



基于强化学习的异构无线网络资源管理算法

冯陈伟,袁江南

(厦门理工学院光电与通信工程学院 厦门 361024)

摘要:为了充分利用各种无线网络的资源,需要实现异构网络的融合,而异构网络的融合又面临接入控制与资源分配的问题。为此,提出一种基于强化学习的异构无线网络资源管理算法,该算法引入 D2D (device-to-device)通信模式,并可以根据终端不同的业务类型、终端移动性及网络负载条件等状态,选择合适的网络接入方式。同时,为降低存储需求,采用神经网络技术解决连续状态空间问题。仿真结果表明,该算法具有高效的在线学习能力,能够有效地提升网络的频谱效用,降低阻塞率,从而实现自主的无线资源管理。

关键词:异构无线网络;接入控制;资源管理;强化学习;Q 学习

doi: 10.11959/j.issn.1000-0801.2015195

Heterogeneous Wireless Network Resource Management Algorithm Based on Reinforcement Learning

Feng Chenwei, Yuan Jiangnan

(School of Opto-Electronic and Communication Engineering, Xiamen University of Technology, Xiamen 361024, China)

Abstract: In order to make full use of the resources of all kinds of wireless network, the integration of heterogeneous network is necessary. However, when it comes to the heterogeneous network integration, the problems of call request access control and resource management emerge. A reinforcement-learning-based algorithm was presented for heterogeneous wireless network resource management. D2D (device-to-device) communication was introduced into the proposed algorithm and the appropriate network for access could be selected according to different traffic types, terminal mobility, network load status and so on. Meanwhile, to reduce the storage requirement, the neural network technology was used to solve the problem of continuous state space. Simulation results show that the proposed algorithm has an efficient learning ability to achieve autonomous radio resource management, which effectively improves the spectrum utility and reduces the blocking probability.

Key words: heterogeneous wireless network, access control, resource management, reinforcement learning, Q-learning

收稿日期:2015-02-16;修回日期:2015-07-28

基金项目:福建省中青年教育科研项目 A 类资助项目(No. JA14233),国家自然科学基金青年科学基金资助项目(No. 61202013),福建省自然科学基金资助项目(No. 2015J01670)

Foundation Items: The Project of Education and Scientific Research of Young Teacher of Fujian (No. JA14233), The Young Scientists Fund of the National Natural Science Foundation of China (No. 61202013), The Natural Science Foundation of Fujian (No. 2015J01670)

论文引用格式:冯陈伟,袁江南. 基于强化学习的异构无线网络资源管理算法. 电信科学, 2015195

Feng C W, Yuan J N. Heterogeneous wireless network resource management algorithm based on reinforcement learning. Telecommunications Science, 2015195



1 引言

未来的无线通信环境是多个异构无线接入技术共存的网络,重叠的网络覆盖、多样的业务需求以及互补的技术特点,使得协调异构无线网络资源变得尤为重要。为此,联合无线资源管理 (joint radio resource management, JRRM)^[1]的方法被提出,用于实现负载均衡^[2]和异构网络选择^[3]。端到端重配置(end-to-end reconfigurability, E2R)^[4]技术的出现,为无线终端提供了动态选择和配置无线工作频率的能力,通过联合接入控制与调度,对无线资源的联合管理更加灵活。

强化学习(reinforcement learning, RL)^[5]是学习者通过与环境互动进行学习的算法。强化学习的目标是学习在每个状态采取何种动作来最大化一个特定的指标。学习者通过反复与控制环境进行交互,利用奖励来评价其性能,从而达到一个最优决策。强化学习广泛应用于机器人和自动控制^[6],同时因其灵活性和自适应性也被引入无线通信系统的资源管理^[7-10]。Q学习是强化学习的一种方法,利用学习者逐步构建一个Q函数,试图估计未来折扣代价,以便于学习者在当前状态采取一定动作。目前,已有文献将Q学习应用于异构无线网络的选择中。参考文献[11]在WiMAX/HSDPA异构网络中,根据系统的负载与无线链路状态,提出基于Q学习的联合网络接入控制算法。参考文献[12]在LTE/WLAN异构网络中,利用Q学习实现网络的智能选择与切换,从而优化用户的体验质量(quality of experience, QoE)。参考文献[13]考虑业务属性对接入控制带来的影响,提出基于Q学习的资源自主联合管理算法,对异构无线网络接入控制进行联合资源优化。参考文献[14]综合考虑业务类型、终端移动性和网络负载条件,提出基于Q学习的异构网络接入选择算法,实现具有适应性的网络接入选择。参考文献[15]提出一种多媒体业务在WCDMA/WLAN异构网络环境下的模糊Q学习接入控制算法,实现准入性估计和驻留时间估计。参考文献[16]利用认知无线电技术,提出基于Q学习的认知无线网络接入方案,使终端根据自身的偏好或者QoS需求,选择最合适的网络接入,同时实现系统的整体能耗开销最优化。参考文献[17]以服务强度和信道占用率等参数作为状态特征,以网络吞吐量、接入失败次数和网络负载均衡度作为即时回报,提出在LTE/WLAN异构网络下的Q学习模型实现接入控制。参考文献[18]提出了由蜂窝和飞蜂窝(femtocell)构成的异

构网络下的Q学习的网络选择方案,将位于不同服务区域的用户的网络选择建模成一个动态演化博弈过程,利用Q学习算法获得均衡来实现网络选择。

然而,许多现有的算法有的不能自主地进行网络接入,有的不能动态地适应变化的无线网络环境。网络应根据实际环境情况进行自主学习,并不断修改其策略实现网络资源的自我管理。针对以上问题,提出一种改进的基于Q学习的联合异构无线网络资源管理算法,引入终端直通(device-to-device, D2D)^[19]通信模式,并根据业务类型、终端移动性和网络负载条件等状态选择合适的网络接入。同时针对系统参数连续性的特点,采用神经网络技术以节约存储空间。仿真结果表明,该算法降低了系统阻塞率并有效地提高了频谱效用。

2 强化学习

针对不断变化的异构无线网络,最简单的收集环境信息的方法就是与环境不断进行交互。强化学习是利用学习者通过不断迭代进行学习的方法,其目标是在每一个状态采取一种动作从而最大化一个特定指标。学习者通过不断与环境进行交互,并且通过从环境中获得的强化信号评估系统性能,从而获取一个优化的控制策略^[20]。

强化学习的基本模型由当前的环境状态集合 $S=\{s_1, s_2, \dots, s_m\}$ 、学习者的动作集合 $A=\{a_1, a_2, \dots, a_n\}$ 、强化信号 R 和学习者的策略 $\pi: s \rightarrow a$ 等基本要素组成。这些要素之间的关系如图1所示^[21]。

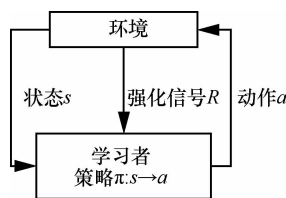


图1 强化学习的基本模型

在 t 时刻,学习者获得环境状态 $s_t \in S$,并根据当前的策略 π 选择动作 $a \in A$,环境受到动作 a 的影响后变为新的环境状态 $s' \in S$,同时产生强化信号 R 并且反馈给学习者。学习者根据强化信号 R 和 s' 更新策略 π ,并进入下一次迭代过程。强化者最终目的是要找到所有状态 s 对应的最佳策略 $\pi^*(s) \in A$,从而最大化期望的长期累积回报,表示为:

$$V^\pi(s) = E \left\{ \sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \mid s_0 = s \right\} \quad (1)$$

其中, $\gamma \in [0,1]$ 为折扣因子。根据 Bellman 最优准则, 式(1)最大值为:

$$V^*(s) = V^{\pi^*}(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P_{s,s'}(a) V^*(s') \right] \quad (2)$$

其中, $P_{s,s'}(a)$ 是状态 s 在动作 a 作用下到达状态 s' 的转移概率, $R(s,a)$ 表示 $r(s,a)$ 的数学期望。 Q 学习是强化学习的一种方法, 能够在未知 $R(s,a)$ 和 $P_{s,s'}(a)$ 的情况下找出最优的策略 π^* 。将策略 π 下的状态与动作 (s,a) 映射为一个 Q 值, 用 $Q^\pi(s,a)$ 表示为:

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s' \in S} P_{s,s'}(a) Q^\pi(s') \quad (3)$$

由式(2)和式(3)得到:

$$V^*(s) = \max_{a \in A} Q^*(s,a) \quad (4)$$

$$\pi^*(s) = \arg \max_{a \in A} Q^*(s,a) \quad (5)$$

Q 学习通过迭代的方式在每个时刻更新 Q 值, 表示如下:

$$Q_{t+1}(s,a) = (1-\alpha)Q_t(s,a) + \alpha (r_t + \gamma \max_{a' \in A} Q_t(s',a')) \quad (6)$$

其中, $\alpha \in [0,1]$ 为学习率。随着 $t \rightarrow \infty$, 若 $Q_t(s,a)$ 能经历无穷多次更新, 且 α 能以某种方式逐渐减小为 0, 则 $Q_t(s,a)$ 将收敛于最优值 $Q^*(s,a)$, 利用式(1)便可获得最优策略 $\pi^{*[22]}$ 。

3 问题映射

系统模型是由一个运营商运营下的 LTE、small cell (小小区) 以及 D2D 构成的单小区异构无线网络, 如图 2 所示。

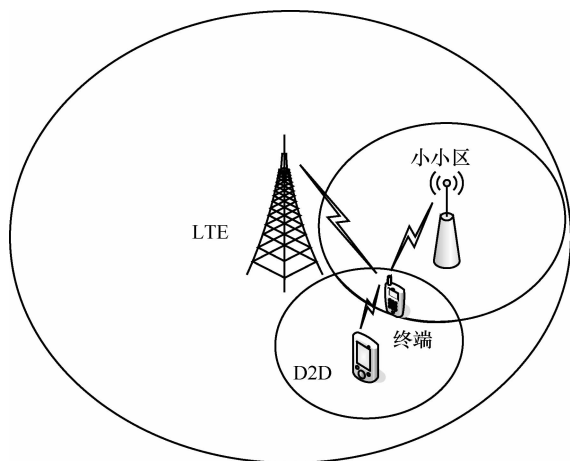


图 2 异构无线网络模型

假设系统中 LTE 基站的控制信令可以覆盖整个小区, 从而可以对所有的会话接入进行控制。考虑业务传输的有

效性, LTE 基站的业务有效传输范围要比其信令传输范围小, 因此在进行网络接入选择时, 需要考虑终端是否在其有效通信范围内。

小小区是运营商用于增强网络覆盖和数据业务提供的一种低成本、低开支的解决方案。D2D 是在蜂窝网络控制下的用户间直连通信方式, 由蜂窝网络进行信令控制, 业务传输在 D2D 用户间直接进行, 不但可以为核心网络分流, 还具有传输速率高、时延低等优点。当用户终端在 D2D 设备有效通信范围内, 并且可以与 D2D 设备直接通信或利用 D2D 设备作为中继进行间接通信时, 静止时用户终端优先选择 D2D 将很大程度地提升网络性能。模型中只考虑两种类型的业务: 语音业务与数据业务。LTE、small cell 与 D2D 有着不同的特点, LTE 网络由于其实时性与连续性的特点更适合语音业务, small cell 网络更适合高带宽的数据业务, D2D 网络由于是直连通信方式, 适合所有业务。然而, 从终端移动性的角度考虑, LTE 网络更适合高速移动终端, small cell 网络更适合低速移动或固定终端, 而 D2D 网络由于其覆盖范围最小, 只适合固定终端。

结合以上模型, 系统中的 JRRM 控制器利用 Q 学习算法, 根据终端访问的业务类型、终端移动性、网络有效覆盖范围、网络负载条件以及是否为 D2D 模式等条件, 选择合适的网络进行接入。在将 Q 学习算法应用于异构无线网络选择前, 需要将系统状态、动作和强化信号等因素映射到实际的接入模型中, 具体的映射过程如下所述。

(1) 状态集合 S

系统状态包括模型中所有网络的有效覆盖范围、负载条件、业务请求类型、终端移动性、直连通信需求等, 因此将状态集合定义如下:

$$S = \{C, D, V, M, L\} \quad (7)$$

其中, $C = \{c_1, c_2, c_3\}$ 表示网络的有效覆盖范围集合, c_1, c_2, c_3 分别表示终端是否在 LTE、small cell 与 D2D 网络的有效通信覆盖范围内, $c_i = 1$ 时表示在对应的第 i 个网络的有效通信覆盖范围内, 反之亦然, 但排除 c_i 均为 0 的情况, 即终端至少在一个网络的覆盖范围内。 $D = \{1, 0\}$ 表示终端是否具有直连通信需求, 1 表示具有直连通信需求, 反之亦然; $V = \{\text{语音业务}, \text{数据业务}\}$ 表示业务请求类型; $M = \{\text{静止}, \text{运动}\}$ 表示终端移动性; L 表示系统中的网络负载条件, 即每个网络中已使用资源与总资源的比例, 可以看作连续变量。



(2) 动作集合 A

JRRM 控制器作为学习者,根据状态和学习经验,为终端选择适当的网络接入并进行后续的资源分配,因此动作集合定义为:

$$A=\{LTE, small\ cell, D2D\} \quad (8)$$

表示选择对应的网络进行接入。

(3) 强化信号 R

强化信号是用于评估在状态 s 下,会话请求被接入后所产生的即时回报。由于网络有效覆盖范围、业务请求类型、终端移动性等因素,不同的网络接入选择及资源分配将会对系统性能产生不同的影响。因此,将强化信号定义如下:

$$R=\begin{cases} \beta \times c_i \times (\eta_v + \eta_w), & \text{执行动作 } a \\ 0, & \text{拒绝会话} \end{cases} \quad (9)$$

其中, η_v 表示业务请求类型与接入网络的匹配系数, η_w 表示终端移动性与接入网络的匹配系数。 c_i 表示终端是否在第 i 个网络的有效通信覆盖范围内。 β 是负载因子,表示所选择网络中的剩余资源与总资源的比例。从强化信号可以看出,如果终端在网络的有效通信范围内,且选择接入的网络能更好地匹配业务请求类型和终端移动性,则获得匹配值将会较大,否则获得的匹配值将较小;同时,强化信号中的负载因子 β 可以及时引导请求会话接入低负载网络,起到一定的负载均衡作用。

为了评价网络性能,体现网络服务用户所获得的收益,定义比例公平的累积频谱效用 U ,表示为:

$$U=\ln\left[\sum_j b_j \times c_j \times (\eta_v + \eta_w)\right] \quad (10)$$

其中, b_j 是为会话 j 所分配的带宽。

4 算法流程

在综合考虑了网络的有效覆盖范围、负载条件、业务类型以及终端移动性等因素后,将 Q 学习算法应用于异构无

线网络的接入选择以及资源分配。基于在线学习的 Q 学习算法包含两个方面:策略更新和 Q 值更新。

(1) 策略更新

为了学习最优决策,学习系统既要尽可能遍历各种动作进行学习,又要根据 Q 值表中的以往信息进行当前状态下的最优动作选择。因此为了使所有状态动作对 (s, a) 都被访问到,同时兼顾效率,每一次迭代,学习者以较大的 $(1-\epsilon)$ 概率选择当前状态下 Q 值最大的动作,以较小的 ϵ 概率在其余动作中随机选择一个。

(2) Q 值更新

记录并获得每个状态下各个动作的 Q 值,根据贪婪算法选择动作,当迁移到下一个状态时, Q 值根据式(6)进行更新。

Q 值的记录可以采用查找表(lookup table)的方式,也可利用神经网络技术^[23]的函数近似的方式。采用查找表的方式较为简单与直接,但仅适用于状态空间较少的情况,而对于状态空间较大甚至连续时,该方法难以实现。因此采用神经网络技术记录 Q 值(如图3所示),多层前馈神经网络(multi-layer feedforward neural network, MFNN)被集成在学习者中,环境状态构成的参数向量作为输入,当前状态下对应所有动作的 Q 值向量作为输出,基于获得的 Q 值向量,采用贪婪算法进行动作的选择,执行动作后得到强化信号,根据式(6)更新 Q 值,更新的 Q 值作为训练序列调整神经网络的权值。

具体的算法流程如下。

(1) 参数初始化。将 Q 值表中的 Q 值都初始化为 0,随机设置神经网络的初始权值,初始化训练队列,初始化折扣因子 γ 、初始学习率 α_0 、初始探索概率 ϵ_0 。

(2) 状态构建。当新的会话到达时, JRRM 控制器收集相关状态信息,包括当前每个网络的网络负载条件以及新会话的业务特征,根据式(7)生成当前状态 s ,并转化为向

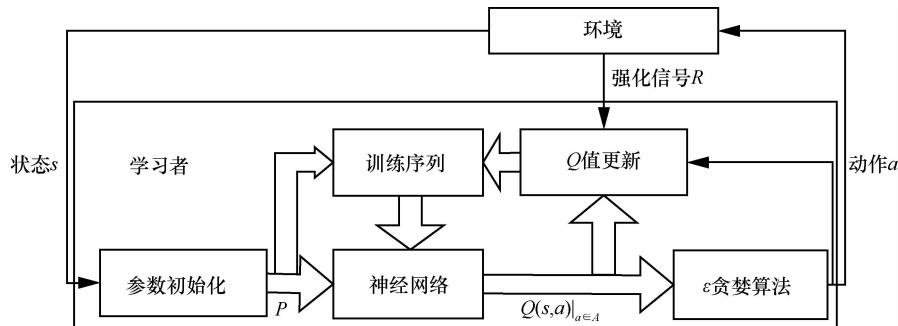


图3 基于 Q 学习的资源管理模型

量 P 作为神经网络的输入。

(3) Q 值获取。向量 P 经过神经网络运算后得到当前状态下各动作对应的 $Q(s,a)$ ，并保存起来作为 Q 值更新使用。

(4) 选择动作。根据神经网络输出的 $Q(s,a)$ ，基于 ε 贪婪策略选择一个动作执行，即选择一个网络接入，被执行的动作缓存起来作为 Q 值更新使用。

(5) 根据式(4)获得强化信号 R 与下一状态 s' 。

(6) Q 值更新。根据式(6)更新 $Q(s,a)$ ，并输入训练队列调整神经网络的权值。

(7) 更新参数。每一次迭代后，学习率 α 和探索概率 ε 需要进行更新。为了满足收敛要求，它们在学习过程中，利用负指数规律逐渐减小为 0。

(8) 返回步骤(2)。

需要注意的是，会话的到达或结束都会影响网络状态，但是 JRRM 控制器在会话结束时不进行任何操作，只有在会话到达时才进行处理。

5 仿真分析

利用如图 2 所示的仿真模型，考察一个由单小区重叠覆盖的 LTE、small cell 和 D2D 构成的异构无线网络。会话到达率服从均值为 180 呼叫/h 的泊松分布，业务持续时间服从均值为 120 s 的指数分布^[13,14]。通过改变用户强度系数来模拟网络的繁忙程度，用户强度系数越大，则网络呼叫数越多。网络中仅有语音业务与数据业务，语音业务的带宽请求为 32 kbit/s，数据业务的带宽请求为 64 ~128 kbit/s。其他仿真参数见表 1，匹配系数的相对大小反映了网络对相应业务及移动性的贡献度程度。

表 1 仿真参数设置

参数		LTE	small cell	D2D
容量/(kbit·s ⁻¹)		1 000	2 000	4 000
业务匹配系数 η_V	语音	5	1	8d
	数据	1	5	8d
移动性匹配系数 η_M	移动	5	1	0
	静止	1	5	5d
其他参数	折扣因子 γ	0.8		
	初始学习率 α_0	0.5		
	初始探索概率 ε_0	0.5		
	d	$\begin{cases} 1, & \text{终端具有 D2D 模式} \\ 0, & \text{其他} \end{cases}$		

仿真评估基于 Q 学习的 JRRM 算法 (Q-learning JRRM algorithm, QLJA) 的阻塞率与频谱效用。作为比较，还仿真了未使用 Q 学习的基于负载均衡的 JRRM 算法 (load balancing JRRM algorithm, LBJA) 和未考虑 JRRM 概念的随机接入算法 (random access algorithm, RAA)。其中，LBJA 是以最小化阻塞率为目标，为会话分配剩余资源最多的网络。

图 4 与图 5 分别给出了不同用户强度系数下各算法的阻塞率和累积频谱效用。如图 4 所示，随着用户强度系数变大，网络中会话接入数逐渐变多，使得阻塞率逐渐变高，因此各算法阻塞率的趋势一致。LBJA 阻塞率最低，因为会话请求总是接入低负载网络从而进行负载均衡，因此各网络都不容易饱和，从而大大降低阻塞率。RAA 阻塞率最高，因为没有考虑网络剩余资源，随机选择网络，很可能选择一个满负载网络进行接入，从而导致阻塞率增加。QLJA 考虑了网络负载条件这一参数，即考虑网络的负载均衡，同时还兼顾网络对业务以及移动性的匹配程度，因

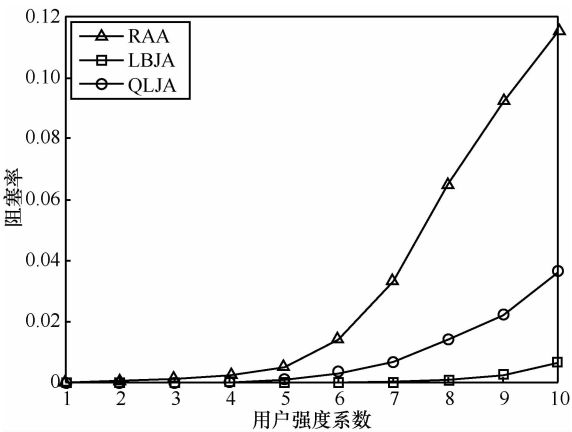


图 4 阻塞率随用户强度系数变化的比较

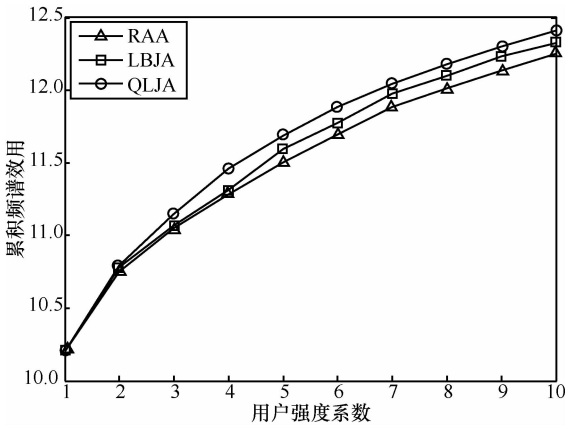


图 5 累积频谱效用随用户强度系数变化的比较



此阻塞率介于 LBJA 与 RAA 之间。

从图 5 可以看出,随着用户强度系数变大,接入网络的呼叫总数也将增大,根据累积频谱效用的表达式,网络将累积所有接入呼叫产生的效益值,使得累积频谱效用逐渐变大,因此各算法累积频谱效用的趋势一致。QLJA 的累积频谱效用最大,这是因为不同业务类型和终端移动类型匹配性得以考虑。虽然 LBJA 的阻塞率低于 QLJA,但是 LBJA 没有考虑业务类型和终端移动性对网络性能的影响,从而导致累积频谱效用比 QLJA 小。RAA 性能最差,一方面原因与 LBJA 一样,都未考虑匹配性问题,另一方面原因是高阻塞率直接影响了累积频谱效用。

图 6 与图 7 分别给出了固定业务强度下,各算法的阻塞率和累积频谱效用随时间变化的情况。如图 6 所示,RAA 与 LBJA 的阻塞率总体较为稳定,而 QLJA 在学习过程中阻塞率逐渐收敛。QLJA 刚开始由于盲目探索导致阻塞率较高,但经过不断地学习,阻塞率迅速收敛到一个可以接受的水平,说明 QLJA 具有很好的“在

线”学习能力。

从图 7 可以看出,QLJA 的累积频谱效用在初期与其他算法基本一样,但随着时间的推移,QLJA 的性能逐渐变好,而且由于频谱效用的“累积”作用,QLJA 与其他算法的性能差异会越来越大。

小区中真实的业务量通常是时变的,为了反映实际通信情况,仿真 QLJA 在两天内的动态业务情况下的阻塞率。图 8 给出了两天内各时段会话到达率情况,第二天的会话到达率与第一天一样。图 9 给出 QLJA 在图 8 所示的业务量下的阻塞率。可以看出,第二天的阻塞率明显好于第一天,这是因为通过在线学习,基于 QLJA 的系统可以有效地利用已学习到的经验来指导后续的策略选择,从而获得性能的提升。

根据上文提及的 D2D 的特性,当存在 D2D 网络并且终端是静止时,无论是语音业务还是数据业务,终端都接入 D2D 网络,D2D 网络对语音业务与数据业务的贡献度一样。因此表 2 与表 3 只比较 3 种算法在 LTE 网络与 small cell 网络中对业务分布的影响,用不同业务的负载比

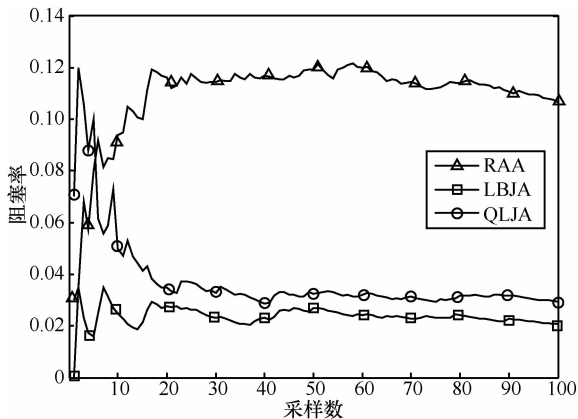


图 6 阻塞率随时间变化的比较

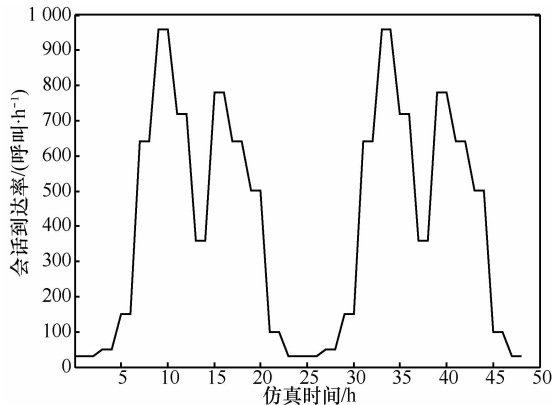


图 8 动态会话到达率

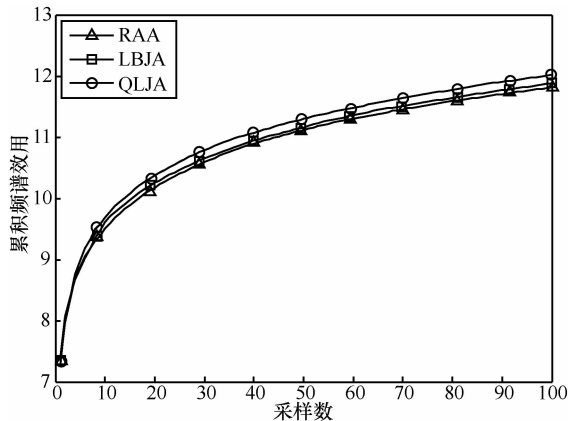


图 7 累积频谱效用随时间变化的比较

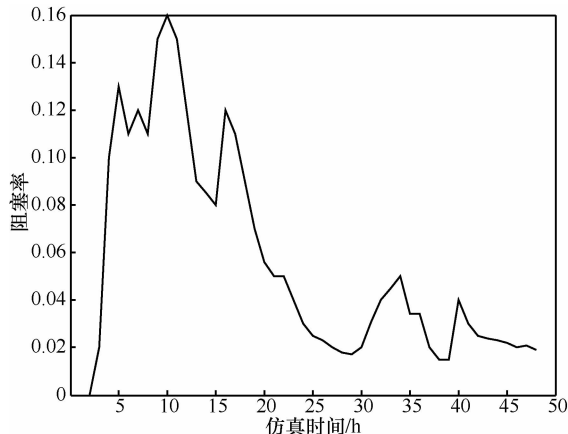


图 9 QLJA 的阻塞率随时间变化情况

表 2 LTE 与 WLAN 中语音和数据负载比例差

网络	RRA	LBJA	QLJA
LTE	0.04	0.05	0.38
small cell	0.03	-0.05	-0.35

表 3 LTE 与 WLAN 网络中静止业务和移动业务负载比例差

网络	RRA	LBJA	QLJA
LTE	0.03	0.07	-0.39
small cell	-0.04	-0.02	0.38

例差表示:

不同业务的负载比例差

$$= \frac{X \text{ 网络中业务类别 } A \text{ 的总量}}{X \text{ 网络中业务类别 } A+B \text{ 的总量}} - \frac{X \text{ 网络中业务类别 } B \text{ 的总量}}{X \text{ 网络中业务类别 } A+B \text{ 的总量}}$$

表 2 表示语音业务所占比例、数据业务所占比例分别在 LTE 与 small cell 网络中的结果,表 3 表示静止业务所占比例、移动业务所占比例分别在 LTE 与 small cell 网络中的结果。从表 2 与表 3 中可以看出,RRA 和 LBJA,不管是从业务类型还是从终端移动性来分析,都不能较好地区分业务。QLJA 使资源得到了优化的配置,LTE 中的语音业务以及移动性业务的比例均远高于数据业务及静止业务;相反地,small cell 中的业务与其相反。这样能使不同接入网充分发挥其技术优势,从而提高系统的总体收益。

6 结束语

在 LTE、small cell、D2D 构成的异构无线网络下,研究基于 Q 学习的资源管理策略。综合考虑网络有效覆盖范围、负载条件、业务类型、终端移动性等因素,JRRM 控制器可以根据网络特性,合理地地为每个到达的会话分配适合的网络进行接入并进行资源分配。同时针对实际系统参数的连续性,引入神经网络技术从而降低存储需求。仿真结果表明,QLJA 能在保证接近低呼叫阻塞率的同时,获得更好的资源配置和更高的系统收益,同时用户感知性能也通过学习收敛过程得以增强。

参考文献

- Luo J, Mukerjee R, Dillinger M, *et al.* Investigation of radio resource scheduling in WLANs coupled with 3G cellular network. *IEEE Communications Magazine*, 2003, 41(6): 108~115
- 3GPP TR 25.881 v5.0.0. Improvement of RRM across RNS and RNS/BSS (Release5). <http://www.3gpp.org>, 2001
- Song Q, Jamalipour A. Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques. *IEEE Wireless Communications*, 2005, 12(3): 42~48
- Zhang Y, Chen J, Zhang P. Autonomic joint radio resource management in B3G environment using reinforcement learning. *Proceedings of the 6th Annual Wireless Telecommunications Symposium (WTS 2007)*, Pomona, California, USA, 2007
- Barto A G. *Reinforcement Learning: an Introduction*. Cambridge: MIT Press, 1998
- Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 1996 (4): 237~285
- Nie J, Haykin S. A Q-learning-based dynamic channel assignment technique for mobile communication systems. *IEEE Transactions on Vehicular Technology*, 1999, 48(5): 1676~1687
- Senouci S M, Beylot A L, Pujolle G. Call admission control in cellular networks: a reinforcement learning solution. *International Journal of Network Management*, 2004, 14(2): 89~103
- Haddad M, Altman Z, Elayoubi S E, *et al.* A nash-stackelberg fuzzy Q-learning decision approach in heterogeneous cognitive networks. *Proceedings of Global Telecommunications Conference (GLOBECOM 2010)*, Miami, Florida, USA, 2010: 1~6
- Simsek M, Czylik A. Decentralized Q-learning of LTE-femtocells for interference reduction in heterogeneous networks using cooperation. *Proceedings of 2012 International ITG Workshop on Smart Antennas (WSA)*, Dresden, Germany, 2012: 86~91
- Saker L, Ben Jemaa S, Elayoubi S E. Q-learning for joint access decision in heterogeneous networks. *Proceedings of Wireless Communications and Networking Conference*, Budapest, Hungary, 2009: 1~5
- Tabrizi H, Farhadi G, Cioffi J. Dynamic handoff decision in heterogeneous wireless systems: Q-learning approach. *Proceedings of 2012 IEEE International Conference on Communications (ICC)*, Ottawa, Canada, 2012: 3217~3222
- 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. *电子与信息学报*, 2008, 30(3): 676~680
- Zhang Y, Feng Z, Zhang P. A Q-learning based autonomic join radio resource management algorithm. *Journal of Electronics and Information Technology*, 2008, 33(3): 676~680
- Zhao Y, Zhou W, Zhu Q. Q-learning based heterogeneous network selection algorithm. *Proceedings of the 2nd World Congress on Computer Science and Information Engineering*, Changchun, China, 2011: 471~477
- Chen Y H, Chang C J, Huang C Y. Fuzzy Q-learning admission control for WCDMA/WLAN heterogeneous networks with multimedia traffic. *IEEE Transactions on Mobile Computing*, 2009, 8(11): 1469~1479



- 16 谭力, 陈亚迷, 李一喆等. 一种基于 Q-学习的认知无线网络接入方案. 新型工业化, 2011(9): 37~44
Tan L, Chen Y, Li Y, *et al.* Q-learning based cognitive wireless network access mechanism. Journal of New Industrialization, 2011(9): 37~44
- 17 陈旋. 基于 Q 学习的 LTE/WLAN 网络接入控制算法研究(硕士学位论文). 哈尔滨工业大学, 2013
Chen X. Call admission control in LTE/WLAN network based on Q-learning (master dissertation). Harbin Institute of Technology, 2013
- 18 Tan X, Luan X, Cheng Y, *et al.* Cell selection in two-tier femtocell networks using Q-learning algorithm. Proceedings of the 16th International Conference on Advanced Communication Technology, Pyeongchang, Korea, 2014: 1031~1035
- 19 Doppler K, Rinne M, Wijting C, *et al.* Device-to-device communication as an underlay to LTE-Advanced networks. IEEE Communications Magazine, 2009, 47(12): 42~49
- 20 Barto A G, Bradtke S J, Singh S P. Learning to act using real-time dynamic programming. Artificial Intelligence, 1995, 72(1): 81~138
- 21 Watkins C J C H, Dayan P. Q-learning. Machine Learning, 1992, 8(3~4): 279~292
- 22 叶培智. 异构无线网络接入选择算法的研究(硕士学位论文). 厦

门大学, 2013

Ye P Z. Research on access selection in heterogeneous wireless networks(master dissertation). Xiamen University, 2013

- 23 Senouci S, Beylot A, Pujolle G. Call admission control in cellular networks: a reinforcement learning solution. International Journal of Network Management, 2004, 14(2): 89~103

[作者简介]



冯陈伟,男,厦门理工学院光电与通信工程学院讲师,主要研究方向为下一代无线通信技术。



袁江南,男,博士,厦门理工学院光电与通信工程学院副教授,主要研究方向为通信信号处理。