DOI: 10.3969/j.issn.1009-6868.2018.02.003
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20180322.1452.010.html

基于强化学习的无线网络智能接入控制技术

The Intelligent Access Control Mechanisms in Wireless Network Based on Reinforcement Learning

严牧/YAN Mu
孙耀/SUN Yao
冯钢/FENG Gang
（电子科技大学，四川 成都 611731）
(University of Electronic Science and Technology of China, Chengdu 611731, China)

中图分类号：TN929.5 文献标志码：A 文章编号：1009-6868 (2018) 02-0010-005

摘要： 介绍了无线网络中的强化学习算法，认为由于强化学习算法与环境交互并动态决策的特点，其对复杂网络环境有着较强的适应能力；然后针对无线网络中的强化学习方法的应用场景做了概述，并给出了两个基于强化学习的无线接入技术案例：毫米波技术的切换技术和 Multi-RAT 接入技术。可以看到：智能的无线接入技术由于具备充分挖掘和扩展无线网络资源的潜力，能够显著提高无线网络用户的体验。

关键词： 未来无线网络；切换；接入控制；强化学习

Abstract: In this paper, the application of reinforcement learning in wireless network is briefly introduced. Due to the characteristics of interacting with environment and dynamic decision making, reinforcement leaning algorithm has strong adaptability to complex network environment. Then the application scenarios of reinforcement learning method in wireless network are summarized, and two cases of wireless access technology based on reinforcement learning are given: handoff policy of mmWave HetNets and multi-rat access control. Intelligent access control of wireless network is powerful in exploiting wireless network resources, which can improve the quality of experiences of mobile users.

Key words: future wireless network; handoff; access control; reinforcement learning

当今社会已经迈入信息经济时代，信息技术已成为推动经济结构向多样化消费和低能耗高效发展的重要驱动力。据思科公司预测，到 2019 年全球移动数据总流量将增长至每月 24.3 EB，接近 2000 年全球互联网总流量的 200 倍[1]。另据全球移动通信系统（GSM）协会分析[2]，到 2020 年全球支撑物联网的机器对机器通信（M2M）连接数将达到 9.8 亿，接近 2000 年全球 M2M 连接数的 14 倍。无线通信网络在面临无线资源趋于枯竭的同时，正在经历着前所未有的高增速无线服务需求与低效率无线服务供给之间的矛盾。

未来无线通信将利用复杂异构网络来支持多样化应用场景，包括连续广域覆盖、热点高容量、高可靠低时延以及低功耗巨连接等。由于用户终端性能和业务需求的不同，用户体验质量（QoE）在不同通信场景也存在极大的差异性。出于成本和兼容性的考虑，未来无线网络将长期处于多网共存的状况，包括 2G、3G、4G、5G、Wi-Fi 等，由于不同网络利用不同的无线接入技术，因而形成了接入技术的差异性。同时，为了进一步提升网络的容量，需要在传统接入站点的基础上引入 Micro、Pico、终端直通（D2D）、移动自组织（Adhoc）及小蜂窝等接入站点，因而形成了对网络的重叠异构覆盖。网络的高密度部署和多网络共存使得复杂异构网络下的无线干扰环境变得更加复杂，并对无线接入网的资源调度和控制管理提出了更高的要求。

传统的无线接入技术在“网络-频谱”的静态匹配关系下对网络进行规划设计和资源配置。设备的接入往往基于某一参数（如信号强弱、区域位置）选择单一接入网络和固定接入站点。由于复杂异构网络中海量用户行为的随机性，不同网络的业务需求呈现出极大的时空动态变化特性。静态的“网络-频谱”匹配使得网络容量无法满足变化的网络业务需求，大大地限制了无线网络的接入能

收稿日期：2018-01-18
网络出版日期：2018-03-22
基金项目：国家自然科学基金（61631005、61471089）；中央高校基本科研基金（ZYGX2015Z005）

力，并导致用户接入体验差等问题。

为根本性地提高无线网络接入能力，必须打破传统的无线资源管理和服务接入控制的僵化机制，研究智能的无线接入理论与技术，充分挖掘和扩展无线网络资源的利用潜力，显著提高无线网络用户的体验。在无线网络中，由于用户行为以及网络的动态性和复杂性，使得接入控制和资源分配是非常具备挑战性的[3]。人工智能（AI）技术，比如机器学习，赋予计算机分析环境并解决问题的能力，并提供了一种有效的方法来处理动态性高、复杂度明显的问题[4]。

1 强化学习在无线网络中的应用

强化学习是一种在非确定环境下做决策的强劲的工具[5]。Google Deepmind 最近所研发的 AlphaGo 以及 AlphaGo Zero 所使用的强化学习在围棋这类动态性明显、环境信息复杂的博弈游戏中表现良好[6]，并且取得较好的成绩。在异构网络接入控制的过程当中，由于网络的动态性导致了决策过程也必然是动态性的，我们需要主体和环境进行频繁交互、感知，从而智能化地协调用户和基站的决策行为。因此，强化学习由于其所具备的特点被我们利用到异构网络的决策过程中也是顺其自然的。

1.1 强化学习的分类

（1）根据强化算法是否依赖模型可以分为基于模型的强化学习算法和无模型的强化学习算法。这两类算法的共同点是通过与环境交互获得数据，不同点是利用数据的方式不同。基于模型的强化学习算法利用与环境交互得到的数据学习系统或者环境模型，再基于模型进行决策。无模型的强化学习算法则是直接利用与环境交互获得的数据改善自身的行为。两类方法各有优缺点：一般来讲基于模型的效率比无模型要高，因为智能体可以利用环境信息；但是有些无法建立模型的任务只能利用无模型强化学习算法，因此无模型强化学习算法更具备通用性。

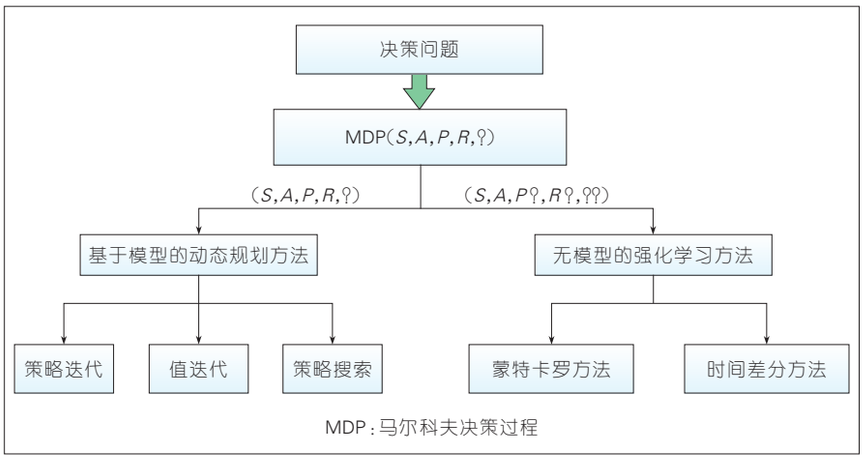
（2）根据策略的更新和学习方法，强化学习算法可分为基于值函数的强化学习算法、基于直接策略搜索的强化学习算法以及 Actor-Critic（AC）的方法。所谓基于值函数的强化学习方法是指学习值函数，最终的策略根据值函数贪婪得到。也就是说，任意状态下，值函数最大的动作为当前最优策略。基于直接策略搜索的强化学习算法，一般是将策略参数化，学习实现目标的最优参数。基于 AC 的方法则是联合使用值函数和直接策略搜索。

（3）根据环境返回的回报函数是否已知，强化学习算法可以分为正向强化学习和逆向强化学习。在强化学习中，回报函数是人为指定的，回报函数指定的强化学习算法称为正向强化学习。很多时候，回报无法人为指定，如无人机的特效表演，这时可以通过机器学习的方法由函数自己学出来回报。

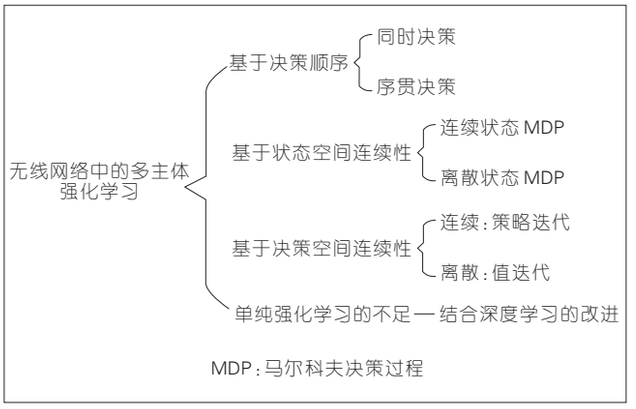
1.2 强化学习在无线网络中的应用

考虑到无线网络的特殊应用场景，在基于图 1 的分类下，我们进一步按照无线网络的特点对强化学习进行分类，如图 2 所示。首先由于受限于网络中有限的频谱资源，用户总是以竞争的关系接入到网络中，那么资源调度、小区切换等考虑用户 QoE 的问题往往可以建模成一个多主体马尔科夫决策过程（MDP）；然后考虑到网络状态空间变化基于时间的连续性或离散性，可将网络决策过程建模为连续时间或者离散时间 MDP，连续时间 MDP 需要决策做到快速反应，尽量做到在线学习；再者，基于网络动作空间的连续性或离散性，有分别基于策略迭代和值迭代的强化学习方法；最后考虑到传统的强化学习方法利用到网络环境中的一些不足，我们可以和深度学习结合起来做一个改进。

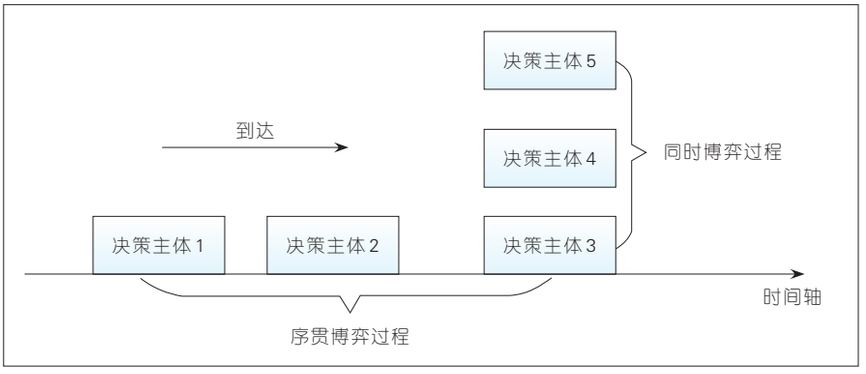
根据做决策的时序先后，我们可以把网络中接入用户的决策分为基于多主体的序贯博弈过程或同时博弈过程，如图 3 所示。具体来说，由于普通的强化学习本身就是基于 MDP 建模，并且解决的是序贯博弈的问题。为了解决同时博弈的问题，我们可以采用 Nash Q-learning 算法[7]。在 Nash Q-learning 的算法中，所有的决策主体在同一个决策时间从一个随机的决策开始去尝试学习它们的最优 Q-value。为了达到这样的目的，每一个主体都通过其他主体的 Q-value 来更新自己的决策，直到达到纳什均衡点。例如：在文献[8]中，作者在认知无线 mesh 网络中考虑在尽可能保证主用户的服务质量（QoS）



▲图 1 基于决策过程的强化学习算法分类



◀图2
无线网络中的多主体强化学习



▲图3 基于决策时间的博弈过程

条件下，为同时接入的次级用户分配功率资源和频谱资源。考虑到次级用户之间的竞争关系（博弈关系），采用了基于多主体的 Nash Q-learning 算法，并得到较好的结果。在决策空间集较小，并且主体数量较少的情况下，Nash Q-learning 是一种很好的用于解决多主体同时博弈的算法。

在无线网络中，经常存在动作（决策）空间过大的现象，例如：在时频资源块分配问题中或者在长期演进（LTE）中非连续接收（DRX）cycle 长度的设置问题中。如果我们把频谱资源或者 cycle 的长度范围划分为较小的决策单元，那么就会使得策略空间异常大，会消耗大量的计算资源。如果我们能通过策略迭代用更平滑的手段去搜索最优策略，会显著增加学习效率，更加贴合无线网络中需求快速决策的特点。

基于状态空间在时间上的连续性或离散性，我们可以把 MDP 建模成连续时间上的 MDP 或离散时间上的 MDP。连续时间 MDP 是基于时间序列连续的马尔科夫过程，其依然具备马尔科夫性。连续时间 MDP 和离散时间 MDP 区别在于时间指标参数从离散的 $T=\{0,1,2...\}$ 改为连续的实数 $T=\{t|t>=0\}$。当我们考虑小时间尺度上的网络问题，由于用户流的不间断涌入，信道质量的无规律变化等都会造成网络状态的频繁波动。因此快速决策就变得尤为重要。这里基于连续空间较好的算法是 AC 算法。AC 较好地平衡了值迭代和策略迭代这两种方法。例如：文献[9]中，作者考虑把基于流量变化下的基站开关操作建模为一个连续状态的 MDP。考虑到用户的接入流量是一个连续变化的过程，那么整个网络的状态也相应具有很强的动态性和连续性。所使用的 AC 算法在该工作中不仅加快了学习速率，TD-error 还具备预测的功能性。

无线网络中，强化学习还可以和深度学习结合起来使用，两者各有优缺点。强化学习本身由于状态空间过大会导致学习时间较长（维度诅咒），在复杂的无线网络环境中，由于网络状态复杂，单纯的强化学习由于算法收敛过慢并不是十分贴合。基于神经网络的深度学习方法，可以利用历史数据对下一时刻的用户行为或者网络状态进行预测。但是，尽管深度学习能够提供较为精准的趋势分析和模式识别，也很难推导出与数据完全匹配的分布函数，在无线网络中带来决策上的明显失误，使得数据失去其应用价值。此外，为了及时保存和处理蜂窝网络数据，基站作为中心控制器需要存储大量的蜂窝网络数据，需要消耗大量的存储和计算资源。因此，我们可以将深度学习利用起来为小时间尺度上的网络决策提供先验信息，从而加速强化学习算法的收敛速度。

2 智能化接入控制案例分析

我们考虑两种智能化的接入控制技术作为案例研究：（1）针对毫米波异构蜂窝网我们提出了一种基于机器学习的智能切换策略，在保证用户服务质量的前提下，减少不必要的切换次数。针对单个用户，在强化学习方法中采用基于置信区间上界（UCB）算法的基站选择策略，可以降低某个用户的切换次数。（2）我们考虑将不同的 QoS 需求的用户接入到蜂窝网和 Wi-Fi 共存的异构网络中。为了在复杂和动态环境中最大化系统吞吐量并且同时满足用户 QoS 需求，我们利用基于多主体强化学习的智能多无线电接入技术，通过动态感知网络环境，来为每个用户分配相应的信道资源。

2.1 基于毫米波技术的智能切换技术

（1）强化学习的奖励函数设计

由于与切换次数相关的奖励函数很难反映，下面我们通过巧妙地设

计奖励函数，来达到最小化 H_n 的目的。我们定义奖励函数为：

$$R_n^k(t)=\int_t^{t_n^k} r_n^k(t)\mathrm{d}t \quad (1)$$

其物理意义是：如果用户（UE）n 选择切换至基站（BS）k，在发生下次切换的这段时间内所传数据的总量。其中 t_n^k 表示的是发生下次切换的时间。可以证明：对 UE n 而言，最小化总的切换次数 H_n 等价于解决上述所提出的具有公式（1）的回报函数的强化学习模型。

（2）估计期望收益

对服务类型相同的 UE，比如服务类型均为 C_n 的 UE，我们让这些 UE 共用一个类回报函数，计作 $R_{C_n}^k(T_{C_n}^k)$，用 C_n 类中的每个 UE 的 $R_n^k(t)$ 来共同更新类回报函数 $R_{C_n}^k(T_{C_n}^k)$。具体的更新定义为：

$$\check{R}_{C_n}^k(T_{C_n}^k+1)=\frac{T_{C_n}^k\check{R}_{C_n}^k(T_{C_n}^k)+\check{R}_n^k(t)}{T_{C_n}^k+1}$$

其中，$T_{C_n}^k$ 代表当前 BS k 被服务类型为 C_n 的用户选中的次数。我们用类回报函数 $R_{C_n}^k(T_{C_n}^k)$ 来作为 BS k 关于服务类型为 C_n 用户的回报函数值就避免了对于某个单个用户而言，无法及时更新 $R_n^k(t)$ 的问题。我们用公式（2）中的方法去估计回报函数的期望。

$$\mathrm{E}[R_n^k(t)]=\begin{cases}\check{R}_{C_n}^k(T_{C_n}^k), if n\in C_n,\\ 0, \text{otherwise}\end{cases} \quad (2)$$

由于处在同一服务类型的用户切换准则相近，在经过一段时间的学习之后，回报函数期望的估计值具有较高的准确性。

（3）基站选择策略

由于强化学习中的定理——探索和利用，我们不能够总是选择当前回报函数期望值最大的基站进行切换。通常，我们用 Regret 来衡量强化学习中的一个策略的优劣程度。Regret 是指所采取的策略与最优策略之间的差距。在我们的这个问题中，UE n 在策略 π 下在执行了 W 次切换后的 Regret 可以表示为：

$$\mathrm{Regret}_\pi(W)=\sum_{\phi(t_n^1)=1}^{W}[R_{\pi^*}(t_n^k)-R_n^k(t_n^k)_\pi] \quad (3)$$

其中，$R_{\pi^*}(t_n^k)$ 代表采用最优策略 π^* 在时刻 t_n^k 所获得的回报。在文献[9]中已经证明所能达到的最优 Regret 是关于切换 W 次数呈对数数量级的。算法 UCB 已经被证明：无论何种形式的 reward 函数都可以实现对数量级的 Regret。UCB 算法的选择策略为：agent 每次选择机器 j^*，其中 j^* 的计算方式为：

$$j^*=\arg\max_j(\bar{x}_j+\sqrt{\frac{2\ln W}{W_j}}) \quad (4)$$

其中，$\bar{x}_j$ 为机器 j 所获得的平均回报值，W_j 代表到目前为止机器 j 被选中的次数，而 W 表示目前为止总的执行决策的次数。

基于 UCB 算法，我们提出了一种目标基站选择策略。我们将基站 j 的 index 设为 $\check{R}_{C_n}^k(T_{C_n}^k)+l\sqrt{\frac{2\ln H_n}{T_{C_n}^k}}$，其中 $l=\max_{k\in A_n,C_n\in C}\check{R}_{C_n}^k(T_{C_n}^k)$，$H_n$ 代表目前为止 UE n 所发生的总的切换次数。因此，基于 UCB 算法，我们提出了 SMART-S 的基站选择策略。UE n 在发生切换后，在可行基站集合 A_n 中选择 BS k^* 进行切换，其中：

$$k^*=\arg\max_k(\check{R}_{C_n}^k(T_{C_n}^k)+l\sqrt{\frac{2\ln H_n}{T_{C_n}^k}}) \quad (5)$$

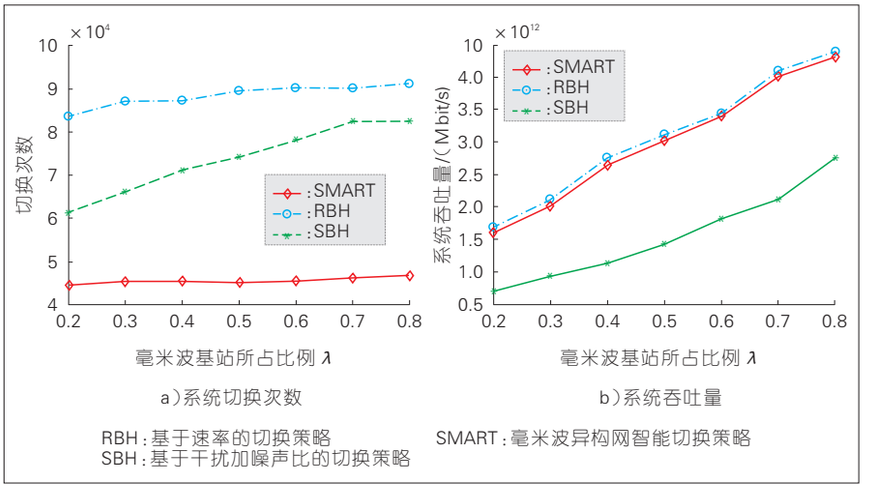
我们考察了毫米波异构网中智能（SMART）切换策略下的性能，并与下面两种传统策略进行了对比：基于速率的切换策略（RBH）是每次用户发生切换时总是选择当前可以提供最大传输速率的基站进行切换；基于干扰加噪声比（SINR）的切换策略（SBH）是用户总是选择可以提供最大信号 SINR 的基站进行切换。图 4 代表了这 3 种切换策略下的系统总的切换次数/系统吞吐量与毫米波小基站（mm-FBS）所占比例 λ 之间的关系。通过图 4 可以看出：我们可以通过较小的系统吞吐量的损失而带来较明显的切换次数的降低。

2.2 Multi-RAT 智能接入技术

为了在复杂和动态环境中最大化系统吞吐量并且同时满足用户 QoS 需求，我们利用基于多主体强化学习方法的智能多无线电接入（SARA）技术，通过动态感知网络环境，来为每个用户分配相应的信道资源。

（1）场景描述

我们研究的场景是蜂窝网小基站（SBS）和 Wi-Fi 热点共存的场景。LTE 下行执行正交频分多址的传输



▲图 4 mm-FBS 所占比例 λ 和系统切换次数/系统吞吐量之间关系图

方式（OFDMA），其频谱资源包含很多的时频资源块（RB），又叫做子信道。在传输的过程中，非连续波段的频谱可以利用传输数据流。为了保护正在进行的会话流，我们假设新到的业务流必须在没有多余频谱资源的情况下进行等待。基站作为中心控制器是能够获取全局的网络信息，包括用户的QoS需求和网络环境信息。由于网络的动态性和跨无线电技术（RAT）的资源调度复杂特性，多无线电技术的聚合需要更加智能化的技术支撑。

（2）基于多主体强化学习的Multi-RAT接入机制

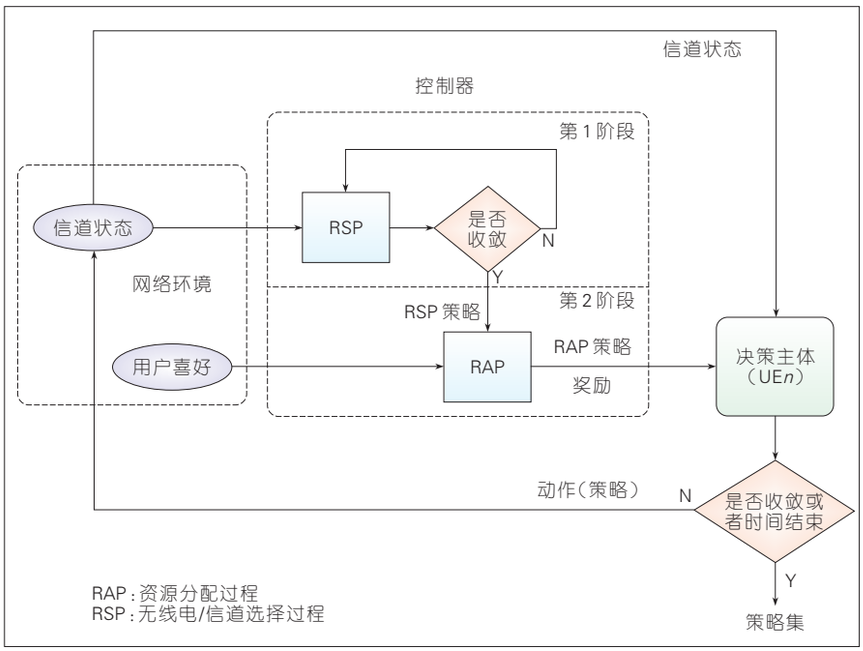
多无线电接入过程是一个多主体的随机过程[9]。在多主体的环境中，我们可以观测到其他所有主体所做的决策已经反馈的回报值。基于该多主体的随机过程，和图5提出的两层决策框架，无线电/信道选择过程（RSP）和资源分配过程（RAP）中分别存在着同时博弈和序贯博弈的过程。我们采取Nash Q-learning算法[9]以及蒙特卡洛树搜索（MCTS）方法[10]来解决这两个博弈的相关问题。

我们把接入过程建模成一个基于半马尔科夫（SMDP）的强化学习模型。具体来说，在我们的工作中有两个决策阶段，如图5所示：第1阶段为RSP，该阶段的目的在于尽可能地去避免碰撞和乱序情况的发生，从而压缩决策空间。当我们的算法收敛后，我们就开始第2阶段——RAP，在该阶段中，基于有限的网络资源和多样的用户喜好，我们考虑去使用有限的信道资源为用户提供合适的服务，并且使得系统平均吞吐量最大化。在这一阶段中，我们假设在蒙特卡洛树搜索中，每一个节点s包含了$\{r(s,a),N(s,a),Q(s,a)\}$的信息，其中$r(s,a)$是即时的奖赏值用来衡量该资源分配决策的好坏，$N(s,a)$是节点的被访问次数，$Q(s,a)$是该节点的Q-value。在决策的搜索过程中，用到了上界信心树搜索（UCT）[11]方法。每个节点所需要满足的是单个用户流的QoS需求，根节点所需要满足的是整个系统的吞吐量的最大化。

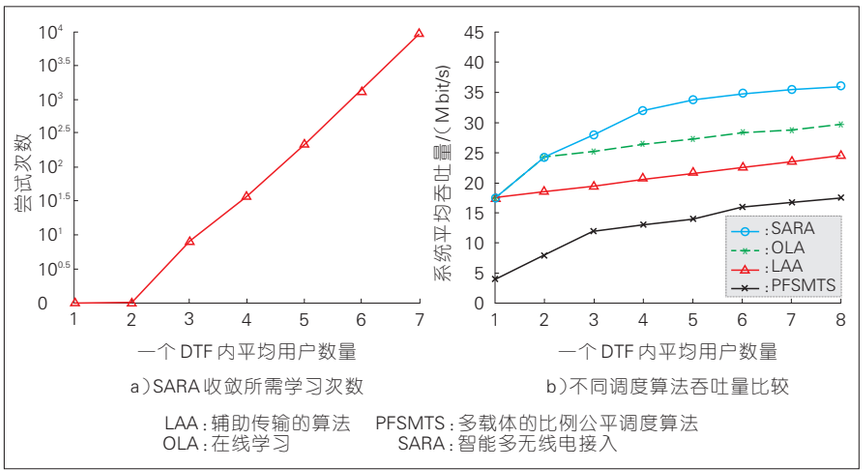
我们使用了下面的一些调度技术作为比较：多载体的比例公平调度算法（PFSMTS）[12]；LTE作为辅助传输的算法（LAA）：在该算法中，Wi-Fi作为流量优先卸载的频段，LTE作为辅助频段；在线学习（OLA）：对SARA中的用户进行流式处理。

从仿真图我们得到的结论分别是：如图6a）所示，SARA所需的收敛学习次数随着用户数量的增加而增加，复杂度也随之上升。考虑到小时间尺度调度特性，我们可以设置在短时间内进行资源调度，这样相应进入用户数量也较少，算法收敛较快，网络性能容易被满足。如图6b）所示，SARA的系统吞吐量性能明显要高于其他的调度算法（当用户数量大于3的时候），这意味着SARA这样的智能化的LTE-WiFi聚合方式可以在动态的环境中明显地提高系统资源的

➡下转第46页



▲图5 基于SMDP的两层决策架构



▲图6 SARA性能仿真结果