无线网络中基于深度 Q 学习的传输调度方案

朱江，王婷婷，宋永辉，刘亚利
（重庆邮电大学移动通信技术重点实验室，重庆 400065）

摘 要：针对无线网络中的数据传输问题，提出一种基于深度 Q 学习（QL, Q learning）的传输调度方案。该方案通过建立马尔可夫决策过程（MDP, Markov decision process）系统模型来描述系统的状态转移情况；使用 Q 学习算法在系统状态转移概率未知的情况下学习和探索系统的状态转移信息，以获取调度节点的近似最优策略。另外，当系统状态的规模较大时，采用深度学习（DL, deep learning）的方法来建立状态和行为之间的映射关系，以避免策略求解中产生的较大计算量和存储空间。仿真结果表明，该方法在功耗、吞吐量、分组丢失率方面的性能逼近基于策略迭代的最优策略，且算法复杂度较低，解决了维灾问题。

关键词：无线网络传输；马尔可夫决策过程；Q 学习；深度学习
中图分类号：TN929.5
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018058

Transmission scheduling scheme based on deep Q learning in wireless network

ZHU Jiang, WANG Tingting, SONG Yonghui, LIU Yali
Key Laboratory of Information and Communication Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065

Abstract: To cope with the problem of data transmission in wireless networks, a deep Q learning based transmission scheduling scheme was proposed. The Markov decision process system model was formulated to describe the state transition of the system. The Q learning algorithm was adopted to learn and explore the system states transition information in the case of unknown system states transition probability to obtain the approximate optimal strategy of the schedule node. In addition, when the system state scale was big, the deep learning method was employed to map the relation between state and behavior to solve the problem of the large amount of computation and storage space in Q learning process. The simulation results show that the proposed scheme can approach the optimal strategy based on strategy iteration in terms of power consumption, throughput, packets loss rate. And the proposed scheme has a lower complexity, which can solve the problem of the curse of dimensionality.

Key words: wireless network transmission, Markov decision process, Q learning, deep learning

1 引言

近年来，随着无线通信技术的不断发展，各种新型的通信网络越来越多，区域化的具有小覆盖范围的通信模型逐步兴起，如应用于室内环境的毫微微网络、无线接入点等。为保证用户的通信质量需求，这些以小型基站和接入点为代表的调度节点需要及时处理大量的数据。因此，对其硬件和软件的处理速度、能耗等的要求就越来越高。在中继调度节点转发数据的过程中，如果节点没有足够的缓存

收稿日期：2017-03-31；**修回日期**：2018-03-16
通信作者：王婷婷，1762089088@qq.com
基金项目：国家自然科学基金资助项目（No.61102062, No.61271260, No.61301122）；重庆市基础与前沿研究计划基金资助项目（No. cstc2015jcyjA40050）
Foundation Items: The National Natural Science Foundation of China (No.61102062, No.61271260, No.61301122), Chongqing Research Program of Basic Research and Frontier Technology (No. cstc2015jcyjA40050)

区空间，那么将会造成数据分组的丢失，而且较差的信道状态也不能实现高效传输，造成能量的浪费。因此，无线网络中的有效传输逐渐成为当前学者的研究热点之一。

针对中继节点的传输调度问题，文献[1]考虑单个发送机，在建立马尔可夫决策过程模型的基础上，通过引入拉格朗日乘子法和黄金分割搜索方法构建了一种无线电网络的传输调度方案，该方案可以在满足缓存区的分组丢失率的约束下，最小化平均功率。文献[2]在 MDP 模型的基础上考虑具有中心节点的无线传输网络，通过 W 学习算法来指导中继节点为其他节点传输数据。文献[3]在图形博弈的基础上引入了学习算法，通过次梯度迭代算法来交换多个代理之间的信息，让单个代理来学习自己周围的环境信息，并据此来指导节点选择信道传输数据。文献[4]同样将无线传输问题描述为 MDP 过程，通过设定新的目标函数来实现延长终端寿命的目的。不过，该文使用策略迭代的方法来对 MDP 的调度问题进行求解，该方法具有较大的求解计算量。现有文献在解决数据传输的问题时，较多关注于单方面的优化目标，没有综合考虑这些方面。此外，在快速时变的环境下，系统在进行求解计算中有大量的信息交互，容易产生维灾难，很难做到快速收敛。目前，针对此类问题的研究较少，文献[1,2]采用状态聚合和行动集缩减的方法来减小系统计算规模，但此种方法需要根据具体问题重新定义状态空间和行动集。

本文在无线网络数据传输问题中，综合考虑数据分组的丢失、能耗以及系统的吞吐量。在实际的无线网络中，合理的假设是环境信息未知，即中继节点不能事先知道环境状态的转移概率。为此，在建模为 MDP 的系统模型中，通过 Q 学习的强化学习方法使中继调度节点对周围环境状态的转移信息进行学习，并对节点的行为进行指导。在状态行为获取中，为了考虑探索与利用的均衡，改进策略选择方法，提出了基于行为评价值 Q 和 $Index(s, a)$ 索引值的综合行为评价方法，以获取更优的状态行为数据。另外，基于强化学习获得的行为数据，使用深度学习方法来构造状态和行为之间的映射关系，达到快速求解的目的。

2 系统模型

在无线网络环境下，数据传输模型如图 1 所示。系统中存在 K 个待传输数据的节点，每个节点对应的缓存区长度为 L，存在一个中继节点可以为 K 个用户传输数据，并假设该节点可用的频谱资源的个数为 M，且各个节点之间和频谱之间相互独立。假设节点的上层数据以到达率为 λ 的泊松分布到达，并存储在对应的缓存区内。每一帧内，中继节点选择一个信道状态较好的信道为一个节点传输数据。当缓存区中的数据满时，如果中继节点没有为它传输数据，那么当下一帧再有数据到达时会造成数据的丢失。

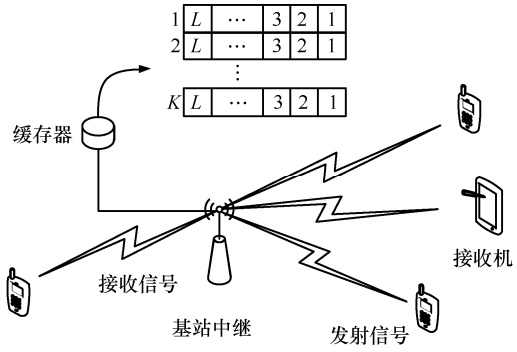


图 1 数据传输模型

2.1 信道状态

定义时间基本单位为帧 T_f，且在每一帧内信道的状态保持不变，信道的状态转移发生在 2 个相邻的状态之间。将块衰落信道状态建模为一阶的有限状态马尔可夫链（FSMC, finite state Markov chain）[5]。在 FSMC 信道状态模型中，存在加性高斯白噪声的信道接收信噪比（SNR, signal-to-noise ratio）服从瑞利分布，其概率密度函数表示为 $p(\rho)=\frac{1}{\overline{\rho}}\exp\left(-\frac{\rho}{\overline{\rho}}\right)$。其中，$\rho>0$，且 $\overline{\rho}=\mathrm{E}(\rho)$ 表示平均接收信噪比。若信噪比门限设为 $\rho_{\mathrm{snr}}=\{\rho_1,\rho_2,\cdots,\rho_N\}$，那么，根据接收信噪比门限，可以将信道划分为多个状态 $C\triangleq\{c_0,c_1,\cdots,c_{N-1}\}$。信道状态概率为

$$p_{\mathrm{C}}(c_n)=\int_{\rho_n}^{\rho_{n+1}}p(\rho)\mathrm{d}\rho \tag{1}$$

信道的状态转移概率为

$$p_{\mathrm{C}}(c_n,c_{n+1})=\frac{N(\rho_{n+1})T_f}{p_{\mathrm{C}}(c_n)},n\in\{0,1,\cdots,N-2\} \tag{2}$$

$$p_{\mathrm{C}}(c_n,c_{n-1})=\frac{N(\rho_n)T_f}{p_{\mathrm{C}}(c_n)},n\in\{1,2,\cdots,N-1\} \tag{3}$$

其中，$N(\rho_n)=\sqrt{\frac{2\pi\rho_n}{\bar{\rho}}}f_{\mathrm{D}}\exp\left(-\frac{\rho}{\bar{\rho}}\right)$，$f_{\mathrm{D}}$ 为最大多普勒频移。那么，系统的信道状态转移概率为 $p_{\mathrm{C}}(c,c')=\prod_{m=1}^{M}p_{c_m}(c_i,c_{i+1}|a_{i,j})$。

2.2 缓存器状态

在每一帧内，每个节点对应缓存区中的数据分组到达服从到达率为 λ 的泊松分布 $p_{d_i}(d_i)=\frac{\exp(-\lambda_{d_i}T_f)(\lambda_{d_i}T_f)^{d_i}}{d_i!}$，$d_i$ 为在每帧内到达的数据量。

定义在时刻 i 节点 k 对应的缓存区中的数据分组数为 $l_{k,i}$，如果每帧到达的数据分组为 $d_{k,i}$，传输的数据分组为 $t_{k,i}$，那么，节点 k 对应的缓存区中的数据量为

$$l_{k,i+1}=\min(l_{k,i}+d_{k,i}-t_{k,i},L) \tag{4}$$

如果用户 k 对应的缓存区的状态转移概率为 $p_{l_k}(l_k,l_k')$，那么，系统中缓存器的状态转移概率为 $p_l(l,l')=\prod_{k=1}^{K}p_{l_k}(l_i,l_{i+1}|a_{i,j})$。

2.3 发送功率

中继节点在通过信道发送数据的时候采用自适应调制（AM, adaptive modulation）的方式(j-QAM) [6~8]，用 $j=1,2,3,4,\cdots$ 来表示选中的方式。通过限定传输方式下的误码率（BER, bit error ratio），可以得到在不同状态和传输方式下满足误码率要求时的最小功率[8]。

当 $j=1$ 时，有

$$p_{\mathrm{BER}}(c_i,j)\leqslant 0.5erfc\left(\sqrt{\frac{\rho_i P(c_i,j)}{WN_0}}\right) \tag{5}$$

当 $j>1$ 时，即 $j=2,3,4,\cdots$ 时，有

$$p_{\mathrm{BER}}(c_i,j)\leqslant 0.2\exp\left(\frac{-1.6\rho_i P(c_i,j)}{WN_0(2^j-1)}\right) \tag{6}$$

其中，WN_0 表示噪声功率。

3 传输调度的 MDP 问题分析

系统中存在 2 个状态对象，分别是节点对应的缓存器的状态和信道状态。系统运行是一个状态转移的过程，系统在当前状态下通过执行某个行为进入到下一个状态。因此，传输调度问题可以建模为 MDP[1,2,4]。

3.1 系统状态转移

定义系统的状态 S 为缓存器和信道的组合状态，即 $S\triangleq B\otimes C$。如果缓存区的长度为 L，那么，单个缓存区的状态个数为 B=L+1。信道的状态个数为 N。

中继节点所采取的行为 A 可以表示为 $a_i=\{a_{i,1},a_{i,2},\cdots,a_{i,K}\}$，当其值 $a_{i,k}=1$ 时，表示在时隙 i 中继节点为节点 k 来传输数据，其中，含有选择的信道 m 和传输方式 j 信息。当 $a_{i,k}=0$ 时，表示不采取任何行动。

在时刻 i，存在一个缓存区和一个信道的情况下，系统状态处于 s_i 时采取行为 a_i 后转移到状态 s_{i+1} 的状态转移概率可以表示为 $p_s(s_i,s_{i+1}|a_i)=p_l(l_i,l_{i+1}|a_i)p_c(c_i,c_{i+1}|a_i)$。那么，整个系统的状态转移概率为

$$p_s(s_i,s_{i+1}|a_i)=\prod_{k=1}^{K}p_{l_k}(l_i,l_{i+1}|a_i)\prod_{m=1}^{M}p_{c_m}(c_i,c_{i+1}|a_i) \tag{7}$$

3.2 系统收益和代价

高效传输是要实现在最大化吞吐量的同时最小化系统发送功率和分组丢失数。如果系统中信息的基本传输速率为 v，那么，在使用不同的传输方式下的传输数据量为 $t=v\times 2^j$。在当前的系统状态 $s_i=\{l_i,c_i\}$ 下选择行为 $a_i=\{a_{i,1},a_{i,2},\cdots,a_{i,K}\}$ $(a_i\in A)$ 后，系统可以获得的最大收益为

$$r_{k,i}=\max(R(s_i,a_i))=\max\left(\sum_{k=1}^{K}v\times 2^j\right) \tag{8}$$

定义缓存区 k 的压力值 $f_{k,i}=\exp(\theta\times l_{k,i})$，其中，$\theta$ 表示缓存区的压力系数。缓存区中的数据越多，接下来到达的数据就越可能因为没有存储空间而丢失。因此，压力值反映了缓存区中的数据量的多少，其值与系统性能成反比，故越小越好。此外，当系统处于某一状态 s_i，采取行为 a_i 时，如果给定比特误码率界限，那么可以得到系统的最小传输功率 $p_{s_i}(s_i,a_i)$ [9]。功率与系统性能成反比。

因此，定义系统的代价 Co 为缓存器的压力值和传输功率的组合，即

$$Co=\left(\sum_{i=1}^{K}f_{k,i}\right)p_{s_i}(s_i,a_i) \tag{9}$$

定义系统的效用为 O，该值与每一帧内传输的数据量成正比，与缓存器的压力值和功耗成反比，可以得到该表达式

$$O=\frac{v\times 2^{j}}{(\sum_{k=1}^{K} f_{k,i})p_{s_i}(s_i,a_i)} \tag{10}$$

4 基于深度 Q 学习的最优行为获取

在本文的 MDP 模型中，中继节点通过学习来获得系统的状态参数。系统的状态和行为是离散的，每执行一次行为，系统的状态就会发生不连续的变化。因此，在求解此 MDP 问题时采用强化学习的方法来指导节点行为[10~12]。当系统规模较大时，往往存在较大的状态行为空间，不宜直接对 MDP 进行求解。这里，使用 QL 算法实现对环境状态信息的学习并获得最优行为。在使用 QL 算法对此问题进行求解时，是一种逐步寻优的过程，所以很难实现行为选择的快速收敛。由于深度神经网络具有较好的泛化能力，几乎可以实现任意非线性函数的逼近功能。所以，在 QL 算法的基础上使用深度神经网络的方法建立状态和行为之间的映射关系，实现问题的加速求解[13~16]，以解决因系统状态规模较大而导致的维灾问题[17~18]。

4.1 QL 算法

Q 学习在与环境交互的过程中，通过不断地试错来找到最优的行为。该最优行为不仅关注立即收益，同时还考虑了未来 n 步的收益情况，用 $V^{\pi}(s_i)$ 表示，即

$$V^{\pi}(s_i)=r_i+\gamma r_{i+1}+\gamma^2 r_{i+2}+\cdots \tag{11}$$

Q 学习算法中的 Q 值是状态和行为的评价值，用立即收益和折扣收益来表示，即

$$Q(s_i,a_i)\leftarrow r_i+\gamma V^{\pi}(s_{i+1}) \tag{12}$$

其中，$\gamma(0<\gamma<1)$ 是折扣系数，表示未来收益对当前行为的影响。Q 学习的目标是为了最大化系统效用，用 O_i 替换 r_i，用 $\max\limits_{a_{i+1}\in A} Q(s_{i+1},a_{i+1})$ 来替换 $V^{\pi}(s_{i+1})$。那么，可得

$$Q(s_i,a_i)\leftarrow O_i+\gamma\max_{a_{i+1}\in A} Q(s_{i+1},a_{i+1}) \tag{13}$$

策略的探索与利用[19]是强化学习中的重要问题。尤其是当系统状态规模较大时，如何有效地选择行为，将直接影响算法的收敛速度以及形同的性能。本文引入基于行为评价值 Q 和 $Index(s, a)$ 索引值的综合行为评价方法，以获取更优的状态行为数据。

当系统处于状态 s_i 时，依据式(14)来选择行为 a_i。

$$a_i\leftarrow\arg\max_{a\in A}(Q(s_i,a)+Index(s_i,a)) \tag{14}$$

其中，Q 表示状态和行为评价值。在 Q 值的基础上，增加行为索引值来选择可以最大化收益所对应的行为，表示为

$$Index(s_i,a)=\varsigma\sqrt{\frac{2\ln n}{T_i(n)}\min\{\frac{1}{4},v_i(n)\}} \tag{15}$$

其中，ς 是一个大于 0 的常量[20]，$T_i(n)$ 表示经过 n 次行为选择之后，行为 a_i 被选择的次数。$v_i(n)$ 表示偏差因子，引入了行为效用值的方差为 $\sigma_i^2(n)$，以反映该值波动性。

$$\sigma_i^2(n)=\frac{\sum_{t=1}^{T_i(n)} r_i^2(t)}{T_i(n)-O_i^2\left(T_i(n)\right)} \tag{16}$$

$$v_i(n)=\sigma_i^2(n)+\sqrt{\frac{2\ln n}{T_i(n)}} \tag{17}$$

基于索引 $Index(s_i,a)$ 的行为选择机制，一方面逐步考虑具有较大效用的行为，体现了利用的特性；另一方面，随着迭代的进行，如果某个行为未被选择或被选择的次数 $T_i(n)$ 很少，那么，在接下来的选择中会偏向于选择该行为，体现了探索的特性。

在确定了执行行为后，中继节点执行行为 a_i，然后计算效用值 O，并按照式(18)更新 Q 值。

$$Q_{i+1}(s_i,a_i)=\begin{cases}(1-\alpha)Q_i(s_i,a_i)+\alpha(O_i+\gamma\max Q_i(s_{i+1},a_{i+1})), & s=s_i \text{ 且 } a=a_i\\ Q_i(s_i,a_i), & \text{其他}\end{cases} \tag{18}$$

其中，$\alpha(0<\alpha\leqslant 1)$ 是状态行为的学习因子，表示为 $\alpha=\frac{1}{(1+T_i(n))}$。

本文 Q 学习方法具体执行过程如算法 1 所示。

算法 1 Q 学习算法

1) 初始化行为访问次数 $T_i(n)=0$

2) 初始化状态行为值 $Q(s_i,a_i)=0$ 及状态行为查询表

3) for $episode_1$=1, I_1 do

4) 初始化行为向量 $\boldsymbol{a}=\{a_1,a_2,\cdots\}$

5) 对于当前状态 s_i

6) for $episode_2$=1, Ite do

7) if $episode_2$=1

8) 随机选择一个行为 a_i

9) end if

10)　if *episode*2>1

11)　$Index(s_i,a)\leftarrow \varsigma\sqrt{\frac{2\ln n}{T_i(n)}\min\{\frac{1}{4},V_i(n)\}}$

12)　根据

$a_i\leftarrow \max_a(Q(s_i,a)+Index(s_i,a))$ 选择行为

13)　end if

14)　执行行为 a_i，获得效用值 O_i，得到下一个状态 s_{i+1}

15)　计算 $\alpha\leftarrow\frac{1}{1+T_i(n)}$

16)　更新 $Q(s_i,a_i)$

17)　更新查询表

18)　end for

19) end for

4.2　算法性能分析

在算法的收敛性分析阶段，本文把最优的 Q 值表示为 $Q^*(s_i,a_i)$。

定理 1　式(10)定义的系统效用函数 O 的值在不同的系统状态下有界。

证明见附录 A。

定理 2　对于收益 O 有界的 Q 值迭代问题，学习因子 $0<\alpha\leqslant 1$，且

$$\sum_{i=1}^{\infty}\alpha_{T_i(n)}=\infty,\sum_{i=1}^{\infty}\alpha_{T_i(n)}^2<\infty,\forall s,a \tag{19}$$

那么，当 $T_i(n)\to\infty,\forall s,a$ 时，有

$$\lim_{i\to\infty}Q_i(s_i,a_i)=Q^*(s_i,a_i) \tag{20}$$

证明　见附录 B[20]。

本文取 $\alpha=\frac{1}{1+T_i(n)}\in(0,1]$，满足式(19)。

4.3　深度行为映射网络

采用多层的栈式自编码（SAE, stacked auto-encoder）深度神经网络模型来建立状态和行为之间的对应关系，以最快获取最优决策行为。模型的结构如图 2 所示。

模型的输入层代表了系统的状态信息，该层的神经元的个数为 $K+M$。输入向量表示为 $\boldsymbol{Input}=[l_1\ \cdots\ l_K\ c_1\ \cdots\ c_M]$，分别代表了系统中的缓存器的状态和信道状态。输出层神经元代表了行为选择信息，输出向量为 $\boldsymbol{Output}=[a(k)\ a(m)\ a(j)]$，分别表示在该系统状态下，通过信道 m，以传输方式 j 为用户 k 发送数据。该层的神经元个数为

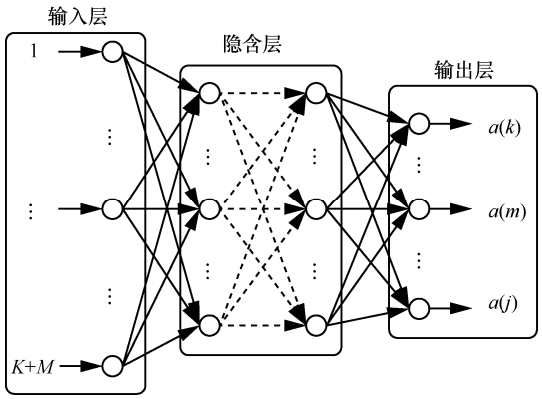


图 2　SAE 网络模型

$K+M+J$。隐含层为多层，节点个数由式(21)确定。

$$n_h=\sqrt{n_i+n_o}+\mathrm{Con} \tag{21}$$

其中，n_i 表示输入层神经元个数，n_o 表示输出层神经元个数，Con 为 1~10 的一个常数。

SAE 模型使用 sigmoid 函数作为传递函数，训练过程中的损失函数为 $L(x)$。

$$L(x)=\arg\min_{x\in(0,1)}\frac{1}{2}\sum_{i=1}^{N}\left\|x^i-f(x^i)\right\|^2+\mu J_{W^l} \tag{22}$$

其中，$J_{W^l}=\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(W_{ji}^l)^2$ 表示链接权值的衰减项，N 为输入的样本总数，μ 为权值矩阵的权重参数，n_l 表示权值矩阵的个数，W_{ji}^l 表示第 l 个权值矩阵，s_l 和 s_{l+1} 分别表示相邻的 2 个隐含层中的神经元节点的个数。使用梯度下降法对参数 W 和偏置 b 进行更新。

5　算法描述与比较

5.1　算法描述

基于深度 Q 学习的策略选择算法流程如图 3 所示。首先，使用 Q 学习算法经过一定时隙的学习获取一部分状态和行为数据，这个过程中不对 SAE 进行训练。随着时间的进行，对于某些状态逐渐找到最优行为，并存储于 Q 查询表中。使用该表中的信息进行有监督地训练 SAE 网络。当系统转移到隐藏状态时，通过 SAE 网络来实现该状态下的行为映射。执行所推荐的行为并更新 Q 值，并将该状态行为信息存储于 Q 查询表中。当后续时刻系统再转移到该状态时，直接通过查询状态行为表来获得可执行的行为。

5.2　算法比较

在本文系统模型中，无线节点个数为 K，缓存

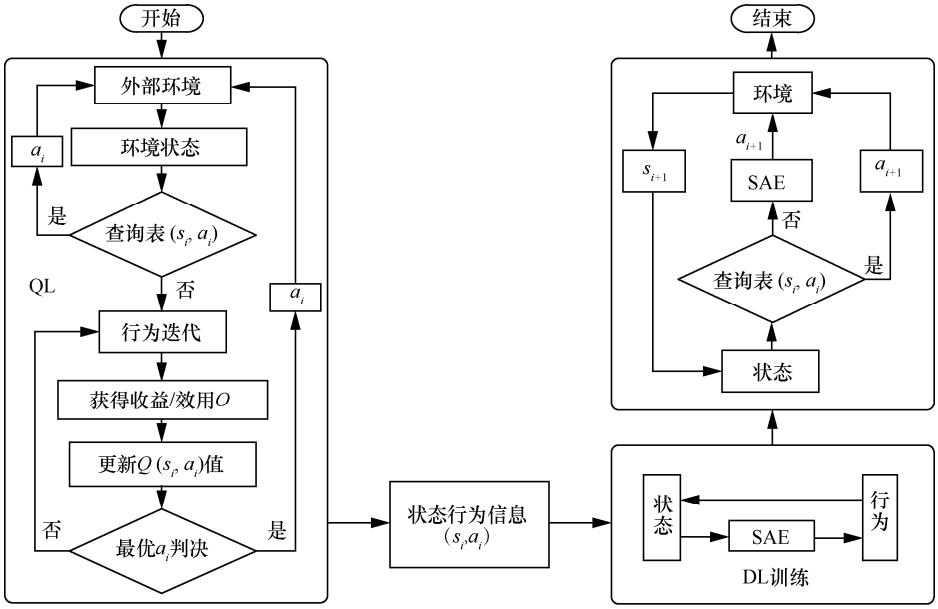


图 3　基于 Q 学习的 SAE 行为获取算法流程

区的长度为 L，信道个数为 M，信道的状态个数为 C，可用的传输方式个数为 J。那么，系统的状态个数为 $S=(L+1)^K C^M$，对于每一个系统状态，可选的行为个数为 $A=KMJ$。因此，系统的规模可以表示为 $D=SA$。

为了验证本文算法的性能，现分别与策略迭代法[4]、W 学习方法[2]和随机选择方法进行比较分析。

1) 策略迭代法

当 MDP 问题被描述为式(23)时，可以通过策略迭代（SI, strategy iteration）法来求解最优决策。

$$V_{n+1}=\max_{a\in A}[r(a)+\gamma p(a)V_n]=r(a_n)+\gamma p(a_n)V_n \tag{23}$$

在求解过程中，SI 法需要预先知道系统的所有状态信息和状态转移概率。然而，当系统的状态空间较大时，需要求解与系统状态等规模的线性方程，即 $S=(L+1)^K C^M$ 个线性方程组，每一次迭代的计算复杂度会达到 $A|S|^2+|S|^3$。所以该算法易陷入维灾问题，在实际问题中并不适用[21]。

2) W 学习法

在 W 学习（WL, W learning）法中，首先使用 Q 学习方法获得 Q 值，然后利用已获得的值进行 W 学习。W 值代表预计收益和实际收益的差值。

$$W_{i+1}(s_i)=(1-\alpha)W_i(s_i)+\alpha(Q_i(s_i,a_i)-(r_i+\gamma\max_{a_i\in A}Q_i(s_{i+1},a_i))) \tag{24}$$

3) 随机选择法

随机选择（RS, random selection）法是在每一个系统状态下，在行动集中随机选择一个行为执行，所以它的计算开销较小。

本文算法只针对系统的当前状态执行算法，不需要过多的环境状态信息。各个算法的计算复杂度比较如表 1 所示。可知，本文算法复杂度较低，且不依赖于系统状态的先验信息。

表 1　　算法复杂度比较

算法	是否依赖先验信息	指数运算	乘除运算	加减运算	比较运算
SI 法	是	0	$D+S$	D	D
本文算法	否	$2A$	$4A$	$2A$	$2A$
W 学习法	否	$2A$	$3A$	$3A$	A
RS 法	否	0	1	0	0

6　仿真分析

本文仿真实验考虑存在 K=5 个待发送数据的无线节点和一个智能中继节点。其中，中继节点可选信道个数为 M=3，设定的信道状态数为 C=4，传输方式的个数为 J=4。无线节点的缓存区长度为 L=5。针对此问题，分别使用上述所描述的 3 种方法和本文方法进行性能对比。仿真的各项参数设置如表 2 所示。结果分别通过图 4~图 8 说明。

表 2　　仿真参数

参数	值描述
信噪比门限/dB	snr=[−6.28, −1.28,1.28]
多普勒频移/Hz	f_D=50
帧长/s	$T_f=2\times10^{-3}$
时隙数 1	$I_1=5\times10^3$
时隙数 2	$I_2=1\times10^3$
噪声功率	1×10^{-3}
缓存区压力系数	θ=0.5
到达率	λ=[0.1,⋯,0.9]
误比特率约束	$BER=10^{-3}$
折扣系数	γ=0.9
索引权重	$\varsigma=\frac{1}{\sqrt{2}}$
MQL 学习因子	$\alpha\in(0,1]$
SAE 各层神经元数	[9,15,15,12]
权值权重	$\mu=3\times10^{-3}$
SAE 学习速率	1×10^{-2}
训练误差精度	$error=1\times10^{-5}$

在 I_1 时隙内是 Q 学习阶段，通过 QL 方法来获得最优状态行为信息并存储于查询表中。

然后，使用这些信息来对 SAE 网络进行有监督的训练，并在接下来的时隙 I_2 内进行对比实验。

图 4 所示的是系统分别处于状态 s_1={l_1=5, c_1=3}、s_2={l_2=9,c_2=4}和 s_3={l_3=12,c_3=2}时，使用 QL 法的行为选择中的 Q 值变化曲线。3 种状态在相同的数据到达率($\lambda=0.1$)下，最终 Q 值逐渐收敛于不同的数值。

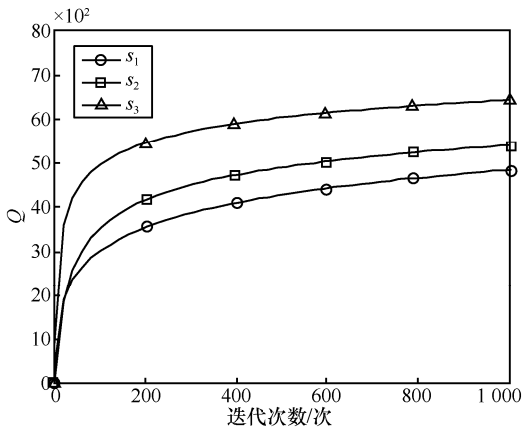


图 4　Q 学习算法在不同状态下 Q 值曲线

在不同的上层数据分组到达率下，不同算法的系统传输的数据量对比如图 5 所示。由图 5 可知，本文算法的数据传输量小于 SI 法的数据传输量，但是优于 W 学习法和 RS 法。从图 5 可以看出，随着数据分组到达率的增大，系统的吞吐量也逐渐增大。当系统中达到的数据分组越来越多时，系统相应的缓存器的压力逐渐增加，这样会使系统的效用减小。因此，在策略寻优时会增加数据的发送量以减小缓存器的压力。到达率对于 RS 法影响较小，因为 RS 法在决策时并不考虑系统的状态信息。

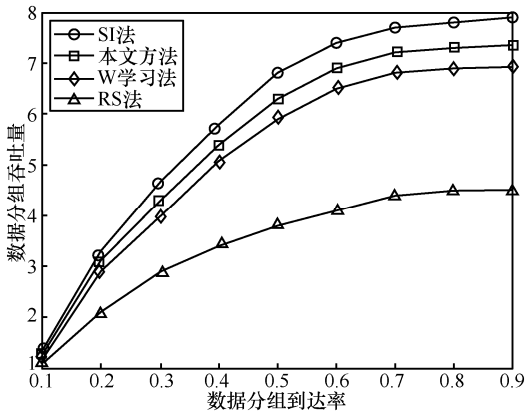


图 5　系统传输数据量对比

图 6 为在不同的数据到达率和不同的算法下系统的平均功耗对比。由图 6 可知，RS 法相对较为平稳。因为 RS 法并不受系统状态信息的影响，所以，数据到达率 λ 对于 RS 传输方式选择基本上没什么影响，另外 3 种方法的功率则受 λ 的影响较大。当系统中的数据量较大时，缓存器的压力较大，迫使中继节点选择更好的传输方式来发送更多的数据以减小缓存器的压力，最终使功率消耗较大。3 种方法的能耗均呈现出先快速增长，随后平缓增长

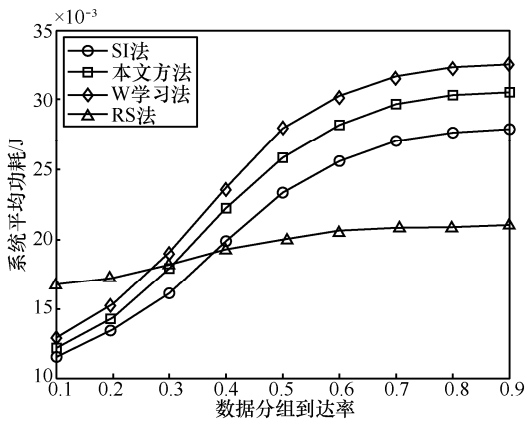


图 6　系统平均能耗对比

的趋势。因为随着缓存器中数据的增多，发送的数据越多，能耗越大，故功率曲线增长越快。因为缓存空间有限，当数据量达到了缓存的最大承受限度后，缓存压力不再增大，因此，最终也趋于平稳。

当系统缓存器中的缓存空间较小时，如果下一帧到达的数据分组个数较多，会因为没有足够的空间而造成数据丢失。系统的平均分组丢失数如图 7 所示，随着到达率的逐渐增大，4 种算法的分组丢失数均逐渐增加。由于 RS 法在选择行为时不考虑功耗和缓存压力，因此，分组丢失数较大。相对来说其他 3 种方法的分组丢失数较小。而本文算法的分组丢失数虽然大于 SI 法获得的最优值，不过仍小于 W 学习法的分组丢失数。

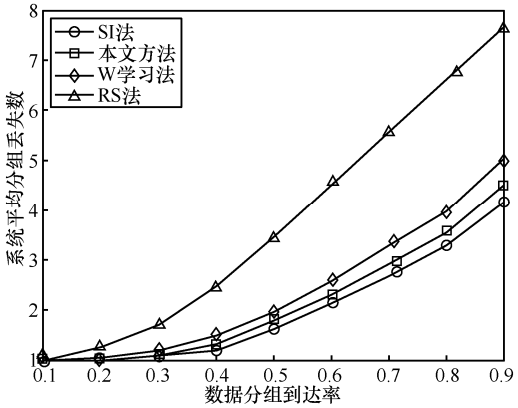


图 7　系统平均分组丢失数对比

系统的平均效用对比如图 8 所示。由图 8 可知，SI 法、本文方法和 W 学习法的效用值均高于 RS 法。RS 法在行为选择时未考虑组成效用函数的各个参量因素。由图 8 可知，本文方法的系统效用值虽然相较 SI 法小，但仍优于 W 学习法的效用值。随着数据分组到达率的增大，这 3 种效用曲线均呈现出先增大后减小的趋势。这是因为当 λ 较小时，系统可以同时选择合适的行为方式来提升效用值。但是当数据量较大时，虽然系统在尽力传输数据但仍然不能做到数据的完全传输，同时在传输数据量较大时功耗也很大。因此，效用曲线呈现出先增大后减小的趋势。

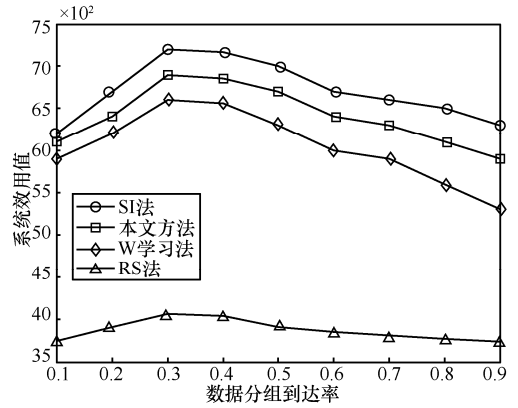


图 8　系统平均效用对比

7　结束语

针对无线网络中的高效传输问题，本文建立了基于 MDP 的系统模型。MDP 模型是一个行为选择以及状态转移模型，通过选择最优的行为来最大化收益。本文提出使用深度 Q 学习的人工智能算法对该 MDP 的传输决策问题进行求解，适合环境状态信息未知，即状态转移概率未知的实际场景。在策略的求解问题中，策略迭代法往往能取得最优的策略，但是该方法易陷入维灾问题，且在求解最优策略的过程中依赖于事先知道的状态转移概率。本文方案是在当前状态下进行求解，不需要过多的环境状态信息；并且，使用深度学习的方法来建立系统状态和行为之间的映射关系，避免了强化学习过程的较大计算量，实现了快速求解，解决了维灾问题。

附录 A　定理 1 证明

证明　式(10)是由 3 个部分组成，分子为收益函数 $r_{i,j}=v\times 2^j, j=1,2,3,\cdots$，表示每一帧能够发送的数据量，是一个有限值。分母部分为系统代价(式(9))。其中，缓存区的压力表达式为 $f_{k,i}=\exp(\rho l_{k,i})$，而缓存中的数据量是有界整数值。因此，$f$ 同样是离散的有限值。由式(5)和式(6)可知，发送的功率与传输方式有关，因此，功率 p 也是离散有限值。所以，系统效用值有界。证毕。

附录 B　定理 2 证明

证明　定义初始函数为 $Q_0(s_i,a_i)$，对于所有的 $s_i\in S$ 和 $a_i\in A$ 均按照式(18)进行更新获得最优值 $Q^*(s_i,a_i)$。

对于函数 $Q^*(s_i,a_i)$、$O(s_i,a_i)$ 和 $Q_0(s_i,a_i)$，使常量 $\varepsilon,\eta,\vartheta,\zeta$ 和 $\gamma,(0<\gamma<1)$ 满足

$$\varepsilon O(s_i,a_i)\leqslant\gamma\max Q^*(s_{i+1},a_{i+1})\leqslant\eta O(s_i,a_i) \tag{25}$$

$$\vartheta Q^*(s_i,a_i)\leqslant Q_0(s_i,a_i)\leqslant\zeta Q^*(s_i,a_i) \tag{26}$$

其中，$0<\varepsilon\leqslant\eta<\infty$ 和 $0\leqslant\vartheta\leqslant\zeta<\infty$，因为最优值是未知的，所以 ε、η、ϑ 和 ζ 的值不能直接获得。因此，需证明对于 $\forall\varepsilon$、η、ϑ、ζ，经过迭代后 $Q(s_i,a_i)$ 可以收敛得到最优。

如果 $0\leqslant\vartheta\leqslant\zeta<1$，那么对于 $\forall i=0,1,\cdots$，性能函数

$Q_i(s_i,a_i)$ 满足

$$(1+\frac{\vartheta-1}{(1+\eta^{-1})^i})Q^*(s_i,a_i)\leqslant Q_i(s_i,a_i)$$
$$\leqslant(1+\frac{\zeta-1}{(1+\varepsilon^{-1})^i})Q^*(s_i,a_i) \tag{27}$$

下面，通过数学归纳法证明式(27)。

当 $i=0$ 时，有

$$Q_1(s_i,a_i)=O(s_i,a_i)+\gamma\max Q_0(s_{i+1},a_{i+1})$$
$$\geqslant O(s_i,a_i)+\vartheta\gamma\max Q^*(s_{i+1},a_{i+1})$$
$$\geqslant(1+\eta\frac{\vartheta-1}{1+\eta})O(s_i,a_i)+\gamma(\vartheta-\frac{\vartheta-1}{1+\eta})\max Q^*(s_{i+1},a_{i+1})$$
$$=(1+\eta\frac{\vartheta-1}{1+\eta})[O(s_i,a_i)+\gamma\max Q^*(s_{i+1},a_{i+1})]$$
$$=(1+\frac{\vartheta-1}{1+\eta^{-1}})Q^*(s_i,a_i) \tag{28}$$

同理可得

$$Q_1(s_i,a_i)\leqslant(1+(\vartheta-\frac{1}{(1+\varepsilon^{-1})}))Q^*(s_i,a_i) \tag{29}$$

于是，当 $i=0$ 时，满足式(27)。

假设当 $i=l-1$，$l=1,2,\cdots$ 时，仍满足式(27)。那么，当 $i=l$ 时，有

$$Q_1(s_i,a_i)=O(s_i,a_i)+\gamma\max Q_0(s_{i+1},a_{i+1})$$
$$\geqslant O(s_i,a_i)+\gamma(1+\frac{\eta^{l-1}(\vartheta-1)}{(1+\eta)^{l-1}})\max Q^*(s_{i+1},a_{i+1})$$
$$\geqslant(1+\eta^l\frac{\vartheta-1}{(1+\eta)^l})[O(s_i,a_i)+\gamma\max Q^*(s_{i+1},a_{i+1})]$$
$$=(1+\frac{\vartheta-1}{(1+\eta^{-1})^l})Q^*(s_i,a_i) \tag{30}$$

同理可得

$$Q_{l+1}(s_i,a_i)\leqslant(1+\frac{\zeta-1}{(1+\varepsilon^{-1})^l})Q^*(s_i,a_i) \tag{31}$$

因此，对 $\forall i=0,1,2,\cdots$ 满足式(27)。

接下来，证明当 $0\leqslant\vartheta\leqslant1\leqslant\xi<\infty$ 时，满足

$$(1+\frac{\vartheta-1}{(1+\eta^{-1})^i})Q^*(s_i,a_i)\leqslant Q_i(s_i,a_i)\leqslant(1+\frac{\zeta-1}{(1+\eta^{-1})^i})Q^*(s_i,a_i) \tag{32}$$

式(32)的左半部分可以通过式(28)和式(30)来证得。对于右半部分，令 $i=0$，有

$$Q_1(s_i,a_i)=O(s_i,a_i)+\gamma\max Q_0(s_{i+1},a_{i+1})$$
$$\leqslant O(s_i,a_i)+\zeta\gamma\max Q^*(s_{i+1},a_{i+1})+$$
$$\frac{\zeta-1}{1+\eta}(\eta O(s_i,a_i)-\gamma\max Q^*(s_{i+1},a_{i+1}))$$
$$=(1+\frac{\zeta-1}{1+\eta^{-1}})Q^*(s_i,a_i) \tag{33}$$

根据数学归纳法可以得到式(32)的右半部分。因此，当 $0\leqslant\vartheta\leqslant\xi<\infty$，可得对于 $\forall i=0,1,2,\cdots$，评价函数满足式(27)。

最后，根据以上结论，对于任意的常量 ε、η、ϑ 和 ξ，由式(27)~式(33)，当 $i\to\infty$ 时，可得式(20)。证毕。

参考文献：

[1] 朱江, 徐斌阳, 李少谦. 一种基于马尔可夫决策过程的认知无线电网络传输调度方案[J]. 电子与信息学报, 2009, 31(8):2019-2023.
ZHU J, XU B Y, LI S Q. A transmission and scheduling scheme based on Markov decision process in cognitive radio networks [J]. Journal of Electronics & Information Technology, 2009, 31(8):2019-2023.

[2] ZHU J, PENG Z Z, LI F. A transmission and scheduling scheme based on W-learning algorithm in wireless networks[C]//8th International ICST Conference on Communications and Networking in China (CHINACOM). 2013: 85-90.

[3] LI H, HAN Z. Competitive spectrum access in cognitive radio networks: graphical game and learning[C]//Wireless Communications and Networking Conference (WCNC). 2010: 1-6.

[4] 林晓辉, 谭宇, 张俊玲, 等. 无线传输中基于马尔可夫决策的高能效策略[J]. 系统工程与电子技术, 2014, 36(7):1433-1438.
LIN X H, TAN Y, ZHANG J L, et al. MDP-based energy efficient policy for wireless transmission[J]. Systems Engineering and Electronics, 2014, 36(7):1433-1438.

[5] WANG H S, MOAYERI N. Finite-state Markov channel-a useful model for radio communication channels[J]. IEEE Transactions on Vehicular Technology, 1995, 44(1): 163-171.

[6] GAO Q, ZHU G, LIN S, et al. Robust QoS-aware cross-layer design of adaptive modulation transmission on OFDM systems in high-speed railway[J]. IEEE Access, 2016, PP(99): 1.

[7] CHEN X, CHEN W. Delay-optimal probabilistic scheduling for low-complexity wireless links with fixed modulation and coding: a cross-layer design[J]. IEEE Transactions on Vehicular Technology, 2016: 1.

[8] LAU V K N. Performance of variable rate bit interleaved coding for high bandwidth efficiency[C]//The Vehicular Technology Conference. 2000:2054-2058.

[9] CHUNG S T, GOLDSMITH A J. Degrees of freedom in adaptive modulation: a unified view[C]// IEEE Transactions on Communications. 2001:1561-1571.

[10] WEI Q, LIU D, SHI G. A novel dual iterative Q-learning method for optimal battery management in smart residential environments[J]. IEEE Transactions on Industrial Electronics, 2015, 62(4):2509-2518.

[11] NI J, LIU M, REN L, et al. A multiagent Q-learning-based optimal allocation approach for urban water resource management system[J]. IEEE Transactions on Automation Science & Engineering, 2014, 11(1):204-214.

[12] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587):484-489.

[13] WEI C, ZHANG Z, QIAO W, et al. An adaptive network-based reinforcement learning method for MPPT control of PMSG wind energy

conversion systems[J]. IEEE Transactions on Power Electronics, 2016: 1.

[14] KIM T, SUN Z, COOK C, et al. Invited-cross-layer modeling and optimization for electromigration induced reliability[C]// Design Automation Conference. 2016:1-6.

[15] COMSA I S, ZHANG S, AYDIN M. A novel dynamic Q-learning-based scheduler technique for LTE-advanced technologies using neural networks[C]// Conference on Local Computer Networks. 2012:332-335.

[16] TENG T H, TAN A H. Fast reinforcement learning under uncertainties with self-organizing neural networks[C]// IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2015:51-58.

[17] KOBAYASHI T, SHIBUYA T, TANAKA F, et al. Q-learning in continuous state-action space by using a selective desensitization neural network[J]. IEICE Technical Report Neurocomputing, 2011, 111: 119-123.

[18] 周文云. 强化学习维数灾问题解决方法研究[D]. 苏州: 苏州大学, 2009.
ZHOU W Y. Research on the curse of dimensionality in reinforcement learning[D]. Suzhou: Soochow University, 2009.

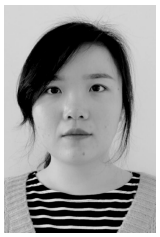
[19] LIU W, LIU N, SUN H, et al. Dispatching algorithm design for elevator group control system with Q-learning based on a recurrent neural network[C]// Control and Decision Conference. 2013:3397-3402.

[20] WEI Q, LEWISF L, SUN Q, et al. Discrete-time deterministic Q-learning: a novel convergence analysis[J]. IEEE transactions on cybernetics, 2016: 1-14.

[21] 李军, 徐玖平. 运筹学:非线性系统优化[M]. 北京: 科学出版社, 2003.
LI J, XU J P. Operations research: nonlinear system optimization[M]. Beijing: Science Press, 2003.

[作者简介]

朱江（1977-），男，湖北荆州人，博士，重庆邮电大学教授，主要研究方向为认知无线电、移动通信、网络安全态势感知。



王婷婷（1993-），女，安徽安庆人，重庆邮电大学硕士生，主要研究方向为网络安全态势感知。

宋永辉（1991-），男，河北邯郸人，重庆邮电大学硕士生，主要研究方向为认知无线电。

刘亚利（1990-），男，河南商丘人，重庆邮电大学硕士生，主要研究方向为认知无线电。