基于半马尔科夫决策过程的虚拟传感网络资源分配策略

王汝言[①②③] 李宏娟*[①②③] 吴大鹏[①②③] 李红霞[④]

[①](重庆邮电大学通信与信息工程学院 重庆 400065)

[②](重庆高校市级光通信与网络重点实验室 重庆 400065)

[③](泛在感知与互联重庆市重点实验室 重庆 400065)

[④](中国联合网络通信有限公司重庆市分公司 重庆 401123)

摘 要：针对传统无线传感网络(WSN)中资源部署与特定任务的耦合关系密切，造成较低的资源利用率，进而给资源提供者带来较低的收益问题，根据虚拟传感网络请求(VSNR)的动态变化情况，该文提出虚拟传感网络(VSN)中基于半马尔科夫决策过程(SMDP)的资源分配策略。定义VSN的状态集、行为集、状态转移概率，考虑传感网能量受限以及完成VSNR的时间，给出奖赏函数的表达式，并使用免模型强化学习算法求解特定状态下的行为，从而最大化网络资源提供者的长期收益。数值结果表明，该文的资源分配策略能有效提高传感网资源提供者的收益。

关键词：虚拟传感网络；资源分配；半马尔科夫决策过程

中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2019)12-3014-08

DOI: 10.11999/JEIT190016

Semi-Markov Decision Process-based Resource Allocation Strategy for Virtual Sensor Network

WANG Ruyan[①②③] LI Hongjuan[①②③] WU Dapeng[①②③] LI Hongxia[④]

[①](*School of Communication and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

[②](*Optical Communication and Network Key Laboratory of Chongqing*, *Chongqing* 400065, *China*)

[③](*Key Laboratory of Ubiquitous Sensing and Networking in Chongqing*, *Chongqing* 400065, *China*)

[④](*Chongqing Branch*, *China Unicom*, *Chongqing* 401123, *China*)

Abstract: The close relationship between resource deployment and specific tasks in traditional Wireless Sensor Network(WSN) leads to low resource utilization and revenue. According to the dynamic changes of Virtual Sensor Network Request(VSNR), the resource allocation strategy based on Semi-Markov Decision Process(SMDP) is proposed in Virtual Sensor Network(VSN). Then, difining the state, action, and transition probability of the VSN, the expected reward is given by considering the energy and time to complete the VSNR, and the model-free reinforcement learning approach is used to maximize the long-term reward of the network resource provider. The numerical results show that the resource allocation strategy of this paper can effectively improve the revenue of the sensor network resource providers.

Key words: Virtual Sensor Network(VSN); Resource allocation; Semi-Markov Decision Process(SMDP)

1 引言

无线传感网络(Wireless Sensor Network, WSN)由部署在监测区域内的大量传感节点，通过无线通信方式形成的自组织网络，实现数据的采集、处理和传输，其被广泛应用于军事、医疗、家居、环境、工业等领域[1,2]。WSN通常面向特定领域和既定任务，单个用户对应单个WSN完成单个任务，已部署的物理资源即使闲置也无法重用于其它任务和其他用户，新的任务每次都需要重新部署资源，导致较低的资源利用率。虚拟化技术可以整合传感网的物理资源，并将其抽象成虚拟资源[3,4]，根据当前网络状态，为不同用户所对应的不同任务

收稿日期：2019-01-07；改回日期：2019-04-16；网络出版：2019-05-22
*通信作者： 李宏娟 ilihj@foxmail.com
基金项目：国家自然科学基金(61871062，61771082)，重庆市高校创新团队建设计划资助项目(CXTDX201601020)

Foundation Items: The National Natural Science Foundation of China (61871062, 61771082), The Chongqing Funded Project of Chongqing University Innovation Team Construction (CXTDX201601020)

按需分配物理资源，将传统传感网物理资源的专有模式转化为共有模式，即多个用户复用WSN物理资源。显然，此种方式这不仅能满足用户的多样化需求，还能通过共享这些物理资源节省WSN资源提供者的开销。

虚拟传感网络(Virtual Sensor Network, VSN)将WSN解耦为无线传感网络基础设施提供者(Wireless Sensor Network Infrastructure Provider, WSNInP)和虚拟传感网络服务提供者(Virtual Sensor Network Service Provider, VSNSP)，其中WSNInP负责部署、管理和维护WSN，而VSNSP向WSNInP租赁传感资源并根据用户需求创建虚拟传感网络请求(Virtual Sensor Network Request, VSNR)。每个VSNR由虚拟传感节点和虚拟链路组成，需要将其分别映射至底层的物理传感节点和物理链路才能为用户提供服务。

本文主要针对多任务驱动的VSN应用场景，使得VSN能同时为多个用户的多个任务提供服务，已部署的传感网资源不再为单个任务所特有，例如森林中的传感器不仅可以监测温度的变化，还可以同时记录动物的行踪，相同的传感网资源可为森林管理者提供不同方面的数据需求。虚拟化使得多个异构的传感器节点可同时存在，而本文在此基础上主要研究如何对新到来的多个任务进行资源分配，在满足用户需求的同时，最大化资源提供者的长期收益。

针对传感网中的资源分配问题，国内外研究人员进行了广泛深入的研究。文献[5]提出了一种应用于无线多媒体传感网的资源分配机制，引入穷举搜索方法获得带有实时性业务需求的最优调度序列，但是该算法缺乏对能量和收益问题的分析。文献[6]提出了软件定义WSN实现资源的集中分配，通过节点的快速重建实现资源的动态分配，在满足服务质量的个体约束下制定了优化问题来最小化能量消耗，但是该方案并未考虑节点重建过程中的资源开销。文献[7]提出了云辅助的软件定义WSN资源分配方案，通过动态联盟合作博弈模型共同完成用户的资源需求且收益共享，以最小化能耗为目标函数并最大化提供商的收益，但是该方案并未给出最优问题的具体解决方法。文献[8]提出一种VSN中的最优资源分配机制，首先建立基于混合整数线性规划的优化框架，旨在最大化共享物理资源的并发任务数量, 但是该机制并没有区分不同重要程度任务的收益差别。

本文提出了基于半马尔科夫决策过程(Semi-Markov Decision Process, SMDP)的VSNR资源分配策略，定义了VSN的状态集、行为集、状态转移概率分布情况，考虑资源分配过程中的能量和服务时间，并将其作为收益的衡量指标对其进行优化，给出最优策略的解决方法，确定特定状态下的行为，从而最大化WSNInP的长期收益。

2　系统模型

如前所述，在同一时刻或者一段时间内，网络中存在多个VSNRs，而VSNRs到达或离开的时间间隔服从任意分布，且下一次对VSNRs的决策仅与当前的决策状态和网络的资源状况有关，与历史的VSNRs决策记录无关，因此本文将接受或拒绝这些带有特定资源量大小的VSNRs的过程建模为基于SMDP的资源分配模型。SMDP突出的是一个决策过程而不是自然过程，系统的进化取决于决策点所做的决策，即在决策点决定采取哪种行为来应对网络中的资源及运行状态的变化。

设网络中的VSNRs集合为$\boldsymbol{VSNRs}=\{\mathrm{VSNR}_1, \mathrm{VSNR}_2, \cdots, \mathrm{VSNR}_m\}$，其对应的资源需求集合为$\boldsymbol{RT}=\{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_m\}$。对于第$i$个VSNR，资源需求表示为$\boldsymbol{R}_i=\{\boldsymbol{R}_{i1}, \boldsymbol{R}_{i2}, \cdots, \boldsymbol{R}_{in}\}$，其具体的资源需求主要包括节点和链路的资源需求，传感网能量受限，且单个传感节点的计算存储能力有限[9]，因此本文主要考虑节点的能量和缓存资源，而链路资源主要考虑数据传输速率即带宽资源。具体地，对于其中第j个虚拟传感节点的资源需求用3元组表示为$\boldsymbol{R}_{ij}=\{R_{ij}(1), R_{ij}(2), R_{ij}(3)\}$，其中(1)代表能量，节点所需的能量主要与数据包大小以及节点间的距离有关，(2)代表缓存资源，(3)代表带宽资源即节点间的资源。从网络中不同的资源种类出发，设所有VSNRs的资源需求集合为$\boldsymbol{RT}=\{\mathrm{RT(Energy)}, \mathrm{RT(Cache)}, \mathrm{RT(Band)}\}$，则第 i个VSNR的资源需求为

$$\boldsymbol{R}_i=\sum_{j=1}^{n}\sum_{k=1}^{3}\boldsymbol{R}_{ij}(k) \tag{1}$$

其中，n表示所有VSNRs中的最大虚拟传感节点数量，若某一VSNR中的节点数小于n，则超过其节点数量的资源需求值为0。若两虚拟传感节点对之间无链路，则其带宽资源需求值为0。根据以上矩阵，所有VSNRs的资源需求可表示为

$$\boldsymbol{RT}=\sum_{i=1}^{m}\boldsymbol{R}_i=\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{3}\boldsymbol{R}_{ij}(k) \tag{2}$$

VSNR对应的资源需求必须成功映射到底层的物理传感网络才能为用户提供服务，设底层物理WSN的资源集合为$\mathbf{PHY}=\{\mathrm{PHY(Energy)}, \mathrm{PHY(Cache)}, \mathrm{PHY(Band)}\}$，在映射过程中需满足的基

本约束条件为：被映射的物理传感节点的能量、缓存以及带宽资源不小于相应的虚拟传感节点的资源需求总和，并且从整体上来看所接受的VSNRs的资源需求总和始终小于WSN的资源总和，具体表示为

$$\left.\begin{array}{l}\sum R_{ij}(k)\le \mathrm{Map}(R_{ij}(k)),\\ \forall i\in[1,m],j\in[1,n],k=\{1,2,3\}\\ \mathrm{RT(Energy)}\le \mathrm{PHY(Energy)}\\ \mathrm{RT(Cache)}\le \mathrm{PHY(Cache)}\\ \mathrm{RT(Band)}\le \mathrm{PHY(Band)}\end{array}\right\} \tag{3}$$

其中，$\mathrm{Map}(R_{ij}(k)),k=\{1,2,3\}$表示将第$i$个VSNR中的第$j$个节点映射至物理传感节点后，该物理传感节点的资源大小，以上约束条件表明虚拟资源请求量小于物理资源量大小是映射成功的基础。

某一时刻，网络中会存在带有具体资源需求的VSNRs到达或者离开，设其到达或离开的速率服从泊松分布，大小分别为λ_{p}和μ_{p}，那么当新的VSNRs到来时，网络选择接受或者拒绝这些VSNRs需根据网络的整体资源情况、运行状况及可能带来的收益等因素综合确定。

3 资源分配策略

马尔科夫决策过程(Markov Decision Process, MDP)有3种类型，包括离散时间MDP、连续时间MDP以及SMDP。其中离散时间MDP的时间和状态都是离散的，状态转移依赖1步转移矩阵；连续时间MDP与离散时间MDP的不同之处在于其状态发生变化的时刻是任意时刻且为连续值，并且相邻状态的停留时间服从指数分布；SMDP是特殊的连续时间MDP，相邻状态的停留时间服从随机分布，并不是指数分布。

本文所研究的资源分配并为用户提供服务的过程中，VSNRs到达和离开的时间点是任意时刻且时间间隔是不固定的，在VSNRs到达后需尽快对其做出决策，因此对于VSNRs的动态变化采用随机分布更符合实际情况，所以本文采用SMDP模型。

3.1 基于SMDP的资源分配策略

(1)状态集：在VSNSP创建VSNRs向WSNInP请求资源并完成映射的资源分配过程中，状态集S包括不同VSNRs的资源需求情况、当前WSN的可用资源情况以及可能发生的事件，其中事件包括VSNR的到来和离开，因此状态集S的表达式为

$$\boldsymbol{S}=\{s|s=(R_1,R_2,\cdots,R_m,M,e)\} \tag{4}$$

根据上节的描述可知，$R_1,R_2,\cdots,R_m$分别为m个VSNR的资源需求量，若K为WSN中的最大资源量，M为WSN中可用的总资源量，则$\sum_{x=1}^{m}R_x\le M$。其中$e\in E=\{A_{R_a},D_{R_i}\}$，$A_{R_a}$表示资源需求量为$R_a$的VSNR的到达，$D_{R_i}$表示资源需求量为$R_i$的VSNR的离开。

(2)行为集：根据状态集，行为集可表示为

$$A=\{-1,0,1\} \tag{5}$$

当新的事件发生时，即存在VSNR的到来或者离开时，网络根据当前的状态决定采取哪种行为。具体地

$$A(s)=\begin{cases}\{-1\}, & e=D_{R_i}\\ \{0,1\}, & e=A_{R_a}\end{cases} \tag{6}$$

其中，$A(s)=-1$表示资源需求量为R_i的VSNR已成功完成并从系统中离开。当新的VSNR到达时，若网络拒绝该请求，则$A(s)=0$，否则接受该VSNR且$A(s)=1$。

(3)状态转移概率：对于VSN中的SMDP，下一个状态出现的概率只依赖于前一个状态以及前状态所采取的行为，但是与历史状态无关，因此将状态转移概率定义为在当前状态s和采取行为$a\in A$时，转变为状态s'的概率[10]。设$\tau(s,a)$表示在采取行为a时从当前状态s转移到下一个状态s'的服务时间，$\sigma(s,a)$表示状态转移过程中网络的平均事件发生率，则$\sigma(s,a)=\tau(s,a)^{-1}$，且其表达式为

$$\begin{aligned}\sigma(s,a)&=\tau(s,a)^{-1}\\ &=\begin{cases}M\lambda_{\mathrm{p}}+\sum_{x=1}^{m}R_x\mu_{\mathrm{p}}-R_i\mu_{\mathrm{p}}, & e=D_{R_i},a=-1\\ M\lambda_{\mathrm{p}}+\sum_{x=1}^{m}R_x\mu_{\mathrm{p}}, & e=A_{R_a},a=0\\ M\lambda_{\mathrm{p}}+\sum_{x=1}^{m}R_x\mu_{\mathrm{p}}+R_a\mu_{\mathrm{p}}, & e=A_{R_a},a=1\end{cases}\end{aligned} \tag{7}$$

其中，$M\lambda_{\mathrm{p}}$用来表示网络中VSNR的到达率，$\sum_{x=1}^{m}R_x\mu_{\mathrm{p}}$表示网络中VSNR的离开率。当资源需求量为$R_i$的VSNR已完成且离开时，所有VSNR占用的物理资源总量变为$\left(\sum_{x=1}^{m}R_x-R_i\right)$，则对应的VSNR事件发生率为$\left(\sum_{x=1}^{m}R_x\mu_{\mathrm{p}}-R_i\mu_{\mathrm{p}}\right)$；同理，当资源需求量为$R_a$的VSNR到达且被接受时，所有VSNR占用的物理资源量变为$\left(\sum_{x=1}^{m}R_x+R_a\right)$，对应的VSNR事件发生率为$\left(\sum_{x=1}^{m}R_x\mu_{\mathrm{p}}+R_a\mu_{\mathrm{p}}\right)$。根据以上分析可知，当网络中存在VSNR的到来或者离开时，一旦网络中的资源状况发生变化，事件

发生率随之变化，且其大小与VSNR的资源量大小息息相关。

考虑不同事件，在状态s时采取行为a转变到状态s'的状态转移概率$P(s'|s,a)$可计算为

(a) $s=(R_1,R_2,\cdots,R_m,M,A_{R_a})$

$$P(s'|s,a=0)=\begin{cases}\dfrac{M\lambda_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,A_{R_a})\\ \dfrac{R_i\mu_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,D_{R_i})\end{cases}\tag{8}$$

$$P(s'|s,a=1)=\begin{cases}\dfrac{M\lambda_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,A_{R_a})\\ \dfrac{R_a\mu_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,D_{R_a})\\ \dfrac{R_i\mu_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,D_{R_i})\end{cases}\tag{9}$$

(b) $s=(R_1,R_2,\cdots,R_m,M,D_{R_i})$

$$P(s'|s,a=-1)=\begin{cases}\dfrac{M\lambda_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,A_{R_a})\\ \dfrac{M\lambda_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,A_{R_i})\\ \dfrac{R_i\mu_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,D_{R_i})\\ \dfrac{R_a\mu_{\rm p}}{\sigma(s,a)}, & s'=(R_1,R_2,\cdots,R_m,M,D_{R_a})\end{cases}\tag{10}$$

(4)奖赏函数：在当前状态为s，采取行为a时VSN的奖赏可表示为

$$r(s,a)=k(s,a)-g(s,a)\tag{11}$$

其中，$k(s,a)$表示网络在事件e发生时，在状态s下采取行为a的收益，该收益包括节省能量和加快VSNR的处理时间带来的效益，其具体计算表达式为

$$k(s,a)=\begin{cases}\omega_{\rm e}\cdot\beta_{\rm e}\cdot(E_{\rm l}-P_{\rm l}\cdot\delta)+\omega_{\rm d}\cdot\beta_{\rm d}\\ \cdot\left(D_l-\dfrac{1}{{\rm Pr}_i\cdot R_a\cdot\mu_{\rm p}}-\delta\right)-\gamma\delta, & e=A_{R_a},\ a=1\\ 0, & e=A_{R_a},\ a=0\\ 0, & e=D_{R_i},\ a=-1\end{cases}\tag{12}$$

其中${\rm Pr}_i$表示该VSNR的归一化重要程度量化值，其取值范围为${\rm Pr}_i\in(0,1)$，任务重要程度越大对应的${\rm Pr}_i$值越大；$\beta_{\rm e}$和$\beta_{\rm d}$表示单位能量和时间转换成收益的单价，可根据用户对资源的服务质量要求和资源提供商定价综合确定，假设单位能量和时间的所消耗资源的定价分别为p_1和p_2，则$\beta_{\rm e}=p_1\cdot[1+\log_2(1+{\rm Pr}_i)]$，$\beta_{\rm d}=p_2\cdot[1+\log_2(1+{\rm Pr}_i)]$；$\omega_{\rm e}$和$w_{\rm d}$分别表示处理VSNR时消耗的能量和时间的权重，则$\omega_{\rm e}+\omega_{\rm d}=1$，其大小取决于收益对能量和时间的偏重，即$\omega_{\rm e}=\beta_{\rm e}/(\beta_{\rm e}+\beta_{\rm d}),\omega_{\rm d}=\beta_{\rm d}/(\beta_{\rm e}+\beta_{\rm d})$；$E_{\rm l}$和$D_{\rm l}$表示处理VSNR时消耗的最大能量和时间，而$P_{\rm l}$表示处理VSNR时单位时间消耗的实际能量，$\delta$表示实际的处理时间；若该VSNR被分配的资源量大小为R_a，那么完成该请求需要的服务时间为$1/R_a\mu_{\rm p}$；γ表示单位时间的成本，则$\gamma\delta$表示接受该请求并将结果发送给被请求者的成本。

$g(s,a)$表示系统在一段连续时间内，从状态s转变到状态s'前处理服务请求所需的成本，其表达式为

$$g(s,a)=c(s,a)\cdot\tau(s,a)\tag{13}$$

其中，$\tau(s,a)$表示在采取行为a时从状态s转变为状态s'前的服务时间，而$c(s,a)=\sum_{x=1}^{m}R_x$表示在这个过程中给VSNR已分配的资源量大小。

为满足实际条件，考虑了一个连续时间的折扣因子α，因此折扣奖赏函数$r(s,a)$的表达式重新定义为

$$r(s,a)=k(s,a)-c(s,a)/[\alpha+\sigma(s,a)]\tag{14}$$

(5)最优策略求解：VSN中基于SMDP的资源分配就是找到最优的策略使得长期期望折扣奖赏值最大化，策略定义为$\pi.S\to a$即$\pi(S)=a$，该最优策略可描述为当VSN处于某种状态时，采取某种行为最大化奖赏值。因此，当事件发生时，在当前状态下网络能采用策略中的行为取得最大的奖赏值，即

$$v_a^{\pi*}(s)=\mathop{\rm Max}_{\pi}E_s^{\pi}\left[\sum_{m=1}^{\infty}e^{-\alpha\sigma_m}r(s_m,a_m)|s_0=s\right]\tag{15}$$

其中，s_0为系统的初始状态，s_m和a_m分别为第m个状态和对应的行为，σ_m表示在状态s_m下单位时间内的平均事件发生率。

根据贝尔曼方程[11]，考虑状态转移概率和折扣率λ，得到

$$v(s)={\rm Max}\left[r(s,a)+\lambda\cdot\sum_{s'\in S}p(s'|s,a)\cdot v(s')\right]\tag{16}$$

其中，$\lambda=\sigma(s,a)/(\alpha+\sigma(s,a)),\lambda\in[0,1]$。由于VSN中的状态是随机的，无法确定未来的奖赏是否与当

前奖赏带来的效益一样，所以引入了折扣率，平衡未来奖赏与当前奖赏之间的差距。特别地，当$\lambda=0$表示不考虑未来奖赏只考虑即时奖赏；当$\lambda=1$表示状态转移是确定的，未来的奖赏可根据当前的状态和行为进行确定。根据式(16)可知，当前状态和行为的最大未来奖赏是当前的直接奖赏加上下一个状态的最大未来奖赏，初始状态时，令等式左右两边的值函数相等。此外，为了进一步均衡连续时间SMDP，引入了参数y，且$y=K\cdot\lambda_{\mathrm{p}}+K\cdot R_a\cdot\mu_{\mathrm{p}}$，式(16)中相关参数的归一化表达式为

$$\widetilde{r}(s,a)=r(s,a)\cdot\frac{\alpha+\sigma(s,a)}{\alpha+y} \tag{17}$$

$$\widetilde{\lambda}=\frac{y}{\alpha+y} \tag{18}$$

$$\widetilde{p}(s'|s,a)=\begin{cases}1-\dfrac{[1-p(s'|s,a)]\sigma(s,a)}{y}, & s'=s\\ \dfrac{[1-p(s'|s,a)]\sigma(s,a)}{y}, & s'\neq s\end{cases} \tag{19}$$

根据式(17)—式(19)，式(16)可重新定义为

$$\widetilde{v}(s)=\mathrm{Max}\left[\widetilde{r}(s,a)+\widetilde{\lambda}\cdot\sum_{s'\in S}\widetilde{p}(s'|s,a)\cdot\widetilde{v}(s')\right] \tag{20}$$

在本文中，考虑有限的状态集和有限的行为集，但是连续时间内到达和离开的VSNR数量可能较大，因此采用值迭代[12]的方法解决大规模状态空间的SMDP，且其终止条件为前后策略的值函数差值小于$\varepsilon(1-\widetilde{\lambda})/2\widetilde{\lambda}$。该算法具体步骤如下：

步骤 1　对于状态s，初始化$\widetilde{v}(s)=0$，设置$\varepsilon=0,k=0$；

步骤 2　根据式(20)，利用当前策略的值函数$\widetilde{v}^k(s)$计算下一次的值函数$\widetilde{v}^{k+1}(s)$；

步骤 3　如果$||\widetilde{v}^{k+1}(s)-\widetilde{v}^k(s)||<\varepsilon(1-\widetilde{\lambda})/2\widetilde{\lambda}$，则转入步骤4；否则$k+1$，转入步骤2；

步骤 4　对任意的$s\in S$，计算最优策略$\pi_\varepsilon(s)\in\arg\max\left[\widetilde{r}(s,a)+\widetilde{\lambda}\cdot\sum_{s'\in S}\widetilde{p}(s'|s,a)\cdot\widetilde{v}^{k+1}(s')\right]$。

3.2 免模型强化学习算法

上节所述的值迭代方法求解最优策略是在一定的条件基础上执行的，该方法必须知道SMDP中的所有信息，包括状态转移概率以及执行该行为后获得的奖赏，但是在实际环境中却并不是能够预先知道的，因此，本文继续考虑强化学习算法[13]求得最优策略。强化学习算法分为基于模型的强化学习和免模型强化学习两种类型，前者需要对环境即网络相关信息进行建模，当学习后的模型非常贴近系统环境时，就可以直接通过一些规划算法来找到最优策略，后者则不需要对系统环境进行建模也能找到最优的策略，因此本文选择免模型强化学习算法。

免模型强化学习算法主要包括4个部分，控制器、系统环境、行为集和奖赏函数。其中控制器是学习的主体，感知系统环境反馈过来的状态、奖赏等信息并进行更新、存储与学习，并对网络的即时情况做出决策采取相应行为，而VSN的系统环境包括传感器节点及其资源状况、以及VSNR的动态变化情况等。

免模型强化学习的本质是通过不断地迭代从而解决贝尔曼优化方程，定义$Q^*(s,a)$是在状态为s采取行为a，并执行最优策略π^*的状态-行为对值函数，那么式(20)中关于奖赏的贝尔曼优化方程更新为

$$Q^*(s,a)=\widetilde{r}(s,a)+\sum_{s'\in S}\widetilde{p}(s'|s,a)\cdot\max_{b\in A}\ Q^*(s',b) \tag{21}$$

结合式(20)和式(21)，则可以看出$\widetilde{v}(s)=\max_{a\in A}Q^*(s,a)$，如果$Q^*(s,a)$已知，那么就可以得到最优的策略$\pi^*\in\arg\max_{a\in A}Q^*(s,a)$。但是，控制器无法获知式(21)中的状态转移概率，所以可以采用下面的状态-动作对值函数[14]并不断迭代得到近似的最优策略

$$\begin{aligned}Q_{m+1}(s,a)=&(1-\alpha_m)Q_m(s,a)\\&+\alpha_m[\widetilde{r}(s,a)+\max\ Q_m(s',b)]\end{aligned} \tag{22}$$

其中，α_m为第m次决策过程中的学习率，它均衡了上节中的值函数和本节中采用无模型强化学习中的值函数，特别地，当$\alpha_m=1$时，得到的更新方法与上节中的贝尔曼方程完全相同。如果控制器能以无限多的次数访问状态-行为对，那么强化学习算法将会收敛到最优[15]。

当有VSNR到达时，控制器可以使用贪心算法做出决策采取行为a^*，并且$a^*\in\arg\max_{a\in A}Q^*(s,a)$，但是控制器也有可能做出其它选择从而收集更多的信息以期望最大化值函数，即控制器以$1-\varepsilon$的概率在当前情况下做出最佳选择执行相应行为，同时以ε的概率选择随机行为。选择随机行为一方面是为了避免每次采样学习都得到同样的序列，另一方面是随机的行为也有可能最大化未来长期奖赏值。本文使用文献[16]中的快速随机梯度搜索方法来衰减ε，因此学习算法能够逐步得到收敛。此外，控制器通过系统环境响应的即时奖赏和状态变化的时间来收集信息进行学习并不断更新$Q(s,a)$，使得该强化学习算法比之前的值迭代方法更加智能。

通过强化学习算法的状态-动作对值函数即式(22)的迭代得到最优策略，由于采用了随机梯度

搜索方法来衰减ε，其时间复杂度为$O(n \times C \times I)$，采用随机梯度搜索方法则$n=1$, C 为单次迭代所需的计算时间，I为迭代次数。假设每次完成式(22)所需的时间为c_i，而式(22)是简单的多项式，完成1次迭代所需的时间较短，随机梯度搜索方法达到收敛所需的迭代次数为$1/\varepsilon$，所以对m个VSNR的状态动作对进行学习所需要的时间$T=m\cdot\sum_{i=1}^{1/\varepsilon}c_i$，即时间复杂度为$O(m/\varepsilon)$。同理，对于空间复杂度，假设完成1次迭代所需的运算次数为J，那么J为常数，得到最优策略需要的运算次数为I/ε，即所需的空间复杂度为$O(m/\varepsilon)$。综上可知，强化学习算法的时间复杂度和空间复杂度均依赖于ε，那么通过设定合理的ε，均衡学习算法所需的精度和迭代次数，可使该学习算法在有限的时间和空间中完成。

4　数值结果分析

本文采用MATLAB仿真平台验证VSN基于SMDP的资源分配策略性能，延伸基于贪婪算法(GA)即只考虑当前奖赏最大化并不考虑长远利益最大化，并将本文所提策略(SMDP)与贪婪算法(GA), CABPA资源分配算法[6]以及Heuristic算法[8]进行对比，验证本文算法的有效性。仿真参数设置如表1所示，以下参数在上文均有定义及解释。

表 1　仿真参数设置表

参数	数值	参数	数值
K	20～30	λ_{p}	1～18
ω_{e}	0.5	ω_{d}	0.5
β_{e}	2	β_{d}	2
E_{l}	20	P_{l}	5
D_{l}	20	δ	2
γ	2	α	0.1

图1描述了不同VSNR到达率即不同λ_{p}随不同资源总量的收益对比图。从图1可以看出，随着资源总量的增加，收益均呈增加的趋势，因为网络中消耗的能量和时间的成本在减少。此外，当资源总量相同时，λ_{p}越小，收益越大，因为需要消耗的能量和时间更少。

图2描述了4种算法在不同资源总量时的收益对比图，其中$\lambda_{\mathrm{p}}=3$。由图2可知，随着最大资源总量的增大，4种算法的收益均有所增加，且本文算法的收益始终高于其它3种算法。因为资源总量的增大，使得同等情况下处理VSNR的能量和时间更少，且本文算法选择恰当的行为最大化长期收益，因此收益会高于其它3种算法。

图3描述了不同算法随VSNR到达率的收益对比图。由图3可知，相同条件下本文基于SMDP的资源分配策略带来的收益始终高于其它3种算法，Heuristic算法旨在最大化共享物理资源的任务数即VSNR的数量从而提高资源利用率，CABPA算法主要考虑能耗因素，GA算法只考虑当前收益最大化，而本文算法能在特定状态下采取合适的行为最大化长期收益。

图4描述了4种不同算法在不同VSNR离开率时的收益对比图。由图4可知，随着VSNR离开率的增加，4种算法的收益均呈增加趋势且增速逐渐放慢，因为在这个过程中资源提供者的成本以$1/\mu$的速率在减小，所以收益以相同的速率在增加。当VSNR离开后，本文的算法更加有效地接受新到达的VSNR，从而带来更多的收益。

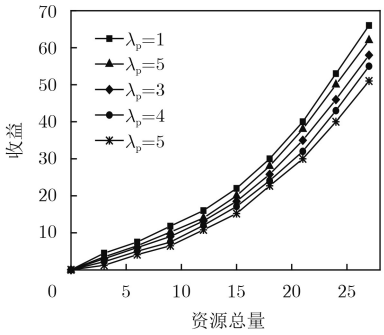

图 1 不同λ_{p}的收益对比

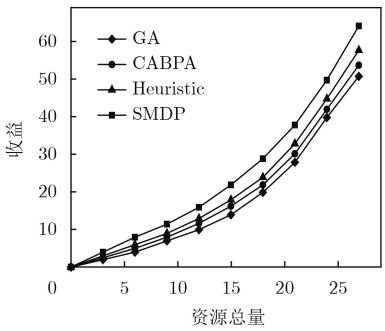

图 2 不同资源总量的收益对比图

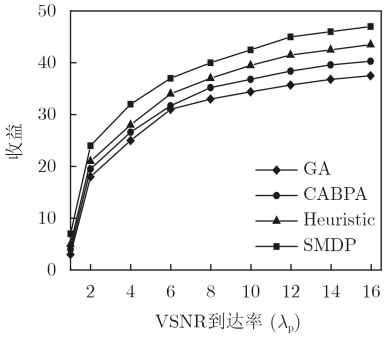

图 3 不同λ_{p}的收益对比图

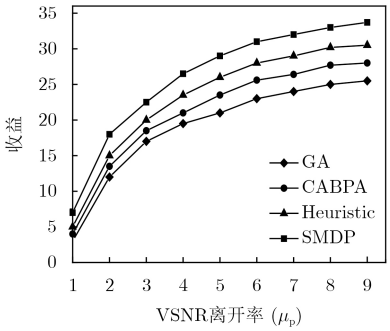

图 4 不同μ_p的收益对比图

图5描述了4种算法在不同VSNR到达率时VSNR的拒绝率。由图5可知，在VSNR到达率较小时，由于网络中有较多的资源，因此VSNR拒绝率总体较低，由于本文采用的SMDP会随机拒绝一部分VSNR所以其拒绝率是最大的，但是随着VSNR到达率继续增加，由于GA算法缺乏对网络整体资源的合理分配，在接受了较多的VSNR后，后期拒绝率会显著增大。

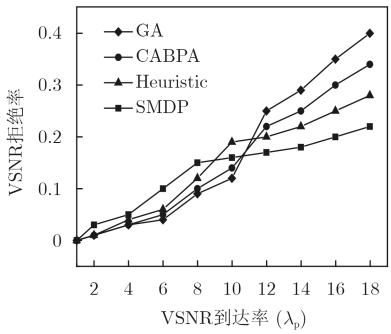

图 5 不同λ_p的拒绝率

5 结束语

根据虚拟化传感网中VSNRs的动态变化情况，本文提出了基于SMDP的VSNRs资源分配策略，定义状态集、行为集，计算不同状态的转移概率分布，给出奖赏函数表达式，并采用免模型强化学习算法求解特定状态下的行为，从而最大化资源提供者的长期收益。传感网中的传感器节点未来会以虚拟化、云化的方法互联、互通、互知，并实现不同任务间的资源共享，提高网络智能化水平，提升用户体验，改善网络中的资源利用率。

参 考 文 献

[1] YETGIN H, CHEUNG K T K, El-HAJJAR M, *et al*. A survey of network lifetime maximization techniques in wireless sensor networks[J]. *IEEE Communications Surveys & Tutorials*, 2017, 19(2): 828–854. doi: 10.1109/COMST.2017.2650979.

[2] WU Dapeng, ZHANG Feng, WANG Honggang, *et al*. Security-oriented opportunistic data forwarding in mobile social networks[J]. *Future Generation Computer Systems*, 2018, 87: 803–815. doi: 10.1016/j.future.2017.07.028.

[3] DELGADO C, CANALES M, ORTÍN J, *et al*. Joint application admission control and network slicing in virtual sensor networks[J]. *IEEE Internet of Things Journal*, 2018, 5(1): 28–43. doi: 10.1109/JIOT.2017.2769446.

[4] WU Dapeng, ZHANG Zhihao, WU Shaoen, *et al*. Biologically inspired resource allocation for network slices in 5G-enabled internet of things[J]. *IEEE Internet of Things Journal*, 2018. doi: 10.1109/JIOT.2018.2888543.

[5] GUO Lei, NING Zhaolong, SONG Qingyang, *et al*. A QoS-oriented high-efficiency resource allocation scheme in wireless multimedia sensor networks[J]. *IEEE Sensors Journal*, 2017, 17(5): 1538–1548. doi: 10.1109/JSEN.2016.2645709.

[6] ZHANG Yueyue, ZHU Yaping, YAN Feng, *et al*. Energy-efficient radio resource allocation in software-defined wireless sensor networks[J]. *IET Communications*, 2018, 12(3): 349–358. doi: 10.1049/iet-com.2017.0937.

[7] HASSAN M M and ALSANAD A. Resource provisioning for cloud-assisted software defined wireless sensor network[J]. *IEEE Sensors Journal*, 2016, 16(20): 7401–7408. doi: 10.1109/JSEN.2016.2582339.

[8] DELGADO C, GÁLLEGO J R, CANALES M, *et al*. On optimal resource allocation in virtual sensor networks[J]. *Ad Hoc Networks*, 2016, 50: 23–40. doi: 10.1016/j.adhoc.2016.04.004.

[9] WU Dapeng, LIU Qianru, WANG Honggang, *et al*. Cache less for more: Exploiting cooperative video caching and delivery in D2D communications[J]. *IEEE Transactions on Multimedia*, 2018. doi: 10.1109/TMM.2018.2885931.

[10] ZHENG Kan, MENG Hanlin, CHATZIMISIOS P, *et al*. An SMDP-based resource allocation in vehicular cloud computing systems[J]. *IEEE Transactions on Industrial Electronics*, 2015, 62(12): 7920–7928. doi: 10.1109/TIE.2015.2482119.

[11] SCHOLLIG A, CAINES P E, EGERSTEDT M, *et al*. A hybrid Bellman equation for systems with regional dynamics[C]. The 200746th IEEE Conference on Decision and Control, New Orleans, USA, 2007: 3393–3398. doi: 10.1109/CDC.2007.4434952.

[12] GOSAVI A. Relative value iteration for average reward semi-Markov control via simulation[C]. 2013 Winter Simulations Conference, Washington, USA, 2013: 623–630. doi: 10.1109/WSC.2013.6721456.

[13] WU Dapeng, SHI Hang, WANG Honggang, *et al*. A feature-based learning system for internet of things applications[J].

IEEE Internet of Things Journal, 2019, 6(2): 1928–1937. doi: 10.1109/JIOT.2018.2884485.

[14] CHEN Yueyun and JIA Cuixia. An improved call admission control scheme based on reinforcement learning for multimedia wireless networks[C]. 2009 International Conference on Wireless Networks and Information Systems, Shanghai, China, 2009: 322–325. doi: 10.1109/WNIS.2009.91.

[15] ABUNDO M, DI VALERIO V, CARDELLINI V, *et al*. QoS-aware bidding strategies for VM spot instances: A reinforcement learning approach applied to periodic long running jobs[C]. 2015 IFIP/IEEE International Symposium on Integrated Network Management, Ottawa, Canada, 2015: 53–61. doi: 10.1109/INM.2015.7140276.

[16] DARKEN C, CHANG J, and MOODY J. Learning rate schedules for faster stochastic gradient search[C]. 1992 IEEE Workshop on Neural Networks for Signal Processing II, Helsingoer, Denmark, 1992: 3–12. doi: 10.1109/NNSP.1992.253713.

王汝言：男，1969年生，教授，博士，研究方向为泛在网络、多媒体信息处理等.

李宏娟：女，1993年生，硕士生，研究方向为虚拟化、无线传感网络.

吴大鹏：男，1979年生，教授，博士，研究方向为泛在无线网络、无线网络服务质量控制等.

李红霞：女，1969年生，高级工程师，研究方向为光无线融合网络、无线传感网络.