基于多智能体强化学习的乳腺癌致病基因预测

刘 健[1, 2] 顾 扬[1, 2] 程玉虎[1, 2] 王雪松[1, 2]

摘 要 通过分析基因突变过程, 提出利用强化学习对癌症患者由正常状态至患病状态的过程进行推断, 发现导致患者死亡的关键基因突变. 首先, 将基因视为智能体, 基于乳腺癌突变数据设计多智能体强化学习环境; 其次, 为保证智能体探索到与专家策略相同的策略和满足更多智能体快速学习, 根据演示学习理论, 分别提出两种多智能体深度 Q 网络: 基于行为克隆的多智能体深度 Q 网络和基于预训练记忆的多智能体深度 Q 网络; 最后, 根据训练得到的多智能体深度 Q 网络进行基因排序, 实现致病基因预测. 实验结果表明, 提出的多智能体强化学习方法能够挖掘出与乳腺癌发生、发展过程密切相关的致病基因.

关键词 乳腺癌, 致病基因, 基因排序, 多智能体强化学习, 演示学习

引用格式 刘健, 顾扬, 程玉虎, 王雪松. 基于多智能体强化学习的乳腺癌致病基因预测. 自动化学报, 2022, **48**(5): 1246−1258

DOI 10.16383/j.aas.c210583

Prediction of Breast Cancer Pathogenic Genes Based on Multi-agent Reinforcement Learning

LIU Jian[1, 2] GU Yang[1, 2] CHENG Yu-Hu[1, 2] WANG Xue-Song[1, 2]

Abstract By analyzing the gene mutation process, it is proposed to use reinforcement learning to infer the process of cancer patients from normal to disease states, and to discover the key gene mutations that lead to the death of patients. Firstly, a multi-agent reinforcement learning environment is designed based on breast cancer mutation data by viewing genes as agents. Secondly, in order to ensure that agents can find the same policy as expert policy and to satisfy more agents for rapid learning, two kinds of multi-agent deep Q networks are proposed based on demonstration learning respectively: Behavioral Cloning-based multi-agent deep Q network and pre-training memory-based multi-agent deep Q network. Finally, we sort genes according to the trained multi-agent deep Q network to achieve pathogenic gene prediction. Experimental results show that the proposed multi-agent reinforcement learning methods can dig out pathogenic genes closely related to the occurrence and development of breast cancer.

Key words Breast cancer, pathogenic genes, gene sorting, multi-agent reinforcement learning, demonstration learning

Citation Liu Jian, Gu Yang, Cheng Yu-Hu, Wang Xue-Song. Prediction of breast cancer pathogenic genes based on multi-agent reinforcement learning. *Acta Automatica Sinica*, 2022, **48**(5): 1246−1258

基因突变是由 DNA 分子中碱基对发生增添、缺失或替换而引起的基因结构变化. 基因突变具有随机性, 是一种可遗传的变异现象. 致病基因突变通过阻止一种或多种蛋白质正常工作扰乱正常发育过程或导致疾病. 癌症是由控制细胞功能的基因突变引起的一系列相关疾病的统称. 导致癌症的基因突变可能遗传自父母, 也可能是人体自身受致癌环境或致癌物质刺激导致细胞分裂时产生的错误. 一般来说, 癌细胞比正常细胞有更多的基因突变. 乳腺癌是世界上最常见的疾病之一, 2018 年新增乳腺癌患者约 20 亿人[1]. 医学领域的多项研究表明, BRCA1、BRCA2 和 PALB2 基因的突变会导致乳腺癌风险增加, 其他与乳腺癌患病风险相关的基因突变包括 ATM、TP53、PTEN 等. 因此, 从乳腺癌组学数据中挖掘出与其密切相关的致病基因对乳腺癌的临床诊断、预后和治疗有着深远意义.

在生物信息学中, 癌症致病基因预测通过基因排序方法实现. 基于网络相似度的基因排序算法通过分析多种基因–疾病网络中的局部、全局信息, 计算基因与疾病之间的相似性, 从而对基因进行排序.

收稿日期 2021-06-27 录用日期 2021-11-26
Manuscript received June 27, 2021; accepted November 26, 2021
国家自然科学基金 (61906198, 61976215, 62176259), 江苏省自然科学基金 (BK20190622) 资助
Supported by National Natural Science Foundation of China (61906198, 61976215, 62176259), Natural Science Foundation of Jiangsu Province (BK20190622)
本文责任编委 黄华
Recommended by Associate Editor HUANG Hua
1. 地下空间智能控制教育部工程研究中心 徐州 221116 2. 中国矿业大学信息与控制工程学院 徐州 221116
1. Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, Xuzhou 221116 2. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116

例如, Kohler 等[2] 提出重启随机游走算法利用网络全局拓扑信息对致病基因进行预测; Xu 等[3] 提出多路径随机游走的网络嵌入模型对异构网络进行致病基因预测. 这些方法过度依赖网络拓扑信息, 不能对网络外的基因进行预测, 且对癌症数据中的噪声比较敏感. 随着机器学习理论的发展, 基于机器学习的基因排序方法利用监督学习或非监督学习方式实现基因预测, 能够挖掘到与癌症相关的致病基因, 被广泛应用于癌症致病基因的预测. 例如 Han 等[4] 将图卷积网络和矩阵分解结合提出一种疾病基因关联任务框架; Natarajan 等[5] 将推荐系统中的归纳矩阵补全用于预测基因与疾病的相关性.

在乳腺癌致病基因预测方面, 自然启发式算法应用较广, 例如粒子群优化 (Particle swarm optimization, PSO)、遗传算法等. Sahu 等[6] 提出一种基于 PSO 的基因选择算法, 首先采用 k 均值聚类方法对数据集进行聚类, 利用信噪比评分对聚类簇中的基因进行排序, 然后从每个聚类簇中收集得分最高的基因生成新的特征子集, 最后将新特征子集作为 PSO 的输入, 生成优化后的特征子集. Malar 等[7] 通过将关联特征选择方法和改进的二进制 PSO 结合选择致病基因, 同时解决了微阵列数据的高维性问题. 为了消除对乳腺癌无意义的基因, AliazKovic 等[8] 将遗传算法用于提取乳腺癌数据中的重要信息, 挖掘与乳腺癌生物过程相关的致病基因. Sangaiah 等[9] 将特征加权和基于熵的遗传算法结合起来, 提出一种乳腺癌致病基因预测的混合方法. Alzubaidi 等[10] 将遗传算法与互信息结合应用于乳腺癌致病基因选择. 通过遗传算法将基于互信息的基因选择算法转化为全局优化算法, 能够有效选择基因. 避免算法陷入局部最优. Alomari 等[11] 结合最小冗余、最大关联算法和花授粉算法来确定包含更多癌症信息的基因子集. Hamim 等[12] 提出一种基于决策树模型的乳腺癌致病基因选择策略, 该策略包括两个阶段: 基于 Fisher 评分的过滤阶段和基于 C5.0 算法的基因选择阶段. Liu 等[13] 为了提高基因选择效率, 将基因评分与深度神经网络产生的基因重要性相结合, 同时考虑癌症亚型间的差异性和亚型内基因间的相关性来选择乳腺癌三阴性亚型的最优致病基因子集. Zhao 等[14] 基于信息熵的不确定性系数被用来定义基因间是否存在逻辑关系, 进而构建基因逻辑网络, 最终通过比较对照组与实验组网络之间的差异程度, 提取乳腺癌致病基因.

上述预测方法都是基于已有癌症组学数据进行基因预测, 这些组学数据来源于对癌症患者的测序. 换言之, 这些方法仅能根据目前已发病患者的基因突变状态来分析基因与癌症之间的关联, 无法预知患者发病前的基因突变状态, 而发病前的基因突变状态与发病基因突变状态之间的差异才是癌症发生的关键.

强化学习[15] 是一类结合了优化控制思想和生命体学习行为的机器学习方法, 其要求待处理的问题环境拥有马尔可夫性质, 即当前状态仅受上一状态的影响, 与其余状态无关. 强化学习希望智能体在指定的状态能够得到让回报最大化的动作, 并通过智能体与环境的交互进行学习, 从而改变特定状态选择某个动作的趋势. 强化学习还是一种拥有自主决策能力的算法, 它使智能体通过在环境中的不断试错得到回报值和下一时刻状态的观测值, 最终学习到一个能够获取较大折扣累积回报的策略. 强化学习已被成功应用于多个研究领域, 例如, 数据驱动控制[16]、多机协同决策[17]、交通控制[18] 等.

本文通过分析基因突变, 发现其过程满足马尔可夫过程, 且基因突变与癌症之间的关联性可以通过强化学习中累计回报函数构建的方式进行计算. 因此, 基于乳腺癌突变数据, 本文设计一套强化学习环境与算法对患者从正常基因突变状态至死亡基因突变状态的过程进行评估、决策, 旨在为癌症致病基因预测提供新思路, 并挖掘出导致乳腺癌死亡状态的致病基因. 实验结果表明, 提出的强化学习算法能够挖掘出与乳腺癌密切相关的致病基因.

1 问题描述

由于基因突变并非确定性事件, 在非人为干涉的前提下, 基因突变可视为一个随机过程. 设任意 t 时刻基因突变状态 (后文简称状态) 为 $\boldsymbol{s}_t$, 下一时刻状态为 $\boldsymbol{s}_{t+1}$, 则在 $t+1$ 时刻状态发生的变化只与 t 时刻的状态有关, 与之前 $0 \sim t-1$ 的状态并无关联, 即

$$P\left(\boldsymbol{s}_{t+1} \left| \boldsymbol{s}_0, \boldsymbol{s}_1, \cdots, \boldsymbol{s}_t\right.\right) = P\left(\boldsymbol{s}_t\right) \tag{1}$$

其中, $P(\cdot)$ 为概率. 基于上述考虑, 可以认为基因突变对应的随机过程为马尔可夫过程.

本文根据乳腺癌患者生存数据中患者的临床信息来定义死亡状态和非死亡状态. 患者生存数据兼有时间和结局两种属性信息. 时间描述的是患者由观察起点至观察终点的时间间隔, 通常称为生存时间. 患者生存数据的结局即为观察终点, 观察终点分为死亡和存活两种, 在生存数据中记为 1 和 0. 在本文中, 如果某患者的观察终点为死亡, 则将该患者在乳腺癌数据中的基因突变状态定义为死亡状态. 值得注意的是, 具有相同基因突变状态的患者, 观察终点并不一定相同, 因此通过定义死亡率来更加精细地对数据进行描述. 若基因突变状态使所有

癌症患者死亡, 则该状态的死亡率为 100%; 若基因突变状态有一定概率导致患者死亡, 例如 100 个患者有相同的状态, 其中有 10 个患者死亡, 则死亡率为 10%. 这里将有概率死亡的基因突变状态统称为死亡状态. 设一个基因与 t 时刻状态 $\boldsymbol{s}_t$ 之间的关联性为 $r(\boldsymbol{s}_t)$, 已有基因排序算法更关注对历史病例数据的数理统计, 通过计算 $r(\boldsymbol{s}_t)$ 的大小来评价某个基因突变与癌症患者之间的联系强弱. 然而这类方法没有充分考虑患者的死亡状态, 且忽视了癌症的发生过程, 比如死亡状态 $\boldsymbol{s}_\alpha$ 虽然死亡率不高, 且 $r(\boldsymbol{s}_t)$ 值较小, 但可能在一定时期内突变成死亡率很高的其他状态, 这类状态 $\boldsymbol{s}_\alpha$ 中的基因与癌症患者死亡之间的应该有很强的关联性. 因此, 对基因与癌症患者之间关联的评估不应只关注状态 $\boldsymbol{s}_t$ 中基因与癌症关联性, 更应从一个正常状态经历漫长基因突变过程至死亡状态的角度, 评估突变基因与某个死亡状态的关联性, 即 $\sum_i r(\boldsymbol{s}_i)$.

乳腺癌突变数据中, 每个患者的所有基因突变状态是一个样本, 每个基因在所有患者上的突变状况是一个特征, 如图 1 所示. 患者的某个基因发生突变, 则记为 1 (图 1 中黑色格子), 不发生突变则记为 0 (图 1 中非黑色格子). 本文构建强化学习环境如下: 将基因作为智能体 (Agent), t 时刻基因突变状况作为状态 $\boldsymbol{s}_t$, 基因突变作为动作 $\boldsymbol{a}_t$, 根据死亡状态的死亡率设计回报函数 $r(\boldsymbol{s}_t)$, 当智能体达到死亡状态时获得最优策略, 停止与环境交互, 给予高回报值. 基因突变数据中的基因数目成百上千, 在一个状态中, 使用单智能体进行强化学习时, 状态−动作空间复杂度极高, 需要大量计算成本. 为此, 考虑利用多智能体深度 Q 网络 (Deep Q network, DQN)[19] 对乳腺癌突变数据进行强化学习. 一方面, 相比于 Q 学习方法, DQN 通过训练更新值函数神经网络的参数, 减小状态高维度对算法训练效果的影响; 另一方面, 使用多智能体进行强化学习, 可降低动作空间复杂度, 大大减少强化学习的计算量.

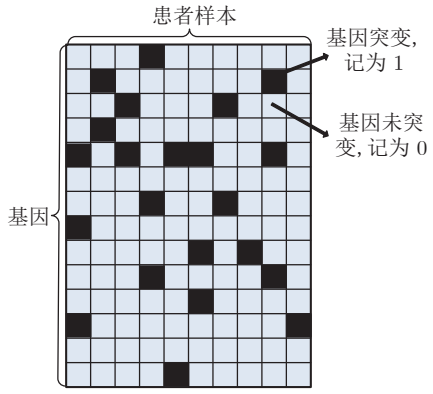


图 1 乳腺癌突变数据

Fig. 1 Breast cancer mutation data

多智能体 DQN 使得学习任务的复杂度减小, 但多智能体的动作维度并没有下降, 智能体探索到最优策略的概率很低. 由于所有死亡状态均来自乳腺癌突变数据, 可将死亡状态作为专家意见指导强化学习过程, 根据演示学习理论, 提出两种多智能体 DQN: 基于行为克隆的多智能体 DQN (Behavioral cloning-based multi-agent DQN, BCDQN) 和基于预训练记忆的多智能体 DQN (Pre-training memory-based multi-agent DQN, PMDQN). 设置探索经验池 B_1 和演示经验池 B_2 两个经验池 , 更好地实现演示学习. 当智能体数量较少时, BCDQN 使智能体在每一步探索时都给出专家意见, 保证 B_1 和 B_2 在状态上同分布, 实现探索策略对专家策略的完全克隆; 当智能体数量较大时, PMDQN 通过预训练将一定数量的专家经验保存在 B_2 中, 再使智能体随机探索填充 B_1, 并通过训练最终实现 B_1 和 B_2 同分布, 这能够使 B_2 中样本之间的相关性下降, 从而加快算法的学习.

2 环境设计

设基因数为 N, 构建一个状态、动作维度都为 N 的状态−动作空间, 则状态空间 S 中任一状态 $\boldsymbol{s}_t = [s_t^1, s_t^2, \cdots, s_t^N]$ 为 N 维二进制向量, 其中 $s_t^k (k = 1, 2, \cdots, N)$ 的取值满足: 基因在 s_t^k 上发生突变则 $s_t^k = 1$, 不发生突变则 $s_t^k = 0$. 动作空间 A 中动作 $\boldsymbol{a}_t = [a_t^1, a_t^2, \cdots, a_t^N]$ 为 N 维二进制向量, 其中 $a_t^k (k = 1, 2, \cdots, N)$ 满足: 基因在 s_t^k 下一状态发生突变则调整 $a_t^k = 1$, 不发生突变则保持 $a_t^k = 0$. 状态间的状态转移 $\boldsymbol{s}_{t+1}$ 满足

$$\boldsymbol{s}_{t+1} = \boldsymbol{s}_t \oplus \boldsymbol{a}_t = \left[s_t^k \oplus a_t^k, \cdots, s_t^k \oplus a_t^k\right] \tag{2}$$

其中, $\oplus$ 为异或运算. 定义汉明距离 D 为:

$$D(\boldsymbol{s}_t, \boldsymbol{s}_{t+1}) = \sum_{i=1}^{N} s_t^k \oplus s_{t+1}^k = \|\boldsymbol{a}_t\|_1 \tag{3}$$

回报函数 $r(\boldsymbol{s}_t)$ 定义为:

$$r(\boldsymbol{s}_t) = \begin{cases} -1 - \eta D(\boldsymbol{s}_t, \boldsymbol{s}_{t+1}), & \text{Alive} \\ -\eta D(\boldsymbol{s}_t, \boldsymbol{s}_{t+1}), & \text{Dead} \end{cases} \tag{4}$$

式中, 设死亡状态 (Dead) 的死亡率为 P_d, 即若状态对应的死亡率不为 0, 则智能体在该状态有 P_d 的概率死亡. 若智能体触发死亡事件, 则停止智能体与环境的交互. 智能体在环境中探索时, 智能体如果存活则给予智能体负的回报, 智能体在环境中存活

的时间越长, 对应的累积回报 $\sum_{i=t}^{\infty}\gamma^{i-t}r(\boldsymbol{s}_i)$ 就越低, 其中, $\gamma\,(0<\gamma<1)$ 为折扣因子. 式 (4) 中的 D 则限制了状态的变化幅度, 以避免违背基因突变的客观规律, 即智能体要想获得更高的回报则必须要用较小动作幅度触发死亡事件. 由于 D 值在 N 足够大情况下会远大于 1, 由霍夫丁不等式可知, 随机变量总和与其期望值之间的偏差上限与随机变量取值区间大小正相关. 因此, 使用常数 $\eta\,(0<\eta<1)$ 限制回报变化幅度, 降低学习任务的复杂度.

3　基于多智能体强化学习的乳腺癌致病基因预测

强化学习目标是找到最优策略 $\pi^*=P(\boldsymbol{a}_t\,|\boldsymbol{s}_t)$, 即最大化期望折扣回报

$$\mathrm{E}\left[\sum_{i=t}^{\infty}\gamma^{i-t}r(\boldsymbol{s}_i)\right] \tag{5}$$

常用的强化学习算法为异步策略的 Q 学习方法[6]. 对于当前的学习问题, Q 学习方法的迭代公式为

$$Q(\boldsymbol{s}_t,\,\boldsymbol{a}_t)=Q(\boldsymbol{s}_t,\,\boldsymbol{a}_t)+\alpha\left(r(\boldsymbol{s}_t)+\gamma\max_{\boldsymbol{a}}Q(\boldsymbol{s}_{t+1},\,\boldsymbol{a})-Q(\boldsymbol{s}_t,\,\boldsymbol{a}_t)\right) \tag{6}$$

从式 (6) 可以看出, Q 学习方法要求智能体使用贪心算法进行动作选择, 从而刚性保证算法的收敛. Q 学习方法倾向于直接估计状态−动作值矩阵. 在所设计的环境中, 状态、动作都是二进制向量, 所以动作空间复杂度为 2^{N+1}, 状态空间复杂度为 2^N. 如果使用 Q 学习方法, 则需要估计复杂度为 2^{2N+1} 的值函数矩阵. Q 学习方法在 N 很大时, 需要耗费大量时间遍历求解值函数矩阵. 为此, 本文选择使用 DQN 通过神经网络训练更新值函数的参数, 减小状态维度对算法训练效果的影响. DQN 的更新目标为

$$Y_t=r(\boldsymbol{s}_t)+\gamma\max_{\boldsymbol{a}}Q(\boldsymbol{s}_{t+1},\,\boldsymbol{a}) \tag{7}$$

相应的损失函数为

$$L(\boldsymbol{\theta}^k)=\mathrm{E}\left[(Y-Q(\boldsymbol{s},\,\boldsymbol{a};\boldsymbol{\theta}))^2\right] \tag{8}$$

其中, $\boldsymbol{\theta}$ 为值函数网络参数. DQN 采用经验回放技术, 训练值函数网络所用的数据需要从环境交互得到的经验信息中随机采样得到, 以消除训练数据之间的相关性, 从而满足深度学习对训练集数据独立同分布的前提条件. DQN 可以高效处理状态−动作空间维度较大的学习问题, 并通过经验回放技术提高经验数据的利用效率.

3.1　多智能体 DQN

本文实验环境如果使用单智能体深度强化学习算法, 则其状态−动作空间复杂度为 2^{2N+1}; 如果使用多智能框架, 则会使 2^{N+1} 的动作空间复杂度变为 $2N$, 整体上的状态−动作空间复杂度则变为 $N2^{N+1}$. 环境所使用的基因数 N 一般很大, 因此 $N2^{N+1}\ll 2^{2N+1}$, 多智能体框架可以大幅降低学习问题的复杂程度, 减少了设计单智能体所需的网络参数.

多智能体强化学习框架如图 2 所示. 首先, 将 $\boldsymbol{s}_t=\left[s_t^1,\,s_t^2,\,\cdots,\,s_t^N\right]$ 输入到具有 N 个智能体的值网络中, 根据 t 时刻每个基因的突变状态, 分别输出动作 a_t^k, 并将输出的 a_t^k 组合成 $\boldsymbol{a}_t$, 进而生成新状态 $\boldsymbol{s}_{t+1}$. 之后, 根据乳腺癌突变数据中患者的死亡状态, 判断是否停止与环境交互, 如果不停止, 则将 $\boldsymbol{s}_{t+1}$ 输入网络继续上述迭代过程.

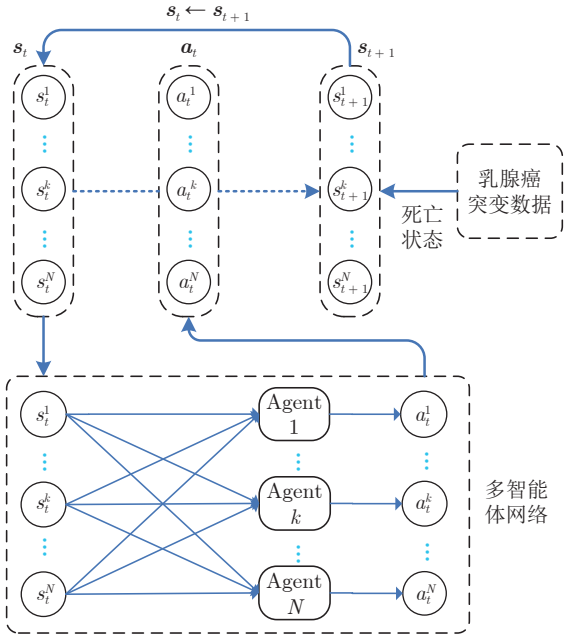


图 2　多智能体强化学习框架 (以第 k 个智能体为例)

Fig. 2　Multi-agent reinforcement learning framework (Take the k-th agent as an example)

每个智能体的更新目标为

$$Y_t^k=r(\boldsymbol{s}_t)+\gamma\max_{a^k}Q^k\left(\boldsymbol{s}_{t+1},\,a^k;\boldsymbol{\theta}^k\right) \tag{9}$$

其中, 第 k 个智能体的动作 a^k 属于各自的动作空间 A^k, $\boldsymbol{\theta}^k$ 则为第 k 个智能体的值函数网络参数. 第 k 个智能体系统的损失函数为

$$L(\boldsymbol{\theta}^k)=\mathrm{E}\left[\sum_{k=1}^{N}\left(Y^k-Q^k\left(\boldsymbol{s},\,a^k;\boldsymbol{\theta}^k\right)\right)^2\right] \tag{10}$$

多智能体 DQN 的伪代码如算法 1 所示.

算法 1. 多智能体 DQN

输入: 最大迭代次数 $I_{\max}$, 折扣因子 γ, 学习率 η, 智能体个数 N.

输出: 网络参数 $\boldsymbol{\theta}^k\ (k=1, 2, \cdots, N)$.

1) 初始化网络参数 $\boldsymbol{\theta}^k\ (k=1, 2, \cdots, N)$;

2) While $I < I_{\max}$;

3) $t=0$;

4) 随机初始化状态 $\boldsymbol{s}_t$;

5) While $t \le t_{\max}$ or 患者死亡;

6) For $k=1:N$;

7) 计算动作: $a_t^k = \arg\max\limits_{a^k} Q^k\left(\boldsymbol{s}_t, a^k; \boldsymbol{\theta}^k\right)$;

8) end For;

9) 环境中应用动作 $\boldsymbol{a}_t = [a_t^1, a_t^2, \cdots, a_t^N]$, 并返回回报 $r\left(\boldsymbol{s}_t\right)$ 和下一时刻状态 $\boldsymbol{s}_{t+1}$;

10) $t \leftarrow t+1$;

11) end While;

12) $I \leftarrow I+1$;

13) For $k=1:N$;

14) 随机采样并更新 $\boldsymbol{\theta}^k$:

$$\boldsymbol{\theta}^k \leftarrow \boldsymbol{\theta}^k + \eta \nabla_{\boldsymbol{\theta}^k} \mathrm{E}\left[\sum_{k=1}^{N}\left(Y^k - Q^k\left(\boldsymbol{s}, \boldsymbol{a}^k; \boldsymbol{\theta}^k\right)\right)^2\right];$$

15) end For;

16) end While.

3.2 多智能体演示学习

本文环境中的基因数目 N 很大, 则对应的动作维度也很大, 这使得智能体通过随机探索找到最优路径的概率很低. 单纯使用多智能体框架也无法完全避免难以探索得到最优路径的问题, 这是因为: 多智能体框架可以使得学习任务的复杂度下降, 但动作的维数并没有下降, 因而随机探索得到最优策略的概率还是很低. 考虑到环境中包含的所有死亡状态和状态转移均已知, 本文将死亡状态视为专家意见, 采用演示学习[20] 方式加快算法的学习.

在计算专家意见对应的回报 $r^e\left(\boldsymbol{s}_t\right)$ 时, 需要考虑死亡概率, 即

$$r^e\left(\boldsymbol{s}_t\right) = \mathrm{E}\left[r\left(\boldsymbol{s}_t\right)\right] = -1 + P_d\left(\boldsymbol{s}^*\right) - \eta D\left(\boldsymbol{s}_t, \boldsymbol{s}^*\right) \tag{11}$$

其中, $\boldsymbol{s}^*$ 为目标状态, $P_d\left(\boldsymbol{s}^*\right)$ 为目标状态的死亡概率. 每个智能体的更新目标为

$$Y_t^{e,k} = r^e\left(\boldsymbol{s}_t\right) + \gamma \max_{a^k}\left(Q\left(\boldsymbol{s}_{t+1}, a^k; \boldsymbol{\theta}^k\right)\right) \tag{12}$$

如果专家意见对应的回报和环境的期望回报 $E\left[r\left(\boldsymbol{s}_t\right)\right]$ 不相符, 值估计将不收敛, 这时专家系统给出的动作 $\boldsymbol{a}^*$ 即为最优动作. 为了更好地实现演示学习, 单独设计一个经验池 B_2 来保存演示经验. 将随机探索得到的经验池 B_1 和演示经验池 B_2 的经验按照 P_s 的概率进行采样, 即用于网络训练的 Batch 有 P_s 的概率从 B_1 采样, $1-P_s$ 的概率从 B_2 采样. 基于值的强化学习问题本质上是对值函数的拟合问题, 所以无论是专家经验还是智能体随机探索得到的非最优解经验, 都需要应用于值迭代.

3.3 基于行为克隆的多智能体 DQN (BCDQN)

启发于行为克隆[21] 思想, 在智能体随机探索的同时, 对应每一步都给出相应的专家意见, 专家意见即为最优策略, 以保证 B_1 和 B_2 在状态上同分布. 算法的每一次迭代训练都会拉近 B_1 和 B_2 之间对应动作的分布差异, 当算法收敛时, B_1 和 B_2 将完全同分布, 从而实现了智能体探索策略对专家策略的完全克隆. BCDQN 的优势是算法会收敛到与专家策略完全相同的策略上.

令 L^o 和 L^e 分别为智能体探索系统和专家演示系统的损失函数, 则有

$$L^o\left(\boldsymbol{\theta}^k\right) = \mathrm{E}_{\boldsymbol{s}\sim\psi, \boldsymbol{a}\sim\varphi}\left[\sum_{k=1}^{N}\left(Y^k - Q^k\left(\boldsymbol{s}, a^k; \boldsymbol{\theta}^k\right)\right)^2\right] \tag{13}$$

$$L^e\left(\boldsymbol{\theta}^k\right) = \mathrm{E}_{\boldsymbol{s}\sim\psi, \boldsymbol{a}\sim\varphi', \varphi'\sim\pi^*(\psi)}\left[\sum_{k=1}^{N}\left(Y^{e,k} - Q^k\left(\boldsymbol{s}, a^k; \boldsymbol{\theta}^k\right)\right)^2\right] \tag{14}$$

其中, ψ 和 φ 分别为探索路径下的状态空间和动作空间. 最终 BCDQN 的损失函数为

$$L\left(\boldsymbol{\theta}^k\right) = P_s L^o\left(\boldsymbol{\theta}^k\right) + \left(1-P_s\right) L^e\left(\boldsymbol{\theta}^k\right) \tag{15}$$

综上所述, BCDQN 的伪代码如下:

算法 2. BCDQN 算法

输入: 最大迭代次数 $I_{\max}$, 折扣因子 γ, 学习率 η, 智能体个数 N, 采样概率 P_s, 初始化探索经验池 B_1 和演示经验池 B_2.

输出: 网络参数 $\boldsymbol{\theta}^k\ (k=1, 2, \cdots, N)$.

1) 初始化网络参数 $\boldsymbol{\theta}^k\ (k=1, 2, \cdots, N)$;

2) While $I < I_{\max}$;

3) $t=0$;

4) 随机初始化状态 $\boldsymbol{s}_t$;

5) While $t \le t_{\max}$ or 患者死亡;

6) For $k=1:N$;

7) 计算动作: $a_t^k = \arg\max\limits_{a^k} Q^k\left(\boldsymbol{s}_t, a^k; \boldsymbol{\theta}^k\right)$;

8) 计算专家动作 a_t^{*k};

9) end For;

10) 环境中应用动作 $\boldsymbol{a}_t = [a_t^1, a_t^2, \cdots, a_t^N]$, 并返回回报

$r(\boldsymbol{s}_t)$ 和下一时刻状态 $\boldsymbol{s}_{t+1}$, 存入 B_1;

11) 环境中应用动作 $\boldsymbol{a}_t^* = [a_t^{*1}, a_t^{*2}, \cdots, a_t^{*N}]$, 并返回回报 $r^e(\boldsymbol{s}_t)$ 和下一时刻状态 $\boldsymbol{s}_{t+1}$, 存入 B_2;

12) $t \leftarrow t+1$;

13) end While;

14) $I \leftarrow I+1$;

15) For $k = 1 : N$;

16) 随机采样并更新 $\boldsymbol{\theta}^k$:

$\boldsymbol{\theta}^k \leftarrow \boldsymbol{\theta}^k + \eta \nabla_{\boldsymbol{\theta}_k} \left(P_s L^o\left(\boldsymbol{\theta}^k\right) + (1-P_s) L^e\left(\boldsymbol{\theta}^k\right)\right)$;

17) end For;

18) end While.

3.4　基于预训练记忆的多智能体 DQN (PMDQN)

随着 N 的增大, BCDQN 中 B_1 和 B_2 状态上同分布反而会使得智能体难以找到最优路径. N 越大, 智能体的随机探索得到最优路径的概率就越低, 经验池里经验向量来自同一条路径的概率就越高, 这间接增加了训练样本间的相关性. 而深度强化学习要求训练样本间要尽可能独立, 所以提出基于预训练记忆的多智能体 DQN (PMDQN) 先使智能体在环境中进行预训练, 并将数量 T 的专家经验保存在 B_2 中, 然后不再对 B_2 进行更新. 随后使智能体进行随机探索填充 B_1, 并继续智能体的训练. 由于最终算法收敛时, B_1 和 B_2 不一定会完全同分布, 因此, 智能体不能保证学习到最优策略. 但 PMDQN 可以使专家经验池提供的样本间的相关性下降, 并加快了算法的学习速度.

这时, 智能体探索系统和专家演示系统的损失函数分别为 L^o 和 L^e, 则有

$$L^o\left(\boldsymbol{\theta}^k\right) = \mathrm{E}_{\boldsymbol{s}\sim\psi,\, \boldsymbol{a}\sim\varphi}\left[\sum_{k=1}^{N}\left(Y^k - Q^k\left(\boldsymbol{s},\, a^k; \boldsymbol{\theta}^k\right)\right)^2\right] \tag{16}$$

$$L^e\left(\boldsymbol{\theta}^k\right) = \mathrm{E}_{(\boldsymbol{s},\,\boldsymbol{a})\sim B_2}\left[\sum_{k=1}^{N}\left(Y^{e,\,k} - Q^k\left(\boldsymbol{s},\, a^k; \boldsymbol{\theta}^k\right)\right)^2\right] \tag{17}$$

最终 PMDQN 的损失函数为

$$L\left(\boldsymbol{\theta}^k\right) = P_s L^o\left(\boldsymbol{\theta}^k\right) + (1-P_s) L^e\left(\boldsymbol{\theta}^k\right) \tag{18}$$

PMDQN 的伪代码如下:

算法 3. PMDQN 算法

输入: 最大迭代次数 $I_{\max}$, 折扣因子 γ, 学习率 η, 智能体个数 N, 采样概率 P_s, 专家经验数量 T, 初始化探索经验池 B_1 和演示经验池 B_2.

输出: 网络参数 $\boldsymbol{\theta}^k$ $(k = 1, 2, \cdots, N)$.

1) While $I < T$;

2) 随机生成状态 $\boldsymbol{s}_t$, 并计算专家动作 a_t^{*k};

3) 环境中应用动作 $\boldsymbol{a}_t^* = [a_t^{*1}, a_t^{*2}, \cdots, a_t^{*N}]$, 并返回回报 $r^e(\boldsymbol{s}_t)$ 和下一时刻状态 $\boldsymbol{s}_{t+1}$, 存入 B_2;

4) 初始化网络参数 $\boldsymbol{\theta}^k$ $(k = 1, 2, \cdots, N)$;

5) While $I < I_{\max}$;

6) $t = 0$;

7) 随机初始化状态 $\boldsymbol{s}_t$;

8) While $t \le t_{\max}$ or 患者死亡;

9) For $k = 1 : N$;

10) 计算动作: $a_t^k = \arg\max\limits_{a^k} Q^k\left(\boldsymbol{s}_t,\, a^k; \boldsymbol{\theta}^k\right)$;

11) end For;

12) 环境中应用动作 $\boldsymbol{a}_t = [a_t^1, a_t^2, \cdots, a_t^N]$, 并返回回报 $r(\boldsymbol{s}_t)$ 和下一时刻状态 $\boldsymbol{s}_{t+1}$, 存入 B_1;

13) $t \leftarrow t+1$;

14) end While;

15) $I \leftarrow I+1$;

16) For $k = 1 : N$;

17) 随机采样并更新 $\boldsymbol{\theta}^k$;

$\boldsymbol{\theta}^k \leftarrow \boldsymbol{\theta}^k + \eta \nabla_{\boldsymbol{\theta}_k} \left(P_s L^o\left(\boldsymbol{\theta}^k\right) + (1-P_s) L^e\left(\boldsymbol{\theta}^k\right)\right)$;

18) end For;

19) end While.

3.5　基于多智能体 DQN 的乳腺癌致病基因排序

通过比较每个基因突变状态 s^k 的值 $F\left(s^k\right)$ 进行乳腺癌致病基因排序. $F\left(s^k\right)$ 可表示为

$$F\left(s^k\right) = \mathrm{E}\left[Q\left(\boldsymbol{s}\,|_{s^k=0},\, a^k = 1; \boldsymbol{\theta}^k\right)\right] + \mathrm{E}\left[Q\left(\boldsymbol{s}\,|_{s^k=1},\, a^k = 0; \boldsymbol{\theta}^k\right)\right] \tag{19}$$

式中, 由于第 k 个智能体从未突变状态 ($s^k = 0$) 到最终突变状态 ($s^k = 1$) 采取的动作为 $a^k = 1$; 从突变状态 ($s^k = 1$) 到最终突变状态 ($s^k = 1$) 采取的动作为 $a^k = 0$, 所以 $F\left(s^k\right)$ 可以用于表示某个基因突变对患者死亡贡献度的高低. 这里默认最终状态为未突变状态 ($s^k = 0$) 时, 对乳腺癌突变基因的分析无意义.

在多智能体框架中, 每一个智能体只处理动作空间为 2、状态空间为 2^N 的强化学习问题, 并使用基于值的强化学习来进行训练, 这时输入为 N 维二进制向量, 输出为 2 维的 Q 值. 这时的多智能框架对神经网络结构的要求不高. 为了加快多智能体的训练速度, 所有 DQN 仅使用单层神经网络, 即第 k 个网络参数 $\boldsymbol{\theta}^k$ 只包含权值向量 $\boldsymbol{w}^k$ 和偏置向量 $\boldsymbol{b}^k$, 则有

$$2^{N-1}\left(\left\|\boldsymbol{w}^k\right\|_1 + \left\|\boldsymbol{b}^k\right\|_1\right) = \mathrm{E}\left[Q\left(\boldsymbol{s}\,|_{s^k=0},\, a^k = 1; \boldsymbol{\theta}^k\right)\right] + \mathrm{E}\left[Q\left(\boldsymbol{s}\,|_{s^k=1},\, a^k = 0; \boldsymbol{\theta}^k\right)\right] \tag{20}$$

由于 $\arg\max_k \left(F\left(s^k\right)\right)$ 与 $\arg\max_k \left(\left\|\boldsymbol{w}^k\right\|_1 + \left\|\boldsymbol{b}^k\right\|_1\right)$ 相等, 所以最终使用下式进行致病基因排序

$$F\left(s^k\right) = \left\|\boldsymbol{w}^k\right\|_1 + \left\|\boldsymbol{b}^k\right\|_1 \tag{21}$$

深度强化学习方法主要通过评估状态–动作值的高低来决定动作: 如果某个基因在式 (21) 中的值越大, 说明智能体在任意状态下发生突变的状态–动作值越大, 即该基因发生突变导致病人死亡的概率越高. 因此, 通过式 (21) 指标可以排序出基因突变与患者死亡之间的关联性. 最后, 根据需求选择排序靠前的 n 个基因作为致病基因.

4 实验结果与分析

4.1 实验设置

本文通过在乳腺癌基因突变数据构建的环境来预测乳腺癌的致病基因. 乳腺癌突变数据和生存数据由 TCGA 数据官网下载得到 (网址: https://portal.gdc.cancer.gov). 深度强化学习的训练时间与环境的状态–动作空间复杂度正相关. 一般环境的状态–动作空间复杂度越高, 需要的神经网络越复杂, 训练时间越长. 受限于实验设备的计算效率, 实验中需要通过一定的规则来限制状态、动作的维度, 因此通过基因突变率来筛选基因数目.

根据乳腺癌突变数据中的基因突变率将实验设置为 2 组: 第 1 组选择基因突变率 $\geq 50\%$ 的基因, 得到 $N = 188$ 个基因, 其中包含 53 种不同的死亡状态; 第 2 组选择基因突变率 $\geq 30\%$ 的基因, 得到 $N = 420$ 个基因, 其中包含 81 种不同的死亡状态. 由于 BCDQN 比 PMDQN 更稳定, 所以 $N = 188$ 时使用 BCDQN 进行训练. 当 $N = 420$ 时, BCDQN 需耗费大量时间进行训练, 为了使算法快速收敛, 使用 PMDQN 进行训练.

本文将基因突变视为多智能体的动作, 若基因突变率太低, 则基因/智能体数目增多, 而死亡状态中突变基因的占比急剧减小, 多智能体很难通过动作学习到死亡状态, 所以选择使用 30%、50% 的基因突变率来确保构建环境所用的基因数满足智能体对乳腺癌死亡状态的学习. 当然, 也可以选择其他突变率的基因数目, 例如突变率 ≥ 40 %, 理论上在合理的基因突变率范围内, 本文提出的算法都能够适用. 不同基因突变率数据集的选择会对实验结果产生影响, 这体现在两个方面: 1) 突变率越低得到的基因数目越大, 状态–动作空间维度也越大, 导致模型收敛速度变慢, 无法学习到最优策略; 突变率越高, 则得到的基因越少, 使得强化学习任务更简单, 且过高突变率的基因使乳腺癌致病基因预测任务无意义. 2) 突变率改变将会产生不同的患者死亡率, 影响智能体完成任务情况. 因此, 在实验设备的允许的情况下, 建议基因突变率的选择范围为 10% ~ 50%.

4.2 实验结果

当 $N = 188$ 时, 使用 BCDQN 进行训练. 多智能体在 53 个死亡状态上的回报值如图 3 所示, 其中, 横坐标表示 episode, 纵坐标表示回报值. 由图 3 可以看出, 所有的策略处于收敛状态, 在每个死亡状态上, 多智能体在每个 episode 都可以取得稳定的回报. 由于策略收敛, BCDQN 可以完成所有学习任务, 具备较好的鲁棒性. 图 4 表示当 $N = 188$ 时, 多智能体完成任务情况 (达到死亡状态), 其中, 横坐标表示 episode, 纵坐标表示完成任务的次数. 图 4 中除 0、1、6、7 四个死亡状态外, 智能体能够稳定学习到死亡状态的最优策略. 智能体在 0、1、6、7 四个死亡状态产生波动是由于这几个死亡状态的死亡率较低 (死亡率分别为 4.60%、9.7%、7.69% 和 9.09%), 使得智能体在上限步数内虽然停留在死亡状态却无法触发死亡事件, 导致智能体无法完全保证稳定学习到最优策略. BCDQN 在状态–动作空间维度较小环境中可以确保找到最优策略. 而在较复杂的状态–动作空间维度中, 若存在充足的专家经验, 则算法一定可以收敛至最优策略, 但需要耗费的训练时间难以估计.

当 $N = 420$ 时, 使用 PMDQN 进行训练. 多智能体在 81 个死亡状态上的回报值如图 5 所示. 除 61、62、67、69、71 五个死亡状态外, 多智能体可在其余所有死亡状态上学习到最高的回报值. 图 6 是当 $N = 420$ 时, 多智能体完成任务情况. 除 61、62、67、69、71 五个死亡状态外, 智能体能够学习到死亡状态的最优策略. 产生这种结果的原因是由于智能体增多导致动作–状态空间复杂度增大, 智能体训练时间不够长, 暂时没有学习到最优策略. PMDQN 虽然保证了采样效率, 提供了大量有效的专家经验, 加快了算法的训练, 却不可避免地会因为环境的太过复杂而遇到专家经验不足的问题. 此时通过专家经验的扩充可在一定程度上的减少这种陷入局部最优现象的发生. 当 $N = 420$ 时, 状态–空间维度较大且复杂, 多智能体在一个情节内经历的轨迹较长, 这也会导致智能体无法探索到上述五个死亡状态. 因此, 也可以尝试利用增强探索的强化学习方法解决此问题.

根据上述结果, 总结 BCDQN 和 PMDQN 的特点和适用情况如下: BCDQN 在状态–动作空间维度较小时, 能够保证智能体探索到与专家策略相

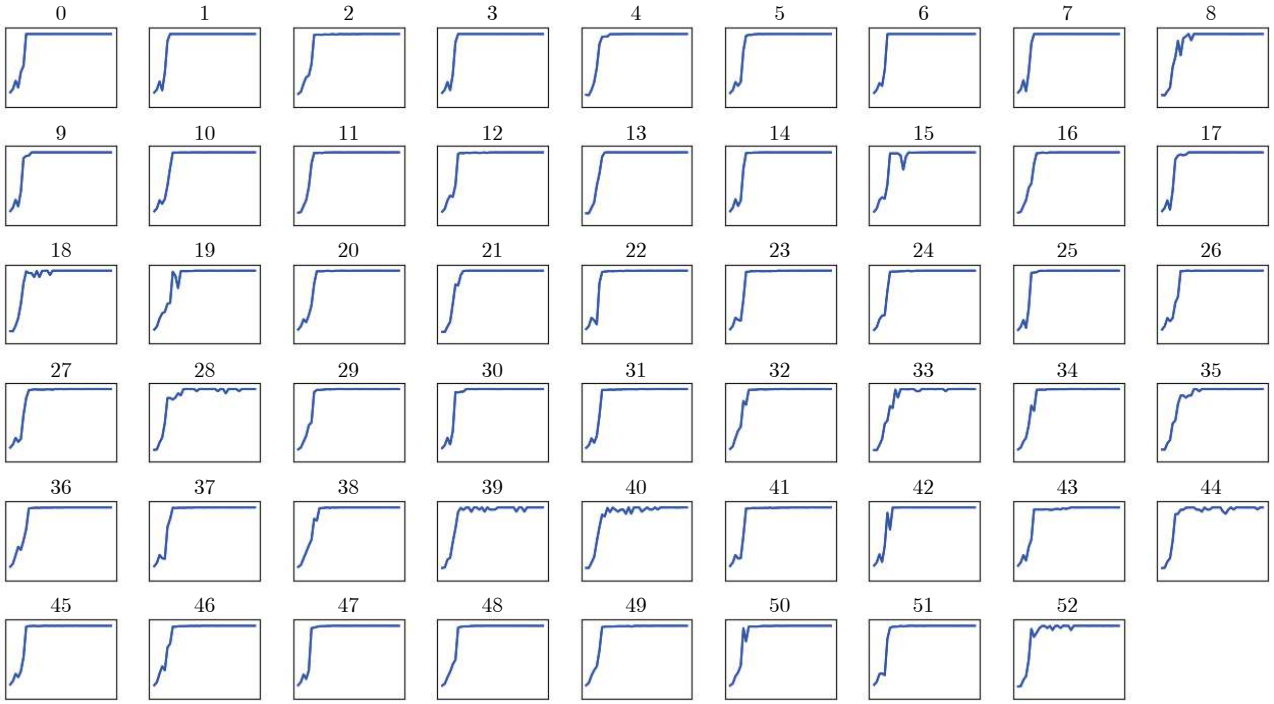


图 3 当 $N = 188$ 时, BCDQN 在 53 个死亡状态上的回报值

Fig. 3 The rewards of BCDQN at 53 death states under the condition of $N = 188$

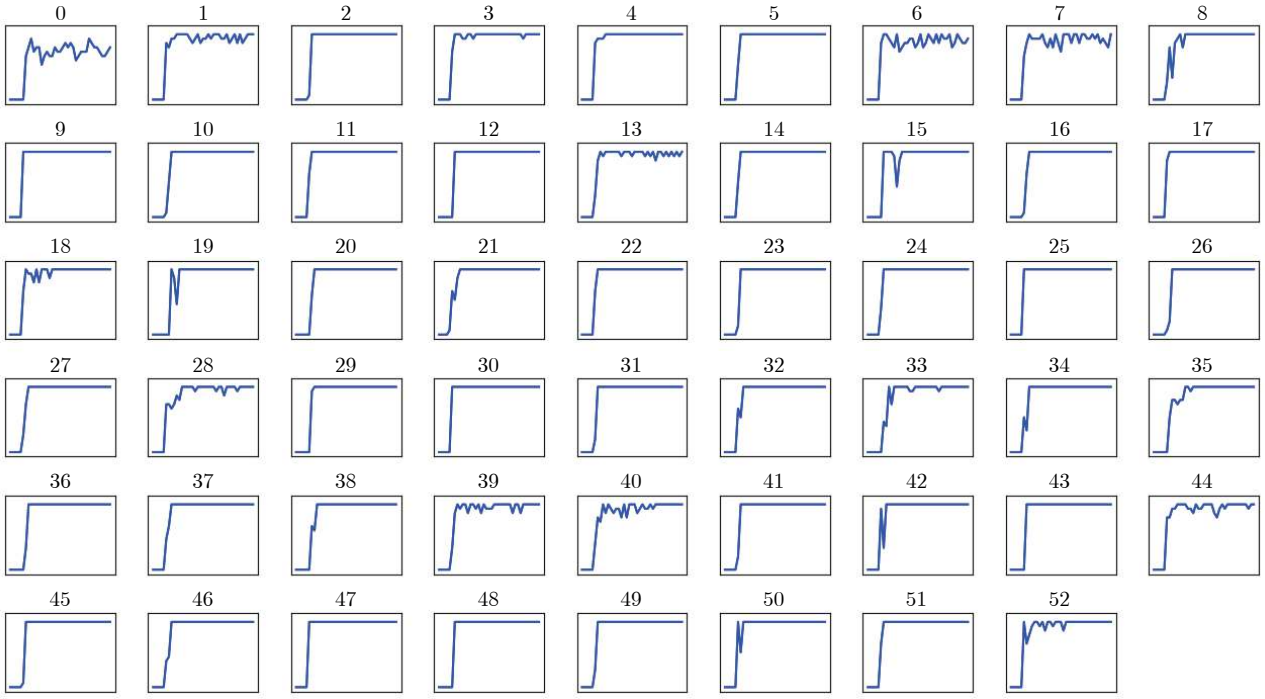


图 4 当 $N = 188$ 时, BCDQN 在 53 个死亡状态上的完成任务情况

Fig. 4 The task completion status of BCDQN at 53 death states under the condition of $N = 188$

同的策略, 稳定找到最优策略; 在状态–动作空间维度大且复杂时, PMDQN 可以减小样本间的相关性, 满足更多智能体快速进行强化学习, 但不能保证智能体学习到最优策略. 综上所述, 在实验设备允许情况下, 建议在 $N < 420$ 时使用 BCDQN, 在 $N \geq 420$ 时使用 PMDQN.

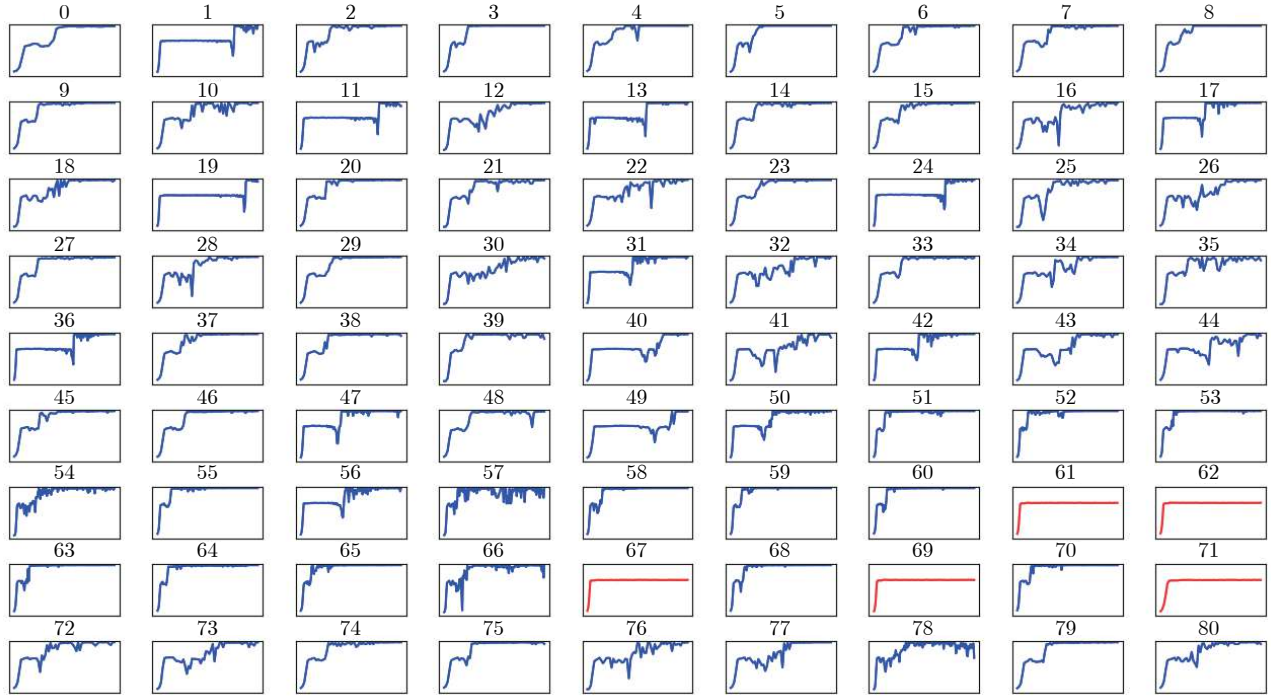


图 5 当 $N = 420$ 时, PMDQN 在 81 个死亡状态上的回报值

Fig. 5 The rewards of PMDQN at 81 death states under the condition of $N = 420$

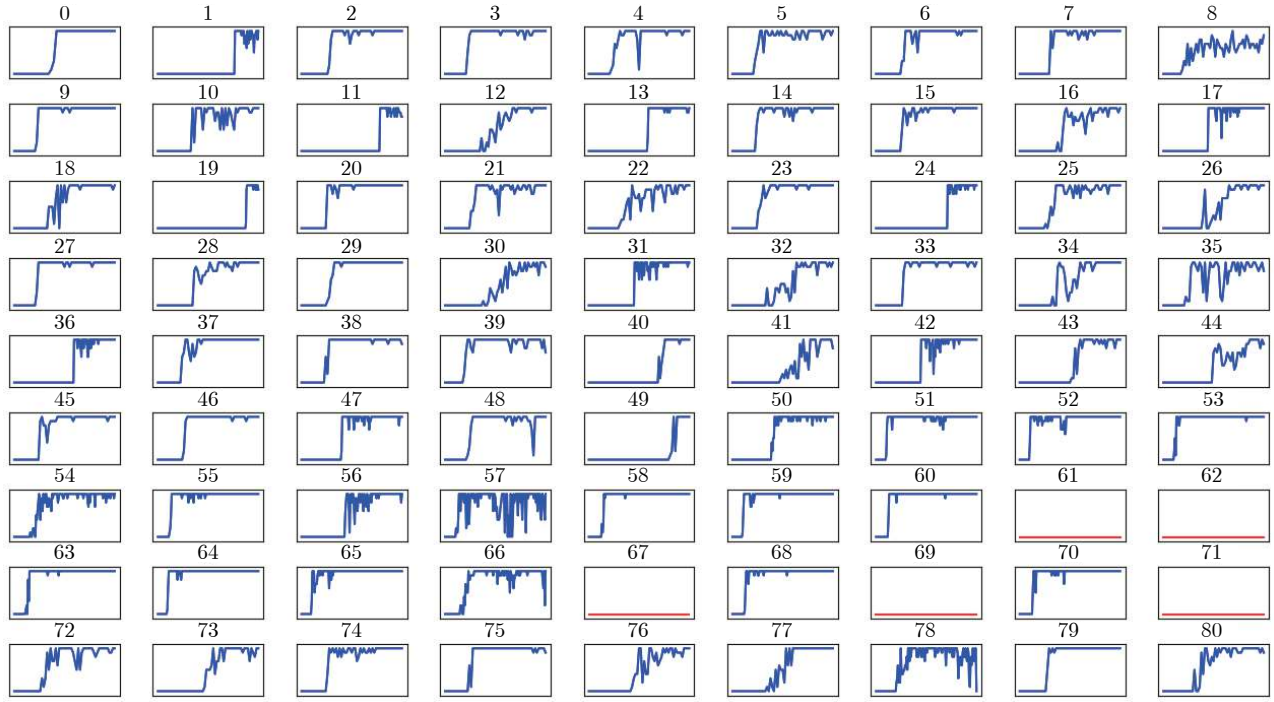


图 6 当 $N = 420$ 时, PMDQN 在 81 个死亡状态上的完成任务情况

Fig. 6 The task completion status of PMDQN at 81 death states under the condition of $N = 420$

4.3 致病基因分析

当$N = 188$和 $N = 420$时, BCDQN 和 PMDQN 预测的前 10 个致病基因如表 1 所示. 在这两种情况下, 预测的致病基因有重叠部分, 例如 TP53、MYC 和 PVT1.

肿瘤抑制基因 TP53 在控制细胞增殖、细胞存

表 1　BCDQN 和 PMDQN 预测的前 10 个致病基因
Table 1　Top 10 pathogenic genes predicted by BCDQN and PMDQN

序号	BCDQN	PMDQN
1	TP53	TP53
2	FAM91A1	PIK3CA
3	TNFRSF11B	TG
4	KCNQ3	HHLA1
5	MYC	ASAP1
6	COL14A1	CASC8
7	CCDC26	SNORA12
8	CCN3	MYC
9	PVT1	PVT1
10	DSCC1	RN7SL329

活和基因组完整性的许多细胞通路中发挥着关键作用. 当细胞经历应激条件 (如 DNA 损伤、缺氧或致癌基因激活) 时, TP53 作为细胞增殖的制动器, 几乎在所有类型的癌症中发生突变. Silwal-Pandit 等[22] 分析了 1 420 名乳腺癌患者体细胞的 TP53 突变, 研究结果表明 TP53 突变谱在乳腺癌中具有亚型特异性和明显的预后相关性. Funda 等[23] 对 257 例转移性乳腺癌患者的 202 个基因进行了高通量测序, 研究表明 TP53 在乳腺癌的三种亚型中都存在显著突变, 且与无复发生存期、无进展生存期和总生存期相关. Han 等[24] 分析了 187 例转移性乳腺癌患者的血液样本, 研究表明 TP53 突变转移性乳腺癌患者的预后明显低于 TP53 野生型患者, 特别是激素受体阳性/表皮生长因子受体 2 阴性和三阴性队列患者. 在 TP53 突变的患者中, DNA 结合域中非错义突变的乳腺癌患者的相关生存率更低.

MYC 是细胞生长、增殖、代谢、分化和凋亡的关键调控因子, 它的扩增或过表达常见于多种恶性肿瘤. 乳腺癌中 MYC 的解除涉及多种机制, 包括基因扩增、转录调节、mRNA 和蛋白质稳定, 这与肿瘤抑制子的缺失和致癌途径的激活相关. Xu 等[25] 报道了肿瘤抑制因子 BRCA1 能够抑制 MYC 的转录和转化活性, 并且 BRCA1 缺失和 MYC 过表达导致乳腺癌的发生, 特别是基底细胞样亚型的乳腺癌. Terunuma 等[26] 发现乳腺癌中 2-羟戊二酸水平升高与 MYC 通路激活之间存在关联, 并在人类乳腺上皮细胞和乳腺癌细胞中 MYC 的过表达和敲低进一步证实了这一关系. Camarda 等[27] 通过靶向代谢组学方法, 发现脂肪酸氧化中间体在 MYC 驱动的三阴性乳腺癌模型中显著上调.

PVT1 在多种恶性肿瘤中高表达, 是潜在的癌基因, 它还可与 MYC 基因相互作用, 通过多种途径参与恶性肿瘤细胞的增殖、凋亡等调控. Cho 等[28] 证明了 PVT1 启动子具有独立于 PVT1 lncRNA 的肿瘤抑制功能, 且 PVT1 启动子 CRISPR 增强了乳腺癌细胞在体内的竞争和生长. Tang 等[29] 报道了 PVT1 在临床三阴性乳腺癌中上调, 并促进 KLF5/beta-catenin 信号通路以驱动三阴性乳腺癌的发生. Wang 等[30] 的研究表明, PVT1 的表达增加与乳腺癌患者的临床分期、淋巴结转移和总生存率有关.

为进一步验证预测得到的致病基因与乳腺癌密切相关, 首先利用 ToppGene 工具 (网址: https://toppgene.cchmc.org/) 进行基因富集分析. 基因富集分析是指将一组基因按照基因组注释信息进行分类的过程, 能够发现基因间是否具有某方面的共性. 基因组注释信息存储于基因注释数据库 (Gene annotation database), 能够帮助理解基因功能, 发现基因与疾病之间的关联等. 本文采用的基因注释数据库是基因本体数据库 (Gene ontology, GO), 其涵盖多种语义分类, 如分子功能、生物学过程、细胞组分等. GO 术语 (GO term) 是 GO 数据库中的基本描述单元, 可描述基因产物的功能, 例如: GO 术语: regulation of DNA biosynthetic process 描述的是一组基因在生物过程中对 DNA 生物合成过程起调节作用.

在富集分析圈图 (图 7 ~ 8) 中, 圆形的左半圆部分表示基因, 右半边表示 GO 术语, 基因与 GO 术语之间有连线表示基因产物与 GO 术语相关, 一个基因与越多 GO 术语相连, 则表示该基因的产物功能越多. 图 7 是在 $N = 188$ 时, 前 10 个致病基因的富集分析圈图, 其中基因 CCDC26 无法与其他基因得到富集结果. 图 7 中的 GO 术语是从富集结果的众多 GO 术语中与乳腺功能密切相关的 15 个 GO 术语, 基因 MYC 与最多数目的 GO 术语相连, 且与多个乳腺癌相关的 GO 术语有关, 表示 MYC 与乳腺癌的发生、发展最为密切, 其次是基因 TP53, 以此类推. 由此可见, 图 7 中的 9 个基因的产物都与乳腺癌的发病过程相关. 虽然 CCDC26 无法与其他基因得到富集结果, 但在文献 [31] 中, CCDC26 作为下调基因, 可在多种癌症的发生过程产生作用, 例如白血病、胶质瘤等.

图 8 是在 $N = 420$ 时, 前 10 个致病基因的富集分析圈图, 本文从富集结果的众多 GO 术语中选择了与乳腺功能密切相关的 18 个 GO 术语. 基因 TP53、MYC、PIK3CA、PVT1 和 TG 与这 18 个 GO 术语相关, 表明与乳腺癌有关联. 虽然基因 HHLA1、ASAP1 与上述 18 个 GO 术语无关, 但与基因 MYC、PVT1、TG 一起与 GO 术语: Human

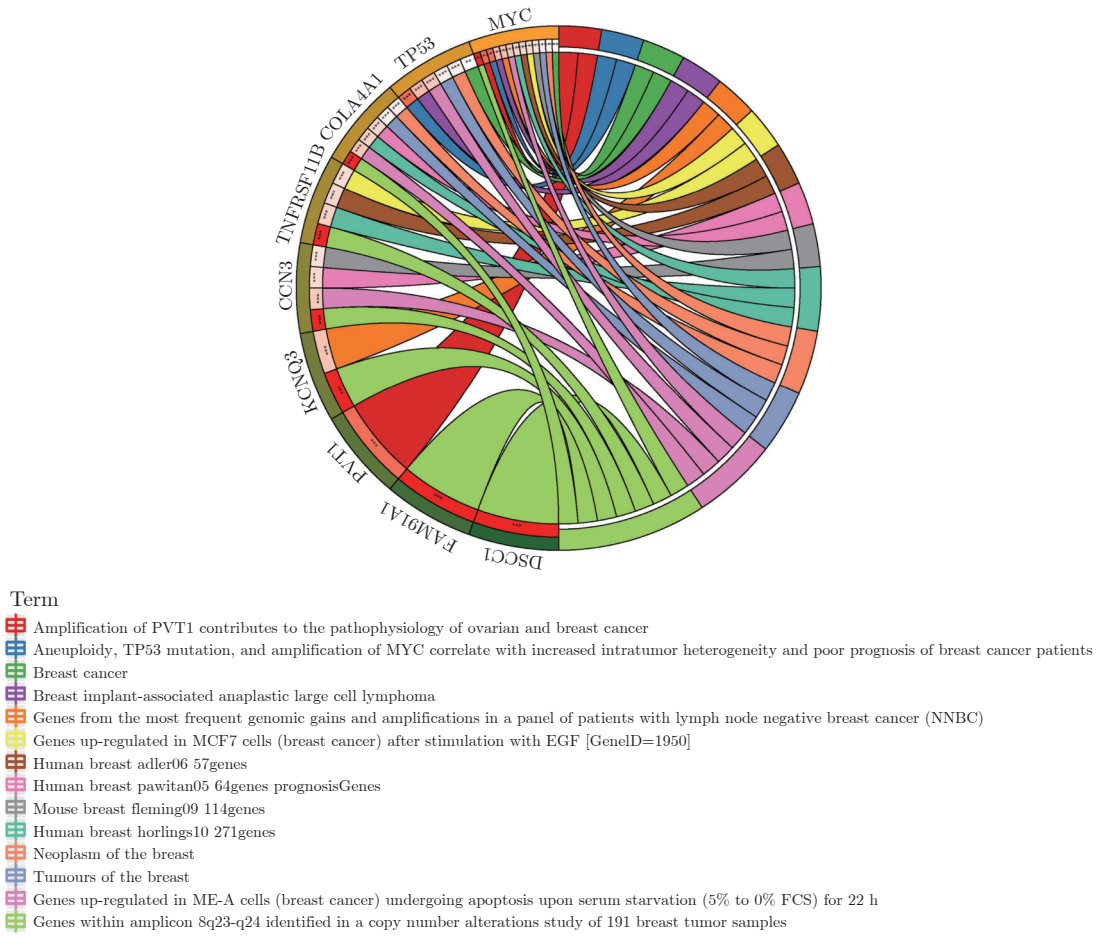


图 7 当 $N = 188$ 时, BCDQN 预测的前 10 个致病基因的富集分析圈图

Fig. 7 The enrichment analysis circle diagram of the top 10 pathogenic genes predicted by BCDQN under the condition of $N = 188$

Leukemia Schoch05 1052genes 相关, 即与白血病相关. 基因 SNORA12 在文献 [32] 中被验证为宫颈癌的 8 个过表达基因之一. 通过 RNA 测序结果, 基因 RN7SL329P 是前列腺癌中前 10 位差异表达的 IncRNAs[33].

值得注意的是, 生命科学是一门实验科学, 由人类在长期的科学探究中不断积累知识逐步完善. 本文预测的部分致病基因现阶段虽与乳腺癌无直接关联, 但都参与了其他癌症的发生过程, 可作为乳腺癌的候选致病基因以待临床验证. 导致乳腺癌风险增加最常见的突变基因 BRCA1、BRCA2 和 PALB2 没有出现在本实验中, 这是由于这些基因的突变率没有达到实验设置要求, 即在 $N = 188$ 和 $N = 420$ 的实验中不包含这些基因. 受篇幅限制, 这里仅提供两种方法预测的前 10 个基因, 排名靠后的基因不再进行分析, 但是, 这并不代表这些基因与乳腺癌无关, 例如, $N = 420$ 的实验结果中, 基因 PIK3CA 排在第 2 位, 但在 $N = 188$ 的实验结果中, 其排在第 23 位.

5 结束语

本文基于乳腺癌突变数据, 构建多智能体强化学习环境, 并根据突变数据特性设计了两种基于演示学习的多智能体 DQN. 借鉴行为克隆思想提出 BCDQN, 将患者死亡状态作为专家信息, 对智能体的每一步探索都给予指导, 最终实现探索经验池与专家经验池完全同分布. 为了满足更多智能体快速进行强化学习, 并减小样本间的相关性, 提出 PM-DQN 通过预训练方式将一定数量的专家经验保存在专家经验池中, 然后令智能体进行随机探索, 加快智能体探索到与专家策略相同的策略. 最后, 通过基因富集分析对预测得到的致病基因进行分析, 实验结果表明, 本文方法能够挖掘出乳腺癌致病基因. 同时, 该算法也挖掘出一些与其他癌症的发生过程相关的基因, 可作为乳腺癌的候选致病基因.

未来的研究工作包括设计癌症连续数据的强化学习环境, 进一步提出适用于连续数据的多智能体强化学习算法.

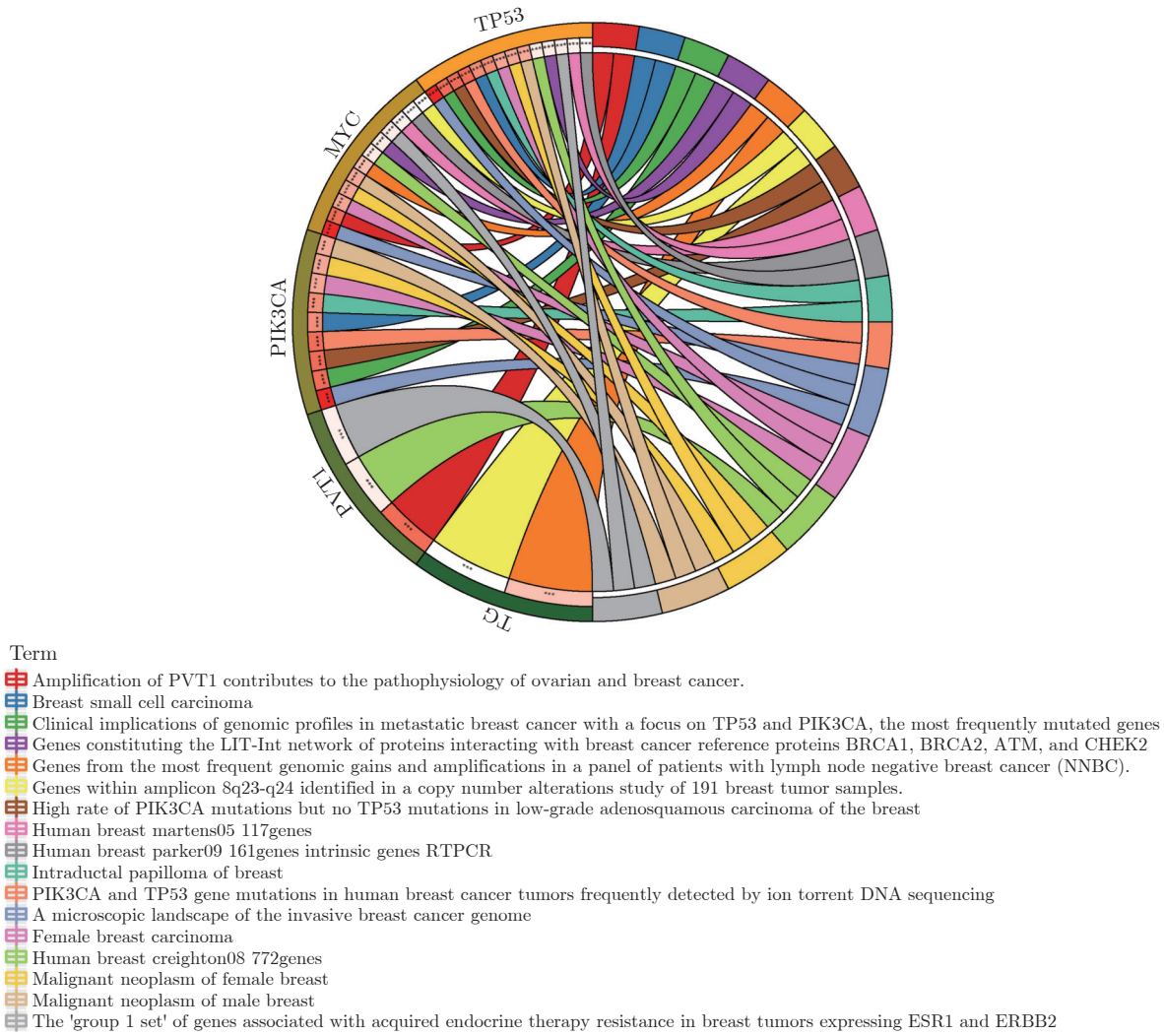


图 8　当 $N = 420$ 时, PMDQN 预测的前 10 个致病基因的富集分析圈图

Fig. 8　The enrichment analysis circle diagram of the top 10 pathogenic genes predicted by PMDQN under the condition of $N = 420$

References

1　Wisesty U N, Mengko T R, Purwarianti A. Gene mutation detection for breast cancer disease: A review. *IOP Conference Series: Materials Science and Engineering*, 2020, **830**(3): 032051

2　Kohler S, Bauer S, Horn D, Robinson P N. Walking the interactome for prioritization of candidate pathogenic genes. *The American Journal of Human Genetics*, 2008, **82**(4): 949−958

3　Xu B, Liu Y, Yu S, Wang L, Dong J, Lin H, et al. A network embedding model for pathogenic genes prediction by multi-path random walking on heterogeneous network. *BMC Medical Genomics*, 2019, **12**: 188

4　Han P, Yang P, Zhao P, Shang s, Liu Y, Zhou J, et al. GCN-MF: Disease-gene association identification by graph convolutional networks and matrix factorization. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, AK, USA: ACM, 2019. 705−713

5　Nagarajan N, Dhillon I S. Inductive matrix completion for predicting gene−disease associations. *Bioinformatics*, 2014, **30**(12): 60−68

6　Sahu B, Mishra D. A novel feature selection algorithm using particle swarm optimization for cancer microarray data. *Procedia Engineering*, 2012, **38**: 27−31

7　Malar B, Nadarajan R, Thangam G J. A hybrid isotonic separation training algorithm with correlation-based isotonic feature selection for binary classification. *Knowledge and Information Systems*, 2019, **59**(3): 651−683

8　AliazKovic E, Subasi A. Breast cancer diagnosis using ga feature selection and rotation forest. *Neural Computing and Applications*, 2017, **28**(4): 753−763

9　Sangaiah I, Kumar A V A. Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: Application to breast cancer prediction. *Cluster Computing*, 2019, **22**(3): 6899−6906

10　Alomari O A, Khader A T, Al-Betar M A, Alyasseri Z A A. A hybrid filter-wrapper gene selection method for cancer classification. In: Proceedings of the 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS). Porto, Portugal: IEEE, 2018. 113−118

11　Alzubaidi A, Cosma G, Brown D, Pockley A G. Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information. In: Proceedings of 2016 International Conference on Interactive Technologies and Games (ITAG). Nottingham, UK: IEEE, 2016. 70−76

12　Hamim M, Moudden I E, Moutachaouik H, Mustapha H. Gene selection for cancer classification: A new hybrid filter-C5.0 approach for breast cancer risk prediction. *Advances in Science,*

Technology and Engineering Systems Journal, 2021, **6**(1): 871−878

13 Liu J, Su R, Zhang J, Wei L. Classification and gene selection of triple-negative breast cancer subtype embedding gene connectivity matrix in deep neural network. *Briefings in Bioinformatics*, DOI: 10.1093/bib/bbaa395

14 Zhao Q, Zhang Y. Ensemble method of feature selection and reverse construction of gene logical network based on information entropy. *International Journal of Pattern Recognition and Artificial Intelligence*, 2020, **34**(2): 2059004

15 Sutton R S, Barto A G. Reinforcement learning. *A Bradford Book*, 1998, **15**(7): 665−685

16 Pang Wen-Yan, Fan Jia-Lu, Jiang Yi, Lewis Frank Leroy. Optimal output regulation of partially linear discrete-time systems using reinforcement learning. *Acta Automatica Sinica*, DOI: 10.16383/j.aas.c190853
(庞文砚, 范家璐, 姜艺, Lewis Frank Leroy. 基于强化学习的部分线性离散时间系统的最优输出调节. 自动化学报, DOI: 10.16383/j.aas.c190853)

17 Shi Wei, Feng Yang-He, Cheng Guang-Quan, Huang Hong-Lan, Huang Jin-Cai, Liu Zhong, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(7): 1610−1623
(施伟, 冯旸赫, 程光权, 黄红蓝, 黄金才, 刘忠, 等. 基于深度强化学习的多机协同空战方法研究. 自动化学报, 2021, **47**(7): 1610−1623)

18 Wu T, Zhou P, Wang B, Li A, Tang X, Xu Z, et al. Joint traffic control and multi-channel reassignment for core backbone network in sdn-iot: A multi-agent deep reinforcement learning approach. *IEEE Transactions on Network Science and Engineering*, 2021, **8**(1): 231−245

19 Volodymyr M, Koray K, David S, Rusu A A, Joel V, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

20 Martinez D, Alenya G, Torras, C. Relational reinforcement learning with guided demonstrations. *Artificial Intelligence*, 2017, **247**: 295−312

21 Torabi F, Warnell G, Stone P. Behavioral cloning from observation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI, 2018. 4950−4957

22 Silwal-Pandit L, Vollan H K M, Chin S F, Rueda O M, McKinney S, Osako T, et al. TP53 mutation spectrum in breast cancer is subtype specific and has distinct prognostic relevance. *Clinical Cancer Research*, 2014, **20**(13): 3569−3580

23 Funda M B, Zheng X, Shariati M, Damodaran S, Wathoo C, Brusco L, et al. Survival outcomes by TP53 mutation status in metastatic breast cancer. *JCO Precision Oncology*, 2018, **2**: 1−15

24 Han B, Yu J, Jia S, Liu X, Liang X, Li H. Prognostic value of the TP53 mutation location in metastatic breast cancer as detected by next-generation sequencing. *Cancer Management and Research*, 2021, **13**: 3303−3316

25 Xu J, Chen Y, Olopade O I. MYC and breast cancer. *Genes and Cancer*, 2010, **1**(6): 629−640

26 Terunuma A, Putluri N, Mishra P, Mathe E A, Ambs S. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *Journal of Clinical Investigation*, 2014, **124**(1): 398−412

27 Camarda R, Zhou A Y, Kohnz R A, Balakrishnan S, Mahieu C, Anderton B, et al. Inhibition of fatty acid oxidation as a therapy for myc-overexpressing triple-negative breast cancer. *Nature Medicine*, 2016, **22**(4): 427−432

28 Cho S W, Xu J, Sun R, Mumbach M R, Carter A C, Chen Y G, et al. Promoter of LncRNA gene PVT1 is a tumor-suppressor DNA boundary element. *Cell*, 2018, **173**(6): 1398−1412

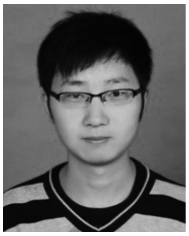
29 Tang J, Li Y, Sang Y, Yu B, Lv D, Zhang W, et al. LncRNA PVT1 regulates triple-negative breast cancer through KLF5/Beta-catenin signaling. *Oncogene*, 2018, **37**(34): 4723−4734

30 Wang Y, Zhou J, Wang Z, Wang P, Li S. Upregulation of SOX2 activated LncRNA PVT1 expression promotes breast cancer cell growth and invasion. *Biochemical and Biophysical Research Communications*, 2017, **493**(1): 429−436

31 Anindya B. Abstract IA13: It takes two to tango: The PVT1-MYC alliance in human cancer. *Cancer Research*, 2016, **76**(6): IA13

32 Roychowdhury A, Samadder S, Das P, Mazumder D I, Chatterjee A, Addya S, et al. Deregulation of H19 is associated with cervical carcinoma. *Genomics*, 2019, **112**(1): 961−970

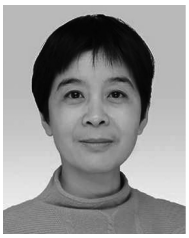
33 Li Z, Teng J, Jia Z, Zhang G, Ai X. The long non-coding RNA PCAL7 promotes prostate cancer by strengthening androgen receptor signaling. *Journal of Clinical Laboratory Analysis*, 2021, **35**(2): e23645



刘 健 中国矿业大学讲师. 2018 年获中国矿业大学博士学位. 主要研究方向为机器学习和生物信息学.
E-mail: liujiansqjxt@126.com
(**LIU Jian** Lecturer at China University of Mining and Technology. He received his Ph.D. degree from China University of Mining and Technology in 2018. His research interest covers machine learning and bioinformatics.)

顾 扬 中国矿业大学博士研究生. 2016 年获中国矿业大学学士学位. 主要研究方向为深度强化学习.
E-mail: guyang@cumt.edu.cn
(**GU Yang** Ph.D. candidate at China University of Mining and Technology. He received his bachelor degree from China University of Mining and Technology in 2016. His main research interest is deep reinforcement learning.)

程玉虎 中国矿业大学教授. 2005 年获中国科学院自动化研究所博士学位. 主要研究方向为机器学习和智能系统.
E-mail: chengyuhu@163.com
(**CHENG Yu-Hu** Professor at China University of Mining and Technology. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2005. His research interest covers machine learning and intelligent system.)



王雪松 中国矿业大学教授. 2002 年获中国矿业大学博士学位. 主要研究方向为机器学习和模式识别. 本文通信作者.
E-mail: wangxuesongcumt@163.com
(**WANG Xue-Song** Professor at China University of Mining and Technology. She received her Ph.D. degree from China University of Mining and Technology in 2002. Her research interest covers machine learning and pattern recognition. Corresponding author of this paper.)