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：基于强化学习框架的脓毒症抗生素多策略推荐模型
作者：王嫄，刘安岐，盛梦茹，侯佳佳，赵婷婷，于琦
DOI：10.13364/j.issn.1672-6510.20230203
收稿日期：2023-10-26
网络首发日期：2024-09-30
引用格式：王嫄，刘安岐，盛梦茹，侯佳佳，赵婷婷，于琦．基于强化学习框架的脓毒症抗生素多策略推荐模型[J/OL]．天津科技大学学报.
https://doi.org/10.13364/j.issn.1672-6510.20230203

DOI：10.13364/j.issn.1672-6510.20230203

基于强化学习框架的脓毒症抗生素多策略推荐模型

王 嫄[1]，刘安岐[1]，盛梦茹[1]，侯佳佳[1]，赵婷婷[1]，于 琦[2]
(1.天津科技大学人工智能学院，天津 300457；2.山西医科大学管理学院，太原 030001)

摘 要：脓毒症是全球几大死亡原因之一，而抗生素是脓毒症治疗的重要一环。近年来，研究人员认为医疗决策问题可以映射为马尔科夫决策过程，并使用强化学习方法进行治疗策略推荐。结合基于值函数和基于策略的强化学习方法构建多策略推荐的模型框架，对脓毒症治疗过程中抗生素的使用进行策略推荐。针对脓毒症患者特征信息划分不同的决策区域，多策略模型进行个性化治疗建议。结果表明：多策略选择模型能够使患者预后良好的情况达到80.32%。通过统计分析决策轨迹和药物作用选择，模型能够提供符合临床实践的合理药物建议，推荐合适的抗生素组合改善患者的预后效果。

关键词：强化学习；医疗决策；脓毒症
中图分类号：TP391 **文献标志码：**A **文章编号：**1672-6510（0000）00-0000-00

A Multi-Policy Recommendation Model for Antibiotic Use in Sepsis Based on Reinforcement Learning Framework

WANG Yuan[1]，LIU Anqi[1]，SHENG Mengru[1]，HOU Jiajia[1]，ZHAO Tingting[1]，YU Qi[2]
(1. College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457, China；
2. School of Management Shanxi Medical University, Taiyuan 030001, China)

Abstract: Sepsis is one of the leading causes of death worldwide, and antibiotics are an important part of sepsis treatment. In recent years, researchers have considered medical decision-making problems as Markov decision processes and used reinforcement learning methods for treatment strategy recommendations. We propose a multi-policy recommendation framework combining value-based and policy-based reinforcement learning methods for antibiotic use in sepsis treatment. Different decision regions are defined based on patient characteristic information, and the multi-policy model provides personalized treatment recommendations. The results show that our multi-policy selection model can achieve a good prognosis for patients in 80.32% of cases. Through statistical analysis of decision trajectories and drug action selection, our model can provide reasonable drug recommendations in accordance with clinical practice, and recommend appropriate antibiotic combinations to improve patient prognosis.

Key words: reinforcement learning; medical decision-making; sepsis

脓毒症是全球几大死亡原因之一，脓毒症的成功治疗通常涉及多种因素的融合[1-3]，包括患者状况和病史，以及抗生素治疗、液体复苏、血管加压药物和机械通气等治疗干预。由于脓毒症主要由细菌或真菌感染引起，选择最合适的抗生素组合和持续时间对于显著改善患者预后至关重要。近年来，研究人员认为医疗决策问题[4-9]可以映射为马尔科夫决策过程，使用强化学习方法进行治疗策略推荐。

目前，抗生素在脓毒症治疗中的使用和持续时间在很大程度上依赖于临床医生的经验和判断[1]，这可能导致治疗不足或过度，最终导致治疗失败或抗生素滥用。因此，需要人工智能指导方法推荐最佳的抗生素治疗方案，提高治疗成功率，并减少医疗资源浪费。近年来，基于强化学习的脓毒症治疗

收稿日期：2023-10-26；修回日期：2024-05-06
基金项目：国家自然科学基金资助(61976156)，天津市科技特派员项目资助(20YDTPJC00560)
作者简介：王 嫄(1989—)，女，山西万荣人，副教授，博士，wangyuan23@tust.edu.cn

建议的相关文章[10-15]主要集中在机械通气、静脉输液和血管升压药方面。虽然指南[1]建议感染或疑似感染的患者及时使用抗生素[16]，但目前还没有准确的工具指导何时停止使用抗生素。本文针对抗生素使用组合和停止时间进行建议，通过使用强化学习和整合医疗知识为临床医生提供脓毒症治疗中的抗生素提供个性化的治疗建议。

强化学习（reinforcement learning，RL）是一种用于学习、预测和决策的方法框架。由于 RL 解决了连续决策问题，考虑了长期奖励问题并优化了策略，因此强化学习可以优化所提出的治疗方案模型，它由智能体与环境之间的互动以及奖励的构成。RL 通过让智力和环境相互作用积累奖励，从而获得最佳策略。

近年来，基于强化学习的医疗策略推荐主要研究方向为非深度强化学习和深度强化学习。非深度强化学习方法在医疗决策领域中被广泛应用。它们通常依赖于手动设计的特征提取器从医学数据中提取关键信息，并基于此进行策略学习和优化[7]。Raheb 等[17]对单一药物采用了传统的强化学习方法用于 2 型糖尿病患者的皮下注射葡萄糖调节，除了考虑血糖水平，还考虑了胰岛素在皮下注射中的延迟作用，并为 24 h 血糖变化设计了奖励函数。Schamberg 等[18]使用基于强化学习的 actor-critic 模型开发了一种全自动麻醉药物给药系统，可以在手术期间自动给药。该研究使用了 3 种不同的奖励函数，并结合了药代动力学/药效学模型训练药物，建立了有针对性的奖励函数，以优化系统性能。Zhang 等[19]设计了一种通用计算框架，为相似临床背景的患者划分决策区域，并使用 Q-Learning 方法推荐每个决策区域的最佳策略，为重症监护室的低血压患者提供治疗建议。与单纯地针对患者信息进行聚类不同，Zhang 等[19]划分的决策区域信息更加透明，这能帮助临床医生清晰地判断某决策区域的患者详细状态，从而判断模型推荐的用药策略是否合理、是否具有借鉴意义。

深度强化学习结合了深度学习的特征提取能力和强化学习的决策能力，进一步提升了诊疗方案推荐的准确性[7]。Zadeh 等[20]提出了一种基于 DQN 的华法林给药模型，该模型使用药物的药代动力学/药效学（PK/PD）模型模拟虚拟患者的剂量反应。上述研究都结合了药理学知识，并建立了相关的奖励功能，但他们的重点仅限于单一药物的决策建议。Peine 等[21]开发和评估了强化学习算法 Vent AI，该算法能够为重症患者提出动态优化的机械通气方案，使用 DQN 算法从次优的 ICU 历史数据中确定给定患者状态下的最佳行动。

强化学习也应用于脓毒症相关的医疗决策问题。脓毒症治疗涉及多种治疗手段，之前的研究通常关注患者的静脉输入和血管加压药的剂量决策[22]。最经典的是 Komorowski 等[23]开发的强化学习 AI 临床医生，该模型利用大量患者数据提取隐含知识，并为脓毒症治疗推荐最佳剂量的静脉输入和血管升压药。该模型为败血症提供了个性化和临床可解释的治疗决策。这项工作强调了强化学习在推荐败血症治疗方案中的潜力，但它利用了最基本的强化学习方法和单一的奖励函数设置，只关注患者的最终结果。Lin 等[24]使用深度确定性策略梯度（DDPG）算法，针对脓毒症患者的静脉输入和血管加压药进行决策推荐，基于 DDPG 算法的医疗决策系统生成的治疗方案更接近于住院脓毒症患者死亡率较低的专业临床医生的治疗方案。

尽管强化学习在脓毒症治疗决策中有许多应用，但目前还没有使用强化学习确定最佳抗生素治疗方案的研究。本研究根据人口统计学特征、基本生命体征、微生物培养结果和实验室结果等划分决策区域，为临床医生提供可视化的患者状态，同时结合基于值函数的 Q 学习（Q-learning）[25-26]和策略梯度（policy gradient，PG）[27-28]方法的各自优势进行推荐，生成的策略再通过策略选择得到各决策区域的最优策略。最优策略选择函数整合了临床知识和脓毒症治疗指南，以确保脓毒症治疗中抗生素组合使用的医学可解释性和临床一致性。

1 强化学习相关原理

预测和优化治疗计划是医学人工智能研究的焦点[29-30]。强化学习是一种机器学习，在决策控制应用中越来越受欢迎，强化学习的优势正在扩展到各个领域，其中包括医疗领域[7-8]。强化学习在医学决策中的应用已经产生了有希望的结果，医疗决策过程可以建模为马尔可夫决策过程[7]（Markov decision process，MDP）。MDP 对患者的环境和轨迹进行建模，近似患者的状态特征和日常用药，以模拟决策过程。MDP 由元组$\{S,A,T,R,\gamma\}$组成，其中：S为一组有限的状态；A为对于给定状态可以采取的一组有限的操作；T为转换矩阵，它包含在状态 s 中采取动作a后，在时间$T+1$转换到状态s'的概率；R为奖励函数，根据每个状态-动作

对，给最终结果的贡献分配一个值；γ 为折现系数，决定了未来奖励相对于当前奖励的重要性。

1.1 Q 学习

Q 学习[25-26]是一种基于值迭代的强化学习算法，旨在学习最优的策略，使智能体从当前状态出发，得到最大的期望总奖赏。通过学习一个价值函数 Q，表示在状态 s 下采取行动 a 所获得的长期奖励的期望，然后使用这个价值函数作为策略评估函数，根据最大 Q 值选择行动 a，进而实现 Q 学习算法。

Q-learning 算法的学习过程如下：首先随机初始化一个 Q 值表，对于每一个状态 s 和可选动作 a，将 $Q(s,a)$ 设为 0。随着智能体与环境的互动，逐渐学习并更新 Q 值表。在每个时间步 t 中，根据智能体的行动和环境的反馈奖励，更新当前状态的 Q 值，即

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\alpha\left(\gamma_{t+1}+\max_a \gamma Q(s_{t+1},a)-Q(s_t,a_t)\right) \tag{1}$$

其中：α 为学习率，γ 为折扣因子，用于平衡即时奖赏和未来奖赏的价值。根据 $Q(s,a)$ 的值可以推断出更好的策略并采取行动，更新状态并对其进行采样，直到学习到一个最优的策略。Q-learning 算法的局限性：一是需要一个足够大的状态空间，以覆盖所有可能的状态；二是 Q_{table} 大小会随着状态空间和动作空间的增大而膨胀，导致计算和存储需求的增加。

1.2 策略梯度

策略梯度[27-28]算法是一种基于策略优化的强化学习算法，它的基本思想是优化策略参数，以最大化期望累积奖励。与 Q-learning 算法的区别在于，PG 算法直接搜索最优策略参数，而不是通过价值函数计算 Q 值，然后再得到最优策略。

PG 算法的基本流程如下：给定策略 $\pi_\theta(a|s)$，其表示在状态 s 下选择动作 a 的概率。首先，使用策略参数 θ 生成若干条轨迹，即执行若干次采样，获得相应的累积奖励。然后，计算损失函数，并将其应用于策略梯度更新，公式为

$$\nabla_\theta J(\theta)=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\, r(\tau)\right] \tag{2}$$

其中：$\tau\sim p_{\{\theta\}}(\tau)$ 表示从策略 π_θ 中得到轨迹 τ，$r(\tau)$ 表示轨迹 τ 的（累积）奖励。

通过这个损失函数，PG 算法可以学习到最优的策略参数 θ，从而获得最优策略。相比于值函数方法，PG 算法更适用于连续空间的问题，因为它可以根据策略直接选择连续动作，而值函数方法需要进行动作离散化处理。智能体从当前状态（决策区域）出发，得到最大的期望总奖赏。通过学习一个价值函数 Q，使用这个价值函数作为策略评估函数，根据最大 Q 值选择行动 a，进而实现 Q 学习算法。

2 多策略推荐模型

2.1 模型框架

为了得到更加透明的患者状态信息和符合医学知识的患者个性化抗生素使用推荐，设计了多策略推荐强化学习模型框架针对脓毒症患者的抗生素使用进行推荐，模型框架整体流程图如图 1 所示，模型框架主要包括三大部分：决策区域划分、强化学习方法池以及最优策略选择。

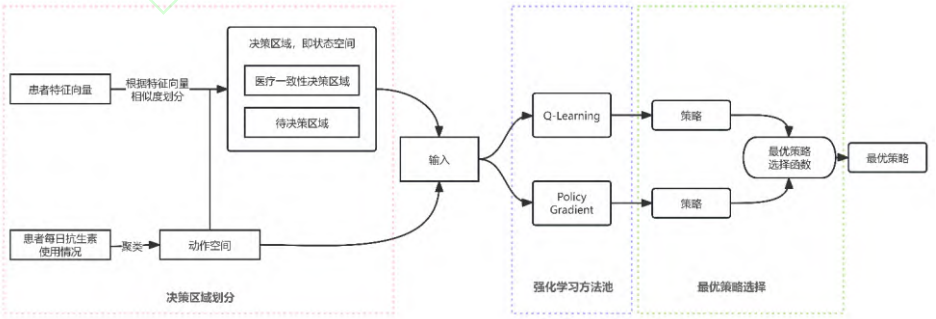

图 1　多策略推荐模型整体流程图

Fig. 1　Overall flowchart of the Multi Policy recommendation model

在决策区域划分的部分，将收集的患者每日特征 $f_{i,j}$ 的相似度进行划分，使用皮尔逊相似系数计算患者相似度，将类内患者间相似度记为 ICS_i，当 $ICS_i >= sim_s$ 时可将当前类作为一个决策区域。根

据相似度划分得到的决策区域集合（DR），对 DR 内的患者对应每日抗生素使用相似情况 SC 进行对比，将用药相似使用抗生素组合在同一聚类类别内的占比大于等于 sim_a，即 $\max(SC_i) >= sim_a$ 的决策区域成为医疗一致性决策区域（MCDR），其余为待决策区域（WDR）。得到决策区域集合后，将决策区域作为状态空间 S 输入到强化学习池中。

强化学习池包括 Q-Learning 和策略梯度算法。经过训练得到两个模型，使用模型生成了抗生素用药策略 π_Q 和 π_{PG}。在 最优策略选择模块，设置了三个参数针对策略 π_Q 和 π_{PG} 进行选择。

2.2 决策区域划分

为了使临床医生能够观察到患者详细的特征信息变化，使用决策区域划分透明化患者信息。每个决策区域内部的患者状态都是相似的，能够将模型预测的患者状态轨迹变化（治疗过程）详细地呈现给临床医生，由临床医生判断策略的可行性。进行决策区域划分，能够更有效地提取和分析患者的特征信息，不仅使策略可见性提高，还使得患者状态的变化更加透明化。这种透明化有助于医生更好地理解患者的病情发展趋势，并为制定更为精确和个性化的治疗方案提供有力支持，进一步增强模型的临床可解释性。图 2 为多策略模型决策区域划分方法流程图。

决策区域划分方法：将筛选得到的训练集的患者每日特征向量作为输入，构建一个递归函数 PolicyDivision。在函数 PolicyDivision 内对输入的状态信息 $f_{i,j}$（特征向量）聚类，数目 k_s。分别计算所得两个类的类内患者状态相似度 ISC_a 和 ISC_b。如果类内相似度小于 sim_s，则将这个类作为输入，循环 b、c 操作，如果类内相似度大于 sim_s，则完成一次划分，所得类为一个未分类的决策区域 DR_i。完成所有特征向量划分，根据各个未分类的决策区域 DR_i 对应的患者当日用抗生素组合类别占比 SC，当某类抗生素组合在这个决策区域内占比超过 sim_a 时 $\max(SC_i) >= sim_a$，将这个决策区域称为医疗一致性决策区域 MCDR，反之，占比没有超过 sim_a 的决策区域称为待决策区域 WDR，再对这些决策区域进行分类。

使用皮尔逊相似度计算 ISC。对于患者特征向量 $f_{i,j}$ 和每日抗生素使用序列 $anti_{i,j}$，使用了 k-means 方法进行聚类。经过划分和分类得到 MCDR 和 WDR 后，再加入两个决策区域，分别对应患者入住 ICU 后的 90 d 内生存情况。

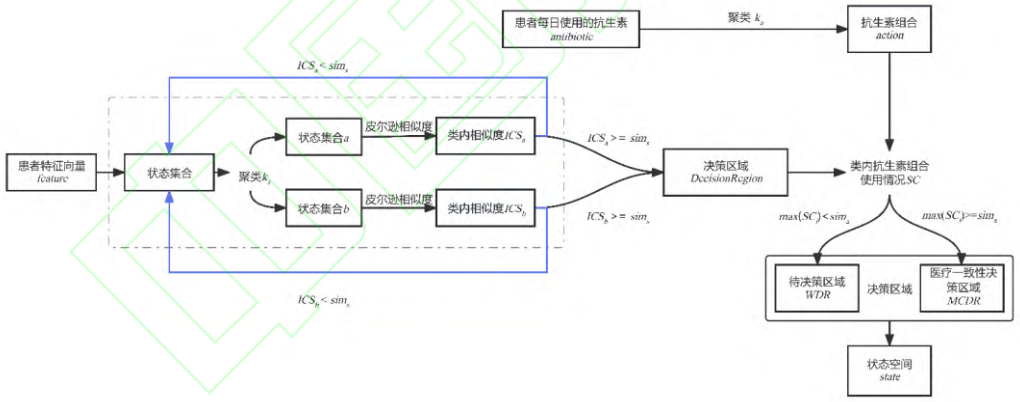

图 2　多策略模型-决策区域划分流程图

Fig. 2　Multi Policy model - decision region division flowchart

2.3 强化学习池

选择 Q-learning[25]和策略梯度[27]算法作为强化学习池。状态空间 S 是划分得到的决策区域集合 DR。动作空间 A 为患者每日抗生素使用序列聚类得到的 k_a 类。对于医疗一致性决策区域 MCDR，认为这类决策区域中超过 sim_a 的医生对其中状态相似的患者使用抗生素的策略相似。将 k_{MCDR} 个医疗一致性决策区域使用的抗生素组合（动作空间）统一为各个医疗一致性决策区域内患者使用最多的抗生素组合。

将训练集中得到的医疗一致性决策区域 MCDR 和待决策区域 WDR 集合 DR 作为状态空间 S、患者每日抗生素使用的聚类结果 AD 作为动作空间输入到 Q-Learning 和策略梯度两种算法模型中，经过训练得到两个策略推荐模型。在两个方法中，根据最终状态（存活和死亡）设置了奖励函数，达到存活对应的决策区域奖励 +100，达到死亡对应的决策区域惩罚 –100。

在 Q-Learning 模型[25]中，首先随机初始化一个 Q 值表 Q_{table}，对于每一个状态 s 和可选动作 a，将 $Q(s,a)$ 设为 0，其中 $s \in DR, a \in AD$。随着智能体 $Agent$ 与环境的互动，逐渐学习并更新 Q_{table}。对于每个时间步 t，根据智能体的行动和环境的反馈奖励，更新当前状态 Q_{value}，即

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\alpha\left(\gamma_{t+1}+\max_a \gamma Q(s_{t+1},a)-Q(s_t,a_t)\right) \quad s \in DR, a \in AD \tag{3}$$

其中：α 为学习率，γ 为折扣因子，用于平衡即时奖赏和未来奖赏的价值。根据 $Q(s,a)$ 的值可以推断出更好的策略并采取行动，更新状态并对其进行采样，直到学习到一个最优的策略 $\pi_{Q-Learning}$。

在 Policy Gradient 方法[27]中，给定策略 $\pi_\theta(a|s)$，其表示在状态 s 下选择动作 a 的概率，其中 $s \in DR, a \in AD$。使用策略参数 θ 生成若干条轨迹，即执行若干次采样，获得相应的累积奖励。计算损失函数并应用于策略梯度更新，公式为

$$\nabla_\theta J(\theta)=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t) r(\tau)\right] \quad s \in DR, a \in AD \tag{4}$$

其中：$\tau \sim p_\theta(\tau)$ 表示从策略 π_θ 中得到的轨迹 τ，$r(\tau)$ 表示轨迹 τ 的（累积）奖励。通过损失函数，策略梯度算法可以通过最小化损失函数学习到最优的策略参数 θ，从而获得最优策略 $\pi_{PolicyGradient}$。

2.4 最优策略选择

使用最优策略选择函数对 Q-Learning 和策略梯度方法推荐的策略进行选择。将每个决策区域统计的中心点作为此决策区域的特征信息。在最优选择函数中，设置了 3 个参数作为选择参考，分别是患者状态平均得分、状态价值得分和轨迹平均价值得分。

表 1 为最优策略选择的参数内容和设置目的。为了使模型关注到患者真实的预后情况，加入了患者最后用药状态（最终状态前一状态）的 SOFA 评分（D_{SOFA}）。在临床中，SOFA 评分能根据患者特征信息评估患者的病情和疾病进展，评分范围从 0 到 24，评分越高表示患者的病情越差。为了直接判断患者的状态值，计算状态价值得分，即完整模拟的轨迹中患者状态 s 对应的 Q 值平均值 $\bar{Q}_{s*}$。状态价值得分与患者整个轨迹状态变化有关，更关注的是轨迹中患者的抗生素组合使用情况，也就是患者的状态-动作对 (s,a) 的 Q 值得分 $\bar{Q}_{(s,a)}$。最优策略选择的评分公式为

$$Score=-W_1 D_{SOFA}+W_2 Q_{s*}+W_3 \bar{Q}_{(s,a)} \tag{5}$$

表 1 最优策略选择的参数设置

Tab. 1 Parameter settings for optimal strategy selection

参数类别	内容	目的	权重
患者真实状态得分	轨迹结束（最后一次用药），患者 SOFA 得分	使最优策略选择更关注患者真实的预后情况	W_1
状态价值得分	完整轨迹中，患者中间状态 $s*$ 的 Q 值平均值	最优策略选择对患者状态价值的直接判断	W_2
轨迹平均价值	完整轨迹中，患者状态-动作对 (s,a) 的 Q 值平均值	最优策略选择对患者用药价值的直接判断	W_3

3　实验结果及分析

3.1 数据集

使用的数据集是重症监护医疗数据集（MIMIC-IV）。该数据集包含波士顿一家主要学术医疗中心 ICU 患者的全面临床信息，包括生命体征、实验室值、药物信息和临床笔记。这些数据已获得 MIMIC-IV 机构审查委员会的许可。从 MIMIC-IV 数据库中提取了 9982 名脓毒症患者的 ICU 数据。

从 MIMIC-IV 数据集中选择了符合脓毒症-3 标准的确诊脓毒症患者，其中排除了新生儿脓毒症患者或在治疗期间转院的患者。从中挑选了 9982 名符合条件的患者。利用了患者的人口统计数据、生命体征和实验室结果的时间序列数据、微生物培养记录、药物和程序以及诊断编码。排除了人口统计学特征、实验室测试中缺失值大于 60%的特征，保留了患者的性别、年龄、体重等口统计学特征，同时对 24 h 内的重复数据进行平均化。根据临床知识和缺失率选择了 37 个实验室项目和 15 个基本生命体征。然后针对这 37 个实验室项目，计算了 24 h 内同一时间间隔测试获取的多次测量数据的平均值。

微生物培养结果指示患者的感染菌株和对不同常用抗生素的易感性，有助于选择最有效的药物。虽然在数据筛选的过程中，微生物培养结果缺失率超过了 80%，但仍保留了培养结果。微生物培养结果包括不同菌株和相应抗生素的敏感性信息，共

2047 种组合。敏感的结果用值1表示，其他结果由值0表示，缺少的数据用–1代替。对于人口统计学特征、实验室数据和基本生命体征，使用正向填充和 k 近邻方法填充缺失值。如果患者的完整特征缺失，使用 k 近邻方法填充。

筛选得到了 30 种脓毒症患者在 ICU 期间使用的抗生素。关注抗生素使用种类和使用时长，因此患者每日用药序列中使用的抗生素标为 1，未使用的抗生素标为 0。由于患者每日使用的抗生素可能有 1 到多种，使用 K-means 聚类方法将患者每日用聚为 30 类。

3.2 评价指标

模型评估的目标是使用由临床数据生成的患者轨迹评估 SAI-DQN 的策略价值。使用了离线策略评估（OPE）中的加权重要性采样 WIS 方法估计策略值在测试集中的真实分布。WIS 是有偏一致性决策估计。强化学习研究经常使用 WIS 方法对决策价值进行估计。将 π_0 定义为行为临床策略，从实际的患者数据中生成，得到策略价值。π_1 定义为 SAI-DQN 的策略。在离线策略估计中，重要性采样是一种纠正策略差异的简单方法，加权估计可以减少其方差。估计目标是 π_1，时间步为 t。$\pi_1(a_t, s_t)$ 为 t 时刻状态 s_t 采取动作 a_t 的策略价值，对应模型训练得到的 $Q_t(s,a)$。

t 时刻，策略 π_0 和 π_1 的重要性比为 $\eta_t = \pi_1(a_t, s_t) / \pi_0(a_t, s_t)$，从开始到时间 t 范围内的累积重要性比率为 $\eta_{1:t} = \prod_{t'=1}^{t} \eta_{t'}$。在测试数据集 D 中，从开始到时间 t 的平均累积重要性比为

$$W_t = \frac{\sum_{i=1}^{|D|} \eta_{1:t}^{(i)}}{|D|} \tag{6}$$

其中是 $|D|$ 轨迹数量，即测试集中患者数量。

因此，每个轨迹 track（患者）的加权重要性采样估计量为

$$T_{wis} = \frac{\eta_{1:E}}{W_E} \sum_{E}^{t=1} \gamma^{t-1} r_t \tag{7}$$

其中：E 是结束轨迹样本（即模拟的患者治疗从开始用药到结束用药的全过程），γ 为折扣因子，r_t 为时刻及时奖励。

测试集 D 中所有轨迹的总估计量为

$$\theta_{wis} = \frac{\sum_{\{k=1\}}^{|D|} T_{wis}^{(k)}}{|D|} \tag{8}$$

其中：k 为测试数据集 D 中的患者，θ_{wis} 就是对测试数据集中的所有患者轨迹的平均估计价值。

3.3 对比方法及实验设置

实验环境为 Python3.9，处理器为 Intel i7-11700k，操作系统为 Linux。

3.3.1 决策区域参数设置

在多策略选择模型的决策区域划分部分，选择 0.6 到 0.8 的患者相似度进行实验，表 2 为不同患者相似度决策区域划分数目。

表 2 不同患者相似度决策区域划分数目

Tab. 2 Number of decision regions for different patient similarities

相似度	决策区域/个
0.6	11
0.65	25
0.7	78
0.75	210
0.8	437

为了直观反映抗生素组合使用相似情况，图 3 为患者相似度在 0.8 时，各决策区域的抗生素组合占比的核密度统计图，其中横坐标为相似动作在同个决策区域内总动作数目占比，纵坐标为概率密度。从图 3 能看出相同抗生素组合在各决策区域内占比大多在 0 到 0.2 之间，只有很少部分处于 0.5 以上。

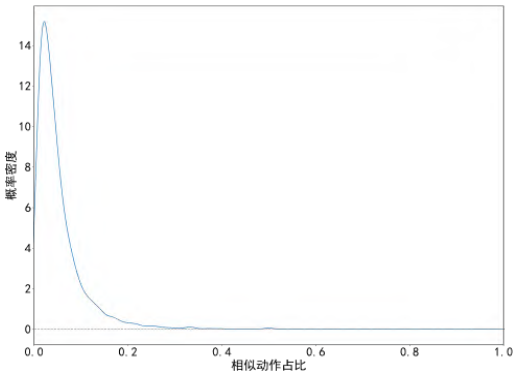

图 3 患者相似度 0.8 时，各决策区域的抗生素组合占比核密度统计图

Fig. 3 Statistical chart of the proportion of antibiotic combinations in each decision region when the patient similarity is 0.8

为了详细分析患者相似度在 0.8 时的统计数据，图 4 为 437 个决策区域内患者数目分布，其中横坐标为患者数目区间，纵坐标为决策区域数量。从图 4 可以看出，只有较少数量的决策区域的患者数量在 500 以上，更多的决策区域患者数目在 0~500。这个结果符合脓毒症患者状态个性化的特点，也表现出划分决策区域能够为脓毒症患者提供个性化治疗方案，也使得模型遵循临床指南。

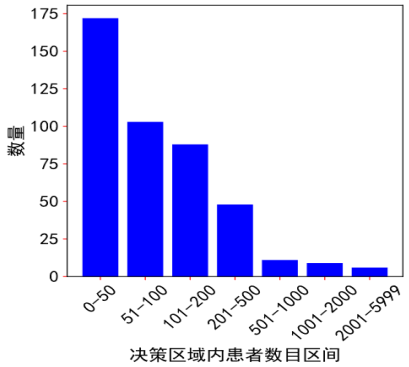

图 4 患者相似度 0.8 时，437 个决策区域内患者数目分布

Fig. 4 Proportional distribution of the number of patients in 437 decision regions when the patient similarity is 0.8

当患者相似度选择为 0.6、0.65、0.7 时，决策区域个数不到 100。由于脓毒症患者状态差别较大、决策区域较小会导致患者状态不相似但处于同一决策区域，这会使模型无法更好的学习脓毒症患者的特征信息，不利于对脓毒症患者的个性化治疗推荐。而当相似度在 0.75 的情况下，虽然决策区域个数达到了 210 个，但是其中每个决策区域经统计只有 3 个区域是超过 50%患者每日使用抗生素的组合相似的情况。统计患者每日用药相似度的目的是选择医疗一致性决策区域，让模型能够学习到临床医生意见较为相同的决策，这些决策是安全可靠的临床经验知识。结合表 2 和图 3，我们可以认为当患者特征相似度为 0.8 时，可以保证模型能够学习到医生诊断脓毒症患者抗生素使用的个性化治疗，并保证每个决策区域有足够的训练数据，所以我们选择 0.8 作为划分决策区域的患者相似度。

3.3.2 对比模型

设置两个对比模型为 Q-Learning 和策略梯度。在对比实验中，两个模型与强化学习池的模型参数设置相同。奖励函数设置为轨迹的最终状态，存活状态奖励+100，死亡状态惩罚-100。

另有对比模型“Clinic”为临床数据的测试集数据。

3.4 结果分析

3.4.1 Q-Learning、Policy Gradient 和多策略模型结果的加权重要性采样对比

为了对比多策略模型与两基础模型在模型性能上的差异，对推荐策略进行了加权重要性采样结果统计分析，其中多策略模型的最优策略选择的参数设置为$W_1:W_2:W_3=4:1:4$。图 5 为 Q-Learning、Policy Gradient 和多策略模型结果的加权重要性采样对比图，横坐标为训练次数，纵坐标为加权重要性采样结果。表 3 为训练次数 50 万次时患者预后情况统计表，其中 Clinic 为测试集的临床数据统计得到的患者结局。

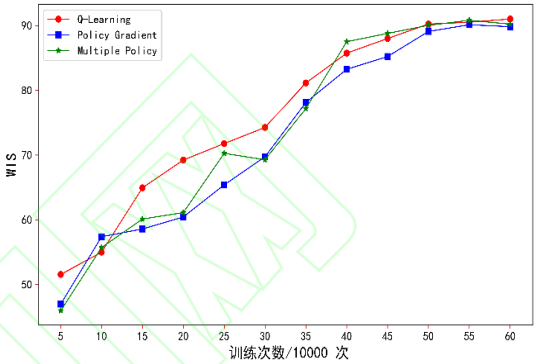

图 5 Q-Learning、Policy Gradient 和多策略模型结果的加权重要性采样对比

Fig. 5 Comparison of Importance Weighted Sampling results for the weighted importance of Q-Learning, Policy Gradient, and Multi-Policy models.

表 3 训练次数 50 万次时，患者预后情况统计表

Tab. 3 Patient prognosis statistical table after 500,000 training iterations

模型	患者结局（/%）	
	预后良好	预后较差
Clinic	60.42	39.58
Q-Learning	79.16	20.84
Policy Gradient	73.16	26.84
Multiple Policy	80.32	19.68

从图 5 中可以看出，多策略模型能够在模型价值方面与两个基础方法差别不大，最后结果都在 90 分左右。但是结合患者预后情况统计表（表 3），多策略模型能够得到比两个基础法更高的预后良好的结果。这符合设置最优选择函数的目的：使模型推荐的结果更偏向于患者中间状态和预后情况更好的结果。

3.4.2 消融实验

为了得到患者最终状态更好的策略，针对最优选择模块的参数设置的消融实验。表 4 为消融实验中的最优策略选择的参数权重设置，参数设置为 的

比例。图 6 为不同最优策略选择参数权重设置下患者预后情况良好的概率，横坐标为训练次数，纵坐标为患者预后结果良好占比。

表 4 消融实验的最优策略选择参数设置

Tab. 4 Optimal policy selection function parameter settings for ablation experiments.

患者状态平均得分	状态价值得分	轨迹平均价值
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	1	1
2	1	2
4	1	4

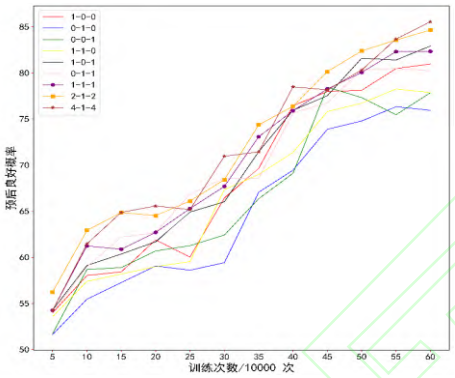

图 6 不同最优策略选择函数参数设置下患者预后结果良好概率

Fig. 6 Probability of favorable patient prognosis depends on the parameter settings of different optimal policy selection functions

从图 6 中可以看出，当训练次数达到 50 万~60 万次时，3 个价值得分权重为2:1:2、4:1:4的情况下，患者预后良好占比能达到 85%左右。同时，在去掉状态价值得分的1:0:1中，结果并未优于2:1:2和4:1:4。设置状态价值得分是为了让模型关注患者的状态变化情况，这对最优策略选择关注患者状态良好与否是有重要意义的，因此不能去掉。相同训练次数下，2:1:2、4:1:4结果更好，这表明设置的针对患者中间状态 SOFA 得分和轨迹（治疗全过程）价值能帮助模型选择对患者预后更好的策略。

最优策略选择设置的 3 个参数充分考虑了临床可解释性和医学可解释性。首先，患者真实状态得分为模型预测患者最后一次用药时患者 SOFA 得分，这结合了临床医学知识对患者状态进行评估。参数患者真实状态得分明确反映患者预后情况，属于医生对患者病情的评估方法。其次，状态价值得分能够反映在模型预测过程中患者状态得分的变化趋势，有助于模型学习到患者状态变化。最后，轨迹平均价值从数据出发直接对模型推荐的药物组合的价值进行判断，有助于模型学习临床数据中医生常用的抗生素治疗组合。因此，这 3 个参数的设置分别从医学知识和临床数据完善模型，参数间相互都是最优策略选择需要参考的项目。

3.4.3 最优策略选择对推荐策略的选择分布可视化

为了分析对比最优策略选择模块对最优策略的影响以及选择是否符合医学知识，对最优策略选择结果的 Q-Learning 和 Policy Gradient 的选择占比进行统计，并选择了 4 个个例进行分析。图 7 为最优策略选择在参数权重为 4:1:4 的情况下对 Q-Learning 和 Policy Gradient 推荐策略的选择分布可视化，横坐标为训练次数，纵坐标为策略分布占比。

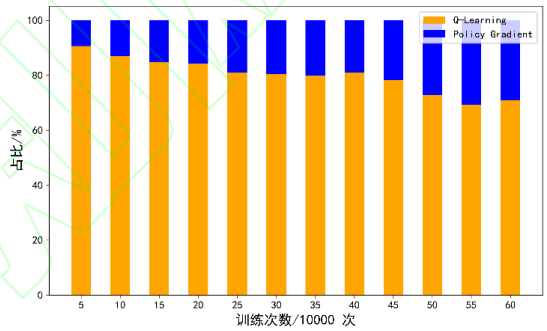

图 7 最优策略选择对 Q-Learning 和 Policy Gradient 推荐策略的选择分布

Fig. 7 Optimal policy selection relies on the choice distribution of recommended policies by Q-Learning and Policy Gradient

随着训练次数增加，经过最优策略选择的 Policy Gradient 推荐的策略占比从不到 10%逐渐增加到 30%左右，这可能是由于基于值函数的方法 Q-Learning 收敛速度快但是训练次数增加可能倒使结果趋同，基于策略的 Policy Gradient 收敛速度慢但是收敛准确性高。这表明最优策略选择在足够多的训练次数后会选择 Policy Gradient 推荐的策略优化 Q-Learning 推荐的策略。

表 5 为起始状态在相同决策区域，最优策略选择了 Q-Learning 推荐的决策的两个个例。患者 A 在临床、Q-Learning 推荐和策略梯度推荐的最后结局为预后良好，但是 Q-Learning 推荐的策略最终状态（最后一次用药时）得分 73.15 高于策略梯度推荐的最终状态得分 70.63。患者 B 在临床和策略梯度推荐的策略结果为预后较差且最终状态得分为

53.19，在 Q-Learning 推荐的策略下预后良好且最终状态得分为 76.35。

表 5 起始状态在相同决策区域，最优策略选择 Q-Learning 推荐的决策

Tab. 5 The initial state is in the same decision region and the optimal policy is to choose the decision recommended by Q-Learning

患者编号	分类	Clinic	Q-Learning	Policy Gradient
A	初始状态得分	30.54	30.54	30.54
	最终状态得分	71.86	73.15	70.63
	预后情况	良好	良好	良好
B	初始状态得分	19.67	19.67	19.67
	最终状态得分	23.13	76.35	53.19
	预后情况	差	良好	差

表 6 为起始状态在相同决策区域，最优选择函数选择了 Policy Gradient 推荐决策的两个个例。利用策略梯度方法推荐的策略，患者 C 在预后良好的情况下，其最终状态得分达到了 81.35，这一分数高于使用 Q-Learning 推荐策略所得到的最终状态得分 80.18。患者 D 的最终状态得分在 Q-Learning 推荐策略下为 53.86 比临床实践有所提高但是预后状况仍较差，在策略梯度推荐的策略下得分 76.31，达到了预后良好的情况。从表 5 和表 6 的结果来看，多策略模型的最优策略选择函数能够选择患者预后状态更好、最终状态得分更高的策略。

表 6 起始状态在相同决策区域，最优策略选择 Policy Gradient 推荐的决策

Tab. 6 Initial state is in the same decision region, and the optimal policy is to choose the decision recommended by Policy Gradient

患者编号	分类	Clinic	Q-Learning	Policy Gradient
C	初始状态得分	26.02	26.02	26.02
	最终状态得分	61.52	80.18	81.35
	预后情况	良好	良好	良好
D	初始状态得分	19.45	19.45	19.45
	最终状态得分	13.13	53.86	76.31
	预后情况	差	差	良好

使用强化学习方法结合对患者状态进行决策区域划分的方法，能够避免聚类带来的患者状态变化细节不可见的情况。使用决策区域划分的方法，保证了每个决策区域内的患者特征的可见性。每个决策区域的中心点与区域内各点的相似度都在 0.8 以上，这保证中心点能够代表这个决策区域的患者特征信息。因此，多策略选择模型能够根据患者的特征信息（决策区域）为患者提供个性化策略。

4 结 语

本研究对脓毒症患者的特征信息进行决策区域划分，并结合 Q-Learning 和 Policy Gradient 两种强化学习方法对脓毒症患者进行个性化的抗生素使用策略推荐，通过最优策略选择函数选取预后更好策略。在决策区域划分和最优策略选择函数中结合临床知识，使多策略选择模型可以在遵循临床和医学指南指导的同时推荐更优策略。多策略推荐模型在改善脓毒症患者的预后情况的同时也能保证患者特征信息对临床医生的高度可视化，为临床医生判断策略的可用性提供有价值的信息。在未来，可以考虑改善对每日抗生素组合的聚类方法，结合更强的强化学习方法完善多策略选择模型。

参考文献：

[1] 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.

[2] 钱建丹,李俊,霍娜,等.《拯救脓毒症运动:2021 年脓毒症和脓毒症休克管理国际指南》感染管理更新要点解读[J].中华传染病杂志,2022,40(7):385-391.

[3] 顾承东. 2012 年《国际严重脓毒症和脓毒症休克治疗指南》解读[C]// 中华医学会急诊医学分会第十六次全国急诊医学学术年会论文集. [出版者不详],2013:160-163.

[4] GREENHILL A T, EDMUNDS B R. A primer of artificial intelligence in medicine[J]. Techniques and innovations in gastrointestinal endoscopy, 2020, 22(2): 85-89.

[5] BENJAMINS J W, HENDRIKS T, KNUUTI J, et al. A primer in artificial intelligence in cardiovascular medicine[J]. Netherlands heart journal, 2019, 27(9): 392-402.

[6] WANG H, PUJOS-GUILLOT E, COMTE B, et al. Deep learning in systems medicine[J]. Briefings in bioinformatics, 2021, 22(2): 1543-1559.

[7] BAMPA M, FASTH T, MAGNUSSON S, et al. EpidRLearn: learning intervention strategies for epidemics with reinforcement learning[C]//International Conference on Artificial Intelligence in Medicine. Cham: Springer International Publishing, 2022: 189-199.

[8] SUN Z, DONG W, LI H, et al. Adversarial reinforcement learning for dynamic treatment regimes[J]. Journal of biomedical informatics, 2023, 137: 104244.

[9] LIU S, SEE K C, NGIAM K Y, et al. Reinforcement learning for clinical decision support in critical care: comprehensive review[J]. Journal of medical internet research, 2020, 22(7): e18477.

[10] PETTIT J F, PETERSEN B K, SILVA F L, et al. Learning sparse symbolic policies for sepsis treatment[R]. Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States), 2021.

[11] LIU X, YU C, HUANG Q, et al. Combining model-based and model-free reinforcement learning policies for more efficient sepsis treatment[C]//Bioinformatics Research and Applications: 17th International Symposium, ISBRA 2021, Shenzhen, China, November 26–28, 2021, Proceedings 17. Springer International Publishing, 2021: 105-117.

[12] 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.

[13] LIN T, ZHANG X, GONG J, 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): 1-12.

[14] ROGGEVEEN L, EL HASSOUNI A, AHRENDT J, et al. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis[J]. Artificial intelligence in medicine, 2021, 112: 102003.

[15] 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.

[16] JU S, KIM Y J, AUSIN M S, et al. To reduce healthcare workload: Identify critical sepsis progression moments through deep reinforcement learning[C]//2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021: 1640-1646.

[17] RAHEB M A, NIAZMAND V R, EQRA N, et al. Subcutaneous insulin administration by deep reinforcement learning for blood glucose level control of type-2 diabetic patients[J]. Computers in biology and medicine, 2022, 148: 105860.

[18] SCHAMBERG G, BADGELEY M, MESCHEDE-KRASA B, et al. Continuous action deep reinforcement learning for propofol dosing during general anesthesia[J]. Artificial intelligence in medicine, 2022, 123: 102227.

[19] ZHANG K, WANG H, DU J, et al. An interpretable RL framework for pre-deployment modeling in ICU hypotension management[J]. NPJ Digital medicine, 2022, 5(1): 173.

[20] ZADEH S A, STREET W N, THOMAS B W. Optimizing warfarin dosing using deep reinforcement learning[J]. Journal of biomedical informatics, 2023, 137: 104267.

[21] 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.

[22] SARIA S. Individualized sepsis treatment using reinforcement learning[J]. Nature medicine, 2018, 24(11): 1641-1642.

[23] KOMOROWSKI M, CELI L A, BADAWI O, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care[J]. Nature medicine, 2018, 24(11): 1716-1720.

[24] LIU R, GREENSTEIN J L, FACKLER J C, et al. Offline reinforcement learning with uncertainty for treatment strategies in sepsis[EB/OL]. 2023-06-01. https://doi.org/10.48550/arXiv.2107.04491.

[25] WATKINS C J C H. Learning from delayed rewards[J]. Robotics and Autonomous Systems, 1995, 15（4）：233.

[26] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[27] SUTTON R S, MCALLESTER D, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation[J]. Advances in neural information processing systems, 1999, 12.

[28] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//International conference on machine learning. PMLR, 2014: 387-395.

[29] QUINN T P, JACOBS S, SENADEERA M, et al. The three ghosts of medical AI: Can the black-box present deliver?[J]. Artificial intelligence in medicine, 2022, 124: 102158.

[30] DAVIDS J, LIDSTRÖMER N, ASHRAFIAN H. Artificial Intelligence in Medicine Using Quantum Computing in the Future of Healthcare[M]//Artificial Intelligence in Medicine. Cham: Springer International Publishing, 2022: 423-446.