УДК 004.94

МОДЕРНИЗИРОВАННАЯ СХЕМА Q – ОБУЧЕНИЯ

Е.А. ШУМКОВ

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

В работе представлен краткий обзор существующих методов реализации одного из подвидов обучения с подкреплением - Q-обучения. Рассмотрен вариант использования нейронных сетей в качестве аппроксиматора Q-таблицы и предложена модифицированная схема реализации Q-обучения.
Ключевые слова: обучение с подкреплением, Q-обучение, нейронные сети, обучение нейросетей.

Q – обучение является одним из основных направлений обучения с подкреплением [0,0,0**Ошибка! Источник ссылки не найден.**]. Алгоритм Q – обучения был предложен в 1989 году Воткинсом (Watkins) [0]. Данный алгоритм использует вместо функции ценности Q – функцию, аргументами которой являются и состояние и действие.

Несколько упростив и опустив математические выкладки, Q – обучение можно описать следующим образом. Пусть имеется агент (возможно программный), который действует в окружающей среде $\overline{S}$ и выполняет какие-либо действия $\overline{A}$ и получает вознаграждение R, т.е. на итерации t агент, оказавшись в определенном состоянии s_i, выполняет действие a_j и получает оценку своего действия в виде подкрепления r_t (возможно отрицательного)[1]. Обычно количество состояний среды (в которое можно включить и параметры состояния агента) намного больше возможных действий агента. При этом в реальных задачах агент не знает обо всех возможных состояниях среды и они появляются во время работы агента. В Q – обучении для сбора информации о результатах выполнения того или иного действия в определенном состоянии среды строится Q – таблица (см. таблица 1), в ячейках которой, обычно простым суммированием, накапливается информации о ценности того или

[1] Существует различная трактовка в присваивании индекса подкреплению – либо «следующее» подкрепление считать текущим, либо индексировать подкрепление как $t+1$. Этот момент важно учитывать в формуле определения ценности стратегии.

иного действия в определенной ситуации. При этом подразумевается, что вначале есть «исследовательский режим» работы агента, когда он многократно выполняет различные действия в максимальном количестве ситуаций и оценивает последствия, которые заносятся в Q - таблицу. Далее в режиме реальной работы, агент с помощью «жадного правила» выбирает с высокой вероятностью то действие, которое ранее приносило максимальное поощрение. В режиме реальной работы Q – таблица также обновляется. По сути, происходит накопление и корректировка опыта агента и этот опыт сохраняется в Q - таблице. Более того, если есть возможность прогнозировать в какое состояние агент попадет, выполняя определенное действие, то можно строить политику поведения агента (стратегию поведения). В Q – обучении, также как и в любом другом обучении с подкреплением, подразумевается, что агент должен выполнять целевую задачу и поступающее подкрепление должно не убывать (часто ставится цель – максимизировать суммарное подкрепление) [0]. Более подробно см., например [0,0].

Данный подход применяется во многих областях и задачах, где заранее неизвестна окружающая среда и ее поведение, либо существует большое количество факторов и учет их всех крайне затруднителен. Также очень перспективно применение Q – подхода в задачах с реальным самообучением.

Учитывая, что в реальных задачах, кроме того, что неизвестно количество состояний внешней среды (или она просто непрерывна) и Q – таблица имеет большой размер, то часто для аппроксимации Q – таблицы используют искусственные нейронные сети. Выделим методы Q – обучения без нейронных сетей: Q–обучение с использованием Хеммингова расстояния; Q–обучение с статической кластеризацией. С использованием нейронных сетей: Q–обучение на основе послойно - полносвязных нейронных сетей прямого распространения; Q–обучение на основе сети Кохонена; Dyna–Q; Competitive MLP (на основе конкуренции)[2]. Также есть нечеткое Q–обучение FQL (Fuzzy Q-Learning) [0,0,0,0].

[2] Обзор методов Q-обучения см. также по Интернет – адресу: http://www.shumkoff.ru/iis/q-learning.php

Использовать нейронные сети для Q – обучения можно несколькими способами. Один из самых простых и используемых показан на Рисунке 1. Но с помощью нейронной сети Кохонена получается более простая схема, но возможно с большим количеством нейронов в слое Кохонена. Алгоритм работы Q – обучения с использованием нейронной сети можно посмотреть в [0].

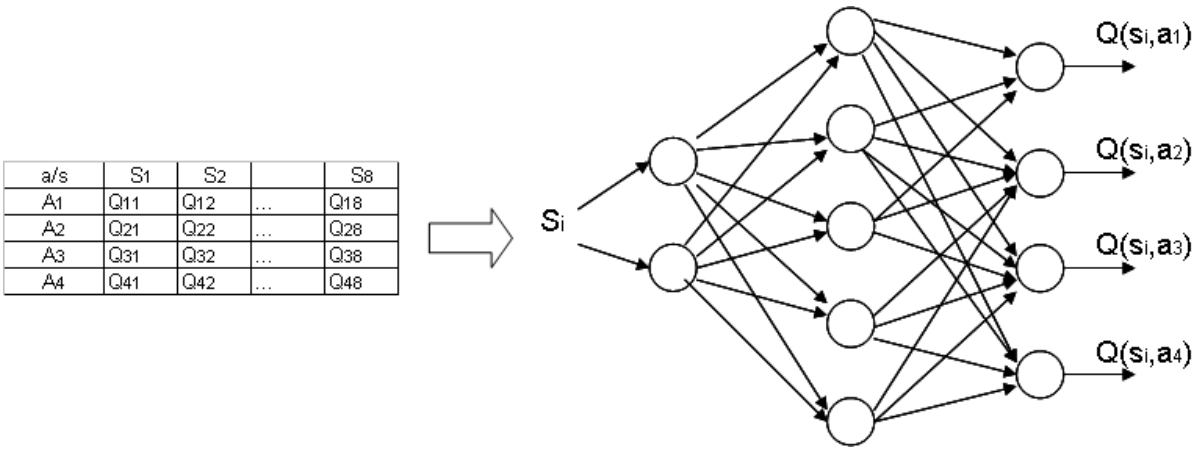

Рисунок 1. Использование нейронной сети для Q-обучения

Как было указано выше - существует несколько разновидностей Q – обучения, вплоть до использования дифференциальных уравнений для расчета Q – фактора. Предложим более общую схему реализации Q – обучения, несколько модернизировав основной подход (см. Рисунок 2). Основные компоненты – Q – матрица, внешняя среда, объект управления сохраняются, но отдельно выделяются блоки: расчета подкрепления, идентификации состояния, обновления Q – матрицы, поиска действия и блок исследования среды. Блок идентификации состояния необходим также для определения новых состояний среды. В случае если найдено новое состояние, то блок обновления Q – матрицы расширяет Q – матрицу и дает команду на блок исследования среда на прорабатывание нового состояния.

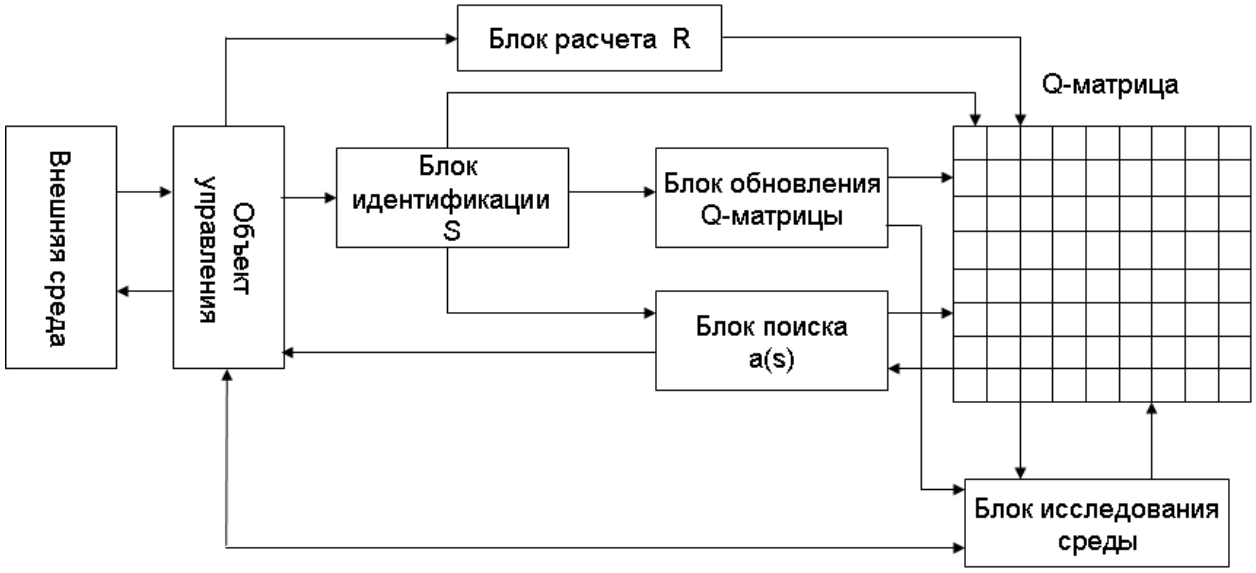

Рисунок 2. Модернизированная схема Q-обучения

В случае, если аппроксиматором Q – матрицы является нейронная сеть, то необходимо также добавить блоки, связанные с работой нейронной сети (блоки поиска структуры и обучения нейронной сети).

Предложенная схема выделяет механизм исследования внешней среды и ее новых состояний в отдельные компоненты. Выделение механизма исследования внешней среды позволяет более детально прорабатывать новые состояния и добавляет возможность планировать «исследовательский режим». Введение блока идентификации позволяет детально анализировать состояние внешней среды и объекта управления и выявлять новые состояния внешней среды. Предложенная схема упрощает конечную реализацию системы управления с использованием Q-обучения для разработчика и позволяет использовать различные механизмы, как идентификации состояний, так и различные методы Q-обучения.

ЛИТЕРАТУРА

1. Ключко В.И., Власенко А.В., Стасевич В.П., Шумков Е.А. Нейросетевые технологии с подкреплением. Краснодар: Изд. ФГБОУ ВПО «КубГТУ», 2012. – 154 с.

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

3. Шумков Е. А. Система поддержки принятия решений предприятия на основе нейросетевых технологий: дис. канд. техн. наук. - Краснодар, КубГТУ, 2004. -158 с.

4. Watkins C. J., Dayan P. Q – learning. Machine Learning, 8:279 – 292, 1992.

5. Watkins C.J.C.H. Learning from Delayed Rewards, Ph.D. Thesis, University of Cambridge, England, 1989

6. Wiering M, Schmidhuber J. HQ – learning. Adaptive behavior, 6(2):219 – 246, 1998.

REFERENCES

1. Kluchko V.I., Vlasenko A.V., Stasevich V.P., Shumkov E.A. Neural network technology with reinforcements. Krasnodar: Pub. FSEI of HPE “KubSTU”, 2012. – 154 p.

2. Satton R.S., Barto E.G. Reinforcement learning. Trans. from english. M.: BINOM: Knowledge Lab. 2012. 399 pp.

3. Shumkov E.A. Making support system of the enterprise solutions based on neural network technology. Thesis fo the degree of PhD (tech.). Krasnodar: KubSTU. 2004. 158 p.

4. Watkins C. J., Dayan P. Q – learning. Machine Learning, 8:279 – 292, 1992.

5. Watkins C.J.C.H. Learning from Delayed Rewards, Ph.D. Thesis, University of Cambridge, England, 1989

6. Wiering M, Schmidhuber J. HQ – learning. Adaptive behavior, 6(2):219 – 246, 1998.

UPGRADED CIRCUIT Q - LEARNING

E.A. SHUMKOV

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

The paper provides an overview of existing methods for the implementation of one of the subspecies of reinforcement learning - Q-learning. Consider using neural networks as approximator Q-table and offered a modified scheme of the Q-learning.
Key words: reinforcement learning, Q-learning, neural networks, neural network training