文章编号：1672-2892(2010)06-0720-06

# 基于 Dec-POMDP 的认知无线电网络频谱接入算法

张迎晓，杨 涛，胡 波，陈光梦

(复旦大学 电子工程系，上海 200433)

**摘 要**：针对认知无线电网络中认知用户(CR)的机会频谱感知及接入问题，提出了一种基于分布式部分可观测马尔科夫决策过程(Dec-POMDP)的多用户频谱接入算法。在该模型框架下，相邻CR用户通过交换接入策略，以区域策略梯度方向为基准，对各个CR用户的接入策略做出调整，从而得到最优联合接入策略。仿真结果表明：该算法有效降低了授权用户的容量损失，提高了空闲频谱的利用效率，能够更有效地做出接入决策。

**关键词**：认知无线电；可观测马尔科夫决策过程；策略梯度；频谱分配

**中图分类号**：TN911.73 **文献标识码**：A

# Decentralized POMDP-based cognitive radio network spectrum access algorithm

ZHANG Ying-xiao, YANG Tao, HU Bo, CHEN Guang-meng

(Department of Electronics Engineering, Fudan University, Shanghai 200433, China)

**Abstract**: In the Cognitive Radio(CR) network, the opportunistic spectrum sensing and access is of paramount importance to the primary user's capacity. This paper propose a multi-user Decentralized Partially Observable Markov Decision Process(Dec-POMDP) CR spectrum access algorithm. In this framework, adjacent CR users exchange access policies with neighbors, and according to policy gradient, each local CR user adjusts its own access policy till the system of CR network obtains an optimum joint access policy. Simulation results show that this algorithm can lower the capacity loss of authorized user efficiently and improve spare spectrum efficiency, moreover, it can make access decision more efficiently.

**Key words**: Cognitive Radio; Partially Observable Markov Decision Process; policy gradient; spectrum allocation

由于无线通信需求的不断增长，作为不可再生资源，无线频谱正变得越来越稀缺，这主要是由传统的固定频谱分配政策造成的。这种分配政策造成了频谱利用的不平衡，有的频段如个人无线通信频段是超负荷的，而有的频段如无线电视广播频段则没有得到充分利用，美国联邦通信委员会的统计数据显示，在目前分配方案下，各频段的频谱利用率从15%到85%不等。近年来，能够实时感知频谱使用状态，动态接入空闲频谱的认知无线电(CR)技术受到了广泛关注，该技术是对传统的软件无线电的扩展，可有效提高频谱利用率[1-4]。

在认知无线电网络中，由于硬件条件的制约，CR用户往往不能得到整个系统的完全观测。对于单个CR用户，无论是对系统频谱状态的感知，还是接入动作对系统状态的影响，结果部分的可观测性是必须考虑的。在文献[5]中，作者提出了一种基于部分可观测马尔科夫决策过程(POMDP)的认知无线电感知和接入算法，这种方案需要状态转移模型的先验知识或者必须对系统状态转移模型进行学习，值迭代搜索最优状态路径方法需要的时间开销较大。而且，单个CR用户独立做出的策略选择只能达到局部最优，对于由若干个CR用户组成的系统而言并非是最优的联合接入策略。为了解决类似的多用户接入问题[6]，文献[7]提出了一种Dec-POMDP。Dec-POMDP为分布式随机控制系统，对相应的动作，每个用户只能得到本地观测，并不能得到系统全局状态的完全观测，用户的联合策略决定了系统的性能。本文在Dec-POMDP框架下建立了多CR用户频谱接入模型，在相邻CR交互合作的基础上，搜索最优的联合接入策略。通过强化学习的方法来逼近Dec-POMDP的联合最优策略。基于策略

**收稿日期**：2010-02-09；**修回日期**：2010-06-25

**基金项目**：国家自然科学基金资助项目(60972024)；教育部博士点基金资助项目(20090071120087)；专用集成电路与系统国家重点实验室开放课题(20080402)

梯度估计，把各个 CR 用户的接入策略参数化，实现 CR 节点的策略更新，以获取最优联合接入策略[8-10]。

## 1　认知网络问题描述和系统模型

### 1.1 问题描述

假设某个频段由 $N$ 个授权用户(主用户)子信道组成，每个子信道对应带宽为 $B_n(n=1,2,\cdots,N)$，主用户对信道的使用时间可分为若干时间片。假设信道的使用状态服从离散马尔科夫过程，$s_i(t)\in\{0,1\}$ 表示时刻 $t$ 子信道 $i$ 的使用状态：0 表示忙，1 则表示空闲，那么，全部子信道的状态信息可以表示为向量：$\boldsymbol{S}(t)=[s_1(t),s_2(t),\cdots,s_N(t)]$。假设在 $T$ 个时间片内，主用户的频谱使用统计信息保持不变。

考虑相邻的 $M$ 个 CR 用户组成一个 ad hoc 工作组，在 $N$ 个子信道中搜索可用频谱。在每个时间片内，CR 用户经过频谱感知，并与相邻 CR 节点交换本地观测信息以及本地接入策略，寻找最优的联合接入策略。

把进行通信的 2 个 CR 用户称为 1 个 CR 用户对，假设 CR 用户对 $j$ 接入了第 $i$ 个子信道，$X_i^p$ 和 $X_i^{cj}$ 分别表示主用户和 CR 用户对 $j$ 的发射信号，$Y_i^p$ 和 $Y_i^{cj}$ 分别表示主用户和 CR 用户对 $j$ 的接收信号，$n_i^p$ 和 $n_i^{cj}$ 分别为主用户和 CR 用户对 $j$ 在该子信道上的信道噪声，$q_i$ 和 $f_{ij}$ 分别是主用户发射机到主用户接收机和 CR 用户对 $j$ 的接收机的信道增益，$g_i$ 和 $h_i$ 分别是 CR 用户对 $j$ 的发射机到主用户接收机和 CR 用户对 $j$ 的接收机的信道增益。当 CR 对主用户信号的检测做出正确判决时，主用户和 CR 用户的接收信号为：

$$\begin{cases} Y_i^p = q_i X_i^p + n_i^p \\ Y_i^{cj} = h_i X_i^{cj} + n_i^{cj} \end{cases} \tag{1}$$

反之，主用户和 CR 用户的接收信号为：

$$\begin{cases} Y_i^p = q_i X_i^p + g_i X_i^{cj} + n_i^p \\ Y_i^{cj} = f_{ij} X_i^p + h_i X_i^{cj} + n_i^{cj} \end{cases} \tag{2}$$

这种情况下，主用户和 CR 用户的信号将互相造成干扰。考虑到认知无线电通信的特点，CR 用户信道接入策略的前提即 CR 用户错误接入造成的授权用户容量损失必须在可接受的范围内。

显然，当 CR 频谱感知正确时，只有信道噪声对 CR 的通信造成干扰；当 CR 频谱感知不正确时，信道噪声和主用户的信号对于 CR 来说都是干扰，同时 CR 信号还会造成主用户信道容量损失，根据香农公式，CR 用户对 $j$ 获得的信道容量为：

$$C_j = B_n\sum_{i=1}^{N} P_{Di}\log\left(1+\frac{h_i^2P_i^{cj}}{\sigma_i^{cj}}\right)+(1-P_{Di})\log\left(1+\frac{h_i^2P_i^{cj}}{\sigma_i^{cj}+f_{ij}^2P_i^p}\right) \tag{3}$$

式中：$N$ 表示该 CR 用户对接入的子信道数量；$P_i^{cj}$ 和 $P_i^p$ 分别表示 CR 用户对 $j$ 和主用户在子信道 $i$ 上的发射功率；$\sigma_i^p$ 和 $\sigma_i^{cj}$ 分别为主用户和 CR 用户对 $j$ 在该子信道上的信道噪声的方差；$P_{Di}$ 表示 CR 用户在子信道 $i$ 上做出正确检测的概率。本文主要研究 CR 用户的接入对授权用户容量损失的影响，因此假设 $P_{Di}$ 为已知，CR 对主用户信号的检测不存在虚警。

各个 CR 节点的总发射功率必须低于某个阈值 $P_{\text{thr}}^c$，以保证主用户接收机的信噪比高于干扰温度门限 $\gamma$：

$$\sum_{j=1}^{M}\sum_{i=1}^{N}P_i^{C_j} \leqslant P_{\text{thr}}^c\left(\sum_{i=1}^{N}P_i^p\right)/\left(P_{\text{thr}}^c+\sum_{i=1}^{N}\sigma_i^p\right)\geqslant\gamma \tag{4}$$

CR 用户做出错误接入时，引起主用户信道容量损失为：

$$\Delta C_{p\text{损失}} = \Delta B\sum_{i=1}^{N}(1-P_{Di})\left(\log\left(1+\frac{q_i^2P_i^p}{\sigma_i^p}\right)-\log\left(1+\frac{q_i^2P_i^p}{\sigma_i^p+g_i^2P_i^{cj}}\right)\right) \tag{5}$$

那么，这 $M$ 个 CR 用户的信道分配问题即为一个带约束条件的优化问题：

$$\text{Max}\,C_{\text{cr}}=\sum_{j=1}^{M}C_j,\text{St.}\begin{cases}\sum_{j=1}^{M}\sum_{i=1}^{N}P_i^{cj}\leqslant P_{\text{thr}}^c\\ \Delta C_{p\text{损失}}\leqslant k\end{cases} \tag{6}$$

式中 $k$ 表示预设的授权用户容量损失阈值。

### 1.2 认知无线电网络的 Dec-POMDP 模型

由 $M$ 个 CR 用户组成的认知无线电 ad hoc 网络，其 Dec-POMDP 模型可表示为元组：

$$\left\langle S, \boldsymbol{A}, P, R, \boldsymbol{\Omega}, O, p_0 \right\rangle \tag{7}$$

式中：$S$ 是授权用户信道集状态的有限集；$\boldsymbol{A}=\{A_i\}$ 是整个 CR 工作组联合接入动作的合集，其中 $A_i$ 表示用户 $i$ 可用的动作集；$P(S'|\boldsymbol{a},S)$ 表示当前状态为 $S$ 时，CR 工作组执行联合接入动作 $\boldsymbol{a}=\{a_1,\cdots,a_N\}$ 后，下一时刻状态为 $S'$ 的概率；$\boldsymbol{\Omega}=\{\Omega_i\}$ 表示联合观测的有限集，其中 $\Omega_i$ 为每个 CR 用户的观测集；$O(\boldsymbol{o}|S,\boldsymbol{a},S')$ 表示 CR 用户工作组在当前状态为 $S$ 的条件下，执行联合动作集 $\boldsymbol{a}$，下一个状态转变为 $S'$ 时，联合观测量为 $\boldsymbol{o}=\{o_1,\cdots,o_N\}\in\boldsymbol{\Omega}$ 的条件概率；$R(S,\boldsymbol{a},S')$ 表示工作组的效用函数，即 CR 的信道容量和 $C_{\text{cr}}$；$p_0$ 表示初始状态分布。

对于 CR 用户 $i$，从时刻 1 到时刻 $t$ 时的本地观测历史向量定义为：

$$\boldsymbol{o}_i^h(t)=\{o_i(1),o_i(2),\cdots,o_i(t)\} \tag{8}$$

定义本地策略 $\omega_i$ 为从本地历史观测向量 $\boldsymbol{o}_i^h(t)$ 到本地动作 $a_i(t)$ 的映射。联合策略 $\boldsymbol{\omega}(t)=\{\omega_1(t),\cdots,\omega_N(t)\}$ 为各个 CR 用户本地策略组成的向量。

结合本文 1.1 节的式(6)，以 CR 工作组的信道容量和 $C_{\text{cr}}$ 作为联合策略的衡量标准。假设初始状态为 $S(0)=s_0$，某个联合策略向量 $\boldsymbol{\omega}(t)$ 在时间 $T$ 内平均效用表示为：

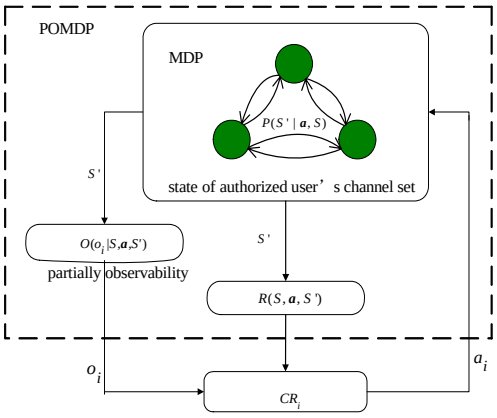



Fig.1 Dec-POMDP model
图 1 分布式 POMDP 模型

$$J(\boldsymbol{\omega},s_0)\coloneqq\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}E_{s_0}^{\boldsymbol{\omega}(t)}\{R(S(t),\boldsymbol{a}(t),S(t+1))\}=\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}E_{s_0}^{\boldsymbol{\omega}(t)}\left\{\sum_{j=1}^{M}C_j(t)\right\} \tag{9}$$

如果对于任意联合策略向量 $\boldsymbol{\omega}(t)$，任意初始状态 $s_0$，有 $J(\boldsymbol{\omega}^*(t),s_0)\geqslant J(\boldsymbol{\omega}(t),s_0)$，那么向量 $\boldsymbol{\omega}^*(t)$ 为最优联合策略。

图 1 为 Dec-POMDP 模型示意图，$CR_i$ 表示 CR 工作组中的用户对 $i$。在 CR 用户的联合接入动作 $\boldsymbol{a}(t)$ 的影响下，授权用户信道集的状态变化为一个马尔科夫决策过程。但是，CR 用户个体无法得到授权用户信道集的真实状态，只能得到本地观测 $o_i(t)$，因此 1.1 节提出的频谱分配优化问题可以通过建立一个 Dec-POMDP 模型来解决。

## 2 基于 Dec-POMDP 的分布式策略搜索

### 2.1 CR 用户状态控制机

与单用户 POMDP 类似[4]，因为不能得到主用户信道集状态的完全观测及其状态转移概率，每个 CR 用户都需要“记忆”历史观测和接入动作，以选择最优接入策略。

本文引入状态控制机作为 CR 用户内部状态转移的控制模型，控制机的内部状态记录了与 CR 用户即时接入决策相关的历史观测和动作[10]。每一个 CR 用户可看作一个状态控制机，表示为元组 $\left\langle I_i,\phi_i,f_i,\theta_i,\mu_i,o_i\right\rangle$。其中，$I_i$ 表示用户 $CR_i$ 的内部状态；$f_i(\cdot)$ 和 $\mu_i(\cdot)$ 分别为内部状态转移概率分布和动作选择概率分布；$\phi_i\in R^{n_{\phi_i}}$，$\theta_i\in R^{n_{\theta_i}}$，$\phi_i$ 和 $\theta_i$ 分别为状态控制机的内部状态转移概率和动作选择概率的参数选择空间；$o_i$ 表示 CR 用户的本地观测；内部状态 $I_i$ 与不同的历史观测和动作有关。本文提出的 CR 用户状态控制机的内部状态转移概率和动作选择概率可通过不断地搜索 $\phi_i$ 和 $\theta_i$ 的参数空间学习得到。

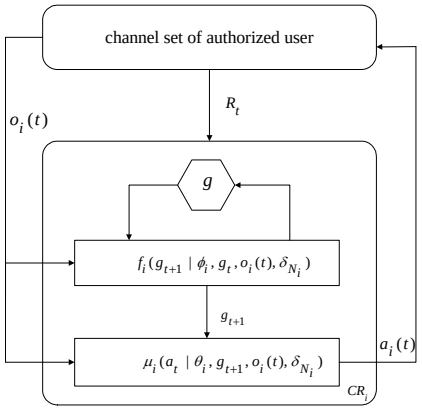



Fig.2 CR user's state machine
图 2 CR 用户状态控制机

CR 系统的分布式特性以及授权用户容量损失的约束条件决定了 CR 用户的策略选择不仅依赖于本地观测和动作选择策略，也需要考虑相邻用户的策略信息，通过这种分布式合作的方法逐渐达到整个 CR 工作组的最优联合策略。因此，把相邻 CR 用户的策略参数考虑进来，定义 CR 用户状态控制机的内部状态转移分布为：

$$f_i(\cdot \mid \phi_i, g_t, o_i(t), a_i(t-1), \delta_{N_i}) \tag{10}$$

定义动作选择概率分布为 $\mu_i(\cdot \mid \theta_i, g_{t+1}, o_i(t), \delta_{N_i})$，其中，$g_t \in I_i$；$o_i$ 是当前本地观测；$a_i(t-1)$ 为前一时刻本地动作；$\delta_{N_i}$ 表示 CR 用户 $i$ 相邻各用户的策略参数集 $\langle \theta_k, \phi_k \rangle$，$k \in \{N_i\}$，$N_i$ 为 CR 用户 $i$ 相邻用户的集合。

通过交互得到了相邻用户的策略集 $\vec{\omega}_{N_i}$，根据式(9)，CR 用户 $i$ 的目标为寻找最优的联合接入策略 $\vec{\omega}(t)$，即 $\{\theta_j, \phi_j\}, j \in \{i \cup N_i\}$，以最大化用户 $i$ 及其相邻用户的区域效用和：

$$C_{\mathrm{loc},i}(t) := \sum_{j=1}^{k} C_j(t) \mid \vec{\omega}_{N_i} \tag{11}$$

式中 $k$ 表示 $i$ 及其相邻 CR 用户集的用户数。

### 2.2 认知无线电的分布式合作策略搜索算法

根据上述分析，CR 用户状态控制机把接入策略参数化表示，同时考虑了相邻 CR 用户的接入策略参数，经过多次迭代计算可以逼近 CR 工作组的联合最优决策。具体算法描述如下：

第 1 步：初始化

$t$=0, $T_n$ 表示预设的迭代次数，$g_t$ 表示用户在 $t$ 时刻的内部状态,

每个用户 $i$ 产生一个随机策略 $\omega_i$，表示为 $\langle \theta_i, \phi_i \rangle$,

第 2 步：

While $t < T_n$,

获得本地观测 $o_i(t)$,

用户 $i$ 与相邻用户 $\{N_i\}$ 相互交换本地策略,

根据 $f_i(\cdot \mid \phi_i, g_t, o_i(t), a_i(t-1), \delta_{N_i})$ 选择下一状态 $g_{t+1}$,

根据 $\mu_i(\cdot \mid \theta_i, g_{t+1}, o_i(t), \delta_{N_i})$ 选择并执行动作 $a_i(t)$,

计算当前策略的即时区域效用和 $C_{\mathrm{loc},i}(t)$,

根据策略梯度方法 local-best-policy( $\omega_i$, $\delta_{N_i}$, $o_t$, $C_{loc,i}(t)$)得到用户 $i$ 的即时最优策略 $\omega_i^*$,

计算最优策略的即时区域效用和 $C^*_{\mathrm{loc},i}(t)$,

设 $Gain_i = C^*_{\mathrm{loc},i}(t) C_{\mathrm{loc},i}(t)$,

向相邻用户 $\{N_i\}$ 广播 $Gain_i$,

Max $Gain$= $\max\limits_{k \in \{i \cup N_i\}} gain_k$,

Winner= $\arg \max\limits_{k \in \{i \cup N_i\}} gain_k$,

如果 $i$ 是 Winner,

　　则更新策略，令 $\omega_i = \omega_i^*$，并广播 $\omega_i^*$ 给相邻用户 $\{N_i\}$,

否则,

　　从 Winner 用户处接受策略 $\omega_{\mathrm{winner}}$，并且更新 $\delta_{N_i}$,

$t$=$t$+1 并返回第 2 步。

在上述算法中，定义策略梯度方向的最优本地策略和用户 $i$ 的当前策略产生的收益差为 $Gain_i$，只有 Gain 最大的用户被允许改变自己的策略到最优策略，其他相邻用户则接受该策略安排。经过多次迭代后，该 CR 工作组可以逼近到最优联合接入策略。其中，策略梯度算法[11]函数 local-best-policy( $\omega_i, \delta_{N_i}, o_t, C_{\mathrm{loc},i}(t)$)具体如下：

折扣因子 $\beta \in [0,1)$，$\alpha_t = \frac{1}{t}$ 表示学习率，设：

$z_0^{\phi_i} = z_{\text{new}}^{\phi_i} = 0$, $z_0^{\theta_i} = z_{\text{new}}^{\theta_i} = 0$； $\Delta_0^{\theta_i} = \Delta_{\text{new}}^{\theta_i} = 0$, $\Delta_0^{\phi_i} = \Delta_{\text{new}}^{\phi_i} = 0$； $\phi_{i,\text{new}} = 0, \theta_{i,\text{new}} = 0$；

其中，$z_0^{\theta_i}, z_{\text{new}}^{\theta_i}, \Delta_0^{\theta_i}, \Delta_{\text{new}}^{\theta_i}, \theta_{i,\text{new}} \in R^{n_{\theta_i}}$； $z_0^{\phi_i}, z_{\text{new}}^{\phi_i}, \Delta_0^{\phi_i}, \Delta_{\text{new}}^{\phi_i}, \phi_{i,\text{new}} \in R^{n_{\phi_i}}$；

$$z_{\text{new}}^{\phi_i} = \beta z_t^{\phi_i} + \frac{\nabla f_i(g_{t+1} \mid \phi_i, g_t, o_i(t), a_i(t-1), \delta_{N_i})}{f_i(g_{t+1} \mid \phi_i, g_t, o_i(t), a_i(t-1), \delta_{N_i})}; \quad z_{\text{new}}^{\theta_i} = \beta z_t^{\theta_i} + \frac{\nabla \mu_i(a_i(t) \mid \theta_i, g_{t+1}, o_i(t), \delta_{N_i})}{\mu_i(a_i(t) \mid \theta_i, g_{t+1}, o_i(t), \delta_{N_i})};$$

$$\Delta_{\text{new}}^{\phi_i} = \Delta_t^{\phi_i} + \frac{1}{t+1}\left[C_{\text{loc},i}(t) z_{\text{new}}^{\phi_i} - \Delta_t^{\phi_i}\right]; \quad \Delta_{\text{new}}^{\theta_i} = \Delta_t^{\theta_i} + \frac{1}{t+1}\left[C_{\text{loc},i}(t) z_{\text{new}}^{\theta_i} - \Delta_t^{\theta_i}\right];$$

$\phi_{i,\text{new}} = \phi_i - \alpha_{t+1}\Delta_{\text{new}}^{\phi_i}$； $\theta_{i,\text{new}} = \theta_i - \alpha_{t+1}\Delta_{\text{new}}^{\theta_i}$；

返回 $\omega_i^* = \left\langle \phi_{i,\text{new}}, \theta_{i,\text{new}} \right\rangle$。

## 3 仿真结果分析

本节通过仿真实验验证了上述分布式合作策略搜索算法的有效性，并在授权用户容量损失、CR 用户信道容量和等方面进行了性能分析与比较。

对于 CR 用户独立检测并随机选择空闲频谱进行通信的接入策略，本文称为随机接入策略。对于处于同一区域系统中的各个 CR 用户，最理想的联合信道接入策略，本文称为最优接入策略。

图 3 给出了在子信道数与检测概率 $P_{Di}$ 分别相同的情况下，随机接入策略与本文的接入算法对授权用户容量损失的影响。因为各个 CR 用户独立进行信道选择，随着 CR 用户数的增加，随机接入策略引起的主用户容量损失越来越大。本文算法由于以主用户容量损失作为信道接入的约束条件，相邻 CR 用户不断交换本地观测信息，通过合作得到较优的联合接入策略，所以主用户的容量损失没有随着 CR 用户数的增加而有较大增长，被控制在较低的水平。

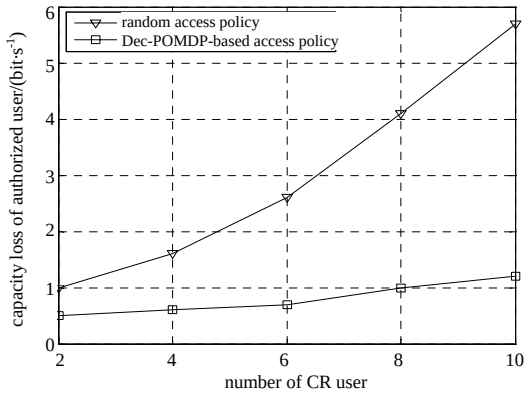



Fig.3 Capacity loss of authorized user
图 3 授权用户的容量损失比较

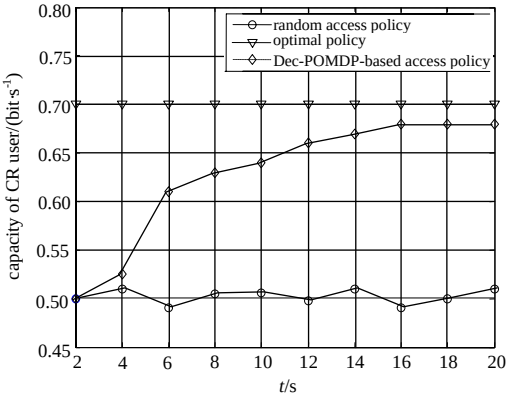



Fig.4 Total capacity of CR users
图 4 CR 用户信道容量和

图 4 给出了不同算法下，CR 系统所有用户信道容量和的比较。本文提出的算法下，CR 用户获得的信道容量高于随机接入策略下获得的容量，随着时间的增加，前者不断逼近于最优接入策略下获得的信道容量，而随机策略下的信道容量变化不大。图 5 给出了检测概率 $P_{Di}$ 对 CR 获得的系统容量和的影响。在随机接入策略、最优接入策略和本文提出的算法下，CR 的信道容量和都随 $P_{Di}$ 的增大而增加，但与随机接入策略相比，本文提出的算法下 CR 信道容量和的增加更大，这表明在本文的接入算法下 CR 对空闲频谱的利用效率更高，这是因为前者随机接入的特点会引起不同 CR 用户间的信道冲突，造成一部分空闲频谱的浪费。

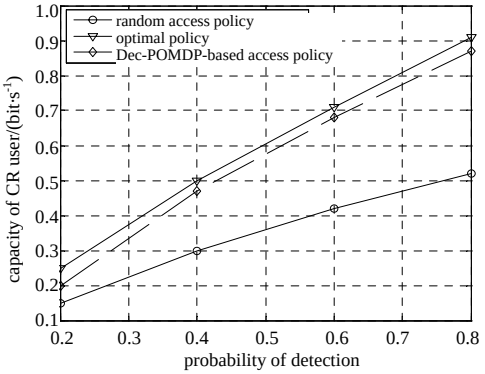



Fig.5 Capacity of CR users by varying probability of detection
图 5 检测概率不同时 CR 用户的信道容量和

## 4　结论

本文以 Dec-POMDP 为基础构造了认知无线电网络频谱分配模型，提出了一种分布式合作的信道接入算法。通过相邻用户的策略交互，CR 用户不再是单独进行接入决策，而是在区域策略梯度的方向上搜索最优联合接入策略，从而减少了信道接入冲突，提高了频谱利用效率。仿真结果表明，该算法有效地降低了 CR 用户引起的授权用户容量损失，提高了空闲频谱的利用效率，随着时间的增加逐渐达到接近于最优接入策略的信道容量。

**参考文献：**


[1] Mitola J. Cognitive radio for flexible mobile multimedia communications[C]// Proc. of 1999 IEEE International Workshop on Mobile Multimedia Communications. San Diego,CA,USA:[s.n.], 1999.

[2] McHenry M. Spectrum Occupancy Measurements[EB/OL]. (2005–08). http://www.sharedspectrum.com/?section=sf_ summary, 2005.

[3] Cabric D,Mishra S M,Brodersen R W. Implementation issues in spectrum sensing for cognitive radio[C]// Proc.38th Asilomar Conference on Signals,Systems and Computers. California University:[s.n.], 2004:772–776.

[4] 陈大海,张健,向敬成. 软件无线电体系结构研究[J]. 信息与电子工程, 2003,1(4):318–322.

[5] Zhao Q,Tong L,Swami A,et al. Chen Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework[J]. Journal of Selected Areas in Communications, 2007,25(3):589–600.

[6] Pynadath D V,Tambe M. The communicative multiagent team decision problem: Analyzing teamwork theories and models[J]. Journal of Artificial Intelligence Research, 2002,16(1):389–423.

[7] Bernstein D S,Givan R,Immerman N,et al. The Complexity of Decentralized Control of Markov Decision Process[J]. Mathematics of Operations Research, 2002,27(4):819–840.

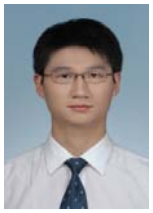
[8] Aberdeen D,Baxter J. Policy-Gradient Algorithms for Partially Observable Markov Decision Process[D]. Australia:Australian National University, 2003.

[9] Jaakkola T,Singh S P,Jordan M I. Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems[J]. Advances in Neural Information Processing Systems, 1995,7:345–352.

[10] Peshkin L M. Reinforcement Learning by Policy Search[D]. USA:Brown University, 2002.

[11] Aberdeen D,Baxter J. Scaling Internal-State Policy-Gradient Methods for POMDPs[C]// Proc. 19th International Conf. on Machine Learning. Sydney,Australia:[s.n.], 2002:1–12.


**作者简介：**




**张迎晓**(1984–)，男，山东省莱芜市人，在读硕士研究生，主要从事认知无线电动态频谱分配方面的研究.email:yingxiao1984@gmail.com.

**杨　涛**(1970–)，男，陕西省汉中市人，副教授，主要从事宽带无线通信理论及信号处理，认知无线电频谱感知及分配等方面的研究.

**胡　波**(1968–)，男，江苏省常州市人，教授，主要从事数字信号处理，电路分析与设计等方面的研究.

**陈光梦**(1950–)，男，上海市人，教授，主要从事数字信号处理、电路分析与设计等方面的研究.