Мосалов О.П., Прохоров Д.В., Редько В.Г. Модели принятия решений на основе нейросетевых адаптивных критиков // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х т. М.: Физматлит, 2004. т.3. С. 1156-1163.

МОДЕЛИ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ НЕЙРОСЕТЕВЫХ АДАПТИВНЫХ КРИТИКОВ*

Мосалов О.П.[1], Прохоров Д.В.[2], Редько В.Г.[3]

Аннотация

Кратко характеризуются нейросетевые адаптивные критики – схемы управления автономных агентов, обеспечивающие принятие решений о тех или иных действиях агента. Работа адаптивных критиков иллюстрируется простым примером Q-критика, моделирующего функционирование агента-брокера, виртуально играющего на бирже. Приводятся результаты компьютерного исследования этой модели Q-критика.

Что такое адаптивные критики

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

Адаптивные критики впервые упомянуты Бернардом Видроу в 1973 году. Он и его коллеги впервые применили понятие "критик" к простой карточной игре и показали, что обучение с критиком позволяет найти оптимальную стратегию игры путём проб и ошибок, без использования учителя. Дальнейшее развитие адаптивные критики получили в работах Ричарда Саттона, Эндрью Барто и особенно Пола Вербоса. Существует целое семейство различных конструкций адаптивных критиков (Adaptive Critic Designs) [Prokhorov et al, 1997].

Основные схемы обучения адаптивных критиков основаны на методе обучения с подкреплением (Reinforcement Learning) [Sutton et.al, 1998]. А именно, рассматривается агент (модельный организм), взаимодействующий с внешней средой (рис.1). В текущей ситуации агент $\mathbf{S}(t)$ выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую

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

[1] Институт оптико-нейронных технологий РАН, г. Москва, Московский физико-технический институт, olegmos_@mail.ru

[2] Ford Research and Advanced Engineering, Ford Motor Company, Dearborn, U.S.A., dprokhor@ford.com

[3] Институт оптико-нейронных технологий РАН, г. Москва, redko@iont.ru

ситуацию $\mathbf{S}(t+1)$ (здесь и далее время предполагается дискретным: $t = 1,2,\ldots$). Подкрепление может быть положительным (награда) или отрицательным (наказание).

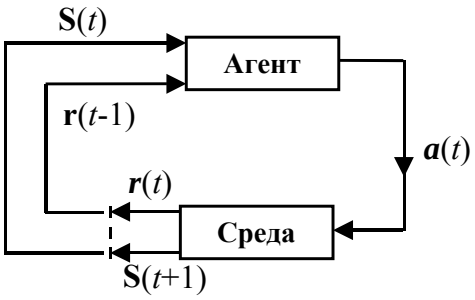

Рис.1. Схема обучения с подкреплением.

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 $\mathbf{S}(t)$ -> $a(t)$ -> $r(t)$ -> $\mathbf{S}(t+1)$ -> $a(t+1)$.

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 $a(t)$. Так как число ситуаций $\{\mathbf{S}_i\}$ и действий $\{a_j\}$ конечно, то в результате обучения формируется матрица $Q(\mathbf{S}_i,a_j)$. Математическое ожидание награды равно:

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

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:

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

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 $1 - \varepsilon$ 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 –жадной" политикой.

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 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 [Workshop, 2002].

Конструкции адаптивных критиков можно рассматривать как развитие моделей обучения с подкреплением на 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 компоненты системы управления целесообразно представить с помощью параметрически задаваемых аппроксимирующих функций (например, с помощью искусственных нейронных сетей), а обучение проводить путем итеративной настройки параметров. В случае аппроксимации с помощью нейронных сетей параметрами аппроксимирующих функций являются веса синапсов нейросети, а обучение производится путем подстройки весов, например, аналогично тому, как это делается в методе обратного распространения ошибки.

Различают схемы Q-критиков и V-критиков [Редько и др., 2004]. 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, которую агент ожидает получить в будущем, если в данный момент он находится ситуации $\mathbf{S}(t)$. В этом случае схема управления дополняется блоком прогноза, и система управления стремиться 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" [Prokhorov et al, 1997].

Модель агента-брокера

Здесь мы рассматриваем модель агента-брокера, который может принимать решения о покупке-продаже акций, играя на бирже. Модель является развитием предыдущих версий нейросетевых моделей агента-брокера [Мосалов и др., 2003, Мосалов, 2004].

Общие предположения модели состоят в следующем:

1) Есть агент, который располагает некоторым количеством ресурсов двух типов: деньги M и некоторое число акций N_A.

2) Внешняя среда определяется временным рядом $X(t)$, $t = 0,1,2,...$; $X(t)$ – стоимость одной акции в момент времени t (строим модель в дискретном времени).

3) Продавая и покупая акции, агент стремится увеличить свой суммарный ресурс

$$R(t) = M(t) + N_A(t)\,X(t). \qquad (5)$$

4) Система управления агента основана на простой версии Q-критика. Q-критик используется при выборе одного из двух возможных действий: а) покупка одной акции, б) продажа одной акции.

5) Ресурс агента меняется в соответствии с изменением количества и стоимости акций. Изменение суммарного ресурса, которое используется как подкрепление $r(t)$ (см. рис.1) в процедуре обучения Q-критика, при переходе от такта времени t к такту t+1 равно:

$$r(t) = \Delta R(t) = N_A(t)\,[X(t+1) - X(t)]\,. \qquad (6)$$

Схема управления агента

Предполагаем, что принятие решения осуществляется с помощью Q-критика (рис. 2) На вход Критика поступают два типа сигналов: 1) сигналы, характеризующие текущую ситуацию $\mathbf{S}(t)$, и 2) сигнал, характеризующий одно из возможных действий a_j (в нашем случае есть только два действия: "покупать" либо "продавать", j =1,2). По этим сигналам Критик делает оценку $Q(\mathbf{S}(t),a_j)$ суммарной награды $U(t) = \Sigma_k \gamma^k r(t+k)$, ожидаемой в будущем для данной ситуации $\mathbf{S}(t)$ для каждого из возможных действий a_j . На основе этих оценок $Q(\mathbf{S}(t),a_j)$ Критик выбирает текущее действие $a(t)$, используя ε-жадное правило.

Обучение Критика происходит методом временной разности, ошибка временной разности $\delta(t)$ определяется выражением (3).

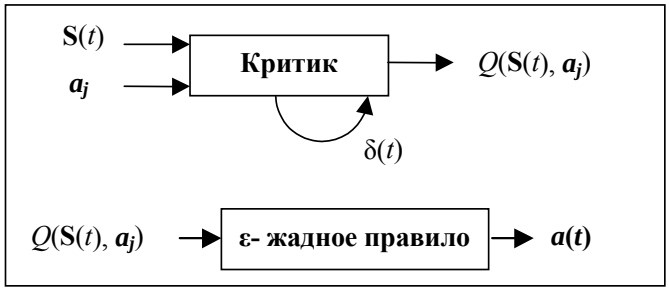

Рис. 2. Схема управления агента на основе Q-критика. Пояснения – в тексте.

В нашей модели оценки $Q(\mathbf{S}(t), a_j)$ вычисляются с помощью нейронной сети (рис. 3). На вход нейронной сети подаются компоненты вектора $\mathbf{S}(t)$ и сигналы, характеризующие действия a_j . Считаем, что в вектор $\mathbf{S}(t)$ входят: а) изменение курса акций $\Delta X(t) = X(t) - X(t\text{-}1)$, б) текущее количество акций агента $N_A(t)$. Сигналы a_j определим как a_1 = -1 для действия «продавать», a_2 = +1 для действия «покупать».

Работа нейронной сети определяется следующими выражениями:

$$\mathbf{x} = \{\mathbf{S}(t), a_j\}, y_k = \text{th}\,(\sum_i W_{ik} x_i), Q = \sum_k V_k y_k\,, \qquad (7)$$

где $\mathbf{x}$ – вход нейронной сети, y_k – выходы нейронов на скрытом слое, W_{ik} – веса нейронов скрытого слоя, V_k – веса выходного нейрона.

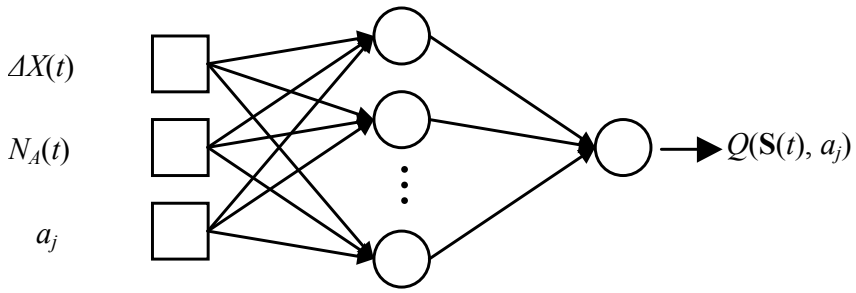

Рис. 3. Нейронная сеть агента.

Для ограничения области возможных ситуаций предполагаем, что число акций агента ограничено: $0 \le N_A(t) \le maxN_A$. В случае если число акций агента $N_A(t)$ становится больше $maxN_A$ или меньше нуля, то оно устанавливается равным случайно выбранному значению из интервала $[0, maxN_A]$. При этом меняется и количество денег $M(t)$ таким образом, что суммарный ресурс агента $R(t)$ остается неизменным.

Выбор действия $a(t)$ на текущем такте осуществляется следующим образом. Для каждого из возможных действий осуществляется работа нейронной сети и вычисляются значения $Q(\mathbf{S},a_1)$ и $Q(\mathbf{S},a_2)$. Далее с вероятностью $1-\varepsilon$ выбирается то действие, которому соответствует максимальное значение Q, с вероятностью ε выбирается произвольное действие ($0< \varepsilon << 1$).

Схема обучения агента

В процессе обучения происходит уточнение оценок $Q(\mathbf{S}(t),a_j)$. Обучение нейронной сети производится методом градиентного спуска. Веса синапсов нейронной сети скрытого и выходного слоев изменяются на каждом такте пропорционально величине ошибки временной разности:

$$\Delta W_{ik} = \alpha\, \delta(t)\, \text{grad}_W Q(\mathbf{S}(t), a(t))\,, \qquad (8)$$

$$\Delta V_j = \alpha\, \delta(t)\, \text{grad}_V Q(\mathbf{S}(t), a(t))\,, \qquad (9)$$

где α – параметр скорости обучения, $\delta(t)$ определяется выражением (3).

Используя (7), легко определить частные производные Q по весам синапсов нейронной сети. При этом формулы (8), (9) принимают вид:

$$\Delta W_{ik} = \alpha\, \delta(t)\, x_i\, (1 - y_k^2)\, V_k\,, \qquad (10)$$

$$\Delta V_k = \alpha\, \delta(t)\, y_k\,. \qquad (11)$$

Результаты моделирования

Модель была реализована в виде компьютерной программы. Моделирование проводилось как на модельных рядах: а) "пила": $X(2m) = 1$, $X(2m+1) = 2$, б) синусоида: $X(k) = 0.5[\sin(2\pi m/T)+1]$ ($m = 0,1,2, \ldots$; $T = 20$), так и на реальных биржевых данных. Для всех компьютерных экспериментов исходные веса нейронной сети задавались случайно, и анализировался процесс обучения Q-критика. Для случая синусоиды и реальных данных в вектор ситуации $\mathbf{S}(t)$ кроме значений $\Delta X(t) = X(t) - X(t\text{-}1)$, $N_A(t)$, дополнительно могли входить $\Delta X(t–1)$ и разность $\Delta X(t) - \Delta X(t–1)$. Типичные параметры расчетов составляли: $\alpha = 0.01$, $\varepsilon = 0.1$, $\max N_A = 5$, число нейронов в скрытом слое равно 6.

Результаты моделирования состоят в следующем. Для очень простой модельной среды – "пила" (для которой пространство возможных ситуаций ограничено) – Q-критик успешно обучается за примерно 5000 тактов времени. При этом в результате обучения агент совершает действие "покупать" при низком курсе акций и действие "покупать" при высоком курсе акций. В более сложной среде – синусоида и реальные биржевые данные – аналогичное обучение происходит, но за значительно большее время – порядка 100000 тактов.

В заключение наметим пути дальнейшей работы над моделью. Основная причина недостаточно эффективного обучения Q-критика в нашей модели обусловлена некоторой неадекватностью восприятия ситуации нейронной сетью при подаче на вход нейронной сети переменной "число акций N_A". Для более корректного восприятия ситуаций целесообразно перейти от переменной "число акций", к переменной "доля капитала в акциях" и к соответствующему изменению схемы покупка-продажа, аналогично тому, как это сделано в [Prokhorov et al, 2001]. Кроме того, представляет интерес переход от схемы Q-критика к более "интеллектуальной" схеме V-критика, в которой явно выделен блок "Модель", предназначенный для прогноза будущих ситуаций. Предварительные исследования такой

модели на основе V-критика действительно продемонстрировали ее эффективность: и для синусоиды и для реальных данных скорость обучения повысилась на порядок.

Список литературы

[Мосалов, 2004] Мосалов О.П. Модель эволюции системы агентов-брокеров // Научная сессия МИФИ – 2004. VI Всероссийская научно-техническая конференция "Нейроинформатика-2004": Сборник научных трудов. Часть 2. М.: МИФИ, 2004. С.138-144.

[Мосалов и др., 2003] Мосалов О.П., Бурцев М.С., Митин Н.А., Редько В.Г. Модель многоагентной Интернет-системы, предназначенной для предсказания временных рядов // V Всероссийская научно-техническая конференция "Нейроинформатика-2003". Сборник научных трудов. М.: МИФИ, 2003. Т.1. С.177-183.

[Редько и др., 2004] 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.

[Prokhorov et al, 1997] Prokhorov D., Wunsch D. Adaptive critic designs // IEEE Trans. on Neural Networks. 1997. Vol. 8. N.5. P.997-1007.

[Prokhorov et al, 2001] Prokhorov D., Puskorius G. and Feldkamp L., "Dynamical Neural Networks for Control" // In: J. Kolen and S. Kremer (Eds.) A Field Guide to Dynamic Recurrent Networks, IEEE Press, 2001.

[Sutton et. al, 1998] Sutton R. and Barto A. Reinforcement Learning: An Introduction. – Cambridge: MIT Press, 1998. See also: http://www-anw.cs.umass.edu/~rich/book/the-book.html

[Workshop, 2002] Workshop "Learning and Approximate Dynamic Programming" (Mexico, April, 2002): http://ebrains.la.asu.edu/~nsfadp/

MODELS OF DECISION MAKING ON THE BASE OF NEURAL NETWORK ADAPTIVE CRITIC DESIGNS

Oleg P. Mosalov, Danil V. Prokhorov, Vladimir G. Red'ko

Institute of Optical Neural Technologies, RAS, Moscow

Ford Research and Advanced Engineering, Ford Motor Company, Dearborn, U.S.A.

Moscow Institute of Physics and Technologies

Abstract

Neural network adaptive critic designs (ACD) are outlined. ACD can be used as control systems of autonomous agents that ensure agent decision making. Operation of ACD is illustrated by an example

of Q-critic that simulates agent-broker functioning. Results of computer simulation of this ACD model are described.