基于 Q-Learning 的认知无线电系统感知管理算法

李 默 徐友云 蔡跃明
(解放军理工大学通信工程学院 南京 210007)

摘 要：认知无线电系统不仅是一个自适应系统，更应该是一个智能系统。该文将智能控制中的 Q-Learning 思想引入到认知无线电系统中，用于解决感知任务在认知用户之间的分配问题，给出了一种基于 Q-Learning 的感知管理算法。该算法在不知道信道状态信息以及不需要对主用户业务进行估计的假设下通过不断地与环境进行交互和学习来给认知用户分配感知任务。仿真表明，该算法能够提高感知效率，并且收敛速度较快，可作为未来认知无线电系统走向智能化的一种尝试。

关键词：认知无线电；Q-Learning；感知任务管理
中图分类号：TN92 **文献标识码**：A **文章编号**：1009-5896(2010)03-0623-06
DOI:10.3724/SP.J.1146.2009.00296

Q-Learning Based Sensing Task Management Algorithm for Cognitive Radio Systems

Li Mo Xu You-yun Cai Yue-ming
(*Institute of Communications Engineering*, *PLA University of Science and Technology*, *Nanjing* 210007, *China*)

Abstract: More than an adaptive system, the cognitive radio system is an intelligent system. The Q-Learning of the intelligent control theory is adopted in the paper, to solve the sensing task allocation problem among cognitive users. And a Q-Learning based sensing management algorithm is proposed. The algorithm allocates sensing tasks to users through times of interaction with the environment and self-learning. The scheme of the paper works without any channel state information and estimation of primary traffic. From the simulation result, the algorithm could improve the sensing efficiency compared to the static allocation algorithm and attain to the convergence in a short time, which could be an attempt to the future intelligent cognitive radio systems.

Key words: Cognitive Radio; Q-Learning; Sensing task management

1 引言

认知无线电(cognitive radio)作为一项解决无线通信系统频谱资源紧张问题的关键技术，近年来引起了人们的广泛关注。作为一种智能无线通信系统，认知无线电系统中的用户能够感知自身周围环境，学习无线环境背景知识，通过实时改变某些操作参数(发射功率、调制方式等)，达到与主用户网络合理共存，共同有效地利用频谱资源的目的[1]。

感知周围环境内的可用频谱资源是实现认知无线电系统所要解决的关键问题之一。有控制中心的认知无线电系统在基站的指导下周期性地进行频谱感知活动。由于认知用户的能量及工作频段受限，并且受到恶劣信道环境的影响，加之主用户信号出现的位置和时间都是随机的，因而不同的认知用户所具有的感知能力是不同的，同时基站也不需要每个用户都进行相同的感知活动。因此，为了使系统更好地运行，同时节省认知用户的资源，提高感知效率，基站需要智能的算法将感知任务分配到各个用户，设计可实现的、高效的频谱感知管理算法，对于从全局对系统进行优化设计是十分重要的。

文献[2，3]研究了认知无线电系统中的感知管理问题。文献[2]讨论了待感知信道在多个无线区域网(WRAN)小区之间的分配，通过提出的频谱跳变方法，使数据传输和信道感知可以同时进行。文献[3]利用了信道的频率选择性衰落特性，将对于某一用户来说遭受深衰落的信道分配给其进行感知，同时将这条信道分配给其他在其上具有良好信道条件的用户进行数据传输。上述这两种研究主要是讨论怎样来协调感知与传输之间的关系，使感知活动不会浪费过多的资源而对传输造成影响。然而由于认知用户感知能力有限，这样做不能保证感知的成功率。并且这两种管理方法需要工作在已知完美的信道状态信息的前提下，文献[2]还需要估计主用户业务的流量和带宽。这不仅需要大量认知链路的信道估计

2009-03-09 收到，2009-09-21 改回
国家 973 计划项目(2009CB3020402)和国家 863 计划项目(2007AA01Z267，2009AA01Z249)资助课题
通信作者：李默 limo8351@gmail.com

过程，而且通常来说认知用户到主用户链路的信道信息不容易得到。并且这样做相当于忽略和回避了认知无线电系统应具有的智能特性，认知无线电系统应该具有与环境进行交互，通过自身的学习得到最优分配策略的能力。

基于上述思想，本文从如何体现认知无线电系统的智能特性考虑它的感知管理过程，可以将这个过程建模为马尔可夫决策过程(Markov decision process)，它研究的就是智能体(agent)不断地与环境交互，根据当前环境所处的状态来决定执行的动作，从环境中获得回报(reward)，并从一个状态转移到另一个状态，积累经验学习最优策略的问题。Q-Learning(Q 学习)作为一种无模型的、无监督的在线强化学习算法(model-free, teacher-free, on-line reinforcement learning)[4]，是解决这类问题的有效途径之一。近年来已有研究将 Q-Learning 用于无线通信系统的资源管理领域[5−11]。文献[5]用 Q-Learning 来解决多小区之间的呼叫接入控制和切换；文献[6]利用 Q-Learning 对业务的速率进行管理，保证它们的 QoS；文献[7]在 HARQ 过程中引入 Q-Learning，用其估计选择最优的调制编码模式时的传输代价；文献[8]利用 Q-Learning 来提高认知无线电系统信号检测的性能；文献[9,10]利用 Q-Learning 派生出来的模糊 Q 学习来解决数据包接入以及均衡实时和非实时业务流量的问题；文献[11]用 Q-Learning 来实现异构 RAT 间自主的联合接纳控制和带宽分配问题。

本文在上述这些研究的启发下，将 Q-Learning 的思想应用于认知无线电系统感知管理问题中，运用 Q-Learning 算法来对感知任务进行合理的分配，使系统的感知效率得到提高。在每个感知周期内，基站通过具有 Q-Learning 功能的感知管理模块为每个用户分配待感知的信道，并根据感知结果确定回报，学习感知活动中的经验，以达到提高感知效率的目的。本文的算法工作在不知道信道状态信息以及主用户业务模型的情况下，仿真结果表明，基于 Q-Learning 的感知管理算法能够提高感知效率，减少漏检的情况，并且收敛较快。

本文的后续安排如下：在第 2 节中，首先给出了系统的模型并做出相应的假设；第 3 节将详细讨论本文给出的基于 Q-Learning 的感知管理算法；第 4 节给出了计算机仿真结果，分析了算法的性能；第 5 节对本文进行了总结。

2 系统模型及假设

本文考虑如图 1 所示的有控制中心的认知无线电系统，小区中有 K 个认知用户。为了避免对主用户信号造成干扰，基站需要周期性地组织认知用户进行感知活动，设该周期为 T。假设将主用户系统的资源分为 N 个信道，这样的信道可以是频点、子载波、扩频码等，基站将由感知管理算法得到的感知任务即待感知的信道，通过无差错的控制信道通知到每一个认知用户。认知用户在接到自己的感知任务后，采用某种感知算法对信道进行信号检测。

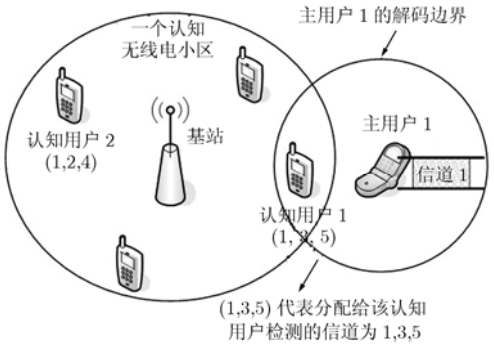

图 1 认知无线电系统示意图

由于本文研究的是对于感知任务即待感知信道的分配，不关心感知算法的性能，因此假设只要某个认知用户处在主用户的解码边界内，则该认知用户以概率 1 感知出该主用户信号存在与否，并且不存在误检的可能，如果检测出这条信道有主用户信号出现，那么叫做一次成功检测；否则将感知不到主用户信号，称作一次漏检。例如图 1 中，主用户 1 正在使用信道 1 进行通话，认知用户 1 正处于主用户 1 的解码边界内，并且当前感知管理算法将信道 1，3，5 分配给认知用户 1 进行检测，那么该用户就可以成功的检测出信道 1 有主用户信号出现；对于认知用户 2，尽管它的感知任务中也包括信道 1，然而它不在主用户的解码边界内，因此它不能检测到信道 1 上的主用户信号，称为一次漏检。在后面的仿真中，令 M 为认知用户在每个感知周期最多能够检测信道的个数，提高 M 值意味着认知用户消耗更多的能量和时间用于信道感知。

本文采用成功检测率及漏检率作为评价系统感知效率的性能指标，成功检测率定义为认知用户成功检测的信道个数与它的感知任务所含信道个数之比，漏检率定义为认知用户漏检的信道个数与它的感知任务所含信道个数之比。

对于主用户的业务模型，假设其业务服从 Poisson(泊松)过程，平均到达率为 λ，业务持续的时间满足均值为 $1/\mu$ 的指数分布。

3 基于 Q-Learning 的感知管理算法

认知用户所处的位置，以及受到能量的限制，

加上主用户的随机移动性，这些因素都使得各个认知用户的检测能力不尽相同，每个用户只能检测到有限个信道的信号，并且基站不能预先知道认知用户当前处在哪些主用户的解码边界以内，也就是说基站没有认知用户检测能力的先验知识。如果简单地为认知用户分配感知任务，如固定为某个用户分配一定的待检测信道，或者随机为每个用户分配待检测的信道，而不考虑与环境的交互，就很容易出现有检测能力的认知用户没有去检测相应的信道，或者出现相反的情况，那么会大大降低检测效率；另一方面，如果使每个用户都获得所有的检测任务，这样尽管可以保证检测性能，避免对主用户的干扰，但其代价是用户需要消耗大量的时间和能量来从事过重的检测任务，不利于数据传输的正常进行。Q-Learning 正是解决这种困境的有力方法之一，它通过不断地进行学习，积累经验，利用环境的回报发现最优的行为序列以获得最优策略。

3.1 Q-Learning 理论

假设一个智能体，它面临的环境是一个有限状态时间离散的动态系统，用 $S=\{s_1,s_2,\cdots,s_n\}$ 表示该系统的状态空间，智能体可采取的动作集为 $A=\{a_1,a_2,\cdots,a_m\}$，$r(s,a)$ 为在当前状态 $s\in S$ 下，智能体采取动作 $a\in A$ 获得的即时回报(reward)。学习算法的任务是学习一个策略，它基于当前状态 s 选择动作 a。那么如何精确地指定哪种策略是此智能体要学习的策略，一个简单的方法是要求此策略对智能体产生最大的累积回报，将累积回报作为评价策略优劣的评估函数。我们知道，当前的回报值及以前的回报值都可以得到，而后续状态的回报则很难得到，因此累积回报就难以计算。而 Q-Learning 用 Q 函数来代替累积回报作为评估函数，正好可以解决这个问题，Q 函数的基本方程为

$$Q(s,a)=r(s,a)+\gamma\max_b Q(s_{\text{next}},b) \tag{1}$$

式中 $s_{\text{next}}\in S$ 为在当前状态 s 和当前动作 a 下系统转入的下一状态；γ 为折扣系数，$0\le\gamma\le 1$，是将未来的回报折算成当前值的因子，γ 值越大未来回报对当前的影响就越大；b 为下一状态 s_{next} 下可采取的动作；$Q(s,a)$ 为智能体在当前状态 s 和当前动作 a 下得到的总计期望回报的估计，也称状态-动作对值。由式(1)可以看出当前的 Q 值是由当前状态和动作下的立即回报加上被 γ 折算的后续状态的 Q 值组成的。因此，Q-Learning 的思想不是去直接学习使累积回报最大的策略，而是通过不断地迭代来优化学习状态-动作对 $Q(s,a)$，通过 Q 值对累积回报进行估计来寻找最优策略。

3.2 基于 Q-Learning 的感知管理算法描述

将 Q-Learning 应用于认知无线电系统的感知管理问题，称基站的感知管理模块为一个智能体模块，建立如图 2 所示的基于 Q-Learning 的基站感知管理智能体模块与环境的交互。感知管理智能体模块在当前状态 s 选择特定动作 a 后，环境反馈奖励值 r，同时观察到系统的下一状态 s_{next}，并学习 $Q(s,a)$ 值，进行下一迭代。那么在基于 Q-Learning 的感知管理算法中，我们需要确定 Q-Learning 算法要素，包括划分状态空间、动作空间、回报函数、搜索策略、初始 Q 函数和折扣系数 γ 等，并确定动作的动态选择以最优化既定系统的性能指标。下面就针对 Q-Learning 应用到感知管理问题所要确定的关键因素具体讨论。

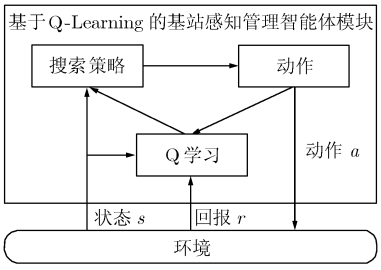

图 2 基于 Q-Learning 的基站感知管理智能体模块与环境的交互

(1)状态空间的选取和划分　状态空间 $S=\{s_1,s_2,\cdots,s_n\}$ 的划分是智能体合理选择动作的基础，正确选取其中的状态变量 s 应使其具备可知性和无后效性，可知性即是使状态的输入必须是智能体可以提取和处理的信息，无后效性则要求前一个状态只对后一个状态有影响，不会因后面的元素而变化不定。感知管理算法考虑的是如何将待感知的信道分配给不同的认知用户，即确定合理的(用户、待感知信道)对的问题，进行分配时既要分配到所有的用户，又不能发生重复分配的问题。我们选择待分配感知任务的认知用户作为状态空间，令状态变量 $s_i=u_i,i=1,2,\cdots,K$，u_i 为认知用户的序号，那么 $S=\{u_1,u_2,\cdots,u_K\}$。这样，在每一次分配时，算法通过一次次的状态转移就可以不重复地为每个用户分配感知任务，并且遍历所有的状态，最终的状态即为吸收状态。

(2)动作空间的选取和划分　感知管理问题的动作规定为给某个认知用户分配的待感知信道，即每个用户的感知任务。由于系统中待感知信道个数一般大于认知用户数，而对于 Q-Learning，每个状态只能选择一个动作，因此我们将 N 个待感知的信道平均分为 K 个信道组，分配的规则尽量使一个信

道组中包含的信道相距较远(对于信道为子载波的系统，这是指子载波的频点不临近)，以增加感知的成功率，如系统中共有 8 个信道，信道序号为 1–8，假设 K=4，4 个信道组可以分为：(1,5)，(2,6)，(3,7)，(4,8)。对于 N/K 不为整数的情况，将其中的 $\lfloor N/K \rfloor \times K$ 个信道分为 K 组($\lfloor\bullet\rfloor$ 为向下取整操作)，剩余信道则随机选择组。令动作变量 $a_j = c_j, j = 1, 2, \cdots, K$，$c_j$ 为第 j 个信道组的序号。这样在进行分配时，为每个状态挑选一个信道组。为避免多个状态选择同样的动作，在每一个分配周期里，被既往状态挑选过的动作将被标记，以免被后续状态重复选择。当系统中认知用户的个数 K 或者 N/K 很大时，就会产生大量的状态-动作对组合，消耗过多的存储空间，这个问题可以通过神经网络得到解决[5]。

(3)回报函数的设计　回报函数 $r(s,a)$ 的设计是基于系统的性能指标的，我们希望的是分配给认知用户的待检测信道是该认知用户能够成功检测的信道，那么这时的回报规定为正回报，其余的情况下回报为零。当前状态下的用户 s 被分配了信道组 a 后通过感知活动得到的回报为

$$r(s,a) = nr \tag{2}$$

其中 n 代表了认知用户 s 成功感知的信道个数，r 为回报权重，在仿真中设为 1。

(4)搜索策略　搜索策略用来平衡“探索”和“利用”。“探索”(explore)和“利用”(exploit)是 Q-Learning 搜索策略的两个重要方面。“探索”使系统尝试未做过的动作，使其有得到更多回报的机会；而在“利用”过程中，系统更倾向于采取先前得到更多回报的动作。本文采用最常见的 ε 贪婪算法(ε -greedy)，在某状态 s 下以小概率 ε 随机选择动作 a，以 $1-\varepsilon$ 选择具有最大 Q 值的动作，即 $a = \arg\max_a Q(s,a)$。

基于 Q-Learning 的感知管理算法流程如下：

(1)随机初始化 Q 值矩阵 $\boldsymbol{Q} = [Q(s,a)]_{K\times K}$，随机初始化状态 s；

(2)对于每一个感知周期，重复下面的过程：

(a)感知管理智能体模块查找 $\boldsymbol{Q}$ 阵，选择具有最大 Q 值的状态作为当前的激活状态 s，$s = \arg\max_{\substack{s\in S\\ a\in A}} Q(s,a)$；

(b)基于当前的状态 $s = u_i$，根据 ε 贪婪算法，选择对应当前状态的动作 $a = c_j$；

(c)对于处于当前激活状态的用户 u_i，对第 c_j 个信道组进行感知活动，将感知结果代入式(2)计算 $r(s,a)$；

(d)根据式(1)更新当前状态 s 下采取动作 a 的 Q 值 $Q(s,a)$，并将 Q 阵中行号为 i 或列号为 j 的 Q 值进行标记，其余的 Q 值不进行更新；

(e)选择 Q 阵中除标记外具有最大 Q 值的状态作为下一个状态，并更新状态 $s \leftarrow s_{\text{next}}$；

(f)回到(b)直到状态 s 为最终状态(或称吸收状态)。

3.3 算法分析

通过上面状态、动作等要素的选取可以看出，本文提出的感知管理算法实际上是通过回报值的设计来指导基站分配感知任务的，尽管选择认知用户作为状态空间没有直接体现出无线环境对状态的影响，但每个用户对应着一种它当前所处的环境(即处于哪些主用户的解码范围内)，环境对状态的影响将直接通过回报值来体现。并且，Q-Learning 中的搜索策略使得基站不会总为某个用户选择一个回报值最大的动作，而是以一定概率选择其他动作，那么当主用户游离出某用户的检测范围，到了另一个用户的检测范围时，这种机制就保证了认知用户能够在不同的感知任务上积累经验，这也就是支持了感知任务的动态管理。

根据主用户业务的模型，业务服从泊松过程，持续时间 t 满足均值为 $1/\mu$ 指数分布，那么在当前一个周期主用户占用某信道的情况下，在本周期继续占用的概率为

$$\begin{aligned} &P(t \ge (n+1)T \mid t \ge nT) \\ &\quad = \frac{P(t \ge (n+1)T, t \ge nT)}{P(t \ge nT)} \\ &\quad = \frac{\int_{(n+1)T}^{\infty} \mu e^{-\mu t}\mathrm{d}t}{\int_{nT}^{\infty} \mu e^{-\mu t}\mathrm{d}t} = e^{-\mu T} \\ &\qquad\qquad n = 1,2,3,\cdots \end{aligned} \tag{3}$$

也就是说当 μ 固定下来时，只要检测周期 T 足够小(主用户还没来得及更换位置和信道)，认知用户采取了动作 a，检测到 a 中某信道有主用户信号出现，那么它在下一个检测周期仍检测到该信道上有主用户信号出现的概率将非常大，在 a 中这样的信道越多，相应的回报 r 就越大，更新后的 Q 值就越大，因此随着时间的推移，采取动作 a 的趋势就越明显。基于 Q-Learning 的感知管理算法的收敛性将在下一节通过仿真得到验证。

4 仿真结果及分析

针对第 2 节给出的认知无线电系统，对提出的感知管理算法进行仿真验证。认知用户的感知周期为 0.5 s。主用户的个数为 10，其业务平均到达率设

为 $\lambda = 300$次/h，业务持续时间的均值为 $1/\mu = 180$ s，平均 10 s 更换一次信道，解码边界的大小为 100 m。主用户和认知用户均在 1000 m×1000 m 的范围内移动，移动速度平均为 5 m/s。

首先通过第 1 组仿真确定折扣系数 γ 与 ε 贪婪算法中 ε 的取值。表 1，表 2 分别给出了漏检率 e 在不同的 γ 或 ε 值下，随着感知周期 t_s 增加的变化。可以看出随着感知周期的增加，漏检率逐渐减小，并且 γ 越大，当前的 Q 值受后续回报的影响就越大，因此当仿真时间较小时，漏检率越大。表 2 中，ε 越大，对非最大 Q 值的动作探索的就越多，因此漏检的情况就越多。综合两表，在下面的仿真中选择 $\gamma = 0.3$，ε =0.2 。

表 1 漏检率 ε 在不同的 γ 值下随着感知周期 t_s 增加的变化情况

γ	t_s					
	50	100	500	1000	5000	100000
0.2	0.3369	0.3081	0.3044	0.2894	0.2997	0.2961
0.3	0.3448	0.3156	0.2909	0.3061	0.3119	0.2905
0.4	0.3673	0.3316	0.2978	0.2952	0.2992	0.2984
0.5	0.345	0.3466	0.2988	0.3091	0.3146	0.2988
0.6	0.3636	0.314	0.3098	0.3192	0.3014	0.2912
0.7	0.39	0.359	0.3314	0.3227	0.312	0.2936
0.8	0.3862	0.3495	0.3331	0.3275	0.3251	0.2981
$e = e(\gamma, t_s)$　$N = 128, K = 8, M = 5$, ε =0						

表 2 漏检率 e 在不同的 ε 值下随着感知周期 t_s 增加的变化情况

ε	t_s					
	50	100	500	1000	5000	10000
0.2	0.3173	0.3158	0.3145	0.296	0.2873	0.2983
0.3	0.3788	0.342	0.3233	0.3112	0.328	0.3295
0.4	0.368	0.3443	0.3492	0.3249	0.3221	0.3257
0.5	0.3577	0.3434	0.3491	0.3457	0.361	0.3602
$e = e(\varepsilon, t_s)$　$N = 128, K = 8, M = 5$, γ =0						

第 2 组仿真给出基于 Q-Learning 的感知管理算法的性能，为了比较的公平性，将本文提出的算法与不利用信道信息和主用户业务模型的静态分配算法(即对于认知用户在每个分配周期都指定其相同的感知任务，或者随机选择感知任务，仿真表明两者的性能接近，均归为静态分配算法中)的性能进行比较。在下面的图中，用“静态”来代表静态分配算法，用“Q-Learning”代表本文提出的算法。图 3 给出了 N=128，K=32，M=5 时两种算法的漏检率以及成功检测率随着仿真时间的变化情况，可以看出基于 Q-Learning 的感知管理算法的漏检性能以及成功检测性能均优于静态分配算法。随着认知用户的检测能力增加，两种算法对信道的检测效率都大大提升，如图 4 所示，但伴随着 M 值的增大，带来的后果是认知用户要消耗更多的能量和时间用于信道感知。

为了验证基于 Q-Learning 的感知管理算法的收敛速度，第 3 组仿真给出了当 N=16，K=4，M=4 时，随机选取的一个认知用户(称为用户 1)在不同的时间段下采取不同动作的比例，如图 5 所示。可以看出，在前 50 和 100 个感知周期内，用户 1 采取的动作还没有明显的趋势，随着感知周期的增加，用户 1 采取各个动作的比例在 500 个周期之后已经相对稳定，达到收敛状态。

5　结束语

本文将智能控制理论中的 Q-Learning 思想引入到认知无线电系统中，考虑如何运用 Q-Learning 算法来对感知任务进行合理的分配，使系统的感知效率得到提高。在每个感知周期内，具有 Q-Learning 模块的基站选择待分配感知任务的认知用户作为 Q-Learning 算法中的状态(state)，将为每个用户分配待感知的信道作为在每个状态下所采取的动作

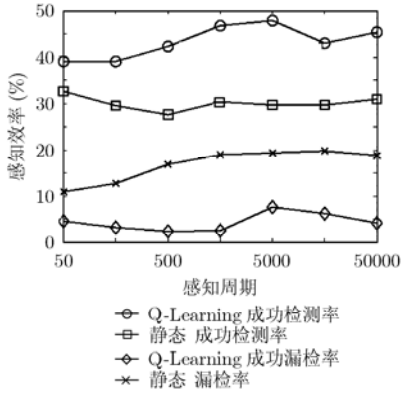

图 3 两种算法的感知效率随时间的变化

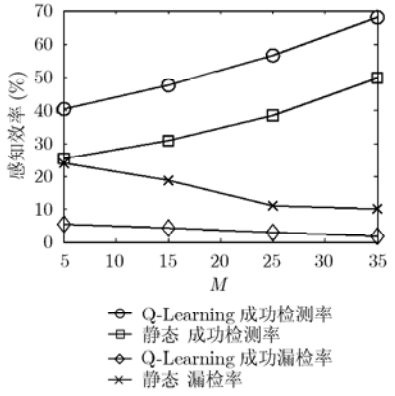

图 4 两种算法的感知效率随 M 的变化

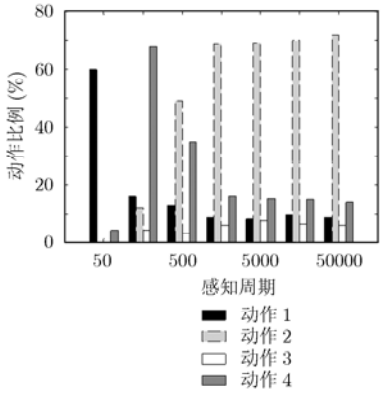

图 5 用户 1 采取的动作随时间的变化情况

(action)，根据收集到的感知结果来确定回报，学习感知活动中的经验，以达到提高感知效率的目的。仿真结果表明，基于 Q-Learning 的感知管理算法工作在不知道信道信息的情况下，与固定分配感知任务相比能够降低漏检率，提高感知效率，并且收敛较快。

参 考 文 献

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

[2] Hu Wen-dong, Willkomm D, and Vlantis G, *et al*.. Dynamic frequency hopping communities for effieient IEEE 802.22 operation[J]. *IEEE Communication Magazine*, 2007, 45(5): 80–87.

[3] Jeong Sang Soo, Jeon Wha Sook, and Jeong Dong Geun. Dynamic channel sensing management for OFDMA-based cognitive radiosystems[C]. Proceeding of VTC 2007, Dublin, 2007: 2646–2650.

[4] Watkin C and Dayan P. Q-Learning [J]. *Machine Learning*, 1992, 8(3): 279–292.

[5] Nie Jun-hong and Haykin S. A Q-Learning-based dynamic channel assignment technique for mobile communication systems[J]. *IEEE Transactions on Vehicular Technology*, 1999, 48(5): 1676–1687.

[6] Chen Yih-Shen, Chang Chung-Ju, and Ren Fang-Chin. Q-Learning-based multirate transmission control scheme for RRM in multimedia WCDMA systems[J]. *IEEE Transactions on Vehicular Technology*, 2004, 53(1): 38–48.

[7] Chang Chung-ju, Chang Chia-yuan, and Ren Fang-ching. Q-Learning-based hybrid ARQ for high speed downlink packet access in UMTS[C]. Proceeding of VTC2007, Dublin, 2007: 2610–2615.

[8] Reddy Y B. Detecting primary signals for efficient utilization of spectrum using Q-Learning[C]. Proceeding of the Fifth International Conference on Information Technology: New Generations, Las Vegas, 2008: 360–365.

[9] Chen Yih-shen, Chang Chung-ju, and Ren Fang-chin. Situation-aware data access manager using fuzzy Q-learning technique for multi-cell WCDMA systems[J]. *IEEE Transactions on Wireless Communications*, 2006, 5(9): 2539–2547.

[10] Nasri R, Altman Z, and Dubreil H. Optimal tradeoff between RT and NRT services in 3G-CDMA networks using dynamic fuzzy Q-Learning[C]. Proceeding of PIMRC'06, Helsinki, 2006: 1–5.

[11] 张永靖, 冯志勇, 张平. 基于 Q 学习的自主联合无线资源管理算法[J]. 电子与信息学报, 2008, 30(3): 676–680.
Zhang Y J, Feng Z Y, and Zhang P. A Q-learning based autonomic joint radio resource management algorithm[J]. *Journal of Electronics & Information Technology*, 2008, 30(3): 676–680.

李 默： 女，1983 年生，博士生，研究方向为认知无线电系统中的资源管理.

徐友云： 男，1966 年生，教授，博士生导师，研究方向为 B3G/4G 关键技术、无线资源管理、认知无线电等.

蔡跃明： 男，1961 年生，教授，博士生导师，研究方向为无线资源管理、协同通信等.