ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/334451667

Quantum Reinforcement Learning

Article *in* Lecture Notes in Computer Science · January 2005

CITATIONS
25

READS
341

3 authors:

Daoyi Dong
Australian National University
323 PUBLICATIONS **5,564** CITATIONS
SEE PROFILE

Chunlin Chen
Nanjing University
186 PUBLICATIONS **3,223** CITATIONS
SEE PROFILE

Zonghai Chen
University of Science and Technology of China
338 PUBLICATIONS **6,041** CITATIONS
SEE PROFILE

All content following this page was uploaded by Zonghai Chen on 14 July 2019.

The user has requested enhancement of the downloaded file.

Quantum Reinforcement Learning

Daoyi Dong, Chunlin Chen, and Zonghai Chen*

Department of Automation, University of Science and Technology of China,
Hefei, Anhui 230027, People's Republic of China
{dydong, clchen}@mail.ustc.edu.cn
chenzh@ustc.edu.cn

Abstract. A novel quantum reinforcement learning is proposed through combining quantum theory and reinforcement learning. Inspired by state superposition principle, a framework of state value update algorithm is introduced. The state/action value is represented with quantum state and the probability of action eigenvalue is denoted by probability amplitude, which is updated according to rewards. This approach makes a good tradeoff between exploration and exploitation using probability and can speed up learning. The results of simulated experiment verified its effectiveness and superiority.

1 Introduction

Learning methods are generally classified into supervised, unsupervised and reinforcement learning (RL). Supervised learning requires explicit feedback provided by input-output pairs and gives a map from input to output. And unsupervised learning only processes on the input data. However, RL uses a scalar value named reward to evaluate the input-output pairs and learns by interaction with environment through trial-and-error. Since 1980s, RL has become an important approach to machine intelligence [1-4], and is widely used in artificial intelligence due to its good performance of on-line adaptation and powerful leaning ability of complex nonlinear system [5, 6]. But there are some difficult problems in applications, such as very slow learning speed, especially for the curse of dimensionality problem when the state-action space becomes huge. Although in recent years many researchers have proposed all kinds of methods to speed up learning, few satisfactory successes were achieved.

On the other hand, quantum technology, especially quantum information technology is rapidly developing in recent years. The algorithm integration, which is inspired by quantum characteristics and quantum algorithms, will not only improve the performance of existing algorithms on traditional computers, but also promote the development of relative research areas such as quantum computer and machine learning. Considering the essence of computation and algorithms, we propose a quantum reinforcement learning (QRL) algorithm inspired by the state superposition principle. Section 2 contains the prerequisite and problem description. In section 3, a novel QRL algorithm is proposed. Section 4 describes the simulated experiments and analyzes their results. Conclusion and remarks are given in section 5.

* corresponding author.

L. Wang, K. Chen, and Y.S. Ong (Eds.): ICNC 2005, LNCS 3611, pp. 686 – 689, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Reinforcement Learning (RL)

Standard framework of RL is based on discrete time Markov decision processes [1]. In RL, the agent is to learn a policy $\pi : S \times \cup_{i \in S} A_{(i)} \to [0,1]$, so that expected sum of discounted reward of each state will be maximized:

$$V_{(s)}^{\pi} = E\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \mid s_t = s, \pi\} = \sum_{a \in A_s} \pi(s,a)[r_s^a + \gamma \sum_{s'} p_{ss'}^a V_{(s')}^{\pi}] \quad (1)$$

$$V_{(s)}^{*} = \max_{a \in As}[r_s^a + \gamma \sum_{s'} p_{ss'}^a V_{(s')}^{*}] \quad (2)$$

$$\pi^{*} = \arg\max_{\pi} V_{(s)}^{\pi}, \qquad \forall s \in S \quad (3)$$

where $\gamma \in [0,1]$ is discounted factor, $\pi(s,a)$ is the probability of selecting action a according to state s under policy π, $p_{ss'}^a = \Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\}$ is probability for state transition and r_s^a is expected one-step rewards.

As for state-action pair value $Q(s,a)$:

$$Q_{(s,a)}^{*} = r_s^a + \gamma \sum_{s'} p_{ss'}^a \max_{a' \in As} Q_{(s',a')}^{*} \quad (4)$$

Let η be the learning rate, and the 1-step update rule of Q-learning is:

$$Q(s_t, a_t) \leftarrow (1-\eta) Q(s_t, a_t) + \eta(r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a')). \quad (5)$$

3 Quantum Reinforcement Learning (QRL)

In quantum information technology, information unit (qubit) is represented with quantum state and qubit is an arbitrary superposition state of two-state quantum system:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (6)$$

where α and β are complex coefficients and satisfy $|\alpha|^2 + |\beta|^2 = 1$. $|0\rangle$ and $|1\rangle$ correspond to logic states 0 and 1. $|\alpha|^2$ and $|\beta|^2$ represent the occurrence probabilities of $|0\rangle$ and $|1\rangle$ respectively when this qubit is measured. The value of classical bit is either Boolean value 0 or value 1, but qubit can simultaneously store 0 and 1, which is the main difference between classical and quantum computation.

Let N_s and N_a be the number of states and actions, then choose numbers *m* and *n*, which satisfy $N_s \le 2^m \le 2N_s$ and $N_a \le 2^n \le 2N_a$. And use *m* and *n* qubits to represent state set $S = \{s\}$ and action set $A = \{a\}$:

$$s : \left[\begin{array}{c|c|c|c} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & & b_m \end{array}\right], \text{ where } |a_i|^2 + |b_i|^2 = 1, \ \ i = 1,2,...m$$

$$a:\begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_n\\ \beta_1 & \beta_2 & & \beta_n\end{bmatrix}, \text{ where } |\alpha_i|^2+|\beta_i|^2=1,\ \ i=1,2,...n$$

Thus they may lie in superposition state:

$$|s^{(m)}\rangle=\sum_{s=00\cdots0}^{11\cdots1}C_s|s\rangle,\ \ |a^{(n)}\rangle=\sum_{a=00\cdots0}^{11\cdots1}C_a|a\rangle \tag{7}$$

where C_s and C_a are complex numbers.

The mapping from states to actions is $f(s)=\pi:S\to A$, and we will get:

$$f(s)=|a_s^n\rangle=\sum_{a=00\cdots0}^{11\cdots1}C_a|a\rangle \tag{8}$$

$|C_a|^2$ denotes the occurrence probability of $|a\rangle$ when $|a_s^n\rangle$ is measured.

The procedural form of QRL is described as follows.

Initialize $|s^{(m)}\rangle=\sum_{s=00\cdots0}^{11\cdots1}C_s|s\rangle$, $f(s)=|a_s^n\rangle=\sum_{a=00\cdots0}^{11\cdots1}C_a|a\rangle$ *and* $V(s)$ *arbitrarily*

Repeat (for each episode)

For all states $|s^{(m)}\rangle=\sum_{s=00\cdots0}^{11\cdots1}C_s|s\rangle$ *:*

(1) Observe $f(s)$ *and get* $|a\rangle$ *;*

(2)Take action $|a\rangle$, *observe next state* $|s'^{(m)}\rangle$, *reward* r

Then update: $V(s)\leftarrow V(s)+\alpha(r+\gamma V(s')-V(s))$

$$C_a\leftarrow e^{\lambda(r+V(s'))}C_a$$

Until for all states $|\Delta V(s)|\le\varepsilon$.

4 Simulation Experiments

To evaluate QRL algorithm in practice, consider the typical rooms with corridor example, gridworld environment of four rooms and surrounding corridors as shown in Fig. 1. From any state the robot (or agent) can perform one of four primary actions: up, down, left and right, and actions that would lead into a blocked cell are not executed. The task is to find an optimal policy which will let the robot move from *S*(4,4) to *G*(8,8) in this $13\times13(0\sim12)$ grid world with minimized cost. In QRL, the action selecting policy is obviously different from traditional RL algorithms, which is inspired by the collapse theory of quantum measurement. And probability amplitudes $|C_a|^2$ (initialized uniformly) are used to denote the probability of an action.

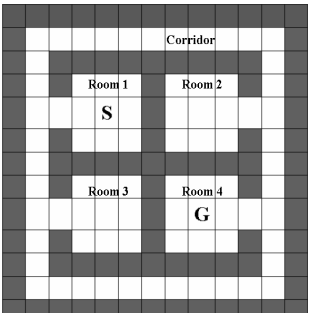

Fig. 1. Rooms with corridor

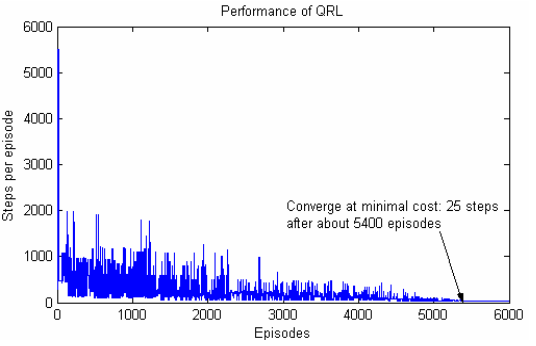

Fig. 2. Performance of QRL

Result and analysis. Learning performance for QRL is plotted in Fig. 2. At the beginning phase this algorithm learns extraordinarily fast, and then steadily converges to the optimal policy that costs 25 steps to the goal G. The results show that QRL algorithm excels other RL algorithms in two main aspects: (1) Action selecting policy makes a good tradeoff between exploration and exploitation. (2) Updating is carried through parallel, which will be much more prominent when practical quantum apparatus comes into use instead of been simulated on traditional computers.

5 Conclusion and Future Work

According to the existing problems in RL algorithms such as tradeoff between exploration and exploitation, low learning rate, QRL is proposed based on the concepts and theories of quantum computation. The results of simulated experiments verified the feasibility of this algorithm and showed its superiority for learning optimal problems with huge state space. With the development of quantum computation theory, the combining of traditional learning algorithms and quantum computation methods will make great change in many aspects such as representation and learning mechanism.

References

1. Sutton, R., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA (1998)
2. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont, MA (1996)
3. Sutton, R.: Learning to Predict by the Methods of Temporal Difference. Mach. Learn. 3 (1988) 9-44
4. Watkins, C., Dayan, P.: Q-learning. Mach. Learn. 8 (1992) 279-292
5. Beom, H.R., Cho, H.S.: A Sensor-based Navigation for a Mobile Robot Using Fuzzy Logic and Reinforcement Learning. IEEE Trans. Syst. Man. Cyc. 25 (1995) 464 -477
6. Smart, W.D., Kaelbling, L.P. Effective Reinforcement Learning for Mobile Robots. Proceedings of the IEEE Int. Conf. on Robotic. Autom. (2002)

View publication stats