Paper

Evaluation of Radio Channel Utility using Epsilon-Greedy Action Selection

Krzysztof Malon

Military University of Technology, Warsaw, Poland

https://doi.org/10.26636/jtit.2021.153621

Abstract—This paper presents an algorithm that supports the dynamic spectrum access process in cognitive radio networks by generating a sorted list of best radio channels or by identifying those frequency ranges that are not in use temporarily. The concept is based on the reinforcement learning technique named Q-learning. To evaluate the utility of individual radio channels, spectrum monitoring is performed. In the presented solution, the epsilon-greedy action selection method is used to indicate which channel should be monitored next. The article includes a description of the proposed algorithm, scenarios, metrics, and simulation results showing the correct operation of the approach relied upon to evaluate the utility of radio channels and the epsilon-greedy action selection method. Based on the performed tests, it is possible to determine algorithm parameters that should be used in this proposed deployment. The paper also presents a comparison of the results with two other action selection methods.

Keywords—cognitive radio, dynamic spectrum access, spectrum monitoring, machine learning, Q-learning.

1. Introduction

With the dynamic development of wireless communications systems, spectrum scarcity has become an increasingly important problem. The vast majority of radio frequency bands are assigned to licensed (primary) users on an exclusive basis. At the same time, by analyzing the use of frequency resources over time [1]–[3], one can identify the so-called spectrum holes. This term refers to frequency bands that are not in use temporarily and may be utilized for transmission by secondary (unlicensed) users. This approach is referred to as dynamic spectrum access (DSA). To apply this concept, cognitive radio (CR) technology is proposed. The functionalities of CR include the ability to receive various information from the surrounding environment, analyze it, make decisions, and perform specific actions. The ability to learn (improve functionality) from previous reactions and from the results obtained is another essential feature of CR.

Implementation of DSA requires constant monitoring of the available radio resources and means that the usefulness of individual frequency ranges (i.e. radio channels) needs to be determined. For secondary users, channels with low occupancy ratios (activity of other users) are the most important ones.

2. Evolution of Radio Channel Utility

The algorithm proposed in this paper for evaluating the utility of radio channels supports DSA and is based on the machine learning method named Q-learning. This technique belongs to the class of reinforcement learning methods in which learning takes place through experimentation. In addition to reinforcement learning, two primary groups of machine learning methods may be distinguished, namely supervised and unsupervised learning [4], [5]. Reinforcement learning is considered to be useful in terms of CR, especially in monitoring and accessing the spectrum in a dynamically changing environment [4]. In the considered solution, reinforcement learning of the single state [6], [7] or stateless type [8], [9] is analyzed. The proposed algorithm does not require any knowledge of the radio environment. It recognizes and learns spectrum usability-related information by relying on the trial and error method [6], [10]. On the other hand, if the state of several frequency channels is known, it may also be used during the initialization step or during the algorithm's operation.

Figure 1 depicts the general scheme of the proposed algorithm that consists of four primary stages repeated as the system is operated. Before the algorithm starts, the Q matrix should be initialized. This matrix consists of channels a and their estimated qualities $Q(a)$.

The first step of the algorithm is the selection of action (a). During this stage, a new channel for sensing (i.e. radio spectrum monitoring) is indicated. Random and cyclic algorithms are the most straightforward and the most popular solution, as they search the entire action space with equal probabilities. In the first case, the action (channel) is selected randomly. The second solution assumes that channels are specified in sequence over a repeated cycle. Another proposal is the epsilon-greedy strategy, which is a greedy policy variant (see Fig. 2). By using this approach, one may exploit (use) the best actions and reduce the effort

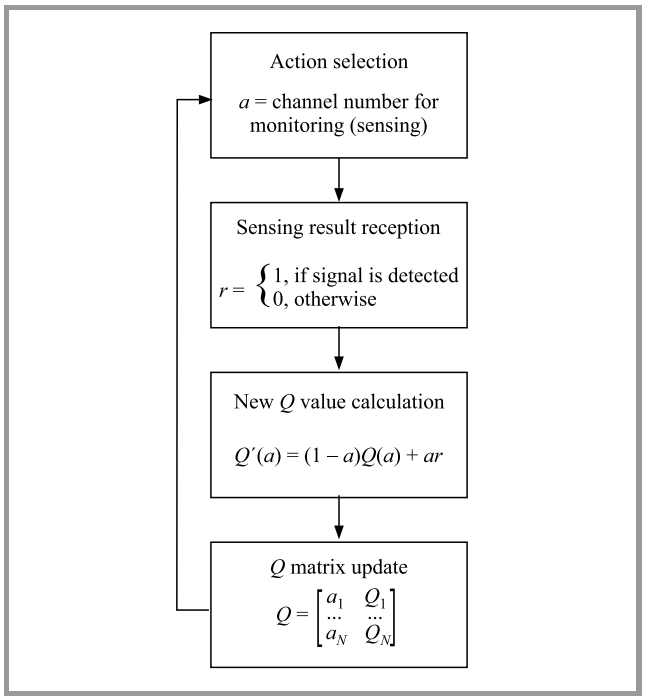

Fig. 1. Radio channel utility evaluation algorithm.

required to explore (search) for others. According to this principle, the following action is chosen:

- randomly with low probability ε,
- or according to the current policy of maximizing rewards with a probability of $1-\varepsilon$.

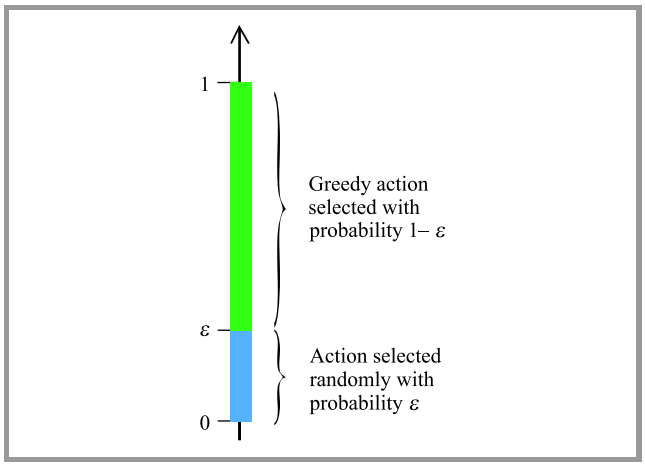

Fig. 2. Epsilon-greedy action selection method.

The ε value determines the probability of taking the greedy action, selecting the best channel for sensing. Otherwise, it also defines the probability $1-\varepsilon$ of performing a random action. This makes it possible to find other channels with good qualities. As shown in Fig. 2, by changing the ε value, a trade-off between exploration and exploitation is reached. An example algorithm for the epsilon-greedy action selection method is shown in Fig. 3. Firstly, a random number p $(p = 0,\ldots,1)$ is generated. Then, its value is compared with the defined ε. Depending on the result, a random action is performed or the best channel is selected.

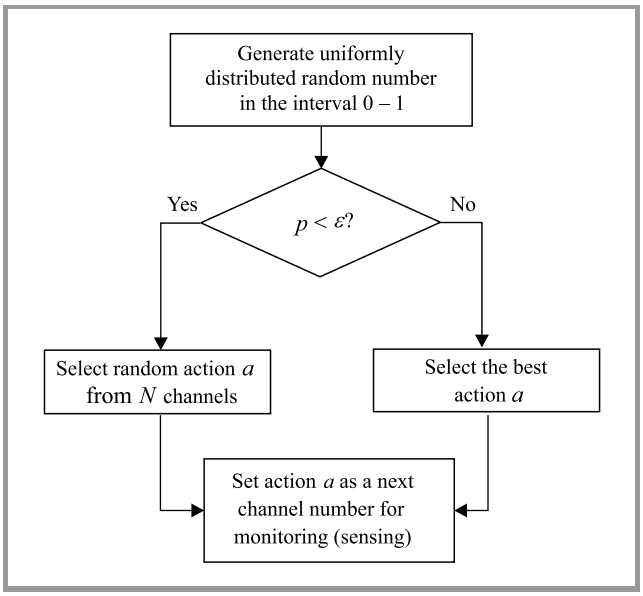

Fig. 3. Epsilon-greedy action selection algorithm.

In the next step, the radio channel utility evaluation algorithm is executed on the selected frequency channel. Two basic sensing approaches may be used in the proposed solution: local spectrum monitoring or cooperative sensing of spatially distributed radio network nodes. The problem of optimizing the placement of sensing elements is considered, inter alia, in [11] and [12]. The use of cooperative spectrum monitoring allows to reduce the severity of the problem of the so-called hidden nodes [13]. In such cases, it is necessary to apply a certain data fusion method, e.g. the Dempster-Shaffer theory [14].
The spectrum monitoring result r is passed to the following step of the algorithm in which the calculation of a new $Q'(a)$ value for the analyzed channel is performed. Both the newly obtained r result and the previous value $Q(a)$ are considered. The significance of new and historical data is defined by the learning rate α. The calculation of the new value $Q'(a)$ is performed using the following relationship:

$$Q'(a) = (1-\alpha)Q(a) + \alpha r\ , \tag{1}$$

where:

- a – selected action,
- $Q(a)$ – Q value for the selected action,
- $Q'(a)$ – new (updated) Q value for the selected action,
- $\alpha \in <0,1>$ – learning rate,
- r – reward.

In the next step, the determined value $Q'(a)$ is used to update the Q matrix, and then the algorithm cycle repeats.

3. Simulations and Results

This section of the paper presents the scenarios, metrics, and simulation results of the proposed algorithm using the epsilon-greedy action selection approach.
To evaluate the proposed algorithm, two base scenarios were prepared. Each scenario consists of radio channels with their occupancy defined over time. To generate spectrum occupancy figures, a statistical approach based on the On-Off model (see Fig. 4) was used [15]. The traffic generated may be interpreted as originating from one or more users. This model assumes two possible states: occupied (On) and not-occupied (Off).

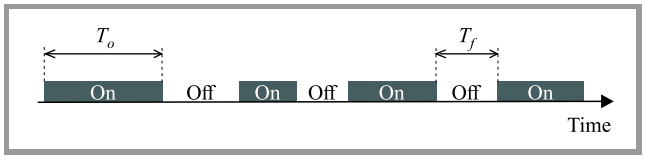

Fig. 4. On-Off spectral occupancy model [15].

In the selected Poisson-exponential model, the arrival procedure is modeled by a Poisson process. The time between successive spectral occupancies T_f (Off periods) is defined by an exponentially distributed random process:

$$f\left(T_f\right) = \frac{1}{\overline{T_f}} \mathrm{e}^{-\frac{T_f}{\overline{T_f}}} , \tag{2}$$

with a mean interarrival time of:

$$E\left[T_f\right] = \overline{T_f} . \tag{3}$$

The duration of the occupancy time T_o (On periods) is also modeled as an exponentially distributed random process given by:

$$f\left(T_o\right) = \frac{1}{\overline{T_o}} \mathrm{e}^{-\frac{T_o}{\overline{T_o}}} , \tag{4}$$

with a mean occupancy time of:

$$E\left[T_o\right] = \overline{T_o} . \tag{5}$$

According to the ITU-R report [16], spectrum resource occupancy *SRO* is the ratio of the number of channels in use (occupied) to the total number of channels in the entire frequency band. *SRO* for multiple channels within a specific time, called the integration time, is calculated as follows:

$$SRO = \frac{N_0}{N} , \tag{6}$$

where: N_0 – number of samples on any channel with a level above the threshold and N – total number of samples taken on all channels during the integration time.
In this case, the integration time is equated with the scenario time. Spectrum resource occupancy *SRO* may be interpreted as the average occupancy of the channels.

Table 1
Scenario 1 parameters

Parameter name	**Parameter value**		
	Even-numbered channels (2, 4, 6, 8, 10, 12)	**Odd-numbered channels (1, 3, 5, 7, 9, 11)**	**All channels**
Simulation time T	10,000		
Number of channels M	6	6	12
Average On time $\overline{T_0}$	10	10	-
Average Off time $\overline{T_f}$	10	30	-
Spectrum resource occupancy *SRO*	0.5	0.25	0.375

Table 2
Scenario 2 parameters

Parameter name	**Parameter value**		
	Even-numbered channels (2, 4, 6, 8, 10, 12)	**Odd-numbered channels (1, 3, 5, 7, 9, 11)**	**All channels**
Simulation time T	10,000		
Number of channels M	6	6	12
Average On time $\overline{T_0}$	40	40	-
Average Off time $\overline{T_f}$	40	120	-
Spectrum resource occupancy *SRO*	0.5	0.25	0.375

For evaluation purposes, two scenarios are considered. Both consist of twelve radio channels for which 10,000 states are defined (Table 1 and Table 2). The channels are divided into two groups with different parameters. Spectrum resource occupancy *SRO* for even-numbered channels is about 0.5, whereas for odd-numbered channels $SRO \approx 0.25$. This means that the overall spectrum occupancy for both scenarios (all channels) equals 0.375. The difference between scenario 1 and scenario 2 is in average On and Off times. Channel state changes in scenario 2 are

four times slower compared to scenario 1, e.g. $\overline{T_o}$ and $\overline{T_f}$ for even-numbered channels in scenario 1 are equal to 10, while for the scenario 2 these parameters are equal to 40.

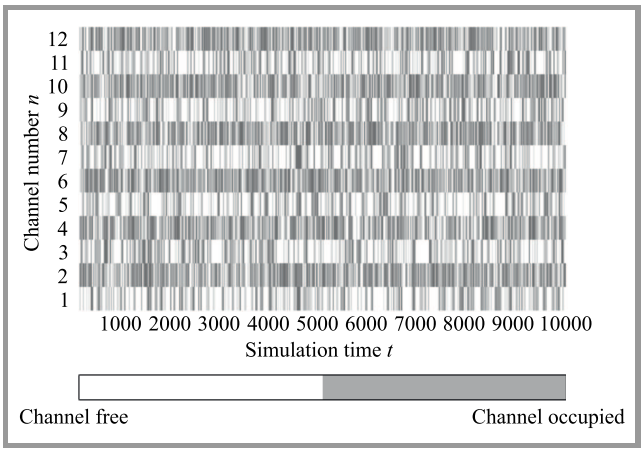

Fig. 5. Radio channel occupancy for base scenario 1.

The generated radio channel occupancy rates are shown in Figs. 5 and 6 for scenario 1 and scenario 2, respectively. The results are consistent with the Poisson-exponential model. White color represents not-occupied states (free), whereas gray is used to identify occupied channels.

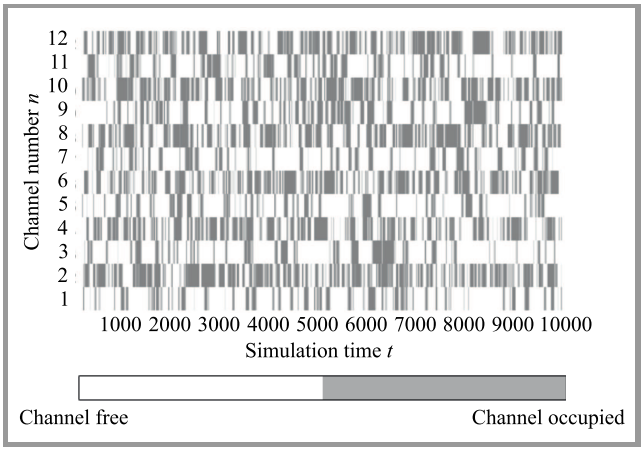

Fig. 6. Radio channel occupancy for base scenario 2.

The two presented scenarios serve as a basis for the simulations (base scenarios). They are used for the first iteration of the simulations performed according to the algorithm illustrated in Fig. 7.

Each scenario is simulated for different ε values, and then the relevant metrics are calculated. After analyzing the obtained metrics, ε value rendering the best results is selected. Consequently, there is a Q matrix and the best radio channel for each time step. In the next stage, information about temporally best channels is used to modify the input scenario, and then the simulations are performed relying on this updated data. That allows the scenarios used in subsequent iterations to be defined based on an increasing spectrum resource occupancy rate.

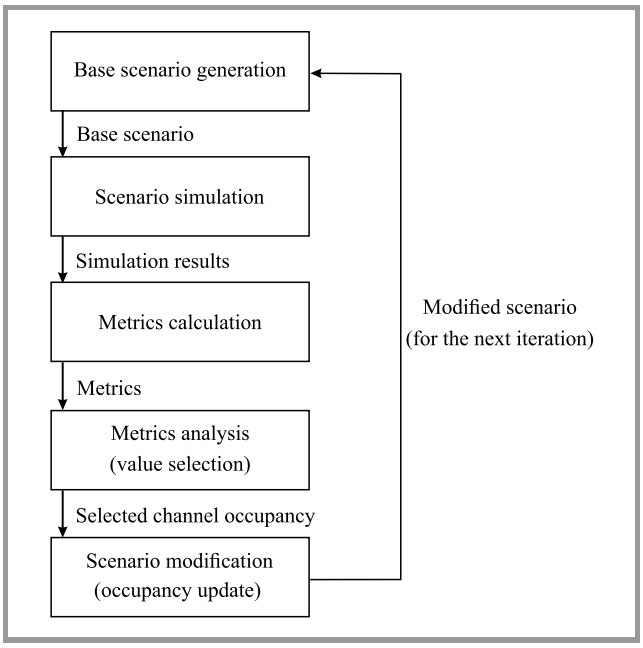

Fig. 7. Simulation algorithm – successive iterations.

3.1. Metrics Used

To evaluate the proposed solution, specific metrics are proposed. The first one is channel utility Utl, defined as:

$$Utl = \frac{N_f}{T} \ , \tag{7}$$

where: N_f – number of samples on the selected channel with a level below threshold (channel not-occupied) and T – total number of samples taken on the selected channel during the scenario time.

The Utl value can vary from 0 to 1. The second metric is spectrum resource occupancy gain SRO_{gain} defined as:

$$SRO_{gain} = SRO_2 - SRO_1 \ , \tag{8}$$

where: SRO_1 – reference spectrum resource occupancy value (occupancy of the scenario prepared for simulation) and SRO_2 – spectrum resource occupancy after simulation (including the occupancy resulting from the use of the best channel determined by the algorithm).

The goal is to obtain the highest SRO_{gain} and therefore the greatest Utl value, as radio resources are then used more efficiently. SRO_1 may be defined in the same way as in Eq. (6):

$$SRO_1 = \frac{N_0}{N} \ . \tag{9}$$

The use of the spectrum, when the system is using the best channels indicated by the proposed algorithm, increases proportionally to the Utl value. In such a case, the not-occupied states of the selected channel N_f change their status to occupied and increase the utilization of frequency resources. Accordingly, SRO_2 may be defined as:

$$SRO_2 = \frac{N_o + N_f}{N} \ . \tag{10}$$

The total number of samples N taken on all channels during the integration time is expressed by:

$$N = MT \ , \tag{11}$$

where M is the number of channels.
The SRO_{gain} may be defined as:

$$SRO_{gain} = \frac{N_o + N_f}{N} - \frac{N_0}{N} = \frac{N_f}{MT} = \frac{Utl}{M} \ . \tag{12}$$

Considering the range of variability of the parameter Utl, the maximum SRO_{gain} is:

$$SRO_{gainMax} = \frac{1}{M} \ . \tag{13}$$

3.2. Results

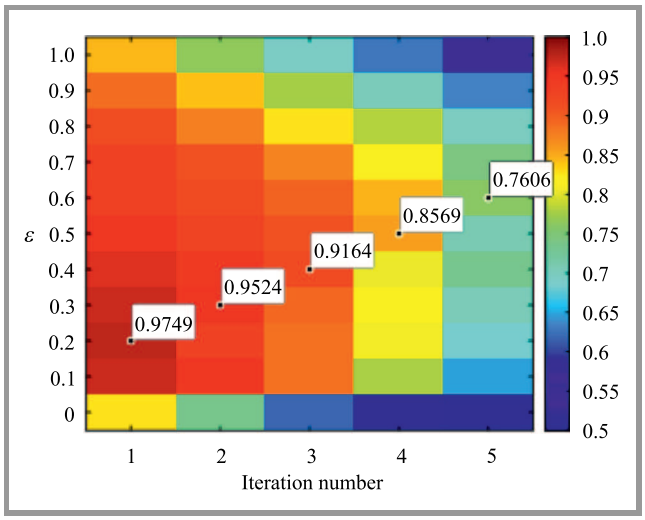

Fig. 8. Utility values for different epsilon ε values in successive iterations of base scenario 1. (see the digital version for color images)

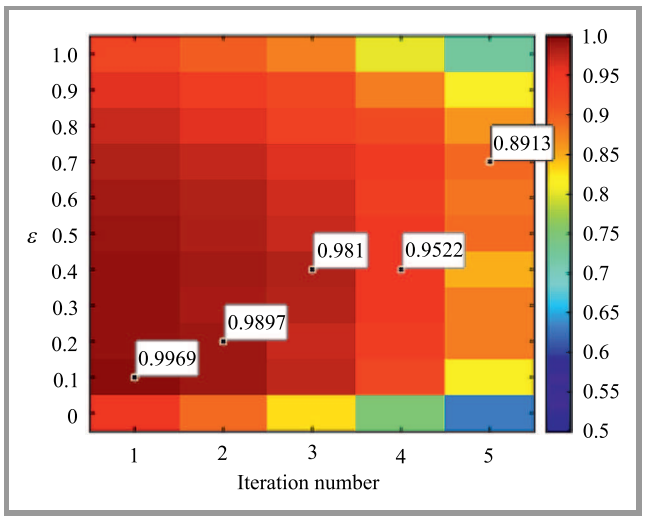

Fig. 9. Utility values for different epsilon ε values in successive iterations of base scenario 2.

Figures 8–11 show Utl, SRO, and SRO_{gain} in successive iterations for both scenarios. Utility values presented in Figs. 8 and 9 are calculated for the first channel in the Q matrix – the best radio channel in each simulation step. Better results are obtained for scenario 2. Here, higher utility values are obtained compared to those for scenario 1 for the same spectrum resource occupancy (iteration number). It is so because of the different channel state changes dynamics. In scenario 1, shorter On and Off times cause frequent channel state changes. The larger the iteration number, the greater value of ε provides the best utility values. It means that for a higher spectrum resource occupancy rate, the epsilon-greedy action selection method should increase exploration.

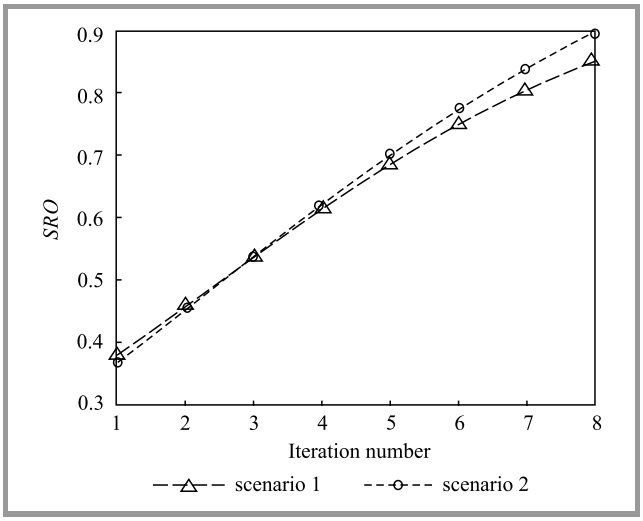

Fig. 10. Spectrum resource occupancy SRO in successive iterations for both scenarios.

Figures 10 and 11 show spectrum resource occupancy and SRO_{gain} for both scenarios. For the first three iterations,

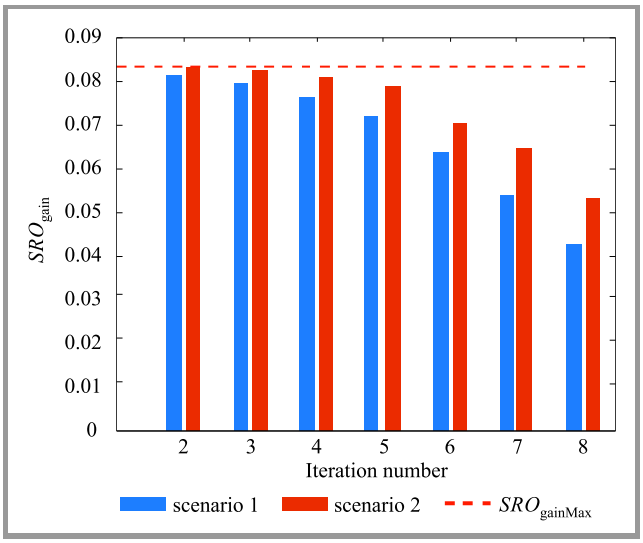

Fig. 11. Spectrum resource occupancy gain SRO_{gain} in successive iterations of both scenarios.

SRO_{gain} values are close to their maximum $SRO_{gainMax}$ (red dashed line in Fig. 11). As mentioned before, better results (higher SRO_{gain}) can be obtained for scenario 2 due to lower channel state changes dynamics. As the iteration number increases, the spectrum occupancy grows, and thus the chances of finding free radio resources (not-occupied channel) decreases. Therefore, the channel utility Utl values shown in Figs. 8 and 9 and the spectrum resource occupancy gain SRO_{gain} presented in Fig. 11 take a lower values with the increase in the iteration number.

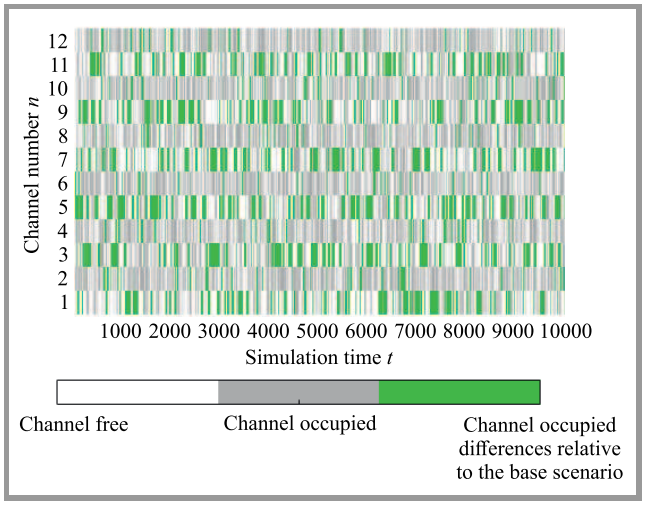

Fig. 12. Radio channel occupancy for the third iteration of base scenario 1.

Figures 12 and 13 show the growth in radio channel occupancy after two iterations, compared to the base scenarios. White and gray colors represent free and occupied states in the base scenario. Green indicates new occupied states resulting from including the temporarily best channels selected by the proposed algorithm. As one may notice, odd-numbered channels are chosen more often because their spectrum resource occupancy is lower. Please refer to Tables 1 and 2 for detailed parameters. This behavior confirms the correct operation of the algorithm identifying temporarily free channels.

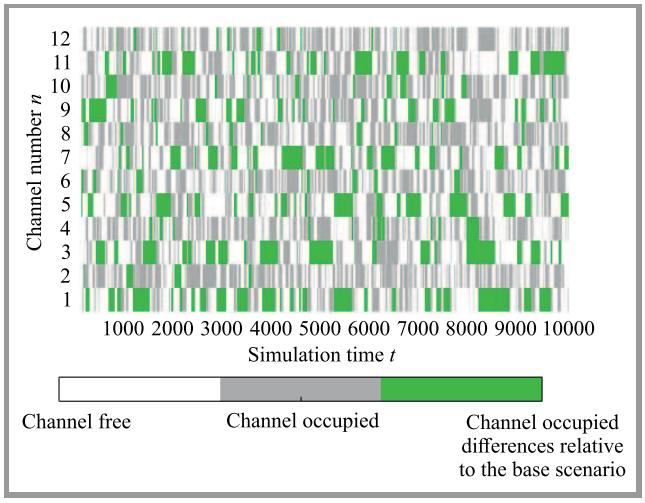

Fig. 13. Radio channel occupancy for the third iteration of base scenario 2.

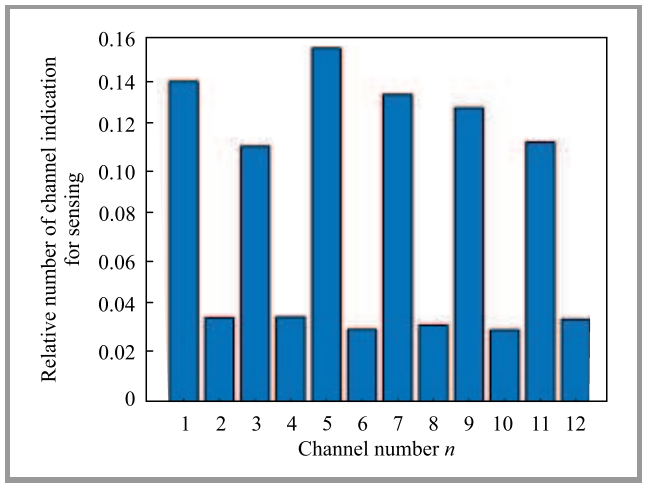

Fig. 14. Radio channels selected by the epsilon-greedy method ($\varepsilon = 0.2$) in the first iteration of base scenario 1.

Figures 14 and 15 depict the effect of the epsilon-greedy action selection method. They show how often particular channels are selected for sensing (monitoring). There are two example results from the scenario 1 simulations. The first one concerns the base scenario (first iteration), when SRO is approx. 0.375. In this case the ε value is set to 0.2, which means that greedy actions are taken with the probability of 0.8 (please refer to Fig. 2). This results in more frequent selection of less busy channels (odd-numbered channels). Figure 15 presents the behavior of the epsilon-greedy action selection method in the fourth iteration. In this situation, the ε value that allows to obtain the best Utl is 0.5 (see Fig. 8). As the spectrum occupancy grows it is needed to increase the exploration to find other free channels. Probabilities of the selection of individual channels are equalized.

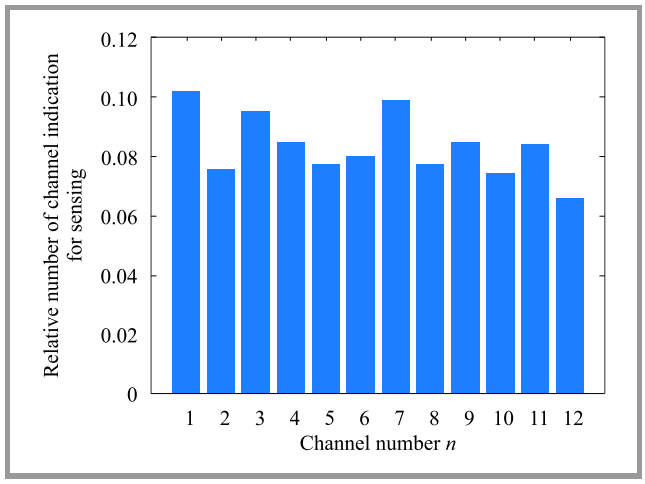

Fig. 15. Radio channels selected by the epsilon-greedy method ($\varepsilon = 0.5$) in the fourth iteration of base scenario 1.

Radio channels utilities in the first and fourth iteration of base scenario 1 are presented in Figs. 16 and 17, respectively. These results compare three action selection methods: epsilon-greedy, random and cyclic. Analysis of the first channel index in the Q matrix (the best one) shows a significant advantage that the epsilon-greedy algorithm has over the other two methods.

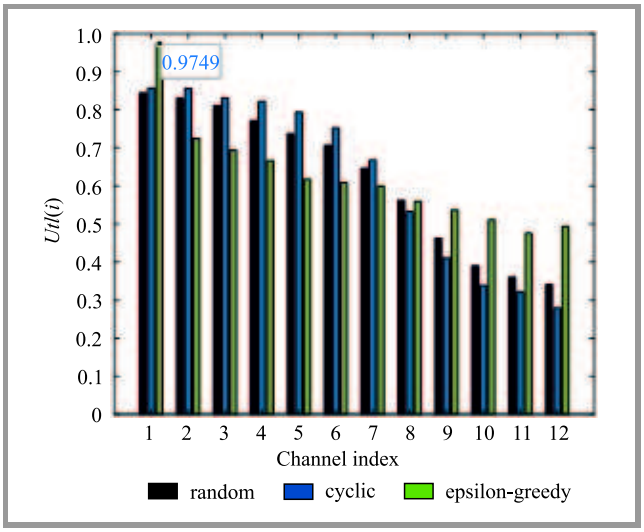

Fig. 16. Radio channel utility rate in the first iteration of base scenario 1.

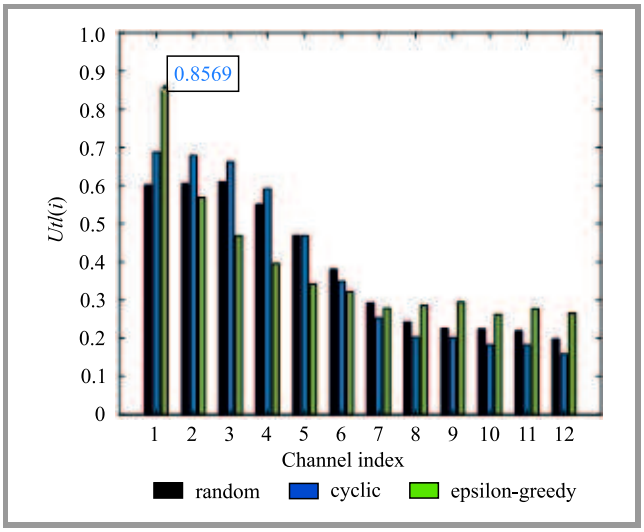

Fig. 17. Radio channel utility rate in the fourth iteration of base scenario 1.

4. Conclusions

This paper presents an algorithm for evaluating the usefulness of radio channels based on a machine learning technique named Q-learning. The proposed algorithm identifies radio channels for sensing using the epsilon-greedy action selection method. This process aims to reach a trade-off between exploration and exploitation of the available radio channels. Based on the results from the process of monitoring frequency resources, individual radio channels are evaluated. As a result, a sorted list of radio channels capable of supporting DSA is generated. An essential feature of the proposed concept is that it does not need to be initialized, meaning it may work in an unknown electromagnetic environment, gradually building its situational awareness. The presented scenarios, metrics, and simulation results show the algorithm's correct operation and the proper choice of the action selection method. The tests performed have identified the crucial ε values that allow to reach the maximum spectrum utilization rate under specific conditions. The epsilon-greedy action selection method is also compared with two other approaches: random and cyclic. It has been shown that the channel utilization rates obtained using the epsilon-greedy approach are much better.

5. Acknowledgment

This work was financed by the Military University of Technology under research project no. UGB/22-854/2021/WAT "Application of selected computer science, communication, and reconnaissance techniques in civilian and military areas".

References

[1] Wireless Innovation Forum, "Dynamic Spectrum Sharing Annual Report – 2014", Document WINNF-14-P-0001, version V0.2.16 [Online]. Available: https://www.wirelessinnovation.org/assets/work_products/Reports/winnf-14-p-0001-v1.0%20dynamic%20spectrum%20sharing%20annual%20report%202014.pdf

[2] M. A. McHenry, D. McCloskey, and G. Lane-Roberts, "New York City spectrum occupancy measurements", *Shared Spectrum Company*, 2005 [Online]. Available: http://www.sharedspectrum.com/wp-content/uploads/4_NSF_NYC_Report.pdf

[3] Shared Spectrum Company, "General survey of radio frequency bands – 30 MHz to 3 GHz", 2010 [Online]. Available: https://www.sharedspectrum.com/wp-content/uploads/2021/01/2010_0923-General-Band-Survey-30MHz-to-3GHz.pdf

[4] E. Biglieri, A. J. Goldsmith, L. J. Greenstein, N. B. Mandayam, and H. V. Poor, *Principles of cognitive radio*. Cambridge: Cambridge University Press, 2012 (ISBN: 9781139236850).

[5] L. E. Doyle, *Essentials of cognitive radio*. Cambridge: Cambridge University Press, 2009 (ISBN: 9780511576577).

[6] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction, second edition*. Cambridge: The MIT Press, 2018 (ISBN: 9780262039246).

[7] K-L. A. Yau, P. Komisarczuk, and P. D. Teal, "Applications of reinforcement learning to cognitive radio networks", in *Proc. IEEE Int. Conf. on Commun. Workshops*, Cape Town, South Africa, 2010, pp. 1–6 (DOI: 10.1109/ICCW.2010.5503970).

[8] N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated Q-learning for robust cognitive spectrum management in LTE cellular systems", *IEEE Transac. on Mobile Comput.*, vol. 15, no. 4, pp. 817–825, 2016 (DOI: 10.1109/TMC.2015.2442529).

[9] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems", in *Proc. of the fifteenth national/tenth Conf. on Artif. Intell./Innovat. Applicat. of Artif. Intell.*, Madison, WI, USA, 1998, pp. 746–752 [Online]. Available: https://www.aaai.org/Papers/AAAI/1998/AAAI98-106.pdf

[10] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios", *IEEE Commun. Surveys & Tut.*, vol. 15, no. 3, pp. 1136–1159, 2013 (DOI: 10.1109/SURV.2012.100412.00017).

[11] K. Malon, P. Skokowski, and J. Łopatka, "Optimization of wireless sensor network deployment for electromagnetic situation monitoring", *Int. J. of Microwave and Wireless Technol.*, vol. 10, no. 7, pp. 746–753, 2018 (DOI: 10.1017/S1759078718000211).

[12] K. Malon, P. Skokowski, and J. Łopatka, "Optimization of the MANET topology in urban area using redundant relay points", *Int. Conf. on Military Commun. and Informat. Systems (ICMCIS)*, Warsaw, Poland, 2018, pp. 1–4 (DOI: 10.1109/ICMCIS.2018.8398720).

[13] P. Skokowski, K. Malon, and J. Łopatka, "Properties of centralized cooperative sensing in cognitive radio networks", in *Proc. XI Conf. on Reconnaissance and Electron. Warfare Systems*, J. Łopatka, Eds. *Int. Society for Optics and Photon.*, vol. 10418, pp. 54–62. Ołtarzew, Poland: SPIE, 2017 (DOI: 10.1117/12.2269996).

[14] P. Skokowski, "Electromagnetic situation awareness building in ad-hoc networks with cognitive nodes", *Military University of Technology*, Warsaw, Poland, 2021 (in Polish, in print).

[15] K. Sithamparanathan and A. Giorgetti, *Cognitive Radio Techniques*. Artech House, 2012 (ISBN: 9781608072040).

[16] ITU-R Report SM.2256-1, "Spectrum occupancy measurements and evaluation", *Int. Telecommun. Union*, 2016 [Online]. Available: https://www.itu.int/pub/R-REP-SM.2256

Krzysztof Malon received his M.Sc. and Ph.D. degrees from the Military University of Technology in 2011 and 2019, respectively. Since 2016 he has been working at the Institute of Communications Systems of MUT. His main research interests are wireless communications, cognitive radio, dynamic spectrum access, and radio spectrum monitoring. Krzysztof Malon has participated in many national and international research projects for the European Defence Agency and the National Centre for Research and Development. Recently, he is also involved in the NATO working group related to the 5G technologies application to NATO operations.

https://orcid.org/0000-0001-9257-1166
E-mail: krzysztof.malon@wat.edu.pl
Institute of Communications Systems
Faculty of Electronics
Military University of Technology
Warsaw, Poland