ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 6,Nov. 2018,pp. 1003－1012
DOI:10.16337/j.1004-9037.2018.06.008

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于两步决策与 ε-greedy 探索的增强学习频谱分配算法

尹之杰　汪一鸣　吴　澄

(苏州大学轨道交通学院,苏州,215131)

摘　要:在认知无线网络中,认知基站需要进行频谱管理来提升非授权用户的服务质量。基站在寻找频谱空洞分配给非授权用户的过程中,需要做出最好的选择,但极可能是局部最优解,从而造成非授权用户频繁的频谱切换和吞吐率的下降。针对此问题,本文提出基于两步决策与探索的集中式增强学习频谱分配算法。通过设计新型状态动作集,认知基站进行信道分配的两步决策,并应用探索模式,解决认知基站在增强学习过程中探索环境和利用经验进行决策的平衡问题,防止决策的局部最优,提升频谱管理的性能。仿真结果表明,该算法在提升非授权用户吞吐率以及降低频谱切换方面明显优于现有的一些频谱分配策略。

关键词:认知无线网络;认知基站;频谱管理;动态频谱接入;增强学习;ε-greedy 探索策略

中图分类号:TN929.5　　**文献标志码**:A

Double-Step Decision Reinforcement Learning Spectrum Management Using ε-greedy Exploration

Yin Zhijie, Wang Yiming, Wu Cheng

(School of Railway Transportation, Soochow University, Suzhou, 215131, China)

Abstract: In cognitive radio network environment, the base station needs to carry out an effective spectrum management policy to guarantee the licensed user's communication and to improve the quality of service of the cognitive radio users at the same time. In the process of allocating spectrum holes to cognitive radio users, the base station faces massive passive channel switching due to the unpredictability of the licensed user and it results in the throughput of cognitive radio users' degradation. To solve this problem, this paper proposes a novel base station-cognitive base station, which contains reinforcement learning model with novel state and action sets. The cognitive base station can perform two-step decision of channel allocation, that is, whether to switch the channel for cognitive radio users and how to select the best channel if the cognitive base station decides to switch, so as to avoid excessive channel switching and improve the throughput of the cognitive radio user. Also, the performance of reinforcement learning spectrum management policy highly depends on the exploration of environment. In this paper, epsilon-greedy exploration method is used to solve the balance problem of cognitive base station in exploring the unknown environment and exploiting the existing knowledge. Simulation results show that the implementation of the epsilon-greedy in each decision step has a remarkable effect on the system performance. Al-

基金项目:国家自然科学基金(61471252)资助项目。
收稿日期:2017-03-26;**修订日期**:2017-09-25

so, we set up the best evaluation of a combination of two-step epsilon so that the proposed method is superior to traditional reinforcement learning spectrum allocation scheme in improving cognitive radio users' throughput and reducing channel switching.

Key words: cognitive radio; cognitive base station; spectrum management; dynamic spectrum access; reinforcement learning; ε-greedy exploration strategy

引　言

随着无线通信业务的不断拓展和增长,频谱资源的匮乏已成为现阶段面临的一个严峻问题。为此,美国联邦通信委员会(FCC)在2002年成立了频谱政策特别工作组,指出现有的固定频谱分配方式已成为无线通信发展的阻碍。随着科技进步以及地区因素变化,这些被固定分配的频段并非全天满负荷运行,甚至有些频段已极少或不再被使用,已造成严重的资源浪费。例如美国已被弃用的电视频段698～806 MHz[1]。针对该问题,Mitola曾在博士论文中提出了认知无线电的概念[2],Haykin对认知无线电做了进一步的研究,提出了在认知无线电中有待发展的各个方面,并指出有效的频谱管理对提升频谱利用率有至关重要的作用[3]。

现阶段频谱管理模型的研究分为集中式和分布式两种[4]。集中式模型由基站独立感知频谱,对频谱空洞统一分配。该模型优点是基站收集全局信息独立工作,不受其他信息干扰,非授权用户不需要具备感知频谱的能力。缺点是基站内部功能复杂,需要强大的计算能力[5]。在分布式模型中,基站与用户协商合作,进行频谱空洞分配。这种模型可以显著降低基站负载,缺点是基站与用户须遵循固有的协商策略,这些策略较难制定[6-7],网络内的非授权用户须具备感知频谱和用户间协作的能力。

在频谱管理模型中,研究的一个重点是信道分配。针对这一问题,研究者提出了大量的方法来提高非授权用户的服务质量[8]。研究普遍选取吞吐率或系统传输成功率作为一种系统性能的评判标准[9]。但在授权用户频发的认知无线网络中,非授权用户需要进行频谱切换以避免干扰其通信,但频繁的频谱切换不仅降低自身的吞吐率,还会造成许多其他的开销[10-14]。所以信道切换次数也应是服务质量的重要评判标准。

增强学习是解决频谱感知、接入和共享问题的一种有效途径。在认知无线网络环境的信道分配过程中应用增强学习已被众多文献证明可以提高非授权用户的成功传输率[15-18]。在具体建立增强学习模型的过程中,有两个关键问题。一是如何定义环境状态和智能体动作。复杂的状态动作集会导致计算量庞大甚至维数灾难[19]。二是智能体如何在探索环境和开采知识之间获得平衡,选择生成问题最优解的最佳度量标准。该问题在机器学习领域已被深入研究[20],但在认知无线电领域中仍值得探讨。

对于上述问题,本文采用集中式频谱管理模型,在对信道分配的研究中,以降低模型难度和提升非授权用户服务质量为目标,提出基于两步决策的新型增强学习认知基站。首先,通过对状态动作集的充分利用,在原有的决策过程中,增加了一次以降低信道切换次数的为目的的决策。该步决策决定认知基站是否需要切换信道提供服务。当决定切换后,再进行第二步信道选择决策。其次,本文引入ε-greedy方法对两步决策进行有效的优化,避免贪婪选择落入局部最优。实验证明,基于此算法的认知基站在提高非授权用户服务质量方面具有有效的作用。

1　认知基站工作方式

1.1　认知基站模型

本文提出一种基于机器学习模型的新型认知基站。其功能是在保障授权用户的通信不受干扰的情

况下,发现并分配频谱空洞给覆盖范围内的非授权用户。采用集中式频谱管理模式的认知基站具有频谱感知,频谱决策和分配的功能。本文研究重点是频谱决策和分配过程,所以认知基站的频谱感知功能被假设为理想,不存在错误感知授权用户行为的可能。在与非授权用户通信的过程中,认知基站采用时槽结构的数据通信方式。在一个时槽 T_{slot} 内,认知基站需在 τ_{sensing} 时间内感知该信道上是否有授权用户的存在。之后根据授权用户的占用情况,在剩余时间 $T_{\text{slot}}-\tau_{\text{sensing}}$ 内做出与非授权用户数据传输、命令其退避等待或者与其在另一条信道上重新建立连接的动作。图 1 描述了认知基站与非授权用户通信的一个时槽内的时间结构分配,图 2 描述了认知基站与非授权用户通信的方式。

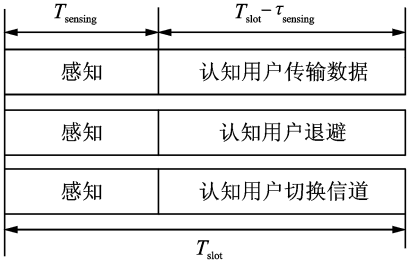


图 1 认知基站与非授权用户通信的一个时槽内部的时间分配结构的 3 种不同情况

Fig. 1 Slot structure of the transmission between cognitive base station and secondary user

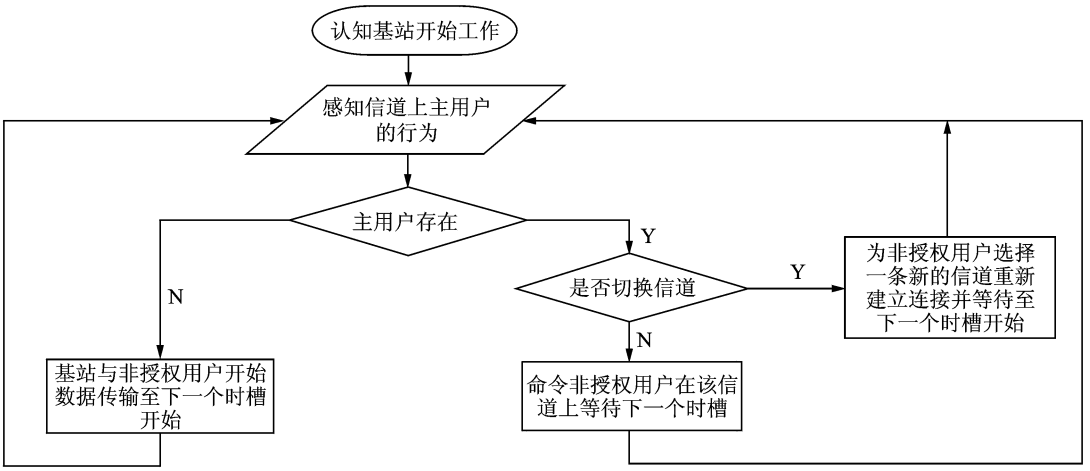


图 2 认知基站与非授权用户通信的方式

Fig. 2 Communication model of cognitive base station

1.2 授权用户模型

授权用户作为授权频段主要使用者,占用信道的时间模型选择对网络环境的真实性有重要的影响。本文中采用连续时间马尔科夫模型描述授权用户对信道占用的情况,其到达或离开授权信道后经过一段指数分布的时间后状态转移

$$\begin{aligned} f(T_{\text{busy}}^{k};\lambda_{\text{busy}}) &= \lambda_{\text{busy}} \mathrm{e}^{-\lambda_{\text{busy}} T_{\text{busy}}^{k}} \\ f(T_{\text{idle}}^{k};\lambda_{\text{busy}}) &= \lambda_{\text{idle}} \mathrm{e}^{-\lambda_{\text{idle}} T_{\text{idle}}^{k}} \end{aligned} \tag{1}$$

式中,T_{busy} 代表授权用户转移到占用状态(Busy)后经过的时间,T_{busy} 代表其转移到空闲状态(Idle)后经过的时间,均服从指数分布。λ_{busy},λ_{idle} 是指数分布参数。授权用户依概率 p,q 进行状态转移的过程如图 3 所示。

1.3 增强学习及在本文中的应用

增强学习提供了一种在学习的过程中进行决策的可能。智能体无须经历大量样本的监督训练之后

才能工作。这样的学习模式更适合在复杂的未知认知无线电环境中应用。

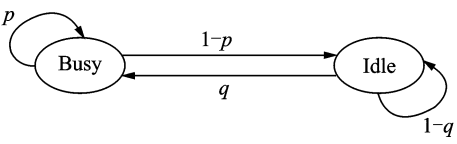


图 3 授权用户状态转移示意图

Fig. 3 Primary user state transition

增强学习的基本模型为$\{S,A,T,R\}$，其中$S=\{s_1,s_2,\cdots\}$，代表环境状态空间，$A=\{a_1,a_2,\cdots\}$代表智能体的动作空间，$T:s*a\rightarrow s'$代表当前状态下，采取动作之后得到的下一状态，$R:s*a*s'\rightarrow r$代表在当前状态s下执行动作转移到状态s'时获得的立即回报值r。

定义状态和动作是集中式频谱管理高效工作的关键。本文以最大化非授权用户吞吐率以及最小化频谱切换次数两个目标进行建模。

首先，将认知基站视作智能体，其覆盖范围视作所处的环境。状态空间S由基站正在提供服务的信道组成

$$ch=\{ch_1,ch_2,ch_3,\cdots,ch_M\} \tag{2}$$

在当前信道上考虑第一步决策，即是否需要更换信道提供服务。对于基站，在时刻的观测状态为

$$s_t=(ch)_t \tag{3}$$

基站在给定时间t时刻的状态下，定义其动作，有

$$a_t=\{k\}_t \tag{4}$$

将 switch_channel 表示为k_1，代表基站更换服务信道，在该时槽内完成状态转移之后，等待后续时槽开始后，重新感知信道的状态。将 stay 表示为k_2，代表认知基站无论授权用户状态如何，均在原信道提供服务。有

$$k_t\in k=\{\text{switch\_channel},\text{stay}\} \tag{5}$$

立即回报值R选取是根据基站的决策对非授权用户服务质量的影响来决定的。立即回报值的给予如下所示：

(1)当基站感知到服务信道ch_x上授权用户活跃，选择动作k_2保持在ch_x上服务，下一状态仍是ch_x，此时槽无法进行数据传输，则给予-1的惩罚值。

(2)当基站在本时槽内没有感知到服务信道ch_x上有授权用户活跃，则进行传输数据，状态转移后仍是ch_x。将给予当前状态ch_x下选择k_2一个$+1$的奖励值。

(3)当基站在感知到服务信道ch_x上授权用户活跃，选择动作，进入第二步决策后更换至信道ch_y提供服务，认知基站的状态转移至下一信道ch_y。此时认知基站与非授权用户在信道ch_y上重新建立连接并等待下一个时槽的开始，感知ch_y授权用户状态。如果活跃，记作一次失败的切换，则给予-2的惩罚值。如果可以传输数据，则记作一次成功的切换，给予$+1$的奖励回报值。设定-2的惩罚回报值是因为认知基站在切换信道之后，仍无法继续传输，将浪费两个时槽的传输时间。

定义完成后，就有$n*2$组状态动作组合。认知基站使用Q表来累计每组状态动作组合的回报值，累计回报值的方法基于下式[21]

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r_t+\gamma\max_a Q(s_{t+1},a_{t+1})-Q(s_t,a_t)\right] \tag{6}$$

式中：s_t是基站在t时刻的服务信道，s_{t+1}是转移之后的信道；a_t代表基站采取的动作；α是学习速率；r_t是立即回报值；γ，$0\leqslant\gamma\leqslant1$是折现因子，是未来的回报值对现在的影响程度。

在决策过程中，智能体依据的是其所维护Q表当中的$Q(s_t,a_t)$，即累计回报值。智能体根据这些值来做出决策π

$$Q_\pi(s_t,a_t)=E_\pi\{R_t\mid s_t=ch_i,a_t=k_t\}=E_\pi\{\sum_t\gamma r_t\mid s_t=ch_i,a_t=k_t\} \tag{7}$$

式中：E_π是在任意时刻智能体在所处信道ch_i上选择动作k_t可获得的立即回报值r_t。智能体决策所期

望的是全局奖励最大化。所以后续动作也应对目前的决策产生影响。由折现因子 γ 控制的目前决策对未来奖励的依赖程度也应列入考虑。

当认知基站感知到在当前信道上有授权用户活跃,便在可行的动作 k_1 和 k_2 中选取基于累计回报值的最优决策,即第 1 步决策。当选择更换信道,目标信道则根据其学习结果选取。即第 2 步决策。本文中的所有累计回报值均以矩阵的形式记录在认知基站之中。第 1 步决策比较当前信道上离开或是停留的累计回报值。第 2 步决策比较在其他信道上停留的累计回报值。这样 Q 表就得以充分利用。

2 ε-greedy 探索

在未知无线环境中,认知基站选择的动作是否最优是不确定的。选择一个局部最优信道而非全局最优信道提供服务,可能会在授权用户突发时,引起非授权用户不必要的滞留或是频谱切换。因此平衡增强学习的探索和利用的至关重要。本文使用 ε-greedy 算法来保证认知基站探索环境的同时也保证决策的质量。应用 ε-greedy 之后的增强学习认知基站在进行第一步决策时,认知基站当前状态 s_t 下进行是否离开当前信道的决策。为防止滞留在局部最优信道,做出第一步决策 π_1 依据

$$\pi_1=\begin{cases}\arg\max\limits_{k_i} Q(ch_i,\ k) & \xi<\varepsilon_1\\ \text{random action from } k & \text{其他}\end{cases} \tag{8}$$

式中 ξ 是一个在 0～1 之间服从均匀分布的随机变量,在每次决策之前随机选取。ε_1,$0\leqslant\varepsilon_1\leqslant1$ 是恒定的探索参数。

当认知基站选择离开当前信道,则需选择切换目标。此时应以一定的概率去随机选择信道以避免贪婪地选择局部最优。做出第 2 步决策 π_2 依据的是

$$\pi_2=\begin{cases}\arg\max\limits_{ch} Q(ch',k_2) & \eta<\varepsilon_2\\ \text{random channel from } ch' & \text{其他}\end{cases} \tag{9}$$

式中 $Q(ch',k_2)$ 是认知基站在所有信道上选择的累计回报值,η 是一个在 0～1 之间服从均匀分布的随机变量,在决策之前随机选取,ε_2,$0\leqslant\varepsilon_2\leqslant1$ 是恒定探索参数。ch' 是不包含当前信道的其余所有信道的集合。当认知基站服务信道上没有授权用户出现时,$Q(s,k_2)$ 会一直增加。其大小可以作为信道优劣的考量。

3 实验及结果分析

为了验证算法的有效性,本节针对算法的每一个模块进行测试。首先选定第一步决策的 ε 参数进行测试,检验第二步决策探索参数对系统性能的影响情况。之后以找出最佳 ε 参数组合为目的,给出对于 ε 值组合的尝试。最后,在确定最佳的 ε 取值组合后,对认知基站进行训练,将本文提出算法的训练结果与随机与轮询分配模型、传统增强学习模型[21]、贪婪的增强学习模型进行比较。

3.1 实验设计

仿真实验平台选择通信网络离散事件模拟器 NS-3。场景是在 1 个认知基站覆盖范围内,有 10 条相同带宽的授权信道,10 条授权信道由 10 个服从连续时间马尔科夫过程的授权用户分别占用,范围内存在 1 个一直有数据待发送的非授权用户。认知基站负责利用空闲的授权信道与非授权用户通信。仿真时间为 2 000 s。服务质量指标设置为吞吐率和信道切换次数。仿真参数见表 1。

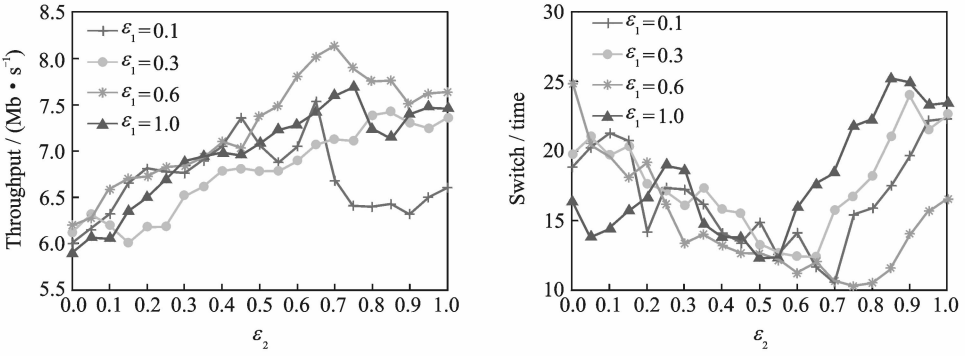
表 1 仿真参数

Tab. 1 Simulation parameters

RL parameter	Value
α	0.01 when $r_t>0$ 0.05 when $r_t<0$
γ	0.01
PU model parameter	**Value**
λ_{busy}	0.02
λ_{idle}	0.02
p_1—p_{10}	0.1,0.2,0.3,0.4,0.5,0.5,0.6,0.7,0.8,0.9
q_1—q_{10}	0.1,0.2,0.3,0.4,0.5,0.5,0.6,0.7,0.8,0.9

3.2 对于参数 ε_1 和 ε_2 的有效性实验

为了验证加入 ε-greedy 探索的必要性，先将参数 ε_1 分别设置为 0.1,0.3,0.6 和 1，观察并比较在不同的 ε_1 下，非授权用户服务质量随 ε_2 的变化情况。结果如图 4(a)和图 4(b)所示。当 $\varepsilon_1=1$ 时，是否切换信道依据贪婪方式选择。此时，可以单独观测参数 ε_2 对系统性能的影响。首先，从图 4(a)和图 4(b)中 $\varepsilon_1=1$ 的曲线可知，吞吐率的峰值出现在 $\varepsilon_2=0.75$ 时，值为 7.63 Mb/s。信道切换最小次数出现在 $\varepsilon_2=0.5$时，平均值为 11.9 次。均优于 $\varepsilon_2=1$ 时的系统性能(7.48 Mb/s，23.6 次)。相同的，观察 $\varepsilon_1=0.1,0.3,0.6$ 时的系统性能曲线，最高吞吐率和最低信道切换次数均没有出现在 $\varepsilon_2=1$ 时。其次，从图 4(a)中可知，$\varepsilon_1=0.6$ 这条曲线明显高于其他曲线，而 $\varepsilon_1=0.1,0.3$ 这两条曲线却普遍低于 $\varepsilon_1=1$。而在图 4(b)中，也反映了相同的情况。当 $\varepsilon_1=0.6$ 时，信道切换次数普遍低于其他 3 条曲线。出现上述情况的原因是贪婪决策可能会导致无法找到全局最优信道，引起非授权用户不必要的停留。并且不恰当的探索参数选择，会导致认知基站决策过于偏向随机或者是贪婪，影响系统的性能。所以，选取合适的探索参数，可以使得全局最优信道更早被发现。



(a) 给定ε_1吞吐率随ε_2变化情况
(a) Throughput indicator varies with ε_2 using fixed ε_1

(b) 给定ε_1信道切换次数随ε_2变化情况
(b) Channel switch times indicator varies with ε_2 using fixed ε_1

图 4 小量贪婪参数对系统性能的有效性实验结果

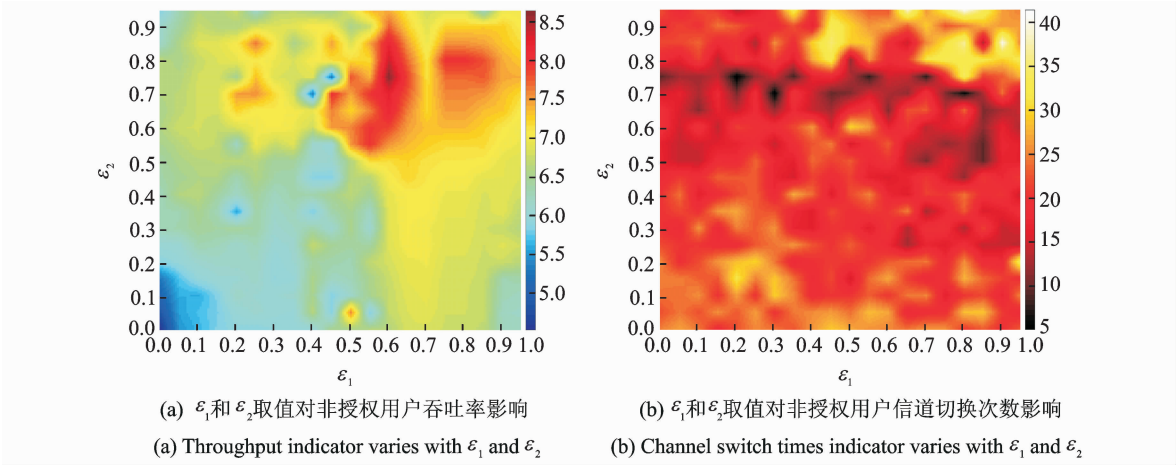
Fig. 4 Experimental results of the validity of epsilon-greedy parameters on system performance

实验结果表明，在有效的探索下，系统的性能会明显优于贪婪决策，而不恰当的探索会降低系统性能。

3.3 选取最佳的取值组合实验

3.2 节对 ε_1 和 ε_2 的有效性进行了单独的分析。从图 4 中可以得知,虽然全局吞吐率最高值出现在 $\varepsilon_1=0.6$,$\varepsilon_2=0.7$ 时,为 8.13 Mb/s,且 $\varepsilon_1=0.6$ 的取值普遍优于其他取值,但依然存在性能劣于其他取值的区间。所以寻找能使系统性能最佳化的参数组合至关重要。因此设置 ε_1 和 ε_2 的取值从 0～1,间隔为 0.05 以测试服务质量。系统吞吐率和信道切换次数随 ε_1 和 ε_2 取值的变化情况如图 5(a),(b)所示。为了能突出较好的取值组合,本文将实验结果绘制成热力图,以便观察最佳性能出现的位置。图 5(a)红色区域是吞吐率出现峰值的位置,位于 $\varepsilon_1=0.6$,$\varepsilon_2=0.75$ 时,吞吐率的较高的区域集中在峰值周围,探索参数相对这一取值增加或减小后,吞吐率均产生下降。图 5(b)中的分布的黑色暗区域是信道切换次数低的区域,集中在 ε_2 取值为 0.5～0.8 左右,离开此区域后,切换次数明显上升,说明 ε_2 取值对其影响有偏重。图 5(a)中的吞吐率峰值区域小于图 5(b)中切换次数低值区域是因为认知基站在过度随机或贪婪的情况下,被迫滞留在局部最优信道,无法获得高吞吐率。

综合图 4 与图 5,选取 $\varepsilon_1=0.6$,$\varepsilon_2=0.75$ 来训练认知基站,可以获得最佳的系统性能。



(a) ε_1和ε_2取值对非授权用户吞吐率影响

(a) Throughput indicator varies with ε_1 and ε_2

(b) ε_1和ε_2取值对非授权用户信道切换次数影响

(b) Channel switch times indicator varies with ε_1 and ε_2

图 5 最佳小量贪婪参数组合的选取实验结果

Fig. 5 Selection experiment results of optimal combination of ε-greedy parameters

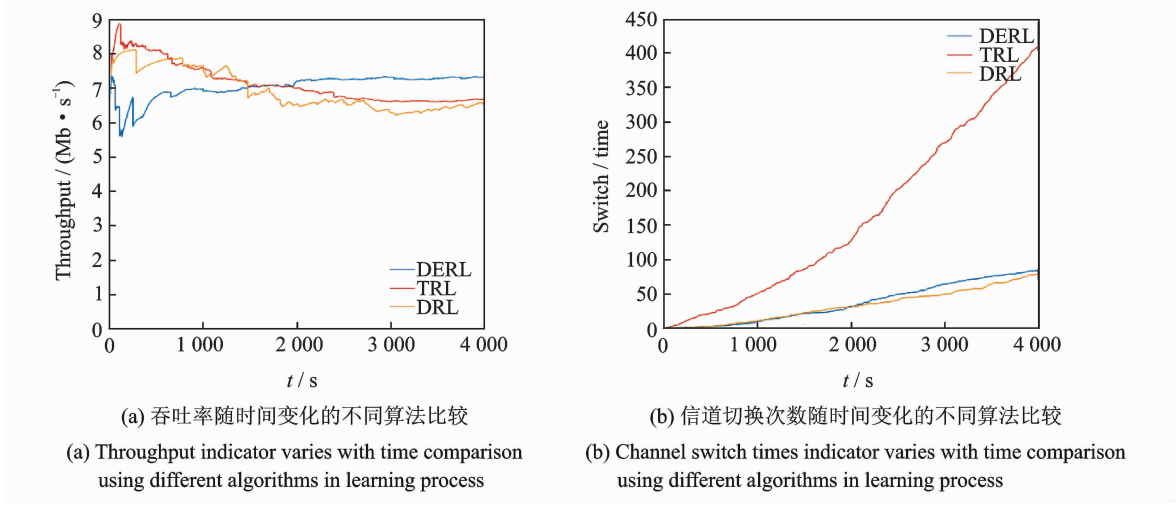
3.4 与其他两种频谱管理策略的性能比较

根据上节实验所得出最佳 ε_1 和 ε_2 取值的组合,对认知基站信道分配进行时长为 4 000 s 的仿真,结果与文献[21]中使用的基于复杂状态动作集的 Q 学习算法和文献[15]与文献[20]中所使用的无状态 Q 学习算法进行比较。文献[21]中所提出的增强学习方式,将智能体环境状态设置为所处信道,但动作却细化到切换至具体的信道。此种方式可以较为精确的规划信道切换路径,却构造了一个平方级的复杂 Q 值矩阵,有待探索的区域非常庞大,且该文献并未提及对状态动作集合的探索问题。而文献[15]中,其智能体可采取的动作为切换信道和切换功率等级,由于本文中假设基站和非授权用户位置相对静止,所以功率等级不发生改变,仅考虑信道切换[20]。本文将两步决策 ε-greedy 增强学习方法命名为 DERL,而文献[15,20]使用的无状态 Q 学习称作 DRL,文献[21]提出的算法称为 TRL。比较结果如图 6(a)和图 6(b)所示。

从图 6(a)可以看出,所比较的 3 种方法 DERL、DRL、TRL 的吞吐率变化过程均可分为两个阶段。第一阶段是学习阶段,采用不同算法的基站,呈现出不同程度的振荡。而在仿真时间达到 1 500 s 左右,进入第二阶段,此阶段性能指标趋向于稳定,由于 DERL 的方法在第一个阶段进行了较好的探索。所以非授权用户的传输被分配到全局最佳信道,吞吐率在经过学习阶段之后有明显的上升。而 DRL 和

TRL 算法进行贪婪决策，导致对信道环境探索的不完全，认知基站在局部最优信道上过早的停留。这样的决策方式，可以较快地使非授权用户获得较高的吞吐率，但由于局部最优信道的授权用户出现更为频繁，导致传输失败的可能性变大，吞吐率在第二阶段出现下降。所以，本文提出的 DERL 算法可以使非授权用户获得优于其余两种算法更好的吞吐率。

从图 6(b)可以较为明显地看出，TRL 算法的非授权用户信道切换次数受庞大的状态动作集影响，增加得非常快。因为没有信道切换保护的 TRL 算法会选择立即离开当前信道。而且 TRL 算法准备离开一条信道之后，会有 9 个信道可以选择，但在每个不同状态均会面临 9 个不同动作。此时，探索的不完备就会导致无法找到全局最优，只能继续试错。DRL 算法使用了简易动作集之后，优化了过多的状态动作对的问题。但贪婪的决策过程使得 DRL 无法稳定地选取全局最优信道。在信道切换保护机制下，DRL 可能会造成不必要的传输堵塞。所以 DRL 虽获得了比本文提出算法 DERL 更低的信道切换次数，但没有获得更高的吞吐率。



(a) 吞吐率随时间变化的不同算法比较

(a) Throughput indicator varies with time comparison using different algorithms in learning process

(b) 信道切换次数随时间变化的不同算法比较

(b) Channel switch times indicator varies with time comparison using different algorithms in learning process

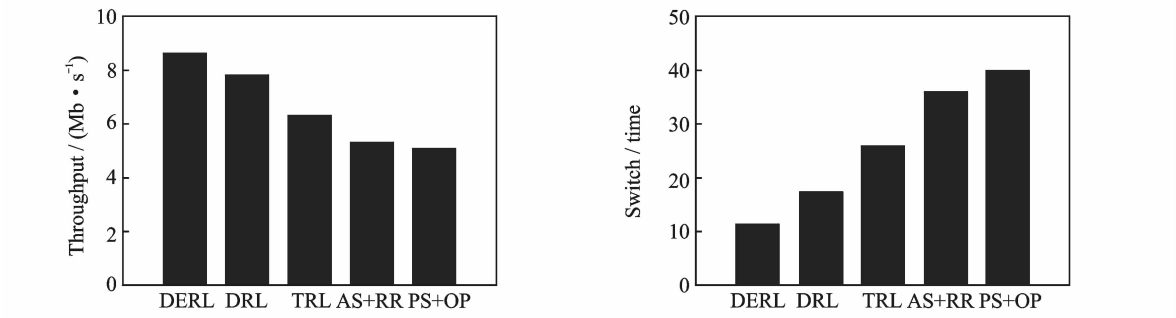
图 6 几种不同算法的训练过程比较结果

Fig. 6 Comparison results of training process of several different algorithms

综上所述，针对较为复杂的认知无线网络环境，构造状态动作集的数量级和决策方式非常关键。本文中探索方式和较为简单的状态动作集，使非授权用户获得了更好的服务质量。

图 7(a)和图 7(b)显示了在仿真时间为 2 000 s 的时间内，本文中提出算法与 DRL，TRL，以及两种基础方法的性能比较。两种基础方法的第 1 步决策分为总是选择切换的称为 AS，和以一定概率 P_r 选择切换，否则退避等待的 PS。第 2 步决策时随机选择信道接入称为 OP，轮询选择信道接入称为 RR。其中，概率切换的参数 P_r 经过测试，本文选取的是可以使非授权用户获得最佳服务质量的概率 $P_r=0.8$。

从图 7(a)和图 7(b)中看出，在认知基站选定最佳探索参数组合之后，通信的吞吐率以及频谱切换次数均优于其他的方法，吞吐率达到了 8.63 Mb/s，信道切换次数为 12 次。无状态 Q 学习模型测试所得结果为 7.83 Mb/s，16 次。由于无状态 Q 学习仅设置智能体可采取的动作，而不设置状态，使得 Q 得到极大的简化。但缺点是在学习时受到惩罚将使其马上采取行动。虽然寻找全局最佳信道的速度较快，但在最佳信道收敛时，一旦与授权用户通信发生冲突，则会立即切换至其他信道。而在文献[21]提出的复杂的状态动作集构建的增强学习模型下，测试结果为 6.59 Mb/s，26 次。面对本文中设置的较为复杂的授权用户模型，TRL 性能大幅下降。因为当信道数量增加从 5 增至 10 条时，其 Q 值则由 25 个



(a) 使用不同频谱管理方法的系统吞吐率

(a) Final throughput comparison between different spectrum management methods after learning process

(b) 使用不同频谱管理方法的系统切频次数

(b) Final channel switch times comparison between different spectrum management methods after learning process

图 7 几种不同算法训练完成后的性能比较结果

Fig. 7 Performance comparison results of several different algorithms after training completion

状态动作组合扩展为 100 个。完备的探索 100 个状态动作组合直至收敛需要很长的时间。所以呈平方级增长的复杂状态动作集不适合应用在复杂的认知无线网络环境中。本文还选取了在认知无线电频谱分配中的两个传统方法与本文中提出的算法进行比较。在与授权用户发生冲突时立即切换并且以轮询方式接入信道的 AS+RR 频谱管理方法吞吐率要略高于概率切换和随机接入方式。对以 $P_r=0.8$ 进行概率切换之后随机选择信道接入的 PS+OP 频谱管理方法进行测试后发现,即使在第一步决策时以概率切换方式做出对频繁切换信道的避免,但该方法信道切换次数仍高于 AS+RR 方法的组合,这也反映出选择目标信道(第二步决策)对信道切换次数的偏重影响。

4 结束语

本文研究了频谱管理中出现的两个重要问题。第 1 个问题是在授权用户频发的环境中,如何避免过多的频谱切换对系统性能造成的危害,并提升系统的吞吐率。第 2 个问题是在应用增强学习到认知无线网络的过程中,如何解决探索以及利用的平衡问题。针对吞吐率和信道切换次数的双目标优化问题,本文给出了一种新型的多用途状态动作集。实验证明,运用该新的状态动作集的认知基站比一些传统的增强学习信道分配方式的认知基站在性能上有较大的提升。针对第二个探索与利用的平衡问题,我们给出了验证 ε-greedy 探索有效性的实验,在与贪婪决策的方法比较的过程中,平衡探索与利用的认知基站性能更好,证明了在认知无线网络中对环境探索的必要性。在两步都应用 ε-greedy 的认知基站性能结果分析中,本文发现了两个 ε 取值分别对不同优化目标的影响有各自的偏重,也找出了一组 ε 值,使得系统的性能相比其他的 ε 取值更为优异。实验结果证明了本文提出的算法在应用到认知无线网络环境的基站中进行频谱管理的有效性。

参考文献:

[1] Flores A B, Guerra R E, Knightly E W, et al. IEEE 802.11af: A standard for TV white space spectrum sharing[J]. IEEE Communications Magazine, 2013,51(10):92-100.

[2] Mitola J, Maguire G Q. Cognitive radio: Making software radios more personal[J]. IEEE Personal Communications, 1999, 6(4):13-18.

[3] Haykin S. Cognitive radio: Brain-empowered wireless communications[J]. IEEE Journal on Selected Areas in Communications, 2005,23(2):201-220.

[4] Salami G, Durowoju O, Attar A, et al. A comparison between the centralized and distributed approaches for spectrum man-

agement[J]. IEEE Communications Surveys and Tutorials, 2011,13(2):274-290.
[5] Peng C, Zheng H, Zhao B Y, et al. Utilization and fairness in spectrum assignment for opportunistic spectrum access[J]. Mobile Networks and Applications, 2006,11(4):555-576.
[6] Akyildiz I F, Lee W, Vuran M C, et al. Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey[J]. Computer Networks, 2006,50(13):2127-2159.
[7] Akyildiz I F, Lee W, Vuran M C, et al. A survey on spectrum management in cognitive radio networks[J]. IEEE Communications Magazine, 2008,46(4):40-48.
[8] Jia J, Zhang Q, Shen X S, et al. HC-MAC: A hardware-constrained cognitive MAC for efficient spectrum management[J]. IEEE Journal on Selected Areas in Communications, 2008,26(1):106-117.
[9] Chowdhury K R, Felice M D, Akyildiz I F, et al. TCP CRAHN: A transport control protocol for cognitive radio ad hoc networks[J]. IEEE Transactions on Mobile Computing, 2013,12(4):790-803.
[10] Feng X, Qu D, Zhu G, et al. Smart channel switching in cognitive radio networks[C]//International Congress on Image and Signal Processing. [S. l.]: IEEE, 2009:1-5.
[11] Kyasanur P, Vaidya N H. Capacity of multichannel wireless networks under the protocol model[J]. IEEE ACM Transactions on Networking, 2009,17(2):515-527.
[12] Kanodia V, Sabharwal A, Knightly E. MOAR: A multi-channel opportunistic auto-rate media access protocol for ad hoc networks[C]// International Conference on Broadband Networks. [S. l.]: IEEE, 2004:600-610.
[13] 陈兵,胡峰,朱琨. 认知无线电研究进展[J]. 数据采集与处理,2016,31(3):440-451.
Chen Bing, Hu Feng, Zhu Kun. Research progress of cognitive radio[J]. Journal of Data Acquisition and Processing, 2016, 31(3):440-451.
[14] 王慧锋,高瞻. 认知无线电网络中基于接收信号强度的定位算法[J]. 数据采集与处理,2014,29(3):465-471.
Wang Huifeng, Gao Zhan. Localization algorithm based on received signal strength compare for cognitive radio networks[J]. Journal of Data Acquisition and Processing,2014,29(3):465-471.
[15] Wu C, Chowdhury K R, Felice M D, et al. Spectrum management of cognitive radio using multi-agent reinforcement learning[J]. Adaptive Agents and Multi-agents Systems, 2010:1705-1712.
[16] Emre M, Gur G, Bayhan S, et al. Cooperative Q: Energy-efficient channel access based on cooperative reinforcement learning[C]//IEEE International Conference on Communication Workshop. IEEE, 2015:2799-2805.
[17] Galindo-Serrano A, Giupponi L. Distributed Q-learning for aggregated interference control in cognitive radio networks[J]. IEEE Transactions on Vehicular Technology, 2010,59(4):1823-1834.
[18] Ahmed A, Amal G, Anis S, et al. Resource allocation for multi-user cognitive radio systems using multi-agent Q-learning [C]//International Conference on Ambient Systems. Networks and Technologies. [S. l.]: [s. n.], 2012.
[19] Tangkaratt V, Morimoto J, Sugiyama M. Model-based reinforcement learning with dimension reduction[J]. Neural Networks the Official Journal of the International Neural Network Society, 2016,84:1-16.
[20] Morozs N, Clarke T, Grace D. Distributed heuristically accelerated Q-learningfor robust cognitive spectrum management in LTE cellular systems[J]. IEEE Transactions on Mobile Computing, 2016,15(4):817-825.
[21] Sutton R, Barto A. Reinforcement learning:An introduction[M]. [S. l.]: MIT Press, 1998.

作者简介:

尹之杰(1992-),男,硕士研究生,研究方向:认知无线电、机器学习。

汪一鸣(1956-),女,教授,博士生导师,中国电子学会高级会员,IEEE 会员,苏州大学通信与信息系统学科带头人之一,研究方向:无线通信网络、认知无线电、超宽带通信等。

吴澄(1978-),男,副教授,硕士生导师,研究方向:认知无线电、图像处理,E-mail: cwn@suda.edu.cn。

(编辑:张彤)