文章编号：2095-4980(2021)04-0573-08

基于多智能体强化学习的动态频谱分配方法

童　乐，梁　涛，张　余*，钱鹏智

(国防科技大学 第六十三研究所，江苏 南京 210007)

摘　要：针对认知无线电网络中多个异质用户具有不同的服务质量(QoS)要求，提出一种基于多智能体强化学习的动态频谱分配方法。该方法从用户满意度角度出发，以用户体验质量(QoE)作为系统的评价指标，构建多个虚拟智能体，模拟多个用户以合作方式与环境进行交互学习，融合各个用户的学习和频谱决策结果，实现频谱资源优化分配。仿真结果表明，在未知主要用户使用规律和信道动态特性条件下，相比基于传统强化学习的动态频谱分配方法，提出的方法能有效提高次用户的 QoE，降低用户间的冲突概率。

关键词：动态频谱分配；体验质量；多智能体强化学习；冲突概率

中图分类号：TN929.5　　**文献标志码：**A　　**doi：**10.11805/TKYDA2021172

Dynamic spectrum allocation method based on multi-agent reinforcement learning

TONG Le，LIANG Tao，ZHANG Yu*，QIAN Pengzhi

(The 63rd Research Institute，National University of Defense Technology，Nanjing Jiangsu 210007，China)

Abstract：Multiple heterogeneous spectrum users require different Quality of Service(QoS) in cognitive radio networks. A dynamic spectrum allocation method is proposed based on multi-agent reinforcement learning. In order to improve the satisfaction of spectrum users, the proposed method is evaluated by the Quality of Experience(QoE) of spectrum users instead of QoS. Multiple virtual agents are established to simulate spectrum users to learn interactively with environment in a cooperative way, and the optimal spectrum allocation can be obtained by integrating their learning and spectrum decision results. Simulation results show that the proposed method can obtain higher QoE performance of secondary users than those methods based on the traditional reinforcement learning. The probability of collision between spectrum users also can be reduced in the proposed method without any information about the usage rules of primary users and dynamic characteristics of channels.

Keywords：dynamic spectrum allocation；Quality of Experience；multi-agent reinforcement learning；collision probability

为应对呈爆炸式增长的频谱需求，动态频谱共享作为缓解频谱资源紧张局面，解决频谱利用不充分的技术之一，近年来得到大量的关注和研究[1]。动态频谱分配是从频谱管理层面实现动态频谱共享的主要方法，其目标在于“满足现有用户优先使用频谱资源及避免对其产生有害干扰的前提下，在现有用户与新用户之间设计灵活的频谱分配策略，有效地将空闲频段分配给新用户，以提高频谱利用效率”[2]。认知无线电网络中，主用户(Primary User，PU)占用某个授权频段时，次用户(Secondary User，SU)必须从该频段退出，动态频谱分配策略既关系到次用户的频谱需求能否得到满足，又关系到主用户的利益是否受到损害，动态频谱分配策略的好坏直接决定着频谱共享的合理性与有效性。文献[3]利用拓扑图关系，将动态频谱分配问题映射为图着色模型，使用干扰图来降低由频谱共享引起的干扰，但只要拓扑结构发生变化，就需要重新进行映射计算，多适用于静态的网络环境。文献[4]提出了基于定价拍卖的频谱分配模型，虽然有效保证了次用户的公平性，但需提前获取频谱

收稿日期：2021-04-22；**修回日期：**2021-05-23

基金项目：国家自然科学基金青年基金资助项目(61801497)；基础加强计划技术领域基金资助项目(2019-JCJQ-JJ-221)

***通信作者：**张　余　email:zhyu63@163.com

空闲信息，对频谱感知能力要求极高。文献[5]采用基于频谱数据库驱动的频谱共享模型，根据地理位置数据库动态调整主要用户的保护边界来提高频谱利用效率，同样对频谱信息的实时更新能力有较高要求。以上方法虽然可以解决动态频谱分配中的频谱利用与用户通信效能和网络通信效能间的约束与优化问题，但存在灵活性差，收敛速度慢，需要较多先验知识的问题，对控制中心与用户间的通信条件和频谱感知的精确性要求比较高，在实际中实现难度大。

随着近年来强化学习等机器学习研究领域的快速发展，基于强化学习的智能动态频谱分配方法逐渐吸引了越来越多研究者的注意。强化学习具备自适应调节能力，不需要进行实时频谱感知，智能体仅需观察环境状态变化，通过学习采取动作后，根据收到的奖励反馈来提高性能[6]，极大地降低了频谱共享的复杂性。文献[7]针对资源受限的认知传感器网络中能效与频谱效率平衡的问题，将强化学习算法用于信道选择和功率控制，提高了系统能效。为解决难以实施频谱感知条件下，信道选择策略低效产生的额外网络资源和延迟，导致网络性能降低的问题，文献[8]提出了基于强化学习的信道选择算法，该算法在降低网络延时和提高吞吐量性能上效果显著。文献[9]考虑频谱切换或等待停留等动态频谱管理操作对吞吐量和延迟等网络性能长期的影响，提出基于行为者与评论家角色转移的智能频谱管理方法，并结合信道利用率、误包率、丢包率和吞吐量等多种业务质量(QoS)需求建立综合函数，利用该函数进行频谱管理的性能分析及优化。文献[10]考虑到多个无线网络系统的共存问题，提出以最大化用户体验质量(QoE)为目标，利用博弈论和强化学习方法分析多网络系统之间的时间共享和资源分配问题。

综合分析现有研究成果发现，现有基于强化学习的动态频谱分配方法较少考虑多智能体的问题，如何充分利用多智能体之间的合作关系依旧是该方向研究的热门问题[11]；另一方面，受传输环境、链路距离及次用户自身服务需求的影响，不同次用户的频段需求具有多样化，而大多数研究只用吞吐量或多个业务质量对频谱分配性能进行评估，考虑用户体验质量较少，用户实际满意度不高。为此，本文针对动态频谱共享网络中存在不同 QoS 需求的多异质用户的频谱分配问题，建立基于多智能体强化学习的动态频谱分配模型，并以 QoE 作为系统的评价指标开展性能分析。针对动态频谱共享网络中存在多个异质次用户的场景，引入 QoE 评价指标，建立以最大化多个次用户的 QoE 总和为目标函数的动态频谱分配模型；利用强化学习来研究动态频谱分配问题，提出基于多智能体强化学习的动态频谱分配方法。通过构建多个虚拟智能体，模拟多个用户以合作方式与环境进行交互学习，融合各个用户的学习和频谱选择决策结果，实现频谱资源的动态自适应优化分配；在强化学习的迭代机制中，既考虑了次用户与主用户的碰撞问题，又考虑了次用户对频谱资源的最低 QoE 需求，保证了动态频谱分配决策的合理性。

1　问题建模

1.1　系统模型

系统模型如图 1 所示，该模型包括 K 个相互独立的正交信道，N 个不同业务需求的次用户与基站(Base Station，BS)间建立通信链路。为简化分析，假设各信道的带宽相等，信道衰减系数服从瑞利衰落分布，当信道空闲时，次用户以 overlay 模式接入[12]。假设次用户包括视频传输和文本传输两大类用户。

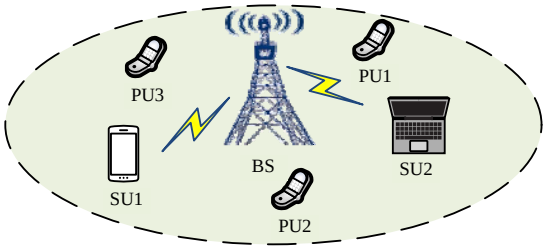

Fig.1 System model
图 1　系统模型

根据用户频谱使用规律，将主要用户的用频行为等效为所有信道的 ON/OFF 行为，其中信道 ON 行为表示信道状态空闲，OFF 行为表示信道被占用。将每个主要用户到达和退出信道的过程构建为离散时间 ON/OFF 两状态的马尔科夫链。假设次用户受到硬件限制，在一个时隙 T 内只能选择一个信道进行数据包传输，信息传输过程包括等待分配阶段 T_{switch}、数据包传输阶段 T_{data} 和学习更新阶段 T_{learning}。

1.2　用户满意度评价模型

目前用户满意度量化方法主要包括两类别法、成对比较法、平均意见得分法(Mean Opinion Score，MOS)，其中国际电信联盟(International Telecommunication Union，ITU)提出的基于 MOS 的用户满意度评价方法应用最为广泛[13]，基于 MOS 的用户满意度评价方法主要是通过用户的 MOS 值来对其系统性能进行量化评价。不失一般性，在此采用基于 MOS 的用户满意度评价方法来构建用户满意度评价模型。参照 ITU 对 MOS 的定义，将 MOS 值划分为 10 个等级，具体值域及含义如表 1 所示。

1.2.1 视频传输

视频传输作为一种实时传输业务，其用户满意度主要由人类视觉系统(Human Visual System，HVS)决定，与网络参数、视频信源编码、视频像素大小等因素有关。为此，综合考虑视频传输速率、信源编码和视频像素大小等要素，采用视频中图像的峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)来定义视频传输 MOS 模型[14]，峰值信噪比φ定义为：

$$\varphi = 10\lg\left(\frac{a_{\max}^2}{MSE}\right) \tag{1}$$

表 1 MOS 值定义表

Table1 Definition of MOS value

MOS value	QoE	degree of damage	satisfaction
10	unconditionally recommend	not damaged	super satisfied
9	extremely recommend	almost no damaged	fully satisfied
8	fully recommend	slightly damaged	very satisfied
7	very recommend	a little damaged	satisfied
6	recommend	less damaged	moderately satisfied
5	moderately recommend	moderately damaged	slightly satisfied
4	slightly recommend	seriously damaged	less satisfied
3	almost not recommend	severely damaged	barely satisfied
2	not recommend	extremely damaged	dissatisfied
1	very not recommend	absolutely damaged	very dissatisfied

式中：$a_{\max}$ 表示图像点颜色的最大值，可表示为 $a_{\max}=2^U-1$，U 表示每个采样点线性脉冲编码调制的位数；MSE 为失真视频与参考视频的均方误差，考虑到由编码方式引起的视频压缩失真以及网络传输中丢包引起的传输失真，将均方误差建模为[14]：

$$MSE = \frac{\theta}{R - R_0} + D_0 + \lambda \cdot P_{\text{loss}} \tag{2}$$

式中：R 为视频的比特率；θ, R_0, D_0 取决于编码的视频序列内容和编码结构的失真参数；λ 为确定压缩视频序列相关的参数；P_{loss} 为端到端丢包率。

若视频或文件传输信号采用 BPSK 调制方式，其相干检测的误比特率(Bit Error Rate，BER)可表示为：

$$P_{\text{b}} = \frac{1}{2}\left(1 - \sqrt{\frac{\overline{\gamma}_{\text{b}}}{1+\overline{\gamma}_{\text{b}}}}\right) \tag{3}$$

式中 $\overline{\gamma}_{\text{b}}$ 为接收机端瞬时信噪比。假设每个 SU 用户数据包长度为 l 比特，则 P_{loss} 则可表示为：

$$P_{\text{loss}} = 1 - (1 - P_{\text{b}})^l \tag{4}$$

当接收到的视频图像的φ大于等于 40 dB 时，其重构的视频序列与原始视频几乎无法区分；当接收到的视频图像的φ低于 20 dB 时，其重构的视频会严重失真[15]。因而，将视频用户峰值信噪比需求的最大值$\varphi_{\max}$ 设置为 45 dB，最小值$\varphi_{\min}$ 设置为 20 dB。当φ值处在中间区域时，人的视觉系统对φ值的变化感受明显；当φ值在很高或者很低的区域变化时，人的视觉系统感受到的视频差异不明显。因此，视频传输用户的 MOS 值表示为：

$$MOS^{\text{v}} = \begin{cases} 1, & \varphi < \varphi_{\min} \\ 9\left(1 - \dfrac{1}{1+\text{e}^{a_1(\varphi - a_2)}}\right) + 1, & \varphi_{\min} < \varphi < \varphi_{\max} \\ 10, & \varphi > \varphi_{\max} \end{cases} \tag{5}$$

式中 a_1, a_2 为视频传输用户 MOS 值的控制参数，主要由用户的最高和最低体验满意度确定。

1.2.2 文件传输

对于文件传输业务，用户满意度与用户所需等待的时间紧密相关，随着等待时间的增长，用户耐心程度逐步降低。因此本文采用有效传输速率，以对数函数的形式定义文件传输用户的 MOS 值。因此，文件传输用户的 MOS 值[16]表示为：

$$MOS^{\text{f}} = \begin{cases} 1, & R < R_{\min} \\ a_3 \lg(a_4 G), & R_{\min} < R < R_{\max} \\ 10, & R > R_{\max} \end{cases} \tag{6}$$

式中：$R_{\min}$ 表示文件传输用户满意度量化所需的最低传输速率；$R_{\max}$ 为文件传输用户满意度量化所需的最高传输速率；a_3, a_4 为文件传输用户的 MOS 值控制参数，主要由用户的最高和最低体验满意度确定；G 为文件传输用户的有效传输速率，可表示为：

$$G = \frac{R(1 - P_{\text{loss}})T_{\text{data}}}{T_{\text{switch}} + T_{\text{data}} + T_{\text{learning}}} \tag{7}$$

1.3 动态频谱分配问题优化模型

动态频谱分配是在考虑不同频谱用户的业务需求前提下，设计合理的频谱分配策略。基于 MOS 值的用户

满意度评价模型将不同类型用户的满意度统一到一个衡量标准，从而为整体衡量动态频谱分配方法的有效性提供了可能。基于此，本文建立以不同用户 MOS 值的和最大化为目标的优化模型来解决动态频谱分配问题，其优化模型可表示为：

$$\max_{\boldsymbol{B}}\left[\sum_{k=1}^{K}\left(\sum_{i=1}^{V}b_{i,k}MOS_{i,k}^{\mathrm{v}}+\sum_{j=1}^{F}b_{j+V,k}MOS_{j,k}^{\mathrm{f}}\right)\right] \tag{8}$$

$$\text{s.t.}\begin{cases} C1: MOS_{i,k}^{\mathrm{v}}\geqslant MOS_{\min}^{\mathrm{v}}, i\leqslant V \\ C2: MOS_{j,k}^{\mathrm{f}}\geqslant MOS_{\min}^{\mathrm{f}}, j\leqslant F \\ C3: V+F=N \\ C4: \sum_{n=1}^{N}b_{n,k}\leqslant 1, \sum_{k=1}^{K}b_{n,k}=1 \\ C5: \boldsymbol{B}=\begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,K} \\ b_{2,1} & b_{2,2} & & \vdots \\ \vdots & & \ddots & \\ b_{N,1} & \cdots & & b_{N,K} \end{bmatrix}_{N\times K},\ b_{n,k}\in\{0,1\} \end{cases} \tag{9}$$

式中：$MOS_{i,k}^{\mathrm{v}}$ 表示将第 k 个信道分配到第 i 个视频传输用户的 MOS 值；$MOS_{j,k}^{\mathrm{v}}$ 表示将第 k 个信道分配到第 j 个文件传输用户的 MOS 值；$MOS_{\min}^{\mathrm{v}}$ 表示视频传输用户的最低满意度门限；$MOS_{\min}^{f}$ 表示文件传输用户的最低满意度门限，V 和 F 分别为视频传输和文件传输的用户数量。约束条件 $C1$ 为视频传输用户的最小 QoE 约束；约束条件 $C2$ 为文件传输用户的最小 QoE 约束；约束条件 $C3$ 表示两类业务的用户数量约束；约束条件 $C4$ 为信道分配约束，每个用户仅占用一个信道，每个信道最多分配给一个用户；约束条件 $C5$ 为分配矩阵 $\boldsymbol{B}$ 的维度约束；$b_{n,k}$ 为信道分配系数，仅当 $b_{n,k}$=1 时表示把第 k 个信道分配到第 n 个用户。通过上述问题建模，本文对由信道分配系数组成的分配矩阵 $\boldsymbol{B}$ 进行优化，从而找到让用户 MOS 值的和最大化的动态频谱分配策略。

2　算法设计

2.1　算法原理

基于多智能体强化学习的动态频谱分配算法以用户满意度为出发点，建立多个虚拟智能体与数据库的历史环境数据交互学习，综合各模拟用户的学习结果，最终生成频谱分配系统的分配策略。对于频谱分配系统，多个模拟用户状态矩阵表示为 $\boldsymbol{S}_l=[\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_k]$，$\boldsymbol{S}_l\in\boldsymbol{S}^{K\times N}$，$\boldsymbol{s}_k$ 表示第 k 个信道内 N 个用户占用情况的状态矢量；频谱分配动作矩阵表示为 $\boldsymbol{A}_l=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n]$，$\boldsymbol{A}_l\in\boldsymbol{A}^{K\times N}$，$\boldsymbol{a}_n$ 表示把 K 个信道分配给第 n 个用户的动作矢量。智能体基于当前状态 $\boldsymbol{S}\in\boldsymbol{S}^{K\times N}$ 和策略 π，执行动作 $\boldsymbol{A}\in\boldsymbol{A}^{K\times N}$，作用于环境，状态由此变化为 $\boldsymbol{S}'\in\boldsymbol{S}^{K\times N}$，同时产生一个强化信号(称为“回报”)$R(s,a)$反馈给智能体；智能体根据强化信号 $R(s,a)$更新其策略，并进入下一轮迭代，其频谱分配算法框架如图 2 所示。

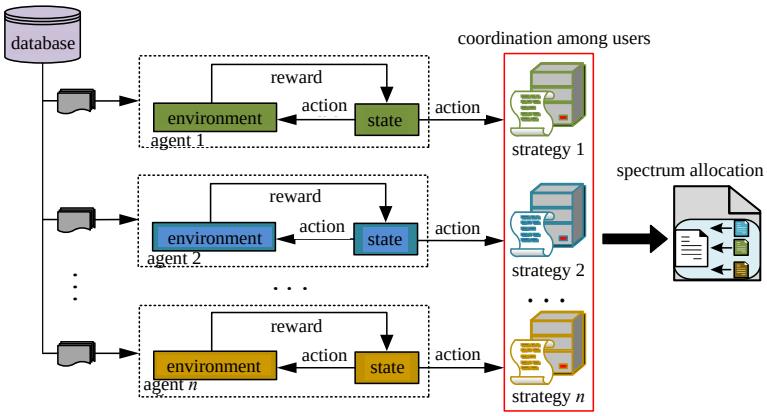

Fig.2 Diagram of proposed algorithm framework
图 2　提出的算法框架图

在算法训练阶段，对于每个模拟用户，其状态空间表示为 $\boldsymbol{s}^{(n)}=\left(s_1^{(n)},s_2^{(n)},\cdots,s_k^{(n)}\right)$，$s_k^{(n)}$ 表示第 n 个模拟用户在第 k 个信道上的状态信息，其动作空间表示为 $\boldsymbol{a}_n=\left(a_1^{(n)},a_2^{(n)},\cdots,a_k^{(n)}\right)$，$a_k^{(n)}$ 表示把第 k 个信道分配给第 n 个模拟用户的动作信息。模拟用户按照序列顺序执行动作信息选择，由于模拟用户之间的动作信息共享机制，单个用户最大化自身效用函数的策略不会与其他用户发生冲突，随后将模拟用户输出的策略进行整合，更新出系统频谱分配策略，并根据用户的反馈信息进行迭代学习，动态调整模拟用户信道选择动作，从而得到用户满意度的和最大的频谱分配策略。为了求解马尔科夫决策过程，基于多智能体强化学习的动态频谱分配算法的 Q 值更新策略参照 Bellman 公式确定，可表示为[17]：

$$Q_{t+1}^{(n)}\left(s_t^{(n)},a_t^{(n)}\right)\leftarrow Q_t^{(n)}\left(s_t^{(n)},a_t^{(n)}\right)+\alpha\left[R^{(n)}+\gamma\max_{a\in A^{(n)}}Q_t^{(n)}\left(s_{t+1}^{(n)},a\right)-Q_t^{(n)}\left(s_t^{(n)},a_t^{(n)}\right)\right] \tag{10}$$

式中：α 为学习率；γ为折扣因子；$R^{(n)}$表示第 n 个模拟用户的回报值；$s_t^{(n)}$表示第 t 时刻，第 n 个模拟用户占用信道的状态信息；$a_t^{(n)}$表示第 t 时刻，第 n 个模拟用户的动作信息；$\gamma\max_{a\in A^{(n)}}Q_t^{(n)}\left(s_{t+1}^{(n)},a\right)$表示智能体未来所得到的评估函数。针对多智能体 Q 表的空间大、难以收敛的问题，基于多智能体强化学习的动态频谱分配算法对 Q 表进行拆解，将多智能体学习问题降维为单智能体学习问题。在迭代过程中，每个智能体单独更新其 Q 值，频谱分配最优策略可以通过所有虚拟智能体的 Q 值的和来表征，即：

$$Q(s,a)=\sum_{n=1}^{N}Q^{(n)}(s,a) \tag{11}$$

频谱分配系统的目标是找到一个频谱分配策略 π，通过频谱分配策略 π 形成的分配矩阵来最大化系统整体的预期累计折扣奖励：

$$V_\pi=\boldsymbol{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t R_{t+1}(\boldsymbol{s}_{t+1},\pi)\right] \tag{12}$$

式中：R_{t+1}表示频谱分配系统 t+1 时刻的回报值；$\boldsymbol{s}_{t+1}$表示 t+1 时刻 K 个信道内 N 个用户占用情况的状态矢量，即 t+1 时刻的最优策略 π^*可表示为：

$$\pi^*=\arg\max_\pi V_\pi=\arg\max_\pi \boldsymbol{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t R_{t+1}(\boldsymbol{s}_{t+1},\pi)\right] \tag{13}$$

2.2 算法实现流程

基于多智能体强化学习的动态频谱分配算法通过调节系统给模拟用户不同的信道选择动作，设计用户反馈的回报值对 $Q(s,a)$进行迭代更新，指导算法朝着用户满意度总和最大来选择频谱分配策略。算法中的“探索–利用”机制使得系统以一定概率去选择频谱分配策略，能在学习前期保证智能体在多个信道上的经验积累；当信道环境发生变化时，这种机制也会使得频谱分配系统自适应地调整分配行为，选取最佳策略。具体算法流程如图 3 所示。

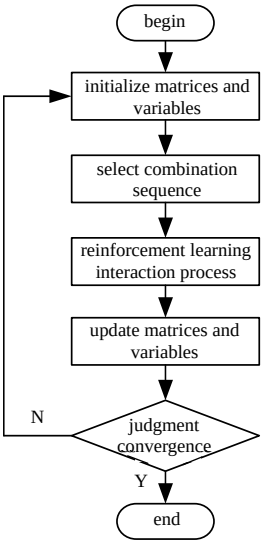

Fig.3 Flow chart of the proposed algorithm

图 3 所提算法流程图

2.3 算法详细步骤

1) 初始化：初始化 $\boldsymbol{Q}$ 矩阵为全零矩阵，设定学习率 α、折扣因子γ和模拟退火算法的初始温度 T_{begin}，再从状态空间 $\boldsymbol{S}^{K\times N}$ 和动作空间 $\boldsymbol{A}^{K\times N}$ 中随机选取一个作为初始值。

2) 组合选择：每一个模拟用户均作为一个智能体，N 个模拟用户按照用户优先级排序则生成 $N!$ 个组合序列，并在每个时隙以 $P_0=1/N!$的概率选取组合序列，按照序列顺序进行迭代学习，从而寻找效益最高的系统分配策略。

3) 交互过程：在 t 时刻的退火温度 $T(t)$ 下，通过查询 $\boldsymbol{Q}$ 矩阵的值，第 n 个智能体在当前状态 $s_t^{(n)}$下从动作空间 $\boldsymbol{a}_n$ 执行动作 $a_i^{(n)}\in\boldsymbol{a}_n$ 作用于环境，其选择概率为：

$$P\left(a_i^{(n)}\middle|s_t^{(n)}\right)=\frac{\exp\left(Q^{(n)}\left(s_t^{(n)},a_i^{(n)}\right)/T(t)\right)}{\sum_{k=1}^{K}\exp\left(Q^{(n)}\left(s_t^{(n)},a_k^{(n)}\right)/T(t)\right)} \tag{14}$$

4) 迭代更新：将执行动作 $a_i^{(n)}$选取的信道分配给用户，同时状态跳转至 $s_t^{'(n)}\in\boldsymbol{s}^{(n)}$，并计算当前 MOS 值作为频谱分配的回报 $R^{(n)}\left(s_t^{'(n)},a_i^{(n)}\right)$，可表示为：

$$R^{(n)}\left(s_t^{'(n)},a_i^{(n)}\right)=\begin{cases}MOS_n, MOS_n\geqslant\xi\\ -1, \quad MOS_n<\xi\end{cases} \tag{15}$$

式中：MOS_n表示第 n 个用户的 MOS 值；ξ为用户满意度的门限值，当获得反馈的 MOS 值低于门限ξ时，则给予系统一个负面的评价反馈；当获得反馈的 MOS 值高于或等于门限ξ时，将其 MOS 值作为回报反馈。回报 R 参照 Bellman 公式更新当前 $Q^{(n)}(s,a)$，再按照反比例函数递减规律，更新模拟退火算法的温度参数 T_t，直至所有用户的 Q 值的和 $Q(s,a)$ 形成稳定最大值。

5) 收敛判决：重复以上步骤，直至退火温度从初始温度 T_{begin} 递减到结束温度 T_{finish}。记录 Φ 时间内的累计回报，当回报趋近稳定时，算法结束。

3　仿真结果与分析

表 2 信道状态转移概率

Table2 State transition probability of channels

channels	1	2	3	4	5	6
$P(0,0)$	0.9	0.9	0.2	0.2	0.8	0.3
$P(1,1)$	0.8	0.3	0.3	0.9	0.2	0.8

为验证所提方法的性能，以文献[18]和文献[19]采用的方法及随机分配方法作为比较对象，其中，文献[18]采用 Q-learning 方法将分配系统的所有分配方案作为策略进行迭代学习；文献[19]采用不考虑用户之间协作的强化学习方法，当单个智能体学习时，其他用户被视为其环境的一部分，发生碰撞时给予一个惩罚机制来进行迭代学习；随机分配方法则是在每个时隙中，随机选择 N 个不同的信道供用户使用。假设所有强化学习算法具有相同的学习率 α=0.1 以及折扣因子 γ=0.9。仿真参数设置如下：信道数量 K=6，次用户数量 N=2(其中 V=1,F=1)。信道空闲与占用的状态转移矩阵如表 2 所示。

考虑到无线信道的动态特性，将每个模拟用户的接收端在每个时隙的信噪比设置为 15~20 dB 均匀分布的随机变量，每个时隙信息传输过程 $T=T_{\text{switch}}+T_{\text{data}}+T_{\text{learning}}$，其中 T_{switch}，T_{data}，T_{learning} 的值分别设为 0.3 ms,9.5 ms 和 0.2 ms，单个数据包长度为 16 bit。考虑到学习前期尽量多地遍历到各种决策，模拟退火算法的初始温度设为 $T_{\text{begin}}=2\ 000$ K，并用反比例函数递减至最终温度 T_{finish}=1 K。用户满意度门限 $\xi=4$，记录时间 $\Phi=50T$，蒙特卡洛实验次数为 1 000 次。

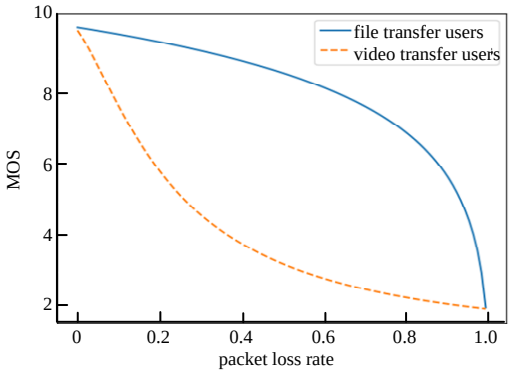

Fig.4 MOS value versus packet loss rate
图 4　丢包率对用户 MOS 值的影响

图 4 给出了相同传输速率下，丢包率对视频传输、文件传输用户的 MOS 值的影响变化曲线。从图 4 可以看出，对于非实时的文件传输用户，传输中轻微的丢包对 MOS 值影响较低，在单位时间内满足一定的数据量即可；当丢包严重时，用户等待时间偏长，将引起用户满意度的急剧下降。对于实时性的视频传输用户，传输中的丢包会直接引起画面的模糊，用户满意度随传输中丢包率的增加大幅度下降，相较于文件传输用户，视频传输用户对丢包率更加敏感。

图 5 和图 6 分别给出了 MOS 值、冲突率随迭代时间的关系。从图 5 可知，随着迭代时间的增加，提出的方法通过与环境的不断交互、学习，引导分配决策选择“较好”的信道优先进行分配，不同类型用户的 MOS 值均得到提高。由于视频传输用户对丢包率更为敏感，整体满意度低于文件传输用户。在 Φ=12 后，文件传输用户 MOS 值稳定在 7.8 左右，视频传输用户 MOS 值稳定在 6.5 左右。从图 6 可以看出，在未知主要用户使用规律和信道动态特性条件下，本文所提方法随迭代次数增加，用户间的冲突率从 48%逐渐降低至 20.5%左右，能有效降低用户间的冲突概率。

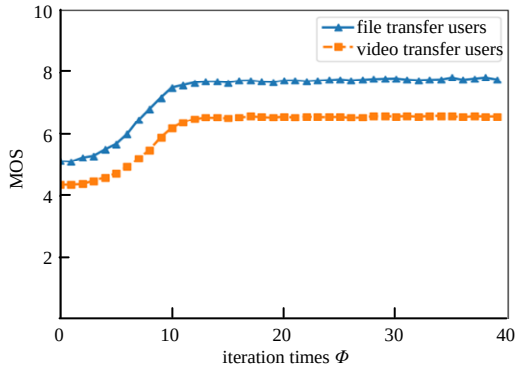

Fig.5 MOS value vs. iteration times of proposed method
图 5　所提方法不同迭代次数下的 MOS 值

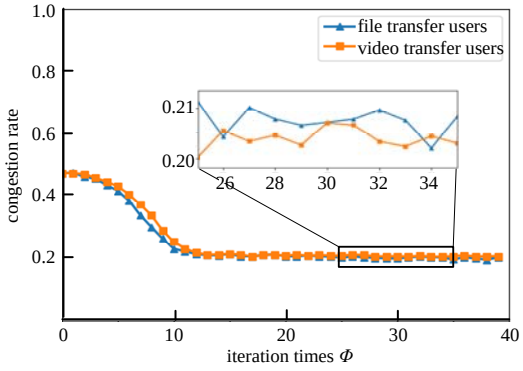

Fig.6 Congestion rate vs. iteration times of proposed method
图 6　所提方法不同迭代次数下的冲突率

图 7 和图 8 为采用不同学习算法后，各算法的用户 MOS 值总和以及用户间冲突概率随迭代时间的关系。图 7 显示，在迭代初期，仅文献[19]的方法中用户间存在相互冲突，MOS 值总和小于随机分配方法，其余强化学习算法与随机分配算法基本相同。随着迭代时间的增加，强化学习方法收敛到各自算法的最优 Q 值，

本文所提方法相比于随机分配方法，MOS 值总和提升 42%。从图 8 可以看出，与随机分配方式相比，强化学习方法能有效降低用户间的冲突率，随迭代时间的增加，各强化学习方法冲突率均低于随机分配方法。

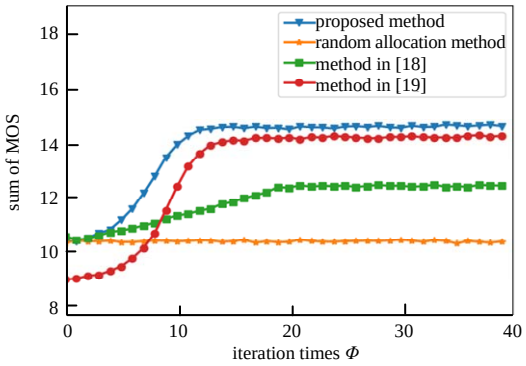

Fig.7 Sum of MOS values vs. iteration times for different methods

图 7 不同方法 MOS 值总和的比较

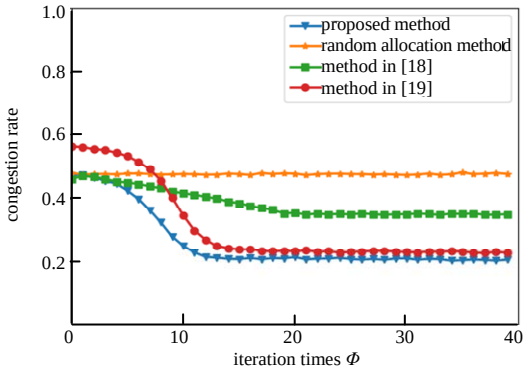

Fig.8 Congestion rate vs. iteration times for different methods

图 8 不同方法冲突率的比较

本文所提方法性能优于文献[18]的方法，其原因在于：本文采用 Q 表拆解的方式，将大状态空间拆解到单个用户，每个用户单独更新其 Q 值，极大减少算法所需的计算时间，且随着 Q 表的缩小，收敛到局部最优的概率大大减小，智能体更容易求解出最优决策。本文提出的方法相比文献[19]的方法性能更优，其原因在于：本文方法在文献[19]方法的基础上，对单个智能体的学习内容进行了优化，智能体间的合作减少了低效的分配决策。在学习过程中，多次不同的用户组合选择，使每个智能体视为同等优先级进行学习，一定程度上避免了因智能体学习环境不同形成的局部最优解。在最终收敛的 MOS 值总和大于文献[19]方法的情况下，收敛所需时间少于该算法。

图 9 为不同均值信噪比下 MOS 值总和比较。由图 9 可见，随着信噪比的增大，4 种动态频谱分配方法的 MOS 值总和都随之增大，但性能差异逐渐明显，本文所提方法优于其他 3 种方法，主要原因在于：随着信噪比的增大，用户接入信道后可获得的传输速率会提高，传输丢包率降低，若此时能够寻找到空闲概率较高的信道进行分配，用户的满意度会有明显的提升。另外，较高的满意度回报给予算法一种强奖励机制，信道质量区分度越大，使用强化学习方法的优势越加明显。

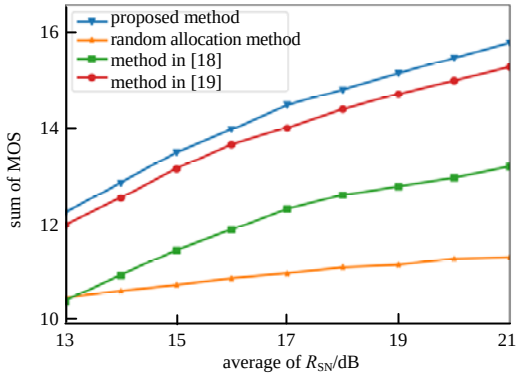

Fig.9 Sum of MOS values vs. average of SNR for different methods

图 9 不同均值信噪比下 MOS 值总和比较

4 结论

本文从提高用户满意度出发，提出一种基于多智能强化学习的动态频谱分配方法。仿真结果表明，该方法相比基于传统强化学习的动态频谱分配方法，提高了用户满意度，降低了冲突率。同时，为解决多智能体强化学习问题提供了一种实现途径，根据群体智能的目标，将问题分成多个小目标进行学习，指导多个智能体进行训练，到达求解目标后，群体智能再重新分配多个小目标进行学习，如此循环往复，最终达到群体智能的最优决策。

参考文献：

[1] DTA B , AB B,MD A,et al. Survey on spectrum sharing/allocation for cognitive radio networks Internet of Things[J]. Egyptian Informatics Journal, 2020,21(4):231–239.

[2] ZHANG Z,XIAO Y,MA Z,et al. 6G wireless networks:vision,requirements,architecture,and key technologies[J]. IEEE Vehicular Technology Magazine, 2019,14(3): 28–41.

[3] DU B,XUE R,ZHAO L,et al. Coalitional graph game for air-to-air and air-to-ground cognitive spectrum sharing[J]. IEEE Transactions on Aerospace and Electronic Systems, 2019,56(4):2959–2977.

[4] FARSHBAFAN M K,BAHOAR M H,KHAIEHRAVENI F. Spectrum trading for Device-to-Device communication in cellular networks using incomplete information bandwidth-auction game[C]// 27th Iranian Conference on Electrical Engineering (ICEE). Yazd, Iran:IEEE, 2019:1441–1447.

[5] BHATTARAI S,PARK J M,LEHR W. Dynamic exclusion zones for protecting primary users in database-driven spectrum sharing[J]. IEEE/ACM Transactions on Networking, 2020,28(4):1506–1519.
[6] KARMAKAR R,CHATTOPADHYAY S,CHAKRABORTY S. Dynamic link adaptation in IEEE 802.11 ac:a distributed learning based approach[C]// IEEE 41st Conference on Local Computer Networks(LCN). Dubai:IEEE, 2016:87–94.
[7] HE J,PENG J,JIANG F,et al. A distributed Q learning spectrum decision scheme for cognitive radio sensor network[J]. International Journal of Distributed Sensor Networks, 2015,11(5):1–10.
[8] OYEWOBI S S,HANCKE G P,ABU-MAHFOUZ A M,et al. An effective spectrum handoff based on reinforcement learning for target channel selection in the industrial internet of things[J]. Sensors, 2019,19(6):1395–1416.
[9] KOUSHIK A M,HU F,KUMAR S. Intelligent spectrum management based on transfer actor-critic learning for rateless transmissions in cognitive radio networks[J]. IEEE Transactions on Mobile Computing, 2017,17(5):1204–1215.
[10] BAIRAGI A K,ABEDIN S F,TRAN N H,et al. QoE-enabled unlicensed spectrum sharing in 5G:a game-theoretic approach[J]. IEEE Access, 2018(6):50538–50554.
[11] LUONG N C,HOANG D T,GONG S,et al. Applications of deep reinforcement learning in communications and networking: a survey[J]. IEEE Communications Surveys & Tutorials, 2019,21(4):3133–3174.
[12] MITOLA J. Cognitive radio for flexible mobile multimedia communications[C]// IEEE International Workshop on Mobile Multimedia Communications. San Diego,CA,USA:IEEE, 1999:3–10.
[13] 林闯,胡杰,孔祥震. 用户体验质量(QoE)的模型与评价方法综述[J]. 计算机学报, 2012,35(1):1–15. (LIN Chuang,HU Jie,KONG Xiangzhen. Overview of QoE models and evaluation methods[J]. Chinese Journal of Computers, 2012,35(1):1–15.)
[14] ZHOU L,WANG X,TU W,et al. Distributed scheduling scheme for video streaming over multi-channel multi-radio multi-hop wireless networks[J]. IEEE Journal on Selected Areas in Communications, 2010,28(3):409–419.
[15] KHAN S,DUHOVNIKOV S,STEINBACH E,et al. MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication[J]. Advances in Multimedia, 2007:6.
[16] HE L,LIU B,YAO Y,et al. MOS-based channel allocation schemes for mixed services over cognitive radio networks[C]// 2013 Seventh International Conference on Image and Graphics. Qingdao,China:IEEE, 2013:832–837.
[17] VLASSIS N. A concise introduction to multi-agent systems and distributed artificial intelligence[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2007,1(1):1–71.
[18] SHIN M,CHUNG M Y. Learning-based distributed multi-channel dynamic access for cellular spectrum sharing of multiple operators[C]// 25th Asia–Pacific Conference on Communications(APCC). Haiphong.Vietnam:IEEE, 2019:384–387.
[19] JIANG H,WANG T,WANG S. Multi-agent reinforcement learning for dynamic spectrum access[C]// IEEE International Conference on Communications(ICC). Shanghai,China:IEEE, 2019:1–6.

作者简介：

童　乐(1997–)，男，湖北省鄂州市人，硕士研究生，主要研究方向为电磁频谱管理.email: tongle63s@163.com.

梁　涛(1961–)，男，博士，研究员，主要研究方向为通信抗干扰.

张　余(1983–)，男，四川省绵阳市人，硕士，副研究员，主要研究方向为电磁频谱技术.

钱鹏智(1994–)，男，硕士，工程师，主要研究方向为空间信息获取与处理.