В.Г. РЕДЬКО, Д.В. ПРОХОРОВ
ИОНТ РАН, г. Москва, Ford Research Laboratory, Detroit, U.S.A.,
redko@iont.ru, dprokhor@ford.com

НЕЙРОСЕТЕВЫЕ АДАПТИВЫЕ КРИТИКИ*

Аннотация

В работе кратко излагаются основные идеи, схемы и уравнения систем управления на базе адаптивных критиков (Adaptive Critic Designs). Конструкции критиков сопоставляются со схемами теории функциональных систем П.К. Анохина, и со схемами проекта "Животное" (развитого в работах М.М. Бонгарда с сотр.).

V.G. RED'KO, D.V. PROKHOROV
IONT RAS, Moscow, Ford Research Laboratory, Detroit, U.S.A.
redko@iont.ru, dprokhor@ford.com

NEURAL NETWORK BASED ADAPTIVE CRITICS

Abstract

Main ideas, schemes and equations of adaptive critic designs are described. Designs of critics are compared with schemes of the P.K. Anokhin's functional systems theory and with schemes of the project "Animal" by M.M. Bongard et al.

1. Обучение с подкреплением

В цикле работ Р. Саттона и Э. Барто был предложен и разносторонне исследован метод обучения с подкреплением (Reinforcement Learning) [1].

Общая схема обучения с подкреплением показана на рис.1.

Рассматривается агент (модельный организм), взаимодействующий с внешней средой. В текущей ситуации агент s_t выполняет действие a_t, получает подкрепление r_t и попадает в следующую ситуацию s_{t+1} (здесь и далее время предполагается дискретным: $t = 1,2,\ldots$). Подкрепление может быть положительным (награда) или отрицательным (наказание).

* Работа выполнена при финансовой поддержке ОИТВС РАН (программа ОИТВС-01 проект № 1.8) и РФФИ (проект № 02-07-90197)

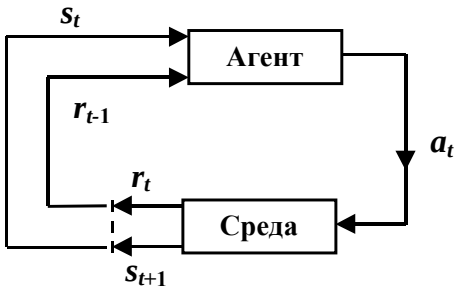

Рис.1. Схема обучения с подкреплением.

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

$$R_t = \sum_{k=0}^{\infty} (\gamma^k r_{t+k}) \quad , \qquad (1)$$

где R_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").

В простейшем случае множества возможных ситуаций $\{s_i\}$ и действий $\{a_j\}$ предполагаются конечными ($i = 1,\ldots N_s$; $j = 1,\ldots N_a$), Для этого случая существует простой метод обучения SARSA, каждый шаг которого соответствует цепочке событий s_t -> a_t -> r_t -> s_{t+1} -> a_{t+1}. Обучение происходит в on-line режиме. В процессе обучения итеративно формируются оценки суммарной величины награды Q (s_t , a_t), которую получит агент, если в ситуации s_t он выполнит действие a_t . Математическое ожидание награды равно:

$$Q(s_t, a_t) = \mathrm{E}\{(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots), \} | s = s_t , a = a_t \quad , \qquad (2)$$

Из (1) и (2) следует $Q(s_t, a_t) = E[r_t + \gamma Q(s_{t+1}, a_{t+1})]$. Ошибку естественно определить так:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) . \qquad (3)$$

Метод SARSA работает следующим образом. Исходные значения компонент матрицы Q (s_i , a_j) произвольны. Затем каждый такт времени одновременно 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 (s_t , a_t): $a^* = \arg\max_a \{ Q(s_t , a_t)\}$,
- с вероятностью ε выбирается произвольное действие, $0 < \varepsilon << 1$.

Такой выбор действия называют "ε –жадной" политикой.

Обучение, т.е. переоценка величин Q (s_t , a_t) происходит в соответствии с оценкой ошибки δ_t – к величине $Q(s_t, a_t)$ добавляется величина, пропорциональная δ_t :

$$\Delta Q(s_t, a_t) = \alpha \delta_t = \alpha [r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)], \qquad (4)$$

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. 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. Методы обучения с подкреплением, адаптивные критики и подобные методы часто называют приближенным динамическим программированием [2].

2. Что такое Критик

Конструкции адаптивных критиков можно рассматривать как развитие моделей обучения с подкреплением на случай, когда как ситуации, так и действия задаются векторами **S** и **A** и изложенная выше схема итеративного формирования матрицы Q (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.

Но понятие Критик имеет и самостоятельное значение. В конструкции агентов на основе адаптивных критиков входят два важных блока: Критик и Актор.

Критик – это блок системы управления, который оценивает качество ее работы.

Актор – блок системы управления, задающий действия этой системы.

Неформально понятие Критик можно пояснить следующим образом. Представим себе, что агент – мобильный робот, который должен добраться до целевой позиции в лабиринте и, по возможности, с минимальными энергозатратами. Предположим, что робот не знает, как добраться до цели (ориентироваться в лабиринте сложно, даже имея карту). Ему на помощь приходит Критик. Пусть у агента имеется не просто карта, а карта, на которой каждой позиции в лабиринте приписана полная стоимость оптимального пути из этой позиции до цели (оптимальная карта). В этом случае задача агента сильно упрощается, так как ему достаточно перемещаться из любой позиции в сторону наибольшего уменьшения стоимости оптимального пути. В этом примере Критик и является такой оптимальной картой. Эта карта формируется у агента итеративно, в процессе его "жизни", и в соответствии с этой картой формируются действия агента.

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 [3].

Ниже мы опишем две простые конструкции адаптивных критиков: Q-критик и V-критик. Обе конструкции используют нейросетевую аппроксимацию характеристик системы управления.

3. Q-критик

Схема Q-критика представлена на рис. 2. Предполагаем, что как Критик, так и Актор представляют собой многослойные персептроны

(такие же, какие используются в методе обратного распространения ошибки) с весами синапсов $\mathbf{W_C}$ и $\mathbf{W_A}$, соответственно.

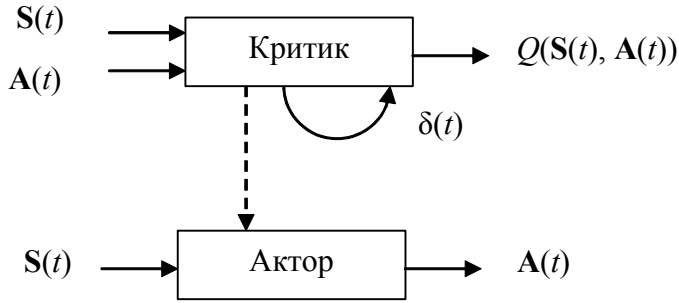

Рис. 2. Схема Q-критика.

Функционирование этой схемы происходит следующим образом. В момент времени t Актор по вектору входной ситуации $\mathbf{S}(t)$ определяет вектор действия $\mathbf{A}(t)$ (команды на эффекторы). Действие $\mathbf{A}(t)$ выполняется, агент получает награду $r(t)$. На входы Критика подаются два вектора: $\mathbf{S}(t)$ и $\mathbf{A}(t)$. По этому составному вектору Критик делает оценку качества $Q(t) = Q(\mathbf{S}(t),\mathbf{A}(t))$ действия $\mathbf{A}(t)$ в текущей ситуации $\mathbf{S}(t)$. Далее происходит переход к следующему моменту времени t+1. Все операции повторяются, в том числе делается оценка значения $Q(t+1)$. После этого определяется ошибка временной разности:

$$\delta(t) = r(t) + \gamma\, Q\,(t+1) - Q\,(t). \qquad (5)$$

Обучение нейросетей выполняется следующим образом:

$$\Delta\, \mathbf{W_C} = \alpha_1\, \mathrm{grad}_{\mathbf{WC}}(Q\,(t))\, \delta(t), \qquad (6)$$

$$\Delta\, \mathbf{W_A} = \alpha_2\, \Sigma_k\, \{[\partial Q\,(t)\, /\partial\, A_k(t)]\, \mathrm{grad}_{\mathbf{WA}}\, A_k(t)\,\}, \qquad (7)$$

где α_1 и α_2 - параметры скорости обучения. Производные по весам синапсов $\mathrm{grad}_{\mathbf{WC}}$ (.) и $\mathrm{grad}_{\mathbf{WA}}$ (.) в (6) и (7), а также $\partial Q\,(t)\, /\partial\, A_k(t)$ в (7) рассчитываются как производные сложных функций, аналогично тому, как это делается в методе обратного распространения ошибки. В формуле (7) учитывается, что нужно брать производные по всем компонентам вектора $\mathbf{A}(t)$ и суммировать по всем этим компонентам.

Смысл изменений весов синапсов по (6),(7) состоит в том, что веса Критика и Актора меняются таким образом, чтобы уменьшить ошибку в оценке ожидаемой суммарной награды (обучение Критика) и увеличить значение самой награды при попадании агента в близкие ситуации (обучение Актора). Обучение идет градиентным методом, общая процедура обучения описана в [3].

4. V-критик

Схема V-критика, использующего модель, представлена на рис. 3.

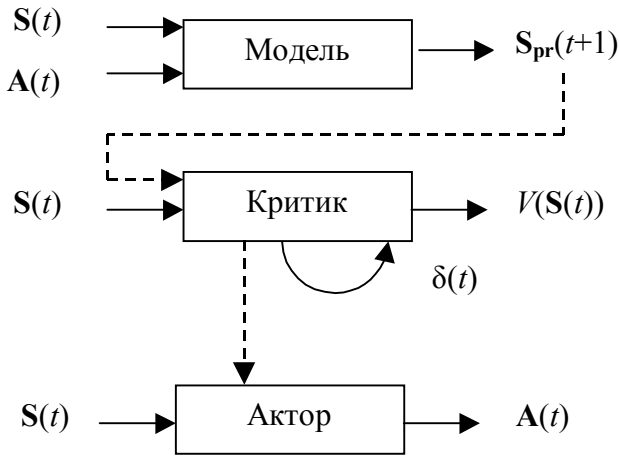

Рис. 3. Схема V-критика.

В этой схеме Критик, в отличие от схемы Q-критика, оценивает качество ситуации $V(\mathbf{S}(t))$ независимо от выполняемого действия. Однако такой Критик содержит Модель, в которой прогнозируется будущее состояние $\mathbf{S}_{\mathbf{pr}}(t+1) = \mathbf{S}_{\mathbf{pr}}(\mathbf{S}(t), \mathbf{A}(t))$ в зависимости от текущего состояния $\mathbf{S}(t)$ и выполняемого действия $\mathbf{A}(t)$ и для прогнозируемого состояния $\mathbf{S}_{\mathbf{pr}}(t+1)$ критик может сделать оценку его качества $V_{\mathbf{pr}} = V(\mathbf{S}_{\mathbf{pr}}(t+1)) = V(\mathbf{S}_{\mathbf{pr}}(\mathbf{S}(t), \mathbf{A}(t)))$.

Предполагаем, что Критик, Актор и Модель представляют собой многослойные персептроны с весами синапсов $\mathbf{W}_{\mathbf{C}}$,$\mathbf{W}_{\mathbf{A}}$ и $\mathbf{W}_{\mathbf{M}}$, соответственно.

Функционирование этой схемы происходит следующим образом. В момент времени t Актор по вектору входной ситуации $\mathbf{S}(t)$ определяет

вектор действия $\mathbf{A}(t)$. Критик делает оценку качества $V(t) = V(\mathbf{S}(t))$ текущей ситуации $\mathbf{S}(t)$. Модель прогнозирует следующее состояние $\mathbf{S_{pr}}(t+1) = \mathbf{S_{pr}}(\mathbf{S}(t), \mathbf{A}(t))$. Критик оценивает качество прогнозируемой ситуации $V_{\mathbf{pr}} = V(\mathbf{S_{pr}}(t+1))$. Действие $\mathbf{A}(t)$ выполняется, агент получает награду $r(t)$. Оценивается ошибка временной разности:

$$\delta(t) = r(t) + \gamma\, V(\mathbf{S_{pr}}(t+1)) - V(\mathbf{S}(t)) . \qquad (8)$$

Обучаются Критик:

$$\Delta\, \mathbf{W_C} = \alpha_1 \operatorname{grad}_{\mathbf{WC}}(V(t))\, \delta(t), \qquad (9)$$

и Актор:

$$\Delta\, \mathbf{W_A} = \alpha_2 \Sigma_k \{[\partial\, V(\mathbf{S_{pr}}(t+1)/\, \partial\, A_k(t)] \operatorname{grad}_{\mathbf{WA}} A_k(t)\} , \qquad (10)$$

$$\partial\, V(\mathbf{S_{pr}}(t+1)/\, \partial\, A_k(t) = \Sigma_j \{[\partial\, V/\, \partial S_{pr\,j}\,]\, [\partial S_{pr\,j}\, /\partial\, A_k(t)]\}. \qquad (11)$$

Производные в (11) берутся в соответствии с формулами нейронных сетей Критика и Модели.

Производится переход к следующему моменту времени $t+1$. Сравниваются прогнозируемая $\mathbf{S_{pr}}(t+1)$ и реальная ситуация $\mathbf{S}(t+1)$. В соответствии с ошибкой этого прогноза Модель тоже может быть обучена, например, обычным методом обратного распространения ошибки.

Обучение Критика состоит в том, чтобы итеративно уточнять оценку качества ситуаций $V(\mathbf{S}(t))$ в соответствии с поступающими подкреплениями.

Обучение Актора состоит в том, чтобы постепенно формировать действия, приводящие к ситуациям с высокими значениями качества.

Смысл обучения Модели – уточнение прогнозов будущих ситуаций.

Отметим, что Критик, оценивающий функцию качества $V(\mathbf{S}(t))$ в этой схеме, аналогичен "Хорошометру" в моделях А.А. Жданова [4].

6. Адаптивные критики, теория функциональных систем П.К. Анохина, проект "Животное" М.М. Бонгарда

Интересно отметить, что Модель и Критик в схеме V-критика в совокупности формируют прямой аналог акцептора результата действия в теории функциональных систем П.К. Анохина [5]. Т.е., у V-критика есть

прогноз в виде параметров результата действия (на выходе Модели) и прогноз качества действия (на выходе Критика), что в совокупности соответствует акцептору результата действия. Однако, работа функциональной системы (ФС) в целом более "интеллектуальна" по сравнению с работой отдельного адаптивного критика, так как процесс обратной афферентации в теории функциональных систем при существенном рассогласовании прогноза и результата может вызвать не просто уточнение прогноза, как в V-критике, а и радикальное изменение работы всей системы управления, например, передачу управления от одной ФС к другой или вообще формирование новой ФС [5,6].

Схема V-критика также аналогична отдельному блоку памяти в схеме управления агентом в проекте "Животное" [7], где в каждом блоке запоминается факт в виде отображения:

$$\{\mathbf{S}(t), \mathbf{A}(t)\} \text{ -> } \{\mathbf{S}_{\mathbf{pr}}(t+1), \mathbf{R}(t)\}, \qquad (12)$$

Здесь, как и выше, $\mathbf{S}(t)$ и $\mathbf{A}(t)$ – текущие ситуация и действие, $\mathbf{S}_{\mathbf{pr}}(t+1)$ – прогнозируемая ситуация. $\mathbf{R}(t)$ – прогнозируемый результат. Такой блок подобен отдельной функциональной системе в теории П.К. Анохина.

Отметим также, что как теория функциональных систем, так и проект "Животное" подразумевают сходную схему обучения, которые включают в себя формирование новых ФС или новых блоков памяти, соответственно.

Для теории функциональных систем такую схему обучения можно представить следующим образом. Вся система управления агента включает в себя структуру, состоящую из отдельных ФС, например, иерархическую структуру ФС, в которой происходит передача управления от одних ФС к другим [8]. Рассматриваемая ФС делает прогноз результата и будущей ситуации. Если прогноз верен, то можно продолжать действовать обычным образом: передавать управление следующим ФС, и совершать действия в соответствии с их указаниями. Но если прогноз оказался неверен, то, во-первых, надо предпринять нестандартные действия (например, отступить и выяснить: "в чем же дело?"), а во-вторых, обучиться и модифицировать систему управления в соответствии с новым опытом [6,9]. Обучение естественно представить, как формирование новой ФС, которая может формироваться методом проб и ошибок, например, из нейронных структур мозга путем эволюционного поиска нейронной сети новой ФС [6], аналогичное тому, которое рассматривается в концепции нейродарвинизма Дж. Эдельмана [10].

В проекте "Животное" процесс обучения так же включает формирование новых блоков памяти, при этом отмечается, что "запоминать имеет смысл только неожиданные либо важные факты" [7].

Основываясь на изложенных в данном разделе схемах обучения, можно предлагать дальнейшие пути развития блочно-иерархических систем управления агентов на базе адаптивных критиков.

Список литературы

1. 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
2. Workshop "Learning and Approximate Dynamic Programming" (Mexico, April, 2002): http://ebrains.la.asu.edu/~nsfadp/
3. Prokhorov D., Wunsch D. Adaptive critic designs // IEEE Trans. on Neural Networks. 1997. Vol. 8. N.5. P.997-1007.
4. Жданов А.А. Метод автономного адаптивного управления // Известия Академии Наук. Теория и системы управления. 1999. N. 5.
5. Анохин П.К. Принципиальные вопросы общей теории функциональных систем // Принципы системной организации функций. – М.: Наука, 1973. См. также: http://www.keldysh.ru/pages/BioCyber/RT/Functional.pdf
6. Анохин К.В. Частное сообщение.
7. Бонгард М.М., Лосев И.С., Смирнов М.С. Проект модели организации поведения – Животное // Моделирование обучения и поведения. – М.: Наука, 1975. С.152-171.
8. 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.
9. Данная идея К.В. Анохина была озвучена М.С. Бурцевым на передаче А.Г. Гордона на НТВ 5 ноября 2002. (Передача В.Г. Редько, М.С. Бурцев "Моделирование происхождения интеллекта").
10. Edelman G. M. Neural Darwinism: The theory of neuronal group selection. Oxford: Oxford University Press, 1989.