Мосалов О.П., Редько В.Г., Прохоров Д.В. Модель агента-брокера на основе нейросетевых адаптивных критиков // Сб. трудов Международной научно-технической конференции «Интеллектуальные системы, IEEE AIS'03», М.: Физматлит, 2004. Т. 1. С. 395-399.

МОДЕЛЬ АГЕНТА-БРОКЕРА НА ОСНОВЕ НЕЙРОСЕТЕВЫХ АДАПТИВНЫХ КРИТИКОВ*

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

Рассматривается модель агента-брокера, виртуально играющего на бирже. Основное внимание уделяется моделированию процесса принятия решений при помощи нейросетевых адаптивных критиков. Цель работы – показать принципиальную возможность использования адаптивных критиков в такого рода задачах.

Введение

В настоящей работе развивается модель агента-брокера, который может принимать решения о покупке-продаже акций, играя на бирже. Принятие решения осуществляется с помощью нейросетевого адаптивного критика [Редько и др., 2004], при этом функция качества действия (action value function), которая используется в этом методе, оценивается аппроксимирующей нейронной сетью. Модель является развитием предыдущих версий нейросетевых моделей агента-брокера [Мосалов и др., 2003, Мосалов, 2004].

Общие предположения модели

1) Есть агент, который располагает некоторым количеством ресурсов двух типов: виртуальные 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).

* Работа выполнена финансовой поддержке РФФИ (проект № 04-01-00179) и РАН (Программа "Интеллектуальные компьютерные системы", проект 2-45)

[1] *119333, Москва, ул. Вавилова, 44, корп. 2, ИОНТ РАН, olegmos_@mail.ru*

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

$$R(t) = M(t) + N_A(t)\, X(t). \qquad (1)$$

4) Система управления агента основана на простой версии Q-критика [Редько и др., 2004]. 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 $r(t)$ [Sutton et.al, 1998] в процедуре обучения Q-критика, при переходе от такта времени t к такту t+1 равно:

$$r(t) = \Delta R(t) = N_A(t)^* \, [X(t+1) - X(t)] \, . \qquad (2)$$

Схема управления агента

Предполагаем, что принятие решения осуществляется с помощью Q-критика (рис. 1). На вход Критика поступают два типа сигналов: 1) сигналы, характеризующие текущую ситуацию $\mathbf{S}(t)$, и 2) сигнал, характеризующий одно из возможных действий a_i (в нашем случае есть только два действия, i =1,2). По этим сигналам Критик делает оценку $Q(\mathbf{S}(t), a_i)$ суммарной награды $U = \Sigma_k \, \gamma^k \, r(t+k)$, ожидаемой в будущем для данной ситуации $\mathbf{S}(t)$ для каждого из возможных действий a_i (γ – дисконтный фактор, $0 < \gamma < 1$). На основе этих оценок $Q(\mathbf{S}(t), a_i)$ Критик выбирает текущее действие a_S , используя ε-жадное правило. Обучение Критика происходит методом временной разности [Sutton et.al, 1998], ошибка временной разности $\delta(t)$ определяется выражением:

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

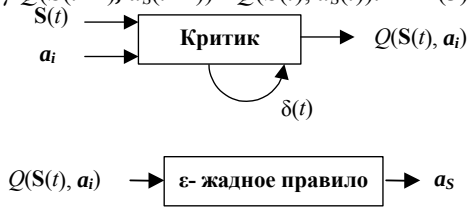

Рис. 1. Схема управления агента на основе Q-критика. Пояснения в тексте.

В нашей модели оценки $Q(\mathbf{S}(t), a_i)$ вычисляются с помощью нейронной сети (рис. 2). На вход нейронной сети подаются компоненты вектора $\mathbf{S}(t)$ и сигналы, характеризующие действия a_i . 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_i определим как $a_1 = -1$ для действия «продавать», $a_2 = +1$ для действия «покупать».

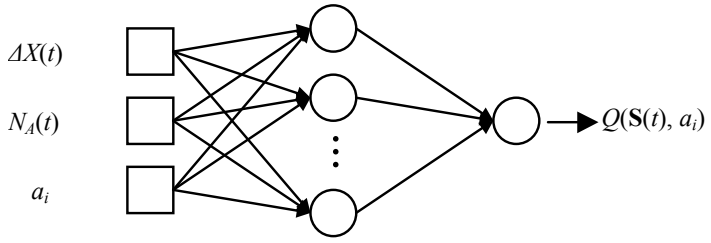

Рис. 2. Нейронная сеть агента.

Работа нейронной сети определяется следующими выражениями:

$$\mathbf{x} = \{\mathbf{S}(t), a(t)\},\ y_j = \text{th}\,(\textstyle\sum_i W_{ij} x_i),\ Q = \textstyle\sum_j V_j y_j\ , \quad (4)$$

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

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 $a_S(t)$ на текущем такте осуществляется следующим образом. Для каждого из возможных действий осуществляется работа нейронной сети, и вычисляются значения $Q(\mathbf{S},a_1)$ и $Q(\mathbf{S},a_2)$, соответственно. Дальше применяется ε-жадное правило:

- с вероятностью $(1 - \varepsilon)$ выбирается то действие, которому соответствует максимальное значение Q,
- с вероятностью ε выбирается произвольное действие $(0 < \varepsilon << 1)$.

Схема обучения агента

Цель обучения – уточнение оценок $Q(\mathbf{S}(t),a_i)$. Обучение нейронной сети производится методом градиентного спуска. Веса синапсов нейронной сети W_{ij} и V_j скрытого и выходного слоев изменяются на каждом такте пропорционально величине ошибки временной разности:

$$\Delta W_{ij} = \alpha\, \delta(t)\, \mathrm{grad}_W Q(\mathbf{S}(t), a_S(t)) , \qquad (5)$$
$$\Delta V_j = \alpha\, \delta(t)\, \mathrm{grad}_V Q(\mathbf{S}(t), a_S(t)) \quad , \qquad (6)$$

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

Используя (4), легко определить частные производные Q по весам синапсов нейронной сети. При этом формулы (5), (6) принимают вид:

$$\Delta W_{ij} = \alpha\, \delta(t)\, x_i\, (1 - y_j^2)\, V_j , \qquad (7)$$
$$\Delta V_j = \alpha\, \delta(t)\, y_j . \qquad (8)$$

Результаты моделирования

Модель была реализована в виде компьютерной программы. Моделирование проводилось как на модельных рядах: а) "пила": $X(2k) = 1$, $X(2k+1) = 2$, б) синусоида: $X(k) = 0.5[\sin(2\pi k/T)+1]$ ($k = 0,1,2, \ldots$; $T = 5, 10, 20, 100$), так и на реальных биржевых данных. Для всех компьютерных экспериментов исходные веса нейронной сети задавались случайно, и анализировался процесс обучения Q-критика. Предварительные результаты моделирования состоят в следующем.

Для очень простой модельной среды – "пила" (для которой пространство возможных ситуаций ограничено) – Q-критик успешно обучается. В более сложной среде – синусоида – обучение происходит, но не всегда стабильно. Этот случай иллюстрируется рис.3,4 (см. ниже). Для реальных биржевых данных только для некоторых удачных реализаций Q-критик находил хорошие решения.

Результаты моделирования для случая синусоиды с периодом $T = 20$ приведены на рис. 3,4. На графиках представлены зависимости изменения суммарного ресурса $R(t)$ и интегральной характеристики действий A(t) от времени. $A(t)$ определялась как сумма индексов I_S действий, выбираемых агентом (полагаем $I_S = 1$ и $I_S = -1$ для действий "покупать" и "продавать", соответственно): $A(t) = \sum_t I_S(t)$. На рис. 3 и 4 представлены зависимости $R(t)$, $A(t)$ в крупном и мелком масштабе, соответственно. Параметры расчетов составляли: $\alpha = 0.01$, $\varepsilon = 0.1$, $\max N_A = 5$, число нейронов в скрытом слое равно 6.

Для приведенного примера видно, что примерно при t = 7000 агент находит приемлемое решение (рис. 3), а при достаточно больших временах поведение агента становится стабильным и действия агента позволяют ему увеличивать ресурс в течение периода колебаний курса акций (рис.4).

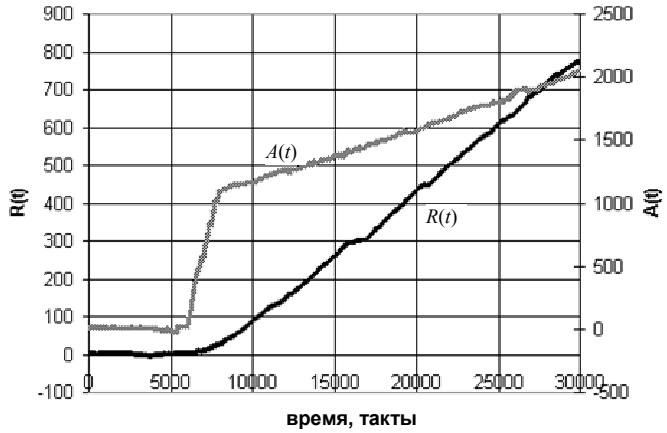

Рис. 3. Глобальные зависимости $R(t)$ и $A(t)$.

В заключение наметим пути дальнейшей работы над моделью. Основная причина недостаточно эффективного обучения Q-критика в нашей модели обусловлена слишком большим пространством возможных ситуаций при задании ситуаций числом акций N_A . Для ограничения пространства 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, целесообразно перейти от схемы Q-критика к более "интеллектуальной" схеме V-критика, в которой явно выделен блок "Модель", предназначенный для прогноза будущих ситуаций.

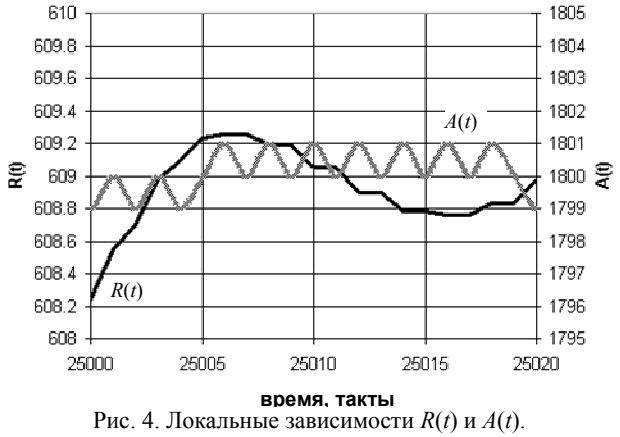

Рис. 4. Локальные зависимости $R(t)$ и $A(t)$.

Список литературы

[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.

[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.

[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, 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.