·智能决策与博弈·

基于多智能体强化学习的分层决策优化方法

□张　倩　李天皓　白春光
[电子科技大学　成都　611731]

[摘　要]　【目的/意义】随着信息技术和人工智能的发展，大数据驱动的辅助决策方法让决策更加科学准确。强化学习作为序贯决策的经典方法，在决策优化方面有着明显的优势。但传统方法无法解决多层次、多目标的决策优化问题，尤其是在长周期决策优化问题中，学习奖励的滞后性严重制约着强化学习的效率。【设计/方法】提出基于多智能体强化学习的分层决策优化方法，应用目标分解的思想解决长期决策优化问题。该方法基于强化学习理论使具有层级关系的多智能体相互合作，利用神经网络进行建模，上层智能体学习目标的分解策略，下层智能体学习完成目标的行动策略，智能体参数交替更新，共同学习完成团队任务的最佳策略，实现决策优化。【结论/发现】在临床医疗决策优化的实验中验证了该方法的有效性与优越性，可为解决长周期序列决策优化问题提供理论与方法支持。

[关键词]　强化学习；多智能体；分层决策；决策优化
[中图分类号] TP18　　[文献标识码] A　　[DOI] 10.14071/j.1008-8105(2022)-1056

Hierarchical Decision Optimization Method Based on Multi-Agent Reinforcement Learning

ZHANG Qian　LI Tian-hao　BAI Chun-guang
(University of Electronic Science and Technology of China　Chengdu　611731　China)

Abstract　[Purpose/Significance] With the development of information technology and artificial intelligence, big data-driven auxiliary decision-making methods are more scientific and accurate. Reinforcement learning, as a classical method of sequential decision making, has obvious advantages in decision optimization. However, the traditional methods cannot solve the multi-level and multi-objective decision-making optimization problem, especially in the long-term decision-making optimization problem, the lag of learning reward seriously restricts the efficiency of reinforcement learning. [Design/Methodology] This paper proposes a hierarchical decision-making optimization method based on multi-agent reinforcement learning. The idea of goal decomposition is applied to solve the long-term decision-making optimization problem. Based on reinforcement learning theory, multi-agent with hierarchical relationship cooperates with each other and uses neural network to build models. The upper agent learns the decomposition strategy of goals, the lower agent learns the action strategy to complete goals. And the agent parameters are updated alternately, learn the best strategy to complete team tasks together and realize decision optimization. [Conclusions/Findings] The effectiveness and superiority of this method are verified in the

[收稿日期]　2022－08－21
[基金项目]　国家重点研发计划资助（2018AAA0101003）.
[作者简介]　张倩（1998–）女，电子科技大学经济与管理学院硕士研究生；李天皓（1995–）男，电子科技大学经济与管理学院博士研究生.
[通信作者]　白春光（1977–）女，电子科技大学经济与管理学院教授、博士生导师. E-mail：Cbai@uestc.edu.cn.

experiment of clinical medical decision optimization, which can provide theoretical and methodological support for solving the long-term sequence decision optimization problem.

Key words　reinforcement learning; multi-agent; hierarchical decision making; decision optimization

引言

决策是指决策主体选择其行为的过程，决策过程的任一环节出现偏差都有可能导致失误，决策辅助支持系统对提高决策科学性和正确性具有重要作用[1~2]。随着机器学习（Machine Learning, ML）、深度学习（Deep Learning, DL）和大数据等技术的发展和成熟，人工智能技术在辅助决策方面也表现出良好的应用前景，可通过挖掘在线医疗评论等为医疗决策提供参考，为政府智能决策提供优化方案等[3~4]。

传统的决策优化方法主要建立数值模型求最优解[5~6]，方法的计算成本高，且模型泛化能力较差，尤其在长周期连续决策问题中往往效果不佳。作为一种智能决策框架，强化学习（Reinforcement Learning, RL）以马尔可夫决策过程（Markov Decision Process, MDP）为理论基础，采用“试错”的方式进行学习，在连续决策过程中寻找解决问题的最佳策略。智能体通过与环境的交互学习经验，并利用过去的经验来改善未来行动的预期结果，在探索与利用的平衡之间实现奖励的最大化，是一种适应性的学习过程。利用强化学习优化决策问题的研究在围棋、电子游戏、医疗决策、军事战略等领域都取得了显著优于人类决策的效果[7~12]。

强化学习被证实在优化重症监护临床决策、电力系统决策控制、职业道路选择推荐等方面发挥出巨大的作用[13~15]。针对自动驾驶汽车在交通中的决策问题，强化学习可以根据道路情况自主决定驾驶行为，进行车道变更的决策[16~17]。在农业方面，强化学习借助天气预报优化水稻灌溉决策，为农作物疾病的最佳治疗方案提供决策支持[18~19]。在商业领域，智能决策支持系统采用强化学习预测物流网络的变化，也可以为金融市场的股票交易策略提供支持[20~21]。在教育方面，强化学习可基于学习者的个人信息和社交资料推荐最佳的学习方式和适合的学习课程[22]，以提高学习质量。在医疗领域，强化学习在支持临床疾病诊断辅助[23]和个性化用药治疗[24]方面展现出明显的优势，可为智慧医疗建设发挥作用。可见，强化学习已经被应用于社会活动的各个方面，在为决策优化提供辅助和支持方面显示出较强的应用潜力。

在社会决策过程中，决策的结果往往由多个参与者共同决定，强化学习使用多智能体建模多主体决策行为[11]。决策者可以应用多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）算法辅助决策，智能体之间通过竞争与合作方式以最大化团队行动的价值，从而改善决策结果[25]。由于现实中多数决策过程的参与者之间存在明显的等级关系，本文应用具有层级关系的多智能体进行建模，即多智能体分层强化学习（Hierarchical Reinforcement Learning, HRL）[26]。作为多智能体合作强化学习的一种特殊结构，分层强化学习采用层次结构克服多智能体强化学习环境的不稳定性，具有解决稀疏奖励和延迟奖励问题的能力[27]。随着多智能体分层强化学习技术的日益成熟，HRL应用于MOOCs课程推荐、自动驾驶辅助决策、机器人控制等多方面都取得了良好的效果[16,28~30]。

本文基于多智能体强化学习提出了分层深度Q网络（Hierarchical Deep Q-network, HDQ）模型，该模型引入两个智能体相互合作进行学习，在分层模型的基础上，引入目标分解的思想，并结合DL模型，通过神经网络对智能体进行建模，让上层智能体学习最佳的目标分解策略，并将分解的最佳目标传递给下层智能体，指导下层智能体采取行动，通过智能体之间的相互合作，实现团队整体的最终决策目标。

一、多智能体强化学习的分层决策优化

（一）强化学习

强化学习是以马尔可夫决策过程为基础的理论框架，阐述了在解决动态决策问题中智能体与环境的交互过程。强化学习可通过其5个主要要素表示成为5元组$<S,A,P,R,\gamma>$，其中S表示状态集合，A定义智能体可采取的动作集合，P为状态转移矩阵，刻画环境状态的动态变化方式，R是智能体采取动作后获得的奖励集合，$\gamma(0\leqslant\gamma\leqslant1)$表示未来奖励对当前累计奖励的折扣率。强化学习将决策主体建模成能与环境进行动态交互和学习的智能体。在时刻$t=1,2,\cdots,T$时，当智能体采取动作$a_t\in A$，环境会以概率$p(s_{t+1}|s_t,a_t)\in P$从当前状态$s_t\in S$转移到下一个状态$s_{t+1}\in S$，此时智能体获得奖励$r_t\in R$。

RL将决策问题形式化为寻找使预期累计奖励最大化的最优策略[31]，其中预期累计奖励可计算如下：

$$r_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} \tag{1}$$

智能体通过最大化预期累计奖励学习最优策略$\pi^*(a|s)$，可表示为：

$$\pi^*(a|s) = \underset{a}{\operatorname{argmax}}\{r_t\} \tag{2}$$

在基于值的RL方法中，智能体以Q值函数衡量状态s_t下动作a_t的价值，即：

$$Q(s,a) = \mathrm{E}_{a\sim\pi(a|s)}(r_t|s_t = s, a_t = a) \tag{3}$$

并采用V值函数衡量在所有动作a_t下状态s_t的价值：

$$V(s) = \mathrm{E}_{a\sim\pi(a|s)}[Q(s,a)] \tag{4}$$

RL的目标是学习一个最佳的Q值函数$Q^*(s,a)$：

$$Q^*(s,a) = \max\{Q(s,\mathrm{a})\}, (s_t = s, a_t = a) \tag{5}$$

由于传统的RL模型在处理高维数据中具有局限性，DL可以与RL相结合实现更好的决策效果。深度Q网络（Deep Q Network, DQN）利用神经网络在高维空间学习中的优势，引入神经网络作为值函数逼近器，计算最大化累计奖励的最优解。DQN采用带参数θ的神经网络估计动作价值$Q(s,a)$，并基于经验回放机制进行学习，通过最小化误差损失不断逼近最优解：

$$L(\theta) = \mathrm{E}_{s,a}[(r + \gamma \max_{a'}\{Q(s',a';\theta')\} - Q(s,a;\theta))^2] \tag{6}$$

其中，θ'表示目标网络的参数，它是周期性地从DQN网络中复制而来，并在一定的迭代过程中保持不变。

但是传统的DQN模型存在高估Q值的问题，容易跳过最优解学习到次优解，导致模型效果不佳。为了缓解这一问题，Dueling DQN[32]在DQN的基础上引入优势函数衡量动作的相对价值，优势函数$A(s,a)$计算如下：

$$A(s,a) = Q(s,a) - V(s) \tag{7}$$

从而将智能体的目标转化为最大化：

$$Q(s,a) = V(s) + \left[A(s,a) - \frac{1}{|A|}\sum_{a} A(s,a)\right] \tag{8}$$

其中$|A|$是智能体动作空间的大小（即动作的个数）。

（二）目标分解与层级决策

强化学习因其特有的马尔可夫特性而在顺序决策中具有较大优势，但在应用于长周期决策优化问题中，短期内无法衡量动作s_t对最终目标G的影响，智能体在多数时间步内的奖励为0，从而造成奖励的稀疏性，且没有奖励引导容易使智能体陷入困境，影响智能体的学习效率。分层强化学习应用具有层级结构的智能体能够解决稀疏奖励问题，智能体通过决策引导其他智能体采取动作。

在复杂任务的解决过程中，决策周期T通常很长，需要在多个决策时间步$(t = 0,1,2,\cdots,T)$依次决策，且决策的有效性和准确性在短期内无法得到验证。本文的做法是采用分解的思想对目标进行细分，化繁为简，分而治之，通过小目标的实现逐步达成最终目标。分解的思想也被用于解决复杂数据集下的机器学习和数据分析问题[33]，表现出优于基线方法的良好效果。如图1所示，智能体的任务是在决策周期T内实现目标G，在目标分解方法下，智能体学习如何将整体目标G分解为各个子目标$g_t(t = 0,1,2,\cdots,T)$，并通过计算状态$s_t(t = 0,1,2,\cdots,T)$与子目标之间的距离$\mathrm{dis}(g_t, s_t)$判断子目标是否完成，此时，智能体的动作定义为在不同的状态下选择子目标，即a_t:=g_t。

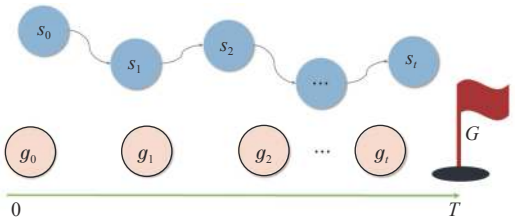

图 1 目标分解过程

本文基于强化学习方法，引入具有层级关系的上层智能体A_1和下层智能体A_2在两个不同的层次等级上进行决策[27]，用于学习实现不同阶段目标的策略，通过智能体的合作决策实现分解的子目标，达成最终目标。上层智能体A_1和下层智能体A_2相互合作，根据环境状态和子目标进行决策，HDQ模型结构如图2所示。智能体A_1学习目标的分解策略π_1，将决策目标G分解为一组子目标$g_1, g_2, \cdots, g_t \in G$，根据对环境状态的观察制定子目标$g_t$，并将其传递给下层智能体$A_2$。智能体$A_2$学习实现子目标$g_t$的行动策略$\pi_2$，根据$A_1$给出的目标$g_t$和环境当前所处的状态$s_t$采取动作$a_t$，随后环境状态依概率$p$转移到下一个状态$s_{t+1}$。在每一时间步$t$内，$A_1$将会根据计算$s_{t+1}$与$g_t$之间的距离$dis(g_t, s_t)$，判断智能体$A_2$对$g_t$的完成情况并给予智能体$A_2$奖励$f_t$。当环境状态发生改变后，智能体$A_1$收到来自环境的奖励$r_t$。智能体$A_1$和智能体$A_2$重复上述分解与行动的过程，直到决策周期$T$终止。奖励$f_t$和$r_t$分别引导智能体$A_1$和智能体$A_2$学习最佳的分解子目标$g_t^*$和实现最终目标的最

佳动作a_t^*。上层智能体A_1和下层智能体A_2的决策过程依次进行，直到决策周期结束$t=T$或整体策略收敛。

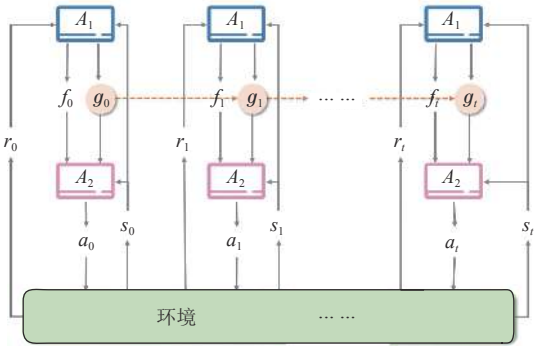

图 2　模型结构

（三）目标优化

上层智能体A_1和下层智能体A_2被建模为带参数θ_1和θ_2的全连接神经网络，均以Dueling DQN作为函数逼近器，学习基于奖励的最佳动作价值函数：

$$Q_1^*(s_t,g_t)=\max\{r_t+\mathrm{E}_{g_t\sim\pi_1}[\Sigma_{t=1}^{\infty}\gamma_1 r_{t+1}|s_t,g_t]\} \tag{9}$$

$$Q_2^*((s_t,g_t),a_t)=\max\{f_t+\mathrm{E}_{a_t\sim\pi_2}[\Sigma_{t=1}^{\infty}\gamma_2 f_{t+1}|(s_t,g_t),a_t]\} \tag{10}$$

其中，γ_1和γ_2分别为智能体A_1和智能体A_2的未来奖励的贴现率，智能体A_2接收来自智能体A_1的子目标作为其状态的一部分，智能体A_1通过最大化动作价值函数选择最佳子目标g_t^*：

$$g_t^*=\mathrm{argmax}\{\mathrm{Q}_1^*(s_t,g_t)\} \tag{11}$$

智能体A_2根据状态和子目标(s_t,g_t)选择动作，并通过最大化动作价值函数选择最佳动作a_t^*：

$$a_t^*=\mathrm{argmax}\{\mathrm{Q}_2^*((s_t,g_t),a_t)\} \tag{12}$$

在每一个时间步t中，智能体A_1和智能体A_2同步学习，通过计算误差损失函数的梯度更新网络参数：

$$\theta_i\leftarrow\theta_i-\alpha_i\nabla L(\theta_i)(i=1,2) \tag{13}$$

其中$\alpha_i(i=1,2)$为梯度下降的步长，即神经网络的学习率，智能体A_1和智能体A_2交替迭代进行学习和参数更新，直至整体策略收敛。

二、实验验证分析

（一）实验背景

脓毒症是由感染引起的危及生命的器官功能障碍，是导致危重患者死亡的主要原因[34]。不同脓毒症患者对治疗措施的个体反应不同[35]，反复住院率高[36]。脓毒症患者的治疗过程是一个长期的、动态的、连续的临床决策过程，对决策质量的要求高，传统的技术方法难以对其进行优化。本文提出的方法一方面可以克服强化学习延迟的奖励和复杂的状态空间容易导致策略的次优性，通过将任务分解为子目标，可以减少探索空间。另一方面，本文方法可以模拟不同等级的医生之间层级指导和合作行为，协同做出治疗决策。我们将结合目前广泛用于研究的医学数据集的MIMIC-IV提取脓毒症患者序列和特征，对患者的治疗决策过程进行优化。

（二）实验设计

1. 数据提取

本文的实验对象为MIMIC-IV数据库中符合Sepsis-3条件[37]的4 800名脓毒症患者。表1显示了原始数据集的汇总，包括存活和死亡患者的比例、平均年龄、男性比例、再入院情况和SOFA评分，其中SOFA评分是脓毒症的顺序器官衰竭评分，根据患者的呼吸系统、血液系统、肝脏系统、心血管系统、神经系统和肾脏系统等六大人体系统相关指标计算而得的分数[37]，是判断脓毒症严重程度的关键指标，与患者死亡有着密切关系[38]。

表 1　患者信息表

	总数	存活患者	病死患者
人数（人）	4 800	3 674	1 126
年龄（岁）	66.11(±16.8)	65.02(±17.1)	69.64(±15.1)
性别（男性）	2 755(57.4%)	2 123(57.8%)	632(56.1%)
再入院	1 165(24.3%)	888(24.2%)	277(24.6%)
SOFA评分	3.66(±2.0)	3.57(±1.9)	3.97(±2.2)

随后，本文提取了患者住院前4小时到住院后72小时的特征，如性别、年龄、体重、SOFA评分、心率、血压、呼吸频率、血氧饱和度、体温、血红蛋白、钾含量、钠含量、凝血酶原时间和血小板数量等在内的45个特征。然后，对每个特征进行一次4小时窗口汇总，使用均值插值方法处理其中的缺失值。其次，使用最大最小归一化方法消除特征之间的量纲，以确保数值在[0,1]区间内。最后，得到了包括45个特征的91 200条可用数据记录，每个患者对应19个历史治疗轨迹。

2. 实验变量与参数选择

（1）状态

状态空间由动态变量和静态变量组成[11]。静态变量包括性别、年龄、体重等信息，动态变量包括患者的生命体征、实验室检查指标和尿液量等数据。由于变量过多容易造成模型的过拟合，影响模型的效果。同时，过多的状态变量容易导致强化学习中的转移矩阵过于稀疏，导致状态转移困难。因此，为了降低数据特征的维数，本文采用了K-means算法对状态进行聚类，以达到降维的目的，避免转

移矩阵的稀疏性。经过聚类处理后，得到了700个不同的状态类别来表示患者的身体状态[39]。

（2）子目标

SOFA评分是衡量脓毒症患者的关键指标，与患者的死亡率密切相关，对于治疗决策的结果有着较大影响。上层智能体A_1学习子目标的分解策略，其动作空间定义为患者的SOFA评分。因此，本文根据数据集中患者的SOFA评分的取值范围对上层智能体的动作进行离散化处理，将其动作空间定义为一维向量，元素取值为[0,18]区间内的整数。因此，在每一个决策的时间步中，上层智能体A_1根据患者状态选择最佳的目标SOFA评分，并将其传递给下层智能体A_2作为子目标。

（3）动作

临床实践中，医生普遍采用血管升压药和静脉输液治疗脓毒症患者。下层智能体A_2在不同的时间步根据状态和子目标选择最优的动作，以学习实现目标的用药策略π_2。动作空间定义为两种药物组成的二维矩阵，分别由血管升压药的最大剂量和静脉注射的总剂量组成，其中血管升压药包括血管升压素、多巴胺、肾上腺素和去甲肾上腺素，静脉注射液包含晶体、胶体和血液制品以及静脉注射抗生素等。药物剂量通过四分位数进行离散化，没有使用药物的患者对应剂量为0。

（4）奖励

奖励函数用于量化行动或目标的有效性。上层智能体A_1获得来自环境的奖励r_t，并通过奖励r_t的大小调整子目标分解策略π_1，由式（14）的分段常数函数组成。状态终止时，即$t=T$，当患者最终存活时获得+15的奖励，当患者最终死亡时获得–15的惩罚。状态未终止时，即$t<T$，当患者状态得到改善时获得+10的奖励，当患者状态出现恶化时获得–10的惩罚。奖励r_t用于引导上层智能体A_1选择对有益于患者生存和状态改善的子目标，改善患者的治疗结果。

$$r_t=\begin{cases}+15(\text{存活})\\-15(\text{死亡})\end{cases}(t=T),r_t=\begin{cases}+10(\text{改善})\\-10(\text{恶化})\end{cases}(t<T) \tag{14}$$

下层智能体A_2获得来自上层智能体A_1的奖励f_t，并通过奖励f_t的大小调整实现目标的行动策略π_2，该奖励基于到达的状态与子目标之间的距离：

$$\text{dis}(g_t,s_t)=|S_{t+1}-g_t| \tag{15}$$

其中S_{t+1}表示A_2采取行动后患者下一时间步的SOFA评分。下层智能体A_2的行动不会受到子目标的严格限制，只要$\text{dis}(g_t,s_t)\leqslant\eta$（$\eta$为常数）便能够获得奖励：

$$f_t=c\times\text{dis}(g_t,s_t)(c<0) \tag{16}$$

（三）实验结果分析

实验环境基于Python 3.6和TensorFlow 1.15，两个智能体均由神经网络进行建模，网络的学习率α设置为0.01。算法基于强化学习建模，强化学习模型的奖励衰减折扣γ设置为0.9，下层智能体的目标阈值η设置为2，模型基于以上参数进行训练。

根据实验设计，本文对照临床医生的用药决策、无层次结构的单智能体Dueling DQN以及DQN作为基准模型，基准模型的所有参数设置和训练迭代次数均相同。通过与基准模型进行对比，评估HDQ模型在决策优化问题中的效果。

在强化学习中，Q值的大小用于衡量模型所选动作的价值高低，图3显示了本文的HDQ模型与Dueling DQN和DQN两个基准模型在训练过程中的Q值比较情况，其中横轴表示模型训练的迭代次数，纵轴表示模型的Q值大小。在模型迭代训练10 000轮后，三个模型的都达到了收敛，学习到了有效的稳定策略。根据图3可见，在训练前期，模型都倾向于选择具有较高Q值的动作。随着训练过程的进行，模型学习调整动作的选择，导致Q值不断减小并最终收敛。同时，根据Q值比较结果也可以发现，Dueling DQN模型的效果优于传统的DQN算法，但与本文的HDQ模型相比还有一定的差距。与基准模型相比，本文的HDQ模型在动作的选择上具有明显优势，在收敛条件下，Q值显著高于基准模型。

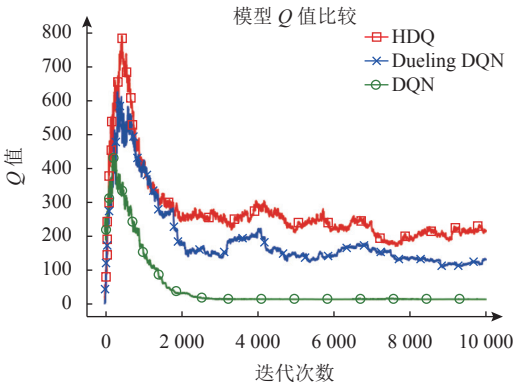

图 3 不同模型Q值比较

死亡率是衡量医疗用药决策策略有效性的重要指标，对患者的治疗结果有着决定性作用，表2列出了不同策略下患者住院死亡率的比较。整体而言，临床医师治疗策略下的患者死亡率是最高的，高达23.5%。对于无分层结构的模型来说，不论是

Dueling DQN还是DQN算法，都能够在临床医师的基础上通过优化决策策略，达到降低患者死亡率的目标，但改进后的Dueling DQN算法在策略优化方面的效果会比DQN更加显著。

表 2　模型死亡率

基准策略	Dueling DQN	DQN	HDQ	临床医师
平均死亡率	16.6%	17. 7%	13.2%	23. 5%

显然，本文提出的HDQ模型在降低患者死亡率方面比临床医师和无分层结构的模型更有优势。虽然其他算法也可以通过推荐药物剂量达到降低患者死亡率的目的，但通过HDQ模型学习的用药策略的死亡率是最低的，比DQN算法低4.5%，比没有层次结构的Dueling DQN结构低3.4%，相比于临床医生的死亡率降低了10.3%。

由此可见，HDQ模型在临床决策优化方面具有显著的优势，这也证明临床医师的决策还有较大的优化空间，无论是本文的分层模型，还是现有的非分层模型，都能实现临床医师策略的优化。

三、结论

针对社会中面临的长期决策优化问题，本文提出了一种基于多智能体强化学习的分层决策优化（HDQ）算法，以目标分解和层级合作的方式实现长周期顺序决策优化。在所提的HDQ算法模型中，具有层级关系的多智能体基于强化学习理论相互合作，上层智能体学习最佳的目标分解，下层智能体学习在子目标指导与约束下完成目标的行动策略，从而共同构建团队任务的最佳策略、实现决策优化。为了检验该模型的决策效率，本文提取MIMIC-IV数据集对脓毒症患者的临床诊疗决策问题进行了分析验证，发现该算法既能避免强化学习延迟奖励和复杂状态空间导致的策略次优性，还能模拟出不同等级医生之间的层级指导和合作行为，进而协同做出优于人类临床医师的治疗决策。

与传统的智能决策算法相比，HDQ算法具有明显的优越性，弥补了传统决策方法模型泛化能力较差、长周期连续决策效率低下的不足，适用于具有连续决策过程的策略优化问题。尽管如此，本文的决策方法在实际运用过程中仍可能存在一定局限性；这是由于该方法作为一种独立学习的方式，采用两个智能体交替学习和更新，下层智能体完成目标的行动策略将高度依赖于上层智能体对子目标分解的合理性。因此，未来可进一步探索消除上层智能体学习训练结果对模型性能产生负面影响的方法，并针对模糊环境下多层次、多目标的决策问题开展研究。

参考文献

[1] LEE D, SEO H, JUNG M W. Neural basis of reinforcement learning and decision making[J]. Annual Review of Neuroscience, 2012, 35: 287.

[2] 任德胜, 郭凤仙. 公共决策失误的根源探析[J]. 电子科技大学学报(社科版), 2006(1): 53-56.

[3] 廖虎昌, 刘凡, 卢柯宇, 等. 基于在线医疗评论的患者行为挖掘及其在医疗决策与管理中的应用综述[J]. 电子科技大学学报(社科版), 2022(3): 1-22.

[4] 肖进, 李春燕, 贾品荣. 人工智能背景下政府治理智能决策优化研究[J]. 电子科技大学学报(社科版), 2021(5): 42-48.

[5] 曲佳莉, 胡本勇. 供应和需求不确定下考虑收益共享合约的供应链决策优化与协调研究[J]. 电子科技大学学报(社科版), 2017(4): 57-61+68.

[6] 程建刚, 李从东. 服务供应链网络优化模型及解法[J]. 电子科技大学学报(社科版), 2009(1): 61-64.

[7] 周瑞朋. 深度强化学习中探索与利用的研究[D]. 贵阳: 贵州大学, 2021.

[8] SILVER D, HUBERT T, SCHRITTWIESER J, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play[J]. Science, 2018, 362(6419): 1140-1144.

[9] RAFATI J, NOELLE D C. Learning representations in model-free hierarchical reinforcement learning[C]. Honolulu: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 10009-10010.

[10] SHORTREED S M, LABER E, LIZOTTE D J, et al. Informing sequential clinical decision-making through reinforcement learning: an empirical study[J]. Machine Learning, 2011, 84(1): 109-136.

[11] LI T, WANG Z, LU W, et al. Electronic health records based reinforcement learning for treatment optimizing[J]. Information Systems, 2022, 104: 101878.

[12] 王帆, 陈楚湘, 王运成. 军事战略智能决策支持系统[J]. 信息系统工程, 2022(5): 128-131.

[13] 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.

[14] PETERS M, KETTER W, SAAR-TSECHANSKY M, et al. A reinforcement learning approach to autonomous decision-making in smart electricity markets[J]. Machine Learning, 2013, 92(1): 5-39.

[15] KOKKODIS M, IPEIROTIS P G. Demand-aware career path recommendations: a reinforcement learning

approach[J]. Management Science, 2021, 67(7): 4362-4383.

[16] YOU C, LU J, FILEV D, et al. Highway traffic modeling and decision making for autonomous vehicle using reinforcement learning[C]. Suzhou: 2018 IEEE Intelligent Vehicles Symposium (IV), 2018: 1227-1232.

[17] WANG J, ZHANG Q, ZHAO D, et al. Lane change decision-making through deep reinforcement learning with rule-based constraints[C]. Budapest: 2019 International Joint Conference on Neural Networks (IJCNN), 2019: 1-6.

[18] CHEN M, CUI Y, WANG X, et al. A reinforcement learning approach to irrigation decision-making for rice using weather forecasts[J]. Agricultural Water Management, 2021, 250: 106838.

[19] BINAS J, LUGINBUEHL L, BENGIO Y. Reinforcement learning for sustainable agriculture[C]. Conifornia: ICML 2019 Workshop Climate Change: How Can AI Help.

[20] RABE M, DROSS F. A reinforcement learning approach for a decision support system for logistics networks[C]. Conifornia: 2015 Winter Simulation Conference (WSC), 2015: 2020-2032.

[21] LI Y, NI P, CHANG V. An empirical research on the investment strategy of stock market based on deep reinforcement learning model[C].Heraklion: 4th International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS), 2019: 52-58..

[22] MADANI Y, EZZIKOURI H, ERRITALI M, et al. Finding optimal pedagogical content in an adaptive e-learning platform using a new recommendation approach and reinforcement learning[J]. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(10): 3921-3936.

[23] LIU Z, YAO C, YU H, et al. Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things[J]. Future Generation Computer Systems, 2019, 97: 1-9.

[24] ZHANG Z. Reinforcement learning in clinical medicine: a method to optimize dynamic treatment regime over time[J]. Annals of translational Medicine, 2019, 7(14): 345.

[25] 闫超, 相晓嘉, 徐昕, 等. 多智能体深度强化学习及其可扩展性与可迁移性研究综述[J]. 控制与决策, 2022, 37(12): 1-20.

[26] BOTVINICK M M. Hierarchical reinforcement learning and decision making[J]. Current opinion in Neurobiology, 2012, 22(6): 956-962.

[27] KULKARNI T D, NARASIMHAN K, SAEEDI A, et al. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation[J]. Advances in Neural Information Processing Systems, 2016, 29: 3682-3690.

[28] ZHANG J, HAO B, CHEN B, et al. Hierarchical reinforcement learning for course recommendation in MOOCs[C].Honolulu: The AAAI Conference on Artificial Intelligence, 2019, 33(1): 435-442.

[29] DUAN J, EBEN L S, GUAN Y, et al. Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data[J]. IET Intelligent Transport Systems, 2020, 14(5): 297-305.

[30] BEYRET B, SHAFTI A, FAISAL A A. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation[C]. Macau: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 5014-5019.

[31] SUTTON R S, BARTO A G. Reinforcement Learning-An Introduction[M]. London: The MIT Press, 1998.

[32] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]. Jeju Lsland: International conference on machine Learning, 2016: 1995-2003.

[33] ZUPAN B. Machine learning based on function decomposition[D].Ljubljana, Slowenia: University of Ljubljana,1997.

[34] KAUKONEN K-M, BAILEY M, SUZUKI S, et al. Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000-2012[J]. The Journal of American Medical Association, 2014, 311(13): 1308-1316.

[35] PRESCOTT H C, ANGUS D C. Enhancing recovery from sepsis: a review[J]. The Journal of American Medical Association, 2018, 319(1): 62-75.

[36] MEYER N, HARHAY M O, SMALL D S, et al. Temporal trends in incidence, sepsis-related mortality, and hospital-based acute care after sepsis[J]. Critical care Medicine, 2018, 46(3): 354.

[37] SINGER M, DEUTSCHMAN C S, SEYMOUR C W, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3)[J]. The Journal of American Medical Association, 2016, 315(8): 801-810.

[38] DE G H-J, GEENEN I L, GIRBES A R, et al. SOFA and mortality endpoints in randomized controlled trials: a systematic review and meta-regression analysis[J]. Critical Care, 2017, 21(1): 1-9.

[39] 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.

编 辑 邓婧