ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.6, June 2007, pp.1491−1500
DOI: 10.1360/jos181491

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

无线多媒体通信网适应带宽配置在线优化算法*

江 琦+, 奚宏生, 殷保群

(中国科学技术大学 自动化系,安徽 合肥 230027)

An Online Adaptive Bandwidth Allocation Optimization Algorithm for Wireless Multimedia Communication Networks

JIANG Qi+, XI Hong-Sheng, YIN Bao-Qun

(Department of Automation, University of Science and Technology of China, Hefei 230027, China)

+ Corresponding author: Phn: +86-551-3620467, E-mail: jiangqi@mail.ustc.edu.cn

Jiang Q, Xi HS, Yin BQ. An online adaptive bandwidth allocation optimization algorithm for wireless multimedia communication networks. *Journal of Software*, 2007,18(6):1491−1500. http://www.jos.org.cn/1000-9825/18/1491.htm

Abstract: The issue of QoS (quality of service) provisioning for adaptive multimedia in wireless communication networks is considered. A reinforcement learning based online adaptive bandwidth allocation optimization algorithm is proposed. First, an event-driven stochastic switching model is introduced to formulate the adaptive bandwidth allocation problem as a constrained continuous-time Markov decision problem. Then, an online optimization algorithm that combines policy gradient estimation by learning and stochastic approximation is derived. This algorithm can online handle the constrained optimization problem efficiently without explicit knowledge of the underlying system parameters. Moreover, this algorithm does not require the computation of performance potentials or other related quantities (e.g. Q-factors), which is necessary in previous schemes, and therefore saves computational cost significantly. Simulation results demonstrate the effectiveness of the proposed algorithm.

Key words: adaptive bandwidth allocation; Markov decision processes; policy optimization; reinforcement learning; stochastic approximation; QoS (quality of service) provisioning

摘 要: 基于强化学习的方法,提出一种无线多媒体通信网适应带宽配置在线优化算法,在满足多类业务不同QoS(quality of service)要求的同时,提高网络资源的利用率.建立事件驱动的随机切换分析模型,将无线多媒体通信网中的适应带宽配置问题转化为带约束的连续时间 Markov 决策问题.利用此模型的动态结构特性,结合在线学习估计梯度与随机逼近改进策略,提出适应带宽配置在线优化算法.该算法不依赖于系统参数,如呼叫到达率、呼叫持续时间等,自适应性强,计算量小,能够收敛到全局最优,适用于复杂应用环境中无线多媒体通信网适应带宽配置的在线优化.仿真实验结果验证了算法的有效性.

* Supported by the National Natural Science Foundation of China under Grant No.60574065 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z114 (国家高技术研究发展计划(863)); the Natural Science Foundation of Anhui Province of China under Grant Nos.050420301, 070412063 (安徽省自然科学基金); the Graduate Student Innovation Foundation of USTC under Grant No.KD2006036 (中国科学技术大学研究生创新基金)

Received 2005-12-24; Accepted 2006-02-23

关键词: 适应带宽配置;Markov 决策过程;策略优化;强化学习;随机逼近;QoS(quality of service)保证
中图法分类号: TP393 文献标识码: A

随着无线技术的发展及综合业务需求的增长,下一代无线通信网提供多媒体无线业务(multimedia wireless services,简称 MWS)服务.MWS 包括语音、视频、多媒体、宽带数据等.MWS 伴随着更高的网络带宽需求,且具有多种 QoS 要求需要满足.无线链路带宽资源紧缺及业务请求到达的高波动性,使得网络提供 QoS 保证更为严峻.适应带宽机制的引入能够有效缓解这一问题,它提供了灵活配置系统带宽资源的功能,通过对系统中正在接受服务的呼叫业务的带宽进行动态配置来适应网络运行的波动.适应带宽机制得到两大国际标准的支持,包括国际标准化组织(ISO)的 MPEG-4 和国际电信联盟(ITU)的 H.263,将在下一代无线通信网中得以广泛应用.有效地适应带宽配置能够在满足各类 MWS 多种 QoS 要求的同时,提高网络资源的利用率.

适应带宽配置策略及其优化算法近年来得到研究者的广泛关注,提出了多种具有不同针对性的带宽配置方案:文献[1]对单一种类业务的适应带宽配置问题进行研究,所提出的算法不能简单推广至多类业务的应用;文献[2]提出一种针对多类业务的公平性算法,各类业务之间的公平通过依据到达率划分带宽来实现,在保证公平性的同时导致所有在线业务的使用带宽降低;文献[3]采用预约式带宽配置方案,通过为特定种类的业务预留一定容量的带宽,以满足其 QoS 的要求,其实质是对特定业务种类的 QoS 的局部优化,结果导致全局的次优,如导致带宽利用率及其他业务 QoS 的降低;文献[4]采用固定预留带宽方案区分不同业务种类的优先级,在建立多维 Markov 决策过程模型进行性能分析的基础上,运用线性规划方法给出最优策略的乘积解,适用于数值计算;文献[5]提出的算法根据网络状态动态调整预留带宽容量以进一步近似全局最优;文献[6]提出一种全带宽配置分析模型,将可配置带宽资源进行总体配置,多类业务的不同 QoS 要求通过优先级的权重设置来实现,适用于进行理论分析;文献[7]提出一种遗传算法来优化带宽配置策略,以系统收益最大为优化目标,而没有考虑相应的 QoS 约束.无线多媒体通信网由于应用环境的复杂性,系统参数往往难以预先精确获取并且具有时变性,使得上述方法在实际应用中具有较大的局限性;文献[8]将适应多媒体的 QoS 保证问题转化为 Markov 决策问题,提出一种基于 Q-学习的带宽配置策略的优化算法.该算法不依赖于系统参数,能够实现在线优化.其不足之处是,只能优化确定型策略,得到的是对最优随机策略的近似值,同时,Q-学习的方法运用于平均性能准则的优化具有较大的计算复杂度.

本文面向复杂应用环境中无线多媒体通信网,将适应多媒体带宽配置问题转化为带约束 Markov 决策问题,提出一种基于强化学习的在线优化算法,有效地实现了系统在参数未知情况下的带宽配置策略优化.该方法与文献[8]的方法相比,具有以下特点:与文献[8]采用的行动基于状态的离散时间分析模型不同,本文构建的事件驱动的连续时间随机切换模型为实时系统的动态特性提供了精确的描述,计算与评估只需在事件发生时刻进行,减少了计算量,且该模型容易推广应用至参数服从一般分布的系统;本文采用基于策略梯度的强化学习方法,不同于文献[8]的基于值迭代的 Q-学习方法,能够有效优化随机策略,更适合处理带约束的优化问题.文献[9]提出了基于离散时间 Markov 决策过程的在线梯度估计与随机逼近算法,在此基础上,本文提出的算法有以下改进和推广:本文基于连续时间 Markov 决策模型,从基于性能势的性能梯度公式[10]出发推导在线梯度估计式,但该梯度估计式并不需要计算各状态的性能势,减少了计算量,提高了算法的实时性;充分利用随机切换模型事件驱动的特性,使得在线梯度估计不依赖系统参数的信息,如各类呼叫的到达率、呼叫持续时间、小区停留时间等,提高了算法的自适应性;同样,通过对模型动态结构特性的利用,保证算法收敛到全局最优,克服了策略梯度法固有的缺陷,即通常情况下只收敛到局部最优,从而保证了算法的有效性.此外,算法考虑了各类业务的不同优先级,并将新呼叫与越区切换呼叫进行区分;能够处理各类 QoS 约束,考虑了 3 种重要的 QoS 指标,其他感兴趣的指标也可用类似方法处理,因而具有较为广泛的适用性.

1 系统模型

1.1 适应带宽配置问题

无线网络带宽资源的紧缺和呼叫业务到达的高波动性导致网络拥塞时常发生,从而降低了 QoS.为了有效缓解这一状况,下一代(3G,4G)无线多媒体通信网在采用微/微微蜂窝(micro/pico)结构提高无线信道带宽的同时,引入更为有效的适应带宽机制.基于分层编码技术,无线多媒体呼叫业务能够在多个层级进行编码,从而具有多个带宽取值.如视频业务可以分为若干层级进行编码,包括一个基本层级和多个扩展层级.基本层级独立编码并提供基本的视频质量,扩展层级结合基本层级进行编码,提供精细的视频细节.这样,一个视频呼叫业务通过分层编码技术可以具有不同的带宽取值.呼叫业务采用不同的带宽编码以适应网络的运行状况,称为适应带宽业务.无线通信网根据网络的运行状况动态调整适应多媒体的带宽,从而缓解网络带宽需求的高波动性.在UMTS(universal mobile telecommunication systems)系统中,为呼叫业务建立的无线承载(radio bearer,简称 RB)能够在呼叫持续过程中进行动态重新配置,通过对 RB 的再配置,呼叫业务的带宽在其生存期内能够进行动态调整.

无线通信网络的用户在网络覆盖范围内自由移动,微/微微蜂窝结构的采用导致用户在小区间频繁切换,加剧了网络带宽需求的波动,导致呼叫阻塞率和切换掉线率增加.适应带宽机制的采用导致呼叫平均配置带宽的降低.无线多媒体通信系统中考虑的 QoS 指标主要有呼叫阻塞率(new call blocking probability,简称 NCBP)、切换掉线率(handoff call dropping probability,简称 HCDP)、平均配置带宽(average bandwidth,简称 AB).NCBP 表示新的呼叫到达而得不到网络服务的概率,HCDP 为呼叫越区切换时被终止服务的概率,AB 是指呼叫业务在接受服务期间所使用的平均服务带宽,通常要求大于特定值.不同种类的业务具有不同的 QoS 要求,实时业务对 NCBP 和 HCDP 要求更高,视频业务对 AB 要求更为迫切,而用户对 HCDP 比 NCBP 更为敏感,各类业务在优先级上需要加以区别.

提高网络资源的利用率以提高网络收益是服务提供者的重要目标,对呼叫业务提供 QoS 保证同样是其不可忽视的紧迫问题,适应带宽配置试图在二者间达到最佳的平衡和折衷.系统通过对在线的呼叫业务进行动态适应带宽配置,在提供 QoS 保证的同时提高网络资源的利用率.当基站的带宽利用处于非饱和状态时,总是尽可能地接受所有到达的呼叫业务并满足其最大带宽需求.而当网络拥塞发生时,NCBP 和 HCDP 上升,QoS 保证机制发挥作用.系统通过降低正在服务的呼叫业务带宽,以接纳新的呼叫和切换转移的呼叫,这同时降低了呼叫业务的平均配置带宽,而受到另一 QoS 指标 AB 的制约.另一方面,当系统中接受服务的呼叫结束或切换至相邻小区时,释放出的带宽将重新配置给正在服务的呼叫业务.适应带宽配置根据系统的运行状况和带宽配置策略决定对哪类呼叫配置何种带宽,在保持 AB 大于设定值的同时,降低 NCBP,HCDP,并使网络资源的利用最大化.适应带宽配置的有效性取决于策略的优劣,带宽配置策略的选取是一个带约束的优化问题,通常称为策略优化(policy optimization,简称 PO).

1.2 随机切换分析模型

考虑无线多媒体通信网中的基站 c,具有固定的带宽容量 B,其服务的呼叫业务种类根据带宽的需求不同分为 K 类,假定其中第 i (i=1,2,…,K)类呼叫业务的带宽可在 $B_i=\{b_{i1},b_{i2},\ldots,b_{ij},\ldots,b_{iN_i}\}$ 中取值,其中 N_i 表示 i 类业务可配置的不同带宽数.假设 i 类业务的新呼叫到达和切换到达服从 Poisson 分布,其到达率分别为 $\lambda_{nci},\lambda_{hci}$.第 i 类业务的呼叫持续时间(call holding time,简称 CHT)服从指数分布,均值为 $1/\mu_i$.此外,还假设呼叫在 c 的保持时间(cell residence time,简称 CRT)服从指数分布,均值为 $1/h$,与呼叫的种类无关.参数 h 的大小反映了切换率的高低.

呼叫在 c 中的信道占用时间为 CHT 与 CRT 较小者,两个随机变量的指数分布的最小值同样是指数分布,第 i 类业务的新呼叫和切换呼叫的信道占用时间满足均值为 $1/\mu_{nci},1/\mu_{hci}$ 指数分布,其中,$\mu_{nci}=\mu_{hci}=\mu_i+h$.

无线多媒体通信网中的带宽配置问题可以通过建立连续时间 Markov 切换模型来描述,构成如下:

设 m_{ij},n_{ij} 分别为系统中第 i 类的新呼叫和切换呼叫以带宽 b_{ij} 接受服务的个数.系统的状态可表示为 $s=(m_{ij},n_{ij},1\le i\le K,1\le j\le N_i)$,所有可能的状态取值构成状态空间:

$$S=\{(m_{ij},n_{ij},\ 1\le i\le K,\ 1\le j\le N_i):\sum_{i=1}^{K}\sum_{j=1}^{N_i}(m_{ij}+n_{ij})\cdot b_{ij}\le B\}.$$

定义呼叫到达或离开系统为事件,4 种类型的事件分别表示为:一个第 i 类新呼叫到达 e^{+nci} 或离开 e^{-nci},一个第 i 类切换呼叫到达 e^{+hci} 及离开 e^{-hci}.定义事件空间 $E=\{e^{+nci},e^{+hci},e^{-nci},e^{-hci},1\le i\le K\}$.

为表述简便,设 $s=(m_{ij},n_{ij},1\le i\le K,1\le j\le N_i),s\in S$,引入以下表示:

$$S^{+nci}=\{(m_{il}+1,m_{ik},n_{ij},l\in\{1,2,\ldots,N_i\},k=1,\ldots l-1,l+1,\ldots,N_i)\in S\},$$
$$S^{+hci}=\{(m_{ik},n_{il}+1,n_{ik},l\in\{1,2,\ldots,N_i\},k=1,\ldots l-1,l+1,\ldots,N_i)\in S\},$$
$$S^{-nci}=\{(m_{il}-1,m_{ik},n_{ij},l\in\{1,2,\ldots,N_i\},k=1,\ldots l-1,l+1,\ldots,N_i)\in S\},$$
$$S^{-hci}=\{(m_{ik},n_{il}-1,n_{ik},l\in\{1,2,\ldots,N_i\},k=1,\ldots l-1,l+1,\ldots,N_i)\in S\}.$$

当一个事件发生时,根据配置策略选取相应的行动.行动定义为 $d_s,s=(m_{ij},n_{ij},1\le i\le K,1\le j\le N_i)$,表示为第 i 类的新呼叫和切换呼叫分别配置 m_{ij} 数量和 n_{ij} 数量的 b_{ij} 带宽.行动在行动集 $D=\{d_s,s\in S\}$ 中取值.设 $\theta_{se}^{d_{s'}}=P(d_{s'}|s,e)$,$d_{s'}\in D,s\in S,e\in E$,表示当系统处于状态 s 时,事件 e 发生,选取行动 $d_{s'}$ 的概率.随机策略 $v:S\times E\to[0,1]^{|D|}$ 可以表示成参数化形式 $\theta=(\theta_{se}^{d_{s'}},d_{s'}\in D,s\in S,e\in E)$,$0\le\theta_{se}^{d_{s'}}\le 1$,$\sum_{d_{s'}\in D}\theta_{se}^{d_{s'}}=1$,相应地,随机 Markov 策略集 Π_{MS} 为

$$\Theta=\{(\theta_{se}^{d_{s'}},d_{s'}\in D,s\in S,e\in E):0\le\theta_{se}^{d_{s'}}\le 1,\ \sum_{d_{s'}\in D}\theta_{se}^{d_{s'}}=1\}.$$

系统在策略 θ 的作用下的运行规律可以用连续时间 Markov 过程(continuous-time Markov decision process,简称 CTMDP)$\{X_t:t\ge 0\}$ 来描述,其中,X_t 表示系统 t 时刻的状态.转移速率矩阵 $A(\theta)=[a_{ss'}(\theta)]$,其中

$$a_{ss'}(\theta)=\begin{cases}\lambda_{nci}\cdot\theta_{se^{+nci}}^{d_{s'}}, & s'\in S^{+nci},\\ \mu_{nci}\cdot\theta_{se^{-nci}}^{d_{s'}}, & s'\in S^{-nci},\\ \lambda_{hci}\cdot\theta_{se^{+hci}}^{d_{s'}}, & s'\in S^{+hci},\\ \mu_{hci}\cdot\theta_{se^{-hci}}^{d_{s'}}, & s'\in S^{-hci},\\ -\sum_{s''\ne s}a_{ss''}(\theta), & s'=s,\\ 0, & \text{otherwise},\end{cases}\quad s,s',s''\in S \tag{1}$$

对于任意的 $\theta\in\Theta$,转移速率矩阵 $A(\theta)$ 不可约,系统的平稳状态分布存在且唯一,记为 $p(\theta)=(p_s(\theta),s\in S)$,其中,$p_s(\theta)$ 表示稳态时处于状态 s 的概率.

设 r_{ij} 为系统提供带宽 b_{ij} 为第 i 类顾客服务时的收益率,则处于状态 s 时,系统的收益率为

$$f^r(s,\theta)=f_s^r(\theta)=\sum_{i=1}^{K}\sum_{j=1}^{N_i}(m_{ij}+n_{ij})\cdot r_{ij},\ s\in S.$$

系统处于状态 s 时,i 类呼叫的阻塞率和掉线率 $f_s^{bi}(\theta)$,$f_s^{di}(\theta)$ 分别为

$$f_s^{bi}(\theta)=1-\sum_{s'\in S^{+nci}}\theta_{se^{+nci}}^{d_{s'}},\ f_s^{di}(\theta)=1-\sum_{s'\in S^{+hci}}\theta_{se^{+hci}}^{d_{s'}}.$$

记 $f_s^b(\theta)=(f_s^{bi}(\theta),i=1,2,\ldots,K)$,$f_s^d(\theta)=(f_s^{di}(\theta),i=1,2,\ldots,K)$,定义系统的性能函数

$$f_s(\theta)=f_s^r(\theta)-\omega_1\cdot f_s^b(\theta)-\omega_2\cdot f_s^d(\theta),$$

其中:ω_1,ω_2 为 K 维向量,其分量的取值反映各类呼叫业务对阻塞率和掉线率 QoS 的不同需求.系统的性能测度定义为

$$\eta(\theta)=\lim_{T\to\infty}\frac{1}{T}E_\theta\left[\int_0^T f(X_t,\theta)\mathrm{d}t\right]=\sum_{s\in S}p_s(\theta)\cdot f_s(\theta).$$

第 i 类呼叫业务处于状态 s 时配置的平均带宽为

$$ab_s^i(\theta)=\sum_{1\le j\le N_i}(m_{ij}+n_{ij})\cdot b_{ij}\Big/\sum_{1\le j\le N_i}(m_{ij}+n_{ij}).$$

第 i 类呼叫业务的平均配置带宽为

$$AB_i(\theta)=\sum_{s\in S}p_s(\theta)\cdot ab_s^i(\theta),\ i=1,2,\ldots,K.$$

由于在时间上稳态分布等于在空间上的稳态分布,故 $AB_i(\theta)$,即为第 i 类顾客在其呼叫持续期间获得的平均配置带宽.记 $AB(\theta)=(AB_i(\theta),i=1,2,\ldots,K)$,平均配置带宽 QoS 要求 $AB(\theta)\ge G$,G 为 K 维常数向量,其分量为对应种类的呼叫业务的平均配置带宽 QoS 要求.这里,$AB(\theta)\ge G$ 表示 $AB_i(\theta)\ge G_i,1\le i\le K$.

综合上述,适应带宽配置的 CTMDP 模型:

$$\{S,D,A(\theta),(\eta(\theta),AB(\theta))\}.$$

在此模型下,适应带宽配置问题为寻找一个最优策略θ^*,在满足平均带宽 QoS 要求的条件下,使得 NCBP 和 HCDP 最小,并同时使得网络的收益最大化.

$$\text{PO}: \max_{\theta\in\Theta}\eta(\theta),$$
$$\text{s.t. } AB(\theta)\geq G.$$

2 在线优化算法

适应带宽配置通过建立随机切换模型,转化为一个带约束的 CTMDP 策略优化问题.无线多媒体通信网适应带宽配置应用环境复杂,系统参数难以预先获取,且具有时变性,这使得动态规划等进行数值计算的优化方法的应用具有一定的局限性,而在线优化的方法具有较高的实际应用价值.

在线优化是指在系统运行过程中不断进行策略改进,随机逼近最优策略,使系统的性能趋于最优.策略改进基于性能灵敏度进行,性能灵敏度有两种表达形式:适用于连续参数空间(如随机策略)的性能关于参数的梯度和适用于离散参数空间(如确定型策略)的对应于两个不同参数的性能之差.性能势[10]是构造性能梯度和性能差的基础,与相对代价向量(relative cost vector)[11]、偏差(bias)[12]之间相差一个常数.性能势能够通过一条样本轨道来估计,强化学习(reinforcement learning,简称 RL)、Q-学习、TD(λ)、神经元动态规划等,是基于样本轨道估计性能势或其他相关量(如 Q-因子)的有效方法.基于性能势的性能梯度公式和性能差公式构成在线优化的基础[13].从性能梯度公式和性能差公式出发,通过性能势的在线估计,有两类方法来实现在线优化:其一是基于性能梯度公式,结合随机逼近算法,沿梯度方向改进策略,使系统性能趋于最优,如文献[9]的在线梯度估计方法、文献[14]的 PA 方法、文献[15]的相似率方法等.该类方法能够优化随机策略,但在通常情况下只能达到局部最优;其二是根据性能差公式,在线比较当前策略与其他策略的性能,通过不断选取优于当前的策略,进行策略迭代,随机逼近最优策略,文献[16]的基于势的在线策略迭代方法属于此类方法.该类方法的优点是收敛速度快,能够保证达到全局最优,局限性为只适用于确定型策略的优化,且需满足状态与行动不相关的前提假设.基于性能势的性能灵敏度公式结合随机逼近算法,为多个应用领域中的在线优化提供了一致化的方法.

适应带宽配置问题是一个带约束的优化问题,其最优策略属于随机策略,适用于上述第 1 类的在线优化方法.在建立随机切换模型的基础上,接下来应用强化学习的方法,基于系统实际运行这一样本轨道,估计性能测度关于参数化策略的梯度,结合随机逼近进行策略改进,实现在线优化.文献[9]中给出了离散时间情况下的在线梯度估计与随机逼近算法,这里推广到连续时间情况,以基于性能势的性能梯度公式为基础,推导在线学习进行梯度估计式,并结合切换模型的结构特点,从自适应性、实时性、有效性 3 方面对算法加以改进,以满足在线优化的要求.

2.1 在线学习估计策略梯度

考虑连续时间 Markov 过程 $X=\{X_t,t\geq 0\}$,状态空间 $S=\{1,2,\ldots,N\}$有限,转移速率矩阵 A 是参数向量$\theta\in\Theta\subset R^K$的函数,即 $A(\theta)=\lfloor a_{ij}(\theta)\rfloor,i,j\in S$,其中的元素 $a_{ij}(\theta),i,j\in S$ 有界、关于θ一阶导数有界及二次可微,且$\forall\theta\in\Theta$,$A(\theta)$不可约.X 的稳态分布 $p(\theta)=(p_1(\theta),p_2(\theta),\ldots,p_N(\theta))$满足平衡方程:

$$p(\theta)e=1,\ p(\theta)A(\theta)=0,\ A(\theta)e=0 \tag{2}$$

其中:$e=(1,1,\ldots,1)^\tau$,上标“τ”表示转置.设性能函数 $f:S\times\Theta\to R$(实数)有界,表示成向量形式为 $f(\theta)=(f_1(\theta),f_2(\theta),\ldots,f_N(\theta))^\tau$.平均性能测度为

$$\eta(\theta)=\lim_{T\to\infty}\frac{1}{T}E_\theta\left[\int_0^T f(X_t,\theta)\mathrm{d}t\right]=\sum_{i=1}^N p_i(\theta)f_i(\theta).$$

定义

$$\left.\begin{aligned} g_i(\theta) &= E_\theta\left[\int_0^{T^{\{i\}}\{i^*\}} (f(X_t^{\{i\}},\theta)-\eta(\theta))\mathrm{d}t\right] \\ g_{i^*}(\theta) &= 0 \end{aligned}\right\} \tag{3}$$

为相应状态的性能势[10],$g(\theta)=(g_i(\theta),1\le i\le N)^\tau$为性能势向量.式中 $X_t^{\{i\}}=\{X_t: X_0=i, t\ge 0\}$ 表示初始状态为 i 的 Markov 过程, $T^{\{i\}}\{i^*\}=\min\{t>0: X_t^{\{i\}}=i^*\}$ 为从初始状态 i 出发首次到达 i^*的时间.容易验证 $g(\theta)$满足 Poisson 方程[10]:

$$A(\theta)g(\theta)=-f+\eta(\theta)e \tag{4}$$

对 Poisson 方程(4)两边关于θ求导,并注意到平衡方程(2),可得到性能关于策略参数θ的梯度公式:

$$\nabla\eta(\theta)=\sum_{i\in S}p_i(\theta)\left(\sum_{j\in S}\nabla a_{ij}(\theta)\cdot g_j(\theta)+\nabla f_i(\theta)\right) \tag{5}$$

考虑到 $a_{ii}(\theta)=-\sum_{j\ne i,j\in S}a_{ij}(\theta), i\in S$ 及 $a_{ij}(\theta)\ge 0, i\ne j, i,j\in S$,式(5)可改写成

$$\begin{aligned} \nabla\eta(\theta) &= \sum_{i\in S}p_i(\theta)\left(\sum_{j\in S_i}\nabla a_{ij}(\theta)\cdot(g_j(\theta)-g_i(\theta))+\nabla f_i(\theta)\right) \\ &= \sum_{i\in S}p_i(\theta)\left(\sum_{j\in S_i}a_{ij}(\theta)\frac{\nabla a_{ij}(\theta)}{a_{ij}(\theta)}(g_j(\theta)-g_i(\theta))+\nabla f_i(\theta)\right) \\ &= \sum_{i\in S}p_i(\theta)\left(\sum_{j\in S_i}a_{ij}(\theta)\cdot L_{ij}(\theta)\cdot(g_j(\theta)-g_i(\theta))+\nabla f_i(\theta)\right) \end{aligned} \tag{6}$$

式中,$S_i=\{j: a_{ij}(\theta)>0, j\in S\}$,$L_{ij}(\theta)=\nabla a_{ij}(\theta)/a_{ij}(\theta)$.

设$\{X_t, t\ge 0\}$为 X 的一条样本轨道,$i^*\in S$ 是一个常返状态,t_m 是第 m 次抵达 i^*的时间.因$\{X_t, t\ge 0\}$在每一时刻 t_m 后的延续在统计意义上以概率 1 等价于从 $t=0$ 开始的过程,称 t_m 为再生时刻,i^*为再生状态,$\{X_t, t\ge 0\}$为一个再生过程,$\{X_t, t_m\le t<t_{m+1}\}$为第 m 个再生周期,周期的长度为 $T_m=t_{m+1}-t_m$. t_m^n 表示在此周期中第 n 次状态转移发生的时刻,两次状态转移的时间间隔为 $T_m^n=t_m^{n+1}-t_m^n$,第 m 个再生周期中发生的状态转移次数用 n_m 表示.对于固定的θ, $\{X_t, t\ge 0\}$在每一再生周期中独立同分布,随机变量 T_m 独立同分布,具有有限均值 $E_\theta[T]$.

根据式(3),可以用式(7)来估计 $g_i(\theta)$.显然, $\hat{g}_{i^*}(\theta)=\hat{g}_{t_m}(\theta)=0$,不需要进行估计.

$$\left.\begin{aligned} \hat{g}_{t_m^n}(\theta) &= \int_{t_m^n}^{t_{m+1}}(f(X_t,\theta)-\hat{\eta}(\theta))\mathrm{d}t=\sum_{k=n}^{n_m}\left(f_{X_{t_m^k}}(\theta)-\hat{\eta}(\theta)\right)\cdot T_m^k \\ \hat{\eta}(\theta) &= \frac{\mathbf{1}}{T_m}\int_{t_m}^{t_{m+1}}f(X_t,\theta)\mathrm{d}t=\frac{\mathbf{1}}{T_m}\sum_{n=\mathbf{1}}^{n_m}f_{X_{t_m^n}}(\theta)\cdot T_m^n \end{aligned}\right\} \tag{7}$$

由式(6)和式(7),可以在第 m 个再生周期得到$\nabla\eta(\theta)$的一个估计:

$$\begin{aligned} \hat{\nabla\eta}_m(\theta) &= \sum_{n=1}^{n_m}(\hat{g}_{X_{t_m^{n+1}}}(\theta)-\hat{g}_{X_{t_m^n}}(\theta))\cdot L_{X_{t_m^n}X_{t_m^{n+1}}}(\theta)+\sum_{n=1}^{n_m}\nabla f_{X_{t_m^n}}(\theta)\cdot T_m^n \\ &= \sum_{n=1}^{n_m}(\hat{\eta}(\theta)-f_{X_{t_m^n}}(\theta))\cdot T_m^n\cdot L_{X_{t_m^n}X_{t_m^{n+1}}}(\theta)+\sum_{n=1}^{n_m}\nabla f_{X_{t_m^n}}(\theta)\cdot T_m^n \end{aligned} \tag{8}$$

由于在不同的再生周期中,随机变量 $\hat{\nabla\eta}_m(\theta)$ 是独立同分布的,可以证明 $\hat{\nabla\eta}_m(\theta)$ 是$\nabla\eta(\theta)$的一个误差有界的无偏估计值.

2.2 随机逼近最优策略

在适应带宽配置的随机切换模型中,性能测度$\eta(\theta)$和平均带宽 $AB(\theta)$关于策略参数θ连续可导.由式(1)、式(6),有

$$\frac{\partial\eta(\theta)}{\partial\theta_{se}^{d_{s'}}}=p_s(\theta)\cdot\left(\frac{\partial a_{ss'}(\theta)}{\partial\theta_{se}^{d_{s'}}}\cdot(g_{s'}(\theta)-g_s(\theta))+\omega_e\right)=p_s(\theta)\cdot(a_{ss'}\cdot(g_{s'}(\theta)-g_s(\theta))+\omega_e),$$

其中,$p_s(\theta)>0,\omega_e>0,a_{ss'}\geq 0$,当且仅当 $g_{s'}(\theta)-g_s(\theta)=-\omega_e/a_{ss'}<0,\partial\eta(\theta)\big/\partial\theta_{se}^{d_{s'}}=0$.而 $g_{s'}(\theta)-g_s(\theta)<0,g_s(\theta)-g_{s'}(\theta)<0$,$s,s'\in S$ 不可能同时成立,故$\forall\theta\in\Theta,\nabla\eta(\theta)\neq 0$.同样地,有$\forall\theta\in\Theta,\nabla AB(\theta)\neq 0$.且有$\eta(\theta),AB(\theta)$关于策略参数$\theta$连续,满足 PO 的策略参数集是连续有界的连通域.这保证了运用梯度方法能够达到全局最优.

系统的实际运行,提供了一条样本轨道,基于这一样本轨道的第 m 个再生周期,由式(8)通过在线学习,得到性能测度$\eta(\theta)$和平均带宽 $AB(\theta)$关于θ的带有随机误差的无偏梯度估计 $\hat{\nabla}\eta(\theta_m),\hat{\nabla}AB(\theta_m)$,进而结合 RM(robbins-monro)随机逼近算法,即可在下一个再生周期的开始时刻进行策略改进.这里采用如下形式的 RM 算法:

$$\theta_{m+1}=\theta_m+\gamma_m\cdot H_m\cdot\hat{\nabla}'_m(\theta_m),$$

其中,γ_m是一个正的步长序列,这里取$\gamma_m=1/m$,以满足 RM 算法收敛对步长序列的要求,即

$$\sum_{m=1}^{\infty}\gamma_m=\infty,\ \sum_{m=1}^{\infty}\gamma_m^2<\infty.$$

H_m 为 K 阶正定矩阵,$\|H_m\|<\infty,\theta_m+\gamma_m\cdot H_m\cdot\hat{\nabla}'_m(\theta_m)\in\Theta$,即将 θ_{m+1} 约束在 $\Theta=\{(\theta_{se}^{d_{s'}},d_{s'}\in D,s\in S,e\in E):0\leq\theta_{se}^{d_{s'}}\leq 1,\sum_{d_{s'}\in D}\theta_{se}^{d_{s'}}=1\}$ 上取值.

$$\hat{\nabla}'_m(\theta_m)=\begin{cases}\hat{\nabla}\eta_m(\theta_m), & \theta_m\in\Theta^{G+}=\{\theta:AB(\theta)\geq G+\varepsilon,\theta\in\Theta\},\\ \hat{\nabla}\eta_m(\theta_m)-\hat{\nabla}AB_m(\theta_m)(\hat{\nabla}\eta_m(\theta_m))^{\tau}\hat{\nabla}AB_m(\theta_m), & \theta_m\in\Theta^{G+\varepsilon}=\{\theta:G\leq AB(\theta)<G+\varepsilon,\theta\in\Theta\},\\ \hat{\nabla}AB_m(\theta_m), & \theta_m\in\Theta^{G-}=\{\theta:AB(\theta)<G,\theta\in\Theta\}.\end{cases}$$

设 $\theta^*=\max_{\theta\in\Theta^{G+}\cup\Theta^{G+\varepsilon}}\eta(\theta)$,在$\Theta^{G+}\cup\Theta^{G+\varepsilon}$和$\Theta^{G-}$中分别选取 $V_1(\theta)=\eta(\theta^*)-\eta(\theta),V_2(\theta)=\max\{AB(\theta),\theta\in\Theta\}-AB(\theta)$作为 Lypunov 函数,可以证明给定初始策略$\forall\theta_0\in\Theta$,由上述算法产生的随机序列$\{\theta_m,m\geq 0\}$,以概率 1 收敛到最优策略,即$\theta_m\to\theta^*,m\to\infty,w.p.1$.

基于系统实际运行这一样本轨道,在线观察 $t_m^n,X_{t_m^n}$,根据式(7)计算 $\hat{\eta}_m(\theta_m)$,运用式(8)计算梯度估计值:

$$\hat{\nabla}\eta_m(\theta_m)=\sum_{n=1}^{n_m}(\hat{\eta}_m(\theta_n)-f_{X_{t_m^n}}(\theta_m))\cdot T_m^n\cdot L_{X_{t_m^n}X_{t_m^{n+1}}}(\theta_m)+\sum_{n=1}^{n_m}\nabla f_{X_{t_m^n}}(\theta_m)\cdot T_m^n,$$

其中的 $L_{X_{t_m^n}X_{t_m^{n+1}}}(\theta_m)$,当 $X_{t_m^n}=s,X_{t_m^{n+1}}=s'$时,其分量$ss'$为$1\big/\theta_{se}^{d_{s'}}$,其余分量为 0.可见上述算法并不依赖系统参数的信息,如呼叫到达率、持续时间等,只要系统状态可观、性能函数已知、QoS 指标给定,即可进行在线优化.在线优化算法的流程如图 1 所示.

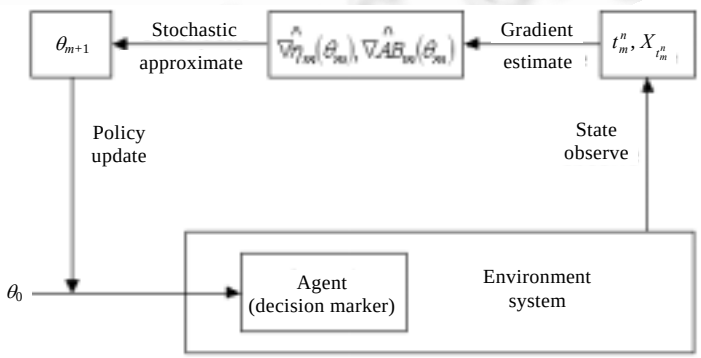

Fig.1 The flowchart of online optimization algorithm

图 1 在线优化算法流程图

3 数值仿真

通过数值仿真对算法的有效性进行验证,包括优化策略的有效性、算法的收敛性和对应用环境的适应性.仿真参数的选取应考虑到在更好地模拟并显示系统过载时适应带宽机制特性的同时,尽可能地简化仿真复杂度.假设小区容量 B=5,有 2 种业务类型,分别具有 2 个层级的适应带宽,呼叫业务的到达率服从 Poisson 分布,其呼叫持续时间与小区停留时间服从指数分布,具体参数见表 1.

Table 1 Simulation parameters

表 **1** 仿真参数

Class−*i*	b_{ij}	f	μ_i	h	G_i
1	5 3	5 4	1.1	0.4	3.7
2	5 2	5 3	0.6	0.4	3.3

假设在到达小区的呼叫中,业务类型 1 和类型 2 各占 50%,其中有 40%为越区切换呼叫.图 2~图 4 给出本文的 O-ABA(online adaptive bandwidth allocation)算法与文献[8]的 Q-ABA(q-learning based adaptive bandwidth allocation)算法应用于不同环境(呼叫到达率)中的网络收益及 QoS 保证情况.从图 2 可以看出,上述两种算法都能确保平均带宽大于要求值,而 O-ABA 比 Q-ABA 具有更高的网络收益.图 3 与图 4 显示两种业务类型的阻塞率与掉线率,可以看出,O-ABA 与 Q-ABA 相比,能够获得更小且较为平稳的阻塞率与掉线率.从图 2~图 4 中也明显可以看出,Q-ABA 由于采用确定性策略来近似最优随机策略,导致在不同的应用环境中的各项性能具有较为显著的波动性,在某些应用环境下能够较好地逼近最优策略,而在其他环境中则有较为显著的差异.

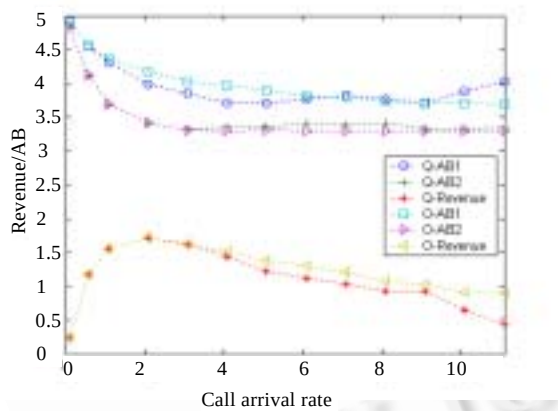

Fig.2 Revenue and AB

图 2 网络收益与平均带宽

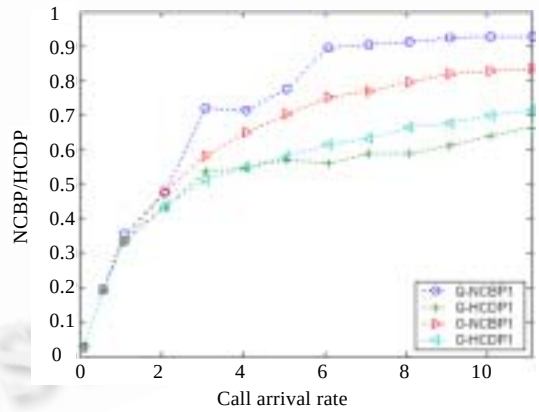

Fig.3 NCBP and HCDP of class 1 calls

图 3 业务类型 1 的阻塞率与掉线率

图 5、图 6 显示 O-ABA 算法在平稳环境中的收敛情况.算法在不同的应用环境中具有较好的收敛性,表现在收敛速度与精度两方面.

图 7 显示 O-ABA 算法在非平稳环境中的适应性.在环境状态发生较为显著的变化时,在线优化算法能够快速响应环境的变化并收敛到最优值.算法对环境显著变化的响应通过滑动窗口法进行检测,并相应调整步长来实现,而对环境状态随时间的微小变化同样具有较强的适应性.

4 结 论

通过建立 Markov 切换模型,将无线多媒体通信网的适应带宽配置问题转化为带约束的策略优化问题,提出一种在线学习估计策略梯度、随机逼近优化带宽配置策略的在线算法.上述在线适应带宽配置优化算法不依赖

于系统的具体参数,如呼叫到达率、呼叫保持时间、小区停留时间等,具有较强的适应性;计算量小,满足实时性的要求;在多种 QoS 约束下,直接优化随机策略,并保证收敛到全局最优,具有较好的优化效果.数值仿真结果进一步验证了上述特性.该算法适用于复杂应用环境中的无线多媒体通信网适应带宽配置的在线优化,具有较高的应用价值.

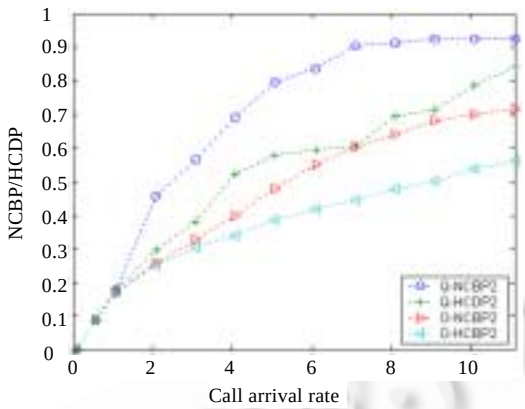

Fig.4 NCBP and HCDP of class 2 calls

图 4 业务类型 2 的阻塞率与掉线率

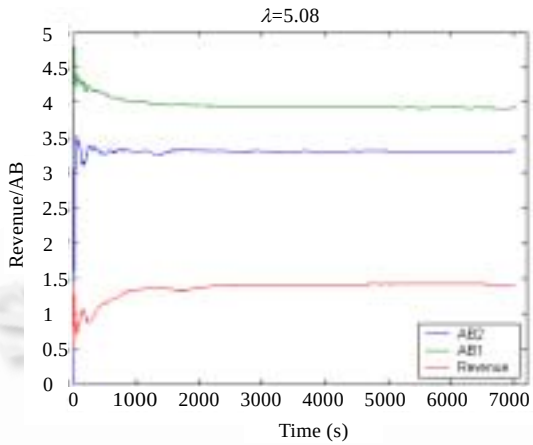

Fig.5 Convergence of the algorithm in stationary environment (λ=5.08)

图 5 算法在平稳环境中的收敛过程(λ=5.08)

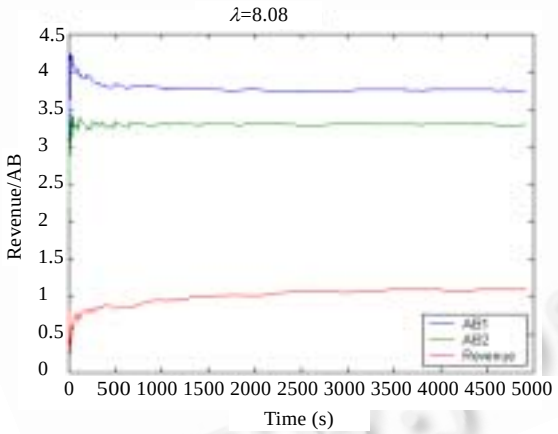

Fig.6 Convergence of the algorithm in stationary environment (λ=8.08)

图 6 算法在平稳环境中的收敛过程(λ=8.08)

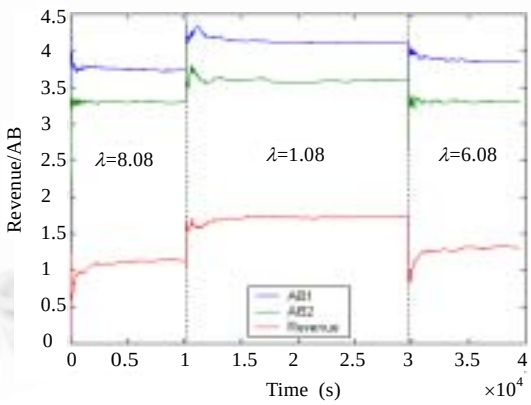

Fig.7 Adaptation of the algorithm in a nonstationary environment

图 7 算法在非平稳环境中的自适应过程

References:

[1] Kwon T, Choi Y, Das S. Bandwidth adaptation algorithms for adaptive multimedia services in mobile cellular networks. Kluwer Wireless Personal Communications, 2002,22(3):337−357.

[2] Xiao Y, Chen C, Wang Y. Fair bandwidth allocation for multi-class of adaptive multimedia services in wireless/mobile networks. In: Proc. of the IEEE 53rd Vehicular Technology Conf. Piscataway: IEEE Press, 2001. 2081−2085.

[3] Chou CT, Shin KG. Analysis of adaptive bandwidth allocation in wireless networks with multilevel degradable quality of service. IEEE Trans. on Mobile Computing, 2004,3(1):5−17.

[4] Nasser N, Hassanein H. Connection-Level performance analysis for adaptive bandwidth allocation in multimedia wireless cellular networks. In: Hassanein H, Oliver RL, Richard GG, Wilson LF, eds. Proc. of the IEEE Int'l Conf. on Performance, Computing, and Communications. Piscataway: IEEE Press, 2004. 61−68.

[5] Wu Y, Bi GG. A measurement based dynamic call admission control scheme in wireless multimedia communication networks. Chinese Journal of Computers, 2005,28(11):1823−1830 (in Chinese with English abstract).

[6] Tang SS, Li W, Kim J. Modeling adaptive bandwidth allocation scheme for multi-service wireless cellular networks. In: Pierre S, Conan J, eds. Proc. of the IEEE Int'l Conf. on Wireless and Mobile Computing, Networking and Communications. Piscataway: IEEE Press, 2005. 189−195.

[7] Jiang AQ, Ye XG, Wu JG. Bandwidth adaptation scheme using genetic algorithm in wireless/mobile networks. Computer Research and Development, 2004,41(9):1453−1459 (in Chinese with English abstract).

[8] Yu F, Wong VWS, Leung VCM. Efficient QoS provisioning for adaptive multimedia in mobile communication networks by reinforcement learning. Mobile Networks and Applications, 2006,11(1):101−110.

[9] Marbach P, Tsitsiklis JN. Simulation-Based optimization of Markov reward processes. IEEE Trans. on Automatic Control, 2001, 46(2):191−209.

[10] Cao XR, Chen HF. Perturbation realization, potentials and sensitivity analysis of Markov processes. IEEE Trans. on Automation Control, 1997,42(10):1382−1393.

[11] Bertsekas DP. Dynamic Programming and Optimal Control. 2nd ed., Belmont: Athena Scientific, 2001.

[12] Puterman ML. Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: John Wiley & Sons, 1994.

[13] Cao XR. The potential structure of sample paths and performance sensitivities of Markov systems. IEEE Trans. on Automatic Control, 2004,49(12):2129−2142.

[14] Chong EKP, Ramadge PJ. Stochastic optimization of regenerative systems using infinitesimal perturbation analysis. IEEE Trans. on Automatic Control, 1994,39(7):1400−1410.

[15] Glynn PW. Likelihood ratio gradient estimation: An overview. In: Thesen A, Grant H, Kelton WD, eds. Proc. of the 19th Winter Simulation Conf. New York: ACM Press, 1987. 90−105.

[16] Fang HT, Cao XR. Potential-Based on-line policy iteration algorithms for Markov decision processes. IEEE Trans. on Automatic Control, 2004,49(4):493−505.

附中文参考文献:

[5] 吴越,毕国光.无线多媒体网络中一种基于测量网络状态的动态呼叫接纳控制算法.计算机学报,2005,28(11):1823−1830.

[7] 姜爱全,叶晓国,吴家皋.无线/移动网络中基于遗传算法的带宽适应方案.计算机研究与发展,2004,41(9):1453−1459.

江琦(1967－),男,安徽歙县人,博士生,主要研究领域为信息通信网络的性能优化.

殷保群(1962－),男,博士,教授,博士生导师,主要研究领域为随机离散事件动态系统性能优化及应用.

奚宏生(1950－),男,教授,博士生导师,主要研究领域为信息通信网络的性能分析与优化.