Int. J. Communications, Network and System Sciences, 2009, 7, 669-674
doi:10.4236/ijcns.2009.27077 Published Online October 2009 (http://www.SciRP.org/journal/ijcns/).

Q-Learning-Based Adaptive Waveform Selection in Cognitive Radar

Bin WANG, Jinkuan WANG, Xin SONG, Fulai LIU
Northeastern University, Shenyang, China
Email: wangbin_neu@yahoo.com.cn

Received June 11, 2009; *revised August* 12, 2009; *accepted September* 20, 2009

ABSTRACT

Cognitive radar is a new framework of radar system proposed by Simon Haykin recently. Adaptive waveform selection is an important problem of intelligent transmitter in cognitive radar. In this paper, the problem of adaptive waveform selection is modeled as stochastic dynamic programming model. Then Q-learning is used to solve it. Q-learning can solve the problems that we do not know the explicit knowledge of state-transition probabilities. The simulation results demonstrate that this method approaches the optimal waveform selection scheme and has lower uncertainty of state estimation compared to fixed waveform. Finally, the whole paper is summarized.

Keywords: Waveform Selection; Q-Learning; Space Division; Cognitive Radar

1. Introduction

Radar is the name of an electronic system used for the detection and location of objects. Radar development was accelerated during World War Ⅱ. Since that time it has continued such that present-day systems are very sophisticated and advanced. Cognitive radar is an intelligent form of radar system proposed by Simon Haykin and it has many advantages [1]. However, cognitive radar is only an ideal framework of radar system, and there are many problems need to be solved.

Adaptive waveform selection is an important problem in cognitive radar, with the aim of selecting the optimal waveform and tracking targets with more accuracy according to different environment. In [2], it is shown that tracking errors are highly dependent on the waveforms used and in many situations tracking performance using a good heterogeneous waveform is improved by an order of magnitude when compared with a scheme using a homogeneous pulse with the same energy. In [3], an adaptive waveform selective probabilistic data association algorithm for tracking a single target in clutter is presented. The problem of waveform selection can be thought of as a sensor scheduling problem, as each possible waveform provides a different means of measuring the environment, and related works have been examined in [4,5]. In [6], radar waveform selection algorithms for tracking accelerating targets are considered. In [7], genetic algorithms are used to perform waveform selection utilizing the autocorrelation and ambiguity functions in the fitness evaluation. In [8], Incremental Pruning method is used to solve the problem of adaptive waveform selection for target detection. The problem of optimal adaptive waveform selection for target tracking is also presented in [9].

In this paper, the problem of adaptive waveform selection in cognitive radar is viewed as a problem of stochastic dynamic programming and Q-learning is used to solve it.

2. Division in Radar Beam Space

The most important parameters that a radar measures for a target are range, Doppler frequency, and two orthogonal space angles. However, in most circumstances, angle resolution can be considered independently from range and Doppler resolution. We may envision a radar resolution cell that contains a certain two-dimensional hypervolume that defines resolution.

Figure 1 is abridged general view of range and Doppler. Range resolution, denoted as ΔR, is a radar metric that describes its ability to detect targets in close proximity to each other as distinct objects. Radar systems are normally designed to operate between a minimum range $R_{\min}$, and maximum range $R_{\max}$. Targets seperated by at least ΔR will be completely resolved in range. Radars use Doppler frequency to extract target radial velocity (range rate), as well as to distinguish moving and stationary

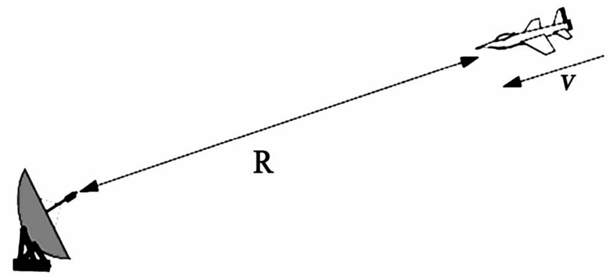

Figure 1. A closing target.

targets or objects such as clutter. The Doppler phenomenon describes the shift in the center frequency of an incident waveform.

In general, a waveform can be tailored to achieve either good Doppler or good range resolution, but not both simultaneously. So we need to consider the problem of adaptive waveform scheduling. The basic scheme for adaptive waveform scheduling is to define a cost function that describes the cost of observing a target in a particular location for each individual pulse and select the waveform that optimizes this function on a pulse by pulse basis.

We make no assumptions about the number of targets that may be present. We divide the area covered by a particular radar beam into a grid in range-Doppler space, with the cells in range indexed by t=1,…,N and those in Doppler indexed by v=1,…,M. There may be 0 target, 1 target or NM targets. So

$$C_{NM}^{0}+C_{NM}^{1}+C_{NM}^{2}+...+C_{NM}^{NM-1}+C_{NM}^{NM}=2^{NM} \tag{1}$$

The number of possible scenes or hypotheses about the radar scene is 2^{NM}. Let the space of hypotheses be denoted by X. The state of our model is X_t=x where $x\in X$. Let Y_t be the measurement variable. Let u_t be the control variable that indicates which waveform is chosen at time t to generate measurement Y_{t+1}, where $u_t\in U$. The probability of receiving a particular measurement X_t=x will depend on both the true, underlying scene and on the choice of waveform used to generate the measurement.

We define $a_{x'x}$ is state transition probability where

$$a_{x'x}=P(x_{t+1}=x'\,|\,x_t=x) \tag{2}$$

We define $b_{x'x}$ is the measurement probability where

$$b_{x'x}(u_t)=P(Y_{t+1}=x'\,|\,X_t=x,u_t) \tag{3}$$

Assume the transmitted baseband signal is $s(t)$, and the received baseband signal is $r(t)$. The matched filter is the one with an impulse response $h(t)$=$s^*(-t)$, so an output process of our matched filter is

$$x(t)=\int s^*(\lambda-t)r(\lambda)d\lambda \tag{4}$$

In the radar case, the return signal is expected to be Doppler shifted, then the matched filter to a return signal with an expected frequency shift v_0 has an impulse response

$$h(t)=s^*(-t)e^{j2\pi v_0 t} \tag{5}$$

The output is given by

$$x(t)=\int s^*(\lambda-t)e^{-j2\pi v_0(\lambda-t)}r(\lambda)d\lambda \tag{6}$$

where v_0 is an expected frequency shift.

The baseband received signal will be modeled as a return from a Swerling target:

$$r(t)=As(t-\tau)e^{j2\pi v_d t}I+n(t) \tag{7}$$

where $s(t,\tau,v_d)=s(t-\tau)e^{j2\pi v_d t}$ is a delayed t and Doppler-shifted v_d replica of the emitted baseband complex envelope signal $s(t)$; I is a target indicator. A approaches a complex Gassian random variable with zero mean and variance $2\sigma_A^2$. We assume $n(t)$ is complex white Gaussian noise independent of A, with zero mean and variance $2N_0$.

At time t the magnitude square of the output of a filter matched to a zero delay and a zero Doppler shift is

$$|x(t)|^2=\left|\int_0^t r(\lambda)s^*(\lambda-t)d\lambda\right|^2 \tag{8}$$

When there is no target

$$r(t)=v(t) \tag{9}$$

So

$$x(\tau_0)=\int_0^{\tau_0} n(\lambda)s^*(\lambda-\tau_0)d\lambda \tag{10}$$

The random variable $x(\tau_0)$. is complex Gaussian, with zero mean and variance given by

$$\sigma_0^2=E\left\{x(\tau_0)x^*(\tau_0)\right\}=2N_0\xi \tag{11}$$

ξ is the energy of the transmitted pulse.

When target is present

$$r(t)=As(t-\tau)e^{j2\pi v_d t}I+n(t) \tag{12}$$

$$x(\tau_0)=\int_0^{\tau_0}\left[As(\lambda-\tau)e^{j2\pi v_d\lambda}+n(\lambda)\right]s^*(\lambda-\tau_0)d\lambda \tag{13}$$

This random variable is still zero mean, with variance given by

$$\begin{aligned}\sigma_1^2&=E\left\{x(\tau_0)x^*(\tau_0)\right\}\\&=\sigma_0^2(1+\frac{2\sigma_A^2\xi^2}{\sigma_0^2}A(\tau_0-\tau,v_0-v))\end{aligned} \tag{14}$$

$A(t,v)$is ambiguity function, given by

$$A(\tau,v)=\frac{1}{\left(\int|s(\lambda)|^2d\lambda\right)^2}\left|\int s(\lambda)s^*(\lambda-\tau)e^{j2\pi v\lambda}d\lambda\right|^2 \tag{15}$$

Recall that the magnitude square of a complex Gaussian random variable $x\sim N(0,\sigma_i^2)$ is exponentially distributed, with density given by

$$y = x^2 \sim \frac{1}{2\sigma_i^2} e^{-\frac{y}{2\sigma_i^2}} \quad \textbf{(16)}$$

We consequently have that the probability of false alarm P_f is given by

$$P_f = \int_D^\infty \frac{1}{2\sigma_0^2} e^{-\frac{x}{2\sigma_0^2}} dx = e^{-\frac{D}{2\sigma_0^2}} \quad \textbf{(17)}$$

And the probability of detection P_d by

$$P_d = \int_D^\infty \frac{1}{2\sigma_1^2} e^{-\frac{x}{2\sigma_1^2}} dx = e^{-\frac{D}{2\sigma_0^2(1+\frac{2\sigma_A^2\xi^2}{\sigma_0^2}A(\tau_0-\tau,\nu_0-\nu))}} \quad \textbf{(18)}$$

In the case when a target is present in cell (τ, υ), assuming its actual location in the cell has a uniform distribution

$$P_d = \frac{1}{|A|}\int_{(\tau_a,\upsilon_a \in A)} e^{-\frac{D}{2\sigma_0^2(1+\frac{2\sigma_A^2\xi^2}{\sigma_0^2}A(\tau_0-\tau,\nu_0-\nu))}} d\tau_a d\upsilon_a \quad \textbf{(19)}$$

where A is the resolution cell centred on (t,v) with volume $|A|$.

3. Q-Learning-Based Stochastic Dynamic Programming

A target for which measurements are to be made will fall in a resolution cell. Another target, conceptually, does not interfere with measurements on the first if it occupies another resolution cell different from the first. Thus, conceptually, as long as each target occupies a resolution cell and the cells are all disjoint, the radar can make measurements on each target free of interference from others.

Define $\pi = \{u_0, u_1, ..., u_T\}$ where T=1 is the maximum number of dwells that can be used to detect and confirm targets for a given beam. Then π is a sequence of waveforms that could be used for that decision process.

We can obtain different π according to different environment in cognitive radar. Let

$$V_t(X_t) = E[\sum_{t=0}^{T} \gamma^t R(X_t, u_t)] \quad \textbf{(20)}$$

where R（X_t,u_t） is the reward earned when the scene X_t is observed using waveform u_t and γ is discount factor. Then the aim of our problem is to find the sequence π^* that satisfies

$$V^*(X_t) = \max_{\pi} E[\sum_{t=0}^{T} \gamma^t R(X_t, u_t)] \quad \textbf{(21)}$$

However, knowledge of the actual state is not available. Using the method of [10], we can obtain that the optimal control policy π^* that is the solution of (21) is also the solution of

$$V^*(\mathbf{p}(0)) = \max_{\pi} E[\sum_{t=0}^{T} \gamma^t R(\mathbf{p}_t, u_t)] \quad \textbf{(22)}$$

where $\mathbf{P}_t$ is the conditional density of the state given the measurements and the controls and $\mathbf{P}_0$ is the a priori probability density of the scene. **P** is a sufficient statistic for the true state X_t. So we need to solve the following problem

$$\max_{\pi} E[\sum_{t=0}^{T} \gamma^t R(\mathbf{p}_t, u_t)] \quad \textbf{(23)}$$

The refreshment formula of $\mathbf{P}_t$ is given by

$$\mathbf{p}_{t+1} = \frac{\mathbf{BAp}_t}{\mathbf{1'LAp}_t} \quad \textbf{(24)}$$

where **B** is the diagonal matrix with the vector $(b_{x'x}(u_t))$ the non-zero elements and **1** is a column vector of ones. **A** is state transition matrix.

If we wanted to solve this problem using classical dynamic programming, we could have to find the value function $V_t(\mathbf{p}_t)$ using

$$V_t(\mathbf{p}_t) = \max_{u_t}(R_t(\mathbf{p}_t, u_t) + \gamma E\{V_{t+1}(\mathbf{p}_{t+1}) \mid \mathbf{p}_t\}) \quad \textbf{(25)}$$

It can also be written in probability form

$$V_t\ (\mathbf{p}_t) = \max_{u_t}(R_t(\mathbf{p}_t, u_t) + \gamma \sum_{\mathbf{p}' \in \mathbf{P}} P(\mathbf{p}' | \mathbf{p}_t, u_t) V_{t+1}(\mathbf{p}')) \textbf{(26)}$$

However, in radar scene, explicit knowledge of target state-transition probabilities are unknown. So directly using Bellman's dynamic programming is very hard. The Q-leaning algorithm is a direct approximation of Bellman's dynamic programming, and it can solve the problem that we do not know explicit knowledge of state-transition probabilities. For this reason, Q-learning is very suitable to be used in the problem of adaptive waveform selection in cognitive radar.

We define a Q-factor in our problem. For a state-action pair $(\mathbf{p}_t, u_t)$,

$$Q(\mathbf{p}_t, u_t) = \sum_{\mathbf{p}' \in \mathbf{P}} P(\mathbf{p}' | \mathbf{p}_t, u_t)[R_t(\mathbf{p}' | \mathbf{p}_t, u_t) + \gamma V_{t+1}] \quad \textbf{(27)}$$

According to (26), (27) we can derive

$$V_t^* = \max_{u_t} Q(\mathbf{p}_t, u_t) \quad \textbf{(28)}$$

The above establishes the relationship between the value function of a state and the Q-factors associated with a state. Then it should be clear that, if the Q-factors are known, one can obtain the value function of a given state from above fomula.

So Q form of Bellman equation is

$$Q(\mathbf{p}_t, u_t) = \sum_{\mathbf{p}' \in \mathbf{P}} P(\mathbf{p}' | \mathbf{p}_t, u_t)[R_t(\mathbf{p}' | \mathbf{p}_t, u_t) + \gamma \max_{u_{t+1}} Q(\mathbf{p}_{t+1}, u_{t+1})] \quad \textbf{(29)}$$

Let us denote the *i*th independent sample of a random variable *X* by S^i and the expected value by *E*(*X*). X^n is the estimate of *X* in the *n* th iteration. So

$$E(X)=\lim_{n\to\infty}\frac{\sum_{i=1}^{n}s^i}{n} \tag{30}$$

$$X^n=\frac{\sum_{i=1}^{n}s^i}{n} \tag{31}$$

We can derive

$$X^{n+1}=(1-\alpha^{n+1})X^n+\alpha^{n+1}s^{n+1} \tag{32}$$

where

$$\alpha^{n+1}=\frac{1}{n+1} \tag{33}$$

So

$$Q(\mathbf{p}_t,u_t)=E[R_t(\mathbf{p}'|\mathbf{p}_t,u_t)+\gamma\max_{u_{t+1}}Q(\mathbf{p}_{t+1},u_{t+1})] \tag{34}$$

where *E* is the expectation operator. We could use this scheme in a simulator to estimate the same Q-factor. Using this algorithm, Equation (29) becomes:

$$\begin{aligned}&Q^{n+1}(\mathbf{p}_t,u_t)\leftarrow(1-\alpha^{n+1})Q^n(\mathbf{p}_t,u_t)\\&+\alpha^{n+1}[R_t(\mathbf{p}'|\mathbf{p}_t,u_t)+\gamma\max_{u_{t+1}}Q^n(\mathbf{p}_{t+1},u_{t+1})]\end{aligned} \tag{35}$$

Obviously, we do not have the transition probabilities in it.

Our Q-learning algorithm is as follows:

Step 1. Initialize the Q-factors to 0. Set *n*=1.

Step 2. For *t*=0,1,…*T*,do step 3-step 6.

Step 3. Simulation action u_t. Let the curren state be $\mathbf{P}_t$, and the next state be $\mathbf{P}_{t+1}$.

Step 4. Find the decision using the current Q-factors:

$$u_t=\arg\max_{u_t}Q_t^{n-1}(\mathbf{p}_t^n,u_t) \tag{36}$$

Step 5. Update Q($\mathbf{P}_t$,u_t) using the following equation:

$$\begin{aligned}&Q^{n+1}(\mathbf{p}_t,u_t)\leftarrow(1-\alpha^{n+1})Q^n(\mathbf{p}_t,u_t)\\&+\alpha^{n+1}[R_t(\mathbf{p}'|\mathbf{p}_t,u_t)+\gamma\max_{u_{t+1}}Q^n(\mathbf{p}_{t+1},u_{t+1})]\end{aligned} \tag{37}$$

Step 6. Find the next state:

$$\mathbf{p}_{t+1}=\frac{\mathbf{BAp}_t}{\mathbf{1'BAp}_t} \tag{38}$$

Step 6. Increment *n*. If *n*<*N*, go to step 2.

Step 7. For each **P** $t\in$P, select

$$d(\mathbf{p}_t)\in\arg\max_{u_t}Q(\mathbf{p}_{t+1},u_{t+1}) \tag{39}$$

The policy generated by the algorithn is $\hat{d}$. Stop.

4. Simulation

In this section, we make three experiments. In order to explain the necessity of waveform selection, we make the curve of measurement probability versus SNR of three waveforms. Curve of uncertainty of state estimation demonstrates validity of our proposed algorithm. We also plot the figure of Q value space versus state and waveform.

We consider a simple situation. The state space is 4× 4. We consider 5 different waveforms where for each waveform u, and each hypotheses for the target x, the distribution of x' is given in Table 1. The discount factor γ=0.9. State transition matrix **A** is given by

$$\mathbf{A}=\begin{bmatrix}0.96&0.02&0.01&0.04\\0.01&0.93&0.03&0.04\\0.02&0.03&0.95&0.02\\0.01&0.02&0.01&0.9\end{bmatrix} \tag{40}$$

Following the approach described in [11,12], linear form of reward function will be adopted:

$$R(\mathbf{p},u)=\mathbf{p}'\mathbf{p}-1 \tag{41}$$

The formula *E*(–*R*) can be considered as the uncertainty in the state estimation. In other words, it can be considered as the tracking errors.

Figure 2 is curve of measurement probability versus SNR of three waveforms. From this curve we can see that with the same SNR, different waveforms correspond to different measurement probability. Generally speaking, the waveform with wide pulse duration corresponds to high measurement probability. From this point of view, the waveform with wide pulse duration is better. However, wide pulse duration means large energy of the transmitted pulse. So we should improve measurement

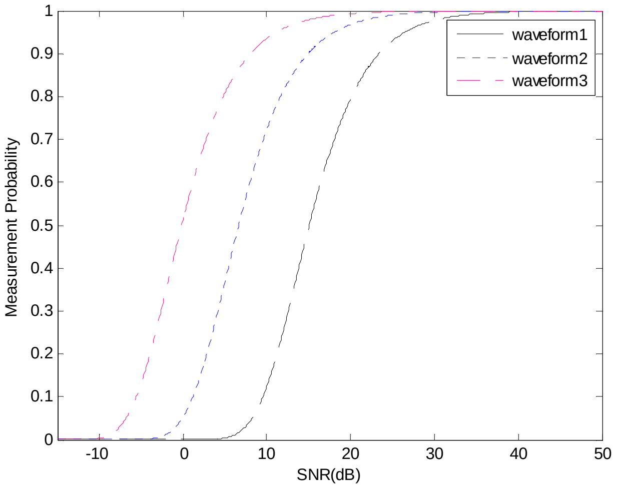

Figure 2. Curve of measurement probability versus SNR of three waveforms.

Table 1. Measurement probabilities for the example scenario.

	x=1	x=2	x=3	x=4
	x’=1,2,3,4	x’=1,2,3,4	x’=1,2,3,4	x’=1,2,3,4
u=1	0.97,0.01 0.01,0.01	0.01,0.01 0.96,0.02	0.01,0.02, 0.01,0.96	0.96,0.01, 0.01,0.02
u=2	0.96,0.01 0.02,0.01	0.02,0.95 0.01,0.02	0.01,0.01, 0.01,0.97	0.02,0.96, 0.01,0.01
u=3	0.94,0.02 0.03,0.01	0.02,0.02 0.01,0.95	0.02,0.96, 0.01,0.97	0.01,0.02, 0.95,0.02
u=4	0.96,0.01 0.01,0.02	0.01,0.02 0.96,0.01	0.97,0.01, 0.01,0.01	0.03,0.95, 0.01,0.01
u=5	0.95,0.02 0.01,0.02	0.01,0.97 0.01,0.01	0.02,0.01 0.96,0.01	0.04,0.94 0.01,0.01

probability through changing waveforms according to different environment and make a balance between the width of pulse duration and the energy of the transmitted pulse. We can also derive measurement probabilities for the example scenario from this curve, as is shown in Table 1.

Figure 3 is curve of uncertainty of state estimation. From this curve we can see that for all the cases, the uncertainty of state estimation is decreasing with time, no matter how the state is changing with time. Compared to a fixed waveform, Q-learning algorithm we proposed has lower uncertainty of state estimation. That means our algorithm will reduce uncertainty in locating targets. Meanwhile our algorithm approaches the optimal waveform selection scheme even though explicit knowledge of state-transition probabilities is unknown.

Figure 4 is the figure of Q value space versus state and waveform. Q value of different state-waveform pair can be obtained in this figure. We can see that the proposed algorithm has lower computational cost.

5. Conclusions

Adaptive waveform selection is an important problem in cognitive radar and the problem of adaptive waveform scheduling can be viewed as a stochastic dynamic programming problem. In this paper, Q-learning-based waveform selecting algorithm is proposed. The advantages of Q-learning over fixed waveform have been shown with simulations. The Q-learning algorithm can minimize the uncertainty of state estimation compared to fixed waveform and approaches the optimal waveform selection scheme. Meanwhile, Q-learning can solve the problems in which explicit knowledge of state-transition probabilities are unknown. Reasearch on alogorithms which approach the optimal waveform selection scheme and has lower computational cost is an important problem.

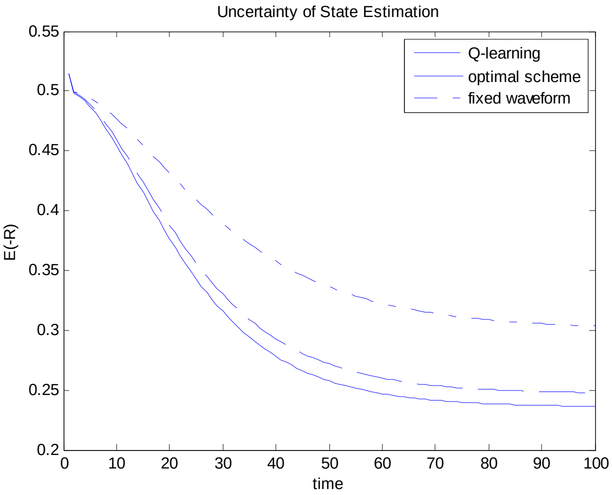

Figure 3. Curve of uncertainty of state estimation.

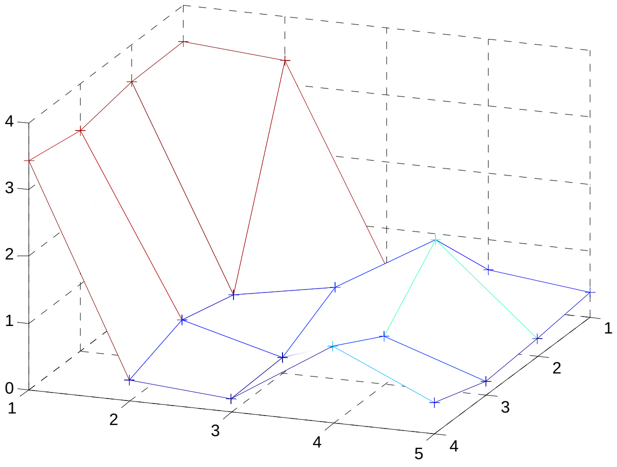

Figure 4. Q value space versus state and waveform.

6. References

[1] S. Haykin, “Cognitive radar: A way of the future,” IEEE Signal Processing Magazine, Vol. 23, No. 1, pp. 30–40, 2006.

[2] C. Rago, P. Willett, and Y. Bar-Shalom, “Detecting-tracking performance with combined waveforms,” IEEE Transactions on Aerospace and Electronic Systems, Vol. 34, No. 2, pp. 612–624, 1998.

[3] D. J. Kershaw and R. J. Evans, “Waveform selective probabilistic data association,” IEEE Transactions on Aerospace and Electronic Systems, Vol. 33, No. 4, pp. 1180–1188, 1997.

[4] Y. He and E. K. P. Chong, “Sensor scheduling for target tracking in sensor networks,” 43rd IEEE Conference on Decision and Control, Paradise, Island, Bahamas, pp. 743–748, 2004.

[5] V. Krishnamurthy, “Algorithms for optimal scheduling of hidden Markov model sensors,” IEEE Trans. on Signal Processing, Vol. 50, No. 6, pp.1382–1397, 2002.

[6] C. O. Savage, and B. Moran, “Waveform selection for maneuvering targets within an IMM framework,” IEEE Transactions on Aerospace and Electronic Systems, Vol. 43, No. 3, pp. 1205–1214, 2007.

[7] C. T. Capraro, I. Bradaric, G. T. Capraro, and T. K. Lue,

"Using genetic algorithms for radar selection," 2008 IEEE Radar Conference, Inc., Utica, NY, pp. 1–6, May 2008.

[8] B. F. La Scala and R. J. Moran Wand Evans, "Optimal adaptive waveform selection for target detection," The International Conference on Radar, Adelaide, SA, Australia, pp. 492–496, Sept. 2003.

[9] La Scala, Rezaeian, and Moran, "Optimal adaptive waveform selection for target tracking," International Conference on Information Fusion, pp. 552–557, 2005.

[10] D. Bertsekas, "Dynamic programming and optimal control," Athena Scientific, Second Edition, Vol. 1, 2001.

[11] V. Krishnamurthy, "Algorithms for optimal scheduling of hidden Markov model sensors," IEEE Transactions on Signal Processing, Vol. 50, No. 6, pp. 1382–1397, 2002.

[12] W. S. Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes," Operations Research, Vol. 39, No. 1, pp. 162–175, 1991.