International Journal of Advance Research and Innovation
Vol. 8(4), Oct-Dec 2020, pp. 48-51
Doi: 10.51976/ijari.842009

Article Info

Received: 29 Jul 2020 | Revised Submission: 20 Oct 2020 | Accepted: 28 Oct 2020 | Available Online: 15 Dec 2020

A New Model of Reinforcement Learning, Algorithms

Vijay Bhandari, Arpana Bhandari**, Ritu Srivastava*** and Kapil Chaturvedi*****

ABSTRACT

RL doen't need prior knowledge, it can autonomously get optional policy with the knowledge obtained by trial-and-error and continuously interacting with dynamic environment. Its characteristics of self- improving and online learning make reinforcement learning become one of intelligent agent's core technologies. In this article, we firstly literature the model and theory of reinforcement learning. Then, we roundly present the main reinforcement learning algorithms, including Sarsa, temporal difference, Q-learning and function approximation. Finally, we briefly introduce some applications of reinforcement learning and point out some future research directions of reinforcement learning.

Keywords: *Reinforcement learning; SARSA; Temporal difference; Q-learning; Function approximation.*

1.0 Introduction

RL has a very long history, but it is not until the late 80s and early 90s that reinforcement learning technology obtains the wide research and application in some fields such as artificial intelligence, machine learning, automatic control and so on [1] .

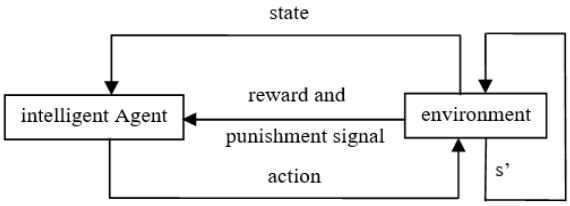
Reinforcement learning is an important machine learning method [2], its learning technology is divided into three types: non-supervised learning, supervised learning and reinforcement learning. Reinforcement learning is an online learning technology which is different from supervised learning and non-supervised learning. The reinforcement signal provided by the environment in reinforcement Learning is to make a kind of appraisal to the action quality of intelligent Agent, but not tell intelligent Agent how to generate the correct action. Because the external environment provides a little information, intelligent Agent must depend on its own experience to learn, by which intelligent Agent obtains the appropriate appraisal value of the environment state and revises own action strategy to adapt to the environment.

In this article, we first literature the model and theory of reinforcement learning. Then, we roundly present the main reinforcement learning algorithms, including Sarsa, temporal difference, Q-learning and function approximation. Finally, we briefly introduce some applications of reinforcement learning and point out some future research directions of reinforcement learning.

2.0 Reinforcement Learning Model

The basic model of reinforcement learning is shown in figure 1. Intelligent Agent can perceive the environment and choose an action to obtain the biggest reward value by continuously interacting with the environment. The interactive interface of intelligent Agent and environment includes action, reward and state.

Figure 1: The Basic Model of RL



When each time reinforcement learning system interacts with the environment, the system first accepts the input of environment state s, and then the output of action a acts the environment according to

**Corresponding author; Computer Science Department,SIRTS Bhopal (E-mail: vijayhomee@gmail.com)*
***Computer Science Department,SIRTS Bhopal (E-mail: arpanabhandari08@gmail.com)*
****Dean, SIRTS Bhopal*
*****Professor, SIRTS Bhopal*

the internal inference mechanism. Finally, the environment changes to new state s' after accepting the action. The system accepts the input of the new state s' and obtains the rewards and punishment signal r of environment for the system. Reinforcement learning system's goal is to learn an action strategy š : S↔ A, the strategy enables the action of the system choice to obtain the largest cumulative reward value of environment [3], it can be defined as formula∧1∨. Where Û is discount factor. The basic theory of reinforcement learning technology is: If a certain system's action causes the positive reward of the environment, the system generating this action lately will strengthen the trend, this is a positive feedback process; otherwise, the system generating this action will diminish this trend.

$$\sum_{i=0}^{\infty} \gamma^{i} r_{t+i}$$

If the environment is Markov, the interaction between the system and the environment may be regarded as Markov decision-making process (MDP), its definition canstate s changes to state s' through action a, is a probability obtained by system when the environment state s changes to state s' through action a.

Because P function and R function is unknown in reinforcement learning environmental mode, the system can only choose the strategy according to the instantaneous reward obtained by trial-and-error each time, but it must consider the uncertainty of the environmental model and long-term goals during selecting action strategies process, therefore, the value function between the strategy and the instantaneous reward can be defined as formula (2), which is used for the choice of the strategy. This formula reflects that the system can obtain expected cumulative reward discount sum if the system follows the strategy however, in fact, reinforcement learning algorithm often uses the iteration approximation method to estimate value function.

$$V^{\pi}(s) \leftarrow \sum_{a \in A(s)} \pi(s,a) \sum_{s'} P^{a}_{ss'} [R^{a}_{ss'} + \gamma V^{\pi}(s')] \quad (2)$$

3.0 Reinforcement Learning Algorithm

Typical reinforcement learning method based on the MDP model includes two kinds: One is model-based method such as SARSA algorithm, in which RL first learns the model knowledge, and then derives the optimal strategy from model knowledge. The other is model- irrelevant method such as the TD algorithm and the Q- learning algorithm, in which RL directly calculates the optimal strategy without model knowledge.

3.1 Sarsa Algo

Sarsa algorithm was proposed by Summery and Niranjan in 1994[4]. In this algorithm, in order to achieve the purpose of the maximum cumulative discount function, the optimal Q of the state-action needs to satisfy formula (3) by the action appraisal function or Q function.

$$Q^{*}(s,a) = \sum_{s \in S} P^{a}_{ss'} [R^{a}_{ss'} + \gamma \max_{a \in A} Q^{*}(s',a)] \quad (3)$$

Sarsa algorithm selects Q value iteration method. According to experience (st, rt, st+1) in each learning step, RLS algorithm can be defined as formula (4).

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \quad (4)$$

3.2 Temporal difference algorithm

TD algorithm was proposed by Sutton in 1988[5]. This algorithm solves the forecast problem according to the time series in the reinforcement learning, and proves its astringency in the simplified conditions, later many scholars carry on the analysis and the improvement of the state. In TD(0) algorithm, value function iterative formula can be defined as follows:

$$V(s_t) \leftarrow (1-\alpha) V(s_t) + \alpha[R^{a}_{ss'} + \gamma V(s_{t+1})] \quad (5)$$

Where, $V(s_t)$ is the reward sum obtained by RLS under the environment state S_t, $V(s_{t+1})$ is reward discount sum obtained by RLS under the environment state St+1. Because TD(0) algorithm converges slowly, an effective method makes the instantaneous reward value be any step back, which is called TD(Ȝ) algorithm. Its formula can be defined as follows:

$$V(s_t) \leftarrow V(s_t) + \alpha[R^{a}_{ss'} + \gamma V(s_{t+1}) - V(s_t)] \cdot e(s) \quad (6)$$

Where, e(s) is the degree of election under the state s, its formula can be defined as follows: if $s=s_k$, then $s_s(k)=1$; else $s_s(k)=0$

$$e(s) = \sum_{k=1}^{t} (\lambda\gamma)^{t-k} . s(k_s) \quad (7)$$

Recursive algorithm can be defined as follows:

$$e(s) = \begin{cases} \gamma\lambda e(s) + 1 & ;s = s_k \\ \lambda\lambda e(s) & other \end{cases} \quad (8)$$

However, to large-scale MDP or continual spatial MDP, TD(Ɜ) is impossible traversal all state space and updates all the state in each time step, therefore, it is difficult to guarantee timeliness.

3.3 Q-learning algorithm

Q-learning algorithm was proposed by Watkins and others [6]. Each state-action corresponds to a related Q value and chooses an action according to Q value in Q- learning algorithm. Q value is defined as follows: RLS carries out the related action and obtains the reward sum Q*(s,a)according to a certain strategy Ê, the basic equation can be defined as follows:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s \in S} P^a_{ss'} . V(s', \pi^*) \quad (9)$$

$$V(s', \pi^*) = \max_{a \in A} Q^*(s,a) \quad (10)$$

$$\pi^*(s,a) = \arg\max_{a \in A} Q^*(s,a) \quad (11)$$

Where, R(s,a) is the instantaneous repayment obtained by performing action a under the state s. Because P function and R function is unknown, R(s,a) obtains state- action value Q(s, a∨through the way of the successive iteration. The initial Q value can be given willfully, thus Q value is updated according to formula (12) after performing an action each time.

$$Q_t(s,a) = \begin{cases} (1-\alpha)Q_{t-1}(s,a) + \alpha[R(s,a) + \gamma \max_{a \in A} Q(s',a)] \\ Q_{t-1}(s,a) \end{cases}$$

$$;s = s_k, a = a_t$$

$$other \quad (12)$$

Q-learning is astringency. Through exploring the state space unceasingly, the Q value trends towards Q* progressively, however. When the state space and decision space are large, Q-learning is impossible traversal all state space. Therefore, it lacks a certain generalization.

3.4 Function approximation algorithm

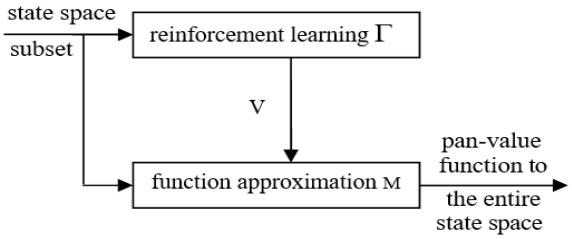
In the RL, function approximation is that the mapping relations S h A ė R ' S h A ė P may use the parameterization function to approach, thus it solves the problems that the typical generalization ability of RL value function is not strong in the continuous space MDP. Structural model is shown in Fig. 2. Supposed the initial value of value function is V0, operator is ī, objective function is V, approximation function is V', function approximation operator is M:V ė V', and thus the sequence of value function generated in the learning process can be defined as formula (13):

$$V_0, M(V_0), \Gamma(M(V_0)), M(\Gamma(M(V_0))), \Gamma(M(\Gamma(M(V_0)))) \quad (13)$$

Iterative formula can be defined as follows:

$$Q(s,a) \leftarrow (1-\alpha)M(Q(s,a)) + \alpha(\gamma + \max V'(s', a')) \quad (14)$$

Figure 2. Function Approximation RL Structure



At present function approximation reinforcement learning method usually uses the supervised learning method, such as state cluster [7], decision tree [8], function interpolation [9] and artificial neural networks [10] and so on; the artificial neural network is the hot spot.

4.0 Reinforcement Learning Application

Reinforcement learning is mainly used in process control, dispatch management, robot control, game competition and information retrieval etc, the widest application is the field of intelligent control and intelligent robot. In the process control field, the typical application example is the inverted pendulum control system [11-14] and ARCHON system [15]; In the dispatch management, the most successful application is Crites and Barton's elevator scheduling problem, they apply a step reinforcement learning algorithm to the operation scheduling[16] including 4 lifts and 10 floors; In the game competition, the most successful application is Backgammon's chess game invented by Gerry through using instantaneous difference algorithm, it uses the TD error to train three BP neural network and achieves a very high level[17]; In the robot field, HeePakBeen uses the fuzzy logic and reinforcement learning to achieve

land- based mobile robot navigation system[18], Wilfriedllg uses reinforcement learning to make hexapod insect robot coordinate their six-legged actions[19], Christopher uses reinforcement learning to control the robot arm action, Mataria uses an improved Q-learning algorithm to make 4 robots perform the foraging task[20]; The information retrieval is mainly used in the Internet-related information collection and information filtering, namely, the user can inquire most cared information from the mass of Web information.

5.0 Conclusions

In recent years, reinforcement learning research has made the breakthrough progress, however, because of the real world complexity, at present there are still many problems to need to further study and solve: First improve the study speed. The slow study speed is a very obvious problem in the Agent field, and seriously affects practical application of multi-agent study. Therefore, how to unify other machine learning techniques (such as neural networks, sign learning techniques and so on) to help the multi-agent system to speed up the learning is an important direction of reinforcement learning research and application. Second, multi-agent coordinated mechanism.

The multi-agent system cooperation needs to introduce the coordinated mechanism to make each agent member's choice maintain consistent.

References

[1] Rummery G, Niranjan M. On-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Cambridge University Engineering Department, 1994.

[2] Sutton R S. Learning to Predict by th methods of temporal differences. Machine Learning, 1988, 3: 9~44.

[3] Watkins C掂Q-Learning [J]掂Machine Learning掌1992掌8 (3)拘279-292掂

[4] Singh S, Jaakkola T, Jordan M I. Reinforcement learning with soft state aggregation. In: Tesauro G, Touretzky D, Advances in Neural Information Processing Systems, 7. Morgan Kaufmann: MIT Press, 1995.361~368.

[5] Crites R H, Barto A G. Elevator group control using multiple reinforcement learning agents. Machine Learning, 1998,33(3),235~262.

[6] McCallum A K. Reinforcement learning with selective perception and hidden State Ph. D. dissertation]. Department CS, University Rochester,1996.

[7] Sutton R S. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In: Touretzky D, Mozer M, Hasselno M, Advances in Neural Information Processing Systems, B. NY: MIT Press, 1996 1038~1044.

[8] Anderson C W. Learning to control an inverted pendulum using neural network [J] . IEEE Control System Magazine , 1989 , 30 (4) :31 – 36.

[9] Whitley D ,Dominic S ,Das R and Aanderson C W. Genetic reinforcement learning for neurocontrol problems [J]. Machine Learning, 1993 ,13 :259 – 284.

[10] Berebji H R. Learning and tuning fuzzy logic controllers through reinforcements [J]. IEEE Trans . on Neural Networks , 1992 , 3 (5) :724 – 740.

[11] Khan E. Reinforcement control with unsupervised learning [A]. Int.Joint Conference on Neural Network [C] ,Beijing ,1992 ,88 – 93.

[12] N.R.Jennings,J.Corera,I.Laresgoti,.H.mamdani,F.Perriolat,P.Skare k and L.Z.Varga.using ARCHON to develop real-world DAI applications for electricity transportation management and Particle acceleration control[J]. IEEE Exert, 1996, 11(6): 60-88, December.

[13] Crites R H and Barto A G. Improving elevator performance using reinforcement learning[A]. In: Touretzky D S ,Mozer M C , and M E H. Advances in Neural Information Processing Systems [M]. Cambridge,MA : The MIT Press ,1995 ,1017 – 1023.