doi:10.3969/j.issn.1000-436x.2013.11.001

未知环境中基于图型博弈和 multi-Q 学习的动态信道选择算法

李方伟，唐永川，朱江
(重庆邮电大学 移动通信技术重庆市重点实验室，重庆 400065)

摘 要：研究了分布式无线网络中，没有任何信息交换、也没有环境变化先验知识情况下的动态信道接入算法。运用图型博弈模型对用户的实际拓扑进行建模分析，证明了此博弈模型存在纯策略纳什均衡并且此纳什均衡是全局最优解。同时，采用 multi-Q 学习求解模型的纯策略纳什均衡解。仿真实验验证了 multi-Q 学习能获得较高的系统容量以及在图型博弈模型中用户的效用主要由节点的度决定，而与用户数量无直接关系。

关键词：动态信道选择；图型博弈；multi-Q 学习；纯策略纳什均衡
中图分类号：TN929.5 **文献标识码**：A **文章编号**：1000-436X(2013)11-0001-07

Dynamic channel selection in unknown environment based on graphical game and multi-Q learning

LI Fang-wei, TANG Yong-chuan, ZHU Jiang
(Chongqing Key Lab of Mobile Communications Technology gqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: For the problem of dynamic channel selection in unknown distributed environment without a priori knowledge and information exchange, multi-Q learning was proposed. The dynamic channel selection problem was formulated the existence of pure strategy Nash equilibrium in graphical game was proved. At the same time, the pure strategy Nash equilibrium was proved to be global optimal solution. Simulation results show that multi-Q learning achieves high system capacity and utility of users in the graphical game are determined mainly by the degree of the node without direct relationship to the number of users.

Key words: dynamic channel selection; graphical game; multi-Q learning; pure strategy Nash equilibrium

1 引言

从当今的无线局域网、军用 ad hoc 网络，到未来的物联网、泛在无线网络，中心控制的网络组织方式已经不再适应网络的复杂性以及灵活性的要求。网络的演进表明，平等、独立、自治的网络用户以大规模分布式方式组织将成为主流。由于缺乏中心控制，如果提前给每个用户分配固定的频谱，无疑是对稀缺频谱资源的浪费。动态资源接入被认为是一种有效的方法去解决资源稀缺问题，因此得到了广泛的关注和研究[1~3]。

文献[4]考虑了平等用户之间的影响，采用博弈定价模型进行频谱接入。文献[5]在分布式网络中通过信息交换采用次梯度算法求解信道接入的混合策略纳什均衡。文献[6]提出了一种基于竞争拍卖的频谱接入方式。在完全信息下，文献[7]通过观察所有用户的效用值和行为策略来更新自己的策略。这些文献都有一个共同的特点就是需要知晓环境的先验信息或用户的行为信息，然而在实际环境下获取这些信息是比较困难的。在不知道用户和环境任何先验概率的情况下，假设任意 2 个用户都存在竞争（干扰）关系，采用 SLA(stochastic learning au-

收稿日期：2013-04-22；修回日期：2013-07-24
基金项目：国家自然科学基金资助项目(61102062, 61301122)；教育部科学技术研究重点基金资助项目(212145)；重庆市科委自然科学基金资助项目(cstc2011jjA1192)；重庆市教委科学技术研究基金资助项目(KJ110503)
Foundation Items: The National Natural Science Foundation of China(61102062, 61301122); The Key Project of Chinese Ministry of Education(212145); The Natural Science Foundation of Chongqing Science and Technology Commission (cstc2011jjA1192); The Science and Technology Research Project of Chongqing Education Commission(KJ110503)

tomata)算法[8]对用户进行动态的频谱分配。但是这种假设过于严格，且在用户增多时，进入不稳定期，系统容量会急剧下降。

本文在缺乏环境先验知识和任何信息交换的未知分布式环境下，采用 p-CSMA[9~11]接入机制来避免用户的冲突。首先针对传统博弈是一个 NP-C 或 NP-hard 问题[9]，采用图型博弈理论将复杂的分布式网络转化为简约的图型博弈模型，用模型中的图型拓扑来表示用户之间博弈的内在结构。然后，证明了在此条件下存在纯策略纳什均衡；同时还证明了此纳什均衡是一个全局最优解，为获取纯策略纳什均衡以及最优解奠定了理论基础。最后，针对此模型的特点，提出了改进的 multi-Q 学习求解纯策略纳什均衡解，提升了系统容量。

2 系统模型

2.1 图型博弈架构

假设 N 个独立自私的用户组成一个分布式无线网络，且任意用户不知道其他用户的存在。所有用户共享 M 个信道资源，信道 m 对用户 i 的质量为 R_m^i。信道质量是对信道衰落、信道噪声、信道可用性等因素的综合考虑。在本文中，它可以表示为比特率(bit/s)，数据分组的传输速率(数据分组/帧)等物理意义，为更好地说明其一般性没有指定其特殊的物理意思[10]。信道质量也是信道对用户的最大效用。信道 m 对除用户 i 之外的其他用户的最大效用为 $R_m^{-i}=(R_m^1,\mathrm{L},R_m^{i-1},R_m^{i+1}\mathrm{L},R_m^n)$，信道质量矩阵 $\boldsymbol{S}\overset{\text{def}}{=}[R_m^i]_{N\times M}$，$1\le M<N$，用户集合 $\mathrm{N}\overset{\text{def}}{=}\{1,\mathrm{L},N\}$，信道集合 $\mathrm{M}\overset{\text{def}}{=}\{1,\mathrm{L},M\}$。用户 i 只能选择一条信道 m 进行传输且信道质量 R_m^i 是时变的。但是，用户不知道信道时变的先验概率和此信道对其他用户的最大效用 R_m^{-i}。同一信道对不同用户的信道质量是不同的，不同信道对同一用户的质量也是不一样的，但是，各信道质量的均值对每个用户都是相同的，即 $(\mathrm{E}[R_m^i]=R)$。这种假设具有一定的实际背景，在 IEEE 802.d/e 标准[12]下，每个用户获得相同的频带宽度。

N 个用户在空间内随机任意分布，它们的干扰（竞争）关系如图 1 所示，虚线框代表每个用户的干扰范围。作者将图 1 复杂的实际拓扑结构抽象为如图 2 所示的图型拓扑结构——用模型中的图型拓扑表示实际环境中博弈的内在结构。$G\overset{\text{def}}{=}(V,E)$，$V$ 表示图中的节点，E 表示图中的边。其中，一个节点表示一个用户，边表示 2 个用户有直接的竞争关系。例如，用户 4 与用户 5 存在相互的干扰，如果它们在同一时刻接入同一信道将存在竞争，反之，用户 4 与用户 6 没有竞争关系，可以同时接入同一信道而不产生竞争（干扰）。d_i 表示节点 i 的度。

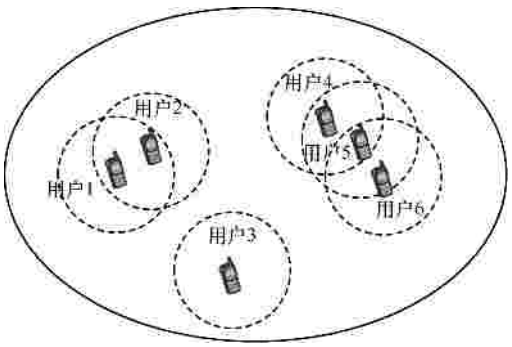

图 1 实际环境下用户的拓扑结构

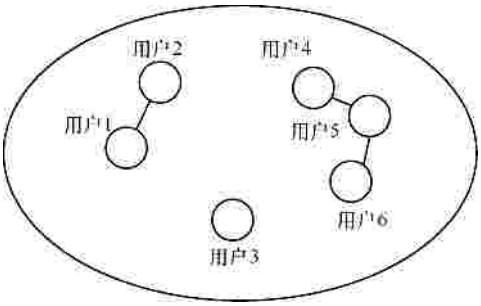

图 2 基于图论的图型博弈架构

由于用户之间没有信息交换，任意用户在动态信道接入时，不知道自己与哪个用户、多少个用户具有直接的竞争关系。用户仅仅知道有多少信道资源和当前信道 m 对自己的最大效用 R_m^i。

定义 1 图型博弈 G：$G\overset{\text{def}}{=}(G,A,u)$，其中，$G$ 表示用户集和用户之间的直接竞争关系；$A=\{1,\mathrm{L},M\}$ 表示用户有效的行为集（共享的信道资源）；u 表示用户的效用函数。在博弈中，每一位自私用户的唯一目的就是最大化自己的效用函数，即

$$G:\max_{a_i\in A_i} u_i(a_i,a_{J_{-i}}),\forall i\in \mathrm{N} \tag{1}$$

若采用传统的博弈模型，任意 2 个用户之间都存在博弈关系，任意用户的优化目标为

$$\max_{a_i\in A_i} u_i(a_i,a_{-i}),\forall i\in \mathrm{N} \tag{2}$$

从式(1)和式(2)可以看出，图型博弈去除了无直

接竞争关系的用户 $l(l \notin J_{-i})$，简化了竞争关系使得信道的重复利用率提高，增加用户获得信道的概率，从而提升系统容量和用户个体的效用。

2.2　信道模型

由于任意用户对其他用户没有任何的先验知识，作者考虑使用 p-CSMA 机制避免冲突的发生。假设用户具有如图 3 所示的传输时隙结构。T_s 表示信道感知和信道选择时隙；T_e 表示竞争传输时隙；T_l 表示用户策略的更新（学习）时隙。其中，竞争传输时隙又分为微控制时隙 d 和数据时隙。在微控制时隙，每一个用户以概率 p_a 传输请求接入信息和接入信道，当有 2 个及其以上用户同时接入信道时，用户将在下一个微控制时隙继续以概率 p_a 接入信道，直到只有一个用户接入信道；当某个用户成功接入信道时，剩余的微控制时隙被当作数据时隙传输数据流。同时，其他的用户将保持沉默直到下一个传输时隙。

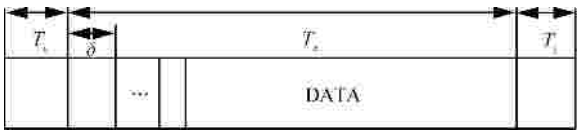

图 3　用户传输时隙结构

3　模型分析

3.1　效用函数

用 a_i 表示用户 i 的行为（信道选择），J_i 表示与用户 i 具有直接竞争关系的用户集合，$c_i \overset{\text{def}}{=} \{n \in J_i : a_n = a_i\}$ 表示在集合 J_i 中与用户 i 选择相同行为的用户集合，s_i 表示与用户 i 选择相同行为的用户数，即 $s_i = |c_i|$。在每次接入竞争中，信道质量是时变的且微控制时隙的个数 $N(s_i)$ 是一个随机数，所以用户 i 的回报也是一个随机效用，即

$$r_i(a_i, a_{-i}) = [(T_e - N(s_i)d)/T_e]R_m^i b_i(s_m) \tag{3}$$

其中，$N(s_i)$ 满足几何分布，$b_i(s_m)$ 满足伯努利分布，即

$$\Pr\{N(s_i) = k\} = p_s(1-p_s)^{k-1}, k \geq 1 \tag{4}$$

其中，$p_s = s_i p_a (1-p_a)^{s_i - 1}$ 表示信道在某个微控时隙被成功接入的概率。

$$\Pr\{b_i(s_i) = x\} = \begin{cases} \dfrac{1}{s_i}, & x = 1 \\ 1 - \dfrac{1}{s_i}, & x = 0 \end{cases} \tag{5}$$

其中，$b_i(s_i) = 1$ 表示用户 i 竞争胜利，获得传输时间；反之，则表示竞争失败。

定义用户 i 的效用为平均回报值，即

$$\begin{aligned} u_i(a_i, a_{-i}) &= \mathrm{E}[r_i(a_i, a_{-i})] \\ &= \mathrm{E}[((T_e - N(s_i)d)/T_e)R_m^i b_i(s_m)] \\ &= \frac{R}{s_i}\mathrm{E}[(T_e - N(s_i)d)/T_e] \\ &= \frac{R}{T_e s_i}(T_e - \frac{1}{p_s}d) \end{aligned} \tag{6}$$

由式(6)可知，当通信的客观条件给定，用户 i 的效用只与 J_i 中选择相同信道的用户数 s_i 有关，而与行为 a_i 无关，并且是关于 s_i 的严格递减函数。

同时，定义系统的效用函数为各个用户的平均回报值之和，即

$$U(\boldsymbol{a}) \overset{\text{def}}{=} \sum_i u_i(a_i, a_{-i}) \tag{7}$$

定义 2　最优策略：策略 $\boldsymbol{a}^{\text{opt}} = (a_1^{\text{opt}}, \mathrm{L}, a_N^{\text{opt}})$ 能使系统容量达到最大值，即

$$\boldsymbol{a}^{\text{opt}} = \arg\max_{\boldsymbol{a} \in \boldsymbol{A}} U(\boldsymbol{a}) \tag{8}$$

本文对系统效用函数进行长期优化，其长期的优化目标函数[13]为

$$p^* \overset{\text{def}}{=} \arg\max_p \left\{ \mathrm{E}_p \left[\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} U(\boldsymbol{a}) \right] \right\} \tag{9}$$

其中，p^* 为长期优化的最优策略，对于 $\forall S$，$p^*(S) @ \boldsymbol{a}^{\text{opt}}$，$T$ 为总传输时间，$\mathrm{E}_p[\cdot]$ 是采纳策略 p 时的期望。从式(9)可以看出，长期优化的目标是寻求整个传输过程中平均系统效用的最大化。

3.2　纳什均衡分析

定义 3　纳什均衡(Nash equilibrium)：如果给定用户的信道选择 $a^* = (a_1^*, \mathrm{L}, a_2^*)$ 是图型博弈的纳什均衡，那么没有任何一位用户可以通过单方面的改变行为来提升自己的效用，即

$$\begin{aligned} &\frac{R}{T_e s_{a_i}^*}(T_e - \frac{1}{p_s^*}d) \\ &\geq \frac{R}{T_e s_{a_i}}(T_e - \frac{1}{p_s}d), \forall i \in \mathrm{N}, a_i \neq a_i^*, a_i \in A \end{aligned} \tag{10}$$

定理 1　本文提出的图型博弈 G 存在纯策略纳什均衡。

证明　由于本文中用户效用函数是一个非线性关系不易找出势函数，所以先假设本文用户的效

用函数：$\bar{u}(a_i,a_{-i})=R-a\,s_i\geq 0$，其中，$a$ 为大于 0 的任意常数。

势函数：$f(a_i,a_{-i})=\sum_i R-\frac{1}{2}\sum_{j\in c_i}a\,s_i$

任意用户 i 单方改变其行为，效用函数的变化量为

$$\bar{u}(\bar{a}_i,a_{-i})-\bar{u}(a_i,a_{-i})=(s_i-\bar{s}_i)a \tag{11}$$

势函数的变化量为

$$\begin{aligned}f(\bar{a}_i,a_{-i})-f(a_i,a_{-i})=&-\frac{1}{2}\sum_{j\in\bar{c}_i}\bar{s}_i a+\frac{1}{2}\sum_{j\in c_i}s_i a-\\&\frac{1}{2}\sum_{j\in\bar{c}_i}\bar{s}_i a+\frac{1}{2}\sum_{j\in c_i}s_i a\\=&(s_i-\bar{s}_i)a\end{aligned} \tag{12}$$

由式(11)和式(12)可得

$$f(\bar{a}_i,a_{-i})-f(a_i,a_{-i})=\bar{u}(\bar{a}_i,a_{-i})-\bar{u}(a_i,a_{-i}) \tag{13}$$

由式(13)可知，此博弈是个潜在博弈[14]。再由潜在博弈性质可知，此博弈必定存在一个纯策略纳什均衡。

假设 $\boldsymbol{a}^*=(a_1^*,\cdots,a_n^*)$ 是一个纯策略纳什均衡，则存在 $s^*=(s_1^*,\cdots,s_2^*)$，满足 $\bar{u}(a_i^*,a_{-i})\geq\bar{u}(a_i,a_{-i})$，即

$$s_i^*\leq s_i \tag{14}$$

由于本文提出的效用函数 $u(a_i,a_{-i})$ 是关于 s_i 的严格递减函数，由式(6)和式(14)可得

$$u(a_i^*,a_{-i}^*)\geq u(a_i,a_{-i}^*),\forall i\in\mathbb{N},a_i\neq a_i^*,a_i\in A \tag{15}$$

由式(10)和式(15)可知，$\boldsymbol{a}^*=(a_1^*,\cdots,a_n^*)$ 也是本文图型博弈 G 的纯策略纳什均衡。

定理 2　在任意两用户都存在竞争的情况下，本文的纯策略纳什均衡是个最优策略。

证明　对纳什均衡定义的分析可知，对任意用户 i 达到纳什均衡有 $s_i^*\leq s_i$。纳什均衡将用户分配到用户数最少的信道，即将所有存在相互竞争的用户均匀地分配到不同的信道上。设 n_j 表示信道 j 上的用户数，$n_j^*=N/M$ 是一个均衡分配。均衡分配的系统容量为

$$\begin{aligned}U(\boldsymbol{a}^*)&=M\frac{N}{M}\frac{R}{T_e\dfrac{N}{M}}\left(T_e-\frac{1}{\dfrac{N}{M}p_a(1-p_a)^{\frac{N}{M}-1}}d\right)\\&=MR-\frac{dR}{T_e}\frac{1}{p_a}\frac{M}{\dfrac{N}{M}(1-p_a)^{\frac{N}{M}-1}}\end{aligned} \tag{16}$$

任意非均衡分配的效用为

$$U(\boldsymbol{a})=MR-\left(\sum_{j\in\mathbb{M}}\frac{dR}{T_e}\frac{1}{p_a}\frac{1}{n_j(1-p_a)^{n_j-1}}\right) \tag{17}$$

由算术平均数不小于几何平均数可得

$$\begin{aligned}\sum_{j\in\mathbb{M}}\frac{1}{n_j(1-p_a)^{n_j-1}}&\geq M\frac{1}{(\prod_{j=1}^{M}n_j)^{\frac{1}{M}}(1-p_a)^{\frac{N}{M}-1}}\\&\geq M\frac{1}{\dfrac{N}{M}(1-p_a)^{\frac{N}{M}-1}}\end{aligned} \tag{18}$$

当且仅当 $n_j=\cdots=n_k=N/M$ 时等号成立。由式(16)~式(18)可得

$$U(\boldsymbol{a}^*)\geq U(\boldsymbol{a}),\forall\boldsymbol{a}\in\boldsymbol{A} \tag{19}$$

其中，$\boldsymbol{A}=A_1\times\cdots\times A_n$ 为联合策略行为空间。由式(8)和式(19)可得定理 2 成立。

4　multi-Q 学习

4.1　multi-Q 学习更新策略

在本文的图型博弈模型中，环境具有未知的特点，因此用户对周围环境的感知中，不知道自己的竞争对手是谁，有多少个竞争对手，它们之间没有任何的信息交换。因此可以用强化学习算法[15,16]实现策略的学习。但是，随着用户数的增多，算法的规模将呈指数增长。由于博弈中的用户都是平等、独立、自治且自私的用户，具有平行性的特点[17]，所以本文最终采用 multi-Q 学习算法。根据模型中用户的特点及大量的实验，本文对基本的 multi-Q 学习做了改进，使其在迭代速度和系统容量上做了权衡，具体的更新步骤如下。

1) 初始化用户 $Q_{t_0}(i,m)=0$。

2) 每个用户以概率 $p_t(i,m)$ 随机选择信道，其中，$p_t(i,m)=\dfrac{Q_t(i,m)}{\sum_m k^{Q_t(i,m)}}$。

3) 用户 i 在选择的信道内以概率 p_a 请求接入信道，并根据式(3)计算当前立即回报值，更新自己的 Q 值

$$\begin{aligned}&Q_{t+1}(i,m)=(1-v_{im})Q_t(i,m)+\\&v_{im}(r_t(a_i,a_{-i})+b\%(t)Q_t(i,m)),a_i=m\\&Q_{t+1}(i,m)=Q_t(i,m)+\\&b(1-\%(t))Q_t(i,m)/(1+t^2)\quad,a_i\neq m\end{aligned} \tag{20}$$

其中，v_{im} 为用户 i 对信道 m 的学习因子，$0<b<1$ 是

步长因子，$\%(t)$ 是归一化立即回报值，其值为

$$\%(t) = r_t(a_i, a_{-i}) / \max_m R_m^i \tag{21}$$

为满足收敛性：$\sum_{i=1}^{\infty} v_{im} = \infty, \sum_{i=1}^{\infty} v_{im}^2 < \infty$。

4) 若对于 $\forall i \in \mathrm{N}$，存在 $p_t(i,m) \geq 0.99$，退出迭代，否则进入步骤 2)。

值得说明的是，较大的 k 值会将较高的概率赋予超出平均 Q 的选择，致使用户利用它所学习到的知识来选择它认为效用最优的策略。相反，较小的 k 值会使其他选择有较高的概率，导致用户探索那些当前 Q 值还不高的动作。一般使 k 随着迭代次数而变化。以使用户在学习的早期可用探索型策略，然后逐渐转换到利用型策略。同时，b 的值能调节 Q 值增加的速度。因此，b 和 k 共同决定了 Q 的更新速度和精确度。

4.2　算法比较

为了验证本文算法收敛的准确性以及实现性能的对比，分别与穷举算法、SLA 算法[8]、随机选择算法进行了比较。

定义 4　穷举法：在网络中存在一个中心节点，此节点收集所有信道对每个用户的效用值；再穷举出所有可能的接入策略；最后比较所有的策略，将最大化系统容量策略发布给每一个用户。

定理 3　穷举法所得到的策略是一个最优策略。

证明　令 $\boldsymbol{a}(\boldsymbol{a} \in \boldsymbol{A})$ 是穷举法的输出策略，则由穷举法的定义可知 $U(\boldsymbol{a}) \geq U(\overline{\boldsymbol{a}}), \forall \overline{\boldsymbol{a}}, \overline{\boldsymbol{a}} \neq \boldsymbol{a}, \overline{\boldsymbol{a}} \in \boldsymbol{A}$，则有：$\boldsymbol{a} = \arg\max_{\boldsymbol{a} \in \boldsymbol{A}} U(\boldsymbol{a})$。再由定义 2 可知，$\boldsymbol{a} = \boldsymbol{a}^{\mathrm{opt}}$。证毕。

虽然穷举法可以得到最优的策略，从而带来最好的系统容量，但是穷举法是不可取的。首先，需要找出每一种可能的信道接入，再找出最大值，其运算复杂度随用户数呈指数增长；其次，需要每个用户向中心节点上传效用矩阵，再将接入策略广播给每个用户，在大规模无线网络中其信息交换量较大；所以，无论是实际操作还是理论仿真都比较难以实现。然而，穷举法的仿真结果可以作为最优策略的性能上界。

随机选择算法就是以相等的概率选择任意信道进行接入。

下面对各算法的计算量和存储代价两方面进行分析。

表 1 比较了各算法的计算复杂度。其中，N 表示用户数，M 表示信道数，s_i 表示选择同一信道的用户数。在通信系统中，对于单个用户来说，其感知能力有限，信道数 M 较小[1]，用户数 N 往往较大。虽然本文的 multi-Q 学习算法计算复杂度接近 SLA 算法的两倍，但由于 M 较小，所以计算量均不大。随机选择算法是固定等概率信道选择，所以几乎没有计算开销。

表 1　各算法计算复杂度比较

算法	指数运算	乘除运算	加减运算	比较运算
穷举法算法	0	$NN^M(s_i+4)$	$3NN^M$	N^M
multi-Q 学习	M	$9M$	$7M$	M
SLA 算法	0	$5M$	$3M$	M
随机选择	0	1	0	0

表 2 表述了各算法主要存储变量以及对应变量的维数。同样，N 表示用户数，M 表示信道数，由于信道数 M 较小，multi-Q 学习算法的存储开销比 SLA 算法的存储开销略大。

表 2　各算法主要存储开销比较

算法	Q 值	学习因子 v	学习步长 b	信道质量矩阵 S	信道选择概率 P
穷举法	—	—	—	$N \times M$	—
multi-Q 学习	$1 \times M$	$1 \times M$	1	$1 \times M$	$1 \times M$
SLA 算法	—	—	1	$1 \times M$	$1 \times M$
随机选择	—	—	—	$1 \times M$	1

注：—表示不需要对应变量。

5　仿真分析

为不失一般性，结合参考文献[10]的例子，假设有 3 条信道且各个信道的平均质量(效用)为 R=1，且 R_m^i 在领域 $U(R, d_{i,m})$ 内均匀分布，其中，$d_{i,m} \in [0.1, 0.3]$ 及 R_m^i 在平均效用附近最大 ±30% 浮动。这里的信道质量(效用)指信道的传输速率，但由上文的叙述可知其还可以指数据分组的传输速率。传输时隙 T_e=90 ms，微控制时隙 d = 5 ms，信道接入请求概率 $p_a = 0.35$，步长 $b = 0.15$，$k = 1.1^t$，并通过 10^4 次实验后取平均得到仿真结果。

图 4 展示了在图型博弈模型中，从 15 bit 平均度为 8 并采用 multi-Q 学习的用户中任意选择一位用户，其信道选择的概率演进。用户信道选择的初始概率为{1/3,1/3,1/3}，大约经过 300 次迭代学习逐渐收敛到{0,1,0}。也就是说算法最终收敛，用户选择信道 2 进行数据传输。

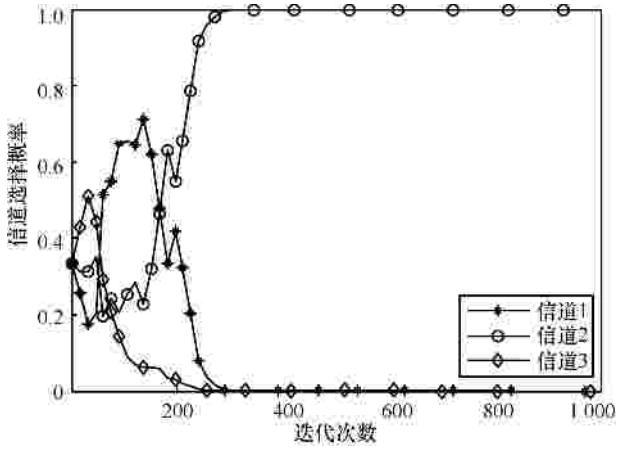

图 4 用户信道选择的概率演进(N=15, d=8)

图 5 展示了传输的过程中，系统容量的变化曲线。一个传输时间表示上文中的一个传输时隙结构。由于穷举法的计算量较大，这里只给出了最终的迭代结果作为系统容量的上界值。从图 5 中可以看出随着传输的进行，multi-Q 学习和 SLA 算法的系统容量逐渐增加。经过 300 次左右的迭代，multi-Q 学习获得了近似最优解。在差不多的迭代次数下 multi-Q 学习能获得比 SLA 算法较高的系统容量。由于随机选择算法是固定等概率信道选择，所以系统容量几乎呈直线。在传输的过程中，multi-Q 学习的系统输出始终高于 SLA 算法和随机选择算法，低于穷举法。因此可以得出 multi-Q 学习是一种长期优化(long term optimal)，在传输的各个阶段均是长期优化方程式(9)的近似最优解。虽然前 300 次策略没有收敛到最优，性能较差，但相对于长期优化而言，这种短时期的较差性能是可以忽略的[18,19]。而且这也符合强化学习基于多次“试错”最终收敛到最优的思想。

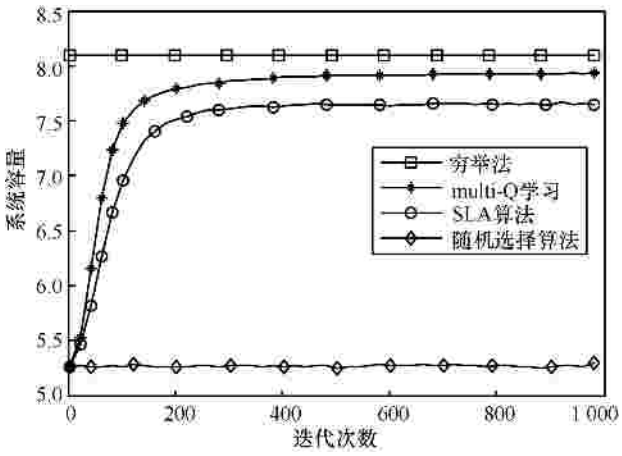

图 5 系统容量与传输时间的关系(N=15, d=8)

从图 6 可以看出系统容量与用户数的关系。采用传统的博弈，系统容量将随着用户数量的增多逐渐降低。这是由于传统的博弈没有考虑用户竞争的实际情况（假设任意 2 个用户均存在竞争），每增加一个人竞争加剧。图型博弈用户之间的竞争只考虑有竞争关系的用户。在用户的度恒定($d=5$)时，系统容量随用户的数量几乎呈线性增长。本文提出的 multi-Q 学习能逼近系统容量的最优值。

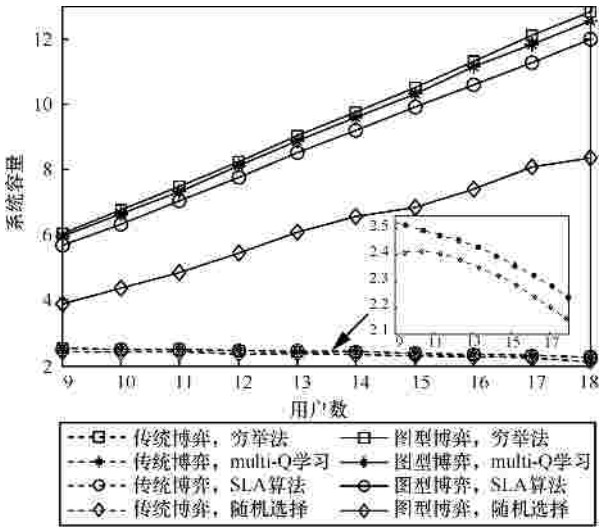

图 6 图型博弈(d=5)与传统博弈系统容量比较

图 7 展示了用户个体的平均效用与用户数的关系。运用图型博弈理论对用户进行建模分析，用户个体的平均效用也能获得近似最优解并随着用户数量的增加基本保持不变，multi-Q 学习能获得较好的个人效用。而传统博弈中个人平均效用将随用户数的增加而明显降低。

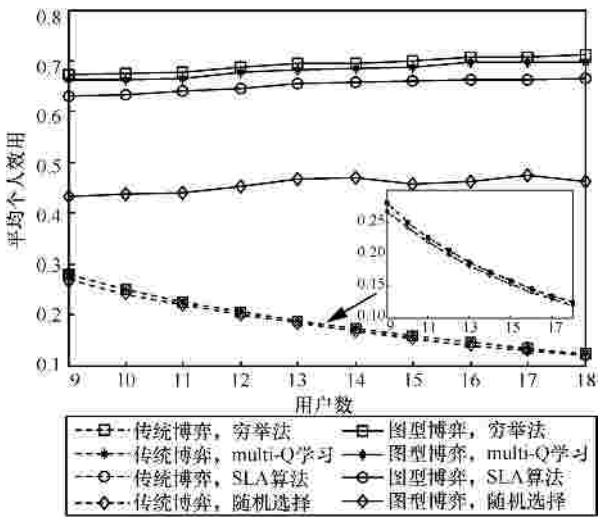

图 7 图型博弈(d=5)与传统博弈个人平均效用比较

图 8 表述了系统容量随用户度的变化趋势。在 15 个用户条件下，系统容量随着用户度的增加而降低。从图中可以看出，本文采用的 multi-Q 学习算法的系统容量接近穷举法系统容量，与另外 2 种算法相比，提升了系统容量。由于用户数是固定的，个人平均效用与系统容量具有相同的趋势。

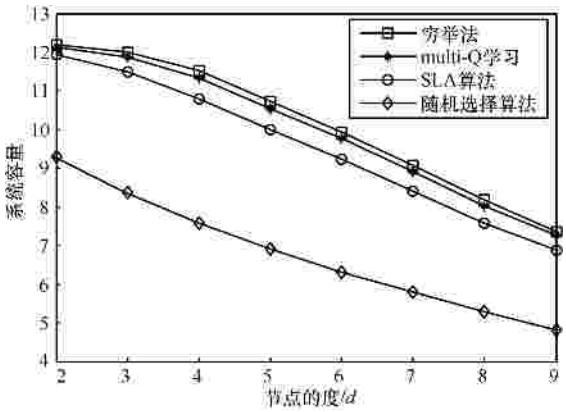

图 8　系统容量比较(N=15)

6　结束语

本文运用图型博弈理论将传统的博弈模型转化为简约的图型博弈，用图型拓扑表示实际用户博弈的内在联系。在缺乏任何信息交换和环境的先验知识的未知环境下，采用 multi-Q 学习求解模型的纯策略纳什均衡。虽然计算量和存储开销都略有增加，但是在迭代次数相近的情况下，能获得长期优化策略的近似最优解，并有较好的系统容量输出。同时，仿真实验验证了在用户度一定的情况下系统的效用将随着用户数的增加而线性增加；个人平均效用主要随用户度的增加而降低，而与用户数量无直接的关系。

参考文献：

[1] JAYAKRISHNAN U, VENUGOPAL V V. Algorithms for dynamic spectrum access with learning for cognitive radio[J]. IEEE Transactions on Signal Processing, 2010, 58(2):750-760.

[2] JIANG C X, CHEN Y, LIU K J, *et al*. Renewal-theoretical dynamic spectrum access in cognitive radio network with unknow primary behavior[J]. IEEE Journal on Selected Areas in Communications, 2013, 31(3):406-416.

[3] SUN Z W, LANEMAN J N. Secondary access policies with imperfect sensing in dynamic spectrum access networks[A]. IEEE International Conference on Communications(ICC)[C]. Ottawa, Canada, 2012. 1752-1756.

[4] **黄丽亚，刘臣，王锁萍. 改进的认知无线电频谱共享博弈模型**[J]. **通信学报**, 2010, 31(2):136-140.
HUANG L Y, LIU C, WANG S P. Improved spectrum sharing l in cognitive radios based on game theory[J]. Journal on Communications, 2010, 31(2):136-140.

[5] NEDIC A, OZDAGLAR A. Distributed Subgradient Methods for Multi-Agent Optimization[R]. LIDS Technical Report 2755, 2007.

[6] WU G G, REN P Y, DU Q H. Dynamic spectrum auction with time optimization in cognitive radio networks[A]. IEEE Vehicular Technology Conference(VTC Fall)[C]. Quebec City, Canada, 2012. 1-5.

[7] HU J L, WELLMAN M P. Multi-agent reinforcement: learning theoretical framework and an algorithm[A]. Proc 15th Inter ional Conference on Machine Learning, Madison, WI[C]. San Francisco, CA, 1998. 242-250.

[8] SASTRY P S, PHANSALKAR V V, THATHACHAR M A L. Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information[J]. IEEE Transactions on Systems, Man and Cybernetics, 1994, 24(5):769-777.

[9] ZHU H, NIYATO D, SAAD W, *et al*. Game Theory in Wireless and Communication Networks Theory, Models, and Application[M]. The United States of America by Cambridge University Press 2012.205- 392.

[10] LI H S, HAN Z. Competitive spectrum access in cognitive radio networks: graphical game and learning[A]. IEEE Wireless Communications and Networking Conference[C]. Sydney, Australia, 2010. 1-5.

[11] **肖遥，周宗仪. 随机接入协议: 研究综述**[J].**通信技术**, 2003, 1:60-63.
XIAO Y, ZHOU Z Y. Random access protocol: a survey[J]. Communications Technology, 2003, 1:60-63.

[12] IEEE 802.16e-2005 and IEEE Std 802.16-2004 /Cor1-2005[EB/OL]. http:// www.ieee802.org/16/.

[13] KARMOKAR A K, DJONIN D V, BHARGAVA V K. POMDP-based coding rate adaptation for type-I hybrid ARQ systems over fading channels with memory[J]. IEEE Transactions on Wireless Communications, 2006, 5(12):3512-3523.

[14] CLUTTON-BROCK T. Cooperation between non-kin in animal societies[J]. Nature, 2009, 462:51-57.

[15] MITCHELL T M. Machine Learning[M]. McGraw-Hill Education (Asia) Co and China Machine Press, 1997.

[16] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8:279-292.

[17] KOK J R, VLASSIS N. Sparse cooperative Q-learning[A]. Proc Twenty-First International Conference on Machine Learning (ICML-04)[C]. Banff, Canada, 2004. 481-488.

[18] ZHOU P, CHANG Y S, JOHN A. Reinforcement learning for repeated power control game in cognitive radio networks[J]. IEE Journal on Selected Areas in Communications, 2012, 30(1):54-68.

[19] LIU X, WANG J L, WU Q H, *et al*. Frequency allocation in dynamic environment of cognitive radio networks based on stochastic game[A]. 2011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)[C]. Beijing, China, 2011. 497-502.

作者简介:

李方伟（1960-），男，重庆人，重庆邮电大学教授、博士生导师，主要研究方向为移动通信理论与技术、信息安全技术等。

唐永川（1989-），男，重庆人，重庆邮电大学硕士生，主要研究方向为移动通信技术、认知无线电。

朱江（1977-），男，湖北荆州人，重庆邮电大学副教授，主要研究方向为认知无线电。