УДК 519.6:004

СТРАТЕГИИ ПРИНЯТИЯ РЕШЕНИЙ В НЕЧЕТКИХ СИСТЕМАХ С ПОДКРЕПЛЯЕМЫМ ОБУЧЕНИЕМ

С.Г. Удовенко, *д.тн., профессор*
Харьковский национальный университет радиоэлектроники
serhii.udovenko@nure.ua
Л.Э. Чалая, *к.тн., доцент*
Харьковский национальный университет радиоэлектроники
Larysa.chala@nure.ua

В докладе рассматривается применение методов подкрепляемого обучения и нечетких методов в компьютерных трединговых системах.

Udovenko S.G., Chala L.E. Strategies of making decision in fuzzy systems with reinforcement learning. In this paper are discussed the using of reinforcement learning and fuzzy methods to computing trading systems.

Ключевые слова: МАШИННОЕ ОБУЧЕНИЕ, СИГНАЛ ПОДКРЕПЛЕНИЯ, НЕЧЕТКАЯ СИСТЕМА, ТРЕЙДИНГОВАЯ СИСТЕМА.
Keywords: MACHINE LEARNING, SIGNAL OF REINFORCEMENT, FUZZY SYSTEM, TRADING SYSTEM.

В последнее время получили распространение управляемые стохастические системы, использующие метод обучения с подкреплением (reinforcement learning (RL)) [1]. Особенностью этого метода является наличие скалярного сигнала подкрепления, который получает агент из внешней среды и который характеризует эффективность системы в данный момент времени. Наиболее распространенным алгоритмом RL-обучения является алгоритм Q-обучения [1]. Для определения оптимальной стратегии здесь используется Q-функция, процедура обновления которой имеет следующий вид:

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

где α – действие, вызывающее переход среды из состояния s в состояние s'; γ – коэффициент нормирования.

Для задания Q-функций могут быть использованы N нейронных сетей типа «многослойный персептрон» (N – число действий α_i), каждая из которых аппроксимирует функцию $Q(s,\alpha_i)$ для действия α_i. После применения действия α_k в состоянии s разность $Q_{t+1}(s,\alpha_k) - Q_t(s,\alpha_k)$ между эволюцией качества для шагов t и $t+1$ может рассматриваться как сигнал ошибки. Для оптимизации ИНС используют алгоритм обратного распространения, минимизирующий следующую функцию:

$$E_t(s,\alpha_k) = \frac{1}{2}\left[Q_{t+1}(s,\alpha_k) - Q_t(s,\alpha_k)\right]^2. \quad (2)$$

В нечеткой версии такого представления для непрерывного пространства состояний и дискретных действий, именуемой Q-FUZ, функция качества реализуется нечеткой системой с N выходами. После выбора функций принадлежности задача обучения состоит в оптимизации правил вывода, позволяющих получить искомые значения s. Рассмотрим возможность расширения Q-FUZ представления для оптимизации нечетких правил вывода Такаги-Сугено (ТС) и его адаптации к задаче принятия решений в системе электронной биржевой торговли. Принцип работы предлагаемого модифицированного алгоритма принятия трейдинговых решений с нечетким RL-обучением (Q-FUZM) состоит в получении множества выводов для каждого нечеткого правила и ассоциации для каждого вывода функции качества, которая будет оцениваться с применением фиксированной функции принадлежности. При настройке по алгоритму Q-FUZM блок нечетких выводов мобильного робота должен корректировать выводы из правил ТС на основе сигналов подкрепления. Задача состоит в аппроксимации функции качества Q нечеткой функцией SIF (System Inference Fuzzy):

$$s \rightarrow y = \hat{Q} = \mathrm{SIF}(s). \quad (3)$$

Если выбрать нечеткую ТС-систему нулевого порядка (ТС0), такая функция определится правилами следующего вида: «Если $s = S_1$, то $y = c_1$; если $s = S_m$, то $y = c_r$», где m – число правил, а прототипы i-го правила S_i определяются как: «x_1 есть A_1^i и... и x_n есть A_n^i». В процессе обучения вывод по каждому правилу выбирается по средним значениям сигналов подкрепления $C_r(i) \in \{1...N\}$. В этом случае результирующий выход определяется как:

$$A(s) = \sum_{i=1}^{N} w_i(s) q\left[i, C_r(i)\right]. \qquad (4)$$

Рассматриваемый алгоритм одношагового нечеткого RL-обучения может быть использован при выборе стратегий в трейдинговой системе. В наиболее простом варианте агент-трейдер использует исходную базу данных типа TS0, определяющую возможные ситуации для желаемого поведения. При этом реализация алгоритма состоит в формировании множества выводов типа «импульс» для каждого правила и ассоциации с каждым выводом функции качества, которая оптимизируется во времени. Целью фазы обучения является определение набора выводов правил, максимизирующих среднее значение сигналов подкреплений. После выбора вывода выбранное действие будет реализовано и показатели качества будут адаптированы функцией полученного подкрепления и новой ситуацией.

Литература

1. Sutton R.S. Reinforcement Learning with Replacing Eligibility Traces / R.S. Sutton // Machine Learning. – 1996.– vol. 22, pp. 123-158.
2. Удовенко С.Г. Гибридные методы машинного обучения в системах управления динамическими объектами / С.Г. Удовенко, А.А. Гришко, Л.Э. Чалая. // Біоніка інтелекту. – 2012. –№1(78) – С.78-84.