天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于强化学习和多决策偏好优化的医疗决策推荐模型
作者：王嫄，盛梦茹，温辉，熊宁，程毅松，吴骏
DOI：10.13364/j.issn.1672-6510.20250151
收稿日期：2025-10-03
网络首发日期：2026-07-09
引用格式：王嫄，盛梦茹，温辉，熊宁，程毅松，吴骏．基于强化学习和多决策偏好优化的医疗决策推荐模型[J/OL]．天津科技大学学报．
https://doi.org/10.13364/j.issn.1672-6510.20250151

DOI：10.13364/j.issn.1672-6510.20250151

基于强化学习和多决策偏好优化的医疗决策推荐模型

王　嫄[1]，盛梦茹[1]，温　辉[1]，熊　宁[1,2]，程毅松[3]，吴　骛[3]
(1. 天津科技大学人工智能学院，天津 300457；2. 天津科技大学机械工程学院，天津 300457；
3. 四川大学华西医院，成都 610041）

摘　要：基于交互优化的强化学习能够有效模拟医生在连续时间点上对患者状态进行观察、干预与反馈。然而，目前用于医疗决策的强化学习架构只能利用显式独立的患者状态信号进行策略学习，未能建模隐式的多诊疗决策偏好复杂关联信息。这导致个体患者信息存在无法利用全局知识信息、模型行为不稳定、诊疗决策反馈不及时的问题。针对此问题，本文在强化学习框架下提出一种多决策偏好优化方法，命名为 DP-DDQN（decision preferences based on double deep Q-network），通过逆强化学习建模隐式的全局决策偏好，应用全局决策偏好更新患者状态表示，将显式的个性化诊疗偏好和隐式的全局决策偏好联合建模为奖励函数指导模型学习策略。结果表明，本文模型在脓毒症抗生素种类剂量推荐任务中可以使患者生存率提升至 84.01%，较临床提高了 17.45%，输出的治疗方案与临床实践规范高度契合，且可以合理建议药物使用剂量。

关键词：逆强化学习；决策偏好；脓毒症；抗生素剂量推荐
中图分类号：TP181；TP399　　**文献标志码：**A　　**文章编号：**1672-6510 (0000)00-0000-00

Medical Decision Recommendation Model via Reinforcement Learning and Multi-Decision Preference Optimization

WANG Yuan[1], SHENG Mengru[1], WEN Hui[1], XIONG Ning[1,2], CHENG Yisong[3], WU Qin[3]
(1. College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China；
2. College of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin 300457, China;
3. Department of Critical Care Medicine, West China Hospital, Sichuan University, Chengdu 610041, China)

Abstract: Interactive optimization-based reinforcement learning (RL) effectively simulates doctors' continuous observation, intervention, and feedback on patient status. However, existing medical decision-making RL architectures only use explicit, independent patient status signals for policy learning, failing to model complex implicit associations among multiple diagnostic and treatment decision preferences. This leads to individual patients being unable to utilize global knowledge, unstable model behavior, and delayed feedback in clinical decisions. To solve this, this paper proposes a multi-decision preference optimization method under the RL framework, named DP-DDQN (decision preferences based on double deep Q-network). It uses inverse reinforcement learning to model implicit global decision preferences, updates patient status representations with such preferences, and jointly takes explicit personalized treatment goals and implicit global decision preferences as the reward function to guide policy learning. Experiments on sepsis antibiotic type and dose recommendation show that the proposed model increases patient survival rate to 84.01%, and 17.45% improvement over the clinical, with

收稿日期：2025–10–03；**修回日期：**2025–12–15
基金项目：国家重点研发计划项目（2022YFC2504500）;天津市自然科学基金资助项目（24JCYBJC00890）
作者简介：王嫄(1989—)，女(汉族)，山西万荣人，副教授，博士，wangyuan23@tust.edu.cn

treatment plans highly consistent with clinical guidelines and reasonable drug dosage suggestions.

Key words: inverse reinforcement learning；decision preference；sepsis；antibiotic dosage recommendation

重症医学场景中，患者因疾病严重程度、个体差异及药物敏感性不同，治疗风险和预后差异显著。传统经验性诊疗方案基于群体平均结果，难以快速实现个性化治疗[1]。重症患者病情复杂多变，需开展抗感染、液体复苏、营养支持及机械通气等综合治疗[2]。现有诊疗多依赖患者临床生理指标等显式独立信息，未充分考量指标间的复杂关联。此外，免疫功能低下患者的经验性抗生素治疗虽然能覆盖常见致病菌，却难以精准应对个体特殊感染[3]。这些问题导致部分患者因治疗不及时或不精准面临更高的死亡风险，因此亟需构建综合多维度信息的模型，实现个性化治疗的快速响应。

强化学习（reinforcement learning，RL）是通过与环境交互，学习最优策略以最大化长期累积奖励的方法[4]。在医疗决策的强化学习建模中，状态输入为患者临床数据及治疗历史反馈，动作输出为时序化用药和治疗手段，环境是患者特征与治疗轨迹集合，奖励则是治疗手段的正负效果反馈。借助奖励机制，模型可持续学习调整决策，以实现长期治疗效果最大化。强化学习为医疗任务建模可以动态适配患者病情波动，通过迭代优化治疗策略，尤其适用于需长期动态调整方案的医疗场景。

近年来，强化学习已应用于抗生素种类组合推荐[5]、静脉输液、血管升压药剂量推荐[6-8]和手术方案选择[9]等医疗研究中，展现出治疗策略推荐的潜在价值。脓毒症的病死率高，病情进展快，患者个体差异大且需动态调整治疗方案[10-11]，因此本文针对脓毒症治疗开展研究。Wang 等[5]针对脓毒症抗生素使用依赖医生经验易导致治疗不足或过度的问题，提出基于治疗指南的深度强化学习模型，将临床知识融入奖励函数指导深度 Q 网络生成个性化抗生素组合建议，该模型能提供符合指南的合理方案并缩短抗生素使用时长，但是该研究未关注剂量问题。Wu 等[12]提出嵌入专家知识的加权对决双深度 Q 网络，用于脓毒症患者静脉输液和血管升压药剂量推荐，为临床提供了可靠辅助工具，但是其奖励函数仅考虑生存率和序贯器官衰竭评估（SOFA）得分，无法全面反映病情复杂性与治疗多目标性。Choi 等[13]提出用于脓毒症治疗策略优化的深度强化学习模型，提升离线环境下的稳定性与准确性，但是该研究仅聚焦静脉注射和血管加压药物剂量，未涉及药物类型等更复杂决策。Barata 等[9]为提升皮肤癌诊断系统性能，提出融合监督学习与多类概率的深度强化学习模型，以专家奖励表的形式整合人类偏好优化诊断决策，但是需专家手动设置，耗时耗力且可能因个人偏见导致奖励值主观性较强。

上述研究表明，医疗决策生成的强化学习方法具有潜在价值，但是只能利用患者生存状态等显式独立的患者信号作为个性化诊疗偏好进行策略学习[6-8]，无法建模脓毒症患者液体复苏量与血压响应的动态关系等隐式多诊疗决策偏好的复杂关联。现有方法聚焦显式独立信息，缺乏多维度数据关联建模，导致模型无法捕捉临床决策关键的隐式关联信息。此外，抗生素不合理使用使耐药菌流行问题日益严峻且缺乏高效普适的解决方案[10]。目前相关研究仅涉及抗生素种类推荐，尚未覆盖剂量优化。

针对上述问题，本文提出一种多决策偏好优化方法，在患者状态表示和奖励建模中，将显式个性化诊疗偏好和隐式的全局决策偏好融入双深度 Q 网络（double deep Q network，DDQN）强化学习框架中，实现脓毒症患者的抗生素种类和剂量组合推荐，命名为 DP-DDQN（decision preferences based on double deep Q-Network，DP-DDQN），该方法通过逆强化学习（inverse reinforcement learning，IRL）建模脓毒症治疗历史数据，挖掘患者状态、治疗方案与结果的隐含复杂关系，形成兼顾显式偏好与全局治疗经验的隐式全局决策偏好，为临床决策提供全面参考。

1　相关算法模型

1.1　强化学习和逆强化学习

强化学习（RL）通过与环境交互学习最优策略，其核心概念是智能体感知自身当前状态并采取行动，环境则反馈奖励信号，智能体的目标是最大化整个交互过程中的累积奖励[4]。逆强化学习（IRL）是强化学习的扩展，指智能体通过学习专家策略轨迹中状态和动作之间的关系，推断并优化奖励函数的过程[14]。该奖励函数是衡量专家决策价值的关

键，不仅可以揭示专家策略、环境状态与动作的内在联系，还可以动态调整特征权重、优化特征表示，进而提升医疗决策的准确性与个性化水平。强化学习和逆强化学习的示意图如图 1 所示。

强化学习和逆强化学习问题可以建模成马尔可夫决策过程（Markov decision process，MDP）。MDP 是描述序贯决策问题的数学框架，可用一个五元组 $<S,A,P,R,\gamma>$表示，其中：S 为状态空间，包含所有状态可能，每个状态 $s\in S$ 表示环境的一个特定配置或情况；A 为动作空间，包含所有动作可能，每个动作 $a\in A$ 表示在某个状态下采取的特定行为；P 为状态转移概率，$P(s'/s,a)$表示在状态 s 下采取动作 a 后，转移到状态 s'的概率；R 为奖励函数，$R(s,a,s')$表示在状态 s 下采取动作 a 并转移到状态 s'后，获得的奖励；γ 为折扣因子，表示未来奖励的折扣，取值范围在[0,1]之间，γ=0 表示只考虑当前奖励，γ=1 表示考虑所有未来奖励。

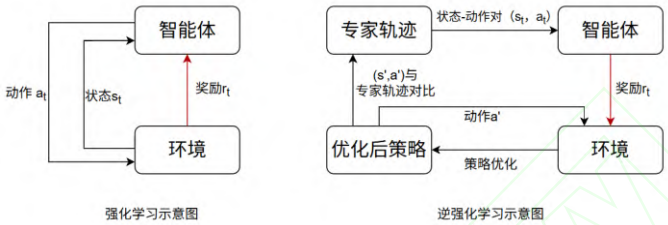

图 1　强化学习和逆强化学习示意图

Fig. 1　IRL and RL Schematic Diagram

强化学习按照学习目标和优化方式可分为基于价值与基于策略两类。基于价值的强化学习通过学习状态或状态-动作对的价值，更新价值函数间接优化策略。典型算法有 Q 学习（Q-learning）算法、SARSA（state-action-reward-state-action）算法、深度 Q 网络（deep Q-network，DQN）、DDQN 等。基于策略的强化学习则直接优化策略函数，以最大化期望回报，典型算法有近端策略优化（PPO）、信任区域策略优化（TRPO）等。DDQN 继承了 DQN 的深度学习能力，可高效处理高维状态空间，适配脓毒症复杂临床环境，且结合经验回放机制打破了数据相关性，为脓毒症临床决策提供可靠依据。逆强化学习主要包括最大边际、最大熵、贝叶斯逆强化学习等方法。最大熵逆强化学习（maximum entropy inverse reinforcement learning，MaxEnt IRL）利用最大熵原理，更适用于连续空间并减轻专家轨迹次优性影响。结合本文的具体任务，选择最大熵逆强化学习[15]。

1.1.1　双深度 Q 网络

在强化学习中，DQN[16]通过神经网络逼近 Q 函数，有效解决了传统 Q 学习在高维状态空间局限性。但在 Q 值更新过程中，动作选择与目标 Q 值计算依赖同一网络，会造成 Q 值过估计问题[17]。为此，DDQN[18]引入“双网络”机制，利用独立的评估网络与目标网络分别负责动作选择和目标 Q 值计算，缓解过估计问题，提升算法的稳定性与收敛性。

算法具体流程为：给定状态空间 S 和动作空间 A，构建结构相同但参数独立的在线网络 Q_θ 和目标网络 Q_{θ^-}，分别用于动作选择和动作价值估计，同时初始化经验回放缓冲区 D、折扣因子 γ 及学习率 η。使用在线网络 Q_θ 选择下一个状态 s_{t+1} 的最佳动作，即

$$\mathrm{a}'=\arg\max_a Q_\theta(s_{\mathrm{t}+1},a) \tag{1}$$

使用目标网络 Q_{θ^-} 计算目标 Q 值，为

$$\mathrm{y_t}=r_t+\gamma Q_{\theta^-}(s_{t+1},a'), \tag{2}$$

其中，$\mathrm{r_t}$ 为 t 时刻的奖励值。再使用均方误差作为损失函数，对 Q_θ 进行梯度下降，为

$$L(\theta)=\frac{1}{N}\sum_{\mathrm{i}}(y_i-Q_\theta(s_i,a_i))^2. \tag{3}$$

通过上述损失函数，可以得到参数 θ 最优值和最优策略。DDQN 通过解耦动作选择与价值评估，缓解了 Q 学习中最大化偏差和 Q 值过估计问题。

1.1.2　最大熵逆强化学习

最大熵逆强化学习（MaxEnt IRL）[19]是 IRL 的变体，旨在从人类专家示范轨迹中反向推断驱动其行为的奖励函数。该奖励函数本质是描述状态-动作对的价值，可将专家治疗行为转化为可计算的价值信号，通过保证专家轨迹累积奖励最大且避免过拟合，建立专家行为与动机的隐含关联。

MaxEnt IRL 核心算法流程：初始化奖励 r_0，并得到策略的熵，为

$$H(\pi(\cdot|s)) = -\sum_a \pi(a|s)\log\pi(a|s) \quad (4)$$

其中：s 为当前状态，a 为当前动作，$\pi(a|s)$ 为策略。熵越大，表示策略在状态 s 下的动作选择越随机多样，反之表示策略越确定。MaxEnt IRL 的核心是最优策略要最大化累积奖励且最大化策略的熵。

基于当前奖励 r，求解最大熵目标下的最优策略 π^*，为

$$\pi^*(a|s) \propto \exp\left(\mathrm{E}[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s,a_0=a]\right) \quad (5)$$

其中：γ 为奖励折扣因子。状态 s 下选择动作 a 的概率与该动作后续能获得的折扣累积奖励的指数成正比，奖励越高，选择概率越大，但仍保留一定的随机性。

奖励的更新需要计算专家示范轨迹 τ 在当前策略 π 下的生成概率，即

$$P(\tau|\pi) = p(s_0)\prod_{t=0}^{T}\pi(a_t|s_t)P(s_{t+1}|s_t,a_t) \quad (6)$$

其中：专家轨迹 $\tau=(s_0,a_0,s_1,a_1,...,s_T,a_T)$，$p(s_0)$为初始状态分布，$P(s_{t+1}|s_t,a_t)$ 为环境状态转移概率，再对其取自然对数，得到对数概率，为

$$\log P(\tau|\pi) = \log p(s_0) + \sum_{t=0}^{T}[\log\pi(a_t|s_t) + \log P(s_{t+1}|s_t,a_t)] \quad (7)$$

其中：环境转移概率 $\log P(s_{t+1}|s_t,a_t)$ 和初始状态概率 $\log p(s_0)$ 是与策略无关的常数，不随奖励 r 变化，策略对数项 $\log\pi(a_t|s_t)$ 会随着奖励 r 的调整而变化。

2　模　型

2.1　任务形式化建模

医疗决策推荐的强化学习的建模基于 MDP 五元组<*S,A,P,R,γ*>映射。状态空间 S 由患者人口学特征、生命基本特征、实验室检查结果构成的动态病情集合，全面反映患者当前病理状态。动作空间 A 包括抗生素种类与剂量、液体复苏方案、机械通气设置等具体操作。状态转移概率 $P(s'|s,a)$表示当前状态 s 下采取动作 a 后病情转移至新状态 s'的条件概率。奖励函数 $R(s,a,s')$为治疗效果反馈，若 s 经 a 转移至 s'后病情好转情况则赋予正向奖励，反之则为负向奖励。在强化学习策略推荐中，MDP 作为输入提供完整的序贯决策问题形式化描述，最终输出模型生成的治疗推荐策略 $\pi(a|s)$。

医疗决策使用连续状态空间的强化学习模型可以更好地处理复杂的医疗数据，提高模型的性能和泛化能力[20]，因此使用长短期记忆网络[21]对患者特征进行映射，得到 36 维特征序列，再将患者住院 90 天内死亡结果作为最终状态（死亡为 0，生存为 1），结合降维后患者特征序列和最终生存情况得到患者状态空间。

另外，从重症医学信息数据集（medical information mart for intensive care，MIMIC-IV）中筛选出临床应用频率最高的 31 种抗生素，涵盖了 β-内酰胺类、大环内酯类、氨基糖苷类等 9 大类抗生素，作为后续构建用药动作空间的基础，并使用 k 均值聚类（k-means）算法对患者用药序列进行聚类，生成患者的动作空间。

2.2　模型框架

模型由两大部分构成，分别是基于医疗知识模式和逆强化学习的多决策偏好建模，以及强化学习框架下多决策偏好优化的医疗决策推荐。模型架构如图 2 所示。

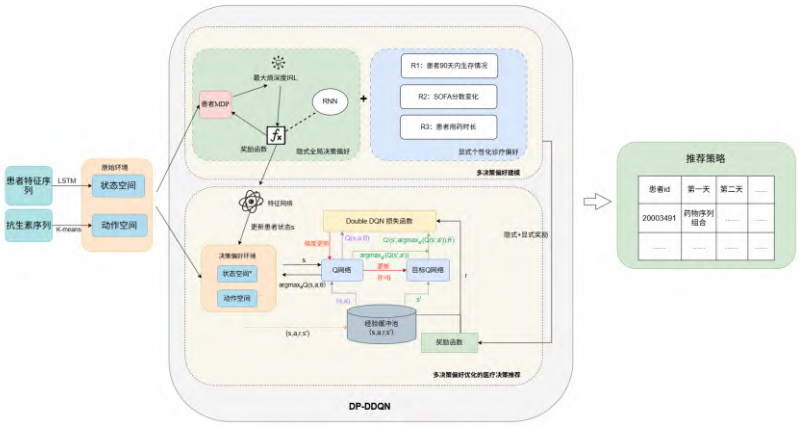

图 2　DP-DDQN 模型框架图

Fig. 2　Framework diagram of DP-DDQN model

多决策偏好建模部分包含显式个性化诊疗偏好和隐式全局决策偏好。显式个性化诊疗偏好关注患者治疗过程中的关键指标。隐式全局决策偏好则通过最大熵逆强化学习建模，学习临床专家的医疗决策，捕捉患者特征与临床治疗方案之间的隐含关系。

多决策偏好优化的医疗决策推荐部分将隐式决策偏好和原始状态空间输入到特征网络中，更新患者状态表示，得到新的状态空间。在模型学习过程中，结合显式和隐式决策偏好，构建多决策偏好的奖励函数，精准反映患者的真实治疗偏好并平衡不同治疗目标，指导模型策略推荐。

2.3　全局决策偏好建模

在脓毒症临床决策过程中，医生的决策偏好通常是基于患者的个体化情况、当前的医学证据以及指南推荐的治疗策略形成的。这些决策偏好在具体实践中表现为医生会根据患者的具体状态而选择不同的用药种类和剂量，例如，当患者状态变严重时，医生会加大某种特定药物的使用剂量。通过逆强化学习模型对患者数据进行学习，可以提取出不同患者状态和用药之间的复杂隐含关系，从而得到隐式的决策偏好。这种决策偏好不仅有助于在实际治疗中提供更个性化、更精确的治疗方案，还可以用来评估和调整现有的治疗策略，确保决策更符合患者的状态。

在全局决策偏好建模中，采用最大熵逆强化学习对患者数据集进行学习，得到患者的隐式决策偏好。这种隐式偏好是指在脓毒症和脓毒性休克的整个诊疗过程中，临床医生基于当前最佳证据、指南和患者个体情况，在一系列关键决策点上优先采取某些策略或行动的总体倾向或原则。具体流程如下：

首先，将原始环境建模成患者 MDP 五元组 $<S,A,P,R,\gamma>$ 输入到最大熵逆强化学习模型中，得到患者轨迹空间 E，其包含多条轨迹，每条轨迹是状态-动作对序列，为

$$\tau=\{(s_1,a_1),(s_2,a_2)...(s_k,a_k)\} \tag{8}$$

其中：(s_k,a_k) 为患者状态-动作对。轨迹空间 E 中每条轨迹的概率为

$$P(\tau|R)=\frac{1}{Z(R)}\exp(\sum_{t=1}^{k}R(s_t,a_t)) \tag{9}$$

其中：$P(\tau|R)$ 表示奖励函数 R 下生成轨迹的概率，$Z(R)$ 为配分函数，确保轨迹概率和为 1。

然后，对患者轨迹概率使用最大似然并代入最大熵分布，为

$$\begin{aligned}L(\theta)&=\sum_{\tau\in E}\log\ p(\tau|\theta)\\&=\sum_{\tau\in E}\log\frac{1}{Z}\exp(R_\theta(\tau))\\&=\sum_{\tau\in E}\log R_\theta(\tau)-M\log\sum_{\tau\in D}\exp(R_\theta(\tau))\end{aligned} \tag{10}$$

其中：M 为轨迹条数；D 为模型当前学习的轨迹空间；$R(\theta)$ 为累积奖励，累积奖励越大，表示模型学习的结果越好，即对患者状态与用药之间复杂隐含关系的学习越准确。θ 为奖励权重参数。

最后，计算奖励函数参数 θ 的梯度得到归一化损失函数，为

$$\nabla\bar{L}=\frac{1}{M}\sum_{\tau\in E}\frac{dR_\theta(\tau)}{d\theta}-\sum_{\tau'\in D}\left[p(\tau'|\theta)\frac{dr_\theta(\tau')}{d\theta}\right] \tag{11}$$

通过减小损失函数优化奖励权重参数 θ，模型不断更新并优化奖励函数 R，即隐式全局决策偏好。由于最大熵逆强化学习中使用神经网络建模奖励函数，因此选择循环神经网络建模奖励函数。循环神经网络能够捕捉到患者状态随时间变化的趋势，以及不同时间点之间患者状态与治疗之间的依赖关系。

2.4　多决策偏好优化

多决策偏好优化通过隐式全局决策偏好更新患者状态表示，结合显式决策偏好构建多决策奖励函数，用于强化学习决策。状态表示优化中，将隐式决策偏好与原始状态空间输入特征网络，得到新状态空间（状态空间*），结合原始动作空间形成决策偏好环境，帮助模型全面捕捉患者状态与治疗动作的隐含关系。在奖励函数建模中，融合显式个性化诊疗偏好与隐式全局决策偏好，构建多决策奖励函数，精准反映患者真实治疗偏好、平衡不同治疗目标，实现更优的医疗决策推荐。

奖励函数包括显式决策奖励（患者最终存活情况、SOFA 分数变化、用药时长）和全局隐式奖励，奖励函数设置见表 1，共构建 4 个奖励函数 R_1、R_2、R_3、R_4，最终综合奖励函数为 R=R_1+R_2+R_3+R_4。其中，R_1 聚焦患者 90 天的生存情况，存活奖励 100、死亡惩罚 100（生存是核心目标）；R_2 依据 SOFA 分数变化（脓毒症关键特征），分数降低奖励 10，分数升高惩罚 10；R_3 参考平均用药时长，超过则惩罚 10，低于则不奖励（避免模型为获奖励缩短用药影响疗效）；R_4 涵盖患者状态与治疗动作的隐含关系，助力模型捕捉最优治疗方案。

表 1 奖励函数设置

Tab. 1 Setting of reward function

奖励函数	目的	内容	设置
R_1	使最优策略选择更关注患者真实的预后情况	患者 90 天内的生存情况	生存：+100；死亡：-100
R_2	使模型改善患者状态和预后	SOFA 分数变化	SOFA 分数减小：+10；SOFA 分数增大：-10
R_3	使模型降低患者的用药时长	用药时长（患者轨迹长度）	大于 7 天：-10
R_4	使模型学习患者状与动作之间复杂关联	通过逆强化学习得到脓毒症治疗全局决策偏好	R=R_{dp}

3 实验结果及分析

3.1 数据集

重症医学信息数据集（medical information mart for intensive care，MIMIC）[22]中包含重症监护室患者生命体征、实验室检查结果、用药信息和病史等临床信息。本研究使用 MIMIC-IV 数据集，从中筛选了符合脓毒症 3 标准的脓毒症患者，并排除年龄小于 18 岁的患者。最终选取 8524 例脓毒症患者，并筛选出包括人口学特征、患者生命体征、实验室指标等 43 个脓毒症患者特征[12]。MIMIC-IV 用作训练和测试数据集。

电子重症监护协作研究数据（emergency intensive care unit，eICU）[23]中包括生命体征、实验室测量、药物、患者病史等信息。本文在 eICU 数据集中选取 2054 例患者，选择与 MIMIC-IV 数据集中重合的特征，包括 23 个实验室项目、8 个生命体征和人口学特征。eICU 作为外部测试数据集。

MIMIC-IV 和 eICU 数据集的脓毒症患者基本信息统计见表 2。

表 2 MIMIC-IV 和 eICU 数据集的脓毒症患者基本信息统计

Tab. 2 Basic information statistics of sepsis patients of MIMIC-IV and eICU dataset

患者	男性占比/%		平均年龄/岁		人数/人		死亡率/%	
	MIMIC-IV	eICU	MIMIC-IV	eICU	MIMIC-IV	eICU	MIMIC-IV	eICU
生存患者	61.57	55.96	66.78	63.32	5728	1544	—	—
死亡患者	55.64	51.18	69.45	65.50	2796	510	—	—
患者总数	59.58	54.77	67.68	63.11	8524	2054	32.80	24.83

3.2 模型评估

选择 3 种强化学习方法 DQN、Q 学习和 SARSA 作为基线模型，并将模型的重要性采样评估、用药时长、Q 值变化、患者预后情况、药物剂量分析、预后与临床数据进行比较。

3.2.1 重要性采样评估

加权重要性采样（weight importance sample，WIS）[24]是一种用于估计策略价值的技术。在强化学习中，WIS 有助于评估某一策略的性能，借助 WIS，强化学习无需实时交互就能评估脓毒症治疗策略。

模型与临床患者（clinical）之间的 WIS 对比如图 3 所示，图中的红色水平线表示临床测试集的平均 WIS 分数，值为 59.20。图 3 展示了 DP-DDQN 模型与 DQN、Q 学习和 SARSA 的 WIS 分数对比，以及 DP-DDQN 在验证集 eICU 上的 WIS 分数。随着训练轮数（Epoch）增加，所有模型的 WIS 分数都呈上升趋势，最终趋于平稳状态，且最终都超过了临床评分，其中 DP-DDQN 模型最明显。在 150 个 Epoch 后，4 个模型基本趋于平稳，且最终 DP-DDQN 的 WIS 分数达到最高。由此可以看出，DP-DDQN 模型比基线模型有更强的预测能力和数据学习能力，学到的最终策略要优于其他 3 个基线模型。对于 DP-DDQN 在 eICU 上的表现，虽然 WIS 分数没有在 MIMIC-IV 上表现那么好，但最终超过临床测试集，表明 DP-DDQN 模型具有较好的泛化能力，在新的数据集上的表现也能超过临床评分。

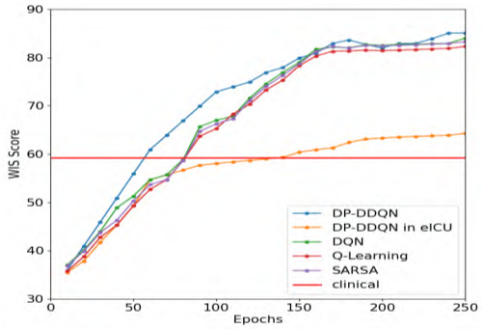

图 3　模型与临床评分之间的 WIS 对比

Fig. 3　Comparison of WIS between the models and clinical patients

3.2.2　用药时长

为了分析模型策略对于患者用药时长的建模情况，比较 DQN、QLearning、SARSA 以及在验证集 eICU 上的患者轨迹长度，结果如图 4 所示。

临床测试集平均患者用药时长为 6.92 天，DP-DDQN 模型最终平均用药时长为 4.45 天。与临床相比，DP-DDQN 模型在减少患者用药时长方面表现最佳，且其用药时长要比其他模型更合理，表明 DP-DDQN 模型得到的患者轨迹更安全，即能减少患者用药时长，同时也不会让患者用药时长过低。相比之下，其他模型虽然也有一定的优化效果，但不如 DP-DDQN 模型显著。

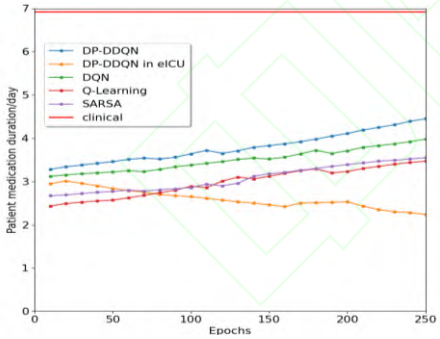

图 4　模型与临床患者之间的用药时长对比

Fig. 4　Comparison of medication duration between the models and clinical patients

3.2.3　模型 *Q* 值变化

在强化学习中，*Q* 值代表的是状态-动作对的期望回报，它帮助智能体评估在特定状态下采取特定动作的优劣。*Q* 值越高表明模型在特定状态采取特定动作时的期望回报越高，这意味着模型在该状态下的决策质量更好，能够获得更多的未来奖励，模型的学习效果越好。为了比较 4 个模型的学习效果，对训练过程中的 *Q* 值进行了可视化，结果如图 5 所示，其中只有 DP-DDQN 模型使用了全局决策偏好。从图 5 中可以看出，DP-DDQN 模型在整个训练过程中表现出最快的增长趋势，*Q* 值最高，为 7.1130。这表明 DP-DDQN 模型在全局决策偏好的指导下能够快速学习并找到高价值的动作，从而在训练过程中学习到更好的决策。DQN 模型次之，Q-learning 和 SARSA 算法的表现较差，可能是由于这两个算法在处理该类高维空间任务存在局限性。

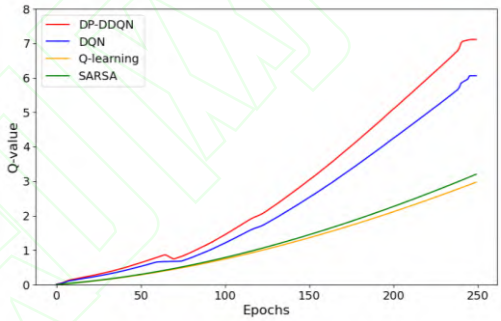

图 5　4 个模型 *Q* 值变化对比

Fig. 5　Comparison of Q-value between four models

3.2.4　患者最终状态以及预后情况

临床测试集与训练后模型生存率对比见表 3。实验结果显示，传统临床方法的存活率为 66.56%，而所有强化学习模型的表现均优于该传统方法，证明强化学习模型能有效提升脓毒症患者的存活率。在 4 个强化学习模型中，本文 DP-DDQN 模型表现最优，存活率达 84.01%，较临床提高了 17.45%，且显著高于其他模型；DQN 模型紧随其后，存活率为 82.65%；SARSA 和 Q 学习的存活率分别为 80.19%和 75.13%，虽然优于临床方法，但是因学习离散患者状态，精度低于采用连续状态空间的 DP-DDQN 和 DQN。这表明机器学习模型潜力突出，DP-DDQN 能更准确识别影响患者预后的关键因素，为临床决策提供有力支持，同时有助于优化医疗资源的分配与利用效率，能为临床医生提供更准确的患者预后信息。

表 3　临床测试集与训练后模型生存率对比

Tab. 3　Comparison of survival rates between clinical test set and models after training

模型	临床方法	Q 学习	SARSA	DQN	DP-DDQN
生存率/%	66.56	75.13	80.19	82.65	84.01

3.3 消融研究

为了比较隐式全局决策偏好以及不同显式的个性化诊疗偏好设置对模型结果的影响，本文对多决策奖励函数进行了消融实验。消融实验不同模型奖励函数设置见表 4。DDQN-1—DDNQ-4 为 3 个不同显式奖励函数之间的组合，DDQN-5—DDQN-7 为隐式全局决策偏好奖励 R_4 与不同显式奖励函数之间的组合，DP-DDQN 模型使用隐式全局决策偏好以及完整的显式奖励函数。

图 6 显示了 DP-DDQN 和 DDQN-1-7 的 WIS 评分和抗生素使用时长统计。消融实验结果表明，DDQN-1、DDQN-2 两个模型相对于其他模型用药时长较短，与其奖励函数的设置对应上，重点关注患者的预后结果，而忽略了患者中间状态变化，这与临床实际治疗过程不符合。DDQN-5 和 DDQN-6 模型在这两个模型基础上加入了隐式全局决策偏好奖励 R4 后，用药时长有所增加，但还是低于 DP-DDQN 模型。DDQN-3 和 DDQN-4、DDQN-7 均加入了用药时长奖励 R3，结果表明用药时长比前面 4 个模型均要高，更符合临床经验，但是 WIS 分数要低于 DP-DDQN 模型。DDQN-5 模型用药时长和 WIS 分数均要低于 DP-DDQN 模型，结合表 4，DP-DDQN 模型中加入了隐式全局决策偏好奖励，该决策奖励可以让模型学习到患者更准确的状态，从而做出更符合临床实践的决策。

表 4 消融实验不同模型奖励函数设置

Tab. 4 Setting of reward function corresponding to ablation experimental model

模型	奖励函数设置	模型	奖励函数设置
DDQN-1	R_1	DDQN-5	R_1、R_4
DDQN-2	R_1、R_2	DDQN-6	R_1、R_2、R_4
DDQN-3	R_1、R_3	DDQN-7	R_1、R_3、R_4
DDQN-4	R_1、R_2、R_3	DP-DDQN	R_1、R_2、R_3、R_4

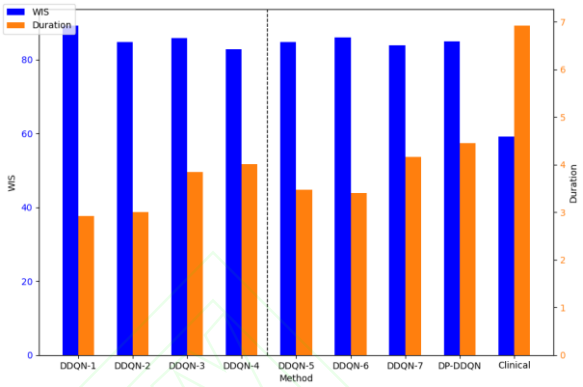

图 6 不同模型与临床患者 WIS 和用药时长对比的消融实验结果

Fig. 6 Results of ablation experiment of comparison of WIS and medication duration between different models and clinical patients

3.4 模型和临床药物相似度对比

使用皮尔逊系数方法计算任意两个患者特征之间的相似度，选择相似度大于 0.8 的两个患者认为是相似患者，并在此基础上计算药物序列相似度。在计算药物相似度中，先将药物的 SMILE 编码转化为摩根指纹[25]，再使用余弦相似度计算，为

$$\text{Similarity}\ (A,B)=\frac{\sum_{i=1}^{n}a_i b_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\sqrt{\sum_{i=1}^{n}b_i^2}} \tag{12}$$

其中：A、B 为剂量加权指纹向量；a_i,b_i>0 表示药物剂量在分子特征上的累积值；n 为指纹位数，n=1024[24,26]。

分别计算临床测试集、DP-DDQN 模型中相似患者药物序列相似度，还对比了临床和 DP-DDQN 模型中相似患者的药物序列相似度，并将药物相似度划分为 10 个区间，结果如图 7 所示。从图 7 中可以看出，临床相似患者的药物序列相似度主要集中在 60%~90%，其中 70%~80%区间的患者占比最高（24.06%）；DP-DDQN 模型预测的相似患者药物序列相似度集中在 80%~100%之间，其中 90%~100%区间的患者占比最高（24.73%），表明 DP-DDQN 模型推荐用药比临床更稳定。临床和 DP-DDQN 模型预测患者药物序列相似主要在 60%~90%，均在 14.5%附近，表明 DP-DDQN 模型可以从临床中学习

给药策略，并能探索出更符合患者状态的用药策略。

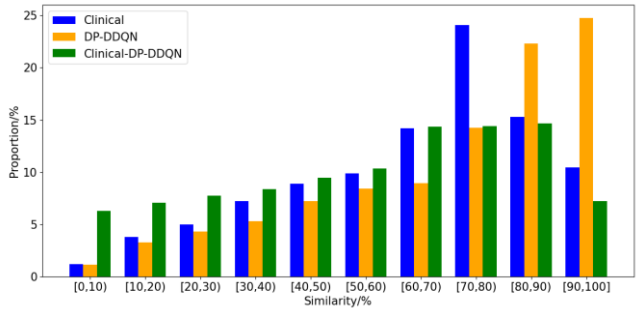

图 7　相似患者临床和模型预测药物相似对比分析

Fig. 7　Comparative analysis of drug similarity predicted by clinical methods and models for similar patients

3.5　临床和模型部分对比

分别计算临床测试集和 DP-DDQN 模型预测的生存或死亡的患者药物序列相似度，结果如图 8 所示。与临床存活的患者对比，DP-DDQN 模型预测患者药物序列相似度主要在 60%~90%区间，表明 DP-DDQN 模型推荐的药物与临床推荐药物具有一致性。

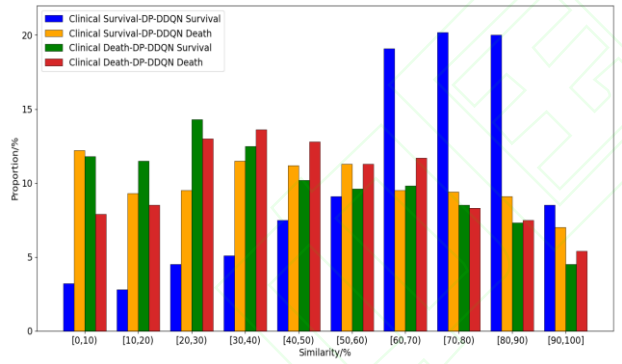

图 8　临床生存与死亡对应的模型用药序列相似度直方图

Fig. 8　Histogram of the similarity of the model medication sequence corresponding to clinical survival and death

对于临床预测将存活但模型预测会死亡的患者，其临床与模型预测之间的用药序列相似度大多落在 0%~60%，表明模型存在一定探索性。对于临床已死亡模型却预测了存活的患者，所推荐的抗生素组合与临床实际使用差异显著，模型推荐的用药序列与临床实际用药序列的相似度主要集中在 0%~70%。

3.6　药物剂量分析

由于本文模型在抗生素种类的基础上加入了剂量推荐，因此将临床测试集抗生素剂量与 DP-DDQN 模型推荐剂量进行对比分析，结果如图 9 所示。横坐标表示 31 种抗生素，左边纵坐标表示每种抗生素平均使用剂量，右边纵坐标（对应绿色折线）表示 DP-DDQN 模型推荐剂量比临床剂量减少的百分比，横坐标按照模型和临床剂量差递减排序。

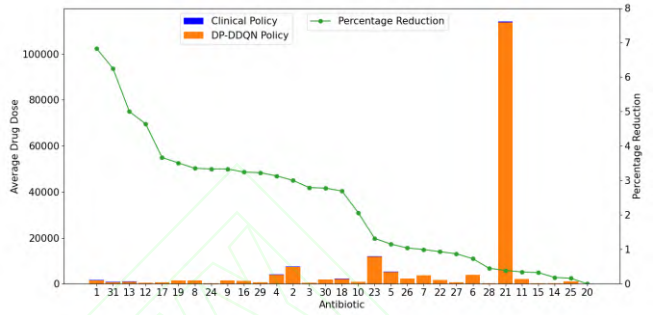

图 9　临床患者与模型预测剂量对比

Fig. 9　Comparison of clinical patient dosages and model-predicted dosages

从图 9 中可以看出，对于大多数抗生素，模型推荐的剂量要低于临床剂量，表明模型对于药物剂量的使用比临床严谨，综合考虑了患者的耐药性风险。DP-DDQN 模型推荐的剂量要比临床使用药物剂量略低。模型推荐剂量相较于临床剂量的减少百分比均未超过 7%（最高为 6.829%）。这种剂量的减少并非单纯的用药量降低，而是 DP-DDQN 模型在保障甚至提升患者预后效果的前提下，具备减少抗生素用药剂量的可能性。DP-DDQN 模型通过对大量患者相关数据的分析，能够识别出较低剂量的抗生素仍可有效维持或优化治疗结局，进而给出相对临床剂量更低的推荐。这一结果为临床治疗提供了一种潜在的精准调整方向：在不损害患者预后的基础上，减少不必要的药物使用，既有助于平衡耐药性风险，也为降低药物副作用、提升治疗安全性提供了可能，可作为临床医生制定治疗决策时的辅助参考。

4　结　论

本文针对目前的医疗决策强化学习架构只基于患者显式且独立的状态信号来进行策略学习问题，且缺少抗生素剂量推荐方面的研究，提出了一种隐式全局决策偏好方法并将其嵌入到强化学习中，为脓毒症患者生成抗生素种类和剂量策略推荐。该框架通过在强化学习中融入脓毒症全局决策偏好，指导模型生成患者的抗生素用药策略。模型与临床策

略对比结果表明，本文算法可以得到更优的脓毒症抗生素策略，将患者生存率从 66.56%提升到了84.01%。在预后效果比临床更好的情况下，降低了患者的用药时长，且模型推荐剂量比临床使用剂量减少 1%~7%，在保证患者预后好的前提下降低了患者的耐药性。未来将使用分层强化学习，并结合患者特征进行安全性约束，以得到更优更安全的抗生素策略推荐。

参考文献：

[1] Schuler M, Bölükbas S, Darwiche K, et al. Personalized treatment for patients with lung cancer[J]. Deutsches Ärzteblatt International, 2023, 120(17): 300.

[2] Artigas A, Carlet J, Ferrer R, et al. 25th International symposium on infections in the critically III patient[J]. Medical Sciences, 2020, 8(1): 13.

[3] Kolar M. Bacterial infections, antimicrobial resistance and antibiotic therapy[J]. Life Basel, 2022, 12(4): 468.

[4] Shakya A K, Pillai G, Chakrabarty S. Reinforcement learning algorithms: a brief survey[J]. Expert Systems with Applications, 2023, 231: 120495.

[5] Wang Yuan, Liu Anqi, Yang Jucheng, et al. Clinical knowledge-guided deep reinforcement learning for sepsis antibiotic dosing recommendations[J]. Artificial Intelligence in Medicine, 2024, 150: 102811.

[6] Lin Tianlai, Zhang Xinjue, Gong Jianbing, et al. A dosing strategy model of deep deterministic policy gradient algorithm for sepsis patients[J]. BMC Medical Informatics and Decision Making, 2023, 23(1): 81.

[7] Innocenti F, Palmieri V, Pini R. Fluids, vasopressors, and inotropes to restore heart-vessel coupling in sepsis: treatment options and perspectives[J]. Anesthesia Research, 2024, 1(2): 128-145.

[8] Zampieri F G, Bagshaw S M, Semler M W. Fluid therapy for critically ill adults with sepsis: a review[J]. JAMA, 2023, 329(22): 1967-1980.

[9] Barata C, Rotemberg V, Codella N C F, et al. A reinforcement learning model for AI-based decision support in skin cancer[J]. Nature Medicine, 2023, 29(8): 1941-1946.

[10] Fleischmann C, Scherag A, Adhikari N K J, et al. Assessment of global incidence and mortality of hospital-treated sepsis[J]. American Journal of Respiratory and Critical Care Medicine, 2016, 193(3): 259-272.

[11] Evans L, Rhodes A, Alhazzani W, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021[J]. Critical Care Medicine, 2021, 49(11): 1063-1143.

[12] Wu Xiaodan, Li Ruichang, He Zhen, et al. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis[J]. NPJ Digital Medicine, 2023, 6(1): 15.

[13] Choi Y, Oh S, Huh J W, et al. Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records[J]. Communications Medicine, 2024, 4(1): 245.

[14] Adams S, Cody T, Beling P A. A survey of inverse reinforcement learning[J]. Artificial Intelligence Review, 2022, 55(6): 4307-4346.

[15] Wulfmeier M, Ondruska P, Posner I. Maximum entropy deep inverse reinforcement learning[EB/OL]. [2025-10-19]. https://doi.org/10.48550/arXiv.1507.04888.

[16] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[17] Anschel O, Baram N, Shimkin N. Averaged-DQN: variance reduction and stabilization for deep reinforcement learning[C]//PMLR. Proceedings of the 34th International Conference on Machine Learning. New York：PMLR, 2017: 176-185.

[18] van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning[J]. Proceedings of the AAAI conference on artificial intelligence, 2016, 30(1): 2094-2100.

[19] Ziebart B D, Maas A L, Bagnell J A, et al. Maximum entropy inverse reinforcement learning[J]. Proceedings of the AAAI conference on artificial intelligence, 2008, 8: 1433-1438.

[20] Liang Dayang, Deng Huiyi, Liu Yunlong. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach[J]. Applied Intelligence, 2023, 53(9): 11034-11044.

[21] Boto-García D. Good results come to those who weight: on the importance of sampling weights in empirical research

using survey data[J]. Current Issues in Tourism, 2024, 27(2): 268-287.

[22] Johnson A E W, Bulgarelli L, Shen Lu, et al. MIMIC-IV, a freely accessible electronic health record dataset[J]. Scientific data, 2023, 10(1): 1.

[23] Pollard T J, Johnson A E W, Raffa J D, et al. The eICU collaborative research database, a freely available multi-center database for critical care research[J]. Scientific Data, 2018, 5(1): 1-13.

[24] Wu Siwei, Pan Zhenxin, Li Xiaojing, et al. Machine learning assisted photothermal conversion efficiency prediction of anticancer photothermal agents[J]. Chemical Engineering Science, 2023, 273: 118619.

[25] Ehiro T. Descriptor generation from Morgan fingerprint using persistent homology[J]. SAR and QSAR in Environmental Research, 2024, 35(1): 31-51.

[26] Pham T, Ghafoor M, Granana-Castillo S, et al. DeepARV: ensemble deep learning to predict drug-drug interaction of clinical relevance with antiretroviral therapy[J]. NPJ Systems Biology and Applications, 2024, 10(1): 48.