УДК 004.8

ОБОБЩЕННАЯ СХЕМА АДАПТИВНЫХ КРИТИКОВ

Е.А. ШУМКОВ

Кубанский государственный технологический университет,
350072, Российская Федерация, г. Краснодар, ул. Московская, 2;
электронная почта: sneveld@rambler.ru

Применение топологий адаптивных критиков в настоящее время являются передовым методом при построении систем управления сложными объектами, действующих в недетерминированной среде. Известно более десятка топологий адаптивных критиков, но ни одна из них не может быть принята, как универсальная. В статье рассмотрены основные топологии адаптивных критиков Q – критик и V – критик, отмечены их достоинства и недостатки. Предложен обобщенный подход к построению нейросетевой топологии адаптивного критика.

Ключевые слова: обучение с подкреплением, ошибка временной разности, адаптивный критик, Q – критик, V – критик, адаптивное поведение, нейросетевая топология управления, нейронная сеть.

Адаптивные критики (англ. adaptive critic design - ACD) являются, пожалуй, самой распространенной, после Q – обучения, реализацией обучения с подкреплением в текущее время. Они ведут свое начало с работы [6]. Отметим весомый вклад американского ученого советского происхождения Д. Прохорова [5]. В настоящее время разработано целое семейство различных конструкций адаптивных критиков.

Адаптивные критики – это схемы управления, которые содержат специальный блок – *Критик*, который оценивает качество работы всей системы [3, 4]. Топология адаптивного критика также содержит *Агента* (объект управления), который выполняет определенные действия в окружающей среде и тем самым взаимодействует с ней. Обычно влияние агента на внешнюю среду не велико, но влияние внешней среды на агента обычно значительно. Схема работы следующая: Агент в текущей ситуации $S(t)$ выполняет действие $a(t)$, получает подкрепление $r(t+1)$ и переходит в следующую ситуацию $S(t+1)$.

$$\ldots \to S(t) \to a(t) \to r(t+1) \to S(t+1) \to a(t+1) \to r(t+2) \to \ldots$$

Под S обычно понимается описание параметров агента, но нередко к ним добавляются параметры внешней среды. Ценность моделей *агент – критик* в том, что стратегия представляется независимо от функции ценности [3].

Введем понятие подкрепления – это оценка действия агента независимой компонентой за определенное время. Подкрепление обычно безразмерная величина. Подробнее про подкреплением см. [3, 4].

Адаптивные критики используют в своей работе ошибку временной разности[1] (далее *ОВР*) [3]. Коротко остановимся на ОВР. Постановка задачи, обычно следующая – допустим, что есть некая многошаговая задача, в которой имеется последовательность состояний. Цель – достичь конечного состояния с заданным показателем качества (подкрепления). При этом реальное значение показателя становится известным только после того, как система достигнет конечного состояния. Системе необходимо научиться давать прогноз на основе текущего состояния. При использовании традиционных методов прогнозирования, система дает прогноз для каждого состояния системы и запоминает их. После достижения конечного состояния, считаются ошибки прогноза на каждом шаге и после этого, каким – либо способом проводятся меры по повышению качества прогноза. Таким образом, система может повысить качество прогноза только после прохода всех состояний и достижения конечной точки.

При использовании метода ОВР в качестве ошибки вычисляется разность между двумя последовательными прогнозами. То есть система может производить коррекцию после каждого перехода.

В задачах с подкреплением, и в адаптивных критиках в частности, необходимо обучить систему последовательному прогнозу функции ценности:

$$V_t = E\{\sum_{k=t}^{\infty} \gamma^{(k-t)} \cdot r_k\} \quad (1),$$

где γ - коэффициент забывания (обычно $0 < \gamma < 1$).

Прогноз V_t может быть скорректирован следующим образом:

$$\Delta V_t = \alpha \cdot [\sum_{k=t}^{\infty} \gamma \cdot r_k - V_t] \quad (2),$$

[1] англ. Time – Difference Error (TDE).

где α - коэффициент скорости обучения. Выражение (2) может быть записано в терминах временной разности между последующими друг за другом прогнозами:

$$\Delta V_t = \alpha \cdot [(r_t + \gamma \cdot V_{t+1} - V_t) + \gamma \cdot (r_{t+1} + \gamma \cdot V_{t+2} - V_{t+1}) + ...] = \alpha \cdot \sum_{k=t}^{\infty} (r_t + \gamma \cdot V_{k+1} - V_k) \cdot \gamma^{(k-t)}$$

(1.3)

Прогноз V_t может быть сгенерирован при помощи аппроксиматора функции V, который снабжен вектором внутренних весов w. Веса такого аппроксиматора могут корректироваться с помощью метода градиентного спуска. Тогда выражение для корректировки весов за все время работы системы примет следующий вид:

$$\Delta w = \sum_{t=0}^{\infty} \Delta w_t = \sum_{t=0}^{\infty} \eta \cdot [\sum_{k=t}^{\infty} (r_k + \gamma \cdot V_{k+1} - V_k) \cdot \gamma^{(k-t)}] \cdot \nabla_w \cdot V_k \qquad (1.4)$$

где η - коэффициент обучения, который включает в себя α. Логичным выглядит использование многослойного персептрона в качестве такого аппроксиматора.

Различают две основные топологии адаптивных критиков – это Q – критик и V – критик.

Рассмотрим подробно Q - критика. В топологии Q – критика предполагается, что и Агент и Критик являются многослойными персептронами, но возможны и другие реализации. При этом Q - критик аппроксимирует Q - таблицу (см. подробнее в [3]). Функционирование Q – критика происходит следующим образом:

1. Агент в момент времени t по вектору входной ситуации $S(t)$ определяет вектор возможного действия $A(t)$;
2. выполняется действие $A(t)$ и Агент получает награду $r(t)$;
3. на входы Критика подаются два вектора $A(t)$ и $S(t)$;
4. по составному вектору $A(t) + S(t)$ Критик вычисляет оценку качества $Q(t) = Q(A(t), S(t))$ действия $A(t)$ в текущей ситуации $S(t)$;
5. происходит переход к следующему моменту времени $t + 1$;

6. повторяются действия 1.-5. и вычисляется оценка значения $Q(t+1)$;
7. вычисляется оценка временной разности по формуле:

$$\delta(t) = r(t) + \gamma \cdot Q(t+1) - Q(t) \quad (1.5)$$

8. обучаются нейронные сети Критика и Агента по формулам:

$$\Delta W_C = \alpha_1 \cdot grad_{WC}(Q(t)) \cdot \delta(t) \quad (1.6)$$

$$\Delta W_A = \alpha_2 \cdot \sum_k \{[\frac{\partial Q(t)}{\partial A_k(t)}] \cdot grad_{WA}(A_K(t))\} \quad (1.7)$$

где α_1 и α_2 - соответственно скорости обучения нейросетей Критика и Агента. Смысл изменения весов – уменьшить ошибку в оценке ожидаемой награды (обучение Критика) и увеличить значение самой награды при попадании Агента в сходные ситуации (обучение Агента) [2]. Таким образом, Q - критик «содержит» в себе таблицу, где столбцы и строки это состояния – действия, а в ячейках лежит уже накопленная история «успехов и неудач», какое подкрепление агент получил, выполняя в i - м состоянии j - е действие (см. Таблица 1).

Таблица 1

Таблица Q - критика

	A_1	**A_2**	**A_3**	**…**	**A_n**
S_1	$\sum R_{11}$	$\sum R_{12}$	$\sum R_{13}$	…	$\sum R_{1n}$
S_2	$\sum R_{21}$	$\sum R_{22}$	$\sum R_{32}$	…	$\sum R_{2n}$
S_3	$\sum R_{31}$	$\sum R_{32}$	$\sum R_{33}$	…	$\sum R_{3n}$
…	…	…	…	…	…
S_m	$\sum R_{m1}$	$\sum R_{m2}$	$\sum R_{m3}$	…	$\sum R_{mn}$

Возможны различные реализации, вплоть до модуля прогнозирования на основе истории поступления подкрепления в i - м состоянии выполнения j - е действие. Отметим, что количество возможных действий в различных состояниях в общем случае неравны и обычно m >> n. Впрочем, данный момент можно разрешить с помощью специальных обозначений и ограничений.

Для Q - критика одна из основных проблем – это процесс исследования, в котором система нарабатывает таблицу 1 [3].

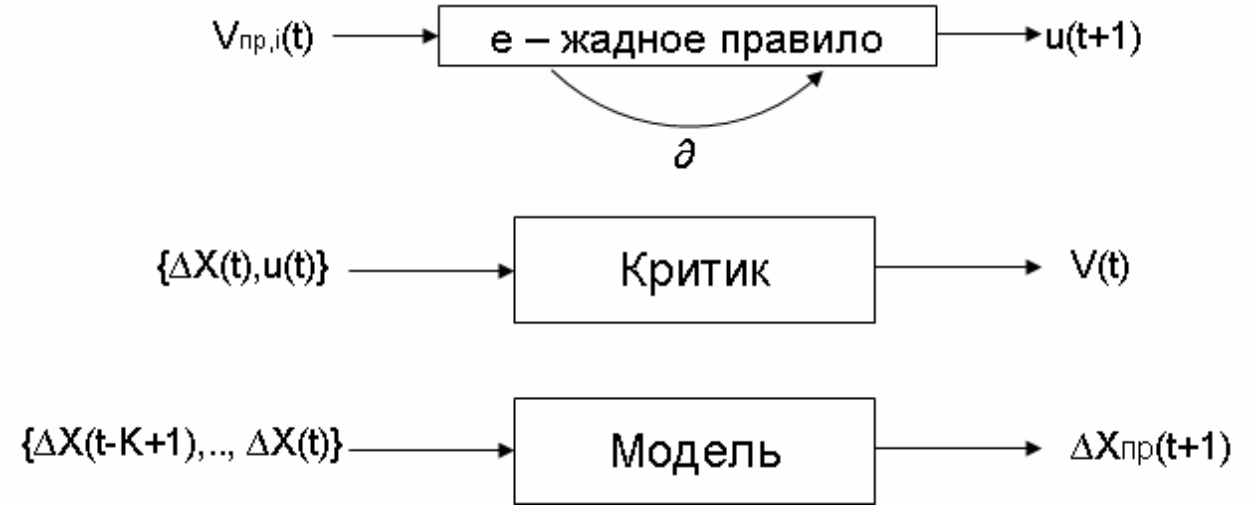

Рисунок 1 - V - критик

Другим распространенным критиком является V – критик, его схема представлена на Рисунке 1.

Согласно [2] V – критик, в отличие от Q – критика, оценивает качество ситуации $V(S(t))$ независимо от выполняемого действия. При этом V – критик содержит так называемую Модель, которая прогнозирует будущее состояние $S_{PR}(t+1) = S_{PR}(S(t), A(t))$ на основании текущих значений $A(t)$, $S(t)$ и возможно их предыдущих значений за несколько последних итераций. Таким образом, V – критик делает оценку качества прогнозного состояния $V_{PR} = V(S_{PR}(t+1))$. Обычно полагают, что в модели V – критика Модель и Критик реализуются на базе многослойных персептронов с алгоритмом обратного распространения ошибки [2].

Кратко рассмотрим конкретный пример работы топологии с критиком. Пусть мы моделируем работу автотрейдера на рынке Forex и у него есть три варианта действия на каждом шаге: а) купить (обозначим как a_1), б) не предпринимать действия (a_2), в) продать (a_3). Купить и продать при условии, что счет позволяет купить и есть, что продавать. Задача Модели в данном случае спрогнозировать значение будущего курса пары валют, т.е. получить значение $\Delta X(t+1)$. Далее, так как у нас три варианта действий, то на Критика

последовательно поступает три пары значений $\{a_1; \Delta X(t+1)\}$, $\{a_2; \Delta X(t+1)\}$, $\{a_3; \Delta X(t+1)\}$ и для каждого из них Критик вычисляет ценность действия. Далее получив три значения ценности, с помощью «жадного» - правила выбирается одно из них и на объект управления соответствующее действие a_i. После отработки выбранного действия критик дообучается используя получившееся значение ОВР.

Здесь есть несколько сложных моментов, в частности – это правильность прогнозирования Моделью и правильность вычисления ценности действия. Так как данные топологии с подкреплением строятся для систем, работающих в недетерминированной среде, то эти моменты существенны, т.к. приходится часто переобучать нейросетевые модели Критика и Модели. Изначально заложено, что нейросеть Критика постоянно дообучается / переобучается на ошибке временной разности.

Предложим обобщенную схему адаптивного критика, которая представлена на Рисунке 2, модифицировав схема из работы [1].

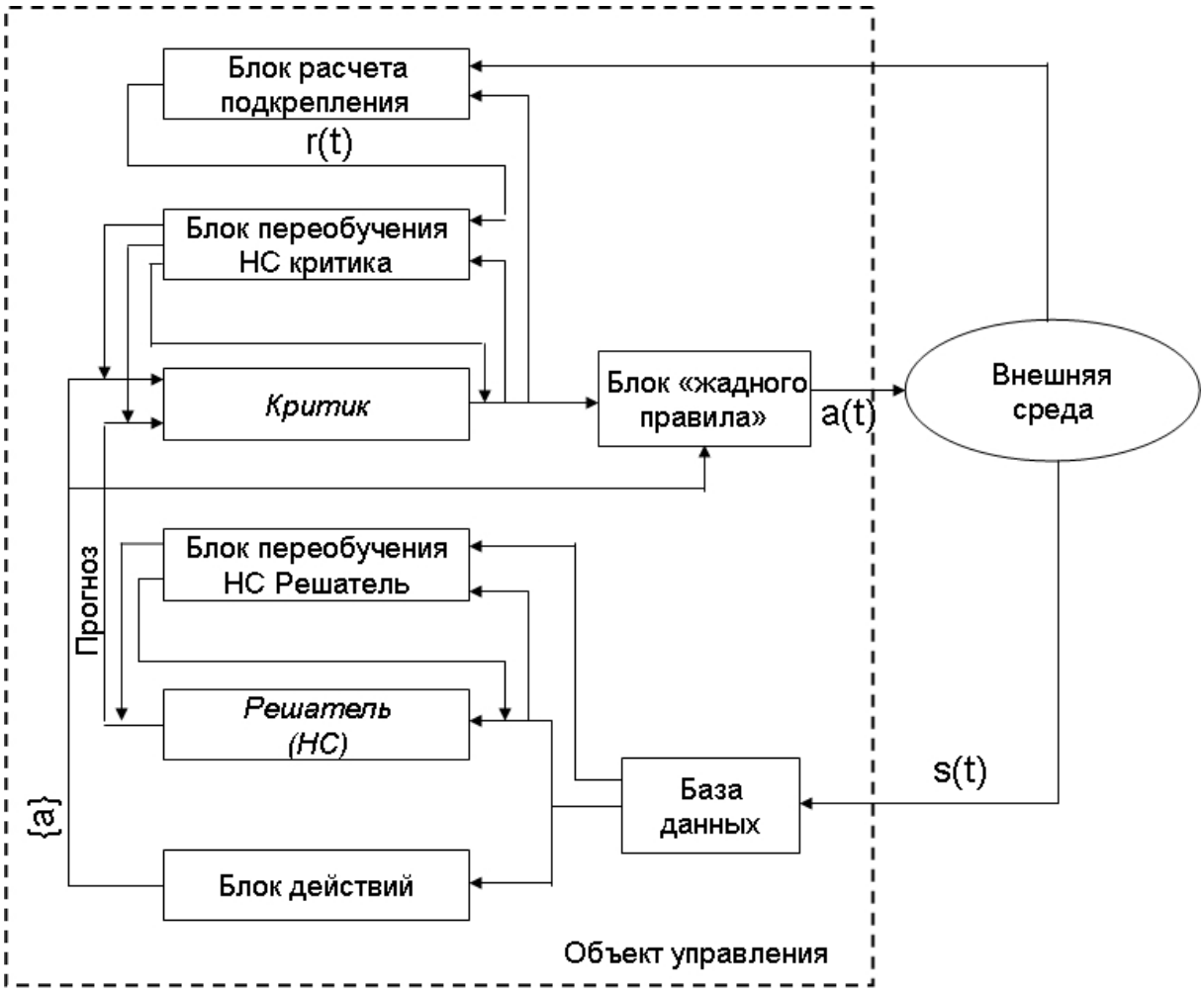

Рисунок 2. Обобщенная схема адаптивного критика

На Рисунке 2 введено обозначение: НС – нейронная сеть.

Дадим последовательность работы схемы. В момент времени *t* объект управления выполняет какое – то действие, которое влияет на внешнюю среду и/или учитывается ею. Внешняя среда принимает действие агента и, возможно, производит свои действия, которые могут повлиять на объект управления. Далее данные о внешней среде поступают в базу данных объекта управления, возможно через сенсорные механизмы и датчики, и далее на Решатель, блок переобучения нейронной сети Решателя и Блок действий. Блок переобучения нейронной сети Решателя принимает решение о необходимости переобучения (или дообучения) нейронной сети Решателя и в случае положительного решения переобучает ее. После переобучения нейронной сети Решателя прогнозируется рабочий параметр объекта управления (или внешней среды) или оценка состояния системы, в зависимости от решаемой задачи.

Параллельно в Блоке действий выбираются возможные действия в данной ситуации (обычно количество возможных действий одинаково для всех ситуаций). Далее и прогнозное значение и выбранные действия попарно подаются на Критика, который оценивает качество ситуации для следующего момента времени для каждой поданной пары. Блок жадного правила выбирает действие на основе данных Критика, которое и отрабатывает объект управления. Параллельно вышеописанному процессу в системе рассчитывается ОВР (Блок расчета подкрепления), на которой переобучается (дообучается) Критик.

Таким образом, в статье рассмотрены основные топологии адаптивных критиков - Q - критика и V - критика. Также предложена обобщенная топология адаптивного критика, которая использует возможности нейронных сетей по дообучению и упрощает конечную реализацию для разработчика.

ЛИТЕРАТУРА

1. Ботин В.А. Адаптивный критик с использованием фильтра Калмана. Дисс. канд. техн. наук. Краснодар: КубГТУ. 2011. 123 с.

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

3. Саттон Р.С., Барто Э.Г. Обучение с подкреплением / пер. с англ. – М.: БИНОМ. Лаборатория знаний. 2012. 399 с.

4. Шумков Е.А. Система поддержки принятия решений предприятия оптово – розничной торговли. Дисс. канд. техн. наук. Краснодар: КубГТУ. 2004. 158 с.

5. Prokhorov D., Wanch D. Adaptive critic designs. IEEE transactions on Neural Networks, September 1997, pp. 997-1007.

6. Witten I.H. An adaptive optimal controller for discrete – time Markov environments. // Information and Control. 1977. 34:286-295.

REFERENCES

1. Botin V.A. Adaptive critic using the Kalman filter. Thesis of cand. tech. sci. Krasnodar: KubSTU. 2011. 123 p.

2. Mosalov O.P., Prokhorov D.V., Redko V.G. Self-learning agents based on adaptive critic // Artificial intelligence. 2004. V.3. pp. 550-560.

3. Satton R.C., Barto E.G. Reinforcement learning / trans. from eng. – M.: BINOM. Laboratory knowledge. 2012. 399 p.

4. Shumkov E.A. Decision support system of the enterprise wholesale – retal.. Thesis of cand. tech. sci. Krasnodar: KubSTU. 2004. 158 p.

5. Prokhorov D., Wanch D. Adaptive critic designs. IEEE transactions on Neural Networks, September 1997, pp. 997-1007.

6. Mosalov O.P., Prokhorov D.V., Redko V.G. Self-learning agents based on adaptive critic // Artificial intelligence. 2004. V.3. pp. 550-560.

GENERALIZED SCHEME OF ADAPTIVE CRITIC

E.A. SHUMKOV

Kuban state Technical University,
2, Moskovskay st., Krasnodar, Russian Federation, 350072
e-mail: sneveld@rambler.ru

The use of adaptive topologies critics are now advanced by the construction of control systems for complex objects of management, acting in non-deterministic environment. There are more than a dozen topologies adaptive critics, but none of them can not be accepted as universal. In the article the basic topology adaptive critics Q - critic and V - critic, noted their advantages and disadvantages. A generalized approach to building an adaptive neural network topology criticism.

Key words: reinforcement learning, temporal difference error, adaptive critic, Q - critic, V - critic, adaptive behavior, neural network topology management, neural network.