A Transmission and Scheduling Scheme Based on W-learning Algorithm in Wireless Networks

Jiang Zhu
Chongqing Key Lab of Mobile Communications Technology
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
zhujiang@cqupt.edu.cn

Zhenzhen Peng
Chongqing Key Lab of Mobile Communications Technology
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
scpzz07@163.com

Fangwei Li
Chongqing Key Lab of Mobile Communications Technology
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
lifw@cqupt.edu.cn

***Abstract*—In this paper, we propose a transmission and scheduling scheme based on W-learning algorithm used in wireless network. With the introduction of W-learning algorithm, we can transmit packets intelligently in wireless networks, namely, reduce package lose under the premise of energy saving. We offer our own system model based on Markov decision process (MDP), in order to prove the efficiency of our scheme, we compare the simulation transmission results of our scheme with the results of several heuristic algorithms and the approximately optimal result of successive approximation method. A state aggregate method and action set reduction scheme are also used in this paper to reduce the number of system states, downsize the calculate amount of iterative process. We prove and simulate that the state aggregate method and action set reduction scheme will have little influence of the efficiency of algorithms while reduce the calculate amount.**

Keywords-W-learning algorithm; intelligent transfer; Markov decision process

I. INTRODUCTION

Wireless communication technology is fast-developing in recent years, the spectrum resources are getting exiguous day by day, and the scale of communication network is also expanding at the same time, so the communication network will become smarter and smarter in the future. Considering learning algorithm can iteratively calculate with little data information, we decide to use learning algorithm to build our own transmission scheme which can transmit packages intelligently. For years, many works studied the intelligent transmit based on learning algorithm. Ref. [1] build a kind of approximate optimal scheme in energy saving sensor network based on reinforcement learning, with the use of learning algorithm, they can reduce their rely on the system transition probability. Paper [2] provides a cross-layer transmission and scheduling scheme based on Markov decision process (MDP). Paper [3] presents a new method for the study of the competition and cooperation relationships among nodes in a network using a CSMA (Carrier Sense Multiple Access) protocol implemented with an exponential backoff process based on self-organized behavior.

We decide to build our transmission scheme based on W-learning algorithm. In order to show that our scheme works, we provide our own system model, hoping that by using our scheme, we can transmit packages intelligently, improve the entire system's utility. Moreover, using W-learning can reduce the calculate amount than the method of successive approximation. This paper also provides a state aggregate method and action set reduction scheme to reduce the required amount of calculation. In the end of this article, we prove that the approximately optimal get from state aggregate method and action set reduction scheme equals to the optimal in the given conditions.

II. SYSTEM MODEL

Fig. 1 shows the system model, exist a wireless node work as a relay node for other wireless nodes to help them transmit packages. In our system, upper data reach nodes present Poisson distribution with the rate of λ, the corresponding buffer of each nodes will store the data and wait to send in the corresponding wireless channel. The relay node can help only one node to send packages in one frame, so it has to decide which node to transmit and the transmit mode by the buffer state and the channel state of each node. We define the transmission channel state of each node as $C \triangleq \{c_0, c_1, c_2, ..., c_n\}$.

A. FSMC Channel and Adaptive Modulation

We assume that the channel is fast fading channel and the estimate of channel state is accurate. The transmission channel of each node can be modeled as ergodic one-order discrete Markov chain with finite state [4]. The SNR in additive white Gaussian noise Rayleigh channel of k nodes distributed exponentially, we define the probability density as $p_{\mathrm{SNR}}(\gamma)=1/\gamma_0^m \exp(-\gamma/\gamma_0^m)$, $\gamma \geq 0$, γ_0^m is the average signal noise ratio. The channel state of every node is decided by delimit the SNR threshold. For the state c_k the probability of channel in this state is $p_c(\mathrm{c}_k)=\int_{\Gamma_k}^{\Gamma_{k+1}} p_{\mathrm{SNR}}(\gamma)d\gamma$, the transition probability of channel state is:

Identify The paper is supported by the National Natural Science Foundation of China (No.61102062), the Key Project of Chinese Ministry of Education (No. 212145), the Natural Science Foundation of Chongqing Science and Technology Commission (No.CSTC2011jjA1192), the Science and Technology Research Project of Chongqing Education Commission (No.KJ110503), the Doctoral Foundation of CQUPT (No.A2010-11).

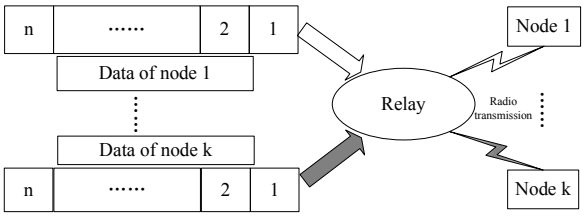

Figure 1 The system model.

$$p_c(\mathrm{c}_k,\mathrm{c}_{k+1})= N(\Gamma_{k+1})T_f / p_c(\mathrm{c}_k), k \in \{0,1,\cdots,K-2\}. \quad (1)$$

$$p_c(\mathrm{c}_k,\mathrm{c}_{k-1})= N(\Gamma_k)T_f / p_c(\mathrm{c}_k), k \in \{1,2,\cdots,K-1\}. \quad (2)$$

$N(\Gamma_k)=\sqrt{2\pi\Gamma_k/\gamma_0^m}\, f_d^m \exp(-\Gamma_k/\gamma_0^m)$, f_d^m is the biggest Doppler frequency shift. We can limit the BER boundary [5] of BPSK, QPSK, 8PSK, 16PSK... For $u_1(m=1, BPSK)$:

$$p_{\mathrm{BER}}(\mathrm{c}_k,u_1) \le 0.5\mathrm{erfc}\left(\sqrt{\Gamma_k P(\mathrm{c}_k,u_1)/WN_0}\right). \quad (3)$$

For u_j, $j \in \{2,3,4\}$ (that is m=2,3,4,$\cdots$), BER $\le 10^{-3}$:

$$p_{\mathrm{BER}}(\mathrm{c}_k,u_j) \le 0.2\exp\left(-1.6\Gamma_k P(c_k,u_j)/WN_0(2^j-1)\right). \quad (4)$$

WN_0 is the power of noise. We can get the minimum power that meets the specific BER in different channel states and transmit mode. We assume that the channel state transition probabilities of nodes in the system are independent of each other [6]. Let $p_{c_i}(c_i,c_i')$ indicates the channel state transition probability of node i then the system channel state transition probability is:

$$p_{ch}(c,c') = p_{c_1}(c_1,c_1')\times p_{c_2}(c_2,c_2')\times\ldots\times p_{c_k}(c_k,c_k'). \quad (5)$$

B. Buffer State

q_i is the data size that reach node i in every frame (the frame size is T_s), the data size that reach buffer in every frame appear Poisson distribution of rate λ_i . So q_i can get by:

$$p_{q_i}(q_i) = \exp(-\lambda_i T_s)(\lambda_i T_s)^{q_i}/q_i!. \quad (6)$$

So if the packages size of buffer is b_i in a frame, the relay node send $a_i(a_i = V\bullet 2^{m_i})$ packet of data. Then the package size of buffer in next frame is:

$$b_i = \min\{b_i - a_i + q_i, L_i\}. \quad (7)$$

Considering that q_i and a_i follow Markov Process, we can know that the state of data size of buffer also follow Markov Process, $p_{b_i}(b_i,b_i')$ is the transition probability of data size in buffer. For the whole system, buffer state is composed of the buffer state of each node which is independent of each other. Then the buffer state transition probability of the whole system is:

$$p_{bu}(b,b') = p_{b_1}(b_1,b_1')\times p_{b_2}(b_2,b_2')\times\ldots\times p_{b_k}(b_k,b_k'). \quad (8)$$

III. MDP AND THE TRANSMISSION SCHEDULES

We can model the transmission schedule as a MDP (Markov decision process) . There are five elements in MDP $\{S, A, P, R, C\}$. S is the state space, A , P , R and C means the action set, state transfer probability matrix, profit and cost respectively.

A. State Space and the State Transfer Probability Matrix

The state space can buffer defined as $S \triangleq B \otimes C$, so the states of buffer and channel states make up the system states. In a frame, if the state is $s \triangleq \{b,c\} \in S$, in the next frame, the state is $s' \triangleq \{b',c'\} \in S$, then the state transfer probability is $p_s(s,s'/a) = p_b(b,b')\times p_c(c,c')$. As there are k nodes in the system, the state transfer probability of every node is $p_s(s,s'/a)$, and they are independent of each other. So the state transfer probability of the whole system can be expressed as (9)*.

B. Action Set

The action set is $A \triangleq \{a_0,a_1,...,a_N\}$. If we choose action a_i , then the relay node will send packages for node i with transmit mode m_i . Transmission mode $m(m=1,2,3,...)$ correspond with BPSK, QPSK, 8PSK, 16PSK… If the packages transmission rate is V . Then the data size we can send is $V\bullet 2^{m_I}$. In the transmission schedule of our paper, the action we hope to get is the one that can get the highest system benefit, so the transmit choice can be described as:

$$r_i = \max(R(s_i,A)), a_i \in A. \quad (10)$$

So for the whole system, transmit action can be described as：

$$a_i \triangleq \{0,0,...,m_i,...,0\}. \quad (11)$$

When the relay node decide to transmit packages for node i , then the transmit mode of node i is m_i , the relay node choose not to transmit packages for the other node, so the transmit mode of other node is 0 .In our system, for state s_i , if we know the BER , Then we can get the minimum power $P_t(s_i,a_i)$ that we need for action a_i by the formula given above [5].

C. Profit and Cost

We define the cost of our system as the power cost by packages transmission and the pressure value of buffer. In state

* $p_s(s,s'/a) = p_{s_1}(s_1,s_1'/a_1)\times p_{s_2}(s_2,s_2'/a_2)\times\ldots\times p_{s_k}(s_k,s_k'/a_k) = \prod_{i=1}^{k} p_{s_i}(s_i,s_i'/a_i) = \prod_{i=1}^{k} p_{b_i}(b_i,b_i')\times\prod_{i=1}^{k} p_{c_i}(c_i,c_i')$. (9)

s_i, the power used by action a_i is $P_t(s_i, a_i)$. The pressure value of buffer is $\exp(gamma \times b_i)$. Then we can define the profit of node i as:

$$R_i = \frac{V \times 2^{m_i}}{P_t(s_i, a_i) \times \exp(gamma \times b_i)}. \quad (12)$$

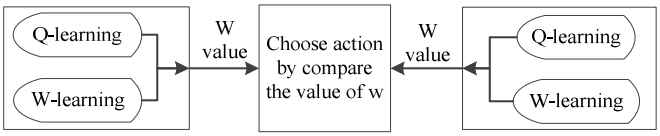

Figure 2 The procedure of W-learning algorithm.

gamma can be used to control the proportion of importance. The system profit can be expressed as:

$$R = R_1 + R_2 + ... + R_k. \quad (13)$$

IV. Algorithm Introduction

A. W-learning Algorithm

Our transmission scheme is built on the base of W-learning algorithm [7]. The procedure of W-learning algorithm is showed in Fig. 2: nodes can learn the w value by themselves, and then the system can make its own decision by compare the w values of nodes. We can see the w value as the lose we get if we don't take the action. The w value can get by:

$$Q_{i+1}(s_i, a) = (1 - \partial) Q_i(s_i, a) + \partial(r_i + \gamma \max Q_i(s_{i+1}, a')). \quad (14)$$

$$\mathrm{W}_{i+1}(s_i) = (1 - \partial_w) W_i(s_i) + \partial_w (Q_i(x, a_i) - (r_i + \gamma \max_{b \in A} Q_i(s_{i+1}, b))). \quad (15)$$

In Q-leaning algorithm, r_i is the profit of iteration, ∂ is the learning factor, define ∂=1/(1+visit) , the value of ∂ is becoming smaller when the visit time grows, in this way, the learning algorithm can convergence eventually. However, in practical, we often cannot wait to begin W-learning until Q-value is learnt, so in order to begin learning W-value with Q-value, we learn W-value by: formula (16)*.

We use $(1 - \partial_Q)^{\omega}$ to control the influence of Q-value on W-value, ∂_Q(∂_Q=1/(1+visit)) is the learning factor of Q-learning, ω can control the convergence speed of W-learning, the smaller ω is the slower convergence speed is.

B. Method of Successive Approximation

In order to evaluate the performance of our scheme, we decide to get the optimal choice by Successive approximation method. Because Successive approximation method cannot get the optimal choice directly, it can only get the near-optimal choice by iteration. Considering the calculate amount, we need to reduce the system state as much as possible. The iteration equation of successive approximation method is [8]:

$$V_{n+1} = TV_n = \max_{f \in F} \{r(f) + \beta P(f) V_n\} = r(f_n) + \beta P(f_n) V_n. \quad (17)$$

C. Calculate Amount

In theory, if the iteration is enough, then we can get the optimal choice by successive approximation method, however, the calculate amount is huge. If the upper data arrival rates of nodes are the same. For successive approximation, the number of states is $|S_1| = ((l+1) \bullet c)^k$, c is the number of channel states, k is the number of nodes. If we let N_1 be the iteration time of successive approximation, A_1 be the number of actions, then the calculate amount is $|T_1| = ((l+1) \times c)^k \times A_1 \times N_1$. While the state number of W-learning algorithm is $|S_2| = (l+1) \times c \times k$, the calculate amount is $|T_2| = (l+1) \times c \times k \times A_2 \times N_2$. N_2 is the iteration time of W-learning algorithm, A_2 is the number of actions. It's easy to know that N_1 is bigger than N_2. In our simulation, we suppose there are 2 nodes, the length of buffer is 5, the successive approximation method channel states are 576, while there are only 24 in W-learning algorithm, A_1 is also lager than A_2 as the action set of successive approximation method has to group the action set of all nodes, so the iteration of Successive approximation method is much larger than W-learning algorithm. Using W-learning algorithm can solve the curse of dimensionality.

V. Solve Markov Decision Process

A. State Aggregate Method

When we begin the iterative operation of successive approximation and W-learning algorithm, if the scale of state space is very large, then the calculate amount will consume huge memory space, slow down the convergence speed of algorithm. If the scale of state space is huge enough, the algorithm may even can not be able to get converged [9]. So in order to reduce the calculate amount, we propose state aggregate method to reduce the scale of state.

The principle of state aggregate method is: if the state of channel is very bad, then there is no need to know the concrete state, the relay node chooses not to transmit packages for it at all. For nodes in the system, if the corresponding channel state is rather bad, then it will take consume a lot of power, however, as the channel state is rather bad, it is rather easy that the transmission will fail. So choose not to transmit packages for the node which channel state is very bad is the best choice. In this way, we can aggregate state, for those states which channel state is $c = 0$, we can get them together [10]. For all the states in the aggregated state, relay node chooses not to transmit packages for them. So for the node, if the channel state is $C \triangleq \{c_1 = 0, c_2 = 1, c_3 = 2\}$, the buffer length is $l = 2$, then the state space matrix is:

$$* \ \mathrm{W}_{i+1}(s_i) = (1 - \partial_w) W_i(s_i) + \partial_w (1 - \partial_Q)^{\omega} (Q_i(x, a_i) - (r_i + \gamma \max_{b \in A} Q_i(s_{i+1}, b))). \quad (16)$$

$$S=\begin{Bmatrix}(0,0),(0,1),(0,2),\\(1,0),(1,1),(1,2),\\(2,0),(2,1),(2,2)\end{Bmatrix}$$

After the state aggregate the state can be changed as: $s_1=\{(0,0),(0,1),(0,2)\}$, $s_2=\{(1,0)\}$, $s_3=\{(1,1)\}$, $s_4=\{(1,2)\}$, $s_5=\{(2,0)\}$, $s_6=\{(2,1)\}$, $s_7=\{(2,2)\}$. In this way, the state scale of successive approximation and W-learning algorithm can be reduced. We can prove that use the state aggregate method to reduce the calculate amount will not influence the performance of algorithm.

Lemma 1: If the state space S is divided into W subspace which have no intersection $S_1,S_2\cdots,S_W$, for action a, if any $s\in S_W$, have $Q(s,a)=\delta(w,a)$, then we can get the optimal strategy by value iteration algorithm based on state aggregate method [9].

Theorem 1: If the state transferring probability of each node is independent and identical, then the near-optimal strategy got from state aggregate method equal to the optimal strategy.

The proof of theorem 1 is: In the solution process of optimal strategy, the iteration equations of successive approximation and W-learning algorithm are formula (17) and (14). As we aggregate the states which channel states are very bad, then $r(f_n)$ and r_i are zero. So for all the state in the aggregate state $S_w=\{(s_1,s_2,...,s_k)\}$, have $V_n(s_k,a)=0$, $Q(s_k,a)=0$, $s_k\in S_w$ from lemma 1, we can know that the strategy we get from value iteration algorithm based on state aggregate method is the optimal strategy.

B. Action Set Reduction Scheme

In the iteration operation, we can also reduce the calculate amount by using action set reduction scheme. In our system, every state has its corresponding action set, action set reduction scheme can reduce the action set with little influence on the performance of algorithms.

Lemma 2: A MDP aimed at minimize the negative earning, in state S, if $i(s,a')<i(s,a)$ and $p_s(s,s'/a)=p_s(s,s'/a')$, $\forall s'\in S$, s' is the next state, and then action a' is better than action a [2].

We can get theorem 2 from lemma 2.

Theorem 2: The strategy get from action set reduction scheme equals to the optimal strategy.

The basic idea of theorem 2 is: If the channel state is rather bad, then the relay node cannot send package successfully in this channel, the transmission will consume huge energy, however it will do no good to the store status of buffer. So if the channel state of node is very bad, then the relay node chooses not to transmit packages, in this way, we can reduce the action set.

C. Space Compression Ratio

In the above figures, the performance of successive approximation and W-learning algorithm have been little influenced by state aggregate method and action set reduction scheme, the line of successive approximation 1, W-learning algorithm 1 and successive approximation, W-learning algorithm overlap basically. We can see that according to the analysis above, successive approximation has $|S_1|=576$ kinds of state, the size of state behavior matrix that has to store is $|T_1|=4032$; the successive approximation 1 has $|S_1'|=361$ kinds of states, the size of state behavior matrix that has to store is $|T_1'|=2377$, W-learning algorithm has $|S_2|=24$ kinds of states, the size of state behavior matrix that has to store is $|T_2|=96$; W-learning algorithm 1 has $|S_2'|=19$ kinds of states, the size of state behavior matrix that has to be store is $|T_2'|=55$. So with the help of state aggregate method and action set reduction scheme, the space compression ratio of successive approximation is 41%, and the ratio of W-learning algorithm is 43%.

VI. Algorithms Compare and Simulation Analysis

In order to evaluate the performance of W-learning algorithm in our system, we define three kinds of heuristic algorithm to compare with our scheme, the optimal strategy is get by successive approximation. Besides, we can see from the simulate result that state aggregate method and action set reduction scheme can reduce the calculate amount with little influence on the performance of algorithms (for simplicity, we call successive approximation SA, W-learning algorithm W-L, heuristic algorithm 1 HA1, heuristic algorithm 2 HA2, heuristic algorithm 3 HA3, the successive approximation and W-learning algorithm after using state aggregate method and action set reduction scheme SA1 and W-L1).

Definition 1: Heuristic algorithm 1 hopes to transmit as many packages as possible, that is the relay node choose to transmit for the node which has more packages in the corresponding buffer.

Definition 2: Heuristic algorithm 2 hopes to transmit packages with lowest energy cost, choose to transmit for the node which has the best channel state.

Definition 3: Heuristic algorithm 3 chooses to transmit for the node which has more packages in the corresponding buffer, the transmit mode is get randomly.

Fig. 3, 4 and 5 show the transmission results of three kinds of heuristic algorithm, successive approximation, W-learning algorithm and the successive approximation, W-learning algorithm after using state aggregate method and action set reduction scheme. As the calculate amount of successive approximation is too huge to get the optimal strategy when there are a lot of states. So we define that there are two nodes in the system, the length of their corresponding buffers are both 5, the upper data reach buffer in the same way. The following pictures show the package lose, energy cost and average utility of those algorithms.

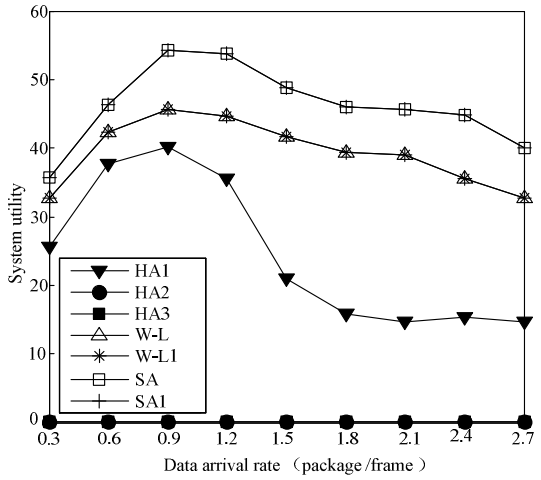

Figure 3 The system utility under different data packet arrival rate.

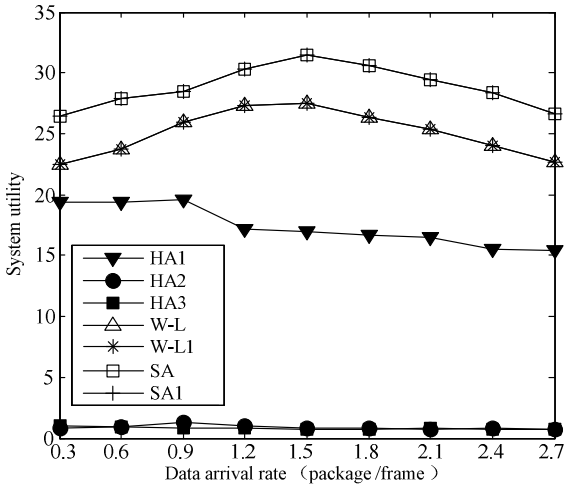

Figure 6 The system utility under different data packet arrival rate.

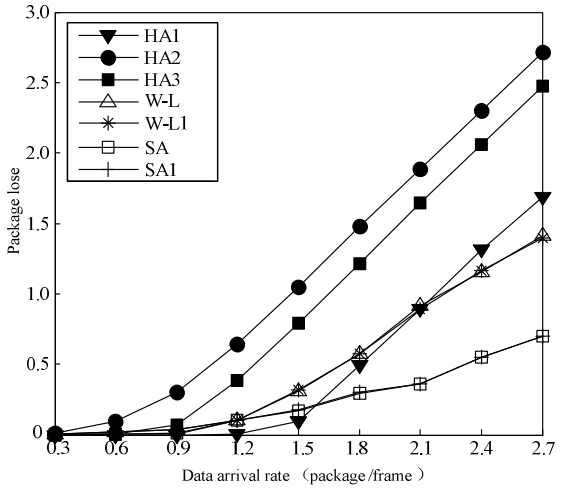

Figure 4 The average package lose under different data packet arrival rate.

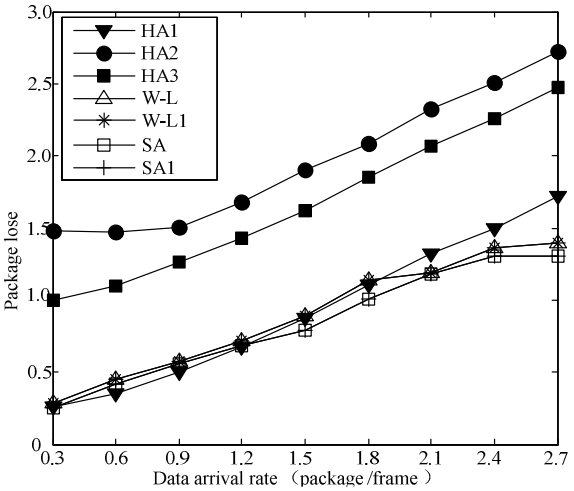

Figure 7 The package lose under different data packet arrival rate.

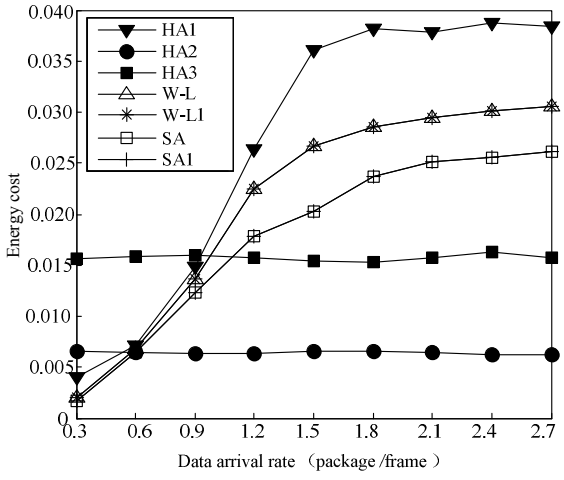

Figure 5 The energy cost under different data packet arrival rate.

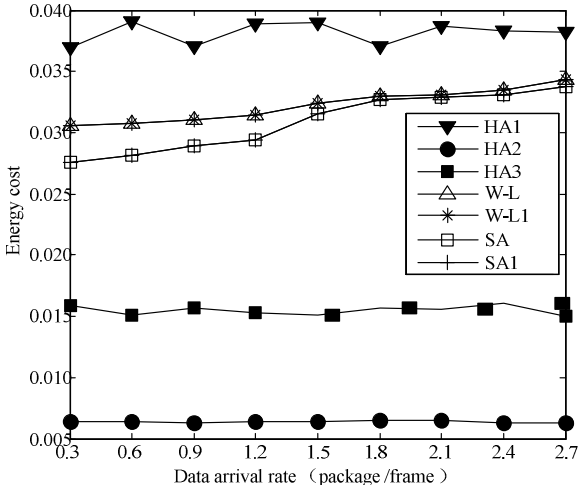

Figure 8 The energy cost under different data packet arrival rate.

From the system utility of algorithm showed in Fig. 3, we can know that W-learning algorithm can get relatively good

system utility. In Fig. 4, W-learning algorithm, successive approximation and heuristic algorithm 1 can transmit packages better than heuristic algorithm 2and 3.When the data arrival rate is low, heuristic algorithm 1 lead to the least package lose, however, with the increase of data arrival rate, W-learning algorithm and successive approximation begin to transmit with the transmit mode that can send the maximum packages, then because our scheme can transmit intelligently, the package lose become smaller than heuristic algorithm1. What's more, package lose in other ways can also show the buffer length, the system with less package lose must have less buffer length than the system with high package lose. So we can see that our scheme can have relatively short buffer length, which also means that the time delay of our scheme is relatively short. Fig. 5 shows the energy cost of algorithms, we can see that heuristic algorithm 2 cost the lowest energy, heuristic algorithm 1 cost the most energy. W-learning algorithm cost less energy than heuristic algorithm 1, successive approximation shows the energy cost of optimal strategy.

When the upper data arrival rate of the two nodes are mutual independent, we can compare the package lose and energy cost of those algorithms. We fix one of the upper data arrival rate as 2.7 data packages in every frame, Fig. 6, 7 and 8shows the system utility, package lose and energy cost of algorithms. We can see that, compared with other algorithms, our scheme can transmit intelligently with energy saving, with the increase of the other node's data arrival rate.

VII. CONCLUSION

In this paper, our scheme introduce W-learning algorithm into wireless communication network, by using our scheme, we can transmit packages intelligently. This paper builds our own system model based on Markov Decision Process. Because of the introduction of W-learning algorithm, we can reduce the calculate amount, save memory space, accelerate the convergence. We also propose and prove that the state aggregate method and action set reduction scheme can reduce the calculate amount with little influence on the performance of algorithm, in the future, we will expand the system to distributed wireless network to study the intelligent transmission with learning algorithm.

REFERENCES

[1] C. Pandana and K. R. Liu, "Near-optimal reinforcement learning framework for energy-aware sensor communications," *Selected Areas in Communications, IEEE Journal on,* vol. 23, pp. 788-797, 2005.

[2] J. Zhu, B. Y. Xu, S. Q. Li, "Optimal and Suboptimal Access and Transmission Polices for Dynamic Spectrum Access over Fading Channels in Cognitive Radio Networks," *Chinese Journal of Electronics,* vol. 17, pp. 726-732, 2008.

[3] Z. Shi, C. C. Beard, and K. Mitchell, "Competition, cooperation, and optimization in multi-hop CSMA networks with correlated traffic," *INTERNATIONAL JOURNAL OF NEXT-GENERATION COMPUTING,* vol. 3, 2012.

[4] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," *Vehicular Technology, IEEE Transactions on,* vol. 44, pp. 163-171, 1995.

[5] S. T. Chung and A. J. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," *Communications, IEEE Transactions on,* vol. 49, pp. 1561-1571, 2001.

[6] A. K. Karmokar, D. V. Djonin, and V. K. Bhargava, "POMDP-based coding rate adaptation for type-I hybrid ARQ systems over fading channels with memory," *Wireless Communications, IEEE Transactions on,* vol. 5, pp. 3512-3523, 2006

[7] M. Humphrys, "W-learning: a simple RL-based society of mind," *Univ. Cambridge, Computer Laboratory, Tech. Rep,* vol. 362, 1996.

[8] D. J. White, "Dynamic programming, Markov chains, and the method of successive approximations," *Journal of Mathematical Analysis and Applications,* vol. 6, pp. 373-376, 1963.

[9] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming: An overview," in *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, 1995, pp. 560-564.

[10] Lin Chuang. Performance Evaluation of Computer Networks and Computer Systems, First Edition, Beijing: Tsinghua University Press, 2001.