文章编号：2095-2163(2022)02-0068-05 中图分类号：TP391 文献标志码：A

基于 TD3 算法的对话策略研究

洪 洲，余承健
（广州城市职业学院 教务处，广州 510408）

摘 要：对话策略是任务型对话系统构建的核心组件，通常被定义为强化学习，通过代理和环境的交互，提升对话策略效率。针对当前任务型的对话系统缺少高质量的标注数据集及模型难于收敛等问题，提出了结合规划的双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient）算法，用以优化对话策略。该算法使用孪生网络结构，采用软更新、策略噪音和延迟学习等方法，有效的改善了过估计问题。实验结果表明，该方法加速了模型的收敛，提升了对话成功率。

关键词：对话系统；强化学习；对话策略；代理

Research on dialogue strategy based on TD3 algorithm

HONG Zhou, YU Chengjian
(Academic Affairs Office, Guangzhou City Polytechnic, GuangZhou 510408, China)

【Abstract】 Dialogue strategy is the core component of task-based dialogue system. It is usually defined as reinforcement learning to improve the efficiency of dialogue strategy through the interaction between agent and environment. In view of the problems that the current task-based dialogue system lacks high-quality annotation datasets and the model is difficult to converge, a double delayed deep deterministic policy gradient algorithm combined with programming is proposed to optimize the dialogue strategy. The algorithm uses twin network structure and adopts soft update, strategy noise and delay learning methods to effectively improve the over estimation problem. The experimental results show that this method accelerates the convergence of the model and improves the success rate of dialogue.

【Key words】 dialogue system; reinforcement learning; dialogue strategy; agent

0 引 言

在人工智能发展时代，许多对话机器人产品逐渐融入了人们的生活[1]。如：阿里的天猫精灵、百度的小度和腾讯的小微等智能语音助理。通过与这些智能对话机器人交互，人们能够获得更便捷的服务。正是由于其广泛的应用前景，工业界和学术界均给予了高度重视和关注。

对话系统分为任务型对话系统和非任务型对话系统。非任务型对话机器人又称闲聊机器人，在开放领域内实现尽可能多轮次对话；任务型对话系统的研究方法主要有两种：多模块级联方法和端到端的方法。多模块级联方法又称为管道方法，各模块功能独立且易于理解，缺点是易导致误差的积累。端到端方法可以将用户的输入直接输入模型，进而得到系统的输出。可以观察到输入的反馈，但可以用于训练的数据难以获取，并且可维护性和解释性较差。

本文聚焦于任务型对话系统的研究。其系统主要模块为：自然语言理解（Nature Language Understanding，NLU）、对话管理（Dialogue Manage，DM）、自然语言生成（Nature Language Generator，NLG）。自然语言理解模块对用户输入的文本进行解析，通常有槽填充、意图识别。对话管理模块的主要功能，是在多轮对话过程中，维护历史信息和当前的状态，并且生成下一轮对话的回复策略，同时也可与外部知识库进行交互。自然语言生成模块的主要功能是，根据对话策略模块的结果及预先定好的规则，生成自然语言形式的回复。

对话管理在整个对话系统中占据重要地位，直接影响整个系统的性能。本文从强化学习的角度出发，提出一种结合规划的双延迟深度确定性策略梯度算法，来优化对话策略，改善模型难以收敛的问题。在代理方面，针对 TD3 算法[2]只适合处理连续空间任务的特点提出了改进，使其能够处理离散空间的数据。在配置环境方面，借助经典 DDQ 模型的思想，将其与 TD3 算法结合。实验结果表明，本文提出的模型能够更快的收敛，取得较好的实验结果。

基金项目：羊城学者科研项目（202032796）。

作者简介：洪 洲（1979-），男，硕士，教授，主要研究方向：智能软件与服务机器人、自然语言处理。

收稿日期：2021-10-18

1　相关工作

任务型对话系统的对话策略主要任务是：根据当前 t 时刻的对话状态 S，在预先定义的动作集 A 中，选择 $t+1$ 时刻的动作 q。对话策略直接决定当前对话任务的优劣，因此，对话策略的设计及其建模过程一直都是研究的热点和难点。当前，主流方法有基于规则的方法、端到端方法和强化学习方法。

1.1　基于规则方法

基于规则的方法利用该领域的专家分析对话流程，并设定预定义的对话状态及对该状态的回复。最有代表性的方法就是有限状态机[3]和槽填充模型[4]。有限状态机的状态转换和流程都是预先设计的，所以其流程可以有效地控制，而且结构清晰。但是，这样的状态机无法移植到另一个领域。对于槽填充模型而言，槽就是对话系统在特定任务中所需要获取的特定信息。如，地点、时间、天气等。对话系统通过当前槽的状态及其优先级，决定下一个动作。对话过程被建模成序列标注，对话顺序是不确定的，获得的回答非常灵活。但是，也可能产生状态爆炸的情况。基于规则的方法，虽能够很好的控制对话的流程，却严重依赖专家制定的领域知识，同时很难迁移到新的领域。

1.2　端到端方法

端到端方法是随着深度学习技术的突破发展而提出的。使用端到端[5]的训练模型，将一个域的序列映射到另一个域。在某一时刻，对话管理根据上一步词序列和一些结构化的外部数据库，选择概率最高的词汇作为下一步的回答。通常选择的模型为编码器-解码器，减少了模块化开发的成本。然而端到端的方法，也受限于对话数据集的获取和标注，无法及时对自身策略进行调整和进行在线学习。

1.3　强化学习

将强化学习[6]的思想应用于对话策略的建模，是当前的主流方法。通过智能体与环境交互过程中的学习，以获得最大化的奖励，其结构如图 1 所示。

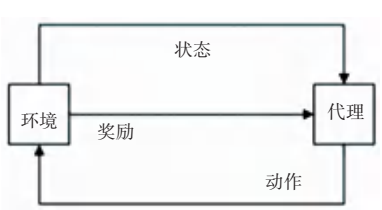

图 1　强化学习模型

Fig. 1　Reinforcement learning model

对话管理的过程可以看作一个马尔科夫决策过程[7]，通常被定义为五元组 $< S,A,P,R,\gamma >$。通过策略 π 实现一个行为与状态之间的映射，策略的取值可以是确定值，也可以是随机值。映射的动作既可以通过一个连续分布函数取值，也可以是离散值。

强化学习的优势：一是无须人工制定规则，且提高了泛化能力；二是可以充分利用状态空间，解决了覆盖率低的问题。近年来，随着深度学习的快速发展，学者们结合深度强化学习的方法用于建模对话策略，其算法性能在一些领域优于人类。如：Li 等人[8]通过与机器人的交互学习对话策略，构建了一个用户模拟器；基于 DQN[9]模型的算法在订电影票的任务上比基于规则的方法正确率更高；Volodymyr 等[10]提出了 BBQN 模型，使用辛普森采样对状态空间进行探索，明显提升了效率；Peng 等人[11]为了解决训练强化学习的代理需要耗费大量资源和时间问题，引入用户模拟器来产生大量用于训练的模拟数据，提出了 DDQ(Deep Dyna-Q)模型；Su 等[12]人对 DDQ 模型进行了改进，引入 RNN 鉴别器，用于区分真实的用户经验和生成的模拟经验，从而过滤掉世界模型生成的低质量训练数据，同时减少了训练过程中对于模拟数据的依赖。从实验结果可见，D3Q 对话管理系统的鲁棒性和泛化能力优于 DDQ 对话管理模型。

2　结合规划的 TD3 算法

对话系统让智能机器人能够使用自然语言的方式与人类沟通，其中任务型对话系统旨在高效沟通并且让用户获取有价值的信息。在这类对话系统中，通常是由一个任务型对话策略，来提供语言上的行为决策。近几年，强化学习广泛应用在对话策略模型的学习上，即从基于语言的人机交互中训练对话策略模型。

本文提出一种基于深度强化学习的算法，来提高对话策略的学习效率，即结合规划的 TD3 算法。整体结构由 5 部分组成，各模块功能如下：

(1)基于 LSTM 的 NLU 模块，用于识别用户的意图和相匹配的语义槽；

(2)根据识别的结果，进行对话状态的跟踪并生成对话状态表述；

(3)对话策略学习：根据对话状态跟踪的结果，选择一个执行的动作；

(4)根据上一步选择的动作转化为对应的自然语言；

(5)世界模型：用于生成模拟的用户行为和奖

励。

本文中的训练是利用预先收集的数据,采用热启动的方式进行的。模型训练过程如图 2 所示,其实现步骤如下:

(1)代理与用户模拟器交互,利用真实对话数据改进对话策略;

(2)使用真实的对话数据更新世界模型;

(3)将更新后的世界模型的模拟经验用于改进对话策略。

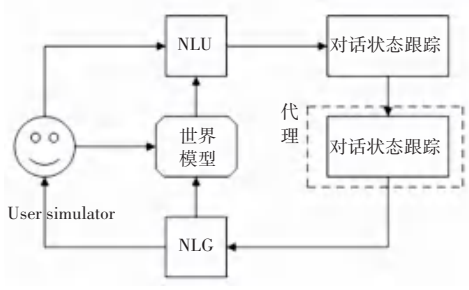

图 2 模型训练过程

Fig. 2 Model training process

2.1 直接强化学习

直接强化学习的目标,是让代理使用用户模拟器的真实对话数据优化对话策略。本文利用改进的 TD3 算法,根据用户模拟器产生的对话状态 s, 由相应的策略选择动作 a, 用户模拟器同时反馈给代理相应的奖励 r, 此时对话状态将更新为 s', 最后将对话经验存储到预先设置的经验回放池,继续循环整个过程,直到结束对话。

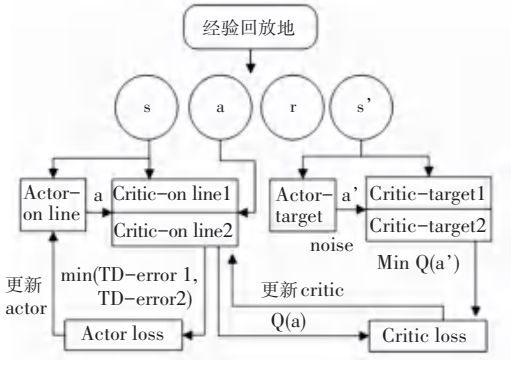

图 3 TD3 算法模型

Fig. 3 TD3 algorithm model

本文中改进的 TD3 算法共有 6 个网络结构,如图 3 所示。该算法采用两个结构完全相同的 critic 网络评估 Q 值,选取较小值作为更新的目标。有效缓解了样本噪声对动作价值估计的影响,以及不准确估计值累加所导致的无法收敛情况。TD3 算法对策略采用延时更新的方法,由于 target 网络与 online 网络参数的更新不同步,则规定 online 网络更新 d 次以后再更新 target 网络,从而减少了误差积累,并降低了方差。TD3 算法采用了一种目标策略的平滑正则化,在 target 网络的动作估计中加入随机噪声 ε, 使得值函数的更新平滑。

在 critic-online 网络中,调节 θ_1、θ_2 的值来最小化均方误差损失函数,优化目标函数如下:

$$y \leftarrow r + \gamma \min_{i=1,2} Q_{\theta'_i}(s',a) \tag{1}$$

$$\theta_i \leftarrow \operatorname{argmin}_{\theta_i} N^{-1} \sum (y - Q_{\theta_i}(s,a))^2 \tag{2}$$

其中, γ 为折扣因子, θ_1、θ_2 和 φ 分别为 critu-online 网络和 Actor-online 网络的随机参数。

原始 TD3 算法用于处理连续空间的数据,使用梯度求最优值,如式(3)所示:

$$\tilde{\mathrm{N}}_{\phi} J(\phi) = N^{-1} \sum \tilde{\mathrm{N}}_a Q_{\theta_1}(s,a)\big|_{a=\pi_{\phi}} \times \tilde{\mathrm{N}}_{\phi} \pi_{\phi}(s) \tag{3}$$

由于本文对话任务数据均为离散数据,因此使用 TD-error 代替梯度计算。表示在当前的环境中,如何选择动作可以获得最大的奖励期望值,并将 actor 网络的输出进行 softmax 计算,使用确定性策略 μ, 选择一个具体的动作值。因此将式(3)改为如下形式:

$$\tilde{\mathrm{N}}_{\theta^{\mu}} J = -N^{-1} \sum (M^{-1} \sum (y - Q_{\theta_i}(s,a)) \times \log(\mathrm{prob}(\mu(s \mid \theta^{\mu})_s)) \tag{4}$$

其中, y 由式(1)可得, $\log(\mathrm{prob}(\mu(s \mid \theta^{\mu})_s$ 为选择某一个动作的概率值。可通过软更新机制更新参数,如下所示:

$$\theta'_i \leftarrow \tau \theta_i + (1-\tau)\theta'_i \tag{5}$$

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi' \tag{6}$$

最后,使用深度学习的 batch 训练方式,迭代更新对话策略的参数。

2.2 规划

在综合规划的步骤中,世界模型产生模拟的对话数据,然后用来训练对话策略。在训练本文的模型时,参数 K 是代理用于执行规划过程的次数。假定世界模型能够准确的模拟用户环境,即在一定程度上增大 K 值来提升对话经验策略。用户模拟器中得到的真实对话经验记为 D^u, 世界模型产生的模拟经验记为 D^s。 虽然规划和直接强化学习都使用改进的 TD3 算法,但是直接强化学习使用的是 D^u 数据,规划使用的是 D^s 数据。

2.3 世界模型

世界模型利用真实的对话数据 D^u 来训练其模型参数。在每一轮对话训练中,世界模型将上一轮的对话状态 s 和上一轮的代理行动 a 作为模型的输入,得到用户的回复动作 a_u、奖励 r 和一个表示对

话是否结束的信号 t。网络结构如图 4 所示。

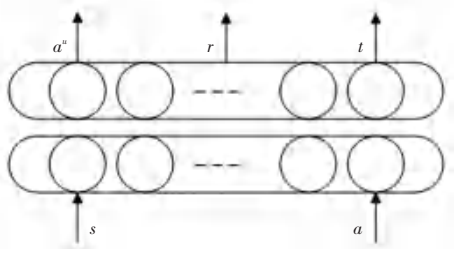

图 4　世界模型结构

Fig. 4　The world model architecture

其中，a_u、r 和 t 的计算公式如下：

$$h = \tanh(W_h(s,a) + b_h) \tag{7}$$

$$r = W_r h + b_r \tag{8}$$

$$a^u = \text{softmax}(W_a + b_a) \tag{9}$$

$$t = sigmoid(W_t h + b_t) \tag{10}$$

其中，(s,a) 表示拼接操作，得到的元组数据 (s,a,r,s') 保存到 D^s 中，用于后续代理的训练。

3　实验设置

3.1　数据集

本文使用的数据集已经通过标注，其中包括语义槽 16 个、对话动作 11 个以及带标记的对话 280 个，对话平均有 11 轮，见表 1。

表 1　意图和语义槽

Tab. 1　Intents and slots

意图	confirm answer, confirm question, deny multiple choice, welcome greeting, inform, thanks, request, not sure, closing
语义槽	closing, distance constrains, start time, greeting, movie name, people, city, price, state, task, complete, date, video theater, ticket, numberofpeople

3.2　基准模型

(1)基于规则的模型：使用基于人工制定规则的对话策略；

(2)DQN 模型：基于 DQN 算法实现；

(3)A2C 模型：使用优势函数代替 Critic 网络中的原始回报，作为衡量选取动作值和所有动作平均值好坏的指标；

(4)TD3 模型：即本文中提出的结合规划改进的 TD3 算法，引入世界模型，并且综合规划步骤进行学习。

3.3　实验设置和评估指标

实验中深度强化学习网络的激活函数选择 tanh 函数。折扣因子 γ 的值设为 0.9，D^u 与 D^s 的大小均为 5 000。规划训练过程中，模拟对话的最大回合数为 40。本文所有实验使用 100 轮对话的进行预训练，即使用热启动的方式。

主要评估指标为成功率、平均回报、平均轮数。假定测试中所有完整的对话数目为 N，成功预定电影票的完整对话为 a，完整对话所获得的总回报为 R，所有完整对话总对话轮数为 s。则：

$$\text{成功率} = a/N \tag{11}$$

$$\text{平均回报} = R/N \tag{12}$$

$$\text{平均轮数} = s/N \tag{13}$$

成功率用来衡量模型的主要性能，评估当前的对话策略的优劣；平均轮数和平均回报展示了系统的鲁棒性，这两个指标表明模型所追求的目标，即在最小的轮次获得最多的回报。

3.4　实验结果分析

本文中 TD3 模型需要学习结构相同，但是参数不同的神经网络有 actor 网络和 critic 网络。实验设置两个网络的参数不需同步更新，critic 网络的打分，决定了 actor 网络的动作。因此，actor 网络的参数更新具有滞后性。DQN 模型只需要学习一种神经网络参数，其效果要优于 A2C 模型。而本文的 TD3 模型要优于 A2C 模型，验证了本文提出模型能更快的收敛，提高对话系统的性能。

在综合规划的步骤中，世界模型可以用于减少代理对用户模拟器的依赖和负面影响。对于不同的 K 值，使得代理的训练结果不一致。世界模型的参数在整个实验的训练过程中也是需要学习的。因此，实验首要任务是寻找最佳 K 值。通过设置不同的 K 值进行训练得出：当 $K = 10$ 时，世界模型生成的模拟经验效果较好，模型训练效果最佳。

表 2 展示了不同模型在 10K 轮对话的最终测试结果。通过分析可以得出：结合规划的 TD3 模型从策略梯度的角度建模对话管理系统，提高了约 20% 的对话系统性能；成功率和平均回报稍优于 DQN 模型所代表的值函数模型；所使用的平均轮数持平。

表 2　不同模型在 10K 轮对话的最终表现

Tab. 2　The final performance of different models in 10K conversations

代理模型	成功率	平均回报	平均轮数
基于规则的模型	0.41	1.98	16.00
DQN 模型	0.81	42.23	15.89
A2C 模型	0.5	8.12	19.23
TD3 模型	0.7	20.23	16.11
DQN 模型 (K = 10)	0.85	54.47	15.52
A2C 模型 (K = 10)	0.55	13.23	18.08
TD3 模型 (K = 10)	0.88	55.56	15.96

综上所述，可以归纳出模型的优点：本文提出的

模型适合于规模较大的离散的对话任务;通过经验回放和用户模型的引入,带来比较好的对话策略学习效果,模型易于收敛。

4 结束语

本文结合规划的 TD3 算法在模型优化和环境建模做出了改进;在代理设置上,使用基于策略梯度的方法建模对话管理,并且使用经验回放和孪生网络结构,加快了模型的收敛,提高了对话性能;在环境设置上,引入了世界模型,减少了用户模拟器在代理训练的负面影响。当然,本文在策略梯度函数的设置上还有待进一步的优化,用户模拟器在功能上可以增加真人的对话数据收集等。

参考文献

[1] WALKER M A. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email [J]. Journal of Artificial Intelligence Research, 2000, 12: 387-416.

[2] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods [C]//International Conference on Machine Learning. PMLR, 2018: 1587-1596.

[3] BRANTING K, LESTER J, MOTT B. Dialogue management for conversational case-based reasoning [C]//European Conference on Case-Based Reasoning. Springer, Berlin, Heidelberg, 2004: 77-90.

[4] YUAN Y, TIAN H, DU B, et al. Research and implementation of frame-based dialogue management model [J]. Computer engineering, 2005, 31(13): 212-214.

[5] BORDES A, BOUREAU Y L, WESTON J. Learning end-to-end goal-oriented dialog[J]. arXiv preprint arXiv:1605.07683, 2016.

[6] ZHAO D, LIU D, LEWIS F L, et al. Special issue on deep reinforcement learning and adaptive dynamic programming [J]. IEEE Transactions on neural networks and learning systems, 2018, 29(6): 2038-2041.

[7] LEVIN E, PIERACCINI R, ECKERT W. Using Markov decision process for learning dialogue strategies [C]//Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). IEEE, 1998: 201-204.

[8] LI X, LIPTON Z C, DHINGRA B, et al. A user simulator for task-completion dialogues [J]. arXiv preprint arXiv: 1612.05688, 2016.

[9] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with Deep Reinforcement Learning [J]. arXiv preprint arXiv: 1312.5602,2013.

[10] LIPTON Z, LI X, GAO J, et al. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[11] PENG B, LI X, GAO J, et al. Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning [J]. arXiv preprint aXiv:1801.06176,2018.

[12] SU S Y, LI X, GAO J, et al. Discriminative Deep Dyna-Q: Robust Planning for Dialogue Policy Learning [J]. arXiv preprint arXiv: 1808.90442,2018.

(上接第 67 页)

[2] 程艳秋. 服务模块化价值网络治理机制对价值创造的影响研究[D]. 南昌:南昌航空大学, 2018.

[3] 苏立悦. 大规模定制下产品服务系统的配置及优化研究[D]. 上海:上海交通大学, 2010.

[4] 肖金花. 基于满意度提升的集体自助养老服务设计研究[D]. 西安:西北工业大学,2016.

[5] CHEN C. CiteSpace II : Detecting and Visualizing Emerging Trends [J]. Journal of the American Society for Information Science & Technology, 2006, 57(3):359-377.

[6] 王贺,高喜银. 基于文献分析的拖拉机振动可视化研究[J]. 中国农机化学报, 2020, 41(11): 95-100.

[7] 李玉林, 杨涛, 王秋月. 客户偏好驱动的产品配置设计过程研究[J]. 机械, 2020(8).

[8] CHEN Z, X MING, WANG R, et al. Selection of design alternatives for smart product service system: A rough-fuzzy data envelopment analysis approach [J]. Journal of Cleaner Production, 2020: 122931.

[9] 吴启飞. 供需交互视角下产品服务系统方案配置研究[D]. 镇江: 江苏大学, 2018.

[10] 赵彩邦. 面向客户需求的加工装备产品服务系统配置研究[D]. 泉州:华侨大学, 2020.

[11] 徐达饶. 基于多层复杂网络的产品服务系统方案配置优化与健壮性研究[D]. 上海:上海大学, 2020.

[12] 隆惠君. 考虑客户感知的产品服务系统配置研究[D]. 上海:上海交通大学, 2015.

[13] 殷振. 基于客户价值的产品服务系统方案设计关键技术研究[D]. 山东:山东大学, 2020.

[14] YANG X, WANG R, TANG C, et al. Emotional design for smart product-service system: A case study on smart beds [J]. Journal of Cleaner Production, 2021(16):126823.

[15] TN A, MS A, YM A, et al. Toward the development of a comprehensive Product-Service System (PSS) evaluation method [J]. Procedia CIRP, 2020, 93:802-807.

[16] AUGUSTO D, CATEN C T, JUNG C F, et al. State of the art on the role of the Theory of Inventive Problem Solving in Sustainable Product-Service Systems: Past, Present, and Future [J]. Journal of Cleaner Production, 2019, 212(MAR.1):489-504.

[17] 郑茂宽. 智能产品服务生态系统理论与方法研究[D]. 上海:上海交通大学, 2018.