УДК 519.62

А.А. Гришко, С.Г. Удовенко

НЕЙРОСЕТЕВАЯ АППРОКСИМАЦИЯ Q-ФУНКЦИЙ В ТРЕЙДИНГОВЫХ СИСТЕМАХ

Аннотация. Работа посвящена разработке системы формирования торговых стратегий, основанных на методах машинного обучения с подкреплением и процедурах настройки аппроксимирующего многослойного персептрона. Рассмотрены традиционные и альтернативные методы аппроксимации таблицы Q-значений при использовании искусственных нейронных сетей. Приведена возможная схема реализации предложенных вычислительных процедур для оптимизации стратегии принятия решений в трейдинговых системах.

Ключевые слова: трейдинговая система, многослойный персептрон, Q-обучение, индикатор, торговая стратегия

Введение

В последнее время получили распространение системы электронной биржевой торговли (трейдинговые системы), предусматривающие возможность максимизации дохода от биржевых операций на основе использования методов машинного обучения [1]. К наиболее эффективным методам машинного обучения, используемым для создания модели финансового рынка следует отнести методы, основанные на обучении с подкреплением (reinforcement learning (RL)). В процессе такого обучения агент компьютерной трейдинговой системы (например, торговый робот) должен по результатам анализа текущих биржевых ситуаций принимать решения даже при неполном знании об этих ситуациях. При этом агент принимает от биржевого рынка скалярный сигнал подкрепления, который является положительным, если его действия предроложительно выгодны трейдеру и отрицательным в противном случае. Задача агента состоит в выработке действий, увеличивающих сумму сигналов подкреплений на длительном интервале времени. Кроме сигналов подкрепления агент также получает информацию относительно текущего состояния биржевого рынка в форме вектора наблюдений. Одним из наиболее известных RL-

 ISSN 1562-9945

методов является метод, основанный на алгоритме Q-обучения, предложенном для частично наблюдаемых марковских процессов в работе [2]. В этом методе для определения оптимальной стратегии используется итеративно обновляемая Q-функция, исходно представляемая Q-таблицей, в которой задаются пары «состояние-действие». Для аппроксимации Q-функций в прикладных задачах обучения с подкреплением используются, как правило, радиально-базисные сети (РБС), многослойный персептрон (МСП), метод CMAC или непосредственная дискретизация пространства состояний [2]. При этом наибольшее распространение получили здесь искусственные нейронные сети на базе МСП.

Представляется целесообразным рассмотреть задачу выбора параметров МСП, используемых для аппроксимации Q-функций, и разработать схему принятия трейдинговых стратегий на основе МСП и RL-алгоритма.

Постановка задачи

Алгоритм Q-обучения с подкреплением в общем случае 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 состояний 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, алгоритм больший вес присваивает будущим доходам.

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

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

где $s \in S$.

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

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

Уравнение (3) предполагает доступными сигнал подкрепления и функцию перехода. Однако, в большинстве практических задач функция перехода недоступна. В работе [3] 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:

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

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

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

Отметим, что методы аппроксимации таблиц перекодировки Q-функций, применяемые для такой процедуры, должны обеспечивать возможность реализации RL-алгоритмов в режиме on-line и использовать приемлемые для решения практических задач биржевой торговли ресурсы памяти. Таким требованиям отвечают нейросетевые методы, в частности, методы, основанные на использовании МСП.

Обучение МСП может упускать относительно ранее полученную информацию, когда новая информация прибывает в систему. Однако МСП удовлетворяет всем критериям и представляет эффективное решение для алгоритмов RL. Целью настоящей работы является

 ISSN 1562-9945

исследование возможности увеличения дохода пользователей трейдинговых систем на основе использования методов машинного обучения с подкреплением. В связи с этим были поставлены следующие задачи:

– провести анализ методов аппроксимации Q-функций на основе МСП, применимых к электронным финансовым рынкам;

– разработать процедуру выбора и коррекции структуры МСП, используемой для аппроксимации Q-функций в RL-алгоритме;

– разработать структуру системы формирования торговых стратегий, основанных на методах машинного обучения с подкреплением и процедурах настройки аппроксимирующего МСП.

Аппроксимация Q-функций с применением многослойного персептрона

Применение МСП для аппроксимации Q-функции (коннекшионистское Q-обучение) имеет следующие преимущества:

– эффективное масштабирование для пространства входов большой размерности;

– возможность обобщения процедуры аппроксимации для больших и непрерывных пространств состояний;

– возможность распараллеливания вычислительного процесса аппроксимации.

Коннекшионистское Q-обучение подразумевает, что представление таблицы Q-функции заменяется моделью МСП, где состояния окружающей среды и Q- значения становятся входами и выходами МСП соответственно.

Важное значение имеет выбор конфигурации МСП и алгоритма его обучения. Для реализации процедуры вида (5) после выбора агентом определенного действия веса для соответствующих выходов должны быть обновлены. Например, если агент выбрал действие a_2 в состоянии s_i, должны быть обновлены все веса для выходов $Q(s_i, a_2)$.

В работе [3] предложено осуществлять коррекцию весов нейронной сети методом «обратного переигрывания» (backward replay). При использовании этого метода веса нейронной сети обновляются только при достижении системой поглощающего состояния. Очевидно, что это вызывает необходимость хранения всех пар «состояние-действие», которые встречаются в системе перед достижением поглощающего состояния. Частично преодолеть эту трудность можно, при-

меняя дисконтирование оценок Q-значений, уменьшая их значимость при обратном удалении от поглощающего состояния. Однако последовательность шагов, которые выполняет система, и, соответственно, получаемые оценки могут оказаться неоптимальными

В настоящей статье в качестве алгоритма обучения МПС, прменяемого для аппроксимации Q-функции в трейдинговой системе, предлагается использовать модифицированный алгоритм Левенберга-Марквардта. Нетривиальной задачей является при этом определение рациональной конфигурации МПС.

Выбор конфигурации многослойного персептрона

При построении модели МСП выбор ее начальной архитектуры определяется опытом разработчика и наличием априорной информации об аппроксимируемой функции. Обычно начальная конфигурация многослойного персептрона выбирается двухслойной с произвольным числом нейронов в слоях. После проведения экспериментов с различными конфигурациями сети выбирается та, которая дает минимальное значение функционала ошибки и требует меньших вычислительных затрат [4]. Для уменьшения времени получения приемлемой нейросетевой модели предлагается следующий подход. Пусть имеется набор нейросетевых моделей, содержащий N МСП с различными архитектурами. Обучение сетей производится на одном и том же наборе обучения $\tau = \{(\underline{x}_i, y_i) : i = \overline{1,n}\} \subset R^d \times R$. Вектор параметров сети обозначим $\underline{w}_j \in R^{p_j}$, а функцию отображения, реализуемую сетью – через $\eta_j(x, \underline{w}_j), j = \overline{1,N}$. В качестве критерия дискриминации нейросетевых моделей выберем среднеквадратичную ошибку минимум которой обеспечивается соответствующей настройкой параметров сети по алгоритму Левенберга-Марквардта.

$$e_j(\underline{w}_j) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \eta_j(\underline{x}_i, \underline{w}_i)], \qquad (6)$$

Пусть для МСП1 и МСП2 значения параметров $\underline{w}^a \in R^{p_1}$ и $\underline{w}^b \in R^{p_2}$ определяют положение локальных минимумов функций ошибки $e_1(\underline{w}^a)$ и $e_2(\underline{w}^b)$ соответственно. Предположим, что значения параметров сети $\underline{w}_1 \in R^{p_1}$ и $\underline{w}_2 \in R^{p_2}$, полученные в ходе обучения, достаточно близки к локальным минимумам, и можно использовать квадратичную аппроксимацию функции ошибки:

$$e_a(\underline{w}_1) = e_a + \frac{1}{2}\tilde{\underline{w}}_1^T H_a \tilde{\underline{w}}_1; \quad e_b(\underline{w}_2) = e_b + \frac{1}{2}\tilde{\underline{w}}_2^T H_b \tilde{\underline{w}}_2, \tag{7}$$

где $e_a = e_1(\underline{w}^a)$, $e_b = e_2(\underline{w}^b)$, $\tilde{\underline{w}}_1 = \underline{w}_1 - \underline{w}^a \in R^{p_1}$, $\tilde{\underline{w}}_2 = \underline{w}_2 - \underline{w}^b \in R^{p_2}$,

$H_a = \nabla^2 e_1(\underline{w}^a)$, $H_b = \nabla^2 e_2(\underline{w}^b)$.

Теперь критерий дискриминации нейросетевых моделей можно сформулировать следующим образом: МСП1 должна быть удалена из набора нейросетевых моделей если:

$$e_a - e_b = [e_a(\underline{w}_1) - e_b(\underline{w}_2)] - \frac{1}{2}[\tilde{\underline{w}}_1^T H_a \tilde{\underline{w}}_1 - \tilde{\underline{w}}_2^T H_b \tilde{\underline{w}}_2] > 0. \tag{8}$$

Таким образом, процедура дискриминации моделей МСП, выбираемых для коннекшионистского обучения, состоит в выполнении следующих операций:

– задаются: набор из N МСП различной архитектуры, максимальное число циклов обучения сети k, период дискриминации $m < k$;

– каждые m циклов обучения проверяются условия (8) и удаляются все МСП, для которых эти условия выполняются.

Критерием останова алгоритма является достижение максимального числа циклов обучения k.

Структура системы определения торговых стратегий

Общая структура системы выбора торговых стратегий с использованием предлагаемого подхода представлена на рис.1.

При пуске системы загруженные данные передаются в модуль анализа текущей ситуации на биржевом рынке, который выбирает наиболее эффективные индикаторы на текущее время и оценивает необходимость коррекции МСП, используемого для определения Q-функций. Когда новые данные поступают в систему с сервера, рассчитываются значения индикаторов и используются для определения текущего состояния окружающей среды. Затем, основываясь на текущем состоянии, модуль RL дает рекомендации для трейдера и обновления Q-значений с применением МСП.

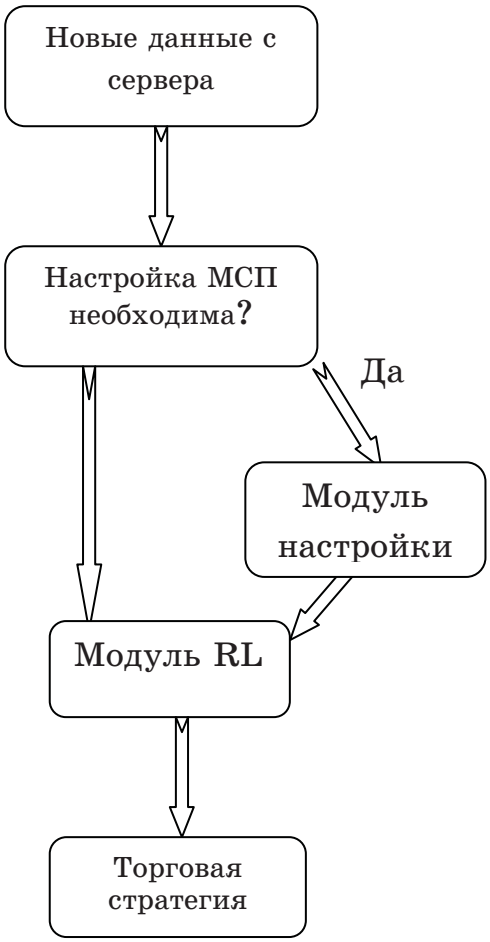

Рисунок 1 - Общая схема выбора стратегии

Результаты тестового моделирования

Программное обеспечение рассмотренной системы ориентировано в основном на работу для торговли на рынке FX. При тестировании трейдинговой системы были, в частности, использованы данные FX-рынка по суточному обменному курсу валют EUR/USD (www.cqg.com). Для оценки эффективности разработанных алгоритмических и программных средств биржевой торговли был выбран коэффициент Kp, характеризующий (в процентном выражении) отношение торговых операций системы с положительным исходом к общему числу проведенных операций. В результате применения модели МСП для RL-обучения в тестируемой системе был получен положительный эффект (увеличение Kp на 2%-4% для различных серий тестовых экспериментов) по сравнению с применением динамических Q-таблиц.

 ISSN 1562-9945

Выводы

Разработанная трейдинговая компьютерная система, основанная на использовании моделей МСП для аппроксимации Q-функций в RL-алгоритме, позволяет принимать эффективные решения по определению стратегий в трйдинговых системах. Результаты тестового моделирования программных модулей системы в онлайн-режиме подтверждают ее работоспособность.

ЛИТЕРАТУРА

1. 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.
2. 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.
3. Watkins S., Dayan P. Q-Learning./ S. Watkins, P. Dayan // In: Machine Laerning 8, Kluwer Academic Publisher, Boston. – 1992.– P. 279-292.
4. Руденко О.Г., Бодянский Е.В. Основы теории искусственных нейронных сетей./ О.Г. Руденко, Е.В. Бодянский// – Харьков: ТЕЛЕТЕХ, 2002. – 317с.