doi:10.11887/j.cn.202004005 http://journal.nudt.edu.cn

频谱感知次序的在线最优选择*

周 敏,王少尉
(南京大学 电子科学与工程学院,江苏 南京 210023)

摘 要:动态频谱接入是解决无线电频谱资源短缺和频谱使用效率低下问题的有效方法,它允许次级用户在授权频谱空闲时动态地接入,以进行数据传输。而频谱感知是实现动态频谱接入的关键挑战之一。由于次级用户的感知能力有限,为了获得更多的频谱接入机会,需要尽快找到频谱空闲概率最大的频段,并研究频谱感知次序问题。考虑到频谱空闲概率对次级用户是不可知的,并且会随时间变化,提出了在线学习框架,把频谱感知次序问题归纳成经典多摇臂赌博机问题,并利用在线学习方法——满意折现汤普森抽样算法处理优化问题。仿真结果表明,和其他算法相比,所提算法可以获得更多的频谱接入机会并且能够跟踪频谱空闲概率的变化。

关键词:动态频谱接入;频谱感知;在线学习;满意折现汤普森抽样
中图分类号:TN92 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2020)04-024-06

听语音 聊科研 与作者互动

Online optimal selection of spectrum sensing order

ZHOU Min, *WANG Shaowei*
(School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China)

Abstract: Dynamic spectrum access is deemed as an effective solution to the radio spectrum scarcity and spectrum usage in efficiency problem, which allows secondary users to access the spectrum dynamically for data transmission when the licensed spectrum is idle. However, spectrum sensing is one of the key challenges for dynamic spectrum access. Since the secondary user was equipped with limited sensing capability, in order to obtain more spectrum access opportunities, the spectrum sensing order problem was investigated to find the frequency band with the highest probability of being idle as soon as possible. Considering that the probability of the spectrum being idle was not available for the secondary users and changes over time, an online learning framework in which the spectrum sensing order problem was formulated as a classical multi-armed bandit problem was proposed, and it was addressed by using an online learning method, referred to as satisficing discounted Thompson sampling. Simulation results indicate that compared with other algorithms, the proposed algorithm yields more spectrum opportunities and can track the changes of the probability of the spectrum being idle.

Keywords: dynamic spectrum access; spectrum sensing; online learning; satisficing discounted Thompson sampling

大部分可用的频谱资源被授权给了特定的应用,如广播电视、移动通信等。在过去的几十年里,移动数据流量快速增长,引发了频谱资源短缺问题。此外,调查表明,授权给特定应用的频谱未被充分利用,部分授权频谱在时间和空间尺度上是空闲的,这导致了频谱的利用效率低下。在不干扰授权用户(主用户)的前提下,认知无线电允许次级用户动态地接入空闲的授权频谱进行数据传输[1-2]。动态频谱接入是解决频谱资源短缺和利用效率低问题的潜在方案[3],引起了广泛关注。

频谱感知是实现动态频谱接入的前提。然而,由于硬件的限制[4],次级用户无法在很宽的频谱范围内进行感知。因此,频谱感知次序在认知无线电系统中是一个重要的问题,次级用户需要在给定的时间内确定优先感知哪个频段,尽快找到空闲信道进行数据传输,从而缩短频谱感知时间,充分利用频谱资源。理想情况下,为了获得更多的频谱接入机会,次级用户应该优先感知频谱空闲概率最高的频段。在信道空闲概率已知的前提下,文献[5]提出了一种按照概率下降次序感知信道的方法。在信道空闲概率未知的情况下,文献[6]采用Q学习方法来提高系统性能。

* **收稿日期**:2019-12-25
基金项目:国家自然科学基金资助项目(61671233, 61801208, 61931023, U1936202)
作者简介:周敏(1995—),女,江苏南通人,博士研究生,E-mail:18351885760@163.com;
王少尉(通信作者),男,教授,博士,博士生导师,E-mail:wangsw@nju.edu.cn

文献[7]提出了一种两阶段学习算法,该算法利用强化学习进行信道选择,减少感知信道的时间。

然而,上述方法无法同时处理以下两种情况:给定频段的频谱空闲概率通常是不可知的;即使可以通过长时间的观察来预测频谱空闲概率,这一概率也随时会发生变化。在这种情况下,可能会使用过时的信息进行决策,导致系统性能的下降。因此,需要建立合理的频谱感知模型并利用有效的算法来跟踪频谱空闲概率的变化。

本文研究了多个频段间的频谱感知次序问题,每个频段的频谱空闲概率不同且其统计规律在时间尺度上是变化的。提出了基于在线学习[8]的频谱感知次序最优选择:利用在线学习的框架,边探索边利用,在利用历史信息进行决策的同时,考虑利用探索获取的新信息对未来决策的影响,待解决的问题可以看作经典多摇臂赌博机问题,并引入满意折现汤普森抽样(Satisficing Discounted Thompson Sampling, SDTS)算法[9-10]处理优化问题。

1　问题描述

本文研究了频谱感知次序问题,也可以说是多频段间的信道选择问题。因为在同一时间、同一地区接入每个频段的用户数不同,所以每个频段的频谱空闲概率不同。此外,由于用户的移动性,空闲概率在一段时间后会发生变化。假设次级用户有多个可能接入的频段,但是由于硬件限制,次级用户无法感知很宽的频谱。同时,假设在给定的时隙内,次级用户只能感知同一频段内的信道。因此,目标是动态选择一个频段进行感知,从而得到更多无主用户活动的信道,以便次级用户传输数据。

1.1　频谱感知次序问题

考虑有 N 个频段且每个频段划分为 C 个互不重叠的信道。研究了 T 时隙内次级用户利用空闲信道得到的总数据吞吐量。频段 i 在时隙 t 的频谱空闲概率记作 μ_i,其中 $i\in\{1,2,\cdots,N\}$。在实际场景中,同一频段内各个信道的空闲概率可能各不相同。假设在时隙 t,频段 i 的第 c 个信道的回报记作 $H_{i,c}$,服从期望为 μ_i 的伯努利分布,其中 $c\in\{1,2,\cdots,C\}$。也就是说,如果信道 c 空闲,则 $H_{i,c}=1$,否则 $H_{i,c}=0$。系统模型如图 1 所示。μ_i 不能提前得知且会随着时间的推移发生变化,因此考虑利用在线学习框架完成频谱感知任务。

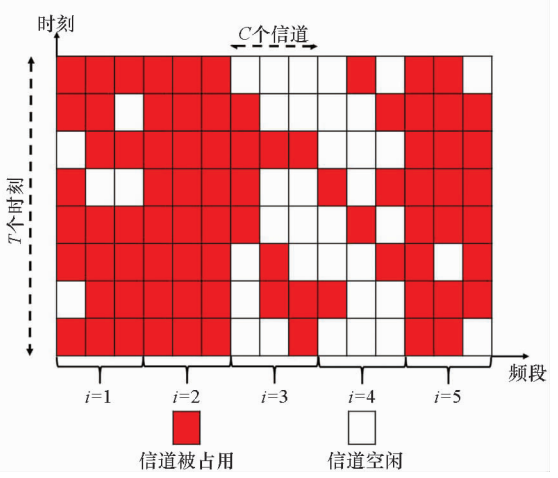

图 1　系统模型示例

Fig. 1　An example of the system model

在时隙 t,次级用户选择频段 i 进行频谱感知,其中 $t\in\{1,2,\cdots,T\}$,$i\in\{1,2,\cdots,N\}$。根据上文 $H_{i,c}$ 的定义,获得的空闲信道数服从二项分布。因此,在时隙 t,频段 i 的期望空闲信道数为 $C\mu_i$。次级用户需要在多个频段间动态选择,从而最大化 T 时隙内总期望空闲信道数目,因此,优化问题可表示为

$$\max_{i\in\{1,2,\cdots,N\}} E\Big[\sum_{t=1}^{T} C\mu_{i(t)}\Big] \tag{1}$$

其中,下标 $i(t)$ 表示在时隙 t,次级用户所选频段的索引。在该优化问题中,要求次级用户提前选择一系列频段使累积空闲信道最多。每个时隙得到的空闲信道数和未知的频谱空闲概率有关,这一概率可以通过在线学习的方式获得。根据上述描述,频谱感知次序问题可以归约为经典多摇臂赌博机问题。

1.2　经典多摇臂赌博机模型

考虑一个有 N 个摇臂(决策项)的赌博机(决策问题)。决策者共需进行 T 次决策。在每个阶段,当选择摇臂 i 后,产生的回报服从期望为 μ_i 的伯努利分布,且该回报立即被决策者观察到。每个阶段选择的摇臂产生的回报是独立同分布的。经典多摇臂赌博机[11]算法根据前 $t-1$ 次决策的回报决定第 t 次决策选择哪一个摇臂。这个问题的目标是最大化 T 时隙内的总期望回报。频谱感知次序问题中供选择的频段可以看作是经典多摇臂赌博机模型中的一个摇臂。为了在 T 时隙内获得更多的数据传输机会,次级用户需要动态选择最优的频段并进行频谱感知,从而最大化总期望空闲信道数目,这等价于经典多摇臂赌博机问题中最大化总期望回报。不同之处在于:经典

多摇臂赌博机问题中,每个摇臂产生的回报是 0 或 1;而优化问题中的回报,即空闲信道数,是随机分布在[0,C]的整数。考虑先将优化问题中的回报缩放至[0,1],再进行一次成功概率为缩放后回报的伯努利实验[12],二值变量 r_i 为伯努利实验的观测结果。伯努利实验的期望回报为 μ_i。经过缩放处理后,可以使用经典多摇臂赌博机的相关算法处理优化问题。

2 满意折现汤普森抽样算法

汤普森抽样算法[13]是处理经典多摇臂赌博机问题的有效算法。它根据过去时隙选择的摇臂及相应回报在当前时隙做出决策。文献[14]给出了汤普森抽样算法的后悔上界和性能保证的理论分析。现有的大多数关于多摇臂赌博机问题的文献的目标是快速收敛到最优决策以减小探索成本。但由于优化问题中每个摇臂产生的回报服从分布的参数是动态变化的,存在达到收敛之前最优决策已经发生变化的情况,也就是说,相对于次优决策,找到最优决策的代价过大,因此,在文献[9]提出的适用于非平稳摇臂的折现汤普森抽样算法的基础上,引入了满意汤普森抽样算法[10]。它的目标是确定一个足够满意或者足够接近最优的决策。

由于 μ_i 未知,满意折现汤普森抽样算法使用 B(1,1),即均匀分布作为其先验分布。在时隙 t,频段 i 先前已接入成功 S_i 次(即 $r_i=1$),接入失败 F_i 次(即 $r_i=0$)。频段 i 回报的后验分布更新为 B(S_i+1,F_i+1),并作为下一时隙决策的先验分布。满意折现汤普森抽样算法的时间复杂度为 $O(NT)$。在每个时隙,次级用户决定选择哪一个频段的具体步骤如算法 1 所示,其中超参数 γ 控制对频段的 S_i(或 F_i)的置信度,防止由于某个频段的 S_i 过大而出现总是选择这个频段的现象,即使该频段不是频谱空闲概率最大的。超参数 ξ 是一个容限参数,表示可以接受的满意决策与最优决策的差距。如果 $\xi=0$,满意折现汤普森抽样与汤普森抽样相同,否则,满意折现汤普森抽样倾向于选择过去时隙已经选择过的频段。特别是当最优决策需要很长时间才能学到而满意决策可以很快学到的时候,满意折现汤普森抽样算法可以快速达到接近最优的性能,而汤普森抽样算法需要继续探索来确定最优决策,会产生较大的性能损失。下面将通过推导展示使用满意折现汤普森抽样方法解决优化问题的合理性。

算法 1 满意折现汤普森抽样

Alg. 1 Satisficing discounted Thompson sampling

已知:参数 $\gamma\in(0,1]$,$\xi\geqslant 0$

1. 对每个频段 $i\in\{1,2,\cdots,N\}$,初始化 $\alpha_i=\beta_i=1$,$S_i=F_i=0$
2. **for** $t=1,2,\cdots,T$ **do**
3. **for** $i=1,2,\cdots,N$ **do**
4. 更新 $S_i=\gamma S_i$,$F_i=\gamma F_i$
5. 根据分布 B(S_i+1,F_i+1)采样 $\theta_i(t)$
6. **end for**
7. 令 $i(t)=\arg\max\theta_i(t)$
8. $\tau^*=\min\{\tau\in\{1,2,\cdots,t-1\}:\theta_{i(\tau)}(t)+\xi\geqslant\theta_{i(t)}(t)\}$
9. **if** τ^* 不为空 **then**
10. $i(t)=i(\tau^*)$
11. **end if**
12. 选择频段 $i(t)$ 并观察回报
13. 进行一次伯努利实验,观察到结果为 $r_{i(t)}$
14. **if** $r_{i(t)}=1$ **then**
15. $S_{i(t)}=S_{i(t)}+1$
16. **else**
17. $F_{i(t)}=F_{i(t)}+1$
18. **end if**
19. **end for**

假设观测向量 $\boldsymbol{R}_i$ 包含频段 i 被选后到当前时隙的所有观测结果 r_i,观测结果的产生方式 1.2 节已经给出。因此,$\boldsymbol{R}_i$ 的似然概率为

$$p_i(\boldsymbol{R}_i|\mu_i)=\mu_i^{S_i}(1-\mu_i)^{F_i} \tag{2}$$

满意折现汤普森抽样算法使用贝塔分布作为参数 μ_i 的先验分布,而贝塔分布是式(2)中似然函数的共轭分布,根据贝叶斯准则,μ_i 的后验分布为

$$p_i(\mu_i|\boldsymbol{R}_i)=\frac{p_i(\boldsymbol{R}_i|\mu_i)p_i(\mu_i)}{p_i(\boldsymbol{R}_i)} \tag{3}$$

其中

$$p_i(\mu_i)=\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\mu_i^{\alpha_i-1}(1-\mu_i)^{\beta_i-1} \tag{4}$$

参数 α_i 和 β_i 决定了贝塔分布的均值和方差。将式(2)和式(4)代入式(3),有

$$p_i(\mu_i|\boldsymbol{R}_i)=\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)p_i(\boldsymbol{R}_i)}\mu_i^{S_i+\alpha_i-1}(1-\mu_i)^{F_i+\beta_i-1} \tag{5}$$

令 $H=\Gamma(\alpha_i+\beta_i)/[\Gamma(\alpha_i)\Gamma(\beta_i)p_i(\boldsymbol{R}_i)]$,有

$$p_i(\mu_i|\boldsymbol{R}_i)=H\mu_i^{S_i+\alpha_i-1}(1-\mu_i)^{F_i+\beta_i-1} \tag{6}$$

由于 $\int x^{\alpha_i-1}(1-x)^{\beta_i-1}\mathrm{d}x=\Gamma(\alpha_i)\Gamma(\beta_i)/\Gamma(\alpha_i+\beta_i)$ 且 $\int p_i(\mu_i|\boldsymbol{R}_i)\mathrm{d}\mu_i=1$,可以得到

$$p_i(\mu_i \mid \boldsymbol{R}_i)=\frac{\Gamma(\alpha_i+\beta_i+S_i+F_i)}{\Gamma(\alpha_i+S_i)\Gamma(\beta_i+F_i)}\mu_i^{S_i+\alpha_i-1}(1-\mu_i)^{F_i+\beta_i-1} \tag{7}$$

这符合参数为 $S_i+\alpha_i$ 和 $F_i+\beta_i$ 的贝塔分布。所以 μ_i 的后验概率可以根据下式更新，即

$$p_i(\mu_i \mid \boldsymbol{R}_i)=\mathrm{B}(S_i+\alpha_i, F_i+\beta_i) \tag{8}$$

3　仿真结果及分析

假设有 5 个待选的频段，每个频段划分为 $C=20$ 个互不重叠的信道。频段 i 的频谱空闲概率 μ_i 服从标准均匀分布，即 μ_i 是从以(0,1)为界的均匀分布中采样得到的。频段 i 的每个信道根据成功概率为 μ_i 的伯努利分布产生回报。通过和其他具有代表性的算法，即折现 ε－贪心算法(Discounted ε-greedy)，折现汤普森抽样[8](Discounted Thompson Sampling，DTS)和折现信心上界算法[15](Discounted Upper Confidence Bound，Discounted-UCB)进行比较，可以看出满意折现汤普森抽样算法的性能有一定的提升。在折现 ε－贪心算法[16]中，每次决策次级用户以 $1-\varepsilon$ 的概率选择当前时隙平均回报最大的频段；以概率 ε 随机选择一个频段。仿真过程中，设置 $\varepsilon=0.1$。在折现信心上界算法中，引入了“信心上界指标”，该指标是回报的经验分布简单函数。次级用户先依次选择每个频段，在 $t>5$ 的任意时隙，选择信心上界指标最大的频段进行频谱感知。为了处理非平稳多摇臂赌博机问题，在标准信心上界算法的基础上，折现信心上界算法利用历史回报计算当前时隙的期望回报时引入了折现因子，使得距离当前时隙较近时隙的回报权重更高。折现 ε－贪心算法采用和折现信心上界算法相同的折现方法计算即时期望回报。

考虑了满意折现汤普森抽样算法中不同大小的容限参数 ξ 对平均每个时隙后悔值的影响。μ_i 在 200 个时隙内保持不变，超参数 γ 等于 1。考虑到每个频段的回报在 0 到 1 之间，ξ 分别取 0.01，0.05，0.10 和 0.15。对于每个频段 i，μ_i 分别是{0.23，0.10，0.28，0.32}。进行了 100 次蒙特卡洛模拟取平均得到仿真结果。平均每个时隙的后悔值定义为后验最优决策和所提算法决策的累积期望回报差值与所经历时隙的比值。如果次级用户可以提前得知 μ_i 的先验信息，那么总是选择频段 5，即 μ_i 最大的频段进行频谱感知是后验最优决策。图 2 展示了不同 ξ 对平均每个时隙后悔值的影响。从图 2 可以看出，对于不同的 ξ 值，平均每个时隙的后悔值都能达到一个较小的值，且随着 ξ 的增大，后悔值会先减小再增大，ξ 设为 0.05 时，后悔值最小。这是因为，当 ξ 较小时，随着 ξ 的增大，算法可以较快地从历史决策中找出和最优决策回报差距较小的满意决策，避免为了寻找回报更大的决策继续探索导致当前后悔的增加，减小了探索成本，后悔值减小；但是超过一定阈值后，ξ 继续增大，符合容限范围的满意决策会更快找到，因而过早停止探索，不断执行和最优决策差距较大的决策而错过回报更大的决策，导致后悔值增大。因此，在后续的仿真过程中，将超参数 ξ 设为 0.05。

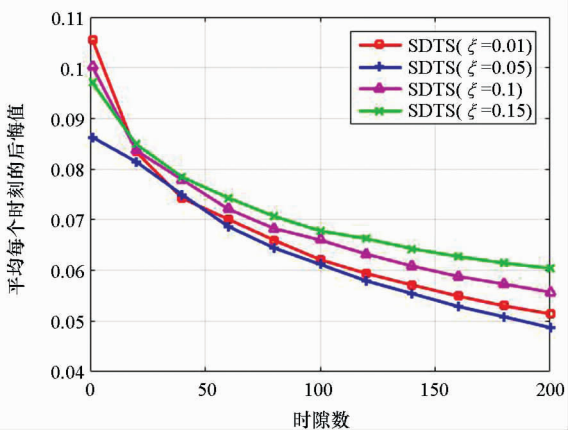

图 2　ξ 对平均每个时隙后悔值的影响

Fig. 2　Effect of ξ on the average per time slot regret

图 3 展示了当 μ_i 在 200 个时隙内保持不变时归一化吞吐量关于时隙的函数，其中归一化吞吐量定义为所用算法得到的总空闲信道数与后验最优决策得到的总空闲信道数的比值。各频段的 μ_i 分别是{0.20，0.04，0.37，0.35，0.06}。如果该信息能够提前得知，那么可以推断 200 个时隙内总是选择频段 3 是后验最优决策。显然，后验最优决策的归一化吞吐量始终为 1。从图 3 可

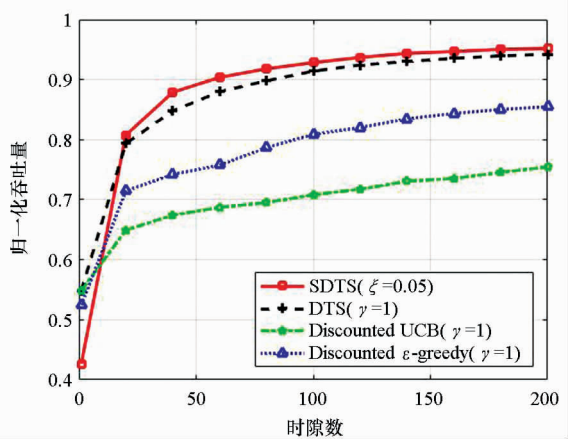

图 3　μ_i 不变时归一化吞吐量关于时隙的函数

Fig. 3　Normalized throughput as a function of the number of time slots with fixed μ_i

以看出，经过一定的时间，根据满意折现汤普森抽样算法进行决策得到的累积空闲信道数只比后验最优决策得到的少 5%，而这一方法不需要知道频谱空闲概率的先验信息，这在实际应用中有很大的优势。此外，和折现汤普森抽样算法、折现信心上界算法、折现 ε - 贪心算法相比，所提算法的吞吐量更大，这是因为它倾向于更快地选择频谱空闲概率最大的频段，从而给次级用户提供更多接入空闲信道的机会，进行数据传输。

研究了非稳态场景下累积空闲信道关于时隙的函数。假设 μ_i 每 200 个时隙变化一次，在 2000 个时隙内共变化 10 次。μ_i 的先验信息如表 1 所示，其中 k 代表 μ_i 第 k 次发生变化。一系列仿真结果显示，超参数 γ 等于 0.99 时得到的空闲信道更多，因此为了应对 0.99 动态变化的场景，超参数 γ 设为 0.99。图 4 给出了 μ_i 变化时累积空闲信道关于时隙的函数。从图 4 可以看出，根据后验最优决策选择频段，也就是总是选择 μ_i 最大的频段进行频谱感知，获得的累积空闲信道最多。但是，在 μ_i 动态变化且对次级用户不可知的情况下，满意折现汤普森抽样算法得到的空闲信道数只比后验最优决策得到的少 9%，且比其他经典算法至少多 4%，这说明每次 μ_i 发生变化时，满意折现汤普森抽样算法都能适应 μ_i 的变化，较快地找到满意决策。

表 1 μ_i 的先验信息

Tab. 1 Prior information of μ_i

k	μ_1	μ_2	μ_3	μ_4	μ_5
1	0.02	0.01	0.32	0.38	0.29
2	0.11	0.29	0.01	0.11	0.17
3	0.16	0.28	0.22	0.01	0.30
4	0.38	0.28	0.32	0.14	0.12
5	0.12	0.17	0.16	0.07	0.27
6	0.17	0.12	0.38	0.28	0.30
7	0.04	0.21	0.38	0.06	0.24
8	0.17	0.04	0.36	0.01	0.04
9	0.21	0.37	0.27	0.38	0.12
10	0.10	0.38	0.34	0.37	0.33

图 5 展示了非稳态场景下平均每个时隙的后悔值关于时隙的函数。这个后悔值反映了后验最优决策和所提算法决策之间的性能差距，可以看作是算法的跟踪误差。如图 5 所示，虽然 μ_i 每隔一定时隙会发生变化，但是经过一定的时隙，每个算法的平均后悔值都会达到一个比较小且相对稳定的值。而且，所提满意折现汤普森抽样算法的平均后悔值比其他算法更小、更稳定。这是因为每次频谱空闲概率发生变化时，满意折现汤普森抽样算法都能够比其他算法更快地找到概率最大的频段，跟踪性能更优。

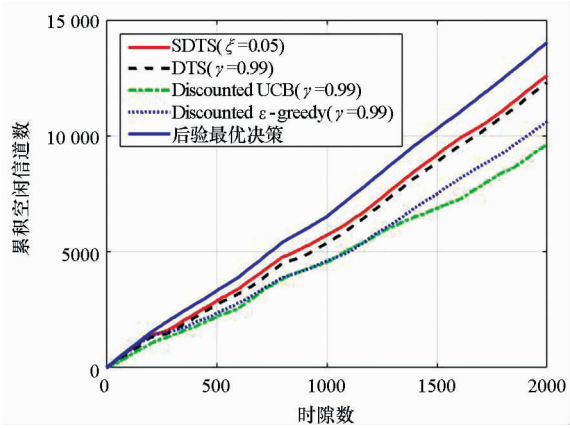

图 4 μ_i 变化时累积空闲信道关于时隙的函数

Fig. 4 Cumulative number of idle channels as a function of the number of time slots with time-varying μ_i

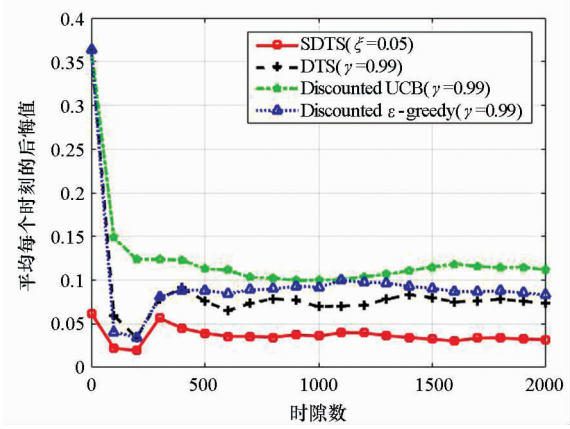

图 5 μ_i 变化时平均每个时隙的后悔值关于时隙的函数

Fig. 5 Average per time slot regret as a function of the number of time slots with time-varying μ_i

4 结论

本文研究了在频谱空闲概率的先验信息不可知且动态变化的情况下，认知无线电中次级用户在多频段间的频谱感知次序选择问题。这个问题被归纳成一个动态的在线学习模型，即多摇臂赌博机问题。在汤普森抽样算法的基础上，提出了一种满意折现汤普森抽样算法处理该问题。仿真结果显示，该算法得到的空闲信道数和后验最优决策得到的相近。与经典的信息上界算法和折现 ε - 贪心算法相比，本文所提算法获得的空闲信道更多。此外，所提算法还能够跟踪频谱空闲概率的动态变化。

参考文献(References)

[1] Dai J Y, Wang S W. Clustering-based spectrum sharing strategy for cognitive radio networks[J]. IEEE Journal on Selected Areas in Communications, 2017, 35(1): 228 – 237.

[2] Wang S W, Ge M Y, Wang C G. Efficient resource allocation for cognitive radio networks with cooperative relays[J]. IEEE Journal on Selected Areas in Communications, 2013, 31(11): 2432 – 2441.

[3] Akyildiz I F, Lee W Y, Vuran M C, et al. Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey[J]. Computer Networks, 2006, 50(13): 2127 – 2159.

[4] Ghasemi A, Sousa E S. Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs[J]. IEEE Communications Magazine, 2008, 46(4): 32 – 39.

[5] Jiang H, Lai L F, Fan R F, et al. Optimal selection of channel sensing order in cognitive radio[J]. IEEE Transactions on Wireless Communications, 2009, 8(1): 297 – 307.

[6] Li H S. Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: a two by two case[C]// Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2009: 1893 – 1898.

[7] Raj V, Dias I, Tholeti T, et al. Spectrum access in cognitive radio using a two-stage reinforcement learning approach[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(1): 20 – 34.

[8] 何斯迈, 金羽佳, 王华, 等. 在线学习方法综述: 汤普森抽样和其他方法[J]. 运筹学学报, 2017, 21(4): 84 – 102.
HE Simai, JIN Yujia, WANG Hua, et al. A survey on online learning methods: Thompson sampling and others[J]. Operations Research Transactions, 2017, 21(4): 84 – 102. (in Chinese)

[9] Zhou M, Wang T Y, Wang S. Spectrum sensing across multiple service providers: a discounted Thompson sampling method[J]. IEEE Communications Letters, 2019, 23(12): 2402 – 2406.

[10] Russo D, Tse D, van Roy B. Time-sensitive bandit learning and satisficing Thompson sampling[J]. arXiv Preprint arXiv: 1704.09028, 2017.

[11] 章晓芳, 周倩, 梁斌, 等. 一种自适应的多臂赌博机算法[J]. 计算机研究与发展, 2019, 56(3): 643 – 654.
ZHANG Xiaofang, ZHOU Qian, LIANG Bin, et al. An adaptive algorithm in multi-armed bandit problem[J]. Journal of Computer Research and Development, 2019, 56(3): 643 – 654. (in Chinese)

[12] Agrawal S, Goyal N. Analysis of Thompson sampling for the multi-armed bandit problem[C]//Proceedings of Conference on Learning Theory, 2012: 39.1 – 39.26.

[13] Chapelle O, Li L H. An empirical evaluation of Thompson sampling[C]// Proceedings of Advances in Neural Information Processing Systems, 2011: 2249 – 2257.

[14] Agrawal S, Goyal N. Further optimal regret bounds for Thompson sampling[C]//Proceedings of Artificial Intelligence and Statistics, 2013: 99 – 107.

[15] Garivier A, Moulines E. On upper-confidence bound policies for non-stationary switching bandit problems[C]// Proceedings of International Conference on Algorithmic Learning Theory, 2011: 174 – 188.

[16] 叶孟宇. 基于多臂赌博机的信道选择[J]. 软件, 2018, 39(4): 196 – 200.
YE Mengyu. Cognitive radio channel selection based on multi-armed bandit algorithm[J]. Computer Engineering & Software, 2018, 39(4): 196 – 200. (in Chinese)