基于强化学习的多基站协作接收时隙 Aloha 网络信道接入机制

黄元康[1]，詹文[1]，孙兴华[2]
（1. 中山大学（深圳），广东 深圳 518107；2. 中山大学，广东 广州 510275）

摘　要：随着物联网（IoT, internet of things）基站的部署愈发密集，网络干扰管控的重要性愈发凸显。物联网中，设备常采用随机接入，以分布式的方式接入信道。在海量设备的物联网场景中，节点之间可能会出现严重的干扰，导致网络的吞吐量性能严重下降。为了解决随机接入网络中的干扰管控问题，考虑基于协作接收的多基站时隙 Aloha 网络，利用强化学习工具，设计自适应传输算法，实现干扰管控，优化网络的吞吐量性能，并提高网络的公平性。首先，设计了基于 Q-学习的自适应传输算法，通过仿真验证了该算法面对不同网络流量时均能保障较高的网络吞吐量性能。其次，为了提高网络的公平性，采用惩罚函数法改进自适应传输算法，并通过仿真验证了面向公平性优化后的算法能够大幅提高网络的公平性，并保障网络的吞吐性能。

关键词：强化学习；物联网；随机接入；多基站网络；时隙 Aloha
中图分类号：TN92
文献标志码：A
doi：10.11959/j.issn.2096−3750.2024.00388

Reinforcement learning-based channel access mechanism for multi-base station slotted Aloha with cooperative reception

HUANG Yuankang[1], ZHAN Wen[1], SUN Xinghua[2]
1. Sun Yat-sen University Shenzhen Campus, Shenzhen 518107, China
2. Sun Yat-sen University, Guangzhou 510275, China

Abstract: With the increasingly dense deployment of base stations in the internet of things (IoT), the importance of interference management becomes ever more pronounced. In IoT environments, devices often employ random access, connecting to channels in a distributed manner. In scenarios involving massive numbers of devices, severe interference may arise between nodes, leading to significant degradation in the throughput performance of the network. To address interference control issues in networks with random access, a multi-base station slotted Aloha network based on cooperative reception was considered, the reinforcement learning techniques was leveraged to design adaptive transmission algorithms that effectively managed interference, optimized network throughput performance, and enhanced network fairness. Firstly, an adaptive transmission algorithm were devised based on Q-learning, which was verified to maintain high network throughput performance under varying traffic conditions through simulation. Secondly, to improve network fairness, the penalty function method was employed to refine the adaptive transmission algorithm. Simulations confirm that the fairness-optimized algorithm significantly enhances network fairness while preserving satisfactory network throughput performance.

Key words: reinforcement learning, internet of things, random access, multi-base station network, slotted Aloha

收稿日期：2023−11−21；**修回日期：**2024−05−16
通信作者：詹文，zhanw6@mail.sysu.edu.cn
基金项目：国家重点研发计划（No. 2023YFB2904100）；深圳市科技计划资助项目（No. RCBS20210706092408010）
Foundation Items: The National Key Research and Development Program of China (No. 2023YFB2904100), Shenzhen Science and Technology Program (No. RCBS20210706092408010)

0　引言

随着物联网（IoT, internet of things）技术的快速发展，IoT设备数量飞速增长。为支持海量机器类设备通信的IoT场景，通信网络需要能够容纳更多的接入设备[1-3]。同时，由于低频频谱资源紧缺，IoT的通信网络通常工作于2.4 GHz的公共频段，而高频电磁波传输距离较短，导致IoT的基站覆盖范围较小，基站部署变得更为密集。在海量接入设备、基站密集部署的网络中，存在多个基站，且不同基站的覆盖范围存在重叠区域的情况，这导致节点遭受错综复杂的干扰[4-6]。面向不同网络流量的场景，如何对海量设备的多基站随机接入网络进行干扰管控，并提高网络性能，以满足各类IoT应用的服务质量要求，是工程界与学术界重点关注的问题。

为了使网络能够灵活支撑海量设备通信，IoT常采用随机接入技术[7-9]，即节点自主决定何时传输，不需要中心基础设施的协调控制。因此，随机接入具有部署简单、开销低的优点，适用于海量设备的IoT场景。典型的随机接入机制包括Aloha、时隙Aloha和载波监听多址接入等，其中，时隙Aloha技术已在IoT通信领域被广泛应用。传统的随机接入技术通常采用退避机制以控制节点的重传。当节点遭遇冲突，则执行退避机制，随机等待一段时间再进行重传。虽然随机接入有效地提高了网络的规模，但在基站部署密集、网络覆盖重叠状况复杂、接入设备数目庞大的场景下，需要对节点的退避参数进行合适设置，才可以保证网络的吞吐量性能，否则竞争导致的干扰将严重影响网络的整体性能[10-14]。然而，传统的理论建模方法需要已知网络的报文输入速率、拓扑结构等，从而获取最优的退避参数。在实际海量设备的IoT场景中，报文输入速率往往是未知的。因此，采用理论建模方法难以实现退避参数的实时调整，难以保证流量动态变化场景下网络的吞吐量性能。

近年来，强化学习已经被广泛应用于随机接入领域。将强化学习技术应用于随机接入，可以对节点的传输进行实时、自适应的调整，从而提高网络的吞吐量性能。文献[15-18]研究了在单基站随机接入网络中采用强化学习技术管控节点传输的策略。文献[15]考虑了水下场景时延处于动态变化的特点，采用Q-学习以智能调控节点的退避时隙，从而减少报文碰撞概率。文献[16-17]设计了基于深度Q-学习的自适应传输算法。文献[16]提出的算法能够基于基站的二元反馈值实时生成传输策略，该算法与指数退避相比，具有更好的吞吐量性能，且适用于不同网络流量的场景。文献[17]提出的算法则面向单基站多信道的网络场景，该算法使节点能够自适应地选择更优的信道进行传输，降低丢包率。文献[18]提出了基于深度强化学习的媒体访问控制层协议，该协议能够使时分多址与Aloha共存的异构网络达到近似最优吞吐量性能。

上述研究仅考虑了单基站网络场景，在密集部署的IoT场景下，网络中通常会存在多个基站，节点间的干扰关系更为复杂，且处于不同基站覆盖区域的节点所受基站服务水平不同，可能会存在吞吐量性能不公平的问题。因此多基站场景的研究更具挑战性，也更具有现实意义。文献[19-22]研究了多小区的随机接入网络场景。文献[19]提出了一种适用于多小区Aloha网络的Q-学习传输算法，该算法中，小区间通过知识迁移相互交换彼此的传输状况，以获取小区间干扰的大小，并在传输策略生成的过程中考虑小区间干扰的影响，该算法能够显著地提高网络的吞吐量性能。文献[20]提出了一种迭代算法，对节点自身的传输功率不断地迭代优化，同时对节点的传输时隙进行合适的选择，从而最大化多基站网络的整体吞吐量。文献[21]考虑了接入节点的功率受限的场景，设计算法优化网络吞吐量并保证节点的发送功率低于限制值。文献[22]针对时延受限的场景，提出了一种基于强化学习的优化算法，该算法可以估计活跃节点的数目，并优化节点的传输策略，提高节点的接入成功概率。

上述研究仅考虑了传统的多小区网络，即节点与相应的基站进行绑定，节点发送的报文只能被绑定的基站进行译码，并且上述研究鲜有关注如何提高网络公平性的问题。本文聚焦于协作接收的多小区网络[23-27]，节点发送的报文被至少一个基站接收，即可成功译码。实际IoT场景中，多网关的远距离无线广域网（LoRaWAN, long range radio wide area network）可被视为一个基于协作接收的多基站网络[28-32]。本文设计了基于强化学习的自适应传输算法，优化了这种网络的吞吐量性能与公平性。

本文针对基于协作接收的多基站时隙Aloha网络，设计了基于Q-学习的自适应传输策略生成算

法。与单基站场景不同，在多基站场景下，一个接入设备可能处于多个基站的覆盖区域内，因此设备发送的报文可能会占用多个基站的信道，且会收到多个基站的反馈。本文设计的算法综合考虑多个基站的反馈值，对节点的传输策略进行自适应调整，降低节点间的干扰，并提高网络吞吐量性能。仿真结果表明，本文提出的算法能够优化网络吞吐量，但代价是网络中存在吞吐量不公平问题。

针对上述公平性问题，本文引入了与公平性相关的限制条件，并通过惩罚函数法对贝尔曼（Bellman）方程进行改进，改进了本文设计的多基站网络分布式强化学习算法，提高了算法的公平性。仿真结果表明，经过公平性优化后的算法能在保证网络吞吐量良好的基础上，大幅地提高公平性。

1　系统模型

假设网络中存在 n_{BS} 个基站，基站的集合记为 $M=\{1,2,\cdots,n_{\mathrm{BS}}\}$。整个网络共包含 n 个节点，节点的集合表示为 $N=\{1,2,\cdots,n\}$。假设每个节点的报文输入过程服从到达速率为 λ 的伯努利过程，每个节点的数据缓存大小为 1（报文），当节点缓存已满时，节点将舍弃新到达的报文。

网络中存在多个基站，基站的覆盖范围会发生重叠，处于重叠区域的节点能够与多个基站进行通信。本文根据节点与基站的关系进行集群划分，考虑一个基站的集合 A，集合 A 中共包含 $|A|$ 个基站，用集群 G_A 表示能够与集合 A 中的所有基站进行通信的所有节点，并假设集群 G_A 包含 n^A 个节点。集群划分如图 1 所示，橙色三角形表示只能与基站 1 进行通信的节点，即集群 $G_{\{1\}}$ 的节点；红色三角形表示能与基站 1 和基站 2 通信的节点，即集群 $G_{\{1,2\}}$ 的节点。

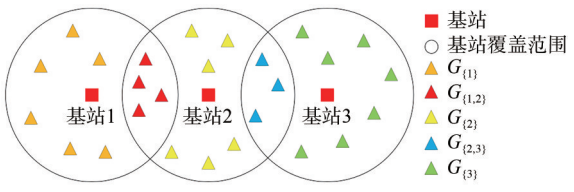


图 1　集群划分

时隙 Aloha 中，时间被离散化成时隙，同时基站与节点需要进行时隙同步，即基站的时隙起点需要与节点的时隙起点对齐。本文采用冲突模型，当基站在某一时隙只收到单个报文时，则基站能够成功地接收到该报文；当基站同时接收到多个报文时，则会发生冲突，基站无法解码报文。本文假设基站采用协作接收，即节点发送的报文被至少一个基站成功接收到即可成功传输。以集群 $G_{\{1,2\}}$ 为例，该集群的节点发送的报文只需要被基站 1 或 2 成功接收到即可成功传输。本文所考虑的场景常见于多网关 LoRaWAN。本文所考虑的系统模型在实际场景中对应于多网关 LoRaWAN。来自其他节点采用相同扩频因子、信道的信号可视为同频干扰。在实际 LoRaWAN 中，网络包含海量设备，且设备以分布式竞争的方式接入信道。当节点传输缺少合适的管控机制，网络中会存在严重的同频干扰问题。本文考虑如下两个性能指标。

1) 网络吞吐量：为整个网络在单个时隙内成功传输报文数目的长期平均值。令 $\lambda_{\mathrm{out},i}$ 表示节点 i 的吞吐量长期平均值，设 $i\in G_A$，本文定义集群 G_A 的集群吞吐量为集群 G_A 在单个时隙内成功传输报文数目的长期平均值 i 表示为

$$\lambda_{\mathrm{out},A}=\sum_{i\in G_A}\lambda_{\mathrm{out},i} \tag{1}$$

网络吞吐量可表示为

$$\lambda_{\mathrm{out}}=\sum_{i\subset N}\lambda_{\mathrm{out},i} \tag{2}$$

2) Jain 公平指数（Jain fairness index）[33]：为衡量网络的公平性性能，本文考虑 Jain 公平指数，计算式为

$$\Phi=\frac{(\sum_{i\subset N}\lambda_{\mathrm{out},i})^2}{n\sum_{i\subset N}(\lambda_{\mathrm{out},i})^2} \tag{3}$$

Jain 公平指数越接近 1，说明节点吞吐量差异越小，网络的公平性越好。

2　基于 Q-学习的多基站传输策略研究

2.1　传输策略

假设节点各自配备强化学习模块，通过执行分布式的强化学习算法，节点能够生成传输策略，实现节点级的干扰管控，并优化网络吞吐量性能。以下基于 Q-学习的多基站传输策略进行说明。

考虑某个节点 i ($i\in G_A$)，记节点 i 在时隙 t 的状态值为 $S_i(t)=\{A_i(t-1), F_i(t-1), B_i(t)\}$，其中，$A_i(t-1)$ 为节点 i 在时隙 $t-1$ 的传输策略，$F_i(t-1)$ 为节点 i 在时隙 $t-1$ 接收到的基站反馈值集合，$B_i(t)$ 为节点 i 在时隙 t 的数据缓存状态。本文将节点 i 在时隙 t 的数据

缓存状态值表示为

$$B_i(t)=\begin{cases}1, & \text{节点}i\text{在时隙}t\text{有待发送报文}\\ 0, & \text{节点}i\text{缓存队列在时隙}t\text{为空}\end{cases} \tag{4}$$

节点i在时隙$t-1$的传输策略为

$$A_i(t-1)=\begin{cases}1, & \text{节点}i\text{在时隙}t-1\text{请求传输}\\ 0, & \text{节点}i\text{在时隙}t-1\text{不请求传输}\end{cases} \tag{5}$$

在时隙$t-1$的末尾，基站将广播应答/否定应答（ACK/NACK, acknowledgment/negative acknowledgment），节点i根据是否接收到基站广播的ACK/NACK获取反馈值，生成回报值并更新Q表。将节点i收到来自基站l的反馈值记为$F_{i,l}(t-1)$，该值反映了时隙$t-1$基站l覆盖网络内的传输情况，$F_{i,l}(t-1)$的表达式为

$$F_{i,l}(t-1)=\begin{cases}1, & \text{节点}i\text{在时隙}t-1\text{接收到ACK}\\ -1, & \text{节点}i\text{在时隙}t-1\text{接收到NACK}\\ 0, & \text{其他}\end{cases} \tag{6}$$

其中，$l\in A$。将节点i在时隙$t-1$所接收到的所有基站反馈值表示为集合$F_i(t-1)=\{F_{i,l}(t-1), l\in A\}$。

本文采用softmax策略进行传输策略的选取。softmax策略中，已知节点i在时隙t的状态$S_i(t)$，选取传输策略A（$A\in\{0,1\}$）的概率为

$$\Pr\{A_i(t)=A|S_i(t)\}=\frac{\exp\{\beta(Q(S_i(t),A))\}}{\sum\limits_{A_1\in\{0,1\}}\exp\{\beta(Q(S_i(t),A_1))\}} \tag{7}$$

其中，$\beta\in(0,\infty)$为策略选择概率生成倾向，β越大，较大Q值策略被选取概率越大。

在时隙t的末尾，基站将根据是否成功接收报文，选择广播或者不广播ACK/NACK。节点根据是否接收到ACK/NACK获取反馈值，生成回报值并更新Q表。以节点i为例（$i\in G_A$），节点i的传输策略对附近所有可通信基站所覆盖的网络都产生了影响，节点i对基站l（$l\in A$）覆盖网络的影响用基站回报值$R_{i,l}(t)$体现。基站回报值$R_{i,l}(t)$有3种取值情况。

1) 如果基站l在时隙t成功接收到报文，则广播ACK，反馈值$F_{i,l}(t)=1$，节点i的传输策略对信道的利用产生正面影响，基站回报值为$R_{i,l}(t)=1$。

2) 如果基站l在时隙t没有接收到任何报文，则不广播ACK/NACK，反馈值$F_{i,l}(t)=0$，此时信道处于空闲状态，基站回报值$R_{i,l}(t)=-1$。

3) 如果基站l在时隙t接收到冲突的报文，则广播NACK，反馈值$F_{i,l}(t)=-1$，若节点i在时隙t不进行传输，即$A_i(t)=0$，节点i的传输策略对信道冲突不负有责任，因此本文设置此时的基站回报值为$R_{i,l}(t)=0$；若节点i在时隙t进行传输，即$A_i(t)=1$，节点i的传输策略对信道的利用产生了负面影响，因此设置基站回报值$R_{i,l}(t)=-1$。

综上所述，基站回报值$R_{i,l}(t)$表示为

$$R_{i,l}(t)=\begin{cases}1, & F_{i,l}(t)=1\\ 0, & F_{i,l}(t)=-1, A_i(t)=0\\ -1, & \text{其他}\end{cases} \tag{8}$$

节点的总回报值为所有基站回报值的期望，表示为

$$R_i(t)=\frac{\sum\limits_{l\in A}R_{i,l}(t)}{|A|} \tag{9}$$

节点i采用贝尔曼方程更新Q表[16]，即：

$$Q(S_i(t),A_i(t))\leftarrow Q(S_i(t),A_i(t))+\alpha(R_i(t)+\gamma\max_A Q(S_{\text{new}},A)-Q(S_i(t),A_i(t))) \tag{10}$$

其中，$l\in N$，$\alpha\in(0,1)$为贝尔曼方程的学习率，$\gamma\in(0,1)$为贝尔曼方程的衰减系数，$S_{\text{new}}=\{A_i(t), F_i(t), B_i(t)\}$。基于Q-学习自适应传输策略如图2所示，上述流程按时间顺序总结如下。

步骤1　节点根据时隙t的状态值生成传输策略，并执行传输策略。

步骤2　节点根据基站发送的ACK/NACK生成反馈值。

步骤3　节点根据反馈值获得回报值，并通过式(10)更新Q表。

步骤4　节点存储时隙t的传输策略与反馈值，用于构成下一时隙的状态值。

在系统的初始化阶段，节点的Q表随机初始化，此后节点执行上述流程步骤，不断进行Q表的更新。最终，每个节点的传输策略能够得到优化，从而实现干扰管控，提高网络吞吐量性能。

2.2　仿真结果

本文仿真基于MATLAB软件实现，每轮仿真运行5.0×10^6个时隙，共运行20轮仿真。如图3、图4所示，曲线表示各轮仿真结果的平均值，阴影表示各轮仿真结果的最大值和最小值所构成的区间。仿真中网络吞吐量的计算式为网络成功传输报文总数与仿真时隙总数的比值，采用式(3)计算Jain公平指数。设置学习率α随仿真时隙自适应调整，令第t个仿真时隙的学习率为$\max(0.01\exp(-10^{-4}t),10^{-6})$，原

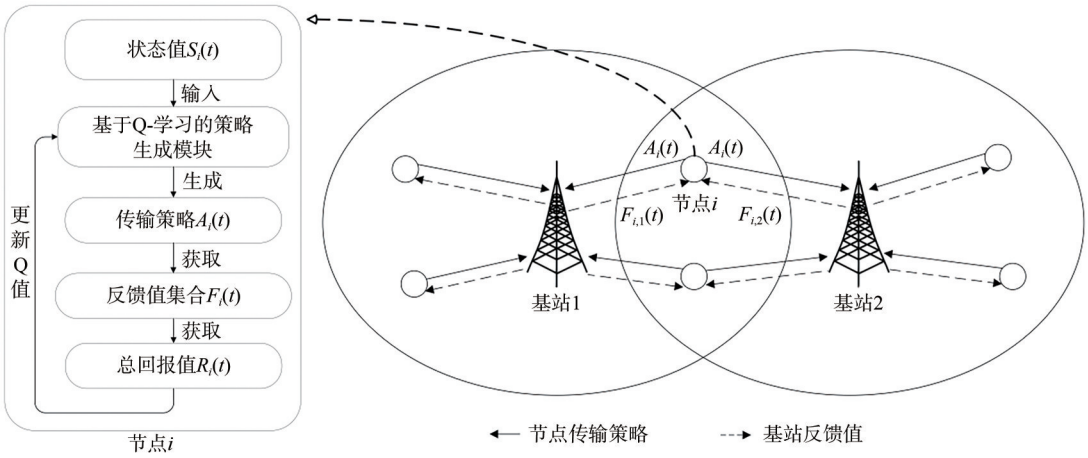
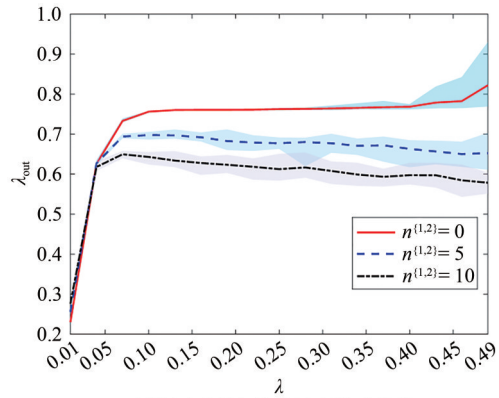


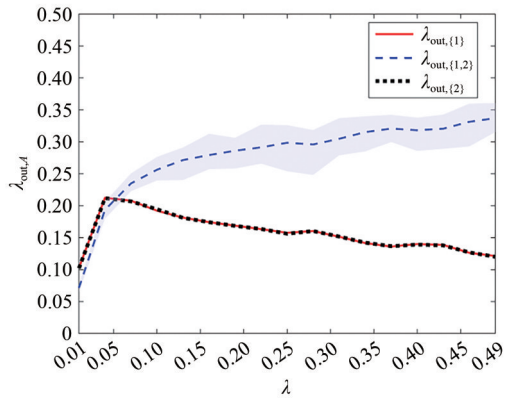
图2　基于Q-学习自适应传输策略

因为采用自适应学习率，能够加快算法收敛速度，提高算法的学习效果[16]。吞吐量性能与输入速率关系曲线如图3所示（n_{BS}=2，$n^{\{1\}}$= $n^{\{2\}}$=15，γ=0.9，β=5），网络公平性与输入速率关系曲线如图4所示（n_{BS}=2，$n^{\{1\}}$= $n^{\{2\}}$=15，γ=0.9，β=5）。

图3(a)为网络吞吐量与输入速率的关系曲线，图3(b)为重叠区域节点数为10时，各集群吞吐量与输入速率的关系曲线。由图3(a)可知，当0.01≤λ≤0.05时，网络吞吐量随着输入速率线性增长，原因为网络流量小，网络不饱和，网络吞吐量等于总输入速率；当λ>0.05且$n^{\{1,2\}}$=0时，网络吞吐量随输入速率增加而缓慢增长；当λ>0.05且$n^{\{1,2\}}\in\{5, 10\}$时，网络吞吐量随输入速率增加而缓慢下降。由图3(b)可知，发生上述现象的原因是当λ>0.05时，$\lambda_{out,\{1,2\}}$随着输入速率增大而缓慢增长，而处于重叠区域的节点会同时占用多个信道的资源，成功发送后也仅

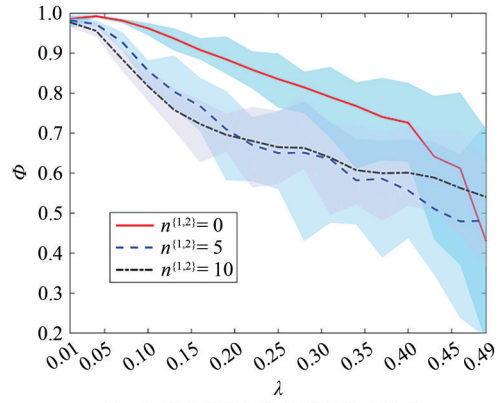


(a) 网络吞吐量与输入速率关系曲线

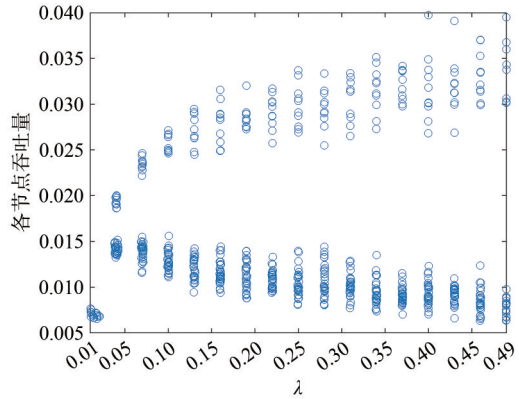


(b) 各集群吞吐量与输入速率关系曲线

图3　吞吐量性能与输入速率关系曲线



(a) Jain公平指数与输入速率关系曲线



(b) 各节点的吞吐量与输入速率关系曲线

图4　网络公平性与输入速率关系曲线

能成功传输一个报文，导致信道资源的浪费，最终使网络整体的吞吐量下降。由上述仿真结果可知，面对不同输入速率λ和不同重叠区域节点数目$n^{\{1,2\}}$，本文设计的Q-学习传输策略生成算法使网络达到较高的网络吞吐量。

图4(a)为Jain公平指数与输入速率的关系曲线，图4(b)为重叠区域节点数目为10的情况下各节点吞吐量与输入速率的关系。由图4(a)可知，随着输入速率的增长，Jain公平指数逐渐降低，网络公平性下降。由图4(b)可知，当输入速率λ<0.1时，各节点的吞吐量相近，网络公平性良好；当λ>0.1时，随着λ的增大，节点的吞吐量差异逐渐增大，网络公平性逐渐下降。随着输入速率λ的增大，节点吞吐量的最大值与最小值差距越来越大。结合图4(a)可知，在λ较大的情况下，随着重叠区域节点数目增多，节点的吞吐量差异增大。当λ=0.49且$n^{\{1,2\}}$=10时，节点吞吐量的最大值约为0.040，最小值接近0.006。综上，本文设计的基于Q-学习自适应传输策略可以取得良好的网络吞吐量性能，但随着输入速率的增大，节点的吞吐量差异增大，网络公平性下降，这一现象随着重叠区域节点数目的增加而加剧。因此，本文在第3节研究如何在保持良好的网络吞吐量性能的同时，提高网络公平性。

3　面向公平性改进的Q-学习传输策略

3.1　公平性改进原理

为了提高自适应传输算法在输入速率较大场景下的公平性，本文采用惩罚函数法对算法进行了改进，核心思想是引入基准吞吐量，并采用惩罚因子改进贝尔曼方程，从而调整节点的传输策略，使每个节点的吞吐量逼近基准吞吐量，最终提高网络的公平性。本文设计的面向公平性改进后的贝尔曼方程为

$$Q(S_i(t),A_i(t)) \leftarrow Q(S_i(t),A_i(t)) + \alpha(R_i(t) + \delta_i + \gamma \max_{A_{\text{new}}} Q(S_{\text{new}},A_{\text{new}}) - Q(S_i(t),A_i(t)) \tag{11}$$

其中，$i \in N$，δ_i为贝尔曼方程的惩罚因子。对比式(10)和式(11)，二者的主要区别在于式(11)中增加了惩罚因子δ_i。δ_i需要根据节点的动态修正系数σ_i，节点i的近期吞吐量$\lambda_{\text{trans},i}(t)$以及基准吞吐量$\lambda_{\text{B}}$进行设置，以对节点的传输策略进行公平性优化。

基准吞吐量λ_{B}是为了提升网络公平性，每个节点应该尽可能靠近的、一致且可达的吞吐量。从系统性能优化角度看，λ_{B}应该尽可能大。多基站的时隙Aloha网络所能达到的理论最大吞吐量为$n_{\text{BS}}\exp\{-1\}$，其中，n_{BS}表示基站数量，$\exp\{-1\}$是在不考虑外环境干扰的情况下，单个基站可以达到的最大吞吐量。因此，单个节点可达的、最大吞吐量为$n_{\text{BS}}\exp\{-1\}/n$。然而，达到上述最大吞吐量的前置条件为所有节点的数据总和输入速率大于或者等于$n_{\text{BS}}\exp\{-1\}$。当数据输入速率λ较小时，$n_{\text{BS}}\exp\{-1\}/n$虽然不可达，但此时网络有可能通过适当调整传输策略使得节点队列处于稳态，即吞吐量等于输入速率。在同构网络场景下，可以将基准吞吐量设置为λ。综上，基准吞吐量被设置为

$$\lambda_{\text{B}} = \min\{n_{\text{BS}}\exp\{-1\}/n, \lambda\} \tag{12}$$

为度量每个节点的吞吐量偏离基准吞吐量的程度，定义节点i的公平性代价值$C_i(t)$为该节点的近期吞吐量与基准吞吐量之差，即

$$C_i(t) = \lambda_{\text{trans},i}(t) - \lambda_{\text{B}} \tag{13}$$

其中，近期吞吐量$\lambda_{\text{trans},i}(t)$为节点$i$在时隙$t-T_{\text{sample}}$至时隙$t-1$的吞吐量。为获取近期吞吐量，每个节点需要存储最近$T_{\text{sample}}$个时隙的基站反馈值。

当$C_i(t)$<0时，节点的吞吐量小于基准吞吐量。此时为达到公平性需求，应当增大节点传输概率。因此，当$A_i(t)$=0时，惩罚因子δ_i需要小于0，以降低迭代后的Q值，减小节点不传输数据的概率；当$A_i(t)$=1时，δ_i需要大于0，增大迭代后的Q值，鼓励节点传输数据；同理，当$C_i(t)$>0时，节点的吞吐量大于基准吞吐量，此时为满足公平性需求，应当降低该节点传输的概率。因此，当$A_i(t)$=0时，惩罚因子δ_i需要大于0；当$A_i(t)$=1时，δ_i需要小于0。综上，本文设置惩罚因子δ_i为

$$\delta_i = \begin{cases} \mu\sigma_i C_i(t), & A_i(t) = 0 \\ -\mu\sigma_i C_i(t), & A_i(t) = 1 \end{cases} \tag{14}$$

其中，μ为全网静态修正系数，σ_i为各节点的动态修正系数，Ω为动态修正系数的学习率，动态修正系数更新规则为$\sigma_i \leftarrow \max\{0, \sigma_i + \Omega C_i(t)\}$，即吞吐量偏离基准吞吐量的程度$C_i(t)$越大，那么动态修正系数越大，算法调整的步长越大。

基于惩罚函数法改进的Q-学习自适应传输策略算法如算法1所示。

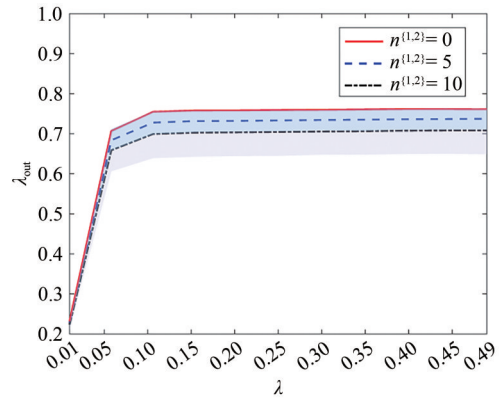
算法1　基于惩罚函数法改进的Q-学习自适应传输策略算法

1) **初始化：**对参数α，γ，T_σ，ρ，μ，σ_i，Ω，T_{sample}，T_{len}进行初始化，t=1。
2) **while** 1 **do**
3)　**for** $i\in N$ **do**
4)　　**if** mod(t, T_{len})=0 **then**
5)　　根据式(12)和式(13)计算$C_i(t)$；
6)　　$\sigma_i\leftarrow\max\{0, \sigma_i+\Omega C_i(t)\}$；
7)　　$\Omega\leftarrow\rho\Omega$；
8)　　**end if**
9)　　**if** mod(t, T_σ)=0 **then**
10)　　重置σ_i，Ω为初始值；
11)　　**end if**
12)　节点i根据softmax策略生成传输策略$A_i(t)$，并执行传输策略；
13)　**end for**
14)　基站广播ACK/NACK，节点获取反馈值；
15)　**for** $i\in N$ **do**
16)　节点i根据反馈值生成回报值；
17)　节点i根据式(11)和式(14)更新Q值；
18)　**end for**
19)　$t\leftarrow t$+1；
20)　**end while**

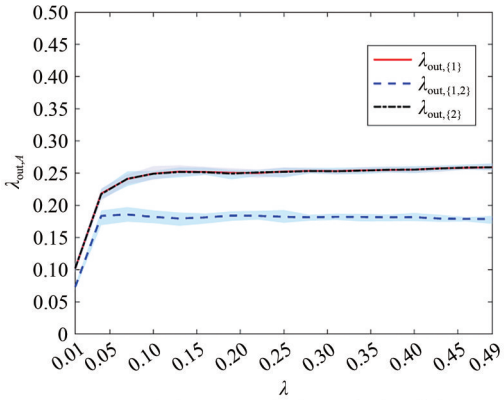
该算法总结了面向公平性改进的Q-学习自适应传输策略的算法流程。算法1的开始，基站对节点的各参数进行初始化，算法1的步骤4)至11)，mod(a,b)=0表示a能够被b整除，其中，步骤4)至8)表示节点每隔T_{len}个时隙，就进行一次$C_i(t)$和σ_i的计算与更新，而$\sigma_i\geqslant 0$，因此$\sigma_i+\Omega C_i(t)<0$时，需要将σ_i设为0。步骤7)中$\rho\in(0,1)$，目的是使学习率递减，从而提高算法的收敛速度。步骤9)至11)表示节点每隔T_σ就将σ_i和Ω重置为初始值，避免σ_i变为0后节点不再进行公平性调整。步骤17)中，节点采用面向公平性改进后的贝尔曼方程式(11)进行Q值更新。

3.2　仿真结果

仿真条件与第2.2节一致，优化后吞吐量与输入速率关系曲线如图5所示，优化后公平性与输入速率关系曲线如图6所示。图5(a)为算法经过公平性优化后，网络吞吐量与输入速率的关系曲线（n_{BS}=2，$n^{\{1\}}=n^{\{2\}}$=15，γ=0.9，β=5，T_{sample}=100，T_σ=1.0×10^4，T_{len}=10，μ=400，Ω=20，ρ=0.6，σ_i=20），图5(b)为算法经过公平性优化后，$n^{\{1,2\}}$=10时各集群吞吐量与输入速率的关系曲线。由图3(a)与图5(a)对比可知，当$\lambda\leqslant$0.05，$n^{\{1,2\}}\in\{0,5,10\}$时，改进后算法的吞吐量性能无明显变化。当$n^{\{1,2\}}\in\{5,10\}$且$\lambda>$0.05时，算法改进后网络的吞吐量性能略有提升，且几乎不随λ的增大而波动。由图5(b)可知，发生上述现象的原因是经过公平性改进后，算法抑制了重叠区域集群的吞吐量增长，从而提高了网络在输入速率较大情况下的公平性和吞吐量性能。



(a) 网络吞吐量与输入速率关系曲线

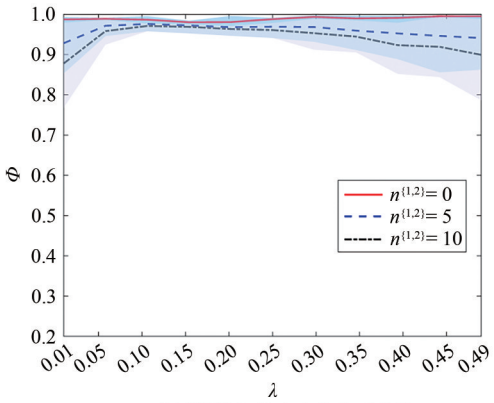


(b) 各集群吞吐量与输入速率关系曲线

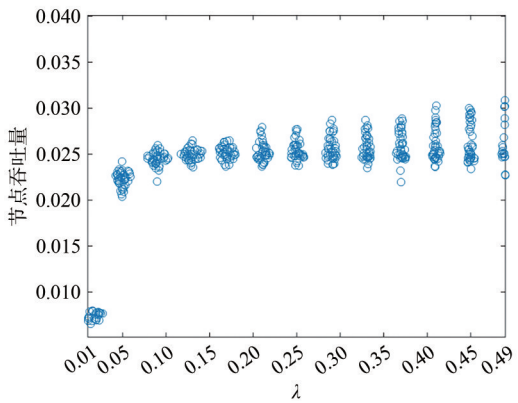
图5　优化后吞吐量与输入速率关系曲线

图6(a)为公平性优化后，Jain公平指数与输入速率的关系曲线（n_{BS}=2，$n^{\{1\}}=n^{\{2\}}$=15，γ=0.9，β=5，T_{sample}=100，T_σ=1.0×10^4，T_{len}=10，μ=400，Ω=20，ρ=0.6，σ_i=20），图6(b)为公平性优化后，重叠区域节点数目为10时各节点吞吐量与输入速率的关系曲线。由图4(a)与图6(a)对比可知，算法经过公平性改进后，Jain公平指数显著提高，始终逼近0.9，网络公平性大幅提升。对比图4(b)和图6(b)可知，算法经过公平性改进后，各节点吞吐量集中分布在区间(0.02,0.03)内，避免部分节点吞吐量跌至0，网络公平性得到了显著提升。

以上仿真结果说明本文提出的公平性优化方法

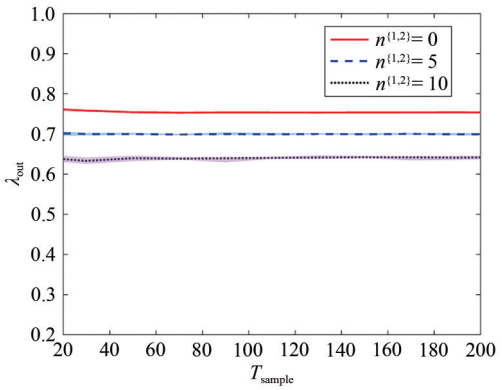


(a) Jain公平指数与输入速率关系曲线

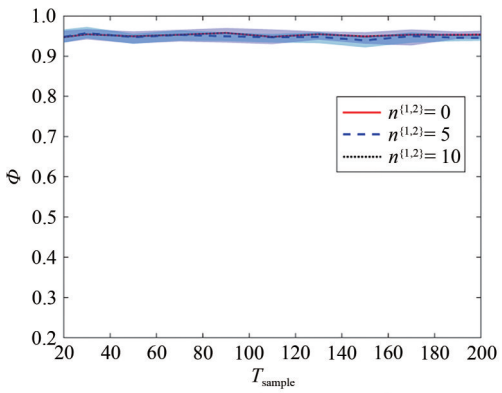


(b) 各节点吞吐量与输入速率关系曲线

图6　优化后公平性与输入速率关系曲线



(a) 网络吞吐量与T_{sample}关系曲线



(b) Jain公平指数与T_{sample}关系曲线

图7　优化后算法性能T_{sample}关系曲线

能够大幅地提高算法的公平性，且仍能保持较高的网络吞吐量。但改进后的算法要求节点存储最近T_{sample}个时隙收到的基站反馈值与缓存状态值，用于计算度量算法运行开销对性能的影响，优化后算法性能T_{sample}关系曲线如图7所示。图7展示了随着T_{sample}增加，网络吞吐量λ_{out}和公平性指标Φ的变化情况（n_{BS}=2，$n^{\{1\}}=n^{\{2\}}$=15，γ=0.9，λ=0.1，β=5，T_σ=1.0×10^4，T_{len}=10，μ=100，Ω=10，ρ=0.6，σ_i=20），其中，T_{sample}为节点需要存储的基站反馈值的时间跨度。由仿真结果可知，吞吐量λ_{out}与T_{sample}弱相关，公平性指数Φ在短暂增加后趋于平稳。这意味着在确保网络性能的前提下，网络可以将T_{sample}设为很小的值（比如T_{sample}=10）。较低的附加成本对实际网络的部署具有重要意义。公平性优化涉及参数见表1。

表1　公平性优化涉及参数

参数	参数意义
α	贝尔曼方程的学习率，$\alpha\in(0,1)$
γ	贝尔曼方程的衰减系数，$\gamma\in(0,1)$
β	softmax策略生成倾向
T_σ	动态修正系数重置间隔
T_{len}	动态修正系数与代价值的更新间隔
T_{sample}	计算近期吞吐量的采样时隙数
Ω	动态修正系数的学习率
ρ	学习率的衰减系数
μ	静态修正系数，$\mu\in(0,\infty)$
$C_i(t)$	节点i在时隙t的代价值
σ_i	动态修正系数，$\sigma_i\in[0,\infty)$
δ_i	节点i的贝尔曼方程惩罚因子

4　结束语

本文设计了基于Q-学习的自适应传输策略生成算法，该算法适用于基于协作接收的多基站时隙Aloha网络，并通过仿真验证了该算法面对不同网络流量时均能保障网络吞吐量性能。为了提高网络的公平性，本文采用惩罚函数算法改进了自适应传输算法，仿真结果显示，经过公平性优化后，仍能保持较高的吞吐量，且Jain公平指数显著提高，节点吞吐量差异缩小，网络公平性显著提升。因此，本文设计的算法对实际LoRaWAN部署具有较高的参考价值。未来工作中，有必要进一步搭建大规模

物联网硬件验证平台，在物理场景中测试算法性能，并考虑更多的性能指标，如网络能效。

参考文献：

[1] BOCCARDI F, HEATH R W, LOZANO A, et al. Five disruptive technology directions for 5G[J]. IEEE Communication Magazine, 2014, 52(2): 74-80.

[2] CHETTRI L, BERA R. A comprehensive survey on Internet of things (IoT) toward 5G wireless systems[J]. IEEE Internet of Things Journal, 2020, 7(1): 16-32.

[3] AL-GARADI M A, MOHAMED A, AL-ALI A K, et al. A survey of machine and deep learning methods for internet of things (IoT) security[J]. IEEE Communications Surveys and Tutorials, 2020, 22(3): 1646-1685.

[4] GUPTA A, JHA R K. A survey of 5G network: architecture and emerging technologies[J]. IEEE Access, 2015, 3: 1206-1232.

[5] VAEZI M, AZARI A, KHOSRAVIRAD S R, et al. Cellular, wide-area, and non-terrestrial IoT: a survey on 5G advances and the road toward 6G[J]. IEEE Communications Surveys and Tutorials, 2022, 24(2): 1117-1174.

[6] AL-TURJMAN F, EVER E, ZAHMATKESH H. Small cells in the forthcoming 5G/IoT: traffic modelling and deployment overview[J]. IEEE Communications Surveys and Tutorials, 2019, 21(1), 28-65.

[7] ZHANG Z, LI Y, HUANG C, et al. User activity detection and channel estimation for grant-free random access in LEO satellite-enabled internet of things[J]. IEEE Internet of Things Journal, 2020, 7(9): 8811-8825.

[8] SHANMUGA SUNDARAM J P, DU W, ZHAO Z. A survey on LoRa networking: research problems, current solutions, and open issues[J]. IEEE Communications Surveys and Tutorials, 2020, 22(1): 371-388.

[9] BELTRAMELLI L, MAHMOOD A, OSTERBETG P, et al. LoRa beyond ALOHA: an investigation of alternative random access protocols[J]. IEEE Transactions on Industrial Informatics, 2021, 17(5): 3544-3554.

[10] ABRAMSON N. The throughput of packet broadcasting channels[J]. IEEE Transactions on Communications, 1977, 25(1): 117-128.

[11] SAYED A H, TARIGHAT A, KHAJEHNOURI N. Network-based wireless location: challenges faced in developing techniques for accurate wireless location information[J]. IEEE Signal Processing Magazine, 2005, 22(4): 24-40.

[12] ZHAN W, SUN X, WANG X, et al. Performance optimization for massive random access of mMTC in cellular networks with preamble retransmission limit[J]. IEEE Transactions on Vehicular Technology, 2021, 70(9): 8854-8867.

[13] ZHAN W, DAI L. Massive random access of machine-to-machine communications in LTE networks: throughput optimization with a finite data transmission rate[J]. IEEE Transactions on Wireless Communications, 2019, 18(12): 5749-5763.

[14] YANG Y, DAI L. Stability region and transmission control of multi-cell Aloha networks[J]. IEEE Transactions on Communications, 2023, 71(9): 5348-5364.

[15] AHMED F, CHO H-S. A time-slotted data gathering medium access control protocol using Q-learning for underwater acoustic sensor networks[J]. IEEE Access, 2021(9): 48742-48752.

[16] JADOON M A, PASTORE A, NAVARRO M, et al. Deep reinforcement learning for random access in machine-type communication[C]//Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC), 2022: 2553-2558.

[17] CUI Q, ZHANG Z, SHI Y, et al. Dynamic multichannel access based on deep reinforcement learning in distributed wireless networks[J]. IEEE Systems Journal, 2022, 16(4): 5831-5834.

[18] YU Y, WANG T, LIEW S C. Deep-reinforcement learning multiple access for heterogeneous wireless networks[C]//Proceedings of the 2018 IEEE International Conference on Communications (ICC), 2018: 1-7.

[19] SHIN K-S, CHOI H-H, LEE H. Knowledge transfer-based multiagent Q-learning for medium access in dense cellular networks[J]. IEEE Wireless Communications Letters, 2022, 11(12): 2542-2545.

[20] WANG S, WANG X, ZHANG Y, et al. Resource allocation in multi-cell NOMA systems with multi-agent deep reinforcement learning[C]//Proceedings of the 2021 IEEE Wireless Communications and Networking Conference (WCNC), 2021: 1-6.

[21] KHAIRY S, BALAPRAKASH P, CAI L X, et al. Constrained deep reinforcement learning for energy sustainable multi-UAV based random access IoT networks with NOMA[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(4): 1101-1115.

[22] YOUN J, PARK J, KIM S, et al. MARL-based random access scheme for delay-constrained umMTC in 6G[C]//Proceedings of the 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), 2023: 1-6.

[23] OGATA S, ISHIBASHI K, FREITAS DE ABREU G T. Optimized frameless aloha for cooperative base stations with overlapped coverage areas[J]. IEEE Transactions on Wireless Communications, 2018, 17(11): 7486-7499.

[24] JAKOVETIC D, BAJOVIC D, VUKOBRATOVIC D, et al. Cooperative slotted aloha for multi-base station systems[J]. IEEE Transactions on Communications, 2015, 63(4): 1443-1456.

[25] YU C-H, HUANG L, CHANG C-S, et al. Poisson receivers: a probabilistic framework for analyzing coded random access[J]. IEEE/ACM Transactions on Networking, 2021, 29(2): 862-875.

[26] WANG R, LI P, CUI G, et al. Cooperative slotted aloha with reservation for multi-receiver satellite IoT networks[C]//Proceedings of the 2018 IEEE/CIC International Conference on Communications in China (ICCC), 2018: 593-597.

[27] MASTILOVIC A, VUKOBRATOVIC D, JAKOVETIC D, et al. Cooperative slotted aloha for massive M2M random access using directional antennas[C]//Proceedings of the 2017 IEEE Interna-

tional Conference on Communications Workshops (ICC Workshops), 2017: 731-736.

[28] GEORGIOU O, PSOMAS C, KRIKIDIS I. Coverage scalability analysis of multi-cell LoRa networks[C]//Proceedings of the 2020 IEEE International Conference on Communications (ICC), 2020: 1-7.

[29] BOUAZIZI Y, BENKHELIFA F, MCCANN J. Spatiotemporal modelling of multi-gateway LoRa networks with imperfect SF orthogonality[C]//Proceedings of the 2020 IEEE Global Communications Conference (GLOBECOM), 2020: 1-7.

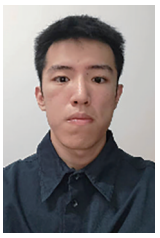
[30] TU L T, BRADAI A, POUSSET Y. Coverage probability and spectral efficiency analysis of multi-gateway downlink LoRa networks[C]//Proceedings of the 2022 IEEE International Conference on Communications (ICC), 2022: 1-6.

[31] OCHOA M N, MAMAN M, DUDA A. Spreading factor allocation for LoRa nodes progressively joining a multi-gateway adaptive network[C]//Proceedings of the IEEE Global Communications Conference (GLOBECOM), 2020: 1-6.

[32] HEUSSE M, CAILOUET C, DUDA A. Performance of unslotted ALOHA with capture and multiple collisions in LoRaWAN[J]. IEEE Internet of Things Journal, 2023, 10(20): 17824-17838.

[33] JAIN R , CHIU D , HAWE W. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems[J]. arXiv preprint, 1998, arXiv: cs-NI/9809099.

[作者简介]



黄元康（1999– ），男，中山大学（深圳）硕士生，主要研究方向为无线通信、物联网和随机接入。

詹文（1990– ），男，博士，中山大学（深圳）电子与通信工程学院副教授，主要研究方向为5G/6G网络和大规模物联网通信。

孙兴华（1985– ），男，博士，中山大学电子与通信工程学院副教授，主要研究方向为下一代无线通信网络、智能通信等。