Temporal Difference Learning Waveform Selection

Bin Wang
Northeastern University, Shenyang, China
wangbin_neu@yahoo.com.cn

Jinkuan Wang, Xin Song and Yinghua Han
Northeastern University at Qinhuangdao, Qinhuangdao, China
sxin78916@mail.neuq.edu.cn

***Abstract*—How to optimally decide or select the radar waveform for next transmission based on the observation of past radar returns is one of the important problems in cognitive radar. In this paper, stochastic dynamic programming model is proposed. Then temporal difference learning method is used to realize the adaptivity of waveform selection. The simulation results show that the uncertainty of state estimation using temporal difference learning is less than that using fixed waveform. Temporal difference learning method approaches the optimal waveform selection scheme but has lower computational cost. Finally, the whole paper is summarized.**

***Index Terms*—waveform selection, stochastic dynamic programming, temporal difference learning, cognitive radar**

I. INTRODUCTION

Radar is the name of an electronic system used for the detection and location of objects. All early radars use radio waves, but some modern radars today are based on optimal waves and the use of lasers. With the development of modern technology, present-day systems are very sophisticated and advanced. We should consider focusing on optimizing the design of the transmitter, not only the receiver. That means there should be a feedback loop from the receiver to the transmitter. This is the core idea of cognitive radar.

Cognitive radar is a new framework of radar system proposed by Simon Haykin in 2006. It can percept external environment real time, select optimal waveform and make transmitted waveform and target environment and information demand of radar working achieve optimum matching, and then multiple performance of searching, tracking, guidance and identification of friend or foe of multi-target can be realized. It builds on three basic gradients: Intelligent signal processing; Feedback from the receiver to the transmitter; Preservation of the information content of radar returns[1]. Now more and more people are doing research in this field.

The obvious difference between cognitive radar and traditional radar is that cognitive radar can select appropriate waveforms according to different radar environment. So how to realize the adaptivity of the transmitter is an important problem in cognitive radar. The design of adaptive transmitter involves adaptive model and adaptive algorithm. Goodman have proposed and simulated a closed-loop active sensor by updating the probabilities on an ensemble of target hypotheses while adapting customized waveforms in response to prior measurement and compared the performance of two different waveform design techniques[2]. In [3], the author focuses on a cognitive tracking radar, the implementation of which comprises two distinct functional blocks, one in the receiver and the other in transmitter with a feed back link from the receiver to the transmitter. In [4], Arasaratnam have successfully solved the best approximation to the Bayesian filter in the sense of completely preserving second-order information, which is called cubature Kalman filters. In [5], Informax principle aimed at maximizing the mutural information is used for designing the transimitted signal waveform. In [6], an extention to the PDA tracking algorithm to include adaptive waveform selection was developed. In [7], 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. 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 [8], [9]. In [10], an adaptive waveform selective probabilistic data association algorithm for tracking a single target in clutter is presented. In [11], radar waveform selection algorithms for tracking accelerating targets are considered. In [12], genetic algorithm is used to perform waveform selection utilizing the autocorrelation and ambiguity functions in the fitness evaluation. In [13], 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

© 2010 ACADEMY PUBLISHER
doi:10.4304/jcp.5.9.1394-1401

[14]. In [15], the author uses ADP method to solve the problem of adaptive waveform selection.

In this paper, under the assumption of range-Doppler resolution cell, stochastic dynamic programming model for adaptive transmitter is proposed. We use temporal difference learning method to realize the adaptivity of waveform selection. The simulation results show the validity of our proposed algorithm.

Ⅱ. RANGE-DOPPLER RESOLUTION CELL

The design of adaptive transmitter in cognitive radar involves adaptive model and adaptive algorithm. We first consider setting up adaptive model. Generally speaking, for a target, the most important parameters that a radar measures are range, Doppler frequency, and two orthogonal space angles. If we envision a radar resolution cell that contains a certain four-dimensional hypervolume, we may assume different targets fall in different resolution cells. That means if a target measured falls in a resolution cell, then another target fall in another resolution cell and does not interfere with measurements on the first. So 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. For a single radar pulse, we may give a general sort of definition by considering the resolution cell to be bounded in range by the compressed pulse's duration, in Doppler by the reciprocal of the transmitted pulse's duration, and in the two angles by the antenna pattern's two orthogonal-plane beamwidths.

For example, if a radar seeks to make measurements on targets resolved in Doppler frequency at the same time, it can provide a bank of matched filters operating in parallel. Each target will excite the filter which is matched to its Doppler frequency, and its response can be used for measurements. Targets resolved in the range coordinate can be seperated with range gates followed by measurements. Thus a radar can perform simultaneous measurements on targets unresolved in angle, provided the targets are resolved in range, or Doppler frequency, or both. However, it is difficult to simultaneously measure targets in angle coordinates. Such measurements require either a bank of main beams or the time-sharing of one main beam among the various targets.

Through the preceding discussion, we can conclude that angle resolution can be considered independently from range and Doppler resolution in most circumstances. When considering this, the resolution properties of the radar in angle are independent of the resolution properties in range and Doppler frequency[16].

We define range-Doppler resolution cell for the waveform selection model.

ΔR denotes range resolution which 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 $R_{\max}$. The distance between $R_{\min}$ and $R_{\max}$ is divided into N range bins, each of width ΔR.

$$N = \frac{R_{\max} - R_{\min}}{\Delta R} \qquad (1)$$

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

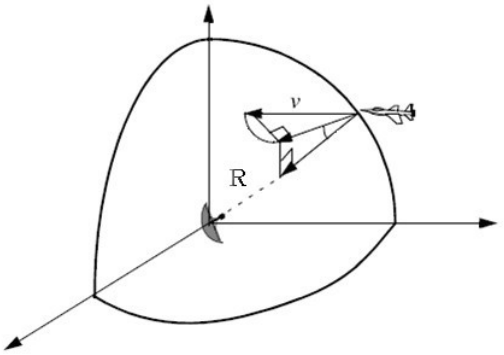

Figure 1. Radar and target.

Figure 1 is a picture of radar and target. In heavy clutter environments, it is an acute problem that we can not obtain good Doppler and good range resolution in a waveform tailoring simultaneously. So we need to consider the problem of adaptive waveform selection and make a trade-off decision between them. The basic scheme for adaptive waveform selection 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 $\tau = 1,...,N$ and those in Doppler indexed by $\upsilon = 1,...,M$. There may be 0 target, 1 target or NM targets. So the number of possible scenes or hypotheses about the radar scene is 2^{NM}. Let the space of hypotheses be denoted by χ. The state of our model is $X_t = x$ where $x \in \chi$. Let Y_{t+1} 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.

© 2010 ACADEMY PUBLISHER

We assume that the evolution of the state is governed by a Markov process and define $a_{x'x}$ is state transition probability where

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

$\mathbf{A} = (a_{x'x})_{x',x\in\chi}$ is the state transition matrix.

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

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

$\mathbf{B}(u_t) = (b_{x'x}(u_t))_{x',x\in\chi}$ is the measurement probability matrix. That means if state of our model is $X_t = x$ and we use the waveform u_t, the probability of measurement $Y_{t+1} = x'$ is $b_{x'x}(u_t)$.

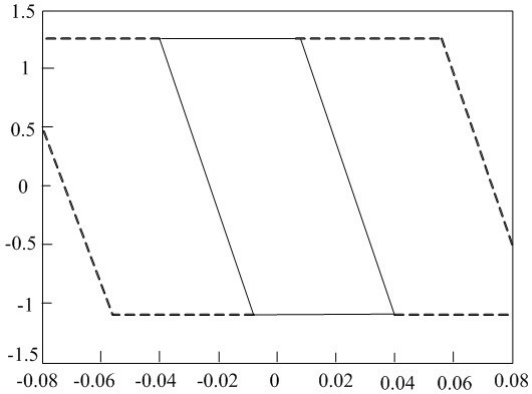

Figure 2. Resolution cell and corresponding parallelogram.

To handle the problem that when a combined waveform is used, we define as a practical resolution cell the parallelogram that contains the resolution cell primitive. Figure 2 is resolution cell and corresponding parallelogram.

Let us consider calculating these probabilities. The matched filter is adpoted in the receiver. 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

$$\begin{aligned} x(t) &= h(t) * r(t) \\ &= \int s^*(\lambda - t) r(\lambda) d\lambda \end{aligned} \tag{4}$$

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 ν_0 has an impulse response

$$h(t) = s^*(-t) e^{j2\pi\nu_0 t} \tag{5}$$

The output is given by

$$x(t) = \int s^*(\lambda - t) e^{-j2\pi\nu_0(\lambda - t)} r(\lambda) d\lambda \tag{6}$$

where ν_0 is an expected frequency shift.

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

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

$$\left|x(t)\right|^2 = \left|\int_0^t r(\lambda) s^*(\lambda - t) d\lambda\right|^2 \tag{7}$$

Following we consider two situations: there is no terget and target is present.

When there is no target

$$x(\tau_0) = \int_0^{\tau_0} n(\lambda) s^*(\lambda - \tau_0) d\lambda \tag{8}$$

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

$$\sigma_0^2 = E\left\{x(\tau_0) x^*(\tau_0)\right\} = 2N_0\xi \tag{9}$$

where ξ is the energy of the transmitted pulse.

When target is present

$$x(\tau_0) = \int_0^{\tau_0} \left[A s(\lambda - \tau) e^{j2\pi\nu_d\lambda} + n(\lambda) \right] s^*(\lambda - \tau_0) d\lambda \tag{10}$$

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

$$\sigma_1^2 = E\left\{x(\tau_0) x^*(\tau_0)\right\} = \sigma_0^2 \left(1 + \frac{2\sigma_A^2 \xi^2}{\sigma_0^2} A(\tau_0 - \tau, \nu_0 - \nu)\right) \tag{11}$$

where $A(\tau, \nu)$ is ambiguity function, given by

$$A(\tau, \nu) = \frac{1}{\left(\int |s(\lambda)|^2 d\lambda\right)^2} \left| \int s(\lambda) s^*(\lambda - \tau) e^{j2\pi\nu\lambda} d\lambda \right|^2 \tag{12}$$

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

$$y = x^2 \sim \frac{1}{2\sigma_i^2} e^{-\frac{y}{2\sigma_i^2}} \tag{13}$$

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

$$P_d = \frac{1}{|A|} \int_{(\tau_a, \upsilon_a \in A)} e^{-\frac{D}{2\sigma_0^2 \left(1 + \frac{2\sigma_A^2 \xi^2}{\sigma_0^2} A(\tau_0 - \tau, \nu_0 - \nu)\right)}} d\tau_a d\upsilon_a \tag{14}$$

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

III. Temporal Difference Learning Waveform Selection

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

$$V_t(X_t) = E[\sum_{t=0}^{T} \gamma^t R_t(X_t, u_t)] \tag{15}$$

© 2010 ACADEMY PUBLISHER

where $R_t(X_t,u_t)$ 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

$$V^*(X_t)=\max_{\pi} E[\sum_{t=0}^{T}\gamma^t R_t(X_t,u_t)] \tag{16}$$

However, knowledge of the actual state is not available. Using the method of [17], we can obtain that the optimal control policy π^* that is the solution of (16) is also the solution of

$$V^*(\mathbf{p}_0)=\max_{\pi} E[\sum_{t=0}^{T}\gamma^t R_t(\mathbf{p}_t,u_t)] \tag{17}$$

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. $\mathbf{p}$ is a sufficient statistic for the true state X_t. So we need to solve the following problem

$$\max_{\pi} E[\sum_{t=0}^{T}\gamma^t R_t(\mathbf{p}_t,u_t)] \tag{18}$$

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

$$\mathbf{p}_{t+1}=\frac{\mathbf{BAp}_t}{\mathbf{1'LAp}_t} \tag{19}$$

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

We write the expected profits using policy π from t onward

$$G_t(\mathbf{p}_t)=E\{\sum_{t'=t}^{T-1}R_{t'}(\mathbf{p}_{t'},u_{t'})+C_T(\mathbf{p}_T)\,|\,\mathbf{p}_t\} \tag{20}$$

where $u_{t'}=U^{\pi}(\mathbf{p}_{t'})$ under policy π.

$G_t(\mathbf{p}_t)$ is the expected total contribution if we are in state $\mathbf{p}_t$ in time t, and follow policy π from time t onward. However, it seems much more natural to calculate V_t recursively using

$$V_t(\mathbf{p}_t)=R_t(\mathbf{p}_t,u_t)+E\{V_{t+1}(\mathbf{p}_{t+1})\,|\,\mathbf{p}_t\} \tag{21}$$

Now we will establish the relationship between the original optimization problem and the optimality equations. Clearly, $G_T(\mathbf{p}_T)=V_T(\mathbf{p}_T)=R_T(\mathbf{p}_T)$. Next, assume that it holds for $t+1,t+2,...,T$. We want to show that it is true for t. This means that we can write

$$\begin{aligned}V_t(\mathbf{p}_t)&=R_t(\mathbf{p}_t,u_t)+E\{V_{t+1}(\mathbf{p}_{t+1})\,|\,\mathbf{p}_t\}\\&=R_t(\mathbf{p}_t,u_t)+E[G_{t+1}(\mathbf{p}_{t+1})]\\&=R_t(\mathbf{p}_t,u_t)+E\{E[\sum_{t'=t+1}^{T-1}R_{t'}(\mathbf{p}_{t'},u_{t'})+R_t(\mathbf{p}_T)\,|\,\mathbf{p}_{t+1}]\,|\,\mathbf{p}_t\}\end{aligned} \tag{22}$$

Because the random variables are discrete and finite, we can obtain that

$$\begin{aligned}&E\{E[G_{t'}\,|\,\mathbf{p}_{t+1}]\,|\,\mathbf{p}_t\}\\&=\sum_{\mathbf{p}_{t+1}\in\mathbf{p}}\sum_{g\in G}gP(G_{t'}=g\,|\,\mathbf{p}_{t+1},\mathbf{p}_t)P(\mathbf{p}_{t+1}=\mathbf{p}_{t+1}\,|\,\mathbf{p}_t)\\&=\sum_{g\in G}\sum_{\mathbf{p}_{t+1}\in\mathbf{p}}gP(G_{t'}=g\,|\,\mathbf{p}_{t+1},\mathbf{p}_t)P(\mathbf{p}_{t+1}=\mathbf{p}_{t+1}\,|\,\mathbf{p}_t)\\&=\sum_{g\in G}g\sum_{\mathbf{p}_{t+1}\in\mathbf{p}}P(G_{t'}=g,\mathbf{p}_{t+1}\,|\,\mathbf{p}_t)\\&=\sum_{g\in G}gP(G_{t'}=g\,|\,\mathbf{p}_t)\\&=E[G_{t'}\,|\,\mathbf{p}_t]\end{aligned} \tag{23}$$

Hence

$$\begin{aligned}&V_t(\mathbf{p}_t)\\&=R_t(\mathbf{p}_t,u_t)+E\{E[\sum_{t'=t+1}^{T-1}R_{t'}(\mathbf{p}_{t'},u_{t'})+R_t(\mathbf{p}_T)\,|\,\mathbf{p}_{t+1}]\,|\,\mathbf{p}_t\}\\&=R_t(\mathbf{p}_t,u_t)+E[\sum_{t'=t+1}^{T-1}R_{t'}(\mathbf{p}_{t'},u_{t'})+R_t(\mathbf{p}_T)\,|\,\mathbf{p}_t]\\&=E[R_t(\mathbf{p}_t,u_t)+\sum_{t'=t+1}^{T-1}R_{t'}(\mathbf{p}_{t'},u_{t'})+R_t(\mathbf{p}_T)\,|\,\mathbf{p}_t]\\&=E[\sum_{t'=t}^{T-1}R_{t'}(\mathbf{p}_{t'},u_{t'})+R_t(\mathbf{p}_T)\,|\,\mathbf{p}_t]\\&=G_t(\mathbf{p}_t)\end{aligned} \tag{24}$$

Using equation (21), we have a backward recursion for calculating $V_t^{\pi}(\mathbf{p}_t)$ for a given policy at π. Now that we can find the expected reward for a given π, we would like to find the best π. That is, we want to find

$$G_t^*(\mathbf{p}_t)=\max_{\pi\in\prod}G_t^{\pi}(\mathbf{p}_t) \tag{25}$$

If the set $\prod$ is infinite, we replace the "max" with "sup". We solve this problem by solving the optimality equations. These are

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

First, we show by induction that $V_t(\mathbf{p}_t)\ge G_t^*(\mathbf{p}_t)$ for all $\mathbf{p}_t\in\mathbf{p}$ and $t=0,1,...,T-1$. Then, we show that the reverse inequality is true, which gives us the result.

We resort again to our proof by induction. Since $V_T(\mathbf{p}_T)=R_t(\mathbf{p}_T)=G_T^{\pi}(\mathbf{p}_T)$ for all $\mathbf{p}_T$ and all $\pi\in\prod$, we get that $V_T(\mathbf{p}_T)=G_T^*(\mathbf{p}_T)$. Assume that $V_{t'}(\mathbf{p}_{t'})\ge G_{t'}^*(\mathbf{p}_{t'})$ for $t'=t+1,t+2,...,T$, and let π be an arbitrary policy. For $t'=t$, the optimality equation tells us

$$V_t(\mathbf{p}_t)=\max_{u\in U}(R_t(\mathbf{p}_t,u)+\sum_{\mathbf{p}'\in\mathbf{p}}P(\mathbf{p}'\,|\,\mathbf{p}_t,u)V_{t+1}(\mathbf{p}')) \tag{27}$$

With the induction hypothesis, $G_{t+1}^*(\mathbf{p})\le V_{t+1}(\mathbf{p})$, so we get

$$V_t(\mathbf{p}_t)\ge\max_{u\in U}(R_t(\mathbf{p}_t,u)+\sum_{\mathbf{p}'\in\mathbf{p}}P(\mathbf{p}'\,|\,\mathbf{p}_t,u)G_{t+1}^*(\mathbf{p}')) \tag{28}$$

© 2010 ACADEMY PUBLISHER

We have that $G_{t+1}^{*}(\mathbf{p}) \geq G_{t+1}^{\pi}(\mathbf{p})$ for an arbitrary π. Also let $U^{\pi}(\mathbf{p}_t)$ be the decision that would be chosen by policy π when in state $\mathbf{p}_t$. Then

$$\begin{aligned} V_t(\mathbf{p}_t) &\geq \max_{u\in U}(R_t(\mathbf{p}_t,u)+\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,u)G_{t+1}^{\pi}(\mathbf{p}')) \\ &\geq R_t(\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))+\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))G_{t+1}^{\pi}(\mathbf{p}') \\ &= G_t^{\pi}(\mathbf{p}_t) \end{aligned} \quad (29)$$

This means $V_t(\mathbf{p}_t) \geq G_t^{\pi}(\mathbf{p}_t)$ for all $\pi \in \prod$.

Next we are going to prove the inequality from the other side. Specially, we want to show that for any $\varepsilon > 0$ there exists a policy π that satisfies

$$G_t^{\pi} + (T-t)\varepsilon \geq V_t(\mathbf{p}_t) \quad (30)$$

To do this, we start with the definition

$$V_t(\mathbf{p}_t) = \max_{u\in U}(R_t(\mathbf{p}_t,u)+\sum_{\mathbf{p}'\in\mathbf{p}}P(\mathbf{p}'\mid\mathbf{p}_t,u)V_{t+1}(\mathbf{p}')) \quad (31)$$

We may let $u_t(\mathbf{p}_t)$ be the decision rule that solves (31). This rule corresponds to the policy π. In general, the set U may be infinite, whereupon we have to replace the "max" with a "sup" and handle the case where an optimal decision may not exist. For this case, we know that we can design a decision rule $u_t(\mathbf{p}_t)$ that returns a decision u that satisfies

$$V_t(\mathbf{p}_t) \leq R_t(\mathbf{p}_t,u)+\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,u)V_{t+1}(\mathbf{p}')+\varepsilon \quad (32)$$

We can prove (32) by induction. We first note that (30) is true for $t=T$ since $G_T^{\pi}(\mathbf{p}_t) = V_T(\mathbf{p}_T)$. Now assume that it is true for $t' = t+1, t+2, ..., T$. We already know that

$$G_t^{V}(\mathbf{p}_t) = R_t(\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))+\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))G_{t+1}^{\pi}(\mathbf{p}') \quad (33)$$

We can use our induction hypothesis which says $G_{t+1}^{\pi}(\mathbf{p}') \geq V_{t+1}(\mathbf{p}') - (T-(t+1))\varepsilon$ to get

$$\begin{aligned} G_t^{\pi}(\mathbf{p}_t) &\geq R_t(\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))+\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))[V_{t+1}(\mathbf{p}')-(T-(t+1))\varepsilon] \\ &= R_t(\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))+\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))V_{t+1}(\mathbf{p}')-\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))[(T-t-1)\varepsilon] \\ &= \{R_t(\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))+\sum_{\mathbf{p}'\in\mathbf{P}}P(\mathbf{p}'\mid\mathbf{p}_t,U^{\pi}(\mathbf{p}_t))V_{t+1}(\mathbf{p}')+\varepsilon\}-(T-t)\varepsilon \end{aligned} \quad (34)$$

Now, using equation (32), we replace the term in brackets with the smaller $V_t(\mathbf{p}_t)$

$$G_t^{\pi}(\mathbf{p}_t) \geq V_t(\mathbf{p}_t) - (T-t)\varepsilon \quad (35)$$

which proves the induction hypothesis. We has shown that

$$G_t^{*}(\mathbf{P}_t) + (T-t)\varepsilon \geq G_t^{\pi}(\mathbf{P}_t) + (T-t)\varepsilon \geq V_t(\mathbf{P}_t) \geq G_t^{*}(\mathbf{P}_t) \quad (36)$$

This proves the result. So we can use $V_t(\mathbf{p}_t)$ to solve our problem[18].

This is stochastic dynamic model of adaptive waveform selection. Then we can use dynamic programming algorithms to solve the problem. Backward dynamic programming algorithm is a basic dynamic programming method. It can be viewed as an optimal adaptive algorithm for waveform selection. When state space and action space are large, it is hard to use this method. Than means we hardly find optimal solution for waveform selection. So approximate solutions are necessary.

Generally speaking, reward function can be different forms according to different problems. It represents the value that we stand in certain place and take some certain action. In the problem of adaptive waveform selection, two forms of reward function are usually used. They are linear reward function and entropy reward function.

Linear reward function is usually used in the circumstance that $R(\mathbf{p},u)$ is required to be a piecewise linear function. The form of this function is simple and easy to calculate. However, it can not reflect the whole value sometimes. The form of linear reward function is

$$R_1(\mathbf{p},u) = \mathbf{p}'\mathbf{p} - 1 \quad (37)$$

Entropy reward function is usually used in the circumstance that $R(\mathbf{p},u)$ is not required to be a piecewise linear function. It comes from information theory. It can reflect the whole value accurately. But it is more complex than linear reward function. The form of entropy reward function is

$$R_2(\mathbf{p},u) = \sum_{x\in\chi} p_x(k)\log(p_x(k)) \quad (38)$$

We can choose different form of reward function according to different problems.

The foundation of approximate dynamic programming is based on an algorithmic strategy that steps forward through time. 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

$$V_t(\mathbf{p}_t) = \max_{u_t\in U}(C(\mathbf{p}_t,u_t)+\gamma E\{V_{t+1}(\mathbf{p}_{t+1})\mid\mathbf{p}_t\}) \quad (39)$$

Assume v is an unbiased sample estimate of the value of being in state $\mathbf{p}_t$ and the policy is π. The definition of v is

$$v_t^n = C_t(\mathbf{p}_t^n,u_t^{\pi}) + C_{t+1}(\mathbf{p}_{t+1}^n,u_{t+1}^{\pi}) + ... + C_T(\mathbf{p}_T^n,u_T^{\pi}) \quad (40)$$

Standard stochastic gradient algorithm is used to estimate the value of being in state X_t

$$V_t^n(\mathbf{p}_t) = V_t^{n-1}(\mathbf{p}_t) - \alpha_n[V_t^{n-1}(\mathbf{p}_t) - v_t^n] \quad (41)$$

The temporal differences is

$$D_\tau = C_\tau(\mathbf{p}_\tau,u_\tau) + V_{\tau+1}^{n-1}(\mathbf{p}_{\tau+1}) - V_\tau^{n-1}(\mathbf{p}_\tau) \quad (42)$$

So

© 2010 ACADEMY PUBLISHER

$$v_t^n = V_t^{n-1}(\mathbf{p}_t) + \sum_{\tau=t}^{T} D_\tau \quad (43)$$

Substituting (43) into (41), we can obtain

$$V_t^n(\mathbf{p}_t) = V_t^{n-1}(\mathbf{p}_t) - \alpha_{n-1} \sum_{\tau=t}^{T} D_\tau \quad (44)$$

The temporal differences are the errors in our estimates of the value of being in state $\mathbf{p}_\tau$. These errors are stochastic gradients for the problem of minimizing estimation error. The discount form is

$$V_t^n(\mathbf{p}_t) = V_t^{n-1}(\mathbf{p}_t) - \alpha_{n-1} \sum_{\tau=t}^{T} (\gamma\lambda)^{\tau-t} D_\tau \quad (45)$$

Through this formular, we can use this method to update the value of V.

Our algorithm is described as follows:

1 Give an initial state $\mathbf{p}_0^1$ and value function approximations $V_t^0(\mathbf{p}_t)$ for all $\mathbf{p}_t$ and t, set $n = 1$.

2 Choose a sample path.

3 For $t = 0,1,2,...,T$, do optimization and simulation. Optimization is to compute a decision $u_t^n = U_t^\pi(\mathbf{p}_t)$ and find the next state using $\mathbf{p}_{t+1} = \dfrac{\mathbf{BAp}_t}{\mathbf{1'LAp}_t}$.

4 Update the value function approximation to obtain $V_t^n(\mathbf{p}_t)$ for all t.

5 If we have not met our stopping rule, increase n and go to the first step. Else, stop.

Ⅳ. 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 different waveforms and measurement probability versus SNR with different targets. Curve of uncertainty of state estimation demonstrates validity of our proposed algorithm. We also plot the figure of value space versus state and waveform.

We adopt linear frequency modulation (LFM) signal in the transmitter. The formular of LFM is

$$s(t) = rect(\frac{t}{T}) e^{j2\pi(f_c t + \frac{K}{2}t^2)} \quad (46)$$

where f_c is carrier frequency and $rect$ is rectangular signal

$$rect(\frac{t}{T}) = \begin{cases} 1, & \left|\frac{t}{T}\right| \le 1 \\ 0, & elsewise \end{cases} \quad (47)$$

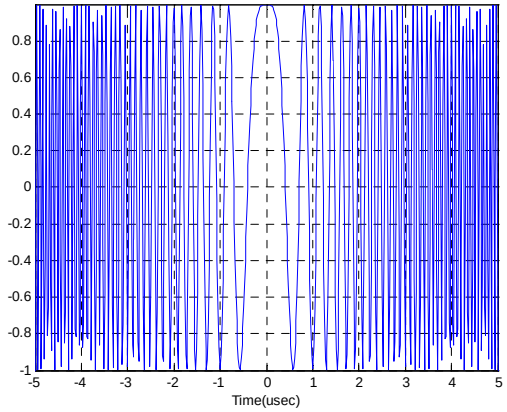

Figure 3. Real part of chirp signal.

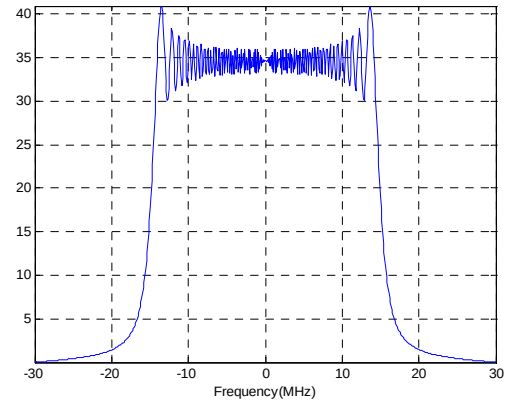

Figure 4. Magnitude spectrum of chirp signal.

Fiure 3 is real part of chirp signal and figure 4 is magnitude spectrum of chirp signal.

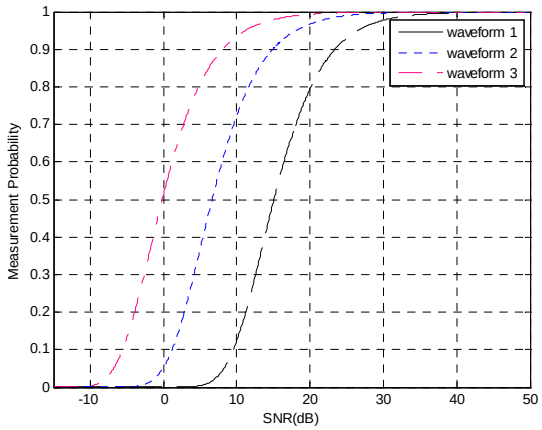

Figure 5. Measure probability versus SNR with three different waveforms.

Figure 5 is curve of measurement probability versus SNR with three different waveforms. From this figure we can see that measurement probability is becoming large with the increase of SNR. Under the same SNR, using different waveform corresponds to different measurement probability. So measurement can be improved through appropriately scheduling waveform in cognitive radar.

© 2010 ACADEMY PUBLISHER

Generally speaking, the wider the pulse duration is, the larger the measurement probability is. However, wide pulse duration means large energy of the transmitted pulse. We should make a balance between pulse duration and energy of the transmitted pulse and appropriately schedule waveform in order to obtain large measurement probability.

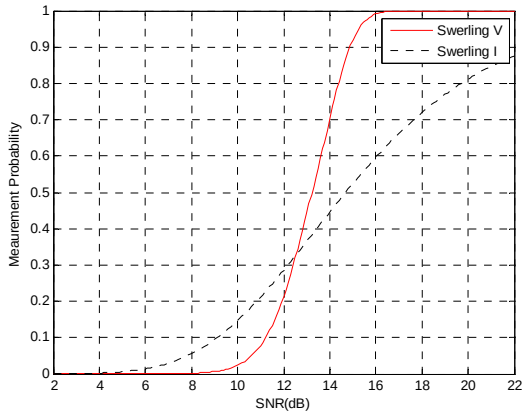

Figure 6. Measurement probability versus SNR with different targets.

Figure 6 is curve of measurement probability versus SNR with different targets. From this figure we can see that under the same SNR, measurement probability is different to different targets. So according to different targets we should select different waveforms. In actuality, path of target is so complex. We should change waveform according to different environment.

We will use the function

$$R(\mathbf{p}) = \mathbf{p}'\mathbf{p} - 1 \tag{48}$$

as the basis for our reward function. The formula $E(1-\mathbf{p}'\mathbf{p})$ can be considered as the uncertainty in the state estimation.

We consider a simple scenario. 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 Ⅰ. The discount factor $\gamma = 0.9$.

TABLE I.

MEASUREMENT PROBABILITIES FOR THE EXAMPLE SCENARIO

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

The matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} 0.96 & 0.02 & 0.01 & 0.01 \\ 0.01 & 0.93 & 0.03 & 0.02 \\ 0.02 & 0.03 & 0.95 & 0.02 \\ 0.01 & 0.02 & 0.01 & 0.95 \end{bmatrix} \tag{49}$$

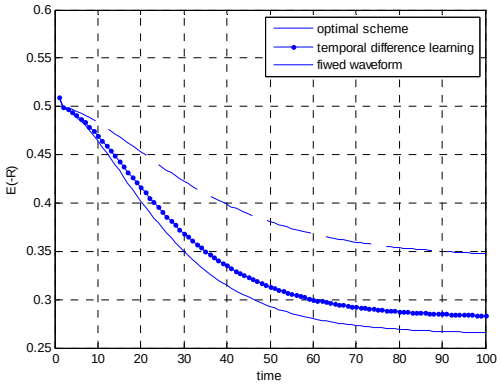

Figure 7. Curve of uncertainty of state estimation

Figure 7 is curve of uncertainty of state estimation. Form this figure, we can see that the uncertainty of state estimation is becoming lower with the increase of time. The uncertainty of state estimation using optimal scheduled waveform is lower than using fixed waveform. The uncertainty of state estimation using temporal difference learning algorithm approaches that of using optimal scheduled waveform. In the simulation, the optimal scheduled algorithm takes 55 seconds, while temporal difference learning algorithm takes 35 seconds. So our algorithm is efficient than the optimal algorithm.

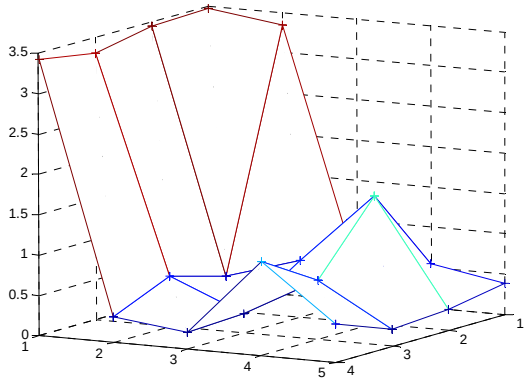

Figure 8. Value space versus state and waveform

Figure 8 is the figure of value space versus state and waveform. Value of different state-waveform pair can be obtained in this figure. We can see that the proposed algorithm has lower computational cost.

Ⅴ. CONCLUSIONS

Adaptive waveform selection is an important problem in cognitive radar and the problem of adaptive waveform selection can be viewed as a stochastic dynamic

© 2010 ACADEMY PUBLISHER

programming problem. In this paper, under the assumption of range-Doppler resolution cell, stochastic dynamic programming model for adaptive transmitter is set up. Then temporal difference learning is used to solve this problem. The simulation results show that the uncertainty of state estimation using temporal difference learning is less than that using fixed waveform. Temporal difference learning method approaches the optimal waveform selection scheme but has lower computational cost. Research on algorithms which approach the optimal waveform selection scheme and has lower computational cost is an important issue.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful comments that helped improve the quality of this paper. This work is supported by The National Natural Science Foundation of China, under Grant no. 60874108 and The Central University Fundamental Research Foundation, under Grant. no. N090604006.

REFERENCES

[1] S. Haykin, "Cognitive radar: a way of the future", *IEEE Signal Processing Magazine*, 2006, 23(1), pp. 30-40.
[2] N. A. Goodman, Phaneendra R. Venkata and Mark A. Neifeld, "Adaptive waveform design and sequential hypothesis testing for target recognition with active sensors", *IEEE Journal of Selected Topics in Signal Processing*, 2007, 1(1), pp. 105-113.
[3] S. Haykin, "Cognition is the key to the next generation of radar systems", *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, 2009. DSP/SPE 2009. IEEE 13th, pp. 463-467.
[4] I. Arasaratnam and S. Haykin, "Cubature Kalman filters", *IEEE Tran. Automatic Control*, 2009, 54(6), 463-467.
[5] S. Haykin, Y. B. Xue and T. Davidson, "Optimal waveform design for cognitive radar", *Asilomar Conference*, 2008.
[6] 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, October 1997.
[7] C. Rago, P. Willett and Y. Bar-Shalom, "Detecting-tracking performance with combined Waveforms", *IEEE Transactions on Aerospace and Electronic Systems*, 1998, 34(2), pp. 612-624.
[8] 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, 2004, pp. 743-748.
[9] V. Krishnamurthy, "Algorithms for optimal scheduling of hidden Markov model sensors", *IEEE Trans. on Signal Processing*, 2002, 50(6), pp.1382-1397.
[10] D. J. Kershawand and R. J. Evans, "Waveform selective probabilistic data association", *IEEE Transactions on Aerospace and Electronic Systems*, 1997, 33(4), pp. 1180-1188.
[11] C. O. Savage and B. Moran, "Waveform Selection for Maneuvering Targets Within an IMM Framework", *IEEE Transactions on Aerospace and Electronic Systems*, 2007, 43(3), pp. 1205-1214.
[12] C. T. Capraro, I. Bradaric, G.T. Capraro and Tsu Kong Lue, " Using genetic algorithms for radar selection", *2008 IEEE Radar Conference*, Inc., Utica, NY, May 2008, pp. 1-6.
[13] B. F. La Scala, W. Moran and R. J. Evans, "Optimal adaptive waveform selection for target detection", *The International Conference on Radar*, Adelaide, SA, Australia, Sept. 2003, pp. 492-496.
[14] B. La Scala, M. Rezaeian and B. Moran, "Optimal adaptive waveform selection for target tracking", *International Conference on Information Fusion*, 2005, pp. 552-557.
[15] B. Wang, J. K. Wang and J. Li, "ADP-based optimal adaptive waveform selection in cognitive radar", *International Symposium on Intelligent Information Technology Applications Workshops*, Shanghai, China, Dec. 2008, pp. 788-790.
[16] P. Z. Peebles, *Radar Principles*, John Wiley & Sons, Inc, 1998.
[17] D. Bertsekas, *Dynamic Programming and Optimal Control*, volume1, Athena Scientific, 2nd edition, 2001.
[18] Warren B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, John Wiley & Sons, Inc, 2007.

Bin Wang was born in Hebei, China, in 1982. He received M.S. degree in communication and information system in Northeastern University in China in 2008. Since March 2008, he has been working for his PhD degree in Northeastern University. His research interests are in the area of cognitive radar and adaptive waveform selection.

Jinkuan Wang received his PhD degree from University of Electro-Communications, Japan, in 1993. He is currently a professor in the School of Information Science and Engineering in Northeastern University, China, since 1998. His main interests are in the area of intelligent control and adaptive array.

Xin Song was born in Jilin, China, in 1978. She received her PhD degree in Communication and Information System in Northeastern University in China in 2008. She is now a teacher in Northeastern University at Qinhuangdao, China. Her research interests are robust adaptive beamforming and wireless communication.

Yinghua Han was born in Jilin, China, in 1979. She received the M.S. and PhD degrees in information science and engineering college from Northeastern University, Shenyang, China, in 2005 and 2008, respectively. Since 2003, she is with Engineering Optimization and Smart Antenna Institute. Her research interests include array signal processing and mobile wireless communication systems.

© 2010 ACADEMY PUBLISHER