研究与开发

基于价值差异学习的多小区 mMTC 接入算法

李昕[1]，孙君[1,2]
（1. 南京邮电大学通信与信息工程学院，江苏 南京 210003;
2. 江苏省无线通信重点实验室，江苏 南京 210003）

摘　要：在 5G 大连接物联网场景下，针对大连接物联网设备（massive machine type communication device，mMTCD）的接入拥塞现象，提出了基于价值差异探索的双重深度 Q 网络（double deep Q network with value-difference based exploration，VDBE-DDQN）算法。该算法着重解决了在多小区网络环境下 mMTCD 接入基站的问题，并将该深度强化算法的状态转移过程建模为马尔可夫决策过程。该算法使用双重深度 Q 网络来拟合目标状态—动作值函数，并采用基于价值差异的探索策略，可以同时利用当前条件和预期的未来需求来应对环境变化，每个 mMTCD 根据当前值函数与网络估计的下一时刻值函数的差异来更新探索概率，而不是使用统一的标准，从而为 mMTCD 选择最佳基站。仿真结果表明，所提算法可有效提高系统的接入成功率。

关键词：大连接物联网；随机接入；强化学习；基站选择
中图分类号：TN929.5
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2022152

Value-difference learning based mMTC devices access algorithm in multi-cell network

LI Xin[1], SUN Jun[1,2]
1. College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Jiangsu Key Laboratory of Wireless Communications, Nanjing 210003, China

Abstract: In the massive machine type communication scenario of 5G, the access congestion problem of massive machine type communication devices (mMTCD) in multi-cell network is very important. A double deep Q network with value-difference based exploration (VDBE-DDQN) algorithm was proposed. The algorithm focused on the solution that could reduce the collision when a number of mMTCDs accessed to eNB in multi-cell network. The state transition process of the deep reinforcement learning algorithm was modeled as Markov decision process. Furthermore, the algorithm used a double deep Q network to fit the target state-action value function, and it employed an exploration strategy based on value-difference to adapt the change of the environment, which could take advantage of both current conditions and expected future needs. Moreover, each mMTCD updated the probability of exploration

收稿日期：2022−01−13；**修回日期：**2022−04−06
基金项目：国家自然科学基金资助项目（No.61771255）；省部级重点实验室开放课题项目（No.20190904）
Foundation Items: The National Natural Science Foundation of China (No.61771255), Provincial and Ministerial Key Laboratory Open Project (No.20190904)

according to the difference between the current value function and the next value function estimated by the network, rather than using the same standard to select the best base eNB for the mMTCD. Simulation results show that the proposed algorithm can effectively improve the access success rate of the system.

Key words: mMTC, RA, reinforcement learning, eNB selection

0　引言

5G 移动通信及未来移动网络（包括物联网）的部署正在推动先进物联网的发展[1–3]，将人与人之间通信拓展到人与物、物与物之间通信，开启万物互联时代[4]。大连接物联网（massive machine-type communication，mMTC）系统的主要挑战是在上行链路中为大量设备设计稳定高效的随机接入方案[5]。特别是，随着 5G 移动通信的发展和大规模物联网场景的出现，数以百计的大连接物联网设备（massive machine type communication device，mMTCD）被连接起来[6]。据 Statista 统计，到 2030 年，全球预计将使用 500 亿台物联网设备[7]，为未来的蜂窝网络设想一个真正的大连接物联网场景。但是，mMTCD 的快速增长为随机接入（random access，RA）带来各方面的挑战[8]。

为了解决蜂窝物联网中的 RA 拥塞问题，3GPP 提出了几种解决方案，包括访问等级限制及其变体、特定于 mMTC 的退避方法、时隙 RA、RA 资源分离和分页 RA 方法[9]。此外，文献[10]中还研究了其他方案，如优先级 RA、基于分组的 RA 和码扩展的 RA。然而，大多数现有的方法适用于集中式系统，并且是被动的，而不是具有高度动态特性的 mMTCD 所需的。因此，目前的研究倾向于强化学习（reinforcement learning，RL）辅助的接入控制方案，因为它更适应于学习系统变化和参数不确定的环境。

在 RL 中，Q-学习因其无模型和分布式特性，更适合 mMTC 场景。文献[11]提出了一种基于协同 Q-学习的拥塞避免方法，该方法利用每个时隙的拥塞水平来设置奖励函数。文献[12]中，每个 mMTCD 通过 Q-学习选择其传输的时隙和传输功率来提高系统吞吐量。文献[13]提出结合 NOMA 和 Q-学习方法，以便在前导分离域中显著分开区域中的设备，既可以重用前导码而不会发生冲突，减少连接尝试。文献[14]提出双 Q-学习算法，该算法可以动态地适应 ACB 机制的接入限制参数。双 Q-学习的实现可以降低传统 Q-学习过高估计 Q 值的风险，避免导致次优性能。文献[15]提出了基于多智能体的智能前导码选择机制，将神经网络与 RL 结合，有效提高了 mMTCD 接入的性能。同时为解决 mMTCD 的接入问题提供了思路。以上方案只关注在单小区接入过程中的拥塞问题，很少关注在基站侧的影响。即使 mMTCD 完成了随机接入过程，也会出现过载和资源分配失败。因此需要设计高效稳定的 eNB 选择方案，以适应 mMTC 的特性，减少网络拥塞和过载。文献[16]中 Q-学习用于选择最佳可用基站，使用吞吐量和时延作为 QoS 测量和奖励。文献[17]在时隙多小区随机接入的场景下，提出了一种基于指数权重探索与开发的 RL 算法，用于选择关联的接入点。但是文献[16-17]不能处理网络密度的增长，因此过载仍然是一个问题。虽然 Q-学习因其分布式和无模型特性而被广泛使用，但是在 mMTC 场景下，其并不能解决网络密度增长带来的挑战，可能会使其 Q 表过大并且增加查表难度导致函数难收敛。因此，本文引入神经网络拟合 Q 表来优化传统用于随机接入的 RL 算法，并使用双重网络提高算法精度。同时，提出一种探索策略，进一步提高算法性能，与现有的基于多小区随机接入的算法相比，本文算法降低了选择 eNB 的随机性，更能选择最优的 eNB。

因此，本文提出了基于价值差异的双重深度 Q 网络（double deep Q network with value-difference based exploration，VDBE-DDQN）算法来解决多小区网络环境下的 eNB 选择问题。通

过双重深度 Q 网络（double deep Q network，DDQN）来学习 mMTCD 到 eNB 的直接映射，同时，当网络参数未知时，引入价值差异探索（value-difference based exploration，VDBE）的研究方法，根据每个 mMTCD 自身学习的情况更新探索概率，使学习过程更符合每个 mMTCD 的需求。在学习过程中，知道的信息越多，越有可能在所知的 eNB 中选择最优 eNB，而不是随机选择。本文所提方法与其他多小区网络选择基站的随机接入方法相比，能够允许大量 mMTCD 的接入，并有效提高系统中 mMTCD 的接入成功率。

1 系统模型

系统模型如图 1 所示，系统模型由位于小区原点的 eNB 和随机分布在其周围的 N 个 mMTCD 组成。本文只考虑 mMTCD 与 eNB 之间的上行链路传输，其中 mMTCD 以信号和数据的形式向 eNB 发送接入请求，eNB 充当数据集中器，并向其覆盖区域的设备广播控制消息。考虑 K 个小区，每个小区由位于其原点的基站所覆盖。区域 1 的 mMTCD 仅由 eNB1 服务，3 个小区重叠区域的设备可以选择 3 个 eNB 中的任一个通信。如果多个集中器可用，那么每个 mMTCD 每次只能选其中一个通信。

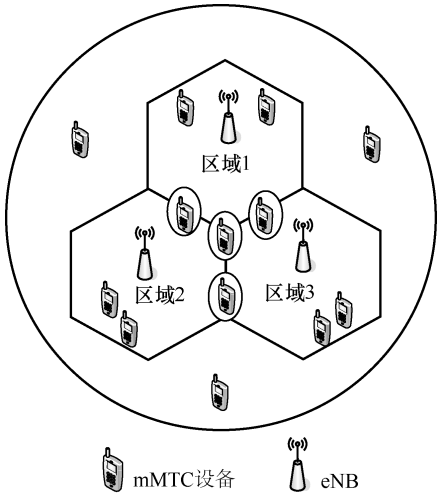

图 1 系统模型

在随机接入之前，设备需要等待前导码来传输分组，前导码被定义为向 eNB 发送数据包的传输机会[18]。如果多个设备同时选择相同的前导码，那么会发生前导码碰撞，使设备无法分配到前导码，无法向 eNB 传输数据包。所以重叠区域的 mMTCD 在发送数据包之前必须选择一个 eNB 进行传输，那么拥塞较少的 eNB 将有更大的可能将前导码分配给 mMTCD，增加成功接入的可能。除此之外，接入失败的设备可以在下一时隙重新请求接入。为了评估在多小区网络下系统的性能，将接入成功率作为性能指标：

$$p=\lim_{T\to\infty}E\left[\frac{\sum_{t=1}^{T}\left(N_{t-1}^{\mathrm{re}-s}+N_t^{s}\right)}{\sum_{t=1}^{T}N_t}\right] \tag{1}$$

其中，N_t^s 代表在时隙 t 成功连接 eNB 的 mMTCD 数量，s 代表设备成功接入，$N_{t-1}^{\mathrm{re}-s}$ 代表在时隙 $t-1$ 接入失败的设备在时隙 t 又成功接入的数量，re 代表设备重新发送，N_t 代表在时隙 t 尝试接入 eNB 的 mMTCD 总数。特别地，定义 $N_t^{\max}$ 为每次最多允许请求接入的 mMTCD 数量，减少设备碰撞。

2 算法描述

2.1 RL 算法描述

在 RL 中，智能体通过试错来学习。智能体与环境交互的学习过程可以被视为马尔可夫决策过程（Markov decision process，MDP），其被定义为一个四元组[S,A,P,R]，分别表示状态集、动作集、状态转移和奖励。RL 框架与 MDP 如图 2 所示，为 RL 中智能体与环境的交互过程以及将其状态转移描述为 MDP。

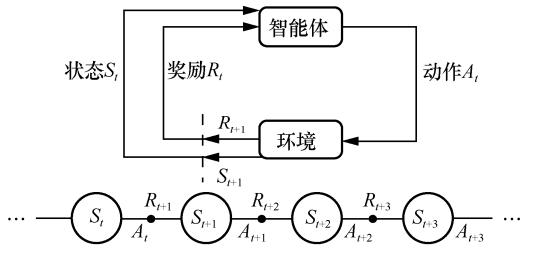

图 2 RL 框架与 MDP

在 RL 框架中，每个 mMTCD 作为一个智能体，不断地与 eNB 交互。在每个离散时隙 $t\in\{1,2,3,\cdots\}$，智能体处于特定状态 $s\in S$，基于当前状态 s 下，智能体选择一个动作 $a\in A$，得到一个奖励值并进入下一个状态 s'。在特定状态下选择动作的决策可以描述为 $\pi(s)=a$。定义 $Q(s,a)$ 为状态—动作值函数，简称 Q 值函数，即在状态 s 选择动作 a 的折扣未来回报。RL 目标是找到一个最优的策略 π^*，使 $Q(s,a)$ 最大。

$$\pi^*=\arg\max_{a\in A}Q(s,a) \tag{2}$$

Q-学习是 RL 最常见的一种，通过对智能体与其环境之间的交互进行抽样观察学习值函数。在时隙 t，先基于当前状态 s，使用 ε-贪婪法按一定概率选择动作 a，得到奖励 r，进入新状态 s'。每个状态下的 Q 值通过以下迭代过程[19]计算：

$$Q(s,a)=Q(s,a)+\alpha\left[r(s,a)+\gamma\max_{a\in A}Q(s',a')-Q(s,a)\right] \tag{3}$$

其中，符号 α 为学习率，γ 为折扣率。Q-学习会建立一个 Q 表来保存状态 s 和将会采取的所有动作 a 和 $Q(s,a)$。但当状态—动作空间较大时，Q 表会非常大，仅通过反复查表很难遍历每个状态，还会使学习速度更慢。为解决此缺点，可将神经网络与 RL 相结合，用神经网络来拟合函数。通过智能体与环境的不断交互使神经网络近似等于最优 $Q(s,a)$。

2.2　VDBE-DDQN 算法

设备每隔一段时间检测当前小区的网络状态，并根据当前网络状态决定是否切换小区来达到更好的接入效果。与算法相关的状态、动作和感知奖励定义如下。

状态 S：状态 s 为 mMTC 设备 $i\in N$，连接它所选择的 eNB。

动作 A_i：当 mMTCD 需发送数据包时，动作 a_i 定义为第 i 个 mMTCD 连接第 k 个 eNB。$a_i(s,s')=1$ 是在状态 s 下，第 i 个 mMTCD 切换到其他 eNB，$a_i(s,s')=0$ 表示保持与当前 eNB 连接。

奖励 R_i：奖励是在 mMTCD 向 eNB 发送接入请求之后获得，需考虑 mMTCD 是否在网络覆盖范围。如果不在覆盖范围，设置 φ=0，此时 mMTCD 向 eNB 发送任何请求都无效，奖励值为 0；如果在覆盖范围，设置 φ=1。此时若 eNB 接收了 mMTCD 的接入请求，则发送确认信息；若 mMTCD 发送数据失败，说明选择同一个 eNB 的设备发生了碰撞冲突。

当设备向同一 eNB 发送接入请求时，eNB 为设备分配前导码，选择相同前导码的设备会发生冲突，导致发送数据失败。mMTCD 间进行协作，通过 eNB 反馈全局信息，使得发送失败的设备知道与自己发生碰撞的设备数量。$\mathbf{CL}=[\mathrm{CL}_1,\mathrm{CL}_2,\cdots,\mathrm{CL}_N]$ 是一 $1\times N$ 的向量，表示每个发送失败的 mMTCD 的碰撞级别。根据碰撞级别，第 i 个发送失败的 mMTCD 的惩罚为：

$$C(i)=\frac{\mathrm{CL}(i)}{N_t} \tag{4}$$

其中，$\mathrm{CL}(i)$ 表示第 i 个发送失败的 mMTCD 的碰撞级别。所以奖励 $r(s,a_i)$ 被定义为：

$$r(s,a_i)=\begin{cases}+1, & \varphi=1,\text{传输成功}\\ -C(i), & \varphi=1,\text{传输失败}\\ 0, & \varphi=0\end{cases} \tag{5}$$

此外，算法将深度神经网络（deep neural network，DNN）代替 Q 表，并将其称为深度 Q 网络（deep Q network，DQN）。在 DQN 中，通过 DNN 对值函数近似，可表示为：

$$Q(s,a_i;\omega)\approx Q^*(s,a_i) \tag{6}$$

其中，ω 是 DNN 的网络参数。此外，Q 值函数是通过从一种状态映射到一种动作来逼近的。利用权重 ω 的 NN 函数逼近器 $Q(s,a_i;\omega)\approx Q^*(s,a_i)$ 作为一个在线网络。DQN 利用目标网络和在线网络来稳定整体网络性能。更新权重，使定义的损失函数最小化，损失函数定义为：

$$L_i(\omega)=E\left[\left(y_i^{\mathrm{DQN}}-Q(s,a_i;\omega)\right)^2\right] \quad (7)$$

目标值表示为：

$$y_i^{\mathrm{DQN}}=r(s,a_i)+\gamma\max_{a_i'\in A_i}Q\left(s',a_i';\omega^-\right) \quad (8)$$

其中，ω^- 表示目标网络的权重。

VDBE-DDQN 算法流程如图 3 所示。动作 a_i 通过 VDBE 策略从在线网络 $Q(s,a_i;\omega)$ 中选择。在初始化时目标网络的结构和权重相同，每经过一段时间的更新，将在线网络的权重赋值给目标网络。同时，为了克服学习不稳定性，采用了经验回放策略。将智能体与环境交互得到的 $(s,a_i,r(s,a_i),s')$ 的转换存储在经验回放池 D 中。在学习过程中，不只使用当前的经验 $(s,a_i,r(s,a_i),s')$，也通过从 D 中随机均匀采样小批量的经验来训练。经验回放通过降低训练实例之间的相关性，保证了最优策略不会被驱动到局部极小值。

此外，由于在 Q-学习和 DQN 方法中使用相同的值来选择和评估动作，Q 值函数可能过于乐

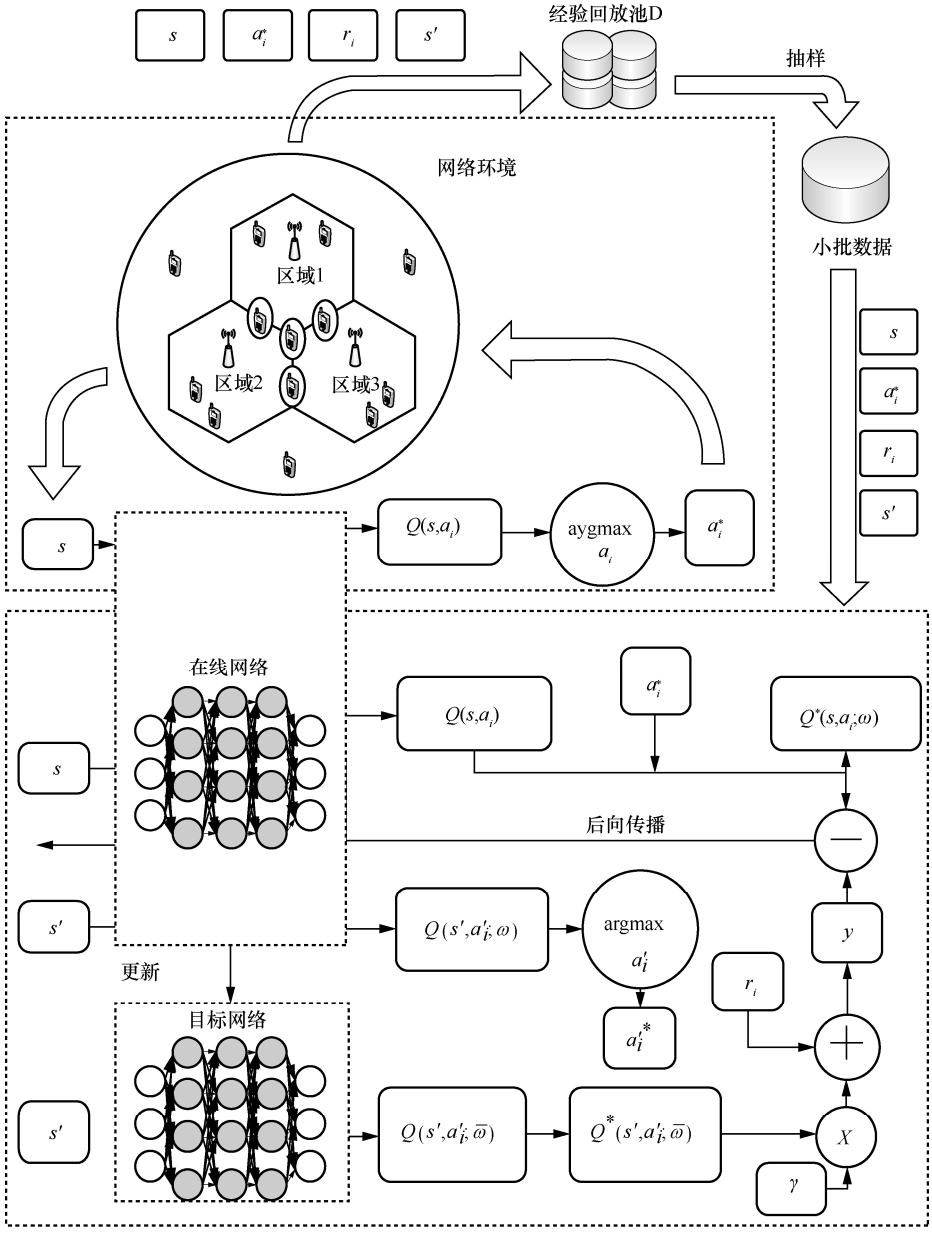

图 3　VDBE-DDQN 算法流程

观。因此，使用 DDQN[20]通过以下定义的目标值替换（8）的值来缓解上述问题：

$$y_i^{\mathrm{DDQN}} = r(s,a_i) + \gamma Q\left(s', \arg\max_{a'\in A_i} Q(s',a_i';\omega);\omega^-\right) \quad (9)$$

更具体地说，下一个状态 s' 利用在线网络和目标网络计算最优值 $Q(s',a_i';\omega)$。然后，利用折扣因子 γ 和当前奖励 $r(s,a_i)$，得到目标值 y_i^{DDQN}。最后，目标值减去在线网络预测的最优值 $Q(s,a_i;\omega)$ 计算误差，得到损失函数，然后后向传播更新权重。损失函数定义为：

$$L_i(\omega) = E\left[\left(y_i^{\mathrm{DDQN}} - Q(s,a_i;\omega)\right)^2\right] \quad (10)$$

梯度下降更新网络权重 ω，如式（11）。每执行 E^- 步之后，更新目标 DQN 参数，使 $\omega^- = \omega$。

$$\omega = \omega - \alpha \frac{\mathrm{d}\omega}{\sqrt{S\mathrm{d}\omega + \lambda}} \quad (11)$$

其中，$\mathrm{d}\omega = \dfrac{\partial L_i(\omega)}{\partial \omega}$，$S\mathrm{d}\omega$ 是损失函数在迭代过程中累积的梯度平方动量，常量 λ 防止分母为 0。

在决策阶段，动作 a_i 通过 VDBE 策略选择，这是一种结合探索与利用模式的策略。在学习开始时，mMTCD 会随机选择可能的动作，即探索模式。否则，将选择处于这一状态时可以提供最大奖励的动作，即利用模式[21]。随着大量 mMTCD 的到来，环境是高度动态的，需要一种非常灵活的算法来有效地利用探索与利用模式的优势。对此，VDBE 策略引入了依赖于状态的探索概率 $\varepsilon(s)$，解决选择问题。$\varepsilon(s)$ 不是全局参数，而是与状态相关的，这意味着每个 mMTCD 具有不同的探索概率。在环境不确定的情况下，应该有更大的概率进行探索，式（12）中的波动函数 $f(s,a_i,\tau)$ 体现了这一点。波动函数 $f(s,a_i,\tau)$ 在每步学习之后计算得到，表示为：

$$f(s,a_i,\tau) = \frac{1-\mathrm{e}^{-\frac{\sqrt{\left|Q_{t+1}^2(s,a_i)-Q_t^2(s,a_i)\right|}}{\tau}}}{1+\mathrm{e}^{-\frac{\sqrt{\left|Q_{t+1}^2(s,a_i)-Q_t^2(s,a_i)\right|}}{\tau}}} \quad (12)$$

通过当前时刻 mMTCD 选择动作获得的 Q 值与下一时刻网络预测的可能选择的最优动作的 Q 值之间的差异，mMTCD 可以根据自身的状态和环境改变学习过程中探索的概率，这样可以使 mMTCD 能够在更贴合自身“需求”的情况下进行学习。在环境不断平稳之后，mMTCD 的学习过程会趋于平稳，前后两个时刻的 Q 值会趋于相等。只要获得更多的信息和知识，就应减少探索。一般来说，获得的信息越多，表明奖励价值的差别很小或没有差别。更新探索概率 $\varepsilon(s)$ 为：

$$\varepsilon_{t+1}(s) = \delta \times f(s,a_i,\tau) + (1-\delta)\times \varepsilon(s) \quad (13)$$

其中，$\tau \in [0,1]$ 为灵敏度因子，确定所选动作对状态相关探索概率的影响。δ 表示当前状态下的动作数。在学习开始时，所有状态的探索概率初始化为 $\varepsilon(s)=1$。算法 1 给出了 mMTCD 在多小区网络环境下选择 eNB 的算法。

算法 1：基于 VDBE-DDQN 的 eNB 选择算法描述

初始化：经验回放池 D、DQN 参数 ω 和目标网络参数更换频率 E^-

主 DQN $Q(s,a_i;\omega)$，参数为 ω；目标 DQN $Q(s',a_i';\omega^-)$，参数 $\omega^- = \omega$

$Q(s,a_i) \leftarrow 0, \forall s \in S, \forall a_i \in A_i$

$\delta(s) \leftarrow 0, \forall s \in S$

for 每个回合数 EP=1,…,M

　for 每个 t=0,…,T

　　每个 mMTCD 观察状态 s

　　以概率 $\varepsilon(s_t)$ 随机选择动作 a_i

　　否则选择在 s 下，$Q(s,a_i;\omega)$ 值最大的动作 a_i

　　mMTCD 向 eNB 发送连接请求,获得 $r(s,a_i)$

　　mMTCD 观察网络状态 s'，更新 $s \leftarrow s'$

　将四元组 $(s,a_i,r(s,a_i),s')$ 存储到 D

　从 D 中随机抽取一批 $(s,a_i,r(s,a_i),s')$ 数据

　　mMTCD 根据式（9）计算 y_i^{DDQN}

mMTCD 对 $\left(y_i^{\mathrm{DDQN}}-Q(s,a_i;\omega)\right)^2$ 执行梯度下降

每执行 E^- 步之后，更新目标 DQN 参数，$\omega^-=\omega$

$\varepsilon(s_t)\leftarrow\varepsilon(s_t)+1$

根据式（13）更新

$t\leftarrow t+1$

end for

end for

3 仿真及结果分析

本文采用半径 500 m、K=3 的小区上行链路系统，可用前导码为 48 个，mMTCD 均匀分布在系统内，其数量服从大小为 500、期望为 10 的泊松分布。也就是说，系统中分布 500 个 mMTCD，每次新激活的数量为 10 个。每次允许接入的 mMTCD 的最大值 $N_t^{\max}$=30 个。同时，在给定的时间内，mMTC 数据包的到达服从 Beta 分布[9]。此外，VDBE-DDQN 是由 1 个输入层（50 个神经元）、3 个隐藏层（64、32 和 32 个神经元）和 1 个输出层（3×16 个神经元）组成的全连接网络，激活函数为 ReLU 函数。在参数更新过程中，使用 RMSProp 优化算法获得最优随机梯度下降[22]。每个回合包括 500 个训练迭代步数，小区覆盖范围内的 mMTCD 接入请求完成后一个回合结束，环境状态等重置。使用 Python 进行仿真，仿真参数的设置见表 1。

表 1 仿真参数设置

仿真参数	参数值
训练回合数	150
训练步数 T	500
批大小	8
经验回放池 D 的大小	500
网络参数更换频率 E^-	50
学习率 α	0.01
折扣率 γ	0.9

由于 DRL 依赖超参数，通过选择合理的超参数，可以实现算法的收敛。图 4～图 6 显示了不同超参数对算法性能的影响。不同学习率 α 下的性能如图 4 所示，使用不同的学习率评估本文方法。α =0.01 和 α =0.001 时都可以达到损失函数的最小值，并且 α =0.01 时网络在迭代 10 步以内就可以收敛，而 α =0.000 1 时网络收敛非常缓慢。因此选择 α =0.01。

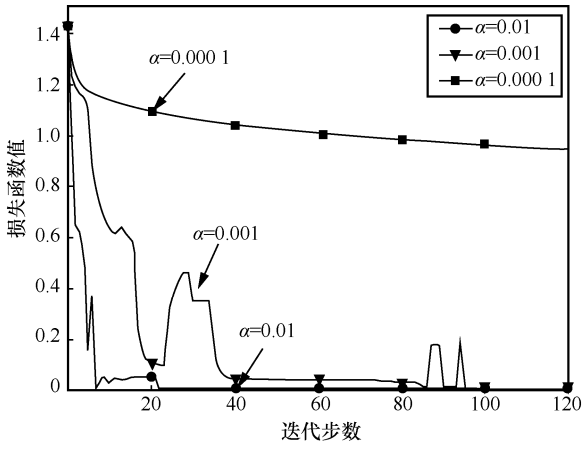

图 4 不同学习率 α 下的性能

不同优化算法的性能如图 5 所示，其针对参数更新的不同优化算法的性能。在学习开始时，这 3 种情况下的训练步数都非常大。随着回合数增加，收敛速度有增加的趋势。RMSProp 优化算法的收敛速度最快。因此，选择 RMSProp 算法更新参数。

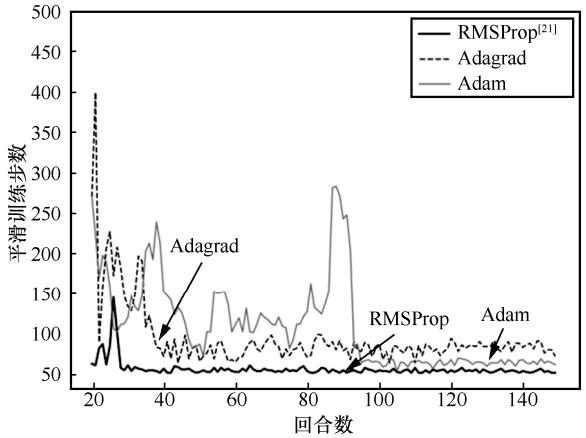

图 5 不同优化算法的性能

不同折扣率 γ 的性能如图 6 所示，其显示了不同折扣率 γ 对算法收敛的影响。γ=0.9 时，回合数在 30 开始收敛，γ=0.99 的曲线波动最大，收敛性能最差。γ 越高，表示希望智能体可以更关注未来奖励，这比关注当前要难得多，导致训练变得缓慢和困难。因此选择 γ=0.9。

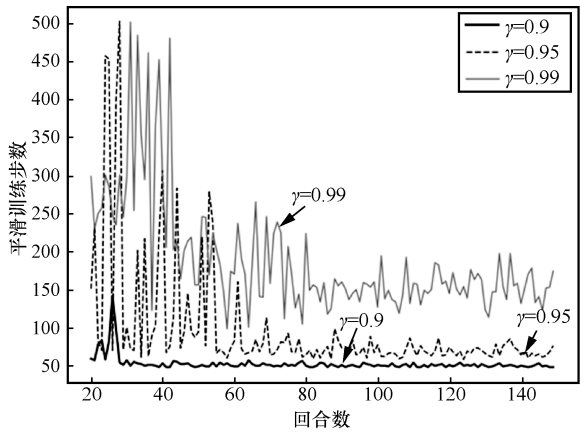

图 6　不同折扣率 γ 的性能

3 种算法性能比较如图 7 所示，其比较了 3 种不同探索策略的 DRL 算法的接入成功率。在学习开始时，3 种算法下的接入成功率都很低。随着训练回合数增加，成功率和收敛速度都有增加趋势。其中，本文算法的成功率最高，其次是 greedy-RL 算法，softmax-RL 算法成功率相对最低。此外，由于系统模型分布着不在小区覆盖范围的设备，因此无法接入基站，影响接入成功率。

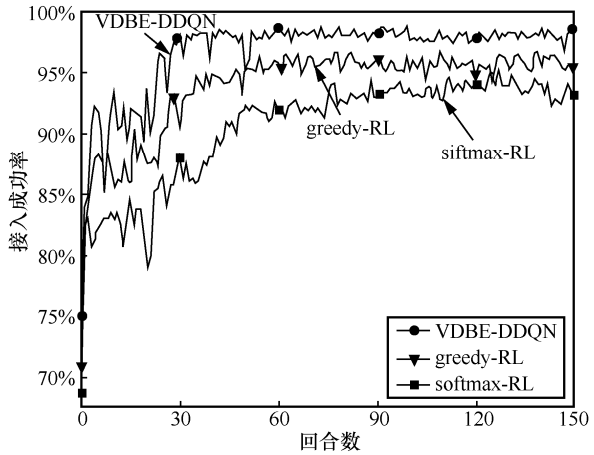

图 7　3 种算法性能比较

不同 $N_t^{\max}$ 值对 3 种算法的性能影响如图 8 所示，其显示了当每次允许接入的 mMTCD 数量最大值 $N_t^{\max}$ 不同时，对 3 种算法的性能影响。随着 $N_t^{\max}$ 值不断地增加，3 种算法的接入成功率也逐渐降低，这是由于当 $N_t^{\max}$ 值增大时，会增加 mMTCD 之间的接入冲突。但是，本文算法相较于其他算法更稳定，接入成功率也更高。

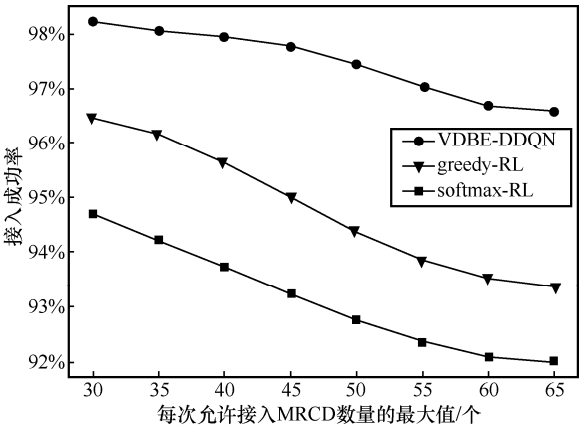

图 8 不同 $N_t^{\max}$ 值对 3 种算法的性能影响

4　结束语

针对大连接物联网场景，本文提出了 VDBE-DDQN 算法来解决在多小区网络环境下 mMTCD 选择 eNB 的问题。首先，本文将此 RL 算法的状态转移过程建模为 MDP，定义了其中的状态、动作和奖励函数，并根据设备之间的协作获得接入失败设备的碰撞级别作为奖励。其次，通过设计的网络来近似值函数，通过不断学习使目标值与值函数无限接近。同时，DDQN 也解决了传统算法对值函数高估的问题。然后，通过 VDBE 方法使每个 mMTCD 有适合自己的探索概率，而不是统一的标准。此外，该算法还能够感知网络环境的变化，调整探索和利用的比值。仿真结果表明，所提方法在接入成功率方面优于其他方法。

参考文献：

[1]　TULLBERG H, POPOVSKI P, LI Z X, et al. The METIS 5G

system concept: meeting the 5G requirements[J]. IEEE Communications Magazine, 2016, 54(12): 132-139.

[2] Latva-aho M, Leppänen K, Clazzer F, et al. Key drivers and research challenges for 6G ubiquitous wireless intelligence[J]. 2020.

[3] BI Q. Ten trends in the cellular industry and an outlook on 6G[J]. IEEE Communications Magazine, 2019, 57(12): 31-36.[LinkOut]

[4] 董石磊, 赵婧博. 面向工业场景的 5G 专网解决方案研究[J]. 电信科学, 2021, 37(11): 97-103.
DONG S L, ZHAO J B. Research on 5G private networking schemes for industry[J]. Telecommunications Science, 2021, 37(11): 97-103.

[5] POPLI S, JHA R K, JAIN S. A survey on energy efficient narrowband internet of things (NBIoT): architecture, application and challenges[J]. IEEE Access, 2018(7): 16739-16776.

[6] NAVARRO-ORTIZ J, ROMERO-DIAZ P, SENDRA S, et al. A survey on 5G usage scenarios and traffic models[J]. IEEE Communications Surveys & Tutorials, 2020, 22(2): 905-929.

[7] ANALYTICS S. Number of Internet of things(IoT) connected devices worldwide in 2018, 2025 and 2030(in billions)[J]. Statista Inc, 2020(7): 17.

[8] SHARMA S K, WANG X B. Toward massive machine type communications in ultra-dense cellular IoT networks: current issues and machine learning-assisted solutions[J]. IEEE Communications Surveys & Tutorials, 2020, 22(1): 426-471.

[9] 3GPP. Study on RAN improvements for machine-type communications:TR 37.868[R]. 2011.

[10] ALI M S, HOSSAIN E, KIM D I. LTE/LTE-A random access for massive machine-type communications in smart cities[J]. IEEE Communications Magazine, 2017, 55(1): 76-83.

[11] SHARMA S K, WANG X B. Collaborative distributed Q-learning for RACH congestion minimization in cellular IoT networks[J]. IEEE Communications Letters, 2019, 23(4): 600-603.

[12] DA SILVA M V, SOUZA R D, ALVES H, et al. A NOMA-based Q-learning random access method for machine type communications[J]. IEEE Wireless Communications Letters, 2020, 9(10): 1720-1724.

[13] TSOUKANERI G, WU S B, WANG Y. Probabilistic preamble selection with reinforcement learning for massive machine type communication (MTC) devices[C]//Proceedings of 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications. Piscataway: IEEE Press, 2019: 1-6.

[14] PACHECO-PARAMO D, TELLO-OQUENDO L. Adjustable access control mechanism in cellular MTC networks: a double Q-learning approach[C]//Proceedings of 2019 IEEE Fourth Ecuador Technical Chapters Meeting. Piscataway: IEEE Press, 2019: 1-6.

[15] BAI J N, SONG H, YI Y, et al. Multiagent reinforcement learning meets random access in massive cellular Internet of Things[J]. IEEE Internet of Things Journal, 2021, 8(24): 17417-17428.

[16] MOHAMMED A H, KHWAJA A S, ANPALAGAN A, et al. Base Station selection in M2M communication using Q-learning algorithm in LTE-A networks[C]//Proceedings of 2015 IEEE 29th International Conference on Advanced Information Networking and Applications. Piscataway: IEEE Press, 2015: 17-22.

[17] LEE D, ZHAO Y, LEE J. Reinforcement learning for random access in multi-cell networks[C]//Proceedings of 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Piscataway: IEEE Press, 2021: 335-338.

[18] MOON J, LIM Y. Access control of MTC devices using reinforcement learning approach[C]//Proceedings of 2017 International Conference on Information Networking (ICOIN). Piscataway: IEEE Press, 2017: 641-643.

[19] LIEN S Y, CHEN K C, LIN Y H. Toward ubiquitous massive accesses in 3GPP machine-to-machine communications[J]. IEEE Communications Magazine, 2011, 49(4): 66-74.

[20] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016: 2094-2100.

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

[22] TIELEMAN T, HINTON G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural networks for machine learning, 2012, 4(2): 26-31.

[作者简介]

李昕（1997– ），女，南京邮电大学通信与信息工程学院硕士生，主要研究方向为大连接物联网设备的随机接入。

孙君（1980– ），女，南京邮电大学副研究员、硕士生导师，主要研究方向为无线网络、无线资源管理和物联网。