# 基于多智能体强化学习的大规模灾后用户分布式覆盖优化

许文俊[1]，吴思雷[1]，王凤玉[1]，林兰[1]，李国军[2]，张治[3]

（1. 北京邮电大学人工智能学院，北京 100876；2. 重庆邮电大学超视距可信信息传输研究所，重庆 400065；
3. 北京邮电大学信息与通信工程学院，北京 100876）

**摘 要**：为了快速恢复大规模受灾用户的应急通信服务，针对接入用户数量众多导致的业务差异性和动态性显著、集中式算法难以扩展等问题，提出了一种基于多智能体强化学习的分布式智简覆盖优化架构。在网络特征层中，设计了考虑用户业务差异性的分布式 k-sums 分簇算法，每个无人机基站从用户需求出发，原生简约地调整局部网络结构，并筛选簇中心用户特征作为多智能体强化学习神经网络的输入状态。在轨迹调控层中，设计了多智能体最大熵强化学习（MASAC）算法，无人机基站作为智能节点以“分布式训练–分布式执行”的框架调控自身飞行轨迹，并融合集成学习和课程学习技术提升了训练稳定性和收敛速度。仿真结果表明，所提分布式 k-sums 分簇算法在平均负载效率和分簇均衡性方面优于 k-means 算法，基于 MASAC 的无人机基站轨迹调控算法能够有效减小通信中断的发生频率、提升网络的频谱效率，效果优于现有的强化学习方法。

**关键词**：应急通信；覆盖优化；多智能体强化学习；分布式训练

**中图分类号**：TN929.5

**文献标志码**：A

**DOI:** 10.11959/j.issn.1000−436x.2022131

# Large-scale post-disaster user distributed coverage optimization based on multi-agent reinforcement learning

XU Wenjun[1], WU Silei[1], WANG Fengyu[1], LIN Lan[1], LI Guojun[2], ZHANG Zhi[3]

1. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Lab of BLOS Trusted Information Transmission, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
3. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

**Abstract:** In order to quickly restore emergency communication services for large-scale post-disaster users, a distributed intellicise coverage optimization architecture based on multi-agent reinforcement learning (RL) was proposed, which could address the significant differences and dynamics of communication services caused by a large number of access users, and the difficulty of expansion caused by centralized algorithms. Specifically, a distributed k-sums clustering algorithm considering service differences of users was designed in the network characterization layer, which could make each unmanned aerial vehicle base station (UAV-BS) adjust the local networking natively and simply, and obtain states of cluster center for multi-agent RL. In the trajectory control layer, multi-agent soft actor critic (MASAC) with distributed-training-distributed-execution structure was designed for UAV-BS to control trajectory as intelligent nodes. Furthermore, ensemble learning and curriculum learning were integrated to improve the stability and convergence speed of training process. The simulation results show that the proposed distributed k-sums algorithm is superior to the k-means in terms of average load efficiency and clustering balance, and MASAC based trajectory control algorithm can effectively reduce communication interruptions and improve the spectrum efficiency, which outperforms the existing RL algorithms.

**Keywords:** emergency communication, coverage optimization, multi-agent reinforcement learning, distributed training

**收稿日期**：2022−02−24；**修回日期**：2022−05−23

**基金项目**：国家重点研发计划基金资助项目（No.2019YFC1511302）；国家自然科学基金资助项目（No.61871057, No.61790553）；中央高校基本科研业务费专项资金资助项目（No.2019XD-A13）

**Foundation Items:** The National Key Research and Development Program of China (No.2019YFC1511302), The National Natural Science Foundation of China (No.61871057, No.61790553), The Fundamental Research Funds for the Central Universities (No.2019XD-A13)

## 0 引言

在发生重大自然灾害后，地面的基础通信设施通常会遭到毁坏而产生通信中断，重要的通信信息被阻绝，危及受灾用户的生命安全，加剧灾后救援的难度。无人机因为具有快速部署、灵活调控等优点，能够通过装备应急基站提供有效的空地视线线路（LoS, line of sight）覆盖受灾区域，在应急通信领域具有广泛的应用前景[1]。随着移动互联网和物联网技术的高速发展，大量数字化机器设备被应用于抢险救灾、智能医疗等应急服务，大量传感器和辅助装置被部署以对灾区状况进行持续监控[1]。因此，服务于 6G 的应急通信网络将面临更大规模、更高密度、更快速度的覆盖需求[2]，并且需要应对大规模用户接入带来的高动态性和未知业务类型[3]。为了应对 6G 背景带来的挑战，“节点极智、网络极简”的智简应急通信网络[4-5]应运而生。通过采取以通信计算融合[6]为代表的智能技术，网络中的节点将成为具备智能的“智慧内生”新型节点，而网络本身的协议结构将趋向于“原生简约”，基于内生智慧驱动打造通信链路与网络组织的随需动态重塑能力。智简应急通信网络将具备针对用户状态动态改变、实时调整网络部署，并根据用户业务差异按需调配网络资源的能力。

传统非智能化的应急通信网络常采用非凸优化方法提升覆盖性能，其中覆盖性能由无人机基站相对地面用户的实时位置主导，需要解决关于无人机基站飞行轨迹的非凸优化问题。Kang 等[7]对多无人机基站多用户的通信场景进行建模，利用迭代吉布斯采样和块坐标下降方法对多无人机基站的飞行轨迹进行联合优化，高效率地提升了网络的最大−最小速率。Yin 等[8]在大规模地面用户场景，利用连续凸逼近方法联合优化了地面分簇和多无人机基站的悬停位置，提升了网络的频谱效率。Zhang 等[9]针对应急通信场景的通信特征与需求，对多无人机基站的功率分配和轨迹优化问题联合建模，最大化应急通信网络的容量。然而，上述传统非智能化的覆盖优化方法需要全部精准的网络环境状态辅助（如用户位置、数据大小、信道状态等）作为待优化非凸问题中的固定参数，在求解过程中保持不变。因此，上述方法只适用于完全静态的网络场景，已知未来时刻的全部网络状态信息和所有用户的业务需求，难以应对大规模灾后用户的动态性与业务差异性。

智能化的深度强化学习方法被视为应对网络动态性的关键技术，配置深度强化学习智能体的无人机基站能够基于实时网络状态时序调控飞行轨迹，以最大化网络长期的性能收益。为了得到最优的覆盖优化策略，深度强化学习智能体需要迭代进行用于拟合动态网络环境的“训练阶段”和用于实时调控无人机基站飞行轨迹的“执行阶段”。不同“训练阶段”和“执行阶段”的实现方式，衍生出了多种基于深度强化学习的覆盖优化方法。文献[10]采用深度强化学习近端策略优化（PPO, proximal policy optimization）算法，提升了单无人机基站的通信速率并减小了飞行能耗。Liu 等[11]利用深度确定性策略梯度（DDPG, deep deterministic policy gradient）算法，在不考虑干扰的情况下对多无人机基站的部署进行了优化。然而，多无人机基站间存在干扰时，单智能体强化学习的学习环境非平稳导致算法难以收敛。为了解决上述问题，Challita 等[12]将博弈论融入回声状态网络（ESN, echo state network），联合优化了多无人机基站的飞行轨迹。不同于文献[12]中基于值函数的强化学习方法，文献[13-14]采用了多智能体深度确定性策略梯度（MADDPG, multi-agent deep deterministic policy gradient）算法，基于策略梯度对动作空间进行泛化，能够连续输出动作精准调控无人机飞行轨迹，避免了维度爆炸的问题[15]。然而，随着应急通信网络的规模增大，以“集中式训练−分布式执行”为框架的 MADDPG 算法的输入维度成倍增加，学习难度呈爆炸式增长，稳定性较差[16]，并且严重受固定架构下集中式训练中心处的灾情影响，难以处理大规模灾后用户的覆盖优化问题。

为了解决上述问题，本文提出了一种分布式智简的大规模灾后用户覆盖优化架构，网络特征层从用户需求本原出发拟合大规模灾后用户的业务差异性，按需重塑用户分簇组网结构，轨迹调控层利用多智能体强化学习技术赋予每个应急无人机基站智能化、分布式决策自身飞行轨迹的能力，提升应急通信网络的总体覆盖性能。本文的主要研究工作如下。

1) 基于多智能体强化学习技术，设计分布式智

简的大规模灾后用户覆盖优化架构。具体地，特征提取层通过自身获取的局部网络环境信息对地面用户执行分布式分簇组网，以特征化的簇中心用户信息作为状态输入多智能体强化学习神经网络，使轨迹调控层能够以小规模维度的状态调控无人机基站的实时轨迹。

2) 提出考虑用户业务差异性的分布式 k-sums 分簇算法，特征化大规模灾后用户状态。首先利用贝叶斯推理在线学习用户的业务差异性，获取用户的传输优先系数。进一步，无人机基站结合局部可获取用户的优先系数和负载信息执行分布式分簇，筛选获取簇中心用户。相比传统分簇算法，分布式 k-sums 分簇算法在平均负载效率和簇间均衡性方面均有性能提升。

3) 提出多智能体最大熵强化学习（MASAC, multi-agent soft actor critic）算法，用于多无人机基站分布式调控自身飞行轨迹。MASAC 以“分布式训练−分布式执行”的框架，融合最大熵理论和集成学习、课程学习技术，改进了现有多智能体深度强化学习方法不稳定、受灾情影响严重的问题，显著降低了应急通信网络的通信中断频率，提升了网络的频谱利用效率。

## 1　系统模型与架构设计

如图 1 所示，受灾区域内存在大规模具有动态性和业务差异性的地面用户，应急通信网络通过部署多架无人机基站接收地面用户的通信信息。假设受灾区域共有 $N$ 个用户，部署了 $M$ 架无人机基站，用户被分为 $M$ 个用户簇分别由各无人机基站恢复通信服务。其中，每个用户簇有一个簇中心用户与无人机基站直接相连，簇内其他用户的信息则会通过簇中心用户转发。用户集合与无人机基站集合分别用 $\mathcal{N}$ 和 $\mathcal{M}$ 表示。描述本文系统环境和覆盖优化算法的参数如表 1 所示。

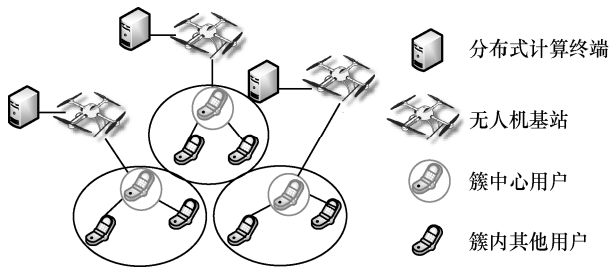



图 1　应急通信网络系统模型

**表 1　系统和算法参数**

| 参数 | 含义 |
|---|---|
| $N, M$ | 受灾用户、无人机基站的数量 |
| $N_j, M_j$ | 无人机基站 $j$ 可观测的用户、无人机基站数量 |
| $\kappa_1, \kappa_2$ | 用户激活状态 Beta 分布的参数 |
| $\mu_i, \sigma_i$ | 用户 $i$ 的传输业务类型分布的均值与标准差 |
| $u_i, u_j$ | 用户 $i$ 、无人机基站 $j$ 对应的簇中心用户 |
| $f_c$ | 地面通信或空地通信的中心频率 |
| $p_i, p_u, p_j$ | 用户 $i$ 、簇中心用户 $u$ 、无人机基站 $j$ 的位置 |
| $P_1, P_2$ | 簇内用户、簇中心用户的传输功率 |
| $L, G$ | 路径损耗、信道增益 |
| $N_0$ | 接收端的噪声功率 |
| $d_i, D_i$ | 用户 $i$ 的新传输任务、总传输任务的大小 |
| $R_i, R_j$ | 用户 $i$ 、无人机基站 $j$ 的频谱效率 |
| $B$ | 地面通信资源块的带宽 |
| $n_{i,u}$ | 用户 $i$ 占用的资源块数量 |
| $N_c$ | 簇内用户负载阈值 |
| $\eta$ | 平均负载效率 |
| $t_0$ | 历史观测数据集存储的帧数 |
| $\lambda$ | 用户优先参数 |
| $\hat{\mu}, \hat{\sigma}$ | 用户优先参数分布的均值、标准差 |
| $y_{i,j}$ | 用户 $i$ 是否处于簇 $j$ 的分簇标识 |
| $g_{i_1,i_2}$ | 用户 $i_1$ 对于用户 $i_2$ 的不相似性度量 |
| $s_t, a_t, r_t$ | 强化学习状态、动作、奖励函数 |
| $\gamma$ | 折扣因子 |
| $\theta$ | 深度神经网络参数 |
| $\alpha$ | 最大熵强化学习温度因子 |
| $\eta$ | 学习步长 |
| $W$ | 集成学习神经网络的组数 |

在大规模应急通信网络中，用户的信息汇聚传输采用簇中心用户做信息转发的优势，在于处理能力、能量损耗和干扰强度 3 个方面。其一，无人机基站的处理能力有限，通过用户分簇能够减少与无人机基站直接相连用户的数目，并有效降低神经网络的维度，避免网络陷于瘫痪；其二，通过减少无人机基站直接接入的用户数目，减少无人机基站的通信能耗和计算能耗，增加无人机基站的持续运行时间；其三，通过用户分簇减少空地通信链路的数目，能够降低空地通信簇间干扰，提升网络整体的通信能力。

本节后续将分别对本文涉及的用户模型、地面传输模型、空地传输模型和覆盖优化架构设计进行详细描述。

### 1.1 用户模型

在真实复杂的应急通信网络环境中，大规模灾后用户呈现出明显的动态性与业务差异性。动态性体现在自身位置实时变化，激活状态具有时间随机性。如果用户在给定时刻处于激活状态，则有新传输任务。用户 $i$ 的激活状态在 $t\in[0,T]$ 时间内服从 Beta 分布

$$f_i(t)=\frac{t^{\kappa_1-1}(T-t)^{\kappa_2-1}}{T^{\kappa_1+\kappa_2-1}B(\kappa_1+\kappa_2)} \tag{1}$$

$$B(\kappa_1+\kappa_2)=\int_0^1 t^{\kappa_1-1}(1-t)^{\kappa_2-1}\mathrm{d}t \tag{2}$$

其中，$\kappa_1$ 和 $\kappa_2$ 是 Beta 分布的参数。值得注意的是，用户的激活状态仅与是否有新的传输任务有关，处于非激活状态的用户仍可以传输上一时刻未被传输的剩余数据，并可能被选为簇中心用户。用户被选为簇中心用户后需要负责转发簇内所有用户的信息和更高的发送功率，由于本文重点关注覆盖优化以恢复大规模用户通信，因此不对用户的能量均衡进行探讨。

用户的业务差异性体现在不同的通信业务服务对速率、时延、安全性等需求各异，本文主要考虑由于业务类型、任务需求不同引起的信息差异性，用户所需传输的数据大小存在明显差异。假设用户 $i$ 在激活时刻 $t$ 的新传输任务数据大小 $d_i(t)$ 服从高斯分布[17]

$$f_d\left(d_i(t)\right)=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{\left(d_i(t)-\mu_i\right)^2}{2\sigma_i^2}\right) \tag{3}$$

其中，$\mu_i$ 和 $\sigma_i$ 是描述用户 $i$ 业务类型传输任务大小的均值和标准差常数，不同时刻的 $d_i(t)$ 由于传输任务的语义变化而产生波动。

### 1.2 地面传输模型

地面大规模受灾用户被划分为 $M$ 个簇，簇数与无人机基站的数目相同，每个用户首先将数据传输至簇中心用户，通过簇中心用户转发将数据传输至无人机基站。用户 $i$ 与簇中心用户 $u_i$ 间的通信采用 sub-6 GHz 频段的地对地通信链接，其中非视距（NLoS, non line of sight）在该无线链路中占主导地位，路径损耗可以依据瑞利衰落信道模型表示为[18]

$$L_{i,u_i}^{\text{ground}}(\text{dB})=37.6\lg\|p_i-p_{u_i}\|+21\lg f_{\text{c}}^{\text{ground}}+58.8 \tag{4}$$

其中，$f_{\text{c}}^{\text{ground}}$ 代表地面通信的中心频率，$p_i$ 和 $p_{u_i}$ 代表用户 $i$ 和簇中心用户 $u_i$ 的位置，$\|p_i-p_{u_i}\|$ 代表用户间的欧氏距离，系数 37.6 和 21 分别代表在非高层建筑城市或郊区场景下路径损耗模型的距离衰减因子和频率衰减因子，常数项 58.8 是考虑了用户间高度差距的附加路径损耗常数。由于簇间用户的距离较远，通过合理的频谱资源分配技术，簇间用户的干扰可以忽略不计，本文不对频谱资源分配进行探讨。用户 $i$ 与簇中心用户 $u_i$ 通信链接的信干噪比（SINR, signal to interference plus noise ratio）可以表示为

$$\text{SINR}_{i,u_i}^{\text{ground}}=\frac{P_1G_{i,u_i}^{\text{ground}}}{N_0} \tag{5}$$

其中，$P_1$ 代表用户的发送功率，$G_{i,u_i}^{\text{ground}}$ 代表用户 $i$ 与簇中心用户 $u_i$ 之间的信道增益，$N_0$ 代表噪声功率。信道增益 $G_{i,u_i}^{\text{ground}}$ 受路径损耗影响，满足

$$P_1G_{i,u_i}^{\text{ground}}(\text{dB})=P_1(\text{dB})-L_{i,u_i}^{\text{ground}}(\text{dB}) \tag{6}$$

用户 $i$ 在时刻 $t$ 传输数据的频谱效率可以表示为

$$R_{i,u_i}(t)=\text{lb}\left(1+\text{SINR}_{i,u_i}^{\text{ground}}(t)\right) \tag{7}$$

用户 $i$ 在时刻 $t$ 的总传输任务大小用符号 $D_i(t)$ 表示，包含时刻 $(t-1)$ 的剩余传输任务大小和时刻 $t$ 的新传输任务大小 $d_i(t)$。规定在初始时刻无剩余传输任务，即 $D_i(-1)=0$，则有

$$D_i(t)=\max\left(0,D_i(t-1)-n_i(t-1)BR_{i,u_i}(t-1)+d_i(t)\right) \tag{8}$$

其中，$B$ 表示地面资源块的带宽大小；$n_i(t)$ 表示用户的负载资源块数目，由总传输任务大小和频谱效率决定

$$n_{i,u_i}(t)=\min\left(N_{\text{c}},\left\lceil\frac{D_i(t)}{R_{i,u_i}(t)}\right\rceil\right) \tag{9}$$

其中，$N_{\text{c}}$ 是资源块负载阈值，以防用户由于低频谱效率而占用过多的频谱资源块。定义评价指标平均负载效率为

$$\eta=\frac{1}{TN}\sum_{t=0}^{T}\sum_{i=0}^{N}\frac{BR_{i,u_i}(t)}{n_{i,u_i}(t)} \tag{10}$$

平均负载效率 $\eta$ 可以有效地评价不同用户动

态性和信息差异性情况下的地面分簇结果。

### 1.3　空地传输模型

应急无人机基站与簇中心用户间的通信采用 sub-6 GHz 频段的空对地通信链接，其中 LoS 在该无线链路中占主导地位。无人机基站 $j$ 与簇中心用户 $u_j$ 间的平均路径损耗可以表示为

$$L_{j,u_j}^{\text{air}}(\text{dB}) = 20\lg\left(\frac{4\pi f_{\text{c}}^{\text{air}} \parallel p_j - p_{u_j} \parallel}{c}\right) + \eta_{\text{LoS}} \quad (11)$$

其中，$f_{\text{c}}^{\text{air}}$ 代表空地通信的中心频率；$p_j$ 代表无人机基站的位置；$c$ 代表光速；$\eta_{\text{LoS}}$ 代表 LoS 的附加空间传播损耗，是一个常量。簇中心用户会对其他无人机基站产生干扰，无人机基站 $j$ 与服务的簇中心用户 $u_j$ 间通信链路的信干噪比为

$$\text{SINR}_j^{\text{air}} = \frac{P_2 G_{j,u_j}^{\text{air}}}{N_0 + \sum\limits_{j' \neq j, j' \in \mathcal{M}} P_2 G_{j,u_{j'}}^{\text{air}}} \quad (12)$$

其中，$P_2$ 代表簇中心用户的发送功率，$G_{j,u_j}^{\text{air}}$ 代表无人机基站 $j$ 与簇中心用户 $u_j$ 之间信道增益。信道增益 $G_{j,u_j}$ 受路径损耗影响，满足

$$P_2 G_{j,u_j}^{\text{air}}(\text{dB}) = P_2(\text{dB}) - L_{j,u_j}^{\text{air}}(\text{dB}) \quad (13)$$

无人机移动带来的多普勒效应可以用现有技术完美补偿，如锁相环技术。无人机基站 $j$ 的频谱效率可以表示为

$$R_j(t) = \text{lb}\left(1 + \text{SINR}_j^{\text{air}}(t)\right) \quad (14)$$

应急通信网络的平均频谱效率可以表示为

$$R(t) = \sum_{j \in \mathcal{M}} \text{lb}\left(1 + \text{SINR}_j^{\text{air}}(t)\right) \quad (15)$$

本文以式(15)的平均频谱效率为优化目标，在考虑无人机基站的最大飞行速度限制、飞行安全性限制和通信中断限制条件下，对优化问题建模

$$\begin{aligned} \mathcal{OP}: & \max_{p_j(t), j \in \mathcal{M}, t=1,\cdots,T} \sum_{t=1}^{T} \sum_{j \in \mathcal{M}} \text{lb}\left(1 + \text{SINR}_j^{\text{air}}(t)\right) \\ \text{s.t. } & \text{C}_1: \left\| p_j(t) - p_j(t+1) \right\| \leqslant V_{\max} \Delta t, \ \ \forall j \in \mathcal{M} \\ & \text{C}_2: \left\| p_j(t) - p_{j'}(t) \right\| > 0, \quad \forall j \neq j' \in \mathcal{M} \\ & \text{C}_3: P_{\text{outage}}(t) \leqslant P_{\text{outage}}^{\max} \end{aligned} \quad (16)$$

其中，$P_{\text{outage}}(t)$ 和 $P_{\text{outage}}^{\max}$ 分别表示时刻 $t$ 网络的通信中断概率和最大通信中断概率限制。优化问题中应急通信网络的平均频谱效率由各无人机基站和簇中心用户之间的信噪比决定，因为空地通信主要为直射路径，所以信噪比的大小由两者之间的距离主导；另一方面，通信中断限制条件 $\text{C}_3$ 也与地面用户分簇和簇中心用户的选择密切相关。因此，大规模多无人机应急通信网络中的轨迹调控问题依赖于地面用户分簇的结果，随着簇中心用户选择的动态变化而调整飞行轨迹。

### 1.4　覆盖优化架构

以上述用户模型和通信模型为基础，应急通信网络的平均频谱效率 $R(t)$ 与无人机基站的位置 $p_j$、簇中心用户的位置 $p_u$、地面分簇结果密切相关。基于此，本文设计了一种分布式智简的大规模灾后用户覆盖优化架构，由网络特征层和轨迹调控层两层结构组成，如图 2 所示。相比于传统的端到端的覆盖优化结构，本文设计的分层级联的覆盖优化结构优势在于：①通过降低无人机基站端的强化学习状态输入维度，降低深度神经网络的规模，减小问题训练的复杂度；②通过分层的设计，空中通信优化和地面通信优化两部分各司其职，在实际工程应用时方便针对性地调整性能与参数，是深度强化学习算法在各产业中落地的常用手段。

具体而言，每个无人机基站配置一个分布式计算终端服务于上述分层的优化架构。在网络特征层中，无人机基站利用局部获取的网络状态信息拟合大规模灾后用户的业务差异性，并依此独立地对局部用户进行分簇组网，筛选簇中心用户特征作为多智能体强化学习的输入状态。在轨迹调控层中，以少量无人机基站间的通信开销作为辅助，利用多智能体强化学习技术应对时序动态的状态输入，无人机基站能够以“分布式训练–分布式执行”的框架自主优化飞行轨迹，以减少通信中断的频率，并最大化网络的频谱效率。需要指出的是，每个时间帧内除了用户信息经簇中心用户中继的信息汇聚传输过程，还需要簇中心用户特征作为强化学习输入，以辅助通信开销的形式传输至无人机基站。

## 2　网络特征层–地面用户分簇

在网络特征层中，地面用户分簇和簇中心用户选择需要应对大规模用户的业务差异性，本节提出一种基于贝叶斯推理的用户差异性学习算法。由于无人机基站难以获取全部大规模用户的信息，因此本节进一步提出了考虑用户差异性的分布式 k-sums 分簇算法，得到平均负载效率更高、簇间数目更均衡的分簇结果。

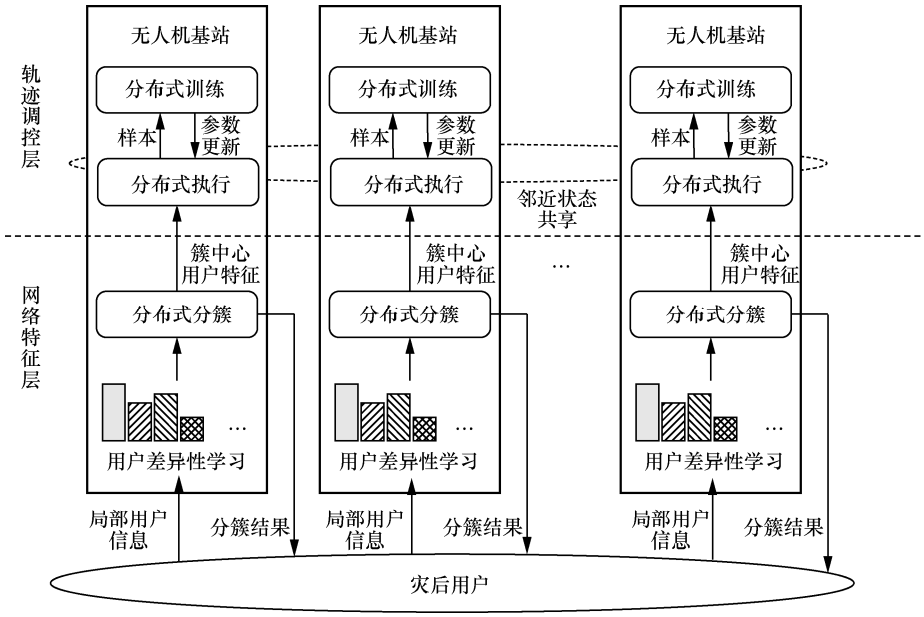



图 2 面向大规模灾后用户的分布式智简覆盖优化架构

### 2.1 用户差异性学习

贝叶斯推理是一种统计机器学习方法，基于贝叶斯公式建立观测量与估计量之间的联系[19]。在用户差异性学习过程中，无人机基站能够获取用户的最近 $t_0$ 帧激活时刻的新任务大小作为观测量 $d_i^*$，对用户优先参数 $\lambda_i$ 进行估计。本文以流量需求大小评价用户业务类型的优先级，其中，优先参数 $\lambda_i$ 表示用户 $i$ 由信息差异性引起的平均流量需求大小在 $[1,\lambda_{\max}]$ 之间的数值表征，旨在为优先级更高的用户分配更高质量的频谱资源。$\lambda_i$ 服从高斯分布，均值和方差分别为 $\hat{\mu}_i$ 和 $\hat{\sigma}_i^2$。假设无人机基站 $j$ 可观测的局部用户数目为 $N_j$，用集合 $\mathcal{N}_j$ 表示，定义向量 $\boldsymbol{d}^*=[d_0^*,d_1^*,\cdots,d_{N_j-1}^*]$，$\boldsymbol{\lambda}=[\lambda_0,\lambda_1,\cdots,\lambda_{N_j-1}]$，$\hat{\boldsymbol{\mu}}=[\hat{\mu}_0,\hat{\mu}_1,\cdots,\hat{\mu}_{N_j-1}]$，$\hat{\boldsymbol{\sigma}}^2=[\hat{\sigma}_0^2,\hat{\sigma}_1^2,\cdots,\hat{\sigma}_{N_j-1}^2]$，其中，$\boldsymbol{d}^*$ 是观测向量，$\boldsymbol{\lambda}\sim\mathcal{N}^{N_j}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\sigma}}^2)$ 是估计向量，$\hat{\boldsymbol{\mu}}$ 和 $\hat{\boldsymbol{\sigma}}^2$ 是参数向量。

贝叶斯推理流程如图 3 所示。首先从先验分布 $\mathcal{N}^{N_j}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\sigma}}^2)$ 中采样获取与观测向量维度数目相同的估计向量 $\boldsymbol{\lambda}$，采样概率 $P(\boldsymbol{\lambda})$ 即先验概率。依据向量 $\boldsymbol{d}^*$ 和 $\boldsymbol{\lambda}$，计算损失函数

$$C(\boldsymbol{d}^*\mid\boldsymbol{\lambda})=-\frac{1}{N_j}\sum_{i=0}^{N_j}\frac{(d_i^*-\lambda_i)^2}{d_i^*\lambda_i} \tag{17}$$

其中，$C(\boldsymbol{d}^*\mid\boldsymbol{\lambda})\in(-\infty,0]$。估计向量 $\boldsymbol{\lambda}$ 对观测向量 $\boldsymbol{d}^*$ 的似然函数可以通过对损失函数进行归一化得到

$$P(\boldsymbol{d}^*\mid\boldsymbol{\lambda})=\frac{\mathrm{e}^{-C(\boldsymbol{d}^*\mid\boldsymbol{\lambda})}}{\int\mathrm{e}^{-C(\boldsymbol{d}^*\mid\boldsymbol{\lambda})}\mathrm{d}\boldsymbol{d}^*} \tag{18}$$

$$P(\boldsymbol{\lambda}\mid\boldsymbol{d}^*)\propto P(\boldsymbol{d}^*\mid\boldsymbol{\lambda})P(\boldsymbol{\lambda}) \tag{19}$$

基于式(19)，对先验概率和似然函数的乘积进行归一化，得到估计向量 $\boldsymbol{\lambda}$ 的后验概率 $\omega$，并更新先验分布的均值与方差

$$\hat{\boldsymbol{\mu}}=\sum_{t=1}^{t_0}\boldsymbol{\lambda}(t)\boldsymbol{\omega} \tag{20}$$

$$\hat{\boldsymbol{\sigma}}^2=\sum_{t=1}^{t_0}\left(\boldsymbol{\lambda}(t)-\hat{\boldsymbol{\mu}}\right)^2\boldsymbol{\omega} \tag{21}$$

基于贝叶斯推理的用户差异性学习算法如算法 1 所示。

**算法 1** 基于贝叶斯推理的用户差异性学习算法

**输入** 观测向量 $\boldsymbol{d}^*$、待优化参数向量 $\hat{\boldsymbol{\mu}}$ 和 $\hat{\boldsymbol{\sigma}}^2$

**输出** 优化后的参数向量 $\hat{\boldsymbol{\mu}}$ 和 $\hat{\boldsymbol{\sigma}}^2$

1) 从先验分布 $\boldsymbol{\lambda}\sim\mathcal{N}^{N_j}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\sigma}}^2)$ 中采取 $t_0$ 组用户优先参数 $\lambda(1),\lambda(2),\cdots,\lambda(t_0)$

2) for $t=1:1:t_0$

3) 根据式(17)计算损失函数 $C\left(\boldsymbol{d}^*(t)\mid\boldsymbol{\lambda}(t)\right)$，表征观测向量与采样的优先参数向量间的差距

4) 根据式(18)对损失函数归一化操作，得到似然函数 $P\left(\boldsymbol{d}^*(t)\mid\boldsymbol{\lambda}(t)\right)$

5) 根据式(19)依照贝叶斯推理计算损失函

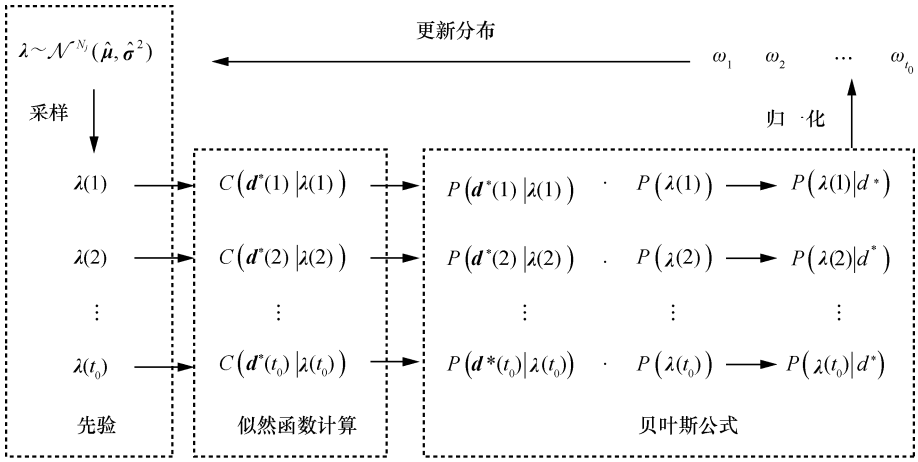



图 3　贝叶斯推理流程

数与似然函数的乘积

6) end for

7) 对所有损失函数与似然函数的乘积进行归一化，得到后验概率 $\omega_1, \omega_2, \cdots, \omega_{t_0}$

8) 根据式(20)、式(21)以后验分布更新待优化参数 $\hat{\boldsymbol{\mu}}$ 和 $\hat{\boldsymbol{\sigma}}^2$

通过算法 1 可以得到每一个用户的优先参数 $\lambda$ 的分布，分簇时按需为存在差异性的用户提供通信服务，通过优先提升 $\lambda$ 更高用户的频谱效率，能够有效减小网络频谱资源块负载。

### 2.2　地面用户分簇

相比于传统的 k-means 算法和谱聚类算法，k-sums 算法[20]具有更低的算法复杂度 $(O(NM))$，在分簇与簇中心用户快速变化时能够高效地执行分簇。同时，k-sums 算法可以有效降低簇内距离并提升簇间用户数目的均衡性。簇内距离和簇间均衡性是评价 k-sums 算法性能的重要评价标准，其中簇内距离与应急通信网络用户间的平均频谱效率性能密切相关，而簇间均衡性与不同无人机基站服务之间的通信负载均衡性能密切相关。综上所述，k-sums 算法能够高效地应对大规模灾后用户的动态性和差异性导致的分簇与簇中心用户快速变化。聚类算法的通用矩阵表达式为

$$\min_{\boldsymbol{Y}\in\mathbb{R}^{N\times M}} \mathrm{Tr}\left(\left(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\right)^{-\frac{1}{2}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{Y}\left(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\right)^{-\frac{1}{2}}\right) \tag{22}$$

其中，矩阵 $\boldsymbol{Y}$ 表示分簇标识矩阵，维度为 $\mathbb{R}^{N\times M}$，当用户 $i$ 处于无人机基站 $j$ 的服务簇内时元素 $y_{i,j}=1$，反之 $y_{i,j}=0$；矩阵 $\boldsymbol{G}$ 表示分簇核矩阵，对于不同的分簇算法，矩阵 $\boldsymbol{G}$ 的定义不同，k-sums 算法采用节点间的邻近不相似性度量，用户 $i_1$ 和用户 $i_2$ 的相似性越小，元素 $g_{i_1,i_2}$ 越大，且仅保留 $N_j$ 个 $g_{i_1,i_2}$ 最小的元素，其他元素用最大不相似性常数替代；运算符 Tr(·) 是矩阵的求迹操作。k-sums 算法为了保证分簇结果的均衡性，对问题式(22)增加限制条件 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\overline{n}\boldsymbol{I}$，其中，$\boldsymbol{I}$ 是单位矩阵，$\overline{n}$ 是任意常数。问题式(22)可以转化为

$$\begin{aligned}&\min_{\boldsymbol{Y}\in\mathbb{R}^{N\times M}} \mathrm{Tr}\left(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{Y}\right)\\ &\text{s.t. } \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\overline{n}\boldsymbol{I}\end{aligned} \tag{23}$$

然而，面向大规模灾后用户，单个无人机基站难以获取全局用户的信息，因此无法计算全局用户间的不相似性度量。若仍采用集中式的分簇方法，会产生大量用户信息的通信开销，因此本文提出分布式的 k-sums 分簇算法，使无人机基站仅利用局部观测信息对大规模灾后用户进行分布式分簇。

分布式的 k-sums 算法的分簇核矩阵 $\boldsymbol{G}$ 采用可观测用户的邻近不相似性度量表示，无人机基站 $j$ 的分簇核矩阵维度为 $\mathbb{R}^{N_j\times N_j}$。而用户之间的不相似度量则用当前时刻用户 $i_1$ 传输至用户 $i_2$ 所需负载资源块数目 $n_{i_1,i_2}$ 与用户优先参数 $\lambda_{i_1}$ 的乘积表征，即

$$g_{i_1,i_2}=n_{i_1,i_2}\lambda_{i_1} \tag{24}$$

如此设计，旨在同时考虑用户传输信息流量需求大小的瞬时特征和长期特征，为存在信息差异性的用户按需分配负载资源块，为业务需求更高的用户提供更优质的资源块，在负载有限的情况下有效降低高优先级用户通信无法被覆盖的概率。值得注意的是，本文分簇核矩阵的设计主要考虑了用户流

量需求差异表现的信息差异性；如果需要考虑其他通信需求差异引起的业务差异性，则需要针对性地改变分簇核矩阵元素的物理意义与之对应。

对于每个无人机基站，分布式的 k-sums 分簇算法仅需得到自身服务的用户簇，因此定义局部分簇标识矩阵 $\boldsymbol{Y}_p \subseteq \mathbb{R}^{N_j \times 2}$，其中 $y_{i,0}$ 表示用户 $i$ 是否处于无人机基站服务的用户簇 $\mathcal{N}_j$ 内。为保证分簇结果用户的均衡性，满足问题式(23)的条件，对于矩阵 $\boldsymbol{Y}_p$ 的元素，有

$$y_{i,0}=\begin{cases}1\ , & i\in\mathcal{N}_j\\ 0\ , & i\notin\mathcal{N}_j\end{cases} \tag{25}$$

$$y_{i,1}=\begin{cases}1 & , \ i\in\mathcal{N}_j\\ \sqrt{\dfrac{N}{M(M-1)}} & , \ i\notin\mathcal{N}_j\end{cases} \tag{26}$$

使局部分簇标识矩阵能够满足全局分簇标识矩阵的限制条件 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\bar{n}\boldsymbol{I}$。此外，无人机基站的可观测用户数目 $N_j$ 需要大于无人机基站服务用户的平均值，即 $N_j>\dfrac{N}{M}$。类似于 k-sums 算法的行迭代方法[20]，依次优化每一个用户的局部分簇标识行向量 $\boldsymbol{y}_i=[y_{i,0},y_{i,1}]$，对于每一个行向量，问题式(23)可以转化为

$$\min_{\boldsymbol{y}_i}\mathrm{Tr}\left(\boldsymbol{Y}_p^{\mathrm{T}}\boldsymbol{G}\boldsymbol{Y}\right)\Leftrightarrow\min_{\boldsymbol{y}_i}\boldsymbol{y}_i^{\mathrm{T}}\tilde{\boldsymbol{Y}}_p^{\mathrm{T}}\boldsymbol{g}_i \tag{27}$$

其中，$\tilde{\boldsymbol{Y}}_p^{\mathrm{T}}$ 是优化前的局部分簇标识矩阵，$\boldsymbol{g}_i$ 是不相似性度量矩阵 $\boldsymbol{G}$ 的列向量。考虑用户差异性的分布式 k-sums 分簇算法如算法 2 所示。

**算法 2** 考虑用户差异性的分布式 k-sums 分簇算法

**输入** 用户的不相似性度量矩阵 $\boldsymbol{G}$，优化前的局部分簇标识矩阵 $\tilde{\boldsymbol{Y}}_p$

**输出** 优化后的局部分簇标识矩阵 $\boldsymbol{Y}_p$

1) 初始化 $\boldsymbol{Y}_p$ 和 $\tilde{\boldsymbol{Y}}_p$，使之互不相同
2) while $\tilde{\boldsymbol{Y}}_p\neq\boldsymbol{Y}_p$
3) $\tilde{\boldsymbol{Y}}_p\leftarrow\boldsymbol{Y}_p$
4) for $i=1:1:N_j$
5) 根据式(27)对分簇标识矩阵 $\boldsymbol{Y}_p$ 进行行迭代优化，得到优化后的 $y_i$
6) end for
7) end while

通过算法 2 的计算结果 $\boldsymbol{Y}_p$，筛选使 $y_{i,0}=1$ 的用户作为无人机基站 $j$ 服务的用户，并选择不相似性度量最小的用户作为簇中心用户，即

$$\min_{i_2\in\mathcal{N}_j,y_{i_2,0}=1}\sum_{i_1\in\mathcal{N}_j}g_{i_1,i_2} \tag{28}$$

基于簇中心用户的特征信息，无人机基站可以实时调整飞行轨迹以优化对地面用户的覆盖，本文将在第 3 节进行深入探讨。

### 2.3 复杂度分析

标准的 k-means 算法需要迭代进行，分配用户到距离最近的簇中心用户、重新计算每个用户簇的分簇中心用户，因此需要计算每个用户到所有分簇中心用户的距离，复杂度为 $O(NM)$。然而标准的 k-means 算法适用范围较窄，只能处理线性可分的数据，并且聚类结果受初始化影响较大。改进的 k-means 算法为了处理非线性可分的数据类型，首先将输入数据非线性地映射至高维空间，然后执行 k-means 算法，计算复杂度为 $O(N^2)$。谱聚类分簇算法使用了用户的近邻图来进行分析，可以处理非线性可分数据，有着更加出色的聚类性能，但是由于先构建邻近图再进行谱分解的操作，计算复杂度较高，达到了 $O(N^2M)$。k-sums 算法的分簇核矩阵采用了邻近不相似性度量，$g_i$ 中大部分取值为相同常数，利用行迭代优化方法计算式(27)的复杂度约为 $O(M)$，算法总体的计算复杂度为 $O(NM)$。相比于 k-sums 算法，本文提出的分布式 k-sums 算法采用了可观测用户的邻近不相似性度量表征分簇核矩阵的元素，矩阵维度由 $\mathbb{R}^{N\times N}$ 降为 $\mathbb{R}^{N_j\times N_j}$，局部分簇标识矩阵的维度也由 $\mathbb{R}^{N\times M}$ 降为 $\mathbb{R}^{N_j\times 2}$，分布式 k-sums 算法的计算复杂度为 $O(2N_j)$。

另一方面，为了在线学习用户的业务差异性，贝叶斯推理算法需要执行 $t_0$ 步计算损失函数 $C\left(\boldsymbol{d}^*(t)\,|\,\boldsymbol{\lambda}(t)\right)$ 和似然函数 $P\left(\boldsymbol{d}^*(t)\,|\,\boldsymbol{\lambda}(t)\right)$ 的操作，其中损失函数的计算复杂度与可观测的局部用户数目 $N_j$ 有关，因此基于贝叶斯推理的用户差异性学习算法总体的计算复杂度为 $O(t_0N_j)$。综上所述，网络特征层，即考虑用户差异性的地面用户分簇的整体复杂度为 $O(t_0N_j)$。

## 3　轨迹调控层−无人机基站调控

传统的无人机基站轨迹优化方法无法处理大规模用户的动态性和长时间维度，而基于单智能体强化学习的调控方法难以应对多架无人机基站导致的非平稳学习环境。基于多智能体强化学习的优化方法可以基于当前时刻的网络环境状态智能决策飞行轨迹，有效解决上述问题。本文提出了一种多智能体最大熵强化学习 MASAC 算法，比现有的多智能体强化学习 MADDPG 算法具有更好的收敛性和稳定性。

### 3.1　基于多智能体强化学习的无人机基站分布式调控设计

针对大规模灾后用户的覆盖优化问题，1.4 节设计了分布式智简的覆盖优化架构，其中网络特征层负责对大规模地面用户进行分簇，筛选簇中心用户的特征信息，作为多智能体强化学习状态输入轨迹调控层。轨迹调控层采用多智能体深度强化学习的方法，用马尔可夫决策过程对轨迹调控问题进行重新建模，将全局优化问题转化为在每一个时刻的强化学习优化目标，基于奖励函数、价值函数的设计能够时序差分渐进地调控无人机基站的飞行轨迹，实现网络频谱效率最大化。因此，基于多智能体强化学习的无人机基站分布式调控设计具体如下。

状态。每个无人机基站提取部分可观测信息作为输入状态，可以特征化为：1) 无人机基站自身的坐标；2) 与地面分簇中心用户的二维相对位置；3) 接收分簇中心用户信息的信噪比大小；4) 与 $M_j$ 个邻近无人机的三维相对位置。

动作。考虑无人机基站在三维空间内可以自由移动，无人机基站的输出动作可以特征化为 $x$ 轴、$y$ 轴、$z$ 轴 3 个方向上的移动速度。

奖励。奖励函数由飞行安全惩罚值、通信中断惩罚值、频谱效率奖励 3 个部分构成，即

$$r_j = R_j(t) - \xi_{\text{collision}} I^j_{\text{collision}} - N_j P^j_{\text{outage}}(t)\xi_{\text{outage}} I^j_{\text{outage}} \tag{29}$$

其中，$P^j_{\text{outage}}(t)$ 和 $R_j(t)$ 分别是瞬时通信中断概率和瞬时网络频谱效率，$\xi_{\text{collision}}$ 和 $\xi_{\text{outage}}$ 是安全性和通信中断的附加惩罚常数，分别用于无人机基站 $j$ 发生碰撞或飞出指定区域 $I_{\text{collision}}=1$ 和发生通信中断 $I_{\text{outage}}=1$ 削减奖励函数的大小，在多智能体强化学习训练过程中减小上述事件发生的概率，更新无人机基站的飞行策略；$\xi_{\text{collision}}$ 和 $\xi_{\text{outage}}$ 是给定的超参数，在优化过程中固定不变。

通信。多智能体强化学习 MASAC 算法需要拟合邻近动作−状态价值函数，奖励在计算过程中也需要邻近无人机基站的通信信噪比与频谱利用效率，因此需要与 $M_j$ 个邻近无人机基站交互部分信息，包括：1) 无人机基站自身的坐标；2) 无人机基站的输出动作；3) 与地面分簇中心用户的二维相对位置；4) 接收分簇中心用户信息的信噪比大小；5) 当前时刻无人机基站的频谱效率。

本节后续将基于上述多智能体强化学习轨迹调控设计介绍本文提出的多智能体最大熵强化学习 MASAC 算法，以及融合集成学习、课程学习技术提升算法的训练稳定性和收敛速度。

### 3.2　多智能体最大熵强化学习 MASAC 算法

面对动态未知的应急通信网络环境，强化学习利用马尔可夫决策过程进行建模，从环境中获取观测值作为状态 $s_t$，依据动作选择策略 $\pi(a_t \mid s_t)$ 输出动作 $a_t$ 调控无人机基站的飞行轨迹，执行动作获取环境交互、通信网络覆盖性能等反馈计算奖励函数 $r_t$，环境状态经过状态转移分布 $p_\pi(s_{t+1} \mid s_t, a_t)$ 转换到下一时刻状态 $s_{t+1}$。强化学习智能体的动作选择策略与状态−动作价值函数 $Q(s_t, a_t)$ 密切相关，表征在状态 $s_t$ 下无人机基站选取动作 $a_t$ 收获的长期折扣累积奖励的期望值，即考虑了长期的应急通信网络覆盖性能。

$$Q(s_t, a_t) = r_t + \gamma \mathrm{E}_{s_{t+1}\sim p_s}\left[V(s_{t+1})\right] \tag{30}$$

其中，$V(s_t)$ 是状态价值函数，用于表征无人机基站从处于状态 $s_t$ 开始能够收益的长期应急通信网络覆盖性能奖励的期望值；$\gamma$ 是折扣因子，当 $0 \leqslant \gamma < 1$ 时，能够保证强化学习策略迭代的收敛性。状态价值函数为

$$V(s_t) = \mathrm{E}_{a_t\sim\pi}\left[Q(s_t, a_t) - \alpha \log \pi(a_t \mid s_t)\right] \tag{31}$$

其中，$\alpha \log \pi(a_t \mid s_t)$ 是熵正则化项。熵正则化项以最大熵强化学习算法[16]为理论基础，配合动作选择策略的优化过程，算法策略输出具有多模特性，可有效应对动态复杂的学习环境，提升算法收敛的稳定性。熵正则化项中的 $\alpha$ 为温度因子，可以通过自调节调整熵正则化项的影响权重。

当网络中存在多个智能体时，智能体 $i$ 仅可以获取局部观测值 $o^i_t$，且环境状态转移受多个智能体

的动作输出同时影响，环境状态转移分布变化为 $p_\pi(s_{t+1}\mid o_t^0,a_t^0,\cdots,o_t^i,a_t^i,\cdots)$，对于智能体 $i$ 学习环境处于非平稳状态，单智能体强化学习算法难以收敛。多智能体强化学习 MADDPG 算法通过获取其他智能体的观测值和输出动作，拟合全局的状态−价值函数 $Q(o_t^i,a_t^i,o_t^{-i},a_t^{-i})$，使智能体 $i$ 的学习环境平稳，其中 $-i$ 表示智能体 $i$ 以外的其他智能体。本文以最大熵强化学习 SAC 算法[21]与多智能体强化学习 MADDPG 算法[15]为基础，为 SAC 算法拟合邻近的状态价值函数，在保证算法收敛性的同时减小通信开销，使算法可以分布式部署。

如图 4 所示，每个 MASAC 智能体由 6 个神经网络与 1 个经验回放池构成。Actor 网络表征动作选择策略 $\pi_{\theta_1^i}$，$\boldsymbol{\theta}_1^i$ 是神经网络参数，输入局部观测状态 $o_t^i$，输出在观测状态下动作输出分布的均值 $\mu_{\theta_1^i}(o_t^i)$ 与标准差 $\sigma_{\theta_1^i}(o_t^i)$ 以表示动作选择策略 $\pi_{\theta_1^i}(a_t^i\mid o_t^i)$。Double Q 网络由 2 个神经网络（Critic1 网络和 Critic2 网络）组成，分别拟合邻近状态−价值函数 $Q_{\theta_2^i}$ 和 $Q_{\theta_3^i}$，神经网络参数分别为 $\boldsymbol{\theta}_2^i$ 和 $\boldsymbol{\theta}_3^i$。拟合 2 个状态−价值函数，可以解决单个 Critic 网络对状态−价值函数的过高估计[22]。Target 网络由 3 个神经网络（Target Actor 网络 $\hat{\pi}_{\theta_4^i}$、Target Critic1 网络 $\hat{Q}_{\theta_5^i}$ 和 Target Critic2 网络 $\hat{Q}_{\theta_6^i}$）构成，神经网络参数分别为 $\boldsymbol{\theta}_4^i$、$\boldsymbol{\theta}_5^i$ 和 $\boldsymbol{\theta}_6^i$。上述 3 个 Target 网络分别是 Actor 网络、Critic1 网络和 Critic 网络的副本网络，但参数更新速率更缓慢，能够提升训练过程的稳定性，加快算法的收敛速度。经验回放池用于记录智能体的样本 $<o_t^i,o_t^{-i},a_t^i,a_t^{-i},r_t^i,o_{t+1}^i,o_{t+1}^{-i}>$，其中，邻近智能体的信息通过相互间的通信获取。训练时智能体从经验回放池中采样，随机获取样本集 $\mathcal{D}$ 用于计算优化目标的梯度。

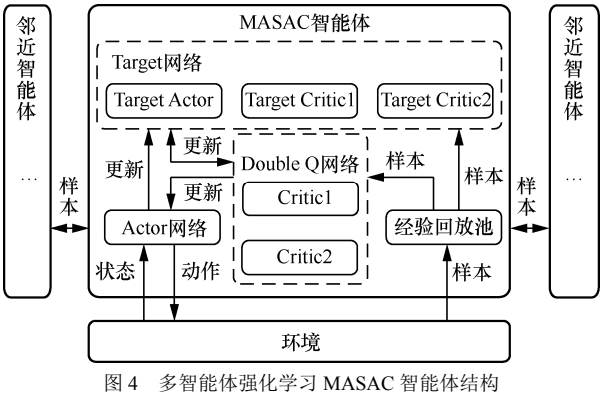



图 4 多智能体强化学习 MASAC 智能体结构

动作选择策略以最大化状态−动作价值函数为目标，因此 Actor 网络的优化目标可表示为

$$J_\pi(\boldsymbol{\theta}_1^i)=\mathrm{E}_{(o_t,a_t)\sim\mathcal{D}}[\alpha\log\pi_{\theta_1}(\hat{a}_t^i\mid o_t^i)-\min_{i=2,3}\mathrm{Q}_{\theta_i}(o_t^i,a_t^i,o_t^{-i},a_t^{-i})] \tag{32}$$

由于 Actor 网络的输出是分布函数而非具体的动作值，在计算优化目标梯度的过程中需要对输出动作数值化表示，因此采用了重参数技巧输出估计动作

$$\hat{a}_t^i=\tanh\left(\mu_{\theta_1}(o_t^i)+\sigma_{\theta_1}(o_t^i)\varepsilon_t^i\right) \tag{33}$$

其中，$\varepsilon_t^i$ 是均值为 0 且与动作输出策略独立的高斯噪声向量。Critic 网络以拟合状态−动作价值函数为目标，因此优化目标可以用时序差分误差表示为

$$J_Q(\boldsymbol{\theta}_{2,3}^i)=\mathrm{E}_{(o_t,a_t,r_t,o_{t+1})\sim\mathcal{D}}\Bigg[\frac{1}{2}\Big(Q_{\theta_{2,3}^i}(o_t^i,a_t^i,o_t^{-i},a_t^{-i})-\big(r_t+\gamma(\min_{j=5,6}\mathrm{Q}_{\hat{\theta}_j^i}(o_{t+1}^i,a_{t+1}^i,o_{t+1}^{-i},a_{t+1}^{-i})-\alpha\log(\hat{\pi}_{\theta_4^i}(a_{t+1}^i\mid o_{t+1}^i))\big)\Big)^2\Bigg]_{a_{t+1}\sim\hat{\pi}_{\theta_4^i}} \tag{34}$$

综合上述优化目标，网络参数更新为

$$\boldsymbol{\theta}_1^i\leftarrow\boldsymbol{\theta}_1^i+\eta_1\nabla_{\theta_1^i}\mathrm{J}_\pi(\boldsymbol{\theta}_1^i) \tag{35}$$

$$\boldsymbol{\theta}_{2,3}^i\leftarrow\boldsymbol{\theta}_{2,3}^i+\eta_{2,3}\nabla_{\theta_{2,3}^i}\mathrm{J}_\mathrm{Q}(\boldsymbol{\theta}_{2,3}^i) \tag{36}$$

$$\boldsymbol{\theta}_{4,5,6}^i\leftarrow\eta_{4,5,6}\boldsymbol{\theta}_{4,5,6}^i+(1-\eta_{4,5,6})\boldsymbol{\theta}_{1,2,3}^i \tag{37}$$

其中，$\eta$ 为神经网络更新步长。智能体通过迭代探索与训练过程，从环境中获取新样本存储于经验回放池、从经验回放池中随机获取批量样本根据式(35)～式(37)训练，使智能体学习到最优的动作输出策略。

### 3.3 集成学习与课程学习

多智能体强化学习算法能够有效地解决多智能体学习环境的非平稳问题，MASAC 算法能够使算法适应复杂动态的环境。然而，多智能体和最大熵强化学习算法都加剧了神经网络的复杂程度，因此，本文应用集成学习[23]和课程学习[24]技术提升算法收敛过程的速度和稳定性。

1) 基于集成学习的稳定收敛技术

本文融入了集成学习技术，自举训练多组神经网络，通过决策过程获取反馈，择劣剪枝、择优继承，避免了灾难性遗忘的影响，增加了算法收敛过程的稳定性。

图 5 详细描述了基于集成学习的稳定收敛技术的实现架构，每个无人机基站装载的智能体会同时训练 $W$ 组神经网络，形成集成学习神经网络集 $\mathcal{W}$。在“分布式训练”阶段，分别从经验回放池中取出 $W$

组独立的样本集 $\mathcal{D}_1,\mathcal{D}_2,\cdots,\mathcal{D}_W$，并训练 $\mathcal{W}$ 中的所有神经网络。在“分布式执行”阶段，智能体从 $\mathcal{W}$ 中随机采样获得一组神经网络 $w$ 决策无人机基站的动作，获取奖励 $r_m$，并更新神经网络 $w$ 的累积奖励 $r_m^{(w)}$

$$r_m^{(w)} = \tau_w r_m^{(w)} + (1-\tau_w) r_m \tag{38}$$

其中，$\tau_w$ 是神经网络的累积奖励的更新步长。

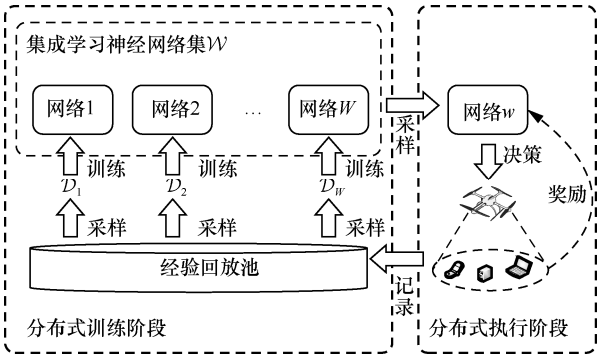



图 5　基于集成学习的稳定收敛技术的实现架构

进一步，更新神经网络集 $\mathcal{W}$ 中最大累积奖励 $r_m^{W\max}$

$$r_m^{W\max} = \max\left(r_m^{W\max}, r_m^{(w)}\right) \tag{39}$$

如果神经网络 $w$ 的累积奖励 $r_m^{(w)}$ 远小于神经网络集的最大累计奖励 $r_m^{W\max}$，则对神经网络 $w$ 采取剪枝操作，并复制 $\mathcal{W}$ 中剩余网络中累积奖励值最大的神经网络作为新的神经网络 $w$。

通过上述集成学习的设计，MASAC 智能体在训练过程中能够剪枝发生了导致巨额性能损失的灾难性遗忘的神经网络，并且择优选择神经网络继承，加速算法的收敛过程。

2) 基于课程学习的加速收敛技术

课程学习按照物理意义将学习任务从易到难划分为多个子任务，并由简入繁地设计每个子任务的奖励函数，降低学习难度，提升算法收敛速度。

运用课程学习的思想，如图 6 所示，将 3.1 节中的奖励函数由简及繁划分为以下 3 个子任务：1) 无人机基站保持飞行在一个固定的区域内；2) 无人机基站通过调整飞行轨迹减小通信服务中断发生，当无人机基站接收分簇中心用户信息的信噪比小于阈值时发生通信中断；3) 无人机基站通过进一步优化飞行轨迹最大化网络的频谱效率。因此，3 个子任务的奖励函数可以分别设计为

$$r_{\mathrm{A}} = -\xi_{\mathrm{collision}} I^j_{\mathrm{collision}} \tag{40}$$

$$r_{\mathrm{B}} = -\xi_{\mathrm{collision}} I^j_{\mathrm{collision}} - N_j P_{\mathrm{outage}}(t) \xi_{\mathrm{outage}} I^j_{\mathrm{outage}} \tag{41}$$

$$r_{\mathrm{C}} = R_j(t) - \xi_{\mathrm{collision}} I^j_{\mathrm{collision}} - N_j P_{\mathrm{outage}}(t) \xi_{\mathrm{outage}} I^j_{\mathrm{outage}} \tag{42}$$

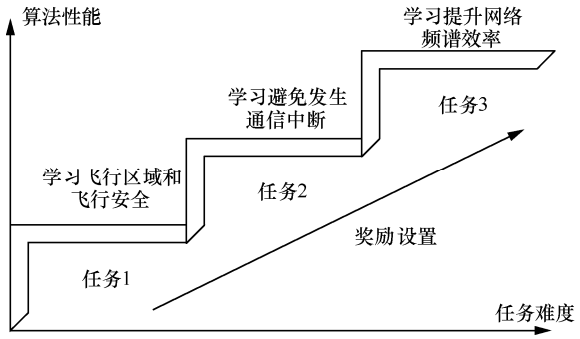



图 6　基于课程学习的加速收敛技术任务划分

值得注意的是，学习更难课程的内容可能会导致神经网络忘记简单课程的学习结果，从而引起灾难性遗忘。在更难课程的奖励设计中，需要包含简单课程的奖励，如式(41)和式(42)所示，并配合集成学习剪枝发生灾难性遗忘的子网络，消除灾难性遗忘的影响。

结合了集成学习、课程学习技术的基于 MASAC 的多无人机轨迹分布式调控算法如算法 3 所示。该算法能够有效降低网络的通信中断频率，最终实现网络频谱效率的提升。

**算法 3**　基于 MASAC 的多无人机轨迹分布式调控算法

1) while $t < T_0$
2) 从环境中获取观测状态 $o_t^i$
3) 从集成学习神经网络集 $\mathcal{W}$ 中随机采样一组神经网络，将观测状态 $o_t^i$ 输入 actor 网络输出动作 $a_t^i$
4) 执行动作与环境交互，获取当前时刻的分簇中心用户信息的信噪比和频谱效率
5) 与邻近无人机通信，计算当前课程任务的奖励大小，更新状态 $o_{t+1}^i$
6) 记录样本 $< o_t^i, o_t^{-i}, a_t^i, a_t^{-i}, r_t^i, o_{t+1}^i, o_{t+1}^{-i} >$ 于经验回放池
7) 根据式(38)和式(39)更新累积奖励 $r_m^{(w)}$ 和最大累积奖励 $r_m^{W\max}$，判断是否进入下一课程学习阶段
8) 如果 $r_m^{(w)}$ 远小于 $r_m^{W\max}$，对神经网络 $w$ 采取剪枝和继承操作
9) for $n = 1:1:W$
10) 从经验回放池中取出一批样本 $\mathcal{D}_n$
11) 根据式(35)～式(37)更新 MASAC 多智能体强化学习神经网络参数

12)　end for

13)　$t=t+1$

14) end while

### 3.4　复杂度分析

在“分布式执行”阶段，每个无人机基站需要获取自身的局部状态信息，并与邻近无人机基站共享，该过程与邻近无人机基站的数目 $M_j$ 呈正相关，因此，这一阶段算法的复杂度为 $O(M_j)$。

在“分布式训练”阶段，每个无人机基站需要更新集成学习神经网络集 $\mathcal{W}$ 中的全部 $W$ 个神经网络，每个神经网络的更新需要计算梯度的次数与从经验回放池中取出的批量样本数目成正比。假设样本数目为 $N_D$，那么，这一阶段算法的复杂度为 $O(WN_D)$。由于邻近无人机基站的数目 $M_j$ 远小于批量样本的数目 $N_D$，因此算法 3 的总体复杂度为 $O(WN_D)$。

### 3.5　面向大规模灾后用户的分布式覆盖优化流程

本文提出的分布式智简的覆盖优化架构可划分为网络特征层和轨迹调控层，其中网络特征层作为多智能体强化学习的特征提取阶段，由基于贝叶斯推理的用户差异性学习算法（算法 1）和考虑用户差异性的分布式 k-sums 算法（算法 2）共同实现，轨迹调控层作为多智能体强化学习的策略实现阶段，由基于 MASAC 的多无人机轨迹分布式调控算法（算法 3）实现。面向大规模灾后用户的分布式覆盖优化的总体流程如图 7 所示。

## 4　仿真分析

本节通过仿真实验评估所提出的基于多智能体强化学习的大规模灾后用户的空中覆盖架构与相应算法的有效性。仿真中应急通信网络系统和算法参数设置如表 2 所示。假设受灾地区在 1 km×1 km 的范围内存在 500 个地面用户，无人机基站的飞行高度变化范围是 100～1 000 m。MASAC 算法中 Actor 网络和 Critic 网络均采用三层全连接层作为隐层，隐层神经元数目分别为 512、256、128。本文在 Python3.7 平台上对所提的基于多智能体强化学习的大规模灾后用户分布式覆盖优化方案进行了性能验证，利用 Numpy 工具包实现了贝叶斯推理和分布式 k-sums 算法，利用 TensorFlow 工具包实现了多智能体强化学习 MASAC 算法，计算机环境为 Windows 10、Intel 7th CPU、GTX 1060。

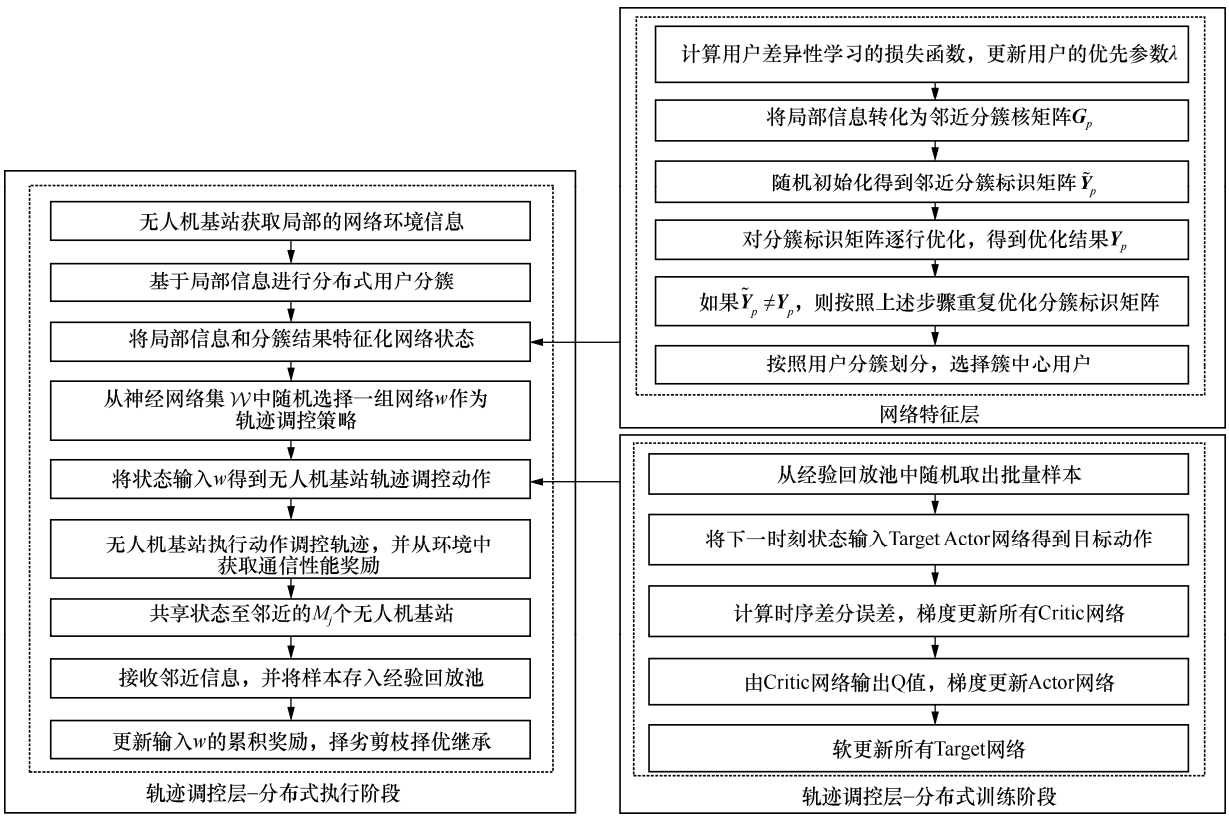



图 7　面向大规模灾后用户的分布式覆盖优化的总体流程

**表 2　应急通信网络系统和算法参数设置**

| 参数 | 取值 | 参数 | 取值 |
|---|---|---|---|
| 用户数目 $N$ | 500 | 观测用户数 $N_j$ | $\frac{2N}{M}$ |
| 邻近基站数 $M_j$ | 3 | 带宽 $B$ /MHz | 1 |
| 空地频率 $f_c^{\text{air}}$ /GHz | 4 | 地面频率 $f_c^{\text{ground}}$ /GHz | 2 |
| 噪声 $N_0$ /(dBm·Hz$^{-1}$) | −174 | 无人机最大速度/(m·s$^{-1}$) | 10 |
| 簇中心功率 $P_1$ /dBm | 100 | 用户功率 $P_2$ /dBm | 20 |
| 负载阈值 $N_c$ | 5 | 总任务时长 $T_0$ /s | 500 |
| 激活参数 $\kappa_1$ | 2 | 激活参数 $\kappa_2$ | 5 |
| 惩罚值 $\xi_{\text{collision}}$ | 500 | 惩罚值 $\xi_{\text{outage}}$ | 10 |
| 步长 $\eta_1$ | 0.001 | 步长 $\eta_{2,3}$ | 0.000 1 |
| 步长 $\eta_{4,5,6}$ | 0.001 | 集成学习维度 $W$ | 10 |
| 折扣因子 $\gamma$ | 0.99 | 经验回放样本数 | 128 |

首先验证底层优化考虑用户差异性的分布式 k-sums 分簇算法的有效性，在不同最大优先参数 $\lambda_{\max}$ 下进行仿真实验，并与 k-sums 算法和 k-means 算法进行对比。图 8 给出了不同分簇算法对簇间用户数目方差的影响。从图 8 中可以看出，所提分布式 k-sums 算法保持了 k-sums 算法的分簇均衡性，当不考虑用户的信息差异性，即 $\lambda_{\max}=1$ 时，分布式 k-sums 算法的簇间用户数量的方差大小与 k-sums 算法基本相同，远小于 k-means 算法；当最大优先参数 $\lambda_{\max}$ 增大时，所提算法由于更关心优先参数更大用户的性能，因此会牺牲一定分簇均衡性，簇间用户数量的方差会有所增大。

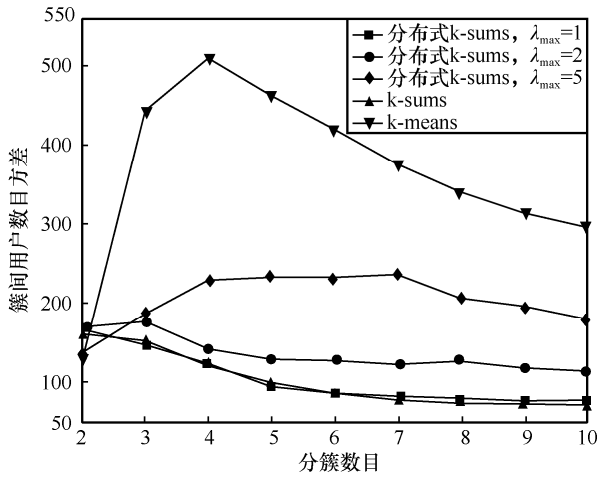



图 8　不同分簇算法对簇间用户数目方差的影响

图 9 给出了不同分簇算法对簇内用户平均负载效率的影响。从图 9 中可以看出，随着分簇数目的提升，平均簇内距离会减小，因此所有分簇算法的平均负载效率均显著提升。当不考虑用户的信息差异性，即 $\lambda_{\max}=1$ 时，所提分布式 k-sums 算法与 k-sums 算法的平均分簇效率相近，整体均好于 k-means 算法。随着最大优先系数 $\lambda_{\max}$ 的增加，通过贝叶斯推理可以学习到用户间的信息差异性，在计算不相似性度量时对优先系数更高的用户赋予更大的权重，从而使平均负载效率提升。综合上述仿真结果，本文通过增加最大优先系数 $\lambda_{\max}$，能够提升流量需求更高用户的通信效率，实现簇内平均负载效率的提升，这验证了所提算法能够有效适应不同优先级的业务。

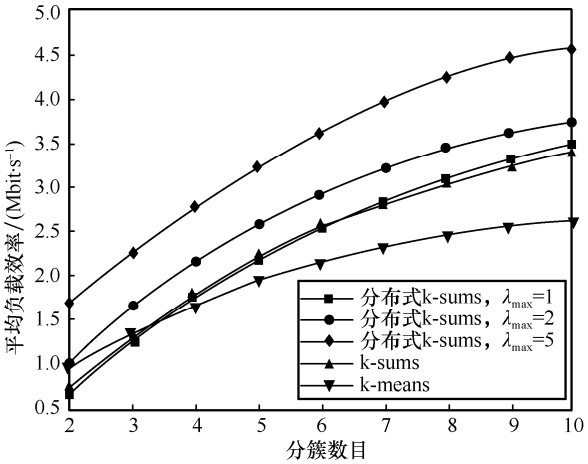



图 9　不同分簇算法对簇内用户平均负载效率的影响

进一步，对本文提出的基于多智能体强化学习的上层空中覆盖优化算法的有效性进行仿真验证。图 10 给出了 MASAC 算法平均累积奖励的收敛性能，在相同的仿真环境下，展示了集成学习和课程学习对 MASAC 收敛速率和稳定性的影响。平均累积奖励是衡量强化学习算法收敛的重要指标[25]，其表示在一个训练轮次内所有时隙得到奖励函数大小的平均值，具体的物理意义由奖励函数的设计决定，本文的平均累积奖励表示一个训练轮次内的平均频谱效率与平均通信中断惩罚、安全性惩罚之和。从图 10 中可以看出，集成学习和课程学习均可以提升算法的收敛速率。然而，集成学习对复杂任务直接学习，仅能收敛到性能一般的局部最优策略；课程学习在学习到任务 1 和任务 2 后会发生灾难性遗忘，收敛性能难以进一步提升。同时，结合集成学习和课程学习的 MASAC 算法能够以更快的收敛速度收敛到更优的策略，同时消除了灾难性遗忘的影响。

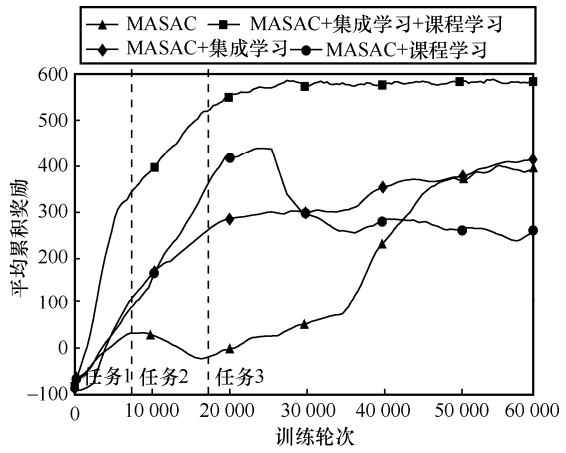



图 10 MASAC 算法平均累积奖励的收敛性能

图 11～图 14 给出了不同强化学习算法对无人机基站轨迹调控学习过程的影响，主要是将所提 MASAC 算法与 MADDPG 算法[13]和 DDPG 算法[11]进行对比。图 11 展示了不同强化学习算法平均累积奖励的收敛性能，图 12～图 14 分别展示了课程学习任务 1～任务 3 的关键指标的变化，即无人机基站飞出指定区域频率、通信中断频率、平均频谱效率。

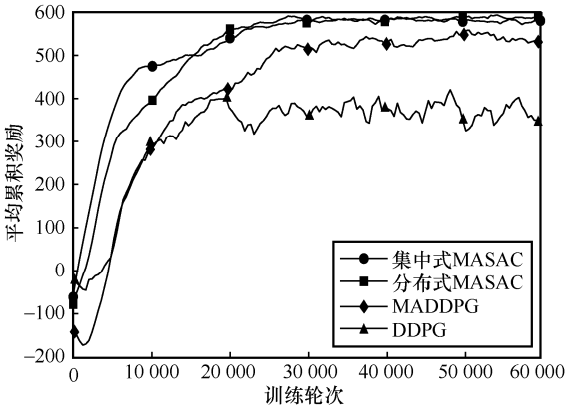



图 11 不同强化学习算法平均累积奖励的收敛性能

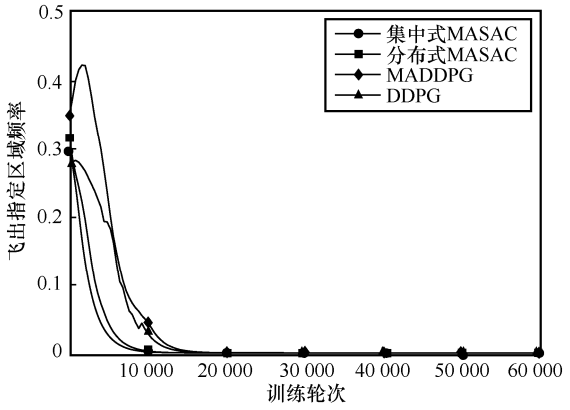



图 12 不同强化学习算法对任务 1–飞出指定区域频率的学习效果

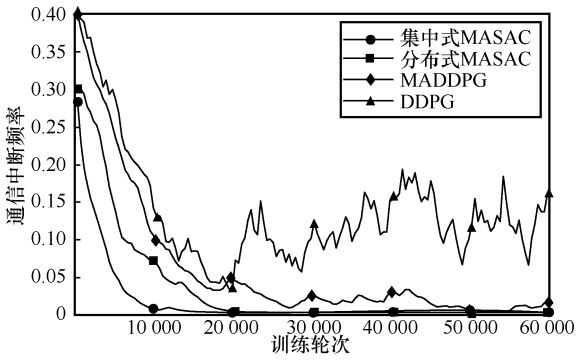



图 13 不同强化学习算法对任务 2–通信中断概率的学习效果

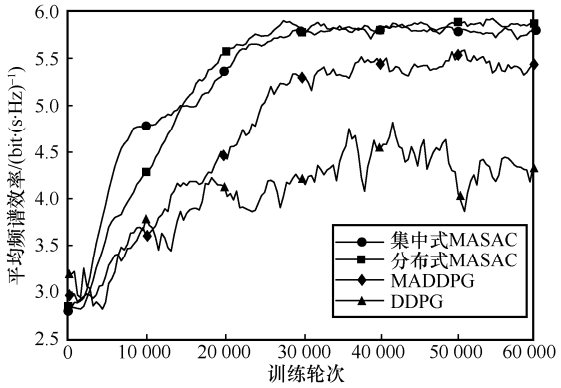



图 14 不同强化学习算法对任务 3–平均频谱效率的学习效果

从图 12～图 14 中可以看出，单智能体强化学习 DDPG 算法能够很快完成任务 1 的学习以飞行在限定的 1 km×1 km 区域内，而难以进一步完成任务 2 和任务 3 的学习。这是由于每个无人机基站飞行区域的策略学习不会影响其他无人机基站的飞行区域，学习环境平稳；而在任务 2 和任务 3 中，无人机基站飞行策略的改变会干扰其他无人机基站的通信，学习环境非平稳。对比多智能体强化学习 MASAC 算法和 MADDPG 算法，2 种算法均可以完成对任务 1 和任务 2 的学习，而 MADDPG 算法由于采用确定性策略算法，收敛性能和稳定性较差，对任务 3 频谱效率的学习效果不如 MASAC 算法。此外，仿真中对获取全局状态的集中式 MASAC 算法和获取邻近状态的分布式 MASAC 算法进行对比。可以看出，分布式 MASAC 算法能够收敛到和全局优化相同的效果，同时因为仅需要获取邻近无人机基站的状态，通信开销大幅减少。

图 15 给出了无人机基站数量对平均频谱效率的影响。从图 15 中可以看出，随着无人机基站数量的增加，学习环境的非平稳性和复杂程度增加，DDPG 和 MADDPG 算法的频谱效率随着无人机基站数量的增加而降低。而本文提出的 MASAC 算法

在无人机基站数量较小时可以通过联合调控无人机基站的飞行轨迹，得到更高的频谱效率，但是随着无人机基站数量的进一步增加，每个无人机基站会受到更多其他无人机基站的干扰，频谱效率下降。此外，对比集中式 MASAC 算法和分布式 MASAC 算法，分布式优化能够得到与全局优化相同的效果，甚至会在无人机数目较多时因为状态输入维度更低、神经网络规模更小而得到性能增益。

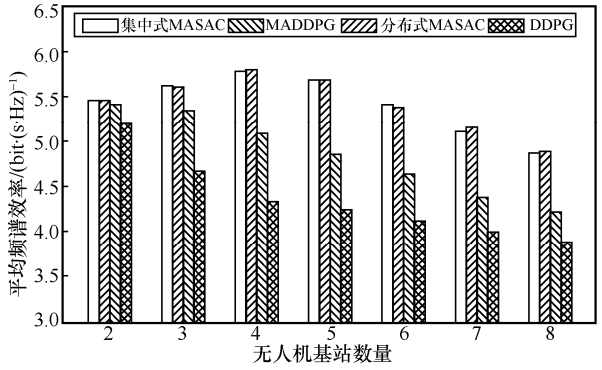



图 15　无人机基站数量对平均频谱效率的影响

## 5　结束语

本文针对大规模灾后用户应急通信恢复提出了分布式智简的空中覆盖优化架构。网络特征层执行用户分簇，并设计了考虑用户差异性的分布式 k-sums 分簇算法。轨迹调控层优化无人机基站飞行轨迹，并设计了基于多智能体强化学习 MASAC 的分布式轨迹调控算法，融合集成学习和课程学习技术提升了收敛速度和效果。由仿真结果可知，所设计的网络特征层算法能够应对用户的动态性和差异性，得到平均负载效率更高的分簇结果；本文所设计的轨迹优化层算法能够应对多无人机基站学习环境的非平稳性，利用邻近观测状态分布式优化各无人机基站的飞行轨迹，减小通信中断频率，提升频谱效率，实现应急网络覆盖性能优化。

本文的研究工作为恢复大规模灾后用户的通信覆盖提供了分布式智简的解决思路，但仍然存在一些局限性，未来的研究工作可以从以下 2 个方向入手：1) 所提算法受多超参数的影响，如邻近无人机基站的数量、无人机基站可观测的用户数量、无人机之间的相关性系数，这些超参数的取值基于规则给定，通过引入深度学习中的注意力机制等方法，上述超参数可以被进一步研究；2) 本文的研究重点聚焦于用户覆盖优化以快速恢复灾区通信，没有考虑实际应用中可能存在的其他问题，包括功率分配、能耗均衡等，未来可以在本文基础上进一步研究多优化目标相互耦合的综合性问题。

## 参考文献：

[1] DEEPAK G C, LADAS A, SAMBO Y A, et al. An overview of post-disaster emergency communication systems in the future networks[J]. IEEE Wireless Communications, 2019, 26(6): 132-139.

[2] GUO H Z, LI J Y, LIU J J, et al. A survey on space-air-ground-sea integrated network security in 6G[J]. IEEE Communications Surveys & Tutorials, 2022, 24(1): 53-87.

[3] ZHOU Y Q, LIU L, WANG L, et al. Service-aware 6G: an intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-260.

[4] ZHANG P, XU W J, GAO H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks[J]. Engineering, 2022, 8: 60-73.

[5] 张平, 许晓东, 韩书君, 等. 智简无线网络赋能行业应用[J]. 北京邮电大学学报, 2020, 43(6): 1-9.
ZHANG P, XU X D, HAN S J, et al. Entropy reduced mobile networks empowering industrial applications[J]. Journal of Beijing University of Posts and Telecommunications, 2020, 43(6): 1-9.

[6] ZHOU Y Q, TIAN L, LIU L, et al. Fog computing enabled future mobile communication networks: a convergence of communication and computing[J]. IEEE Communications Magazine, 2019, 57(5): 20-27.

[7] KANG Z Y, YOU C S, ZHANG R. 3D placement for multi-UAV relaying: an iterative Gibbs-sampling and block coordinate descent optimization approach[J]. IEEE Transactions on Communications, 2021, 69(3): 2047-2062.

[8] YIN S X, LI L H, YU F R. Resource allocation and basestation placement in downlink cellular networks assisted by multiple wireless powered UAVs[J]. IEEE Transactions on Vehicular Technology, 2020, 69(2): 2171-2184.

[9] ZHANG Y X, CHENG W C. Trajectory and power optimization for multi-UAV enabled emergency wireless communications networks[C]//Proceedings of International Conference on Communications Workshops. Piscataway: IEEE Press, 2019: 1-6.

[10] LI X, WANG Q, LIU J, et al. Trajectory design and generalization for UAV enabled networks: a deep reinforcement learning approach[C]// Proceedings of Wireless Communications and Networking Conference. Piscataway: IEEE Press, 2020: 1-6.

[11] LIU X, LIU Y W, CHEN Y. Reinforcement learning in multiple-UAV networks: deployment and movement design[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8): 8036-8049.

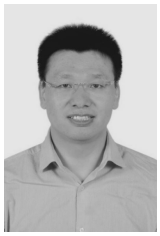
[12] CHALLITA U, SAAD W, BETTSTETTER C. Interference management for cellular-connected UAVs: a deep reinforcement learning approach[J]. IEEE Transactions on Wireless Communications, 2019, 18(4): 2125-2140.

[13] ZHAO N, LIU Z H, CHENG Y Q. Multi-agent deep reinforcement learning for trajectory design and power allocation in multi-UAV networks[J]. IEEE Access, 8: 139670-139679.

[14] QIN Z Q, LIU Z H, HAN G J, et al. Distributed UAV-BSs trajectory optimization for user-level fair communication service with mul-

ti-agent deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2021, 70(12): 12290-12301.
[15] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. arXiv Preprint, arXiv: 1706.02275, 2017.
[16] HAARNOJA T, TANG H R, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[J]. arXiv Preprint, arXiv: 1702.08165, 2017.
[17] 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.
[18] 3GPP. Technical specification group (TSG) RAN WG4; RF system scenarios: TR 25.942 v2.1.3[S]. 2000.
[19] WANG L T, SUN L T, TOMIZUKA M, et al. Socially-compatible behavior design of autonomous vehicles with verification on real human data[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 3421-3428.
[20] PEI S, NIE F, WANG R, et al. Efficient clustering based on a unified view of k-means and ratio-cut[J]. Advances in Neural Information Processing Systems, 2020, 33: 14855-14866.
[21] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning. New York: PMLR, 2018: 1861-1870.
[22] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 2094-2100.
[23] DONG X B, YU Z W, CAO W M, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14(2): 241-258.
[24] NARVEKAR S, PENG B, LEONETTI M, et al. Curriculum learning for reinforcement learning domains: a framework and survey[J]. arXiv Preprint, arXiv: 2003.04960, 2020.
[25] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Massachusetts: MIT Press, 1998.

[作者简介]



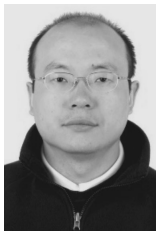
**许文俊**（1982− ），男，安徽安庆人，博士，北京邮电大学教授、博士生导师，主要研究方向为 B5G/6G 智能无线网络、语义智能通信网络、无人机通信及组网、认知无线网络等。

**吴思雷**（1997− ），男，北京人，北京邮电大学硕士生，主要研究方向为智能无线通信、分布式系统设计、机器学习、强化学习等。

**王凤玉**（1992− ），女，山西朔州人，博士，北京邮电大学讲师，主要研究方向为无线人工智能、通信感知一体化、智简通信等。

**林兰**（1996− ），女，河北衡水人，北京邮电大学博士生，主要研究方向为应急通信、NOMA、非凸优化方法、深度强化学习等。

**李国军**（1978− ），男，四川资阳人，博士，重庆邮电大学教授、博士生导师，主要研究方向为复杂恶劣环境超视距无线通信与网络。



**张治**（1977− ），男，河北安平人，博士，北京邮电大学副教授、硕士生导师，主要研究方向为移动通信、电子信号处理、通信系统设计等。