RESOURCE MINIMIZATION DRIVEN SPECTRUM SENSING POLICY

Jan Oksanen, Jarmo Lundén and Visa Koivunen

SMARAD CoE, Department of Signal Processing and Acoustics
School of Electrical Engineering, Aalto University
Email: jhoksane@wooster.hut.fi, jrlunden@wooster.hut.fi, Visa.Koivunen@hut.fi

ABSTRACT

In this paper a reinforcement learning-based distributed sensing policy is proposed for cognitive radio networks. The proposed sensing policy is controlled by a fusion center that employs action-value learning to focus the search for idle frequencies to those parts of the spectrum that persistently provide a high data rate. The fusion center learns the local sensing performances of the secondary users and attempts to minimize the number of assigned users for sensing under a constraint on the global detection probability. A heuristic polynomial time algorithm iteratively employing the Hungarian method is proposed for finding a feasible assignment that minimizes the number of active sensors. Simulation results show that the proposed algorithm is able to find near-optimal solutions in practise significantly faster than an exact branch-and-bound search.

Index Terms— Cognitive Radio, Sensing policy, Reinforcement learning, Hungarian method, Multi-armed bandit

1. INTRODUCTION

Cognitive radio (CR) is a promising new technology for supplying radio spectrum for future wireless services in an agile manner. CRs explore the radio spectrum in the hope of finding idle frequencies which they could exploit without interfering with the primary users (PUs). In order to identify such spectral opportunities the CRs need to sense the radio spectrum. The problem of searching for idle frequency bands over a large bandwidth can be viewed as a restless multi-armed bandit problem. The spectrum may be seen to consist of N_B subbands (arms) that provide stochastic throughputs (rewards) $R_t(i), i = 1, ..., N_B$, at time t with means $\mu(i) = \mathrm{E}[R_t(i)]$. The CR network, consisting of N_S secondary users (SUs), then selects a set of subbands $\mathcal{B}, |\mathcal{B}| = L \leq N_B$, to sense so that the expected sum rate is maximized. After choosing the subbands to be sensed, the CR network decides which SUs are assigned for sensing.

When lacking a parametric model for the dynamics of the PU activity, or if the model order may change abruptly, it is obviously not possible to find an optimal sensing policy. In such cases reinforcement learning (RL) methods become attractive. RL methods reinforce good actions by selecting more likely those actions that have recently provided large rewards. In a spectrum sensing policy actions correspond to selecting the frequency bands to be sensed and the SUs to do the sensing. The rewards for these actions are selected such that they reflect the achievable data rates from the subbands and the sensing performances of the SUs. An important feature of reinforcement learning methods is the *exploration-exploitation* trade-off, which emerges here when the CR has to decide whether to exploit the seemingly best frequency bands at the moment or to explore other bands in the hope of finding even better ones.

This paper extends our previous work in [1] by developing a practical heuristic algorithm for finding feasible sensing assignment for identifying idle spectrum. The contributions of this paper are the following. A novel reinforcement learning-based sensing policy with heuristic minimization of sensing resources is proposed. The proposed policy balances between exploitation and exploration using the ϵ-greedy method [2]: with probability ϵ (where typically ϵ is small) the policy goes into exploration phase and with probability $1 - \epsilon$ the policy goes into exploitation phase. Exploration and Exploitation phases will be elaborated in section 2. The sensing policy employs action-value learning to focus the search for idle frequencies to those parts of the spectrum that persistently provide high data rate. The sensing policy assigns the SUs for sensing such that the number of active SUs is minimized under a constraint on the global sensing performance. For the minimization problem a heuristic polynomial time algorithm that provides near-optimal solutions in practical scenarios is developed. The proposed algorithm works iteratively in a greedy manner by assigning one SU per subband using the Hungarian algorithm [3]. It is demonstrated that the heuristic algorithm finds near-optimal sensing assignments significantly faster than an exact branch-and-bound search.

The literature on spectrum sensing strategies for CR has obtained attention during the past few years. In [4, 5] parametric spectrum sensing policies are derived using the formalism of partially observable Markov decision processes (POMDPs). In [5] a closed form Whittle index policy for Markovian rewards was derived and shown to be optimal under certain conditions. In [6] a single-user RL-based sensing policy was proposed using soft max action selection. However, none of these works consider the effect of the sensing assignment in cooperative sensing. In [7] a sensor-mission assignment problem is considered with additive gains obtained by sensing assignment in a cooperative scenarios. However, in cooperative sensing the additivity assumption does not hold necessarily.

This paper is organized as follows. In section 2 the proposed sensing policy is presented and the sensing assignment problem is formulated. Section 2 also introduces a polynomial time algorithm for solving the sensing assignment problem. Section 3 provides simulations for the proposed heuristic algorithm and for the sensing policy. The paper is concluded in section 4.

2. THE PROPOSED SENSING POLICY

The proposed sensing policy is a cooperative action-value policy maintained at the fusion center (FC). The FC makes a global decision about the state of the spectrum using a fusion rule that combines the local binary decisions that individual SUs have communicated. In this paper the simple OR-rule is employed but any other fusion rule could be used as well. The sensing policy uses ϵ-greedy method

[2] for balancing between exploration and exploitation. The policy is managed by the FC that tracks two kinds of action-values: the $Q(b)$-values for sensing subband b and the $Q(s,b)$-values for assigning SU s to sense subband b. A natural way to define the reward $r_t(b)$ for selecting subband b to be sensed is the obtained throughput:

$$r_{t+1}(b) = \begin{cases} R_{t+1}(b), & \text{if } b \text{ is accessed and free,} \\ 0, & \text{if } b \text{ is occupied,} \end{cases} \quad (1)$$

where $R_{t+1}(b)$ is the instantaneous throughput at subband b at time $t+1$. It is assumed that the SU who has been granted the permission to access the band will feed back an estimate of the achieved throughput. For example, this may be based on the estimate of the channel quality between the two communicating SUs.

The reward for assigning SU s to sense subband b is:

$$r_{t+1}(s,b) = \begin{cases} d_{t+1}(s,b), & d_{t+1}(\mathrm{FC},b) = 1 \\ Q_t(s,b), & d_{t+1}(\mathrm{FC},b) = 0, \end{cases} \quad (2)$$

where $d_{t+1}(s,b)$ denotes the local decision by SU s for subband b and $d_{t+1}(\mathrm{FC},b)$ denotes the corresponding global decision at the FC. In this paper decision 1 means that the band is considered to be occupied and 0 that the band is considered to be idle. Hence, the rewards $r_{t+1}(s,b)$'s depend on the SUs' local detection probabilities assuming that the global decision at the FC is reliable.

After receiving the feedback from the SUs the $Q(b)$- and $Q(s,b)$-values are updated according to [2]

$$Q_{t+1}(b) = Q_t(b) + \alpha_1[r_{t+1}(b) - Q_t(b)], \quad (3)$$

$$Q_{t+1}(s,b) = Q_t(s,b) + \alpha_2[r_{t+1}(s,b) - Q_t(s,b)], \quad (4)$$

where α_1 and α_2 ($\alpha_1, \alpha_2 \in]0,1]$) are the step sizes.

2.1. Exploration

In order to gain information about the qualities of all parts of the spectrum and the local sensing performances the CR need to do exploration. In the exploration phase the spectrum is sensed according to pseudorandom patterns with fixed diversity order D (that is the number of SUs simultaneously sensing the same band) so that all subbands and all combinations of D spectrum sensors are considered in minimum time [8]. This diversity guarantees reliable decisions at the FC. Using pseudorandom frequency hopping provides quick scanning over the spectrum of interest with minimal control signaling, thus being extremely suitable for exploring the spectrum. The frequency hopping code design allows for trading off scanning speed and diversity (and consequently detector performance) in an elegant manner.

2.2. Exploitation

The utilization of the obtained information about the band qualities and local sensing performances is referred as exploitation. In the exploitation phase the FC selects the set of bands $\mathcal{B}$ to be sensed and a corresponding sensing assignment with some desired properties. For convenience time index t has been dropped in the rest of the paper. The exploitation phase is divided into two stages:
Stage1: *Select the set of subbands $\mathcal{B}$ with the largest $Q(b)$'s.*
Stage2: *Find a feasible sensing assignment for $\mathcal{B}$ using $Q(s,b)$'s.*
In [1] we showed that the update of $Q(s,b)$-value in (3) given the reward function in (2) converges to the local detection probability of SU s at band b provided that the probability of error (the probability of missed detection or false alarm) at the FC is low.

One possible criterion for choosing a sensing assignment is the total number of SUs assigned for sensing, which affects the SUs power consumption. In order to conserve the SUs' batteries we want to minimize the number of assigned SUs while pursuing to guarantee the desired level of detection performance. Denoting the set of SUs as $\mathcal{S}$ the sensing assignment problem (SAP) can be formulated as

$$\begin{aligned} \min_{x(s,b)} \quad & \sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}} w(s)x(s,b) \quad (5) \\ \text{s.t.} \quad & P_{FC}(Q(s,b),\mathbf{X}) \geq P_{d,target}, \forall b \in \mathcal{B} \\ & \sum_{b\in\mathcal{B}} x(s,b) \leq K(s), \forall s \in \mathcal{S} \\ & x(s,b) \in \{0,1\}. \end{aligned}$$

In this paper Neyman-Pearson detectors are used, that maximize the detection probability under a constraint on the false alarm rate and the false alarm rate is not included in (5) as a separate constraint. In (5) $P_{FC}(Q(s,b),\mathbf{X})$ is the estimated detection probability at the FC at band b and $P_{d,target}$ the minimum probability of detection that the FC is allowed to have. $K(s)$ is the number of bands SU s can sense simultaneously and $w(s)$ is the cost of user s that may be used to favor certain sensing assignments. The costs may be chosen, for example, according to the SUs' battery charge so that if a SU has a low battery charge it may be given relatively large costs compared to the other SUs. The unknown $N_S \times L$ binary sensing assignment matrix is denoted as $[\mathbf{X}]_{s,b} = x(s,b)$, where $x(s,b) = 1$ if SU s is assigned to sense subband b and $x(s,b) = 0$ otherwise.

Generally, (5) is an NP-hard problem. Using a branch-and-bound (BB) type algorithm the worst case running time is $2^{N_S L}$ (corresponding to the case when pruning of the search tree is not possible). However, in practical scenarios with many SUs, the probability that there exists multiple optimal or near-optimal assignments is high. In such cases heuristic search algorithms are likely to find reasonably good sensing assignments. In [9] a polynomial time algorithm is proposed for finding sensing assignments that minimize the probability of missed detection at the FC. In each round the algorithm in [9] assigns SUs to sense the subbands using the Hungarian method until all SUs have been assigned. In this paper, however, the minimum number of SUs that are able to meet a desired detection probability is found. To this end, the Hungarian method is employed iteratively to assign SUs to subbands one by one until a feasible solution is found.

2.2.1. An algorithm for solving the SAP

In this paper we propose a heuristic algorithm for solving the SAP in (5). Here the simple OR-rule is used but any other fusion rule could be applied as well. The listing of the proposed iterative Hungarian algorithm (IH) is shown in algorithm 1. The algorithm takes as inputs the $Q(s,b)$-values, $\mathcal{B}$, $\mathcal{S}$, $w(s)$ and $P_{d,target}$, and outputs the binary sensing assignment matrix $\mathbf{X}$. At time instant $t = 0$ the $Q(s,b)$-values are initialized randomly between 0 and 1. The basic idea of the algorithm is to in each round to assign one SU to each subband in $\mathcal{B}$ using the Hungarian method [3]. The Hungarian method is a strongly polynomial time algorithm for finding maximum (or minimum) weight matching (e.g. assigning workers with certain qualifications for different jobs so that the total quality of work is maximized with the constraint that each job is assigned to exactly one worker). Here the SUs assigned for each subband are the ones that increase the sum of weighted probabilities of detection the

Algorithm 1 Iterative Hungarian method for solving the SAP in (5) when the OR-rule is used at the FC.

- Step 1: initialize: $\mathbf{X} = \mathbf{0}_{N_S \times L}$.
- Step 2: initialize: $Q'(s,b) = Q(s,b), b \in \mathcal{B}$.

while a feasible solution has not been found **do**

// Solve the maximum weight matching using the Hungarian method.

$$[\mathbf{X}']_{s,b} = \arg\max_{x'(s,b)} \sum_s \sum_b w(s) Q'(s,b) x'(s,b)$$
$$\text{s.t.} \sum_s x'(s,b) = 1, \forall b \in \mathcal{B}$$
$$x'(s,b) \in \{0,1\}.$$

// Update the sensing assignment.
$\mathbf{X} = \mathbf{X} + \mathbf{X}'$.
// Calculate the obtained detection probabilities for the bands.
$P_{FC}(b) = 1 - \prod_s (1 - Q(s,b))^{x(s,b)}$.
// Re-weight the $Q'(s,b)$-values
$Q'(s,b) = Q(s,b)(1 - P_{FC}(b))$.
// Set $Q(s,b)$-values that cannot be assigned anymore to $-\infty$.
if $P_{FC}(b) \geq P_{d,target}$ **then**
  $Q'(s,b) = -\infty, \forall s$.
end if
if $\sum_b x(s,b) = K(s)$ **then**
  $Q'(s,b) = -\infty, \forall b$.
end if
// Check if all $Q'(s,b)$ are -∞.
if $Q'(s,b) = -\infty, \forall s, b$ **then**
  if $P_{FC}(b) \geq P_{d,target}, \forall b$ **then**
    $\mathbf{X}$ is feasible. Return $\mathbf{X}$.
  else
    Remove band $\tilde{b}$ from $\mathcal{B}$ and go to step 1.
  end if
end if
end while

most. At those bands where the current sensing assignment achieves the desired sensing performance $P_{d,target}$ the $Q(s,b)$-values are set to $-\infty$. This guarantees that none of the remaining SUs will be assigned to those subbands. At those bands where the current sensing assignment does not yet meet the detection performance constraint the $Q(s,b)$-values are re-weighted according to how good the sensing performance would be with the current iterated sensing assignment. Similarly, for SUs that have already been assigned to sense $K(s)$ subbands the $Q(s,b)$-values are set to $-\infty$. As soon as a feasible sensing assignment is found the algorithm is stopped and $\mathbf{X}$ is returned. If no feasible sensing assignment is found the algorithm removes one subband $\tilde{b}$ from $\mathcal{B}$ and tries to find a feasible sensing assignment for the new set of bands $\mathcal{B} \backslash \tilde{b}$. There are many ways for selecting the subband to be removed. In this paper we remove the band that essentially has the smallest product of Q-values $\tilde{b} = \arg\min_{b \in \mathcal{B}} Q(b)[1 - \prod_S (1 - Q(s,b))]$, which reflects the product of the mean throughput of the band and the detection probability of the band if all SUs would be sensing it.

3. SIMULATIONS

3.1. Performance of the heuristic algorithm

We compare the performance of the proposed heuristic iterative (IH) search algorithm to the performance of an exact branch-and-bound algorithm using the OR-rule. For the BB algorithm the SAP is written as a linear binary integer program (BIP) [1]. Random $Q(s,b) \sim$ U$[0,1]$ values are generated for $N_s = 4$ SUs and $L = 1, 2, 3$ subbands with $P_{d,target} = 0.9$, $K(s) = 1$ and $w(s) = 1$. Only the cases where there exists at least one feasible assignment for L bands are considered (consequently, the BB-algorithm always finds an optimal sensing assignment). The algorithms were run using Matlab-software on a 3.00 GHz processor. Table 1 shows the ratios of each corresponding measure given in the left most column for the IH- and the BB-search. Variable T_r denotes the time to find a feasible solution for sensing L subbands and Min the corresponding number of active sensors. Variable $N_{B,s}$ denotes the number of bands sensed on average and $\#(N_{B,s} = L)$ the number of cases where a feasible assignment was found for exactly L bands (i.e. the number of times IH did not have to remove any band from $\mathcal{B}$). For all results, except for the mean run time T_r, the performance of the proposed heuristic algorithm is the better the closer the value is to 1. Respectively, the closer T_r is to 0, the faster the heuristic algorithm is compared to the exact BB search. The results were averaged over 1000 random sets of $Q(s,b)$'s for which there existed at least one feasible solution for sensing L bands. From the sensing policy point of view the most important measures are the first three rows, i.e. the run time, the number of active sensors and the number of bands sensed. It can be seen that the heuristic algorithm is able to find sensing assignments that are close to the optimal solution significantly faster. In table 2 we show the corresponding results for a harder problem when $Q(s,b) \sim [0, 0.9]$, $N_S = 10$ and $L = 1, ..., 5$ while keeping the other parameters the same as before. For this case the results for Min, $N_{B,s}$ and $\#(N_{B,s} = L)$ are slightly worse, but still reasonably good compared to the gain in the mean run time T_r.

	$L=1$	$L=2$	$L=3$
T_r	0.071	0.058	0.038
Min	1	1.002	1.0003
$N_{B,s}$	1	0.982	0.988
$\#(N_{B,s} = L)$	1	0.965	0.965

Table 1. Performance ratio of the IH- and BB-algorithms when $Q(s,b) \sim$ U$[0,1]$. The assignments found by the IH-algorithm are very close to the exact minimized assignment with a significant reduction in the computation time. T_r denotes the runtime of the algorithm when a feasible solution has been found for sensing L subbands and Min the corresponding number of active sensors. Variable $N_{B,s}$ denotes the number of bands sensed on average and $\#(N_{B,s} = L)$ the number of cases where a feasible assignment was found for exactly L bands.

	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$
T_e	0.13	0.03	0.004	0.0009	0.0006
Min	1	1.01	1.01	1.02	1
$N_{B,s}$	1	1	0.99	0.96	0.93
$\#(N_{B,s} = L)$	1	1	0.96	0.84	0.63

Table 2. Performance ratio of the IH- and BB-algorithms when $Q(s,b) \sim$ U$[0, 0.9]$. The assignments found by the IH-algorithm are close to the exact minimized assignment with a significant reduction in the computation time.

3.2. Performance of the sensing policy

We consider the throughput of the CR network and the mean number of active sensors when $N_S = 6$, $N_B = 10$, $L = 2$, $K(s) = 1$, $w(s) = 1$, $P_{d,target} = 0.9$ and $D = 2$. We consider the sensing policy with both the IH- and BB-algorithms. The parameters

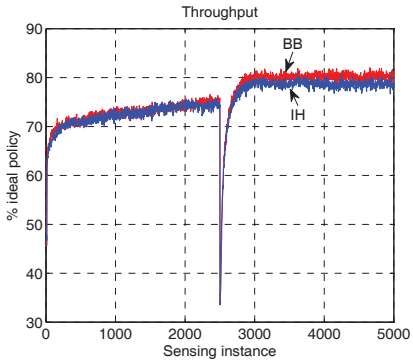

Fig. 1. Obtained throughput relative to an ideal policy for the proposed heuristic minimization (IH) and the exact minimization (BB). It can be seen that the performance of the policy with the heuristic method is roughly the same as with the exact method.

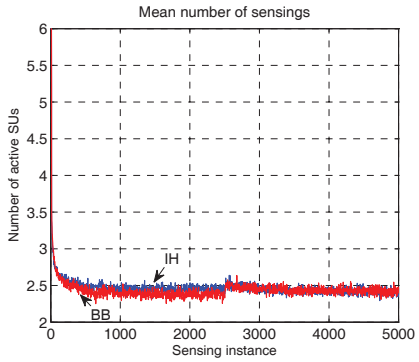

Fig. 2. The average number of assigned sensors by the policy using the proposed heuristic algorithm (IH) and the exact search (BB). It can be seen that the number of active sensors is not affected by using the proposed heuristic method.

for the action-value updates are $\epsilon = 0.1$ and $\alpha_1 = \alpha_2 = 0.1$, where α_1 is the step size for the $Q(b)$-values and α_2 the step size for the $Q(s, b)$-values. The mean SNRs of different SUs at different bands are normally distributed with mean 0 dB and standard deviation 9 dB. The local detectors are Neyman-Pearson energy detectors using 50 samples and the false alarm rate at the FC is set to 0.01. The PU activities at the bands are assumed to be independent two-state Markov processes with an "idle" and "occupied" state. The mean stationary throughputs of the bands are $\mu(b) = [6.4, 0.6, 2.3, 3.7, 0.7, 0.8, 0.9, 2.0, 8.7, 0.4]$. To illustrate the sensing policy in a non-stationary environment at time instant $t = 2500$ the throughputs are randomly permuted among the subbands and the mean SNRs are randomly permuted among the SUs at a given band. Both algorithm satisfy the detection probability constraint in steady state. Figure 1 shows the average throughput of the CR network compared to an ideal, genie aided, policy. An ideal policy always knows the states and throughputs of the bands and is therefore able to select the best available bands for sensing. It can be seen that the policy is able to provide high throughput and to quickly re-adapt to the changes in the environment. Furthermore, the obtained throughput is approximately the same for the exact search and the heuristic search, while the runtime of the heuristic algorithm was on average only 5 % of the runtime of the BB-algorithm. Note that here it is assumed that the throughput is not affected by the computation time of the optimal or near-optimal action, as it would be in practice and thus favoring the IH-algorithm even more.

Figure 2 shows the average number of active sensors assigned by the policy using the IH- and BB-algorithm. With both algorithms the policy assigns on average roughly 2.5 SUs for sensing. The results indicate that the IH algorithm is able to find minimal assignments most of the time.

4. CONCLUSIONS

In this paper a centrally-controlled RL-based cooperative sensing policy has been proposed for CR. The policy learns to focus the search for idle frequency bands to those parts of the spectrum that persistently provide high data rate. Furthermore, the policy learns about the individual sensing performances of different SUs and uses this information to minimize the sensing resources. For the minimization problem a heuristic polynomial time algorithm that iteratively employs the Hungarian method has been proposed. The simulation results have shown that the proposed algorithm is able to find near-optimal solutions in practical scenarios significantly faster than an exact branch-and-bound type search. In dynamic environment it is important to save time from the policy computation to the actual exploitation of the spectral opportunities.

5. REFERENCES

[1] J. Oksanen, J. Lundén, and V. Koivunen, "Reinforcement learning based sensing policy optimization for energy efficient cognitive radio networks," *Neurocomputing – Special Issue on Machine Learning for Signal Processing 2010*, vol. 80, no. 0, pp. 102 – 110, March 2012.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 1998.

[3] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Res. Logistics Q.*, vol. 2, pp. 83–97, 1955.

[4] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality and performance," *IEEE Trans. Wireless Commun.*, pp. 5431–5440, Dec. 2008.

[5] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, pp. 5547–5567, Nov. 2010.

[6] U. Berthold, F. Fu, M. van der Schaar, and F. Jondral, "Detection of Spectral Resources in Cognitive Radios Using Reinforcement Learning," in *Proc. DySPAN, Chicago, IL*, Oct. 2008, pp. 1–5.

[7] M. P. Johnson, H. Rowaihy, D. Pizzocaro, A. Bar-Noy, S. Chalmers, T. La Porta, and A. Preece, "Sensor-mission assignment in constrained environments," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 11, pp. 1692 –1705, Nov. 2010.

[8] J. Oksanen, V. Koivunen, J. Lundén, and A. Huttunen, "Diversity-based spectrum sensing policy for detecting primary signal over multiple frequency bands," in *Proc. ICASSP, Dallas, TX*, Mar. 2010, pp. 3130–3133.

[9] Z. Wang, Z. Feng, and P. Zhang, "An Iterative Hungarian Algorithm Based Coordinated Spectrum Sensing Strategy," *IEEE Commun. Lett.*, vol. 15, no. 1, pp. 49–51, Jan. 2011.