Мосалов О.П., Прохоров Д.В., Редько В.Г. Самообучающиеся агенты на основе нейросетевых адаптивных критиков // Искусственный интеллект. 2004, Т.3. С. 550-560.

Мосалов О.П.
Московский физико-технический институт, Россия, *olegmos_@mail.ru*

Прохоров Д.В.
Ford Research and Advanced Engineering, Ford Motor Company, Dearborn, U.S.A., *dprokhor@ford.com*

Редько В.Г.
Институт оптико-нейронных технологий РАН, Москва, Россия, *redko@iont.ru*

Самообучающиеся агенты на основе нейросетевых адаптивных критиков*

Аннотация. В статье содержится краткое введение в теорию нейросетевых адаптивных критиков и построена конкретная модель агента-брокера на основе так называемого V-критика. Проведены серии компьютерных экспериментов, которые продемонстрировали принципиальную применимость нейросетевых адаптивных критиков в финансово-экономических задачах.

Введение

Уже около 20 лет идут активные исследования в области нейронных сетей – сетей из искусственных нейроноподобных элементов, которые реализуют различные алгоритмы обработки информации, предназначенные для распознавания образов, ассоциативной памяти, кластеризации образов и т.д. [1,2,3].

Основная функция нейронных сетей в живых организмах – обеспечение управления поведением организма. И во многих случаях формирование поведения происходит путем прямого взаимодействия с внешней средой, без присутствия учителя, путем самообучения. Можно ли создать конструкции нейронных сетей, обеспечивающих управление поведением в отсутствии учителя? Сравнительно недавно, во второй половине 1990-х годов такие конструкции – так называемые нейросетевые адаптивные критики – были разработаны, и в настоящее время ведется активное их исследование. Функционирование адаптивных критиков основано на хорошо известном методе обучения с подкреплением [4]. Данная работа посвящена анализу применения адаптивных критиков к задачам формирования принятия решений агентом-брокером.

1. Обучение с подкреплением и адаптивные критики

* Работа выполнена при финансовой поддержке РФФИ (проект № 04-01-00179) и ОИТВС РАН.

Адаптивные критики – это схемы управления, которые содержат специальный блок – Критик, оценивающий качество работы всей системы управления.

Адаптивные критики разработаны и исследованы в работах Бернарда Видроу [5], Ричарда Саттона, Эндрью Барто [4,6], Пола Вербоса [7], Данила Прохорова, Дональда Вюнша [8,9]. Существует целое семейство различных конструкций адаптивных критиков (Adaptive Critic Designs) [8].

Основные схемы обучения адаптивных критиков основаны на методе обучения с подкреплением (Reinforcement Learning) [4]. В этом методе рассматривается агент (модельный организм), взаимодействующий с внешней средой (рис. 1). В текущей ситуации $\mathbf{S}(t)$ агент выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию $\mathbf{S}(t+1)$. Здесь и далее время предполагается дискретным: $t = 1,2,\ldots$ Подкрепление $r(t)$ может быть положительным (награда) или отрицательным (наказание).

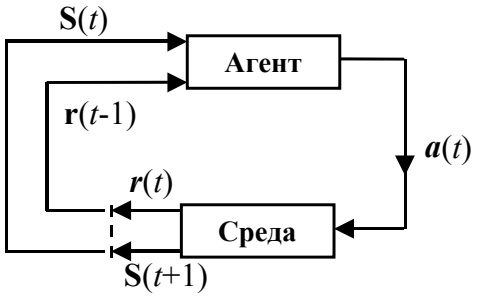

Рис.1. Схема обучения с подкреплением.

Цель агента – максимизировать суммарную награду, которую можно получить в будущем в течение длительного периода времени. Агент оценивает суммарную награду с учетом коэффициента забывания:

$$U(t) = \sum_{k=0}^{\infty} \gamma^k \, r(t+k), \qquad (1)$$

где $U(t)$ – оценка суммарной награды, γ – коэффициент забывания, $0 < \gamma < 1$. Коэффициент забывания учитывает, что чем дальше агент «заглядывает» в будущее, тем меньше у него уверенность в оценке награды («рубль сегодня стоит больше, чем рубль завтра»).

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 агент, если в ситуации $\mathbf{S}(t)$ он выполнит действие $a(t)$. Математическое ожидание награды равно:

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

Из (1) и (2) следует $Q(\mathbf{S}(t), a(t)) = E\,[r(t) + \gamma\, Q(\mathbf{S}(t+1), a(t+1))]$. Ошибку естественно определить так:

$$\delta(t) = r(t) + \gamma\, Q(\mathbf{S}(t+1), a(t+1)) - Q(\mathbf{S}(t), a(t)). \qquad (3)$$

Величина $\delta(t)$ называется ошибкой временной разности.

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

- в момент t с вероятностью 1 - ε выбирается действие с максимальным значением $Q(\mathbf{S}(t), a_j)$: $a(t) = \arg\max_a \{Q(\mathbf{S}(t), a_j)\}$
- с вероятностью ε выбирается произвольное действие, $0 < \varepsilon << 1$.

Такой выбор действия называют «ε–жадным правилом».

Обучение, т.е. переоценка величин $Q(\mathbf{S}, a)$ происходит в соответствии с оценкой ошибки $\delta(t)$ – к величине $Q(\mathbf{S}(t), a(t))$ добавляется величина, пропорциональная ошибке временной разности $\delta(t)$:

$$\Delta Q(\mathbf{S}(t), a(t)) = \alpha\, \delta(t) = \alpha\, [r(t) + \gamma\, Q(\mathbf{S}(t+1), a(t+1)) - Q(\mathbf{S}(t), a(t))], \qquad (4)$$

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 итераций, причем оценки эффективности тех или иных решений, соответствующие предыдущим шагам процесса, переоцениваются с учетом знаний о возможных будущих шагах. Обучение с подкреплением, адаптивные критики и подобные методы часто называют приближенным динамическим программированием [10].

Конструкции адаптивных критиков можно рассматривать как развитие моделей обучения с подкреплением на тот случай, когда ситуации (и, возможно, действия) задаются векторами и изложенная выше схема итеративного формирования матрицы $Q(\mathbf{S}_i, a_j)$ не работает. В этом случае характеристики системы управления целесообразно представить с помощью параметрически задаваемых аппроксимирующих функций (например, с помощью искусственных нейронных сетей), а обучение проводить путем итеративной оптимизации параметров. В случае аппроксимации с помощью нейронных сетей, параметрами аппроксимирующих функций являются веса синапсов нейросети, оптимизация производится путем подстройки весов, например, аналогично тому, как это делается в методе обратного распространения ошибки.

Различают схемы Q-критиков и V-критиков [11]. В схемах Q-критиков блок Критик делает оценку величины суммарной награды $Q(\mathbf{S}(t), a(t))$, которую агент ожидает получить в будущем, если он в данной ситуации $\mathbf{S}(t)$ выполнит определенное действие $a(t)$. Т.е. происходит оценка качества того или иного действия в известной ситуации (аналогично формированию таких оценок в методе SARSA).

В схемах V-критиков блок Критик делает оценку качества ситуации $V(\mathbf{S}(t))$, т.е. оценку ожидаемой величины суммарной награды в этой ситуации. В этом случае схема управления дополняется блоком прогноза, и система управления стремится выбирать те действия, которые, согласно прогнозу, приведут к ситуациям $\mathbf{S}(t+1)$ с наибольшими оценками $V(\mathbf{S}(t+1))$.

Существуют и более сложные (и часто более эффективные практически) схемы критиков, основанные на оценках производных функции критерия качества по переменным состояния системы «среда-агент» [8,9].

Подчеркнем, что обучение нейросетевых адаптивных критиков является, на самом деле, самообучением: нет учителя, который говорит, какое действие нужно выполнить в той или иной ситуации. Напротив, обучение происходит путем самостоятельного взаимодействия с внешней средой, при котором агент получает только поощрение или наказание. Отметим, что существуют нейросетевые схемы самообучения и без использования критиков, см. [9,12].

В следующем разделе построена и исследована простая модель агента-брокера на основе V-критика.

2. Модель агента-брокера на основе V-критика

2.1. Описание модели

Общие предположения модели состоят в следующем:

1. Есть агент, который располагает некоторым количеством ресурсов двух типов: виртуальными деньгами и акциями. Сумма этих ресурсов составляет общий капитал агента $C(t)$. Состояние агента характеризуется переменной $u(t)$ – доля акций в общем капитале агента.
2. Внешняя среда определяется временным рядом $X(t)$, $t = 0,1,2,...$, где $X(t)$ – курс акций на бирже в момент времени t .
3. Агент стремится увеличить свой капитал $C(t)$, изменяя значение $u(t)$.
4. Система управления агента содержит блок Модель, который служит для прогнозирования изменения курса акций $\Delta X(t+1) = X(t+1) - X(t)$ для следующего такта времени.
5. Система управления содержит блок Критик, который оценивает качество ситуации $V(\mathbf{S}(t))$. Ситуация $\mathbf{S}(t)$ задается вектором $\{\Delta X(t), u(t)\}$; $\Delta X(t) = X(t) - X(t-1)$.
6. Система управления содержит ε–жадное правило, которое используется для выбора одного из возможных двух действий:

а) $u(t+1) = 0$ – перевести весь капитал в деньги,

б) $u(t+1) = 1$ – перевести весь капитал в деньги.

Общая схема системы управления агента представлена на рис. 2.

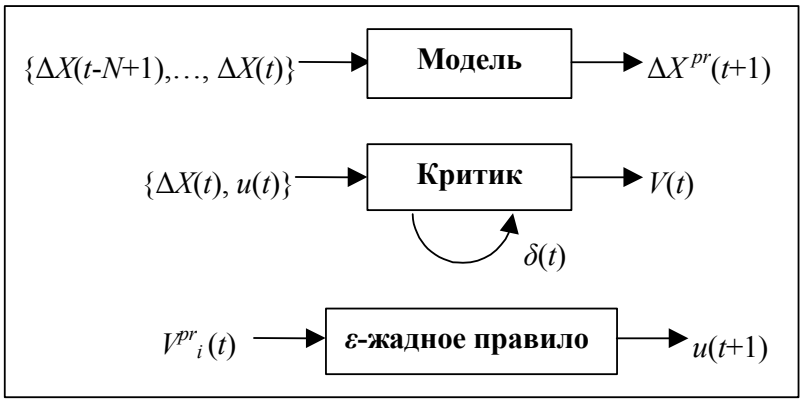

Рис. 2. Схема управления агента на основе V-критика.

Блок Модель предназначен для прогнозирования изменения курса временного ряда. На вход этого блока подается N предыдущих значений изменения курса $\Delta X(t\text{-}N\text{+}1), \ldots, \Delta X(t)$, на выходе формируется прогноз изменения курса в следующий такт времени $\Delta X^{Pr}(t\text{+}1)$.

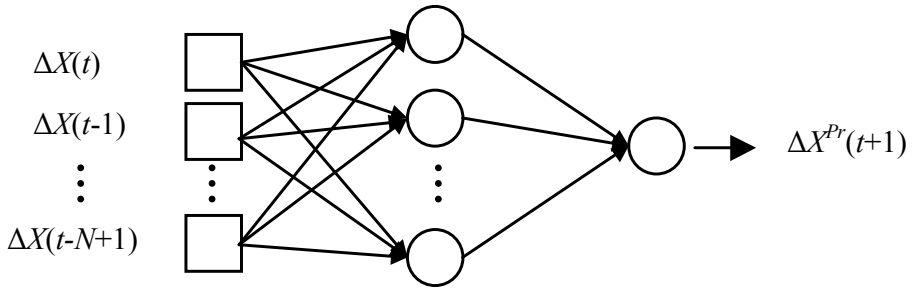

Рис. 3. Нейронная сеть блока Модель.

Блок Модель представляет собой двухслойную нейронную сеть (рис. 3), работа которой описывается формулами:

$$\mathbf{x^M} = \{\Delta X(t\text{-}N\text{+}1), \ldots, \Delta X(t)\},$$
$$y^M_j = \text{th}\,(\textstyle\sum_i W^M_{ij}\, x^M_i),$$
$$\Delta X^{pr}(t\text{+}1) = \textstyle\sum_j V^M_j\, y^M_j. \qquad (5)$$

где $\mathbf{x^M}$ – входной вектор, $\mathbf{y^M}$ – вектор выходов нейронов скрытого слоя, W^M_{ij} и V^M_j – веса синапсов нейронной сети, $\Delta X^{pr}(t\text{+}1)$ – значение на выходе сети, которое трактуется как прогноз изменения курса временного ряда на момент времени t+1, $\Delta X(t) = X(t) - X(t\text{-}1)$.

Модель обучается обычным методом обратного распространения ошибки, т.е. производится минимизация функционала ошибки:

$$E = 0.5\,(\Delta X^{pr}(t\text{+}1) - \Delta X(t\text{+}1))^2. \qquad (6)$$

$$V^M_j(t\text{+}1) = V^M_j(t) - \alpha^M\, \partial E / \partial V^M_j, \qquad (7.1)$$
$$W^M_{ij}(t\text{+}1) = W^M_{ij}(t) - \alpha^M\, \partial E / \partial W^M_{ij}\,, \qquad (7.2)$$

где α^M – параметр скорости обучения Модели.

Из (5) и (6) следует, что

$$\partial E / \partial V^M_j = (\Delta X^{pr}(t\text{+}1) - \Delta X(t\text{+}1))\, y^M_j\,, \qquad (8.1)$$
$$\partial E / \partial W^M_{ij} = (\Delta X^{pr}(t\text{+}1) - \Delta X(t\text{+}1))\, V^M_j\, (1 - (y^M_j)^2)\, x_i\,. \qquad (8.2)$$

Формулы (7), (8) определяют изменения весов синапсов в процессе обучения нейронной сети Модели.

Блок Критик предназначен для оценки $V(\mathbf{S})$ суммарной награды $U(t) = \Sigma_k \gamma^k r(t+k)$ для текущей и прогнозируемой ситуаций.

Блок Критик также представляет собой двухслойную нейронную сеть (аналогичную представленной на рис. 3), работа которой описывается формулами:

$$\mathbf{x}^{\mathbf{C}} = \mathbf{S}(t) = \{\Delta X(t), u(t)\},$$
$$y^C_j = \text{th}\,(\textstyle\sum_i W^C_{ij}\, x^C_i),$$
$$V(t) = V(\mathbf{S}(t)) = \textstyle\sum_j V^C_j\, y^C_j. \qquad (9)$$

где $\mathbf{x}^{\mathbf{C}}$ – входной вектор, $\mathbf{y}^{\mathbf{C}}$ – вектор выходов нейронов скрытого слоя, W^C_{ij} и V^C_j – синапсы нейронной сети, $V(t)$ – значение на выходе сети, которое трактуется как оценка качества данной ситуации $\mathbf{S}(t)$.

При подаче на вход Критика вектора $\mathbf{S}^{\mathbf{pr}}_i(t+1) = \{\Delta X^{pr}(t+1), u_i(t+1)\}$ на выходе формируется оценка ожидаемой суммарной награды V^{pr}_i $(t+1)$ прогнозируемой ситуации $\mathbf{S}^{\mathbf{pr}}_i(t+1)$ для каждого из возможных действий. При этом полагаем $u_1(t+1) = 0$, $u_2(t+1) = 1$ (первое действие соответствует переводу всего капитала в деньги, второе – в акции, см. также ниже пункт «Выбор действия»).

Критик обучается с помощью временной разности

$$\delta(t) = r(t) + \gamma\, V(t) - V(t\text{-}1), \qquad (10)$$

где $V(t)$ и $V(t\text{-}1)$ – оценка суммарной награды для ситуаций $\mathbf{S}(t)$ и $\mathbf{S}(t\text{-}1)$, соответственно, $r(t)$ – изменение логарифма суммарного ресурса (для удобства переходим к логарифмической шкале оценки суммарного ресурса агента, $R(t) = \log C(t)$ [12]):

$$r(t) = \log C(t) - \log C(t\text{-}1), \qquad (11)$$

в соответствии с формулами:

$$V^C_j(t+1) = V^C_j(t) + \alpha^C\, \delta(t)\, \partial V / \partial V^C_j\,, \qquad (12.1)$$
$$W^C_{ij}(t+1) = W^C_{ij}(t) + \alpha^C\, \delta(t)\, \partial V / \partial W^C_{ij}\,. \qquad (12.2)$$

α^C – параметр скорости обучения Критика. Смысл формул (12) состоит в том, веса нейронной сети подстраиваются так, чтобы минимизировать ошибку временной разности $\delta(t)$.

Из (9) следует, что

$$\partial V / \partial V^C_i = y^C_j, \qquad (13.1)$$
$$\partial V / \partial W^C_{ij} = V^C_j\, (1 - (y^C_j)^2)\, x^C_i. \qquad (13.2)$$

Формулы (12), (13) определяют изменения весов синапсов в процессе обучения нейронной сети Критика.

Выбор действия. В каждый момент времени может быть выбрано одно из двух действий: а) перевести капитал в наличные деньги ($u_1(t+1) = 0$), б) перевести капитал в акции ($u_2(t+1) = 1$).

Выбор действия $u(t+1)$ на текущем такте осуществляется следующим образом. Для каждого из возможных действий Критик дает оценку качества $V^{pr}_i(t+1) = V(\mathbf{S}^{pr}_i(t+1))$ ожидаемой ситуации $\mathbf{S}^{pr}_i(t+1)$. При этом на вход Критика наряду с полученным с выхода Модели значением $\Delta X^{pr}(t+1)$ подается соответствующее действию значение $u_i(t+1)$, см. первую формулу в (9). Далее применяется ε-жадное правило: с вероятностью 1 - ε выбирается то действие, которому соответствует максимальное значение $V^{pr}_i(t+1)$, с вероятностью ε выбирается произвольное действие ($0 < \varepsilon << 1$).

Динамика изменения ресурса. Пусть доля акций в суммарном капитале на предыдущем такте равна $u(t-1)$, а на текущем – $u(t)$, а суммарный капитал на предыдущем такте равен $C(t-1)$. Тогда суммарный капитал на текущем такте определяется следующим образом:

$$C(t) = C(t-1)\ \{1 + u(t)\ \Delta X(t) / X(t-1)\}\ [1 - J\ |u(t) - u(t-1)|]. \qquad (14)$$

Множитель в фигурных скобках в (14) соответствует изменению капитала в результате роста/падения курса акций; множитель в квадратных скобках – затратам агента на покупку/продажу акций.

Далее для ресурса используется логарифмическая шкала: $R(t) = \log C(t)$.

2.2. Результаты моделирования

Изложенная модель была реализована в виде компьютерной программы на языке Java и исследовалась путем численного моделирования. Расчеты проводились для трех вариантов входного ряда $X(t)$: двух модельных – «пилы» ($X(2k+1) = 1,\ X(2k+2) = 2,\ k = 0, 1, \ldots$) и синусоиды $X(t) = 0.5\ (1 + \sin(2\pi\ t/20))$, а также для реальных финансовых данных.

Реальные финансовые данные $X(t)$ представляли собой отношение курса доллара США к швейцарскому франку (цены закрытия пятиминутных интервалов, котировки рынка Forex, 1998-2001 гг.), усредненные по окну в 48 отсчетов. Мы рассматриваем такой ряд $X(t)$ как модельный курс акций на бирже.

При моделировании параметры основного (опорного) варианта расчета составляли: число входов Модели $N = 10$, число нейронов скрытого слоя нейронной сети Модели и нейронной сети Критика $N_h^M = N_h^C = 20$, коэффициент затрат на конвертирование $J = 0.0$, коэффициенты обучения Модели и Критика $\alpha^M = \alpha^C = 0.01$, параметр ε-жадной политики $\varepsilon = 0.1$, величина коэффициента забывания $\gamma = 0.9$. Такой набор параметров рассматривался как опорный, анализировалось также влияние изменения некоторых параметров относительно опорного варианта на функционирование агента (см. ниже).

Пример результатов моделирования для отдельного агента для реальных финансовых данных представлен на рис. 4.

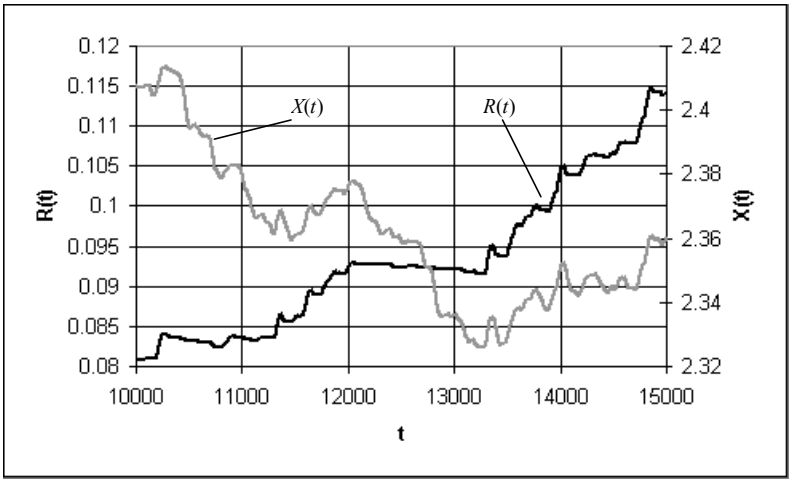

Рис. 4. Зависимость ресурса агента $R(t)$ и курса акций $X(t)$ от времени, $t \in [10000, 15000]$.

Видно, что агент находит естественную стратегию – переводить капитал в акции при росте курса и переводить капитал в наличные деньги при падении курса.

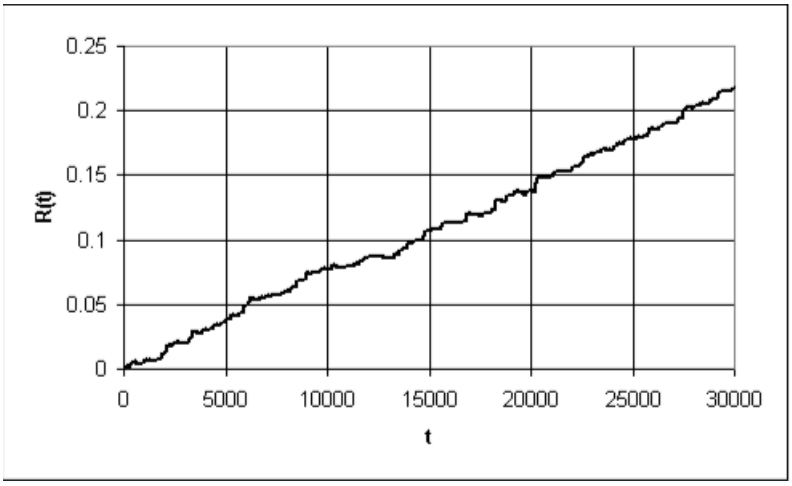

Рис. 5. Усредненная зависимость ресурса агентов от времени $R(t)$.

Усредненная по 100 агентам зависимость $R(t)$ показана на рис. 5. Рис. 5 демонстрирует, что найденная стратегия обеспечивает стабильный рост ресурса агента.

Был проведен анализ влияния изменения наиболее критических параметров на работу V-критика относительно опорного варианта. При введении эффективных затрат на конвертирование денег и акций скорость возрастания ресурса уменьшалась. Например, при $J = 10^{-5}$ конечное значение ресурса, полученное агентом после 30000 тактов времени, составляло 0.17 вместо 0.22 для $J = 0$. При упрощении нейронной сети блока Критик (при $N_h^C = 10$) его работа существенно ухудшается, в результате чего ресурс агента растет медленнее. При упрощении же нейронной сети блока Модель ($N_h^M = 10$) получаются зависимости $R(t)$, практически совпадающие с представленной на рис. 5. Последнее можно проинтерпретировать следующим образом. Нейронная сеть Модели обучается

формировать отображение определенной зависимости, которая задается извне, а Критик должен сам найти заранее неизвестную стратегию поведения агента. Т.е. задача, которую решает Критик, существенно сложнее задачи, которая стоит перед Моделью, поэтому уменьшение числа нейронов в сети Критика более критично по сравнению с нейронной сетью Модели.

Для проверки эффективности работы процедуры обучения V-критика было проведено ее сравнение с работой метода SARSA (см. раздел 1). В качестве ряда, задающего курс акций, была взята синусоида $X(t) = 0.5\,(1 + \sin(2\pi\, t/20))$. При расчете при $t = 10000$ вероятность выбора случайного действия уменьшалась до нуля, т.е., в ε-жадном правиле полагалось $\varepsilon = 0.1$ при $t < 10000$ и $\varepsilon = 0$ при $t > 10000$.

В данном случае для метода SARSA рассматривались две возможных ситуации: $\Delta X(t) > 0$ и $\Delta X(t) < 0$ и два возможных действия: $u(t+1) = 0$ и $u(t+1) = 1$. Таким образом, матрица Q имеет размерность 2x2, а значения ее элементов определяют то, насколько выгодно в данной ситуации принять то или иное решение.

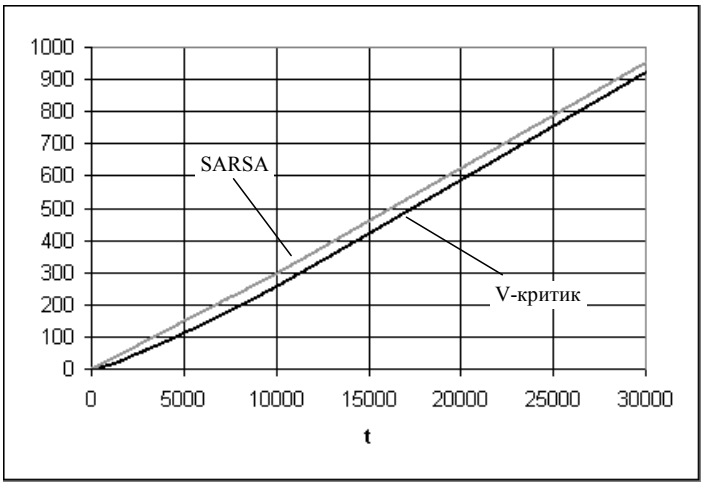

Рис. 6. Ресурс $R(t)$, полученный при работе V-критика и при работе метода SARSA, $t \in [0, 30000]$.

Полученные результаты таковы: V-критик обучается медленнее SARSA, в результате чего в начале (при $t \in [0, 30000]$) ресурс, получаемый с помощью SARSA, больше (рис. 6). Затем, обучившись, V-критик начинает работать эффективнее и при $t \in [100000, 130000]$ полученный с его помощью ресурс уже больше, чем у SARSA (рис. 7).

То, что V-критик работает эффективнее метода SARSA, связано с тем, что V-критик может использовать прогноз, формируемый блоком Модель, для принятия решения, а в методе же SARSA ситуации жестко определены (задаются только знаком изменения курса акций $\Delta X(t)$ на данном такте). На рис. 8 показано изменение подкрепления $r(t)$ за период синусоиды (20 тактов) для V-критика (черная кривая) и для метода SARSA (серая кривая). Видно, что V-критик предвидит начало падения курса, успевает перевести акции в деньги и за счет этого получает большее суммарное подкрепление, чем метод SARSA.

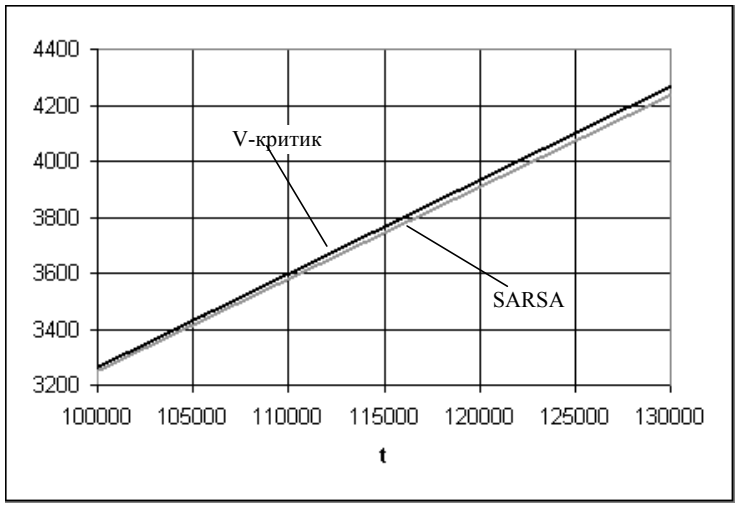

Рис. 7. Ресурс $R(t)$, полученный при работе V-критика и при работе метода SARSA, $t \in [100000, 130000]$.

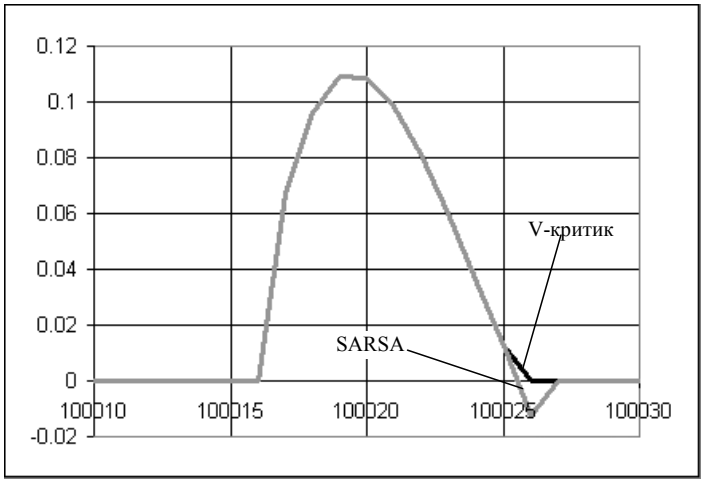

Рис. 8. Преимущество работы V-критика над работой метода SARSA. Представлена зависимость величины подкрепления $r(t)$ от времени. Видно, что прогноз, который делает V-критик, позволяет ему своевременно (в момент времени $t = 100025$) перевести акции в деньги.

В то же время, понятно, что V-критик мог бы предвидеть не только начало падения, но и начало роста курса акций. Рассмотрим такой алгоритм (оптимальный для синусоиды при $J = 0$): переводить капитал в акции, когда прогнозируемое изменение курса положительно и переводить капитал в деньги, когда прогнозируемое изменение курса отрицательно. На рис. 9 показано изменение подкрепления $r(t)$ за период синусоиды (20 тактов) для V-критика (черная кривая) и для оптимального алгоритма (серая кривая).

Таким образом, из двух возможных улучшений по сравнению с методом SARSA, которые V-критик мог бы в принципе найти для рассматриваемого модельного ряда (в начале и в конце роста курса акций), в нашем расчете V-критик находит только одно. Этот факт можно проинтерпретировать

следующим образом. Самостоятельное обучение путем стохастического поиска с подкреплением (которое и осуществляет V-критик) имеет и свои недостатки: сложно найти решение всех возможностей сразу с помощью одной простой конструкции.

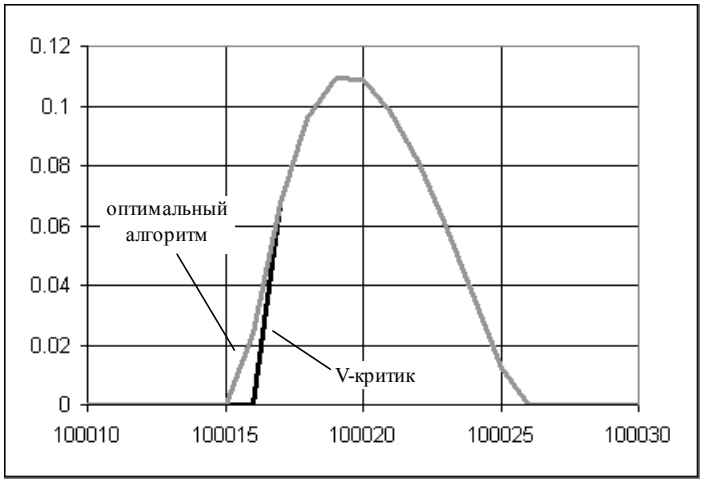

Рис. 9. Недостаток работы V-критика по сравнению с оптимальной стратегией. Представлена зависимость $r(t)$ для V-критика и оптимальной стратегии.

Заключение

Итак, продемонстрировано, что агенты, основанные на простых схемах нейросетевых адаптивных критиков, способны самообучаться и находить естественную стратегию в рассмотренных случаях.

Разработанные модели адаптивных критиков являются достаточно простыми и универсальными и могут быть положены в основу разработок разнообразных систем адаптивного управления и принятия решения.

Есть дальнейшие перспективы развития моделей на основе адаптивных критиков за счет включения в схемы адаптивных критиков рекуррентных нейронных сетей, модели желаемого поведения, нейронной сети формирования действий (так называемого Контроллера) и т.п. [8,9,12].

Авторы благодарны Н.А. Митину за предоставление данных используемого финансового ряда.

Литература

1. Уоссермен Ф. *Нейрокомпьютерная техника. Теория и практика*. М.: Мир, 1992. 238 с.

2. Потапов А.Б., Али М.К. Нелинейная динамика обработки информации в нейронных сетях // *Новое в синергетике: Взгляд в третье тысячелетие* (Под ред. Г.Г. Малинецкого и С.П. Курдюмова). М.: Наука, 2002. С. 367-426.

3. Rumelhart D.E., Hinton G.E., Williams R.G. Learning representation by back-propagating error // *Nature*. 1986. Vol.323, N.6088, pp. 533-536.

4. Sutton R., Barto A. *Reinforcement Learning: An Introduction.* – Cambridge: MIT Press, 1998.

5. Widrow B., Gupta N., Maitra S. Punish/Reward: Learning with a Critic in Adaptive Threshold Systems // *IEEE Transactions on Systems, Man and Cybernetics,* 1973. Vol. 3, no.5, pp. 455-465.
6. Barto A.G., Sutton R.S., Anderson C.W. Neuronlike elements that can solve difficult learning control problems // *IEEE Transactions on Systems, Man, and Cybernetics*, 1983. Vol. 13, pp. 835-846.
7. Werbos P.J. Approximate dynamic programming for real-time control and neural modeling // In: *Handbook of Intelligent Contro*l, White and Sofge, Eds., Van Nostrand Reinhold, 1992. pp. 493 – 525.
8. Prokhorov D.V., Wunsch D. Adaptive critic designs // *IEEE Trans. Neural Networks*, 1997. Vol. 8. No.5, pp.997-1007.
9. Prokhorov D.V. Backpropagation through time and derivative adaptive critics: a common framework for comparison // In J. Si et al. (Eds.), *Learning and Approximate Dynamic Programming*, IEEE Press, 2004 (in press).
10. Workshop "Learning and Approximate Dynamic Programming" (Mexico, April, 2002): http://ebrains.la.asu.edu/~nsfadp/
11. 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. См. также: http://wsni2003.narod.ru/RFFI/rvgpdv.pdf
12. Prokhorov D.V., Puskorius G., Feldkamp L. Dynamical Neural Networks for Control // In: J. Kolen and S. Kremer (Eds.) *A Field Guide to Dynamic Recurrent Networks*, IEEE Press, 2001.

Mosalov O.P.
Moscow Institute of Physics and Technology, Russia, *olegmos_@mail.ru*

Prokhorov D.V.
Ford Research and Advanced Engineering, Ford Motor Company, Dearborn, U.S.A., *dprokhor@ford.com*

Red'ko V.G.
Institute of Optical Neural Technologies, Russian Academy of Science, *redko@iont.ru*

Self-learning agents on the base of adaptive critic designs

Abstract. The paper includes a short survey of neural adaptive critic designs and description of the original model of agent-broker based on V-critic scheme. The results of computer simulations of the agent-broker model are described. The simulations demonstrated the applicability of neural adaptive critic designs for solving certain financial problems.