УДК 519.62

ОПРЕДЕЛЕНИЕ СТРАТЕГИЙ В ТРЕЙДИНГОВЫХ СИСТЕМАХ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

А.А. Гришко, С.Г. Удовенко, Л.Э. Чалая

ХНУРЭ, г. Харьков, Украина, udovenko@kture.kharkov.ua

В работе проведен анализ методов машинного обучения, применимых к компьютерным системам биржевой торговли. Предложен и протестирован модифицированный метод формирования торговых стратегий, основанный на использовании *Q*-обучения. Новый подход позволяет пользователю трейдинговой системы непрерывно получать, анализировать и использовать для принятия решений рыночную информацию.

ТРЕЙДИНГОВАЯ СИСТЕМА, *Q*-ОБУЧЕНИЕ, ТЕХНИЧЕСКИЙ АНАЛИЗ, ИНДИКАТОР, ТОРГОВАЯ СТРАТЕГИЯ

Введение

Развитие информационных технологий на основе применения методов вычислительного интеллекта способствовало появлению компьютерных систем биржевой торговли (трейдинговых систем). Возможности, которыми раньше располагали только крупные инвестиционные компании и банки, стали доступны широкому кругу пользователей сети Интернет, связанных с проблемами сбережения, накоплении и инвестирования свободных денежных средств. К наиболее перспективным путям решения этих проблем можно отнести торговлю на бирже через Интернет.

Последние исследования показывают эффективность применения в электронной биржевой торговле методов машинного обучения [1, 2]. Рассмотрим основные принципы, используемые при машинном обучении: обучение по результатам контроля, неконтролируемое обучение и обучение с использованием сигналов подкрепления. В первом случае обучение основано на анализе входного вектора и связанного с ним выходного вектора. Контроль основан на том, что в течение обучения требуемые выходные векторы задаются извне и используются как эталонные данные. Такой контроль наиболее часто применяется для распознавания образов, аппроксимации функций и прогнозировании. При неконтролируемом обучении единственной информацией, доступной обучающей системе является набор наблюдаемых входных образцов. В этом случае система обучается без вмешательства учителя, то есть она основана исключительно на использовании обучающих выборок. Зачастую, несмотря на большой объем обучающих данных, здесь не удается достичь желаемых результатов из-за трудоемкости обучения или сложности поставленной цели.

Принцип машинного обучения, который является наиболее перспективным для создания модели финансового рынка, – обучение с подкреплением (reinforcement learning (RL)). Суть такого обучения сводится к следующему: агент трейдинговой системы (например, торговый робот) должен исследовать текущие биржевые ситуации и принимать решения даже при неполном знании об этих ситуациях. Единственная обратная связь, получаемая агентом от биржевого рынка – скалярный сигнал подкрепления, который является положительным, если его действия выгодны трейдеру, и отрицательным в противном случае. Задача агента – выбрать свои действия, чтобы увеличить сумму сигналов подкреплений на длительном интервале времени [3]. Кроме сигналов подкрепления агент также получает информацию относительно текущего состояния биржевого рынка (в форме вектора наблюдений).

Для решения поставленных в настоящей статье задач модифицируем RL-метод, основанный на алгоритме *Q*-обучения, предложенном для частично наблюдаемых марковских процессов в работе [3].

Суть модификации состоит в расширении возможностей исходного алгоритма для его работы в он-лайн режиме на основе данных, получаемых из финансового рынка.

1. Постановка задачи

Целью настоящей работы является разработка и тестирование методов определения стратегий трейдера на электронной бирже на основе использования *Q*-обучения. В связи с этим были поставлены следующие задачи:

– разработать метод *Q*-обучения с подкреплением, применимый к электронным финансовым рынкам;

–протестировать процедуру формирования торговых стратегий, основанных на предложенном методе.

Для тестирования оценки эффективности разрабатываемых алгоритмических и программных средств используем данные международного межбанковского валютного рынка FX (Foreign Exchange Market), доступные индивидуальным пользователям.

2. Модифицированный метод *Q*-обучения

Задача *Q*-обучения с подкреплением может быть описана марковским процессом решений, который идентифицирует дискретный набор состояний окружающей среды S и выполняет одно из возможных действий из множества A. В ответ на действие a_t в момент t при текущем состоянии среды s_t агент системы получает ответный сигнал подкрепления $r_t = r(s_t, a_t)$ от окружающей среды, после чего окружающая среда переходит в новое состояние $s_{t+1} = \delta(s_t, a_t)$. В алгоритме используются функции перехода $\delta(s_t, a_t)$. Функции перехода и подкрепления зависят только от текущего состояния и действий и не зависят от предыдущих состояний и действий.

Задача базового алгоритма *Q*-обучения – определить и реализовать стратегию $\pi : S \to A$, основанную на текущем состоянии s_t; это может быть записано, как $\pi(s_t) = a_t$. Обычно требуется найти стратегию, соответствующую максимальному значению длительной суммы сигналов подкрепления. Чтобы формализовать это, введем функцию $V^{\pi}(s_t)$, которая является суммой всех сигналов, полученных алгоритмом, стартующим из состояния s_t с последующей стратегией π:

$$V^{\pi} \leftarrow \sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i} \ , \qquad (1)$$

где r_{t+i} – последовательность сигналов подкрепления; γ $(0 \le \gamma \le 1)$ – коэффициент дисконтирования, который определяет текущую оценку будущих доходов. При $\gamma = 0$ алгоритм отдает предпочтение максимизации текущего дохода. При приближении коэффициента дисконтирования к 1, больший вес алгоритм присваивает будущим доходам.

Оптимальная стратегия, максимизирующая полный доход, начиная с любого состояния, может быть представлена в виде:

$$\pi^* \leftarrow \arg\max_{\pi} V^{\pi}(s) \ , \qquad (2)$$

где $s \in S$.

Алгоритм реализации такой стратегии должен максимизировать сумму непосредственного дохода и значение функции подкрепления, уменьшенное коэффициентом дисконтирования:

$$a^* \leftarrow \arg\max_{a} \left[r(s,a) + \gamma \cdot V^*(\delta(s,a)) \right]. \qquad (3)$$

Уравнение (3) предполагает доступными сигнал подкрепления и функцию перехода. Однако в большинстве практических задач функция перехода недоступна. В работе [4] был впервые предложен одношаговый алгоритм *Q*-обучения, не использующий непосредственно функцию перехода. В этом алгоритме для определения оптимальной стратегии используется *Q*-функция, итеративную процедуру обновления которой можно представить в следующем виде:

$$Q_{t+1}(s,a) \leftarrow r + \gamma \cdot \max_{a \in A} Q(s',a) \ , \qquad (4)$$

$$Q_{t+1}(s,a) = \\ = Q_t(s,a) + \alpha \cdot (r + \gamma \cdot \max_{a' \in A} Q_t(s',a') - Q_t(s,a)) \ , \qquad (5)$$

где a – действие, вызывающее переход среды из состояния s в состояние s'; α $(0 \le \alpha \le 1)$ – коэффициент нормирования значений *Q*-функции.

В практических задачах алгоритм (4), (5) часто усиливают с помощью метода временной разности (Temporal Difference (TD)), что способствует ускорению обучения [5]. Идея TD-обучения состоит в использовании разности между двумя последовательными предсказаниями конечного ожидаемого результата вместо разности между текущим предсказанием и фактическим ожидаемым результатом.

Рассмотрим задачу применения TD-обучения для оценки функции (1), связанной с математическим ожиданием суммы будущих дисконтированных доходов:

$$V_t = E\left[\sum_{k=t}^{\infty} \gamma^{(k-t)} \cdot r_k \right], \quad \forall t \ . \qquad (6)$$

Для получения разности между действительными доходами и их математическими ожиданиями будем использовать следующее обновляющее правило:

$$\Delta V_t = \alpha \cdot \left[\sum_{k=t}^{\infty} \gamma^{(k-t)} \cdot r_k - V_t \right], \qquad (7)$$

где α – коэффициент обучения.

Оценку V_t можно осуществлять, используя аппроксиматор функции V^* с набором внутренних весов w. Эти весовые коэффициенты можно корректировать по методу градиентного спуска:

$$\Delta w_t = \eta \cdot \left[\sum_{k=t}^{\infty} \left[r_k + \gamma \cdot V_{k+1} - V_k \right] \cdot \gamma^{(k-t)} \right] \cdot \nabla_w \cdot V_t \ , \qquad (8)$$

где η – коэффициент обучения ($\eta = \eta(\alpha)$).

Изменение общего веса за все время работы системы (обновляющее правило) определится следующей суммой:

$$\Delta w_t = \left[r_t + \gamma \cdot V_{t+1} - V_t \right] \cdot \sum_{k=0}^{t} \eta \cdot \gamma^{(t-k)} \cdot \nabla_w \cdot V_k \ , \qquad (9)$$

где сумма с приращением обновляется на каждом шаге t.

В классическом методе *Q*-обучения *Q*-значения представляются таблицей с парами «состояние-действие». Практические задачи обычно характеризуются большим числом таких пар, что делает невозможным использование табличного представления *Q*-функций в большинстве случаев. Альтернативой таблицам перекодировки являются аппроксимирующие функции.

Рассмотрим возможность комбинирования *Q*-алгоритма с TD(λ)-алгоритмом с целью повышения скорости процесса обучения. Отметим, что метод *Q*-обучения, основанный на обновлении(5), получает весьма ограниченную информацию от системы и выполняет модификации, основанные только на значениях смежных состояний. В TD(λ)-алгоритме принимаются во внимание и предыдущие зафиксированные значения состояний. Заменим *Q*-таблицу а дифференцируемой *Q*-функцией, имеющей некоторый набор внутренних параметров.

В соответствии с (5) TD-ошибка для *Q*-обучения представляется следующей зависимостью:

$$TD_{err} = Q_{t+1}(s,a) - Q_t(s,a) = \\ = (r + \gamma \cdot \max_{a' \in A} Q_t(s',a') - Q_t(s,a)). \qquad (10)$$

Уравнение (10), определяющее ошибки временной разницы в терминах *Q*-значений, можно использовать как основу комбинированного *Q*/TD-обучения. Заменяя *Q*-значениями *V*-значения в уравнении (8), получаем:

$$\Delta w_t = \eta \cdot \left[r_t + \gamma \cdot \max_{a \in A} Q_{t+1} - Q_t \right] \cdot \nabla_w \cdot Q_t, \qquad (11)$$

где $\nabla_w \cdot Q_t$ – градиент дифференцируемой *Q*-функции.

Отметим, что методы аппроксимации таблиц перекодировки, применяемые в комбинированном *Q*/TD-обучении должны обеспечивать возможность реализации алгоритмов в режиме on-line и иметь приемлемые потребности в памяти. Таким требованиям отвечают нейросетевые методы, в частности, методы, основанные на использовании многослойного персептрона (MLP)[5].

Рассмотрим многошаговый модифицированный алгоритм *Q*-обучения. Очевидно, что оценки *Q*-функций являются неточными, пока состояния финансового рынка не будут неоднократно наблюдаться. Последовательные уточнения динамики рынка могут быть учтены путем применения многошагового уточняющего правила следующего вида:

$$\Delta w_t = \eta \cdot \left[r_t + \gamma \cdot Q_{t+1} - Q_t \right] \cdot \sum_{k=0}^{t} (\gamma \cdot \lambda)^{t-k} \cdot \nabla_w \cdot Q_k \,. \quad (12)$$

Для аппроксимации таблиц перекодировки эффективным является применение специального метода коррекции весов MLP, именуемого методом «обратного переигрывания» (backward replay (BR))[6]. В соответствии с этим методом веса обновляются только при достижении системой поглощающих состояний. Использование BR-процедур предполагает необходимость хранения всех пар «состояние – действие», которые встречаются до достижения системой поглощающего состояния.

Альтернативой BR-методу может служить применение интерактивного обучения (IL) для аппроксимации *Q*-таблиц. Идея IL-метода состоит в обновлении аппроксимирующей функции на каждой итерации до достижения системой конечного состояния. Следует отметить, что рассмотренные основные и вспомогательные процедуры *Q*-обучения нашли практическое применение в основном для поиска управляющих стратегий в технических системах (в частности, для управления роботом). Использование методов *Q*-обучения в трейдинговых системах имеет свои специфические особенности.

Для работы описанного *Q*/TD-алгоритма на электронной бирже в составе трейдинговой системы необходимо в реальном времени получать значения доступных индикаторов, помогающих обнаружить тенденции изменения основных показателей биржевого рынка. Индикаторы можно разделить на две большие группы: трендовые индикаторы и осцилляторы. Трендовые индикаторы применяются при анализе трендовых рынков и неэффективны, когда тренд отсутствует. Осцилляторы, наоборот, плохо работают на трендовых рынках и хорошо, когда тренд отсутствует. Ниже приведен пример формирования оптимальных стратегий с использованием алгоритма *Q*-обучения, иллюстрирующих принцип работы трейдерной системы, основанной на получении сигналов подкрепления от финансового рынка.

3. Пример *Q*-обучения на финансовом рынке

На рис. 1 рассматривается упрощенная торговая ситуация, где трейдер имеет два индикатора финансового рынка и по сигналам этих индикаторов он должен принимать решения. В этой упрощенной ситуации предполагается, что каждый индикатор может советовать совершить либо покупку (buy), либо продажу (sell). Индикаторы нейтральной позиции, советующие отказаться от торговых транзакций, здесь отсутствуют. Так как два индикатора могут принимать лишь по два значения каждый, то финансовый рынок (среда) может иметь 4 возможных состояния (state). В любом из этих состояний среды трейдер может принять два решения - купить или продать. Очевидно, что набор действий, доступных в каждом из состояний, может бать представлен следующим образом:

$$A(state\ i) = \{buy, sell\}, i = 1,2,3,4 \,.$$

Любое действие, осуществленное трейдером, приводит к переходу от одного состояния окружающей среды к другому в соответствии с некоторым набором вероятностей, определенным рыночной ситуацией.

Состояние окружающей среды характеризуется значениями индикаторов, и в любом состоянии трейдер должен выбрать действие «buy» или действие «sell». Для понимания процедуры обучения,

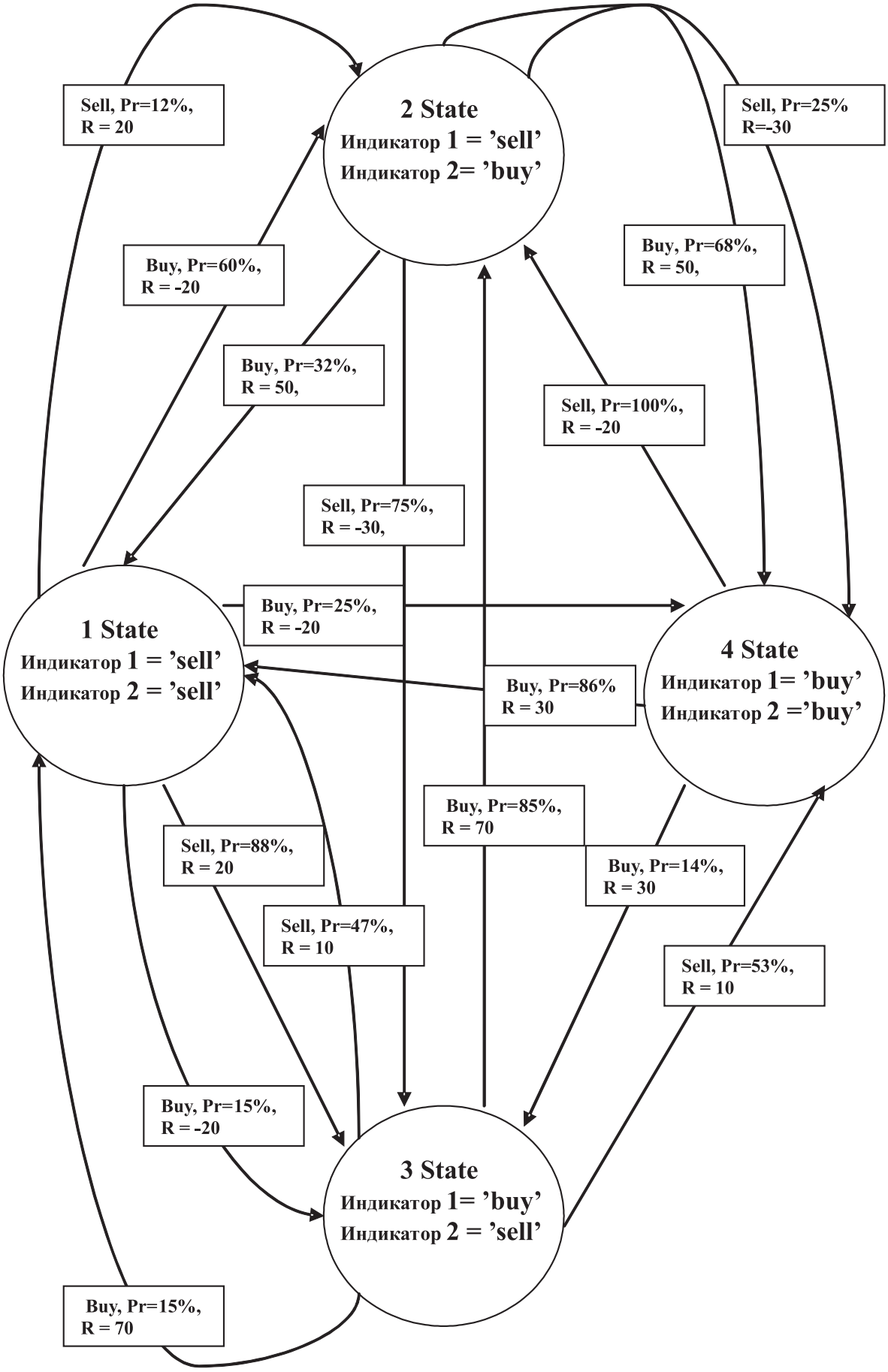

Рис. 1. Пример *Q*-обучения на финансовом рынке

представленной на рис. 1, рассмотрим, например, прямоугольник в левом верхнем углу этого рисунка, содержащий данные: «sell, Pr = 12%, R = 20». Это означает, что здесь переход из состояния 1 в состояние 2 будет иметь место, если трейдер в состоянии 1 выберет действие «sell», и что это решение приведет к переходу в состояние 2 с вероятностью 0,12, при котором трейдер получит сигнал подкрепления R = 20. Отметим, что решение «sell» в состоянии 1 может также привести к переходу в состояние 3 с вероятностью 0,88. Суммарная вероятность этих двух переходов равна 1, что указывает на нулевую вероятность перехода в состояние 4 при решении «sell» в состоянии 1.

Значения сигналов подкреплений и вероятностей переходов на рис.1 были выбраны произвольно, поскольку рассматриваемый пример имеет лишь иллюстративный характер. На практике эти значения изменяются в соответствии с динамической природой рынка, но на рис.1 они зафиксированы для понимания ключевых особенностей *Q*-обучения. Предположим, что начальные значения $Q_0(s,a)$ являются нулевыми и что коэффициент дисконтирования γ=0,85. Допустим также, что имеет место следующая последовательность «состояние – действие»:

$$(state1, sell) \to (state3, buy) \to (state3, buy) \to$$
$$(state2, sell) \to (state3, sell) \to (state4, buy) \to$$
$$(state3, buy) \to (state2, buy) \to (state1, buy) \to$$
$$(state2, sell) \to (state3, ...) \to ...$$

Рассмотрим для такой последовательности процедуру пересчета *Q*-значений по описанному выше методу *Q*-обучения. Отметим, что здесь каждый шаг *Q*-обучения осуществляется по следующей рекуррентной зависимости:

$$Q_{t+1}(s,a) = Q_t(s,a) +$$
$$+\alpha \cdot (r + \gamma \cdot \max_{a' \in A} Q_t(s',a') - Q_t(s,a)),$$

где α – коэффициент обучения, принятый здесь равным 0,15.

Последовательность первых 4 шагов модификаций имеет вид:

1) $(state1, sell, 20, state3)$

$$Q_1(state1,\ sell) = Q_0(state1,\ sell) +$$
$$+0{,}15\left[20 + 0{,}85 \max_a Q_0(state3,\ a) - Q_0(state1,\ sell)\right] =$$
$$-0 + 0{,}15\left[20 + 0{,}85 \max\{0{,}0\} - 0\right] = 0 + 0{,}15 * 20 = 3;$$

2) $(state3, buy, 70, state2)$

$$Q_2(state3,\ buy) = Q_1(state3,\ buy) +$$
$$+0{,}15\left[70 + 0{,}85 \max_a Q_1(state2,\ a) - Q_1(state3,\ buy)\right] =$$
$$= 0 + 0{,}15\left[70 + 0{,}85 \max\{0{,}0\} - 0\right] =$$
$$= 0 + 0{,}15 \cdot 70 = 10{,}5;$$

3) $(state2, sell, -30, state3)$

$$Q_3(state2,\ sell) = Q_2(state2,\ sell) +$$
$$+0{,}15\left[-30 + 0{,}85 \max_a Q_2(state3,\ a) - Q_2(state2,\ buy)\right] =$$
$$= 0 + 0{,}15\left[-30 + 0{,}85 \max\{10{,}5{,}0\} - 0\right] =$$
$$= 0 + 0{,}15 * (-30 + 0{,}85 \cdot 10{,}5) = -3{,}16;$$

4) $(state3, sell, 10, state4)$

$$Q_4(state3,\ sell) = Q_3(state3,\ sell) +$$
$$+0{,}15\left[10 + 0{,}85 \max_a Q_3(state4,\ a) - Q_3(state3,\ sell)\right] =$$
$$= 0 + 0{,}15\left[10 + 0{,}85 \max\{0{,}0\} - 0\right] = 0 + 0{,}15 \cdot 10 = 1{,}5.$$

Ниже приведена *Q*-таблица, сформированная после четвертого шага.

Таблица 1

Q-таблица после 4 шагов обучения

$Q_4(s,a)$	Buy	Sell
state 1	0	3
state 2	0	–3,16
state 3	10,5	1,5
state 4	0	0

Последовательность следующих пяти шагов пересчета *Q*-значений имеет вид:

$$(state4, buy) \to (state3, buy) \to (state2, buy) \to$$
$$(state1, buy) \to (state2, sell) \to (state3, ...) \to ... .$$

Q-таблица, сформированная после девятого шага, приведена ниже.

Таблица 2

Q-таблица после 9 шагов обучения

$Q_9(s,a)$	Buy	Sell
state 1	0,55	3
state 2	7,88	-4,71
state 3	19,43	1,5
state 4	5,84	0

Стратегия *Q*-обучения после 9 шагов определяется следующим уравнением:

$$\pi^*(s) = \arg\max_a Q_9(s,a).$$

Определим последовательность действий трейдера в соответствии с этой стратегией:

– если система находится в состоянии «state1», то выбирается действие «sell», так как 3 > 0,55;

– если система находится в состоянии «state2», то выбирается действие «buy», так как 7,88 > -4,71;

– если система находится в состоянии «state3», то выбирается действие «buy», так как 19,43 > 1,5;

– если система находится в состоянии «state4», то выбирается действие «buy», так как 5,84 > 0.

В статической окружающей среде, рассматриваемой в этом примере, *Q*-значения в конечном счете сошлись бы к фиксированным значениям и, таким образом, трейдер мог бы наиболее эффективно использовать для торговли значения *Q*-таблиц только после их стабилизации. В реальных трейдерных системах, когда окружающая среда является динамической, трейдеру, кроме анализа сходимости *Q*-значений, приходится использовать некоторые дополнительные критерии, чтобы определить наиболее целесообразный момент использования значений *Q*-таблиц для торговли.

Выводы

Результаты, полученные при моделировании торговых стратегий по данным валютного рынка FX, подтверждают работоспособность и перспективность применения машинного обучения с подкреплением в трейдинговых системах. Практическая реализация принципов электронной торговли с использованием рассмотренных выше процедур, осуществленная авторами настоящей статьи, состоит в следующем:

– составлен банк *Q*-обучающих алгоритмов с возможностью программного его пополнения пользователем;

– предложен вычислительный алгоритм, основанный на использовании обучающей модели с RL, и дополнительные средства, позволяющие принимать оперативные решения относительно входа в рынок и выхода из рынка;

– разработан программный модуль *Q*-Trayder, основанный на алгоритме, который позволяет принимать действия, имитирующие процесс торговли. Модуль может соединяться с сервером брокера, постоянно загружать данные с сервера, анализировать ситуацию на рынке, а затем обеспечивать формирование рекомендаций по ведению торговли.

Список литературы: 1. *Люггер, Д.* Искусственный интеллект: стратегии и методы решения сложных проблем [Текст] / Д.Люгер // М.: Издательский дом «Вильямс». – 2003. – 864 с. **2.** “Computational Learning Techniques for Intraday FX Trading Using Popular Technical Indicators / M. Dempster, T. Payne, Y Romahi, G. Thompson // IEEE Transactions on Neural Networks, Vol. 12, 4, July 2001. – P. 744-754. **3.** *Dempster, M.* Intraday FX trading: An evolutionary reinforcement learning approach. Intelligent data engineering and automated learning / M.Dempster, Y.Romahi // Proceedings of the IDEAL 2002 International Conference. – 2002. – P. 347-358. **4.** *Sutton, R.* Learning to Predict by the Methods of Temporal Differences, Machine Learning, 3, pp. 9-44, 1998. **5.** *Барский, А.Б.* Нейронные сети: распознавание, обучение, принятие решений [Текст] / А.Б. Барский. – М.: Финансы и статистика, 2004. – 320 с. **6.** *Hryshko A.* An Implementation of Genetic Algorithms as a Basis for a Trading System on the Foreign Exchange Market / A. Hryshko, T. Downs // Proceedings of the 2003 Congress on Evolutionary Computation. – 2003. – P. 1695-1701.

Поступила в редколлегию 19.02.2010 г.

УДК 519.62

Визначення стратегій в комп’ютерних трейдингових системах на основі методів машинного навчання / А.О. Гришко, С.Г. Удовенко, Л.Е. Чала // Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 1 (72). – С. 18–23.

У статті розглядаються стратегії трейдера у системах біржової торгівлі, що використовують методи машинного навчання з підкріпленням. Запропоновані алгоритми дозволяють перевищити за своїми характеристиками інші алгоритми машинного навчання, які використовуються на фінансових ринках. Розглянуті принципи трейдингової торгівлі засновані на розроблених методах. Методи реалізовано програмно та протестовано.

Табл. 2. Іл. 1. Бібліогр.: 6 найм.

УДК 519.62

Strategies of computing markets systems based on machine learning methods / A.A. Hryshko, S.G. Udovenko, L.E. Chalaya // Bionics of Intelligence: Sci. Mag. – 2010. – № 1 (72). – P. 18–23.

This paper is devoted to application of machine learning for computing markets systems. The algorithms developed in the proposed learning system is shown to outperform other machine learning algorithms used on financial markets. The principles of trading markets based on the development of novel trading strategies are considered. The new methods are implemented in software. The results of testing of methods are presented.

Tab. 2. Fig. 1. Ref.: 6 items.