基于云推理模型的深度强化学习探索策略研究

李晨溪*① 曹 雷① 陈希亮① 张永亮① 徐志雄① 彭 辉① 段理文②
①(解放军理工大学指挥信息系统学院 南京 210007)
②(浙江大学机械工程学院 杭州 310027)

摘 要：强化学习通过与环境的交互学得任务的决策策略，具有自学习与在线学习的特点。但“交互试错”的机制也往往导致了算法的运行效率较低、收敛速度较慢。知识包含了人类经验和对事物的认知规律，利用知识引导智能体(agent)的学习，是解决上述问题的一种有效方法。该文尝试将定性规则知识引入到强化学习中，通过云推理模型对定性规则进行表示，将其作为探索策略引导智能体的动作选择，以减少智能体在状态-动作空间探索的盲目性。该文选用 OpenAI Gym 作为测试环境，通过在自定义的“CartPole-v2”中的实验，验证了提出的基于云推理模型探索策略的有效性，可以提高强化学习的学习效率，加快收敛速度。

关键词：云推理；深度强化学习；知识；探索策略
中图分类号： TP18 **文献标识码**： A **文章编号**：1009-5896(2018)01-0244-05
DOI: 10.11999/JEIT170347

Cloud Reasoning Model-based Exploration for Deep Reinforcement Learning

LI Chenxi① CAO Lei① CHEN Xiliang① ZHANG Yongliang①
XU Zhixiong① PENG Hui① DUAN Liwen②
①(*Institute of Command Information System*, *PLA University of Science and Technology*, *Nanjing* 210007, *China*)
②(*College of Mechanical Engineering*, *Zhejiang University*, *Hangzhou* 310027, *China*)

Abstract: Reinforcement learning which has self-improving and online learning properties gets the policy of tasks through the interaction with environment. But the mechanism of “trial-and-error” usually leads to a large number of training episodes. Knowledge includes human experience and the cognition of environment. This paper tries to introduce the qualitative rules into the reinforcement learning, and represents these rules through the cloud reasoning model. It is used as the heuristics exploration strategy to guide the action selection. Empirical evaluation is conducted in OpenAI Gym environment called “CartPole-v2” and the result shows that using exploration strategy based on the cloud reasoning model significantly enhances the performance of the learning process.

Key words: Cloud reasoning; Deep reinforcement learning; Knowledge; Exploration strategy

1 引言

作为解决序贯决策问题的机器学习方法，强化学习(RL)[1]通过与环境的不断交互学得策略，使得累积奖赏能够最大。传统的解决决策问题的方法(如专家系统、贝叶斯网络、影响图、决策图等方法)，一般都需要针对具体问题进行建模，在建模的诸多环节中引入了一定的主观因素，这些因素的准确性和合理性在很大程度上影响着决策的质量。强化学习则通过智能体(agent)与任务环境(environment)的“交互试错”直接学得策略，不需要人为地构建推理模型，同时对样本数据(带标签的样本)的需求也较小，因而具有更强的适用性和通用性。近年来逐步兴起的深度强化学习[2,3](DRL)，利用深度神经网络强大的非线性表示能力，实现了端到端的学习，在诸如游戏、控制等领域取得了突破性进展。但强化学习由于试错而导致的学习效率低的问题仍然存在。如何在有限的计算资源下，引导智能体高效地探索未知空间，解决好强化学习探索与利用之间的矛盾，是我们面临的一个重要难题。

收稿日期：2017-04-18；改回日期：2017-09-30；网络出版：2017-11-01
*通信作者：李晨溪 streamorning@qq.com
基金项目：中电集团重点预研基金(6141B08010101)，中国博士后科学基金(2015T81081, 2016M602974)，江苏省自然科学青年基金(BK20140075)
Foundation Items: The Advanced Research of China Electronics Technology Group Corporation (6141B08010101), China Postdoctoral Science Foundation (2015T81081, 2016M602974), The Jiangsu Natural Science Foundation for Youths (BK20140075)

深度强化学习算法在最初普遍采用了简单的探索策略，如ε-greedy 贪心策略、Boltzmann 探索策略等。2016 年以来，研究者对深度强化学习的探索策略进行了大量的研究。Osband 等人[4]提出了 Bootstrapped DQN 方法，Bellemare 等人[5]提出了基于伪计数的内在动机方法，Houthooft 等人[6]提出了基于环境模型信息增益的 VIME 探索策略。

托马斯•达文波特曾指出，“知识是行动和决策的依据和指南”[7]。利用已有的经验知识引导智能体探索状态动作空间，是加快强化学习收敛速度，学得有效策略的一种重要方法。根据使用的知识类型的不同，可以将现有的方法分为：使用控制性知识、使用实例性知识和使用规则性知识的探索策略。

在强化学习过程中，使用控制性知识主要是指利用一些算法(如 A*算法等)对学习中已产生的轨迹数据进行学习与归纳，进而对智能体的动作选择进行指导。Santos 等人[8]基于 Dyna-Q 框架提出了 Dyna-H 算法，使用 A*算法作为启发式函数，为算法中规划部分的动作选择提供指导。

基于实例性知识的探索策略主要指利用已有的相似案例或轨迹样本，结合案例推理[9]或迁移学习的方法构造启发式函数，改进探索策略。

基于规则性知识的探索策略是指利用规则性的先验领域知识，以“IF-THEN”形式表示的命题规则(propositional rule)或一阶规则(first-order rule)来引导智能体的动作选择。Kuhlmann 等人[10]结合语料库，将自然语言形式的先验知识转化为形式化的 IF-THEN 规则，而后在不同的状态下，对智能体的动作选择提出指导性建议或者禁止性建议。上述工作均是采用基于定量数值的精确规则。比如，在足球机器人比赛中的规则：“如果在 8 m 范围内没有对方球员，则智能体继续控球”。而在实际应用中，规则性知识往往更多的是具有不确定性的定性知识，这也是上述方法的局限性。

针对目前基于规则性知识的探索策略多是运用精确(定量)规则的现状，本文提出了一种基于云推理模型的深度强化学习探索策略，依据云推理模型对定性的规则知识进行表达，以真实体现客观世界知识的模糊性与随机性，引导强化学习的动作选择。

云推理模型[11]自 1998 年由李德毅院士提出以来，在不确定性知识推理方面得到了广泛的应用。其以云模型正向云发生器为基础建立前件云与后件云，构建规则发生器，实现对单条件或多条件规则知识的表达。

本文将云推理模型与深度强化学习相结合，将云推理模型作为强化学习的启发式探索策略，引导智能体的学习，加快策略的收敛速度。本文方法与只使用规则知识进行推理的算法(如云推理、模糊推理等)相比，由于强化学习只是利用知识来引导探索，而不是完全基于知识进行推理，因而其不要求定性知识的完备性和知识表达参数的准确性。在实际的应用中，该方法也更具有通用性和实用性。本文分别从理论和实验上验证了本文方法的有效性。

2　基于云推理的启发式探索策略

本文基于云推理的知识表示方法，对静态的定性规则知识进行表示，提出了一种构建强化学习启发式探索策略的方法。

本文以 Nature DQN[3]算法为例，介绍基于云推理模型的深度强化学习探索策略，具体算法如表 1 所示。首先根据具体任务环境构建规则知识集，然后确定每条定性规则的前件云与后件云参数，并对其他超参数进行初始化。强化学习开始后，在探索策略动作选择阶段(步骤 5)，以ε的概率采纳云推理建议的动作，以$1-\varepsilon$的概率按贪心策略选择动作。在初始学习阶段，由于智能体还未学得有效的策略，云推理启发式探索策略就显得更为重要，因而ε的值应当较高；而随着学习的进行，在学得的策略逐渐完善的情况下，由于采用的知识的不完备性，可以适当逐渐减小ε。在步骤 5 中，云推理给出的动作是连续的数值，因而对于离散动作的问题，应进行取整操作。在后面的步骤中，即完全按照 DQN 算法进行，步骤 6 和步骤 7 是 DQN 算法的经验回放机制，步骤 10 是目标网络分离机制。

在给出基于云推理的启发式探索策略基础上，本文尝试从理论上分析了探索策略对强化学习的影响。Singh[12]定义了最优策略值函数与当前策略值函数在状态s下的差值为：

$$L_{\overline{Q}}(s)=Q^{*}\left(s,\pi^{*}(s)\right)-\overline{Q}\left(s,\pi_{\overline{Q}}(s)\right) \tag{1}$$

其中的动作选择即为探索策略，最优策略值函数中的动作为$a^{*}=\pi^{*}(s)$，当前策略值函数中的动作为$\overline{a}=\pi_{\overline{Q}}(s)$。损失函数$L_{\overline{Q}}(s)$即为在 Q-Learning 学过程中的值函数与理想情况下的最优值函数的差值。每个状态下的$L_{\overline{Q}}$越小，代表学习效率越高，收敛速度越快。从迭代角度来看，在学习的每个步骤中，动作选择时的$\pi_{\overline{Q}}(s)$越接近$\pi^{*}(s)$，即选择的动作越接近最优动作，则迭代中的$\overline{Q}$就越接近Q^{*}，即损失函数$L_{\overline{Q}}(s)$就会越小。由此可以发现，在学习的初始阶段，由于贪心策略效果较差，使用基于知识的探索策略将发挥巨大的作用，且越接近最优策略(即知识规则越完备)，对学习过程的贡献就越大。

表 1 基于云推理模型的启发式 DQN 探索策略

算法 1 基于云推理模型的启发式 DQN 探索策略

输入：定性规则知识集

输出：策略 π

步骤 1 云推理模型参数初始化：对每条规则 i 初始化前件云 $(Ex,En,He)_{Ai}$ 与后件云 $(Ex,En,He)_{Bi}$ 参数；设置 ε 的初始值 $\varepsilon_{\text{start}}$ 和最终值 ε_{end} ,及其下降速率 $\Delta\varepsilon$ ，令 $\varepsilon=\varepsilon_{\text{start}}$

步骤 2 DQN 参数初始化：初始化记忆存储单元 D(容量为 N)，初始化 $Q(s,a;w)$ 网络，初始化目标网络 $\widehat{Q}(s,a;w^{-})$，参数 $w^{-}=w$ ，设置折扣系数 γ 。

步骤 3 **For** episode=1,2,$\cdots$, MaxEpisode **do**

步骤 4 **For** t=1,2,$\cdots$, MaxStep **do**

步骤 5 使用基于云推理模型的启发式探索策略，选择动作：

If rand() $<\varepsilon$

调用云推理模型算法，根据当前状态 s_t 得到云推理模型输出的动作 a_{advice}

计算每个前件云 CG_{Ai} 在当前状态 s 下的确定度 μ_i

IF 只有 1 个确定度 $\mu_i>0$ (1 条规则被激活)

使用后件云 CG_{Bi} 发生器生成云滴 (a_i,μ_i) 的动作值 a_i

ELSE 有 2 个确定度 $\mu_i,\mu_j>0$ (2 条规则被激活)

使用后件云 CG_{Bi} 发生器生成云滴 (a_i,μ_i) 动作值 a_i

使用后件云 CG_{Bi} 发生器生成云滴 (a_j,μ_j) 动作值 a_j

根据 a_i 与 a_j 生成虚拟云，求取虚拟云期望 $Ex_v=\dfrac{a_i\sqrt{-2\ln\mu_j}+a_j\sqrt{-2\ln\mu_i}}{\sqrt{-2\ln\mu_i}+\sqrt{-2\ln\mu_j}}$

得到云推理的动作

$$a_{\text{advice}}=\begin{cases}a_i, & \text{当 1 条规则被激活}\\ Ex_v, & \text{当 2 条规则被激活}\end{cases}$$

对于连续动作问题：$a_t=a_{\text{advice}}$

对于离散动作问题：$a_t=\text{Math.Round}(a_{\text{advice}})$

同时减小 ε : $\varepsilon=\varepsilon-\Delta\varepsilon$

else

使用贪心策略：$a_t=\arg\max\text{Q}(s_t,a_t)$

步骤 6 执行动作 a_t 后，观察环境得到新的状态 s_{t+1} 和奖赏 r_t，将 (s_t,a_t,r_t,s_{t+1}) 数据存储于记忆单元 D

步骤 7 从 D 中 miniBatch 随机采样若干 (s_t,a_t,r_t,s_{t+1})

步骤 8 设置 i 目标标签

$$y_i=\begin{cases}r_i, & \text{达到终止状态}\\ r_i+\gamma\,\text{argmax}_{a'}\,\widehat{Q}(s_{i+1},a';w^{-}), & \text{其他}\end{cases}$$

步骤 9 $\left(y_i-Q(s_i,a_i;w)\right)^2$ 随机梯度下降更新网络参数 w

步骤 10 每 C 步将 $\widehat{Q}$ 与 Q 同步

步骤 11 **End For**(step)

步骤 12 **End For**(episode)

3 实验与分析

3.1 实验环境介绍

本文在 OpenAI Gym 的“CartPole-v0”和“CartPole-v1”基础上，自定义“CartPole-v2”作为实验环境。“CartPole-v2”的任务与“CartPole-v0”和“CartPole-v1”相同，均是对底部可移动的平板施加不同方向的力，使直杆偏角保持在 $\pm12°$ 之间，且平板左右最大位移不超过 ±2.4，每坚持 1 步，环境奖励 1 分。所不同的是，前两个版本只有 2 个离散动作，每回合最大步长分别是 200 和 500，而自定义的“CartPole-v2”有 5 个离散动作，且最大步长为 1000。“CartPole-v2”的难度较前两个版本有一定提高。环境参数对比如表 2 所示。

3.2 实验结果及分析

本文实验分为两部分，第 1 部分直接使用云推理模型对“CartPole-v2”环境进行测试；第 2 部分将云推理模型作为探索策略，验证本文所提出的方法的有效性。

3.2.1 云推理模型 在“CartPole-v2”实验中，云推理模型使用的输入状态，暂只考虑角度 θ ；输出则是对应的离散动作 $a\in\{0,1,2,3,4\}$ 。

对于上述问题，我们可以根据日常经验很容易地总结出定性控制规则：(1)如果直杆正向偏角很大，使用很大的正向力；(2)如果直杆正向偏角较大，使用较大的正向力；(3)如果直杆偏角为零，力为零；(4)如果直杆负向偏角较大，使用较大的负向力；(5)如果直杆负向偏角很大，使用很大的负向力。

使用云模型对上述定性规则进行表示，前件云和后件云参数如表 3 所示。其中表示正向和负向偏角很大的云为半升半降云，其余为完整的高斯云。这里，云模型的参数均是直接根据经验设定而来，并没有经过更多地调整和修正。

本文将随机控制作为基准，将精确控制作为对比实验。精确控制根据定性控制规则，人为给定前提条件，当符合某一条件时，即采取对应的动作。精确控制的参数设定为：

(1)如果 $\theta\le-0.75\text{unit}$， $a=0$ ；

(2)如果 $-0.75\text{unit}<\theta\le-0.25\text{unit}$, $a=1$ ；

(3)如果 $-0.25\text{unit}<\theta\le0.25\text{unit}$, $a=2$ ；

(4)如果 $0.25\text{unit}<\theta\le0.75\text{unit}$, $a=3$ ；

(5)如果 $\theta\ge0.75\text{unit}$, $a=4$ 。

对随机控制、精确控制和云推理模型分别进行 10 次实验，每次实验执行 2000 个回合，每个回合最大步长为 1000。对每次实验求取均值，实验结果如表 4 所示。从实验结果可以看出，云推理模型和

表 2　OpenAI Gym 环境的状态和动作空间

环境	状态空间	动作空间	最大步长
CartPole-v0	4 维连续空间位移： $x\in[-2.4,2.4]$ 速度： v 角度： $\theta\in\left[\frac{-12\cdot\pi}{180},\frac{12\cdot\pi}{180}\right]$ 角速度： $\dot{\theta}$	1 维离散动作： $\text{action}\in\{0,1\}$ 对应环境中的实际作用力： $F\in\{-10,10\}$	200
CartPole-v1			500
CartPole-v2		1 维离散动作： $\text{action}\in\{0,1,2,3,4\}$ 对应环境中的实际作用力： $F\in\{-10,-5,0,5,10\}$	1000

表 3 云推理模型参数

	前件云参数	后件云参数
负向偏角很大	$(-1\text{unit},0.030,0.003)$	$(0,0.220,0.020)$
负向偏角较大	$(-0.5\text{unit},0.020,0.002)$	$(1,0.200,0.018)$
偏角为零	$(0,0.020,0.002)$	$(2,0.150,0.010)$
正向偏角较大	$(-0.5\text{unit},0.020,0.002)$	$(3,0.200,0.018)$
正向偏角很大	$(1\text{unit},0.030,0.003)$	$(4,0.220,0.020)$

$\text{unit}=12\pi/180$

表 4 随机控制、精确控制和云推理模型的得分

实验	随机控制	精确控制	云推理
1	26.06	120.68	148.11
2	25.74	120.88	146.50
3	26.24	121.00	147.23
4	26.05	120.79	146.66
5	25.80	120.61	146.95
6	25.63	120.79	146.89
7	25.86	120.36	147.92
8	25.75	120.96	146.25
9	25.49	120.90	147.24
10	25.85	120.85	146.52
均值	25.85	120.78	**147.02**

精确控制都能达到 100 分以上，高于随机控制。同时，云推理模型比精确控制高 20%左右，但距离满分 1000 还有较大差距。由于本文的重点不在于研究云推理模型的参数设置与优化，因而接下来的实验就采用当前参数的云推理模型和精确控制模型。

3.2.2 基于云推理模型的深度强化学习　本文选用 DQN[2], Nature DQN[3], Double DQN[13]算法，分别使用 ε -greedy、精确控制和云推理模型作为强化学习的探索策略进行测试。各个算法在“CartPole-v2”环境下独立训练 10 次，每次训练不超过 2000 个回合。每训练 10 个回合，使用学到的策略测试 10 个回合，求取平均累积奖赏，作为当前策略的评价指标。测试使用的策略不包含随机探索或者规则知识。平均累积奖赏收敛到 1000 分或训练达到 2000 个回合后结束训练。DQN, Nature DQN, Double DQN 算法均采用相同的超参数：含 1 个隐层(20 个神经元)的全连接神经网络，Replay Memory 大小为 10000, Mini Batch 大小为 32，优化方法采用 Adam 梯度下降算法，学习率为 0.001。Nature DQN, Double DQN 中，每训练 200 步更新一次 Target 网络。设置 $\varepsilon_{\text{start}}=0.5$, $\varepsilon_{\text{end}}=0.01$, $\Delta\varepsilon=4.9\times10^{-5}$。精确控制和云推理的探索策略使用相同的 ε。

实验结果如表 5 所示，表中数字表示达到满分 1000 分所需要训练的回合数，“2000-”表示 2000 个回合后仍然未能达到满分。“N-DQN”和“D-DQN”代表 Nature DQN 和 Double DQN 算法。“-Pre”代表精确控制探索策略，“-Cloud”代表云推理探索策略。从实验数据可以看出，在每个算法中加入精确控制和云推理模型都能起到加速训练的效果。在当前超参数条件下，使用 ε -greedy 贪心策略的强化学习算法训练都需要较多的回合数，甚至在 2000 个回合下无法收敛；而使用基于精确控制和云推理模型的探索策略则都能较快收敛。精确控制与云推理策略相比较，虽然在个别实验中，精确控制训练的回合数比较少，但总体来讲，云推理策略具有更稳定、更明显的加速效果。

4　结束语

本文提出了一种基于云推理模型的启发式探索策略，在强化学习过程中使用定性规则知识引导智能体的动作选择，使得智能体能很快的学到有效策略，减少探索的盲目性。与单纯使用定性规则进行控制决策所不同的是，本文提出的算法只是利用定性规则作为引导，因而其不要求定性规则的完备性和知识表达参数的准确性。在实际应用中，该算法也更具有通用性与实用性。本文结合深度强化学习算法，在“CartPole-v2”环境中进行了实验验证。实验结果表明，基于云模型的启发式探索策略较之贪心策略和精确控制策略相比，训练效率稳步提升。

在本文研究工作的基础上，今后的工作应在更多环境中进行验证，同时进一步讨论不同质量的规则知识对收敛效率影响的量化研究。

表 5 基于云推理模型的启发探索策略所需回合数

实验	DQN	DQN -Pre	DQN -Cloud	N-DQN	N-DQN -Pre	N-DQN -Cloud	D-DQN	D-DQN -Pre	D-DQN -Cloud
1	1770	360	790	980	1740	940	2000-	830	390
2	1320	1100	600	940	920	640	1870	570	570
3	1970	1190	370	1280	630	740	800	670	650
4	2000-	290	660	2000-	380	670	780	1700	570
5	2000-	1790	500	1780	530	790	800	620	600
6	880	840	670	870	920	1240	700	690	570
7	1290	530	430	2000-	810	910	1030	940	470
8	1880	1400	330	500	820	170	720	1160	600
9	2000-	390	610	1830	790	580	680	1180	390
10	1940	890	420	680	560	250	440	1040	300
最小值	880	**290**	330	500	380	**170**	440	570	**300**
最大值	1970	1790	**790**	1830	1740	**1240**	1870	1700	**650**
均值	1578	878	**538**	1107	810	**693**	868	940	**511**

参 考 文 献

[1] SUTTON R S and BARTO A G. Reinforcement Learning: An Introduction[M]. MA: MIT Press, 1998: 3–24. doi: 10.1109/TNN.1998.712192.

[2] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Playing Atari with deep reinforcement learning[OL]. https://arxiv.org/abs/1312.5602v1, 2013.12.

[3] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529–533. doi: 10.1038/nature14236.

[4] OSBAND I, BLUNDELL C, PRITZEL A, *et al*. Deep exploration via bootstrapped DQN[C]. Proceedings of the 29th Neural Information Processing Systems, Barcelona, 2016: 4026–4034.

[5] BELLEMARE M, SRINIVASAN S, OSTROVSKI G, *et al*. Unifying count-based exploration and intrinsic motivation[C]. Proceedings of the 29th Neural Information Processing Systems, Barcelona, 2016: 1471–1479.

[6] HOUTHOOFT R, CHEN X, DUAN Y, *et al*. VIME: Variational information maximizing exploration[C]. Proceedings of the 29th Neural Information Processing Systems, Barcelona, 2016: 1109–1117.

[7] DAVENPORT T H, PRUSAK L, and PRUSAK L. Working Knowledge: How Organizations Manage What They Know [M]. Boston: Harvard Business School Press, 1997: 1–24. doi: 10.1145/347634.348775.

[8] SANTOS M and BOTELLA G. Dyna-H: A heuristic planning reinforcement learning algorithm applied to role-playing game strategy decision systems[J]. *Knowledge-Based Systems*, 2012, 32(8): 28–36.

[9] BIANCHI R A C, ROS R, and MANTARAS R L D. Improving reinforcement learning by using case based heuristics[C]. Proceedings of the International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, Burlin, 2009: 75–89.

[10] KUHLMANN G, STONE P, MOONEY R, *et al*. Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer[C]. Proceedings of the 19th National Conference on Artificial Intelligence Workshop on Supervisory Control of Learning and Adaptive Systems, California, 2004: 30–35.

[11] LI Deyi, CHEUNG D, SHI Xuemei, *et al*. Uncertainty reasoning based on cloud models in controllers[J]. *Computers & Mathematics with Applications*, 1998, 35(3): 99–123.

[12] SINGH S P. Learning to solve Markovian decision processes [D]. [Ph.D. dissertation], University of Massachusetts, Amherst, 1994: 66–72.

[13] HASSELT H V, GUEZ A, and SILVER D. Deep reinforcement learning with double Q-learning[C]. Proceedings of the 30th AAAI Conference on Articial Intelligence, Phoenix, 2016: 2094–2100.

李晨溪： 男，1989 年生，博士生，研究方向为指挥信息系统工程、强化学习.

曹 雷： 男，1965 年生，教授，研究方向为指挥信息系统工程.

陈希亮： 男，1985 年生，讲师，研究方向为指挥信息系统工程.

张永亮： 男，1982 年生，讲师，研究方向为指挥信息系统工程.