи между ФС и сами ФС могут меняться в процессах обучения и эволюции. Это соответствует принципам метаобучения, предложенным в вычислительном интеллекте [28]. Тем не менее, необходимо согласование функционирования всех элементов системы управления аниматом и это накладывает довольно серьезные ограничения на ее архитектуру. Следовательно, важны дальнейшие исследования по данному перспективному проекту, в том числе и на основе опыта разработки архитектуры на базе нейросетевых адаптивных критиков.

Заключение

Информатика и биология – науки 21 века [29]. Направление исследований «Адаптивное поведение» – область на стыке этих наук. В этом направлении ведутся как фундаментальное изучение архитектур и принципов функционирования систем управления биологических и модельных организмов, так и биологически инспирированные прикладные работы в ряде приложений: искусственный интеллект, робототехника, модели адаптивного поведения в социально-экономических системах. Естественно ожидать, что это направление исследований будет иметь серьезные перспективы развития.

Литература

1. Meyer J.-A., Wilson S.W. (Eds.). From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior**.** The MIT Press: Cambridge, Massachusetts, London, England. 1991.
2. От моделей поведения к искусственному интеллекту. Серия «Науки об искусственном» (под ред. Редько В.Г.). М.: УРСС, 2006.
3. Nolfi S., Baldassarre G., Calabretta R., Hallam J., Marocco D., Miglino O., Meyer J-A, Parisi D. (Eds.). From Animals to Animats 9: Proceedings of the Ninth International Conference on Simulation of Adaptive Behaviour. LNAI. 2006. Volume 4095. Berlin, Germany: Springer Verlag.
4. Непомнящих В.А. Поиск общих принципов адаптивного поведения живых организмов и аниматов // Новости искусственного интеллекта. 2002. No 2. С. 48-53.
5. Donnart J.Y., Meyer J.A. Learning reactive and planning rules in a motivationally autonomous animat // IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 1996. V. 26. No 3. PP. 381-395.
6. Wilson S.W. The animat path to AI // In: [1]. PP. 15-21.
7. Цетлин М.Л. Исследования по теории автоматов и моделирование биологических систем. М.: Наука, 1969.
8. Бонгард М.М., Лосев И.С., Смирнов М.С. Проект модели организации поведения – «Животное» // Моделирование обучения и поведения. М.: Наука, 1975. С. 152-171.
9. Гаазе-Рапопорт М.Г., Поспелов Д.А. От амебы до робота: модели поведения. М.: Наука, 1987.
10. Holland J.H. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: The University of Michigan Press, 1975 (1st edn.). Boston, MA: MIT Press, 1992 (2nd edn.).
11. Holland J.H., Holyoak K.J., Nisbett R.E., Thagard P. Induction: Processes of Inference, Learning, and Discovery. Cambridge, MA: MIT Press, 1986.
12. Sutton R., Barto A. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 1998.
13. Krichmar J.L., Seth A.K., Nitz D.A., Fleischer J.G., Edelman G.M. Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions // Neuroinformatics, 2005. Vol.3. No 3. PP. 197-221.
14. Marocco D., Nolfi S. Origins of communication in evolving robots // In: [3]. PP. 789-803.

15. Непомнящих В.А., Попов Е.Е., Редько В.Г. Бионическая модель адаптивного поискового поведения // Изв. РАН. Теория и системы управления, 2008. No 1. С. 58-66.
16. Растригин Л. А. Адаптация сложных систем. Рига: Зинатне, 1981.
17. 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.
18. Red'ko V.G., Prokhorov D.V., Burtsev M.S. Theory of functional systems, adaptive critics and neural networks // International Joint Conference on Neural Networks, Budapest, 2004. PP. 1787-1792.
19. Si J., Barto A., Powell W., Wunsch D. (Eds.). Learning and Approximate Dynamic Programming: Scaling Up to the Real World, IEEE Press and John Wiley and Sons, 2004.
20. 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.
21. Rumelhart D.E., Hinton G.E., Williams R.G. Learning representation by back-propagating error // Nature. 1986. Vol. 323. No 6088. PP. 533-536.
22. 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.
23. Prokhorov D., Puskorius G., Feldkamp L. Dynamical neural networks for control // In: Kolen J. and Kremer S. (Eds.). A Field Guide to Dynamical Recurrent Networks. New York: IEEE Press, 2001. PP. 257-289.
24. 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.
25. Baldwin J.M. A new factor in evolution // American Naturalist, 1896. Vol. 30. PP. 441-451.
26. 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.
27. Red'ko V.G., Anokhin K.V. et all. Project "Animat Brain": Designing the animat control system on the basis of the functional systems theory // In: M.V. Butz, O. Sigaud, G. Pezzulo, G. Baldassarre (Eds.), Anticipatory Behavior in Adaptive Learning Systems: From Brains to Individual and Social Behavior. LNAI 4520, Berlin, Heidelberg: Springer Verlag. 2007. PP. 94-107.
28. Duch W. Towards comprehensive foundations of computational intelligence // In: Duch W., Mandziuk J. (Eds.). Challenges for Computational Intelligence, 2007. Berlin, Heidelberg: Springer Verlag. PP. 261-316.
29. Редько В.Г. Информатика и биология – науки 21-го века. Что на стыке? // Информационные процессы. 2007. Т.7. No 3. С. 214-247.

Редько Владимир Георгиевич. Заместитель директора Центра оптико-нейронных технологий НИИ системных исследований РАН. В 1971 году окончил Московский физико-технический институт. Доктор физико-математических наук. Имеет 120 печатных работ, в том числе 2 монографии. Область научных интересов: проблема происхождения интеллекта человека, эволюционная кибернетика, модели адаптивного поведения, нейроинформатика.