计算机科学
COMPUTER SCIENCE

融合Dead-ends和离线监督Actor-Critic的动态治疗策略生成模型

杨莎莎, 于亚新, 王跃茹, 许晶铭, 魏阳杰, 李新华

引用本文

杨莎莎, 于亚新, 王跃茹, 许晶铭, 魏阳杰, 李新华. 融合Dead-ends和离线监督Actor-Critic的动态治疗策略生成模型[J]. 计算机科学, 2024, 51(7): 80-88.

YANG Shasha, YU Yaxin, WANG Yueru, XU Jingming, WEI Yangjie, LI Xinhua. Dynamic Treatment Regime Generation Model Combining Dead-ends and Offline SupervisionActor-Critic [J]. Computer Science, 2024, 51(7): 80-88.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000138

融合 Dead-ends 和离线监督 Actor-Critic 的动态治疗策略生成模型

杨莎莎　于亚新　王跃茹　许晶铭　魏阳杰　李新华
东北大学计算机科学与工程学院　沈阳 110169
医学影像智能计算教育部重点实验室(东北大学)　沈阳 110169
(1692080148@qq.com)

摘　要　强化学习对数学模型依赖性低,利用经验便于架构和优化模型,非常适合用于动态治疗策略学习。但现有研究仍存在以下问题:1)学习策略最优性的同时未考虑风险,导致学到的策略存在一定的风险;2)忽略了分布偏移问题,导致学到的策略与医生策略完全不同;3)忽略患者的历史观测数据和治疗史,从而不能很好地得到患者状态,进而导致不能学到最优策略。基于此,提出了融合 Dead-ends 和离线监督 Actor-Critic 的动态治疗策略生成模型 DOSAC-DTR。首先,考虑学到的策略所推荐的治疗行动的风险性,在 Actor-Critic 框架中融入 Dead-ends 概念;其次,为缓解分布偏移问题,在 Actor-Critic 框架中融入医生监督,在最大化预期回报的同时,最小化所学策略与医生策略之间的差距;最后,为了得到包含患者关键历史信息的状态表示,使用基于 LSTM 的编码器解码器模型对患者的历史观测数据和治疗史进行建模。实验结果表明,DOSAC-DTR 相比基线方法有更好的性能,可以得到更低的估计死亡率以及更高的 Jaccard 系数。

关键词:动态治疗策略;Dead-ends;Actor-Critic;状态表征
中图分类号　TP399

Dynamic Treatment Regime Generation Model Combining Dead-ends and Offline Supervision Actor-Critic

YANG Shasha,YU Yaxin,WANG Yueru,XU Jingming,WEI Yangjie and LI Xinhua
College of Computer Science and Engineering,Northeastern University,Shenyang 110169,China
Key Laboratory of Intelligent in Medical Image,Northeastern University,Shenyang 110169,China

Abstract　Reinforcement learning has low dependence on mathematical models,and it is easy to construct and optimize models by using experience,which is very suitable for dynamic treatment regime learning. However,existing studies still have the following problems:1)risk is not considered when learning strategy optimality,resulting in certain risks in the learned policy; 2)the problem of distribution deviation is ignored,resulting in learning policies completely different from the doctor's policy; 3)the patient's histo-rical observation data and treatment history are ignored,thus failing to obtain a good patient status and thus failing to learn the optimal policy. Based on this,DOSAC-DTR,a dynamic treatment regime generation model combining dead-ends and offline supervision actor-critic,is proposed. First,considering the risk of treatment actions recommended by the learned policies,the concept of dead-ends is integrated into the actor-critic framework. Secondly,in order to alleviate the problem of distribution offset,physician supervision is integrated into the actor-critic framework to minimize the gap between learned policies and doctors' policies while maximizing the expected return. Finally,in order to obtain a state representation that includes critical patient historical information,a LSTM-based encoder decoder model is used to model the patient's historical observation data and treatment history. Experiments show that DOSAC-DTR has better performance than the baseline approach,resulting in lower estimated mortality rates and higher Jaccard coefficients.

Keywords　Dynamic treatment regime,Dead-ends,Actor-Critic,State representation

1　引言

近年来,随着电子健康记录(Electronic Health Records,EHRs)在数量和多样性上的快速增长,如何有效地利用EHRs来生成动态治疗策略(Dynamic Treatment Regime,DTR)受到了人们的普遍关注[1-2]。由于DTR强化学习不

到稿日期:2023-10-19　返修日期:2024-03-27
基金项目:国家自然科学基金(62373084)
This work was supported by the National Natural Science Foundation of China(62373084).
通信作者:于亚新(yuyx@mail.neu.edu.cn)

依赖精确的数学模型，无需准确匹配治疗结果与过程，且以最大化预期回报为学习目标的思想非常适合改善长期治疗结果[3]。因此，将 DTR 问题建模为强化学习问题是当前研究的热点。但是，目前的 DTR 强化学习在医疗领域应用中仍存在一些实际问题。

首先，如何使用强化学习方法从离线数据中学习最优的 DTR 是一个重要的问题。目前常用的方法有 SARSA(on-policy)[4]和 D3QN(off-policy)[5]，两者的共性是将 online 环境下的方法直接应用到离线的 DTR 学习中。但是，直接使用 on-policy 算法从离线数据中学习策略未考虑构建环境是否准确，而 off-policy 算法则并未考虑分布偏移问题。为缓解分布偏移问题，Wang 等提出了 SRL-RNN(Supervised Reinforcement Learning with Recurrent Neural Network)方法[6]，通过评价信号和指标信号共同更新演员网络，但该方法并没有考虑药物的剂量问题，每种药物只有吃和不吃两种选项。Kaushik 等则将 CQL(Conservative Q-Learning)应用于脓毒症最佳治疗策略的学习[7]，但分布偏移问题仍然严重。最近，Fujimoto 等提出了 TD3＋BC 算法[8]，在最大化预期回报的同时，最小化和行为策略之间的差距，并在 D4RL 数据集上取得了很好的效果，该思想值得借鉴。

其次，如何在 DTR 中充分考虑患者的历史信息也是一个非常重要的问题。如 Matthieu 等将患者的当前观测数据离散化为 750 个互斥状态[4]，但是离散化的状态很难捕捉微小的症状波动。Yin 等先将观测数据离散化，然后使用 LSTM 建模患者的观测数据历史和治疗史[9]，但同样地，该方式并不能捕捉到微小的症状波动。Raghu 等和 Kaushik 等直接将患者的当前观测数据当作状态[5,7]，并没有考虑患者的观测数据历史和治疗史。Wang 等使用 LSTM 将患者的观测数据历史进行建模[6]，但是并没有考虑患者的治疗史。

另外，目前的大多数 DTR 研究更多的是关注如何生成最优 DTR[4-7]，鲜有考虑学到的策略所推荐的治疗行动是否是高风险的。相对而言，Fatemi 等详尽地研究了离线环境下基于强化学习推荐治疗方案的风险性问题[10]，提出了 Dead-ends 概念和相应的安全条件，通过求解治疗行动导致消极结果的概率(包括 3 种情况发生的概率之和，即患者进入 Dead-ends 的概率、患者立即死亡的概率、即使在未来所有步骤中都采用了最优治疗但患者仍然死亡的概率)，避免了 DTR 的高风险问题。然而，DQN 算法在求解强化学习问题时，需要对某个状态下所有动作的 Q 值进行估计，但有些状态动作对在离线数据集中是不存在的，从而导致 Q 值估计不准确。此外，Fatemi 等只考虑了应该要避免的高风险的治疗，但没有考虑应该选择的治疗方式。

为了解决上述问题，本文提出了融合 Dead-ends 和离线监督 Actor-Critic 的动态治疗策略生成模型(Dynamic Treatment Regime generation based on Dead-ends & Offline Supervised Actor-Critic，DOSAC-DTR)。首先，考虑学到的策略所推荐的治疗行动的风险性，使用 Actor-Critic 框架，并融入 Dead-ends 概念(对应于两个特殊的 Critic 网络 D-Critic 以及 R-Critic)。其次，为了缓解分布偏移问题，在 Actor-Critic 框架中融入医生监督，在最大化预期回报的同时，最小化所学策略与医生策略之间的差距。最后，为了得到包含患者关键历史信息的状态表示，使用基于 LSTM 的编码器解码器模型对患者的历史信息(患者的历史观测数据和治疗史)进行建模，从而学习到包括患者关键历史信息的状态表征。综上，演员网络的最终目标是最大化评论家网络输出的长期目标、最小化所学策略和医生策略之间的差距以及最大化 D-Critic 网络和 R-Critic 网络输出的值(分别对应于产生消极结果的可能性的相反数以及产生积极结果的可能性)。

综上所述，本文的主要贡献如下：

1)提出在 Actor-Critic 框架中融入 Dead-ends 概念，避免了基于学到的策略所推荐的治疗行动的风险性。使用两个特殊的 MDP，对应两个特殊的 Critic 网络，即 D-Critic 网络和 R-Critic 网络，两个网络的输出值分别对应于产生消极结果的可能性的相反数以及产生积极结果的可能性，且这两个网络与评论家网络和演员网络一起联合训练。

2)提出在 Actor-Critic 框架中融入医生监督以及患者关键历史信息，缓解了分布偏移问题且得到了包含患者关键历史信息的状态表示。在最大化预期回报的同时，还最小化了与医生策略之间的差距。使用单独训练的基于 LSTM 的编码器解码器模型对患者的历史信息进行建模。

3)进行可扩展性实验设计与验证。在公开数据上进行验证，实验结果表明本文方法相比基线方法有更低的估计死亡率，且与医生的策略更加接近。

2 相关工作

2.1 强化学习和 DTR

近年来，深度强化学习快速发展，并取得了一定的成功(西洋双陆棋游戏[11]、AlphaGo[12]、热气流滑翔[13])。目前，强化学习在医疗领域也得到了广泛的关注。强化学习算法已经被应用于解决重症监护病房中的 DTR 问题(脓毒症中的液体复苏和血管加压药[4]、机械通风[9]、一般的药物推荐[6]等)。

为了学习最优的 DTR，Matthieu 等提出了人工智能临床医生[4]，使用 K-Means 算法将状态空间离散化为 750 个互斥的状态，并使用离线的 EHRs 数据学习转移矩阵，最后直接应用 SARSA 学习脓毒症的最佳治疗策略。然而，他们并没有证据证明学习到的转移矩阵是否对患者状态数据拟合良好[14]。Raghu 等提出使用全连接 D3QN(Dueling Double DQN)直接从连续的生理数据状态空间和离散的动作空间中学习脓毒症的最佳治疗策略[5]。Liang 等提出将 D3QN 算法与情景控制相结合[15]，修改网络训练的损失函数，从记忆中主动锁定过去类似的治疗策略，然后利用这些好的策略来监督当前策略的优化，从而提高数据的利用率。Yin 等提出了一个强化学习框架 DAC(Deconfounding Actor-Critic)，并将其应用于机械通风的个性化治疗决策[9]，将患者的观测数据离散化，并使用 LSTM 建模患者的观测数据历史和治疗史，引入了患者重采样模块和混杂平衡模块，以减小学习到的策略的偏差。然而，上述方法都是直接将 online 强化算法应用于离线环境中，并没有考虑分布偏移问题。

2.2 强化学习和分布偏移

为缓解分布偏移问题，Kaushik 等提出将离线强化学习算法 CQL 应用于脓毒症最佳治疗策略的学习[7]，直接从离线数据中进行策略的学习，然而分布偏移问题仍然严重。Yu 等提出采用 SAC(Supervised Actor-Critic)算法来解决 ICU 通气和镇静剂量的问题[16]，SAC 在最大化预期奖励的同时最小化与医生策略的差距，以保持患者的稳定。但他们并没有说明如何根据离线数据来构建环境以及构建是否准确。Wang 等提出了 SRL-RNN[6]，并将其应用于 ICU 中更一般的药物推荐，基于 DDPG 算法，直接从离线数据中学习策略，通过评价信号和指标信号共同更新演员网络，以学习到最优的动态治疗方案。但该方法并没有考虑药物的剂量问题，每种药物只有吃和不吃两种选项。最近，Fujimoto 等提出了 TD3+BC 算法[8]，该算法在最大化预期回报的同时，最小化与行为策略之间的差距(即 MSE 损失)，并在 D4RL 数据集上取得了很好的效果，该思想值得借鉴。

除此之外，上述方法在学习最优动态策略时，未考虑患者的历史信息或者考虑不足。Matthieu 等和 Yin 等将患者观测数据离散化[4,9]，不能捕捉微小的症状波动。Raghu 等、Kaushik 等、Liang 等直接将患者的当前观测数据当作状态[5,7,15]，未考虑患者的历史信息。Wang 等只考虑了患者的观测数据历史，没有考虑患者的治疗史[6]。同时，上述方法只是考虑了学习到的策略的最优性，并没有考虑学到的策略所推荐的治疗行动是否是高风险的。

2.3 强化学习和高风险治疗

医疗领域是一个高风险领域，在学习 DTR 时，在考虑策略的最优性的同时，也要考虑学到的策略所推荐的治疗行动是否是高风险的。

在 online 环境中，最近的研究工作通过约束参数不确定性来限制高风险行为[17]，通过直接约束智能体优化过程[18]，或者通过改进基线策略来避免不安全行为[19]。然而，在 online 环境中，智能体可以与环境进行交互，以获取更多的数据信息。这些工作不适用于离线环境。

Fatemi 等在探索的背景下提出了 Dead-ends 概念和相应的安全条件[20]，学习代理需要经历每个来自每个状态的各种行动过程，通过这些过程学习到最优行动。在此基础上，Fatemi 等将其扩展到脓毒症的治疗决策问题上[10]，将其建模为强化学习问题，通过两个 Q 网络和两个特殊的奖励函数设计，使价值函数有了特殊的意义，即分别对应于产生消极结果的可能性(包括 3 种情况发生的可能性之和，即患者进入 Dead-ends 的可能性、患者立即死亡的可能性、即使在未来所有步骤中都采用了最优治疗但患者仍然死亡的可能性)的相反数和产生积极结果的可能性，并根据患者的当前健康状态来确定哪些治疗增加了产生消极结果的可能性，从而得到要避免的治疗。然而，该方法并没有考虑应该要选择的治疗方式，且对于离线数据中没有的治疗，估计是不准确的。

3 变量定义和问题描述

3.1 变量定义

表 1 列出了本文用到的符号及其描述。

表 1 符号

Table 1 Symbols

变量	描述
$\mathcal{O}$	观测空间
$\mathcal{S}$	状态空间
$\mathcal{A}$	动作空间
o_t	患者在时间步 t 时的观测数据
s_t	患者在时间步 t 时的状态
s_d	医疗 Dead-ends 状态
a_t	患者在时间步 t 时医生采取的治疗行动
$\hat{o}_t$	解码器预测的患者在时间步 t 时的观测数据
$\hat{a}_t$	演员网络推荐的治疗行动
r	奖励函数
M_d, r_d	D-Critic 网络对应的 MDP 和奖励函数
M_r, r_r	R-Critic 网络对应的 MDP 和奖励函数
$V(s), Q_w(s,a)$	r 估计的价值函数和动作价值函数
$V_d(s), Q_{w_d}(s,a)$	r_d 估计的价值函数和动作价值函数
$V_r(s), Q_{w_r}(s,a)$	r_r 估计的价值函数和动作价值函数
$\mu_\theta(s)$	演员网络学习到的确定性策略
D_{raw}	患者的轨迹数据
D	由 $(s_t, a_t, r_t, s_{t+1}, done)$ 组成的集合
z	$z=1$ 表示患者最终死亡，$z=0$ 表示患者最终出院
$\gamma, \gamma_d, \gamma_r$	分别表示 Q_w, Q_{w_d} 和 Q_{w_r} 对应的折扣系数，其中 γ_d, γ_r 都等于 1
τ	超参数，远小于 1，用于目标网络的更新
α, β	用于演员网络训练的 3 个模块之间的平衡
$w*, w'*, \theta, \theta'$	网络的可学习参数

定义 1(医疗 Dead-ends 状态 s_d) 当患者在状态 s_d 时，无论未来的治疗行动 $\{a_d, \cdots, a_T\}$ 如何，都会导致 $z=1$。

定义 2(患者的轨迹数据 D_{raw}) 患者的轨迹数据表示为 $D_{raw}=\{(o_1^i, a_1^i, \cdots, o_{T_i}^i, a_{T_i}^i)\}_{i=1}^n$，其中 $o_t \in \mathcal{O}$，$a_t \in \mathcal{A}$，n 表示患者的个数。

定义 3(DTR) DTR 指根据患者的观测数据史 $O_t=\{o_1, o_2, \cdots, o_t\}$ 和既往治疗史 $A_{t-1}=\{a_1, a_2, \cdots, a_{t-1}\}$ 确定时间点 t 的治疗 $\hat{a}_t$。

3.2 问题描述

本文将 DTR 问题建模为有限时间步长的马尔可夫决策过程(Markov Decision Process, MDP)，它由观测空间 $\mathcal{O}$、状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$ 和奖励函数 $r: \mathcal{S} \times \mathcal{A} \times \to \mathbb{R}$ 组成。

将患者的轨迹数据 D_{raw} 输入基于 LSTM 编码器中得到包含患者关键历史信息的状态 s_t，并使用解码器预测患者下一个时间步的观测数据 $\hat{o}_{t+1}$，使用 MSE 损失联合训练编码器和解码器。训练完成后，将患者的轨迹数据以四元组集合 $D=\{(s_{t_i}, a_{t_i}, r_{t_i}, s_{t_i+1}, done)\}_{i=1}^k$ 的形式进行存储。其中 $k=\sum_{i=1}^n T_i$，$done$ 表示 s_{t_i} 是否是最后一个状态 s_{T_i}。使用离线数据 D，根据奖励函数 r, r_d, r_r 来更新 $Q_w(s,a)$，$Q_{w_d}(s,a)$，$Q_{w_r}(s,a)$。并通过最大化预期回报($Q_w(s,a)$)、最小化与医生策略的差距以及最小化产生消极结果的可能性($Q_{w_d}(s,a)$)、最大化导致积极结果的可能性($Q_{w_r}(s,a)$)来学习一个确定性策略 $\mu_\theta(s)$，并将学到的策略输出。

4 融合 Dead-ends 和离线监督 Actor-Critic 的动态治疗策略生成模型 DOSAC-DTR

4.1 总体架构

本文提出的融合 Dead-ends 和离线监督 Actor-Critic 的

动态治疗策略学习(DOSAC-DTR)模型的总体架构如图 1 所示，主要由 4 个核心模块组成：演员网络 Actor、评论家网络 Critic、D-Critic 网络以及 R-Critic 网络。演员网络用于根据患者的状态输出治疗方案。评论家网络用于估计价值函数。D-Critic 用于确定产生消极结果的可能性(包括 3 种情况发生的可能性之和，即进入 s_d 的可能性、导致患者立即死亡的可能性、即使在未来所有步骤中都采用了最优治疗但患者仍然死亡的可能性)。R-Critic 用于确定产生积极结果的可能性(与消极结果相反)。同时也使用了 4 个目标网络(Actor$_{target}$，Critic$_{target}$，D-Critic$_{target}$ 和 R-Critic$_{target}$)来提高训练的稳定性和收敛性。为了对患者的观测数据历史和治疗史进行建模，以得到包含患者关键历史信息的状态表示，使用基于 LSTM 的编码器解码器模型对患者状态进行表征，如图 2 所示。

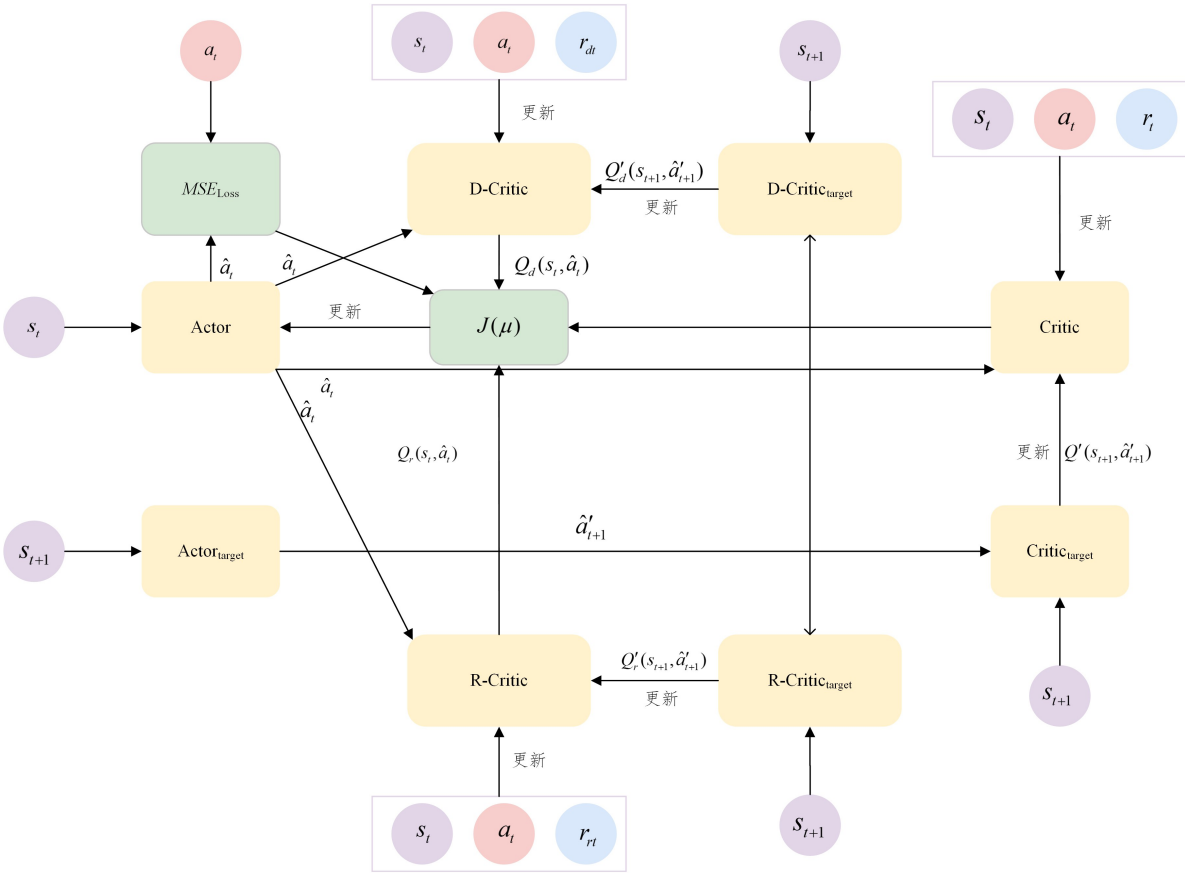


图 1　总体框架图

Fig. 1　Overall architecture

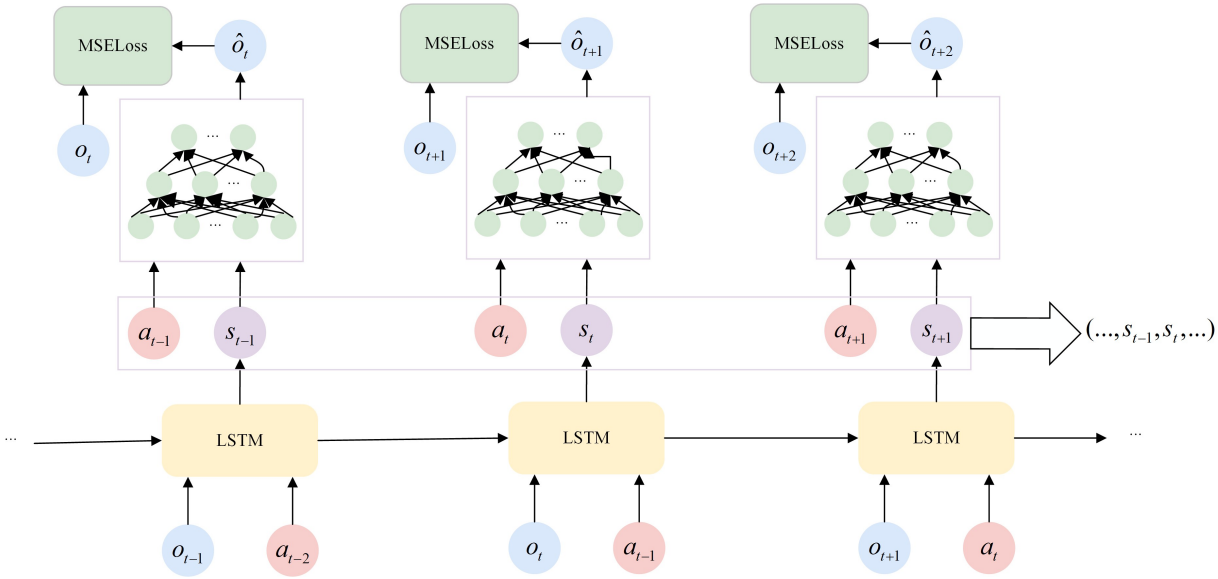


图 2　状态表征模块

Fig. 2　State representation module

4.2 基于 LSTM 的患者状态表征

如图 2 所示，使用基于 LSTM 的编码器解码器模型对患者状态进行表征[21]。编码器使用 LSTM 来实现，学习编码函数，如式(1)所示：

$$s_t = f(O_t, A_{t-1}) \tag{1}$$

其中，$O_t = \{o_1, o_2, \cdots, o_t\}$，$o_t \in \mathcal{O}$，表示患者的观测数据历史；$A_{t-1} = \{a_1, a_2, \cdots, a_{t-1}\}$，$a_t \in \mathcal{A}$，表示患者的治疗史。解码器使用全连接网络来实现，学习解码函数，如式(2)所示，根据编码器生成的状态 s_t 和时间步 t 时的治疗 a_t，预测时间步 $t+1$ 时的观察数据 $\hat{o}_{t+1}$。

$$\hat{o}_{t+1} = g(s_t, a_t) \tag{2}$$

通过真实数据和预测数据之间的 MSE 损失来联合训练编码器和解码器。

$$L = MSE(o_{t+1}, \hat{o}_{t+1}) \tag{3}$$

4.3 评论家网络更新

评论家网络使用全连接神经网络来学习动作价值函数 $Q_w(s,a)$，它的输入是患者的状态 s_t 和医生治疗行动 a_t，输出是在状态 s_t 下采取治疗 a_t 之后产生的预期回报，表示为 $Q_w(s_t, a_t)$。通过最小化式(4)所示的损失函数来更新网络参数。

$$J(w) = \mathbb{E}_{s_t, a_t \sim D}[(Q_w(s_t, a_t) - y^t)^2] \tag{4}$$

其中，

$$y^t = r(s_t, a_t) + \gamma Q'_{w'}(s_{t+1}, \mu'_{\theta'}(s_{t+1})) \tag{5}$$

其中，$Q'_{w'}$ 表示目标评论家网络，$\mu'_{\theta'}$ 表示目标演员网络，w'，w，θ 和 θ' 是神经网络的可学习参数，如果 s_t 是最后一个状态，则 $Q'_{w'}(s_{t+1}, \mu'_{\theta'}(s_{t+1})) = 0$。

4.4 D-Critic 网络和 R-Critic 网络的更新

D-Critic 网络和 R-Critic 网络是两个独立训练的全连接网络，分别学习具有特殊意义的动作价值函数 $Q_{w_d}(s,a)$ 和 $Q_{w_r}(s,a)$，网络的输入是患者的状态 s_t 和医生治疗行动 a_t，输出是在状态 s_t 下采取治疗 a_t 之后产生消极结果的可能性的相反数或产生积极结果的可能性，表示为 $Q_{w_d}(s_t, a_t)$ 和 $Q_{w_r}(s_t, a_t)$。

根据文献[10]，D-Critic 网络和 R-Critic 网络分别对应两个特殊的 MDP 过程 M_D 和 M_R。M_D 的奖励函数 r_d 设计为：如果 $z=1$，则最后一个时间步对应的奖励为 -1，其余为 0。M_R 的奖励函数 r_r 设计为：如果 $z=0$，则最后一个时间步对应的奖励为 $+1$，其余为 0。

通过最小化式(6)所示的损失函数来更新 D-Critic 网络参数。

$$J(w_d) = \mathbb{E}_{s_t, a_t \sim D}[(Q_{w_d}(s_t, a_t) - y_d^t)^2] \tag{6}$$

$$y_d^t = r_d(s_t, a_t) + \gamma_d Q'_{w_d'}(s_{t+1}, \mu'_{\theta'}(s_{t+1})) \tag{7}$$

同理，更新 R-Critic 网络参数：

$$J(w_r) = \mathbb{E}_{s_t, a_t \sim D}[(Q_{w_r}(s_t, a_t) - y_r^t)^2] \tag{8}$$

$$y_r^t = r_r(s_t, a_t) + \gamma_r Q'_{w_r'}(s_{t+1}, \mu'_{\theta'}(s_{t+1})) \tag{9}$$

其中，$Q'_{w_d'}$ 表示目标 D-Critic 网络，$Q'_{w_r'}$ 表示目标 R-Critic 网络，w_d，w_d'，w_r 和 w_r' 是神经网络的可学习参数。如果 s_t 是最后一个状态，则 $Q'_{w_d'}(s_{t+1}, \mu'_{\theta'}(s_{t+1})) = 0$，$Q'_{w_r'}(s_{t+1}, \mu'_{\theta'}(s_{t+1})) = 0$。

4.5 演员网络更新

演员网络使用全连接网络来学习确定性策略 $\mu_\theta(s)$，演员网络的输入是患者的状态 s_t，输出是治疗行动 a_t'，表示为 $\mu_\theta(s_t)$。演员网络的更新需要最大化预期回报、最小化与医生策略之间的差距、最小化产生消极结果的可能性以及最大化产生积极结果的可能性。将这些信息结合起来，则网络的目标是最大化式(10)所示的目标函数。

$$J(\mu_\theta) = J_1(\mu_\theta) + (-\alpha J_2(\mu_\theta)) + \beta J_3(\mu_\theta) \tag{10}$$

其中，$J_1(\mu_\theta)$ 表示最大化预期回报，$J_2(\mu_\theta)$ 表示最小化与医生策略的差距，$J_3(\mu_\theta)$ 表示最小化产生消极结果的可能性和最大化产生积极结果的可能性(即最大化 D-Critic 网络和 R-Critic 网络输出的 Q 值)。α 和 β 是参数，用于 3 个模块之间的平衡。通过梯度上升的方法更新网络的参数。

$$\theta = \theta + \varepsilon(\nabla_\theta J_1(\mu_\theta) + (-\alpha \nabla_\theta J_2(\mu_\theta)) + \beta \nabla_\theta J_3(\mu_\theta)) \tag{11}$$

其中，$\varepsilon \in [0,1]$ 是网络的学习率。

$J_1(\mu_\theta)$ 表示最大化累积奖赏值，即：

$$J_1(\mu_\theta) = \mathbb{E}_{s_t \sim D}[V^{\mu_\theta}(s_t)] = \mathbb{E}_{s_t \sim D}[Q_w(s_t, \mu_\theta(s_t))] \tag{12}$$

其中，D 表示医生策略下患者的数据。$J_1(\mu_\theta)$ 对参数 θ 求导，得到 $\nabla_\theta J_1(\mu_\theta)$，如式(13)所示：

$$\begin{aligned}\nabla_\theta J_1(\mu_\theta) &\approx \mathbb{E}_{s_t \sim D}[\nabla_\theta Q_w(s,a)|_{s=s_t, a=\mu_\theta(s_t)}] \\ &= \mathbb{E}_{s_t \sim D}[\nabla_\theta \mu_\theta(s)|_{s=s_t} \nabla_a Q_w(s,a)|_{s=s_t, a=\mu_\theta(s_t)}]\end{aligned} \tag{13}$$

其中，$J_2(\mu_\theta)$ 表示最小化网络预测的治疗行动 $\mu_\theta(s)$ 与医生推荐的治疗行动 a_t 之间的差距，使用 MSE 损失来实现，如式(14)所示：

$$J_2(\mu_\theta) = \mathbb{E}_{s_t \sim D}\left[\frac{1}{M}\sum_{m=1}^{M}(a_{t,m} - \mu_\theta^m(s))^2\right]|_{s=s_t} \tag{14}$$

其中，M 表示治疗药物的种类，$a_{t,m}$ 表示医生在时间步 t 采取的第 m 个治疗的值，通过链式法则对参数 θ 求导，得到 $\nabla_\theta J_2(\mu_\theta)$，如式(15)所示：

$$\nabla_\theta J_2(\mu_\theta) = \mathbb{E}_{s_t \sim D}\left[\frac{2}{M}\sum_{m=1}^{M}(\hat{a}_{t,m} - a_{t,m})\nabla_\theta \mu_\theta(s)\right]|_{s=s_t} \tag{15}$$

其中，$\hat{a}_{t,m}$ 表示演员网络推荐的治疗行动。

$J_3(\mu_\theta)$ 表示最小化产生消极结果的可能性和最大化产生积极结果的可能性，而 D-Critic 网络和 R-Critic 网络的输出 $Q_{w_d}(s,a)$ 和 $Q_{w_r}(s,a)$，分别对应于产生消极结果的可能性的相反数或产生积极结果的可能性。因此最小化产生消极结果的可能性和最大化产生积极结果的可能性等价于最大化 D-Critic 网络和 R-Critic 网络输出的 Q 值，如式(16)所示：

$$\begin{aligned}J_3(\mu_\theta) &= \mathbb{E}_{s_t \sim D}[V_d^{\mu_\theta}(s_t) + V_r^{\mu_\theta}(s_t)] \\ &= \mathbb{E}_{s_t \sim D}[Q_{w_d}(s_t, \mu_\theta(s_t)) + Q_{w_r}(s_t, \mu_\theta(s_t))]\end{aligned} \tag{16}$$

$J_3(\mu_\theta)$ 对参数 θ 求导，得到 $\nabla_\theta J_3(\mu_\theta)$，如式(17)所示：

$$\begin{aligned}\nabla_\theta J_3(\mu_\theta) &\approx \mathbb{E}_{s_t \sim D}[(\nabla_\theta Q_{w_d}(s,a) + \\ &\quad \nabla_\theta Q_{w_r}(s,a))|_{s=s_t, a=\mu_\theta(s_t)}] \\ &= \mathbb{E}_{s_t \sim D}[\nabla_\theta \mu_\theta(s)(\nabla_a Q_{w_d}(s,a) + \\ &\quad \nabla_a Q_{w_r}(s,a)|_{s=s_t, a=\mu_\theta(s_t)})]\end{aligned} \tag{17}$$

4.6 目标网络更新

每个网络都有一个对应的目标网络，用于提高训练的稳定性和收敛性，目标网络参数的更新计算式如式(18)—式(21)所示：

$$\theta' = \tau\theta + (1-\tau)\theta' \tag{18}$$

$$w' = \tau w + (1-\tau)w' \tag{19}$$

$$w'_d=\tau w_d+(1-\tau)w'_d \tag{20}$$

$$w'_r=\tau w_r+(1-\tau)w'_r \tag{21}$$

4.7 网络优化

本文采用两种措施来对网络进行优化[22]。

1)策略延迟更新：演员网络的更新频率比 3 个评论家网络的低，如评论家网络更新两次，演员网络更新 1 次，这样做的目的是先让评论家网络学好，才能更好地指导演员网络更新。

2)目标策略平滑：在构造评论家网络的目标值($Q'_{w'}(s_{t+1},\mu'_{\theta'}(s_{t+1}))$等)时，对目标动作 $\mu'_{\theta'}(s_{t+1})$等加入噪声，以帮助评论家网络学习。

4.8 DOSAC-DTR 算法

算法 1 给出了 DOSAC-DTR 算法的伪代码，主要包括编码器-解码器模型的训练和 DTR 策略的学习。其中编码器-解码器模型的训练为第 1－9 行，主要用于建模患者的观测数据历史和治疗史，得到患者的状态表征；DTR 策略的学习为第 10－22 行，主要是对 4 个核心网络以及 4 个目标网络进行联合训练更新。

算法 1 DOSAC-DTR

输入：患者轨迹数据 D_{raw}

输出：学到的策略 μ_θ

1. 初始化编码器-解码器的网络参数
2. REPATE
3. For t=1,T do
4. 将患者的观察数据 o_t 和治疗行动 a_{t-1}输入编码器中，得到患者状态 s_t
5. 将患者状态 s_t 和治疗行动 a_t 输入编码 器中，得到下一个时间步的预测观察数据 $\hat{o}_{t+1}$
6. 根据 MSE 损失(见式(3))，使用梯度下降更新网络参数
7. End For
8. End REPATE
9. 使用编码器，将患者轨迹数据以 $(s_t,a_t,r_t,s_{t+1},done)$的形式存储
10. 初始化网络参数 $\theta,\theta',w,w',w_d,w_d',w_r,w_r'$，演员网络更新频率 c，以及评论家网络的更新次数 $c_1=0$
11. REPATE
12. 采样患者数据$(s_t,a_t,r_t,s_{t+1},done)$
13. 得到 $\mu'_{\theta'}(s_{t+1})$并加入噪声
14. 根据式(6)、式(7)更新 D-Critic 网络
15. 根据式(8)、式(9)更新 R-Critic 网络
16. 根据式(4)、式(5)更新 Critic 网络
17. $c_1=c_1+1$
18. If c_1 mod c then
19. 根据式(10)－式(17)更新演员网络
20. End If
21. 根据式(18)－式(21)更新目标网络
22. End REPATE

5 实验结果与分析

5.1 数据集与环境配置

本文使用大型公开的真实世界数据集 MIMIC-III 进行实验[23]，构造了两个数据集，即 Sepsis 和 Ventilation。

1)Sepsis。根据文献[4]，本文提取 18 岁以上的脓毒症患者数据，他们在入院后首次在 ICU 就诊时就有脓毒症发病(根据脓毒症 3 标准)。对每位患者提取 44 维的观察数据，包括人口统计、生命体征和实验室值。观察数据汇总为 4h 的窗口，当多个数据点出现在一个窗口时，记录平均值或总和，对于缺失数据，使用 KNN 进行插补，如果没有类似的观测数据，则使用总体均值进行填充，并对所有数据进行归一化。本文学习了输入液体、血管加压剂的 DTR 策略，治疗动作空间被离散为 5×5 共 25 个动作。根据文献[5,24]，使用 SOFA 来计算奖励，奖励函数如式(22)所示：

$$r(s_t,a_t,s_{t+1})=-0.025\,\mathbb{I}(s_{t+1}^{SOFA}=s_t^{SOFA}\,\&\,s_{t+1}^{SOFA}>0)-0.125(s_{t+1}^{SOFA}-s_t^{SOFA}) \tag{22}$$

其中，$\mathbb{I}$ 表示指示函数，在最后一个时间步，如果 $z=1$，则给予＋15 的奖励，否则给予－15 的奖励。

2)Ventilation。根据文献[9,25]，本文提取所有接受有床通气超过 24h 的成年患者，对每位患者提取 48 维的观察数据，包括人口统计、生命体征和实验室值，并对提取的观测数据进行预处理(与 Sepsis 数据集一样)。本文学习了呼吸末正压(PEEP)、吸入氧分数(FiO2)和理想体重调整潮汐量(Vt)的 DTR 策略，将动作空间离散为 7×7×7＝343 个动作。奖励函数设计为：在最后一个时间步，如果 $z=1$，则给予－15 的奖励，如果 $z=0$，则给予＋15 的奖励，其余时间步奖励为 0。

5.2 评价指标

根据文献[5-6,26-27]，本文使用估计死亡率和 Jaccard 系数作为评价指标来评估模型的性能。

1)估计死亡率

具体来说，本文根据模型得到测试集中所有状态动作对的 Q 值，并将这些值离散为不同的区间，即$\{Q_1,Q_2,\cdots,Q_g\}$，其中 g 表示划分的区间数。每个区间对应的死亡率为$\{Mr_1,Mr_2,\cdots,Mr_g\}$，每个区间对应状态动作对的数量表示为$\{count_1,count_2,\cdots,count_g\}$，每个区间对应的死亡患者的状态动作对的数量表示为$\{Mcount_1,Mcount_2,\cdots,Mcount_g\}$。具体来说，对于一个患者进入 ICU 的例子，如果 $y=1$，假设该患者轨迹的所有状态动作对对应的 Q 值所在的区间集合为$\{Q_1,Q_2,Q_3\}$，则 $Mcount_1+1,Mcount_2+1,Mcount_3+1$。则死亡率的计算式为：

$$Mr_l=\frac{Mcount_l}{count_l},l=1,2,\cdots,g \tag{23}$$

根据死亡率$\{Mr_1,Mr_2,\cdots,Mr_g\}$，得到图 3 和图 4 所示的曲线。基于该曲线，计算出测试集中所有状态-动作对的 Q 值的均值，并将其对应区间的死亡率作为估计死亡率来评估模型的性能。

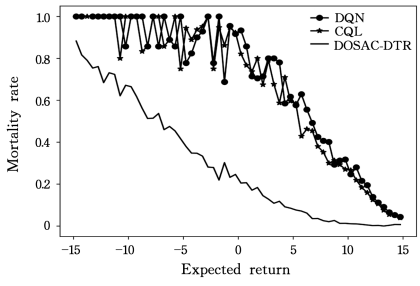


图 3 Sepsis 数据集上死亡率-预期回报曲线

Fig. 3 Mortality-expected return curve on Sepsis

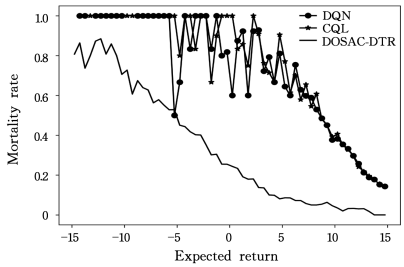


图 4 Ventilation 数据集上死亡率-预期回报曲线

Fig. 4 Mortality-expected return curve on Ventilation

2)Jaccard 系数

本文使用 Jaccard 系数来衡量不同策略与医生策略之间的相似性。测试集中医生的治疗行动可以表示为 $B=\{a_1, a_2, \cdots, a_j\}$，本文方法推荐的治疗行动可以表示为 $B_p=\{\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_j\}$，其中 j 是测试集中状态-动作对的个数，将 a_1 和 a_1' 等当作一个只有一个元素的集合，则 Jaccard 系数为：

$$Jaccard(B,B_p)=\frac{\sum_{l=1}^{j}(a_l==\hat{a}_l)}{j} \tag{24}$$

5.3 对比模型

为了验证 DOSAC-DTR 模型的性能，将其与 DQN[28]、DQN+编码器、CQL[7]、DDPG[29]、SDDPG 这 5 个模型进行对比，5 个模型的基本信息如下。

1)DQN：DQN 训练一个 Critic 网络，使用该网络估计在某个状态下所有动作的 Q 值，并最终选取 Q 值最大的动作。

2)DQN+编码器：在 DQN 的基础上，使用状态表征模块得到患者的状态表示，然后训练 Critic 网络。

3)CQL：CQL 是一种离线强化学习算法，在 Q 值的基础上增加一个正则化项，学习一个保守的 Q 函数。CQL 训练 Critic 网络，最终选取 Q 值最大的动作。

4)DDPG：基于 Actor-Critic 框架，Actor 网络学习确定性策略，Critic 网络学习价值函数，通过最大化预期回报来联合训练。

5)SDDPG：在 DDPG 的基础上加入医生监督，最大化预期回报的同时最小化与医生策略之间的差距。

5.4 性能测试与分析

5.4.1 总体性能

表 2 列出了在 Sepsis 和 Ventilation 数据集上本文方法与基线方法的对比。

表 2 整体性能对比

Table 2 Overall performance comparison

算法	Sepsis		Ventilation	
	估计死亡率/%	Jaccard	估计死亡率/%	Jaccard
DQN	5.21	0.090	14.33	0.004
DQN+编码器	5.15	0.184	14.04	0.005
CQL	5.10	0.098	14.26	0.004
DDPG	5.95	0.109	17.25	0.001
SDDPG	5.81	0.315	16.63	0.087
DOSAC-DTR	**3.51**	**0.362**	**13.74**	**0.126**

与基线方法对比的综合评价结果表明，本文方法的性能有明显提升。在 Sepsis 数据集上，估计死亡率可以达到 3.51%，Jaccard 系数可以达到 0.362。在 Ventilation 数据集上，估计死亡率可以达到 13.74%，Jaccard 系数可以达到 0.126。

DQN、DQN+编码器、CQL3 种方法的目标是最大化预期回报，它们需要估计所有的状态-动作对的值，因此在最终的结果中，其预期返回值甚至可以达到最大值 15，这就造成了对 Q 值的严重高估，即对数据集中不存在的状态-动作对的值的高估，最终也造成结果的不准确。虽然预期返回值很大，但是估计死亡率和 Jaccard 系数都没有本文方法好。DDPG 不需要估计数据集中不存在的状态-动作对的值，从一定程度上缓解了 Q 值估计的不准确性，但是它并没有考虑医生的监督，因此结果也不如本文方法。SDDPG 在 DDPG 的基础上加入了医生的监督，性能有了一定的提升，但是由于只考虑了最优性而没有考虑风险性，因此效果也不如本文方法。同时从 DQN 与 DQN+编码器两种方法的对比结果可以看出，考虑患者的历史观测数据和治疗史可以学习到更好的策略，进一步降低估计死亡率，提高 Jaccard 系数。

5.4.2 策略有效性

图 3 和图 4 分别给出了在 Sepsis 数据集和 Ventilation 数据集上预期回报(Q 值)和死亡率之间的关系。可以看出，本文方法相比 DQN 和 CQL 具有更明显的负相关。对于 DQN 和 CQL，不管患者最终是死亡还是出院，对应的 Q 值都较大，这就导致在[−15,0]之间对应的死亡率在 1 附近波动(在 Q 值较小的地方对应的死亡患者的状态-动作对数量为个位数，而在 Q 值较大的地方对应数量是两位数和三位数)。而本文方法在 Q 值较小的地方对应的死亡患者的状态-动作对数量为两位数或三位数，而在 Q 值较大的地方对应数量是个位数，而对于出院患者，则相反。

图 5 给出了 Sepsis 中的两种治疗 IV 和 VP 上的死亡率-学到的策略与医生策略之间的关系，可以看到，对于不同方法学习到的策略，当与医生的策略差距为 0 时，死亡率最低。但是本文方法的曲线的相关性更明显。

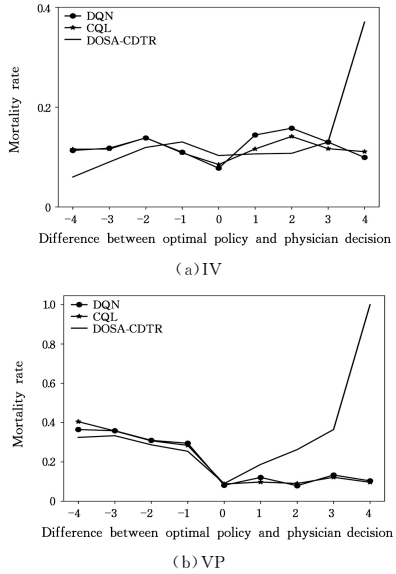
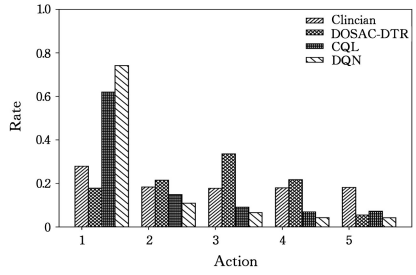


图 5 死亡率-学到的策略与医生策略的不同曲线

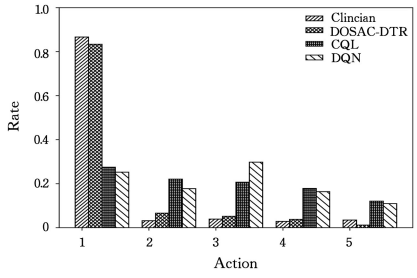
Fig. 5 Different curve of mortality rate vs. learned policy and physician policy

图 6 给出了 Sepsis 测试集中的两种治疗 IV 和 VP 上的

动作分布，可以看到，相比基线方法，本文方法学到的策略与医生策略更为相似。



(a)IV



(b)VP

图 6 测试集上医生策略和学到的策略的动作分布

Fig. 6 Action distribution of physician policy and learned policy on the test set

5.4.3 参数影响

表 3 列出了在 Sepsis 数据集上参数 α 和 β 对实验结果的影响，可以看出，当 $\alpha=4,\beta=1$ 和 $\alpha=2,\beta=15$ 时取得最优结果，获得了最高的 Jaccard 系数和最低的估计死亡率。这证明本文方法不仅可以显著降低估计死亡率，还能提高与医生策略的相似度。

表 3 参数影响

Table 3 Parameter influence

β	α	估计死亡率/%	Jaccard	β	α	估计死亡率/%	Jaccard
1	1	5.54	0.360	7	1	5.69	0.363
1	2	4.27	0.360	7	2	5.98	0.363
1	4	3.51	0.362	7	4	6.29	0.362
2	1	5.74	0.362	15	2	3.68	0.365
2	2	6.11	0.359	15	4	3.68	0.361
2	4	6.17	0.360				

5.4.4 消融实验

表 4 列出了消融实验结果，DOSAC-DTR-O 表示去掉网络优化措施，DOSAC-DTR-DR 表示去掉 D-Critic 网络和 R-Critic 网络。实验结果表明了网络优化措施以及 D-Critic 网络和 R-Critic 网络的有效性。

表 4 消融实验结果的比较

Table 4 Comparison of ablation results

算法	Sepsis		Ventilation	
	估计死亡率/%	Jaccard	估计死亡率/%	Jaccard
DOSAC-DTR-L	5.71	0.304	15.31	0.123
DOSAC-DTR-O	5.73	0.355	15.33	0.126
DOSAC-DTR-DR	4.88	0.357	14.62	0.124
DOSAC-DTR-DR-O	5.81	0.315	16.63	0.087
DOSAC-DTR-DR-O-BC	5.95	0.109	17.25	0.001
DOSAC-DTR	3.51	0.362	13.74	0.126

结束语 本文提出了 DOSAC-DTR 模型，融合 Dead-ends 和离线监督 Actor-Critic 方法，并考虑了患者的历史观测数据和治疗史。首先，单独训练基于 LSTM 的编码器解码器模型，对患者的历史观测数据和治疗史进行建模，从而学习到包括患者关键历史信息的状态表征。其次，将 Dead-ends 概念和医生的监督融入 Actor-Critic 框架中，演员网络的目标是最大化评论家网络输出的长期目标、最小化与医生推荐的治疗之间的 MSE 损失以及最大化 D-Critic 网络和 R-Critic 网络输出的值。4 个网络联合训练，以学习 DTR。实验结果表明，本文提出的 DOSAC-DTR 模型有更好的性能。

参 考 文 献

[1] RIACHI E, MAMDANI M, FRALICK M, et al. Challenges for Reinforcement Learning in Healthcare[J]. arXiv: 2103.05612, 2021.

[2] CORONATO A, NAEEM M, DE PIETRO G, et al. Reinforcement learning for intelligent healthcare applications: A survey [J]. Artificial Intelligence in Medicine, 2020, 109: 101964.

[3] YU C, LIU J, NEMATI S, et al. Reinforcement learning in healthcare: A survey[J]. ACM Computing Surveys (CSUR), 2021, 55(1): 1-36.

[4] MATTHIEU K, LEO A C, OMAR B, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care[J]. Nature Medicine, 2018, 24(11): 1716.

[5] RAGHU A, KOMOROWSKI M, AHMED I, et al. Deep reinforcement learning for sepsis treatment[J]. arXiv: 1711.09602, 2017.

[6] WANG L, ZHANG W, HE X, et al. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2447-2456.

[7] KAUSHIK P, KUMMETHA S, MOODLEY P, et al. A conservative Q-learning approach for handling distribution shift in sepsis treatment strategies[J]. arXiv: 2203.13884, 2022.

[8] FUJIMOTO S, GUS S. A minimalist approach to offline reinforcement learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 20132-20145.

[9] YIN C, LIU R, CATERINO J, et al. Deconfounding Actor-Critic Network with Policy Adaptation for Dynamic Treatment Regimes[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 2316-2326.

[10] FATEMI M, KILLIAN T W, SUBRAMANIAN J, et al. Medical Dead-ends and Learning to Identify High-risk States and Treatments[C]// Advances in Neural Information Processing Systems 34. 2021.

[11] TESAURO G. Programming backgammon using self-teaching neural nets[J]. Artificial Intelligence, 2002, 134(1/2): 181-199.

[12] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[13] REDDY G, CELANI A, SEJNOWSKI T J, et al. Learning to soar in turbulent environments[J]. Proceedings of the National

Academy of Sciences, 2016, 113(33): E4877-E4884.
[14] JETER R, JOSEF C, SHASHIKUMARS, et al. Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care? [J]. arXiv: 1902.03271, 2019.
[15] LIANG D, DENG H, LIU Y. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach[J]. Applied Intelligence, 2023, 53(9): 11034-11044.
[16] YU C, REN G, DONG Y. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units[J]. BMC Medical Informatics and Decision Making, 2020, 20(3): 1-8.
[17] THOMASP S. Safe reinforcement learning [R]. University of Massachusetts Libraries, 2015.
[18] THOMAS P S, CASTRO DA SILVA B, BARTO A G, et al. Preventing undesirable behavior of intelligent machines [J]. Science, 2019, 366(6468): 999-1004.
[19] LAROCHE R, TRICHELAIR P, DES COMBES R T. Safe policy improvement with baseline bootstrapping[C]// International Conference on Machine Learning. PMLR, 2019: 3652-3661.
[20] FATEMI M, SHARMA S, VAN SEIJEN H, et al. Dead-ends and secure exploration in reinforcement learning[C]// International Conference on Machine Learning. PMLR, 2019: 1873-1881.
[21] TAYLOR W K, HAORAN Z, JAYAKUMAR S, et al. An empirical study of representation learning for reinforcement learning in healthcare[C]// Machine Learning for Health. PMLR, 2020: 139-160.
[22] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]// International Conference on Machine Learning. PMLR, 2018: 1587-1596.
[23] JOHNSON A E W, POLLARD T J, SHEN L, et al. MIMIC-III, a freely accessible critical care database[J]. Scientific Data, 2016, 3(1): 1-9.
[24] NANAYAKKARA T, CLERMONT G, LANGMEAD C J, et al. Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment[J]. PLOS Digital Health, 2022, 1(2): e0000012.
[25] PEINE A, HALLAWA A, BICKENBACH J, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care[J]. NPJ Digital Medicine, 2021, 4(1): 1-12.
[26] WENG W H, GAO M, HE Z, et al. Representation and reinforcement learning for personalized glycemic control in septic patients[J]. arXiv: 1712.00654, 2017.
[27] ZHANG Y, CHEN R, TANG J, et al. LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 1315-1324.
[28] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[29] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [J]. arXiv: 1509.02971, 2015.

YANG Shasha, born in 2000, postgraduate. Her main research interests include reinforcement learning and dynamic treatment regime.

YU Yaxin, born in 1971, Ph.D, associate professor. Her main research interests include data mining and social network.

(责任编辑:喻藜)