Wireless Networks (2020) 26:4507–4516
https://doi.org/10.1007/s11276-020-02353-9

Joint optimization of power control and time slot allocation for wireless body area networks via deep reinforcement learning

Lili Wang[1] · Ge Zhang[1] · Jun Li[1] · Gaoshang Lin[1]

Published online: 10 May 2020

Abstract
E-healthcare system based on wireless body area network (WBAN) promises to produce potential benefits in health-care industry. A major issue of such an on-body networked system is the energy efficiency, that is, how to improve the reliability and effectiveness of physiological data transmission with the energy constraints of tiny wireless sensors. Motivated by this, we consider an individual WBAN scenario, focusing on finding an adaptive time slot allocation and power control scheme to maximize the average energy efficiency for implementing the task of health monitoring. We formulate the maximization problem with latency and sensors' energy budget constraints as a markov decision process (MDP). As a solution, we propose a deep reinforcement learning-based scheme to make a sequence decision for the MDP, which jointly optimizes power control and slot allocation. Simulation results show that the proposed scheme is energy efficient and has a good convergence.

Keywords WBAN · Energy efficiency · Markov decision process · Deep reinforcement learning

1 Introduction

In recent years, wireless body area network (WBAN) has attracted wide attention of scholars due to its efficiency and convenience in health-care industry. As a branch of the wireless sensor network (WSN), a typical WBAN system consists of multiple medical sensors worn or implanted on the body that monitor various physiology signals, as well as a coordinator which receive data from sensors by short-range on-body wireless communications for further data processing and analysis [1]. Compared with conventional WSN, WBAN shows many similarities, such as energy restriction and self-organizing network. Meanwhile, some more prominent characteristics must be noted, including strict latency and power constraint, reliable data transmission of vital signs and time-varying wireless channel [2].

Frequent body movement and various posture changes induce path loss and possible severe shadowing effects. The actual measurements show that the deep fading duration of WBAN channel usually ranges from 10 to 400 ms [3], which greatly exceeds traditional WSNs. Moreover, in order to reduce the electromagnetic radiation to human body, the transmission power of sensors must be strictly limited and sensitive to channel changes [4]. IEEE 802.15.6 stipulates that the maximum transmission power of nodes in WBAN is 1 MW (0 dBm). Meanwhile, the suitable value for dedicated medical nodes is lower than 0.1 MW (-10 dBm) [5]. In view of the above, it is a challenge to solve the power control problem for building a WBAN, because reliable communication usually means higher transmission power while the opposite is true for human safety and energy restriction.

Besides, the time-varying characteristic of WBAN channel also has a significant impact on time slot allocation. A static Time Division Multiple Access (TDMA) is a simple but mature mechanism for slot allocation [6]. However, a conventional TDMA scheme for WBAN may cause energy waste, as the time slot may be allocated to a channel with high path loss leading to a great quantity of data retransmissions [7]. Based on the analysis, it is more reasonable for a sensor to temporarily store its current data in local buffer when channel state is unsatisfactory. However, considering the long duration of depth attenuation in

✉ Lili Wang
liliwang@njust.edu.cn

[1] School of Automation, Nanjing University of Science and Technology, Nanjing 210094, Jiang Su, China

WBAN [3], the queue length and the transmission latency will increase substantially if channel state is unsatisfactory. High latency is intolerable for many health monitoring businesses. Hence, it is necessary to propose a slot allocation strategy to minimize energy consumption and transmission delay.

Plenty of solutions to energy efficiency issues have been proposed, e.g. power control and resources allocation for WBAN. The authors in [8] propose a power control protocol via link state estimation in WBAN. The studies in [9, 10] measure and study the variation of on-body channel and indicate that the change of path loss in WBAN can be described as a Markov process. The study in [11] investigates the energy-efficient scheduling problem with Lyapunov function and proposes an adaptive scheduling mechanism, however the solution of Lyapunov function needs strict assumptions which may not exist in practical application. In [12], the sampling problem of on-body sensors is considered as a Markov decision process (MDP) and a viterbi-based context aware mobile sensing (VCAMS) mechanism is proposed by solving the corresponding dynamic programming (DP) of the MDP formulations, which is accurate but has large time complexity. Similar methods are used in other slot allocation systems. The authors in [13] consider unmanned aerial vehicle as a mobile edge server and formulate the maximizing throughput of task offloading problem as a semi-MDP (SMDP), then solve the SMDP by some Reinforcement Learning (RL)based algorithms, which has different system and model with this paper but provides us with enlightenment in solving RL problem. The study in [14] proposes a sensor access control scheme for WBANs based on reinforcement learning that enables the coordinator to choose the access time and transmit power of the sensors based on the state that consists of the signal-to-interference plus noise ratio, the transmission priority, the battery level, and the transmission delay of the sensors. Compared with the benchmark schemes, the proposed scheme increases the overall utility of the sensors.

Different from the existing works, we propose a scheme of adaptive slot allocation and power control based on TDMA mechanism for WBAN. Precisely, we study the problem of sensed data transmission between the medical sensors and the coordinator with the goal of maximizing the average energy efficiency of the network, while satisfying the constraints of QoS. In order to address this problem, time slot allocation and power control are jointly optimized. Firstly, we consider the problem as a model-free MDP, then convert the optimization problem into maximizing the average reward of MDP during processing period. Next, we propose the deep reinforcement learning (DRL)-based scheme to find the optimal strategy, which uses Reinforcement Learning (RL) and Neural Network (NN) together to deal with the explosion of state space and improves some extent performance. Simulation results demonstrate that the proposed scheme can maximize the average reward to increase the energy efficiency, and can converge quickly as well.

The contributions of this paper are as follows:

- We formulate the optimization problem of WBAN as an MDP with combination state spaces including channel state, queue state and energy state, which takes the various factors into account to solve joint multi-state Markov problem efficiently.
- The proposed DRL-based scheme has two NNs with same structure but different parameters, which makes the proposed scheme efficient and converge quickly.

2 System model

We think about a single WBAN system with star topology, which is consisted of N sensors and one coordinator as Fig. 1. To avoid transmission interference among sensors, a periodic TDMA protocol is used. That is to say, the time horizon is separated into multiple discrete equal-length slots, indexed by $t \in \{0, 1, 2, \ldots, T\}$, and the coordinator can communicate with only one sensor in a slot.

Because of the heterogeneity of sensors, various data rates are usually set up for different sensors in a practical WBAN, with $a_i(t)$, $i \in \{1, 2, \ldots, N\}$ representing the data rate of the ith node in slot t. Then we define $B_i(t)$ as the

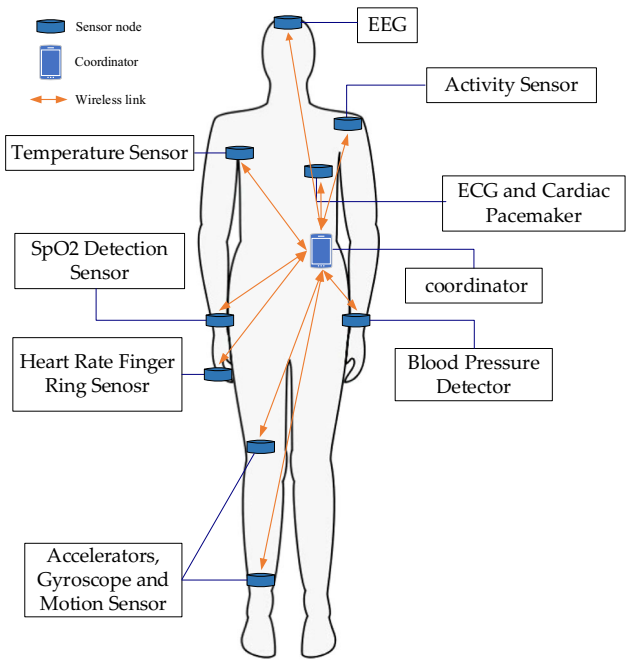

Fig. 1 System model of a WBAN

quantity of packets that are sent from the ith sensor to the coordinator with transmission power $p_i^{tx}(t)$ in slot t. Conversely, unsent packets are temporarily stored in the local buffer whose capacity denoted by $Q_{\max}$.Thus, at the beginning of slot t, the quantity of packets waiting to be transmitted in the ith sensor's buffer can be denoted by $l_i(t) \in [0, Q_{\max}]$.Since the MDP is adopted in the paper, we introduce Markov variable $c_i(t)$ to represent the channel path loss of the ith sensor in slot t [10].

To enable the data uploading for each sensor, every slot is divided into *Decision Phase* and *Transmission Phase* as Fig. 2 illustrated. In the *Decision Phase*, the coordinator chooses a sensor to access to the wireless channel according to current path loss and the quantity of data in each sensor's buffer. Then, the chosen sensor is adjusted to an appropriate transmission power based on the channel state before entering the *Transmission Phase*. Here the QoS of the system includes the system average delay $D(t)$ and the fairness index $J(t)$. A drawback of the decision-making mechanism is that the coordinator is likely to choose sensors with small path loss, which will lead to a massive data accumulation in the other sensors, and results in a larger average delay and lower system fairness. Hence, it is necessary to satisfy the QoS, i.e. $D(t) \leq D_{\max}$,$J(t) \geq J_{\min}$,where $D_{\max}$ and $J_{\min}$ denote maximum tolerant system delay and minimum fairness index respectively. Recall that too small transmission power makes a low signal-to-noise ratio which may cause transmission failure whereas too large transmission power causes energy waste and compromises user's security. So we assume the current transmission power $p_i^{tx}(t)$ can be changed within a fixed range $P_{\min} \leq p_i^{tx}(t) \leq P_{\max}$. In addition, our optimization objective is the average energy efficiency $\bar{\eta}$ which is defined as the number of data packets transmitted with unit energy.

3 Problem formulation

The problem of maximizing average energy efficiency subjected to the QoS of the WBAN can be formulated as an MDP. So we can achieve our goal by maximizing average reward in one episode of MDP. In the following of this section, we define the state space, action space, reward function of MDP and the objective function of this system model.

3.1 State space

The set of state is $S = \{s(t)|t = 0, 1, 2, \ldots, T\}$ for each sensor. For the ith sensor, its state can be indicated as a 3 tuples $s_i(t) = \{c_i(t), l_i(t), e_i(t)\}$, the three elements are described in detail below.

(1) Path loss state $c_i(t)$ represents path loss from the current sensor to the coordinator. The path loss for WBAN is affected by the distance between receiver and transmitter according to Friis formula in free space. In addition, the influence of the environment around the body and the shadowing phenomenon also bring certain degree of signal loss. For the ease of exposition, the path loss is regarded as constant within one slot considering the slot duration is quite short. Further, we model the path loss as a finite state Markov variable according to [10]. It can be quantified to Z values as

$$c_i(t) \in \{C_1, C_2, \ldots C_Z\}, \quad i \in \{1, 2, \ldots, N\} \tag{1}$$

(2) Data quantity state $l_i(t)$ means the number of unsent data packets in the ith sensor's buffer. Define $x_i(t) \in \{0, 1\}$ as a binary variable which equals 1 if the ith node is selected and 0 if it is not selected. The data quantity state at the beginning of slot t + 1 is

$$l_i(t+1) = l_i(t) - x_i(t)B_i(t) + a_i(t) \tag{2}$$

where the initial state $l_i(0)$ is 0 and $B_i(t)$ represents the quantity of data packets transmitted successfully which can be calculated by the following formula derived from [11],

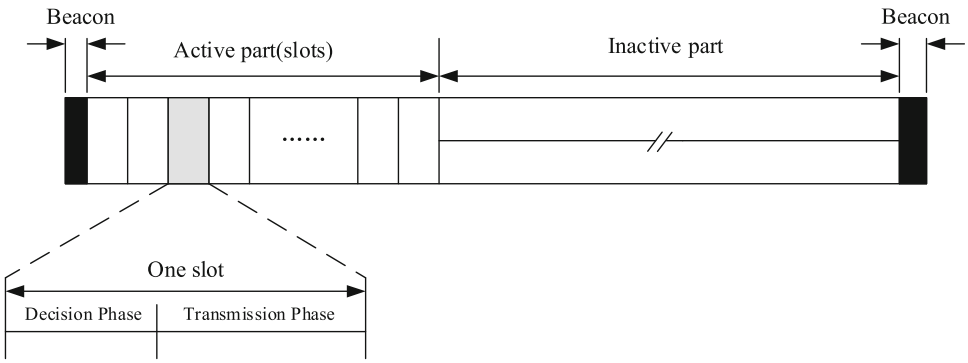

Fig. 2 Structure of a TDMA frame

$$B_i(t)=\left(1-\frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{p_i^{tx}(t)}{p_n c_i(t)}}\right)\right)^K\frac{R}{K} \tag{3}$$

where p_n is the noise power, K (bit) is the length of a packet, R is the bit rate and $\mathrm{erfc}(\cdot)$ is the complementary error function.

(3) Energy state $e_i(t)$ represents total energy of the ith node consumed from slot *0* to slot t, namely

$$e_i(t)=\sum_t x_i(t)p_i^{tx}(t)\tau \tag{4}$$

where τ is the duration of one slot.

3.2 Action space

There are N sensors can be selected in a slot and each sensor has M transmission power modes, so the action space A which contains $N * M$ elements in slot t can be described as

$$A=\left\{d_{i,j}(t)|i=1,2,\ldots,N,\quad j=1,2,\ldots,M\right\} \tag{5}$$

where $d_{i,j}(t)$ indicates the ith sensor is selected to access the channel with the jth power mode.

3.3 Reward function

The goal of the MDP is finding the best sequence decision $d_{i,j}(t)$ to maximize average reward. In this paper, our objective is energy efficiency $\eta(t)$ and main constraints are system delay $D(t)$ and fairness index $J(t)$, the gains from the three elements are described in detail as follows.

(1) Energy efficiency The optimization object in this paper is energy efficiency $\eta(t)$ which is defined as

$$\eta(t)=\sum_{i=1}^{N}\frac{x_i(t)B_i(t)}{p_i^{tx}(t)\tau} \tag{6}$$

Higher energy efficiency means greater reward, so the gain $U(\eta(t))$ can be given by

$$U(\eta(t))=\frac{\varepsilon_1}{1+\exp\left(\frac{\varepsilon_2}{\eta(t)}\right)} \tag{7}$$

It is a variant of the sigmoid function, which represents the corresponding reward that can be obtained with the change of energy efficiency. ε_1 and ε_2 are scaling parameter which adjust the order of magnitude of the reward, and we choose ε_1=0.92 and ε_2=10^7 in this paper.

(2) System delay According to Little's Law [15], the average delay of the system is dominated by the quantity of data packets and the effective data rate. Thus, it can be modeled as

$$D(t)=\frac{\sum_{i=1}^{N}l_i(t+1)}{\sum_{i=1}^{N}\left(a_i-x_i(t)B_i(t)\right)} \tag{8}$$

WBAN is a typical delay-sensitive network whose timeliness of data transmission is of great significance to the health-care applications, so the delay gain is a normal value if it meets delay constraint whereas the gain is set to a very small value if it exceeds permissible delay, namely

$$\lambda(D(t))=\begin{cases}\beta_1, & D(t)\le D_{\max}\\ \beta_2, & D(t)>D_{\max}\end{cases} \tag{9}$$

where β_1 and β_2 is the value of gain. In order to obtain satisfactory convergence, we choose β_1= 1 and β_2=0.001 based on a series of simulations.

(3) System fairness index The fairness refers to relative fairness depending on data rate, which means the sensor with large data rate requirement has greater probability to be selected [3]. Considering the data transmission during an observation window containing several consecutive slots Δt, the ratio of collected data and transferred data can be calculated as

$$\zeta_i(t)=\frac{F_i(t)}{G_i(t)}=\frac{\sum_{t=t-\Delta t}^{t}x_i(t)B_i(t)}{\sum_{t=t-\Delta t}^{t}a_i(t)} \tag{10}$$

According to [16], Jain index $J(x_1,x_2,\ldots,x_n)$ defined by R. Jain is often used as evaluation criterion of fairness in communication systems

$$J(x_1,x_2,\ldots,x_n)=\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n\cdot\sum_{i=1}^{n}x_i^2} \tag{11}$$

where n is the number of users, and x_i is the throughput of the ith user. Replace the x_i in original definition with the ratio of collected data and transferred data, the Jain fairness index $J(t)$ is updated as

$$J(t)=\frac{\left(\sum_{i=1}^{N}\zeta_i(t)\right)^2}{N\sum_{i=1}^{N}\zeta_i^2(t)} \tag{12}$$

The gain of $J(t)$ can be obtained as

$$W(J(t))=\frac{\varepsilon_3}{1+\exp(J_{\min}-J(t))} \tag{13}$$

where ε_3 is scaling parameter. In order to obtain satisfactory convergence, we choose ε_3=0.9 in this paper based on a series of simulations.

Converting the considered indicators into a unified order of magnitude by the sigmoid function, we can combine the gains of the above three parts. The total reward in slot t is calculated as a self-defined function

$$R(t)=\lambda(D(t))(U(\eta(t))+W(J(t))) \tag{14}$$

3.4 The formulation of objective function

The purpose of the paper is to find a slot allocation and power control scheme to improve average energy efficiency with constrains of delay and fairness. The average energy efficiency is a long-term expectation of energy efficiency and can be given by

$$\bar{\eta}=\mathrm{E}(\eta(t)) \tag{15}$$

where $\mathrm{E}(\cdot)$ is expectation function.

The initial problem can be formulated as

$$\begin{aligned} &\max_{x_i(t),p_i^{tx}(t)} \bar{\eta}, \\ s.t.\ &D(t) \le D_{\max}, \\ &J(t) \ge J_{\min}, \\ P_{\min} \le\ &p_i^{tx}(t) \le P_{\max}, \\ &e_i(t) \le E_i, \\ &l_i(t) \le Q_{\max} \end{aligned} \tag{16}$$

where E_i denote the total energy of the ith sensor.

Once we further model the problem as an MDP, the average reward $H(t)$ is expected to represent the average energy efficiency. It is defined as the ratio of the accumulated rewards to the number of slots T, namely

$$H(t) = \frac{\sum_{t=1}^{T} R(t)}{T} \tag{17}$$

Then we formulate the maximization energy efficiency problem into making the sequence decision which can obtain maximal $H(t)$ in an episode, that is to say the objective function in (16) can be transformed as

$$\max_{x_i(t),p_i^{tx}(t)} H(t) \tag{18}$$

4 DRL-based scheme of maximizing system energy efficiency

Reinforcement learning (RL) is a framework that obtains target reward through interaction between agents and environment. In this section, the conventional RL scheme is expounded to get optimal state-action value of this MDP. Afterwards, a DRL-based scheme of maximizing system energy efficiency where RL and Neural Network (NN) are applied jointly is presented to replace the conventional one.

4.1 RL scheme

RL method can be used to solve our problem based on MDP and how to seek out the optimal sequence of decision $(\pi^* : S \to A)$ to maximize the long-term average reward is the focus of the issue [17]. The most classic model-free algorithm is Q-learning that uses Q-value to approximate the optimal state-action value. When the current slot is t, agent is in the state $s \in S$ and take an action $a \in A$ then get to next state $s' \in S$ and obtain the reward r_t, in this case the Q-value is updated as follows

$$Q_{t+1}(s,a) = (1-\alpha)Q_t(s,a) + \alpha\left[r_t + \gamma \max_{a'} Q_t(s',a')\right] \tag{19}$$

where $Q(s,a)$ is the sum of discount reward when agent is in the state s and takes the action a, $\alpha \in [0,1]$ is the learning rate, and $\gamma \in [0,1]$ is discount factor that decides the dependence on current Q-value. ε-greed policy is used to choose the action in the process of iteration until the Q-value converges, then the optimal policy can be obtained as

$$\pi(s) = \arg\max_{a \in A} Q(s,a) \tag{20}$$

4.2 DRL-based scheme of maximizing system energy efficiency

Traditional RL methods use a table to store and manage state-action pairs, which may cause capacity problem and huge management cost for the explosion of huge state space in practical engineering applications. To solve this problem, we propose DRL-based algorithm which deals with the huge state space with NN [18].

The key of DRL-based algorithm is approximating state-action value with a 3-layers NN. In addition, the training process is accomplished through collaborative updates of two NNs which have an identical architecture but distinct parameters i.e. weights and biases of NNs. One is called evaluation network that inputs state s and outputs the evaluation of state-action value

$$Y_{evaluation} = Q(s,a;\omega) \tag{21}$$

Another is called target network that inputs the next state and outputs the target of state-action value

$$Y_{t\arg et}^{DQN} = r_{t+1} + \gamma \max_{a'} Q(s',a';\omega') \tag{22}$$

The parameter ω of evaluation network is updated at every step but ω' of target network is updated by copying

from ω of evaluation network every episode. After every step, ω is updated by following formula

$$L(\omega) = E\left[(Y_{t\arg et}^{DQN} - Y_{evaluation})^2\right]$$
$$\nabla_\omega L(\omega) = E\left[Y_{t\arg et}^{DQN} - Y_{evaluation}\right]\nabla_\omega Y_{evaluation} \quad (23)$$

where ∇ is the loss function. and ∇ is the gradient function.

In order to accelerate the convergence of the algorithm, the system state i.e. 3-tuples $s_i(t) = \{c_i(t), l_i(t), e_i(t)\}$ is pre-processed according to the data quantity before entering a slot. According to (8), if the requirement of system delay is satisfied, the 3-tuples is simplified to a 2-tuples $s_i(t) = \{c_i(t), e_i(t)\}$, that is to say, the influence of data quantity is neglected to reduce the difficulty of training. Otherwise, the original 3-tuples is processed. The pseudo-code is outlined in Algorithm 1.

Algorithm 1 describes the processing steps of DRL-based scheme. Before training, some parameters i.e. ω and ω' of NN, relay memory D and Q-value Q are initialized. Then some training episodes are started (lines 3 to 19). The first step of each training episode is initializing the state s and reward R, $H(t)$ of MDP (lines 4 and 5). Then the training of NN which is described in detail below is repeated in the next steps (lines 6 to 15). In each step of training, the agent selects the action a according to ε-greed policy (line 7) and then a new state s' and corresponding reward r are obtained (line 8). According to the optimization method mentioned above, the agent judges whether the state s' can be simplified (line 9). Based on the above values, the agent stores experience tuples (s, a, r, s_0) in relay memory D for subsequent training (line 10). Next the agent samples the minibatch experience tuples randomly from D to update parameters ω according to Eq. (20), (21) and (22) to train the NN (lines 11 and 12). Finally, the agent updates $R = R + r$,$s = s'$ and $t = t + 1$ to repeat the above steps until the first node's energy is exhausted and an episode ends (lines 13, 14 and 15). After an episode ends, the agent updates the parameters ω' and calculates the average reward $H(t)$, then enters the next episode until episode is over (lines 16 to 19). In the DRL-based scheme, we can see that a sensor with larger Q-value has a higher probability to be selected to use channel, which depends on its current three states.

Algorithm 1 DRL-based Scheme of Maximizing System Energy Efficiency

1: Initial parameter ω and ω' of NN, relay memory D;
2: Initial Q-value $Q = 0$;
3: **repeat**
4: Initial $s = s(0)$;
5: Initial total reward R = 0 and average reward $H(t)$=0;
6: **repeat**
7: Take action a by ε-greed policy:
choose random a with probability ε;
or choose $a = \arg\max_a Q(s, a; \omega)$;
8: Get reward r and next state s';
9: If delay is satisfied, simplify s' as $\{c_i(t), e_i(t)\}$;
10: Store the experience tuples (s, a, r, s') in D;
11: Sample minibatch experience tuples from D;
12: Update parameter ω with (21), (22), (23);
13: Set $R = R+r$;
14: Set $s = s'$, $t = t+1$;
15: **until** (the first node's energy is exhausted)
16: Update parameter ω';
17: Set $T=t$ and $H(t) = R / T$;
18: Enter the next episode;
19: **until** (episode is over)

5 Simulations and results

A WBAN consisting of N sensors and one coordinator is generated to evaluate our scheme by simulations. The date rates of sensor is a random variable which is subject to the Poisson distribution with 3 packets per slot as the mean. The bits a packet contains is K = 250 bits, transportation bit rate is R = 250 kbps and the slot length is τ = 10 ms. According to the very common RF transceiver CC2530 whose transmission power is programmable to vary from − 28 dbm to 4.5 dbm. We choose 4 typical power $p = \{-10\text{ dBm}, -12\text{ dBm}, -14\text{ dBm}, -16\text{ dBm}\}$ in this simulations. Considering the research results of IEEE 802.15.6 task group on on-body channel [3], the corresponding path loss is quantified as 4 values $c = \{47\text{ dB}, 49\text{ dB}, 51\text{ dB}, 53\text{ dB}\}$. Markov chain is used to simulate the process of channel change, that is, if the channel is in a certain path loss state in the current slot, it will shift to other path loss state in the next time slot according to the corresponding transition probability in its state transition matrix. In addition, Gaussian noise power is set to be $p_n = -70$ dBm. If no otherwise specified, the parameters of DRL-based algorithm are as follows, discount factor γ=0.99 and learning rate $\alpha = 0.001$. Two algorithms compared with DRL in this paper are round

robin (RR) slot allocation algorithm and optimal channel quality (OCQ) slot allocation algorithm [19].

5.1 Analysis of loss function

As shown in Fig. 3, the initial values of $L(\omega)$ in DRL-based scheme with three different learning rates are all 1. Since then $L(\omega)$ falls sharply by continuous training, and tends towards stability around the 300th episode. We can see the DRL-based scheme with learning rate $\alpha = 0.001$ has the best convergence property as it has the minimal convergence value, while $\alpha = 0.01$ has the worst performance convergence property. Considering the general case of DRL and the simulation results, we set the learning rate to be $\alpha = 0.001$ in the subsequent simulations.

5.2 Analysis of reward and energy efficiency

Figure 4 illustrates the average reward when we increase the number of episodes when the quantity of sensors is $N = 8$. RR and OCQ maintain a larger reward $H(t)$ at the beginning of the training which is about 0.65 and 0.56 respectively because they are static scheduling strategy and do not require training. DRL started with reward only 0.2 but rise fast with the varying of episode and exceeds RR after about 80 episodes, it goes to steady around 0.8 at 250 episodes. The changing trend of energy efficiency is consistent with that of reward, and the maximal average energy efficiency is about 3.6 * 10^7 packets per Joule.

5.3 Analysis of QoS

Figure 5 illustrates the average system delay with different sensors. The average system delay increases with the increase of sensor number. This can be explained according

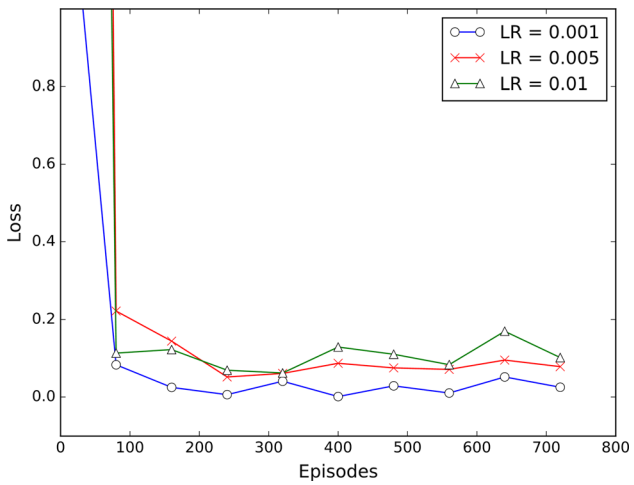

Fig. 3 Convergence performance Loss function $L(\omega)$ for DRL-based scheme

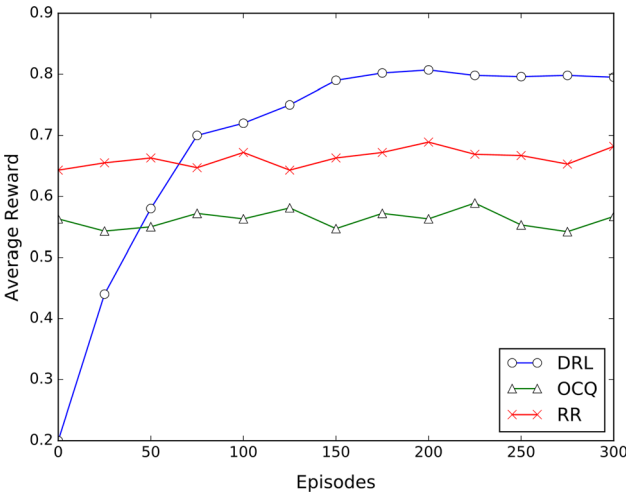

Fig. 4 Average reward $H(t)$ in different episodes

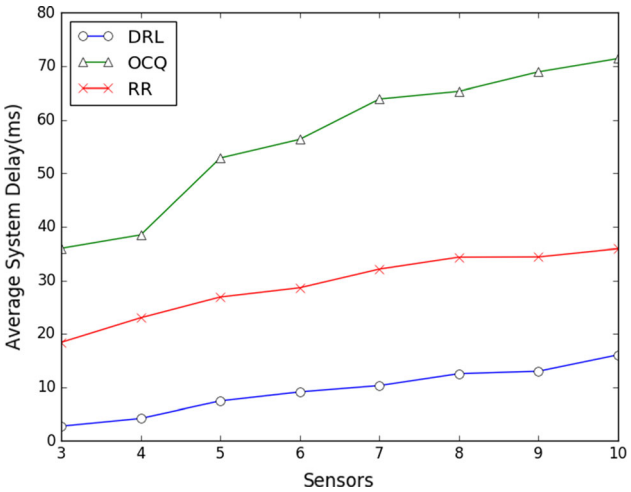

Fig. 5 Average system delay with different sensors

to Little's Law that the quantity of data received increases while the effective data rate keep constant, resulting in more data backlog in the buffers and greater system delay. It is obvious that DRL can keep the minimum system delay, followed by the RR and OCQ has a poor effect in terms of system delay.

Figure 6 illustrates the Jain index with different sensors. As the increase of sensor number the Jain index decreases and the effect between different algorithms is significant. DRL has the highest fairness index. Although OCQ can guarantee transmission with the best channel every time, the cost is a sharp drop in fairness index. It is well understood that sensors closer to the sink node tend to have better channel quality and the probability of being selected is greatly increased, causing the remote and occluded sensors to lose transmission opportunities. The fairness index of RR is between the above two algorithms.

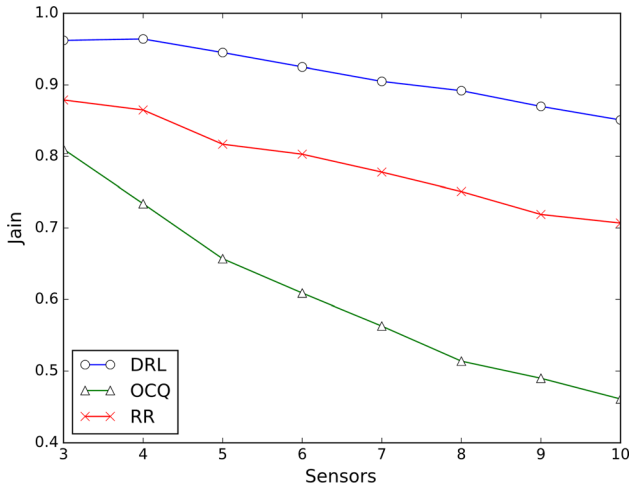

Fig. 6 Jain index with different sensors

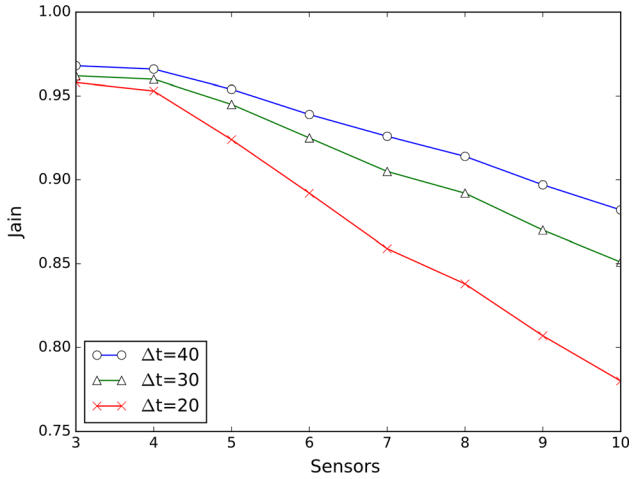

Fig. 7 Jain index with different Δt for DRL-based scheme

Figure 7 shows the effect of different window length Δt. on Jain index with DRL algorithm. It shows a similar vary to Fig. 6 that Jain index decreases with the increase of sensor number because the state quantity of each slot increases exponentially as the sensor number increases which leads to the increase of selection complexity and the decrease of fault tolerance rate. In addition, as Fig. 7 illustrates the Jain index falls with the decrease of Δt, which indicates that a short observation window meaning strict fairness requirement causes a low fairness index. It is consistent with the practical significance of Jain index.

Figure 8 shows the effects of parameters packet length and transmission bit rate on average reward when the quantity of sensors is $N = 8$. It is obvious from the curves that the increase of the packet length will reduce system reward because it will consume more energy for transmitting a packet. What's more, the rise of transmission bit rate can increase system reward because it allows more

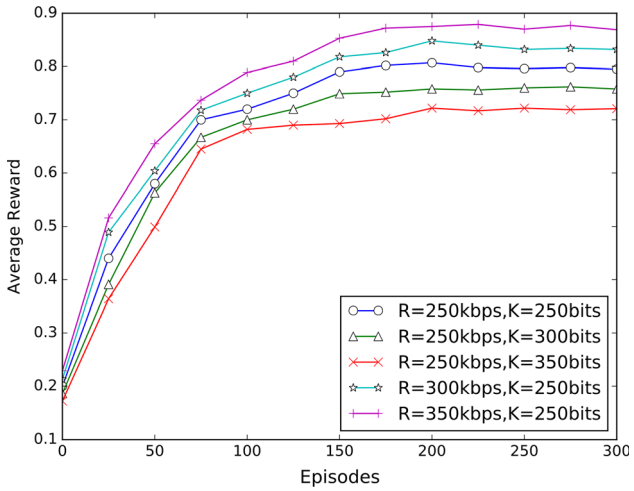

Fig. 8 Average reward with different packet length and bit rate

data packets to be transmitted with the same energy in one slot.

Furthermore, it should be noted that the DRL-Based scheme is specially proposed for WBANs. Although the scheme is effective for WBANs which is a kind of small-scale WSN, it can't guarantee that the effect of the scheme will not degenerate if it is transplanted to a traditional large-scale WSN. This is mainly because the DRL-Based scheme is a centralized approach based on neural network, but the convergence speed of the scheme in a large WSN that has hundreds of nodes will decrease. Therefore, we suggest that in a large-scale network, applying a distributed learning method will be more appropriate, unless this large-scale network is applied in situations with low real-time requirements.

6 Conclusions

This paper aims to improve the average energy efficiency and the QoS for a wireless body area network. We transform the joint optimization problem into a Markov decision problem whose goal is to find a strategy to get maximal reward. To solve the problem, a scheme for power control and time slot allocation based on deep reinforcement learning is proposed. In the scheme, the coordinator schedules the sensors based on certain state parameters such as data quantity, path loss and energy state. Focusing on the joint optimization problem, the gains of energy efficiency, system delay and system fairness index are designed to grow with the learning reward of the Markov decision model. The coordinator executes the online learning and updates the reward in each running time. In this way, the DRL-based scheme can adapt the environment changing. The simulation results indicate the

scheme has a good convergence performance, and obtains the maximal energy efficiency of $3.6 * 10^7$ packets per Joule with the minimal system delay and a Jain index of more than 0.9, which shows it can fully satisfy the QoS requirements. Also, the scheme exhibits better performance than RR and OCQ schemes in terms of energy efficiency, system delay and system fairness. In the future we intend to implement the scheme on a real health monitoring platform.

Acknowledgements Funding was provided by the Fundamental Research Funds for the Central Universities (Grant No. 30918011329).

References

1. Movassaghi, S., Abolhasan, M., Lipman, J., Smith, D., & Jamalipour, A. (2014). Wireless body area networks: A survey. *IEEE Communications Surveys & Tutorials, 16*(3), 1658–1686.
2. Salayma, M., Al-Dubai, A., Romdhani, I., & Nasser, Y. (2017). Wireless body area network (WBAN): a survey on reliability, fault tolerance, and technologies coexistence. *ACM Computing Surveys (CSUR), 50*(1), 3.
3. Yazdandoost, K., & Sayrafian, K. (2009). *Channel model for body area network (Ban)*. IEEE p802. 15-08-0780-09-0006. IEEE 802.15 working group document.
4. Hayajneh, T., Almashaqbeh, G., Ullah, S., & Vasilakos, A. V. (2014). A survey of wireless technologies coexistence in WBAN: analysis and open research issues. *Wireless Networks, 20*(8), 2165–2199.
5. Davenport, D. (2009). *Medwin physical layer proposal*. IEEE 802.15 document repository. IEEE 802.15-09-0328-01-0006.
6. Kaur, T., & Kumar, D. (2016). TDMA-based mac protocols for wireless sensor networks: A survey and comparative analysis. In *2016 5th international conference on wireless networks and embedded systems (WECON)* (pp. 1–6). IEEE.
7. Salayma, M., Al-Dubai, A., Romdhani, I., & Nasser, Y. (2017). New dynamic, reliable and energy efficient scheduling for wireless body area networks (WBAN). In *2017 IEEE international conference on communications (ICC)* (pp. 1–6). IEEE.
8. Kim, S., & Eom, D. S. (2014). Link-state-estimation-based transmission power control in wireless body area networks. *IEEE Journal of Biomedical Health Informatics, 18*(4), 1294–1302.
9. Goswami, D., Sarma, K. C., & Mahanta, A. (2015). Experimental determination of path loss and delay dispersion parameters for on-body UWB WBAN channel. In *IEEE international conference on signal processing*.
10. Chaganti, V. G., Hanlen, L. W., & Lamahewa, T. A. (2011). Semi-markov modeling for body area networks. In *2011 IEEE international conference on communications (ICC)* (pp. 1–5). IEEE.
11. Li, H., Yang, B., Yu, W., Guan, X., Gong, X., & Yu, G. (2014). Joint sleep scheduling and opportunistic transmission in wireless body area networks. In *The 26th Chinese control and decision conference (2014 CCDC)* (pp. 1886–1891). IEEE.
12. Taleb, S., Hajj, H., & Dawy, Z. (2018). VCAMS: Viterbi-based context aware mobile sensing to trade-off energy and delay. *IEEE Transactions on Mobile Computing, 17*(1), 225–242.
13. Li, J., Liu, Q., Wu, P., Shu, F., & Jin, S. (2018). Task offloading for UAV-based mobile edge computing via deep reinforcement learning. In *2018 IEEE/CIC international conference on communications in China (ICCC)* (pp. 798–802). IEEE.
14. Chen, Guihong, Zhan, Yiju, Sheng, Geyi, Xiao, Liang, & Wang, Yonghua. (2019). Reinforcement learning-based sensor access control for WBANs. *Access IEEE, 7,* 8483–8494.
15. Kleinrock, L. (1975). Queueing systems. Vol. I: Theory. *IEEE Transactions on Communications, 65*(6), 990–991.
16. Jain, R. K., Chiu, D.-M. W., & Hawe, W. R. (1984). *A quantitative measure of fairness and discrimination*. Hudson, MA: Eastern Research Laboratory, Digital Equipment Corporation.
17. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge: MIT Press.
18. Anschel, O., Baram, N., & Shimkin, N. (2017). Averaged-DQN: Variance reduction and stabilization for deep reinforcement learning. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 176–185). JMLR.org.
19. Umehira, M., Fujita, S., Gao, Z., & Wang, J. (2013). Dynamic channel assignment based on interference measurement with threshold for multibeam mobile satellite networks. In *2013 19th Asia-Pacific conference on communications (APCC)* (pp. 688–692). IEEE.

Lili Wang received her Ph.D. degree in Control Science and Engineering from Nanjing University of Science and Technology, Nanjing, China, in 2015. She is now an assistant professor in Nanjing University of Science and Technology. Her current research interests include Wireless Ad hoc Network and Coordination Control.

Ge Zhang received the B.E. in Electrical Engineering from Nanjing University of Science and Technology, Nanjing, China, in 2017. Currently, he is working as a graduate student in Nanjing University of Science and Technology, Nanjing. His research interests include Wireless Body Area Network, Reinforcement Learning.

Jun Li received Ph.D. degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. From June 2009 to April 2012, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia. From April 2012 to June 2015, he is a Research Fellow at the School of Electrical Engineering, the University of Sydney, Australia. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include network information theory, channel coding theory, wireless network coding, resource allocations in cellular networks.

Gaoshang Lin received the B.E. in automation from Fuzhou University, Fuzhou, China, in 2019. Currently, he is working as a graduate student in Nanjing University of Science and Technology, Nanjing. His current research interests include Wireless Ad hoc Network and Reinforcement Learning.