基于双延迟深度确定性策略梯度学习机制的标签多伯努利传感器管理策略

张新迪[①] 陈 辉*[①] 张虹芸[①] 连 峰[②] 张光华[②] 阴志鹏[③]

[①](兰州理工大学自动化与电气工程学院 兰州 730050)
[②](西安交通大学自动化科学与工程学院 西安 710049)
[③](河北新天科创新能源技术有限公司 张家口 075000)

摘 要：针对复杂不确定性环境下的多目标跟踪问题，提出了一种基于双延迟深度确定性策略梯度(Twin-Delayed Deep Deterministic Policy Gradient algorithm, TD3)学习机制的标签多伯努利滤波器(Labeled Multi-Bernoulli, LMB)传感器管理方法。首先，联合深度强化学习和有限集统计(Finite Set Statistics, FISST)理论，基于LMB多目标运动感知结果构建信念空间，将传感器管理任务转化为该空间内的连续动作决策问题，使智能体能够根据当前多目标信念状态选择移动传感器的下一步观测动作。其次，为精准量化传感器配置对感知效能的增量贡献，引入柯西-施瓦茨(Cauchy-Schwarz, CS)散度动态评估候选策略驱动下的多目标概率密度信息增益，并据此设计信息驱动型奖励函数。通过TD3网络的双评论家架构、目标策略平滑与延迟更新机制，智能体能够在连续动作空间内学习观测配置，实现对传感器管理的闭环决策。实验结果表明，该方法能够改善LMB滤波器的状态估计性能和传感器轨迹连续性。

关键词：多目标跟踪；传感器管理；深度强化学习；标签多伯努利滤波器

中图分类号：TP274 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-08

DOI: 10.11999/JEIT260045 **CSTR**: 32379.14.JEIT260045

1 引言

多目标跟踪(Multiple Target Tracking, MTT)旨在杂波、漏检及目标数目变化等不确定条件下估计目标数量与运动状态，是雷达监视和态势感知中的重要问题[1–3]。随着认知雷达和自适应资源管理技术的发展，传感器可实时调节波束指向、驻留时间及平台运动等参数，主动改善观测条件[4–6]。传感器管理通常可建模为部分可观测马尔可夫决策过程(Partially Observable Markov Decision Process, POMDP)[7–8]，但传统求解方法在高维状态空间和连续动作空间中面临较大的计算负担。因此，如何结合多目标后验信息实现连续、在线的传感器决策，仍是需要解决的问题。

现有多目标传感器管理方法主要包括任务驱动和信息驱动两类[9]。任务驱动方法直接以跟踪误差或任务性能为优化目标。例如，Jones等人利用广义最优子模式分配(Generalized Optimal Sub-Pattern Assignment, GOSPA)指标评价高斯伯努利滤波器的候选动作[10]，Gostar等人则以多伯努利滤波器的势估计误差选择传感器策略[11]。信息驱动方法通过比较预测密度与更新密度之间的信息差异评价候选动作。Ristic等人采用Rényi散度构造概率假设密度滤波器的传感器控制奖励[12]，Gostar等人将柯西-施瓦茨(Cauchy-Schwarz, CS)散度用于标签多伯努利滤波器的传感器控制[13]。这类方法能够利用多目标后验评价观测信息，但通常需要在预设的有限动作集合中逐一搜索。当动作变量具有连续性或维数增加时，离散搜索难以同时兼顾控制精度和在线计算效率。

近年来，深度强化学习(Deep Reinforcement Learning, DRL)被应用于传感器路径规划、平台引导、雷达功率分配及多目标雷达资源管理等任务[14–16]。这些研究表明，DRL能够利用数据学习复杂环境中的控制策略，但部分方法仍采用单目标状态、任务特定指标或多个单目标指标的组合构造奖励，难以充分表达杂波、漏检和量测来源不确定条件下的整体多目标后验。此外，早期确定性策略梯度方法容易受到价值估计偏差和训练波动影响。双延迟深度确定性策略梯度(Twin-Delayed Deep Deterministic Policy Gradient, TD3)通过双评论家取较小值、目标策略平滑和延迟策略更新缓解价值高估问题[17]，适合连续传感器动作的学习。

收稿日期：2024-09-27；改回日期：2026-06-30；网络出版：2026-07-13
*通信作者： 陈辉 huich78@hotmail.com
基金项目：国家自然科学基金(62163023, 61873116)，甘肃省科技计划项目(25JRRA058, 25ZYJA040)，2024年度甘肃省重点人才项目，2023年度甘肃省军民融合发展专项资金

Foundation Items: The National Natural Science Foundation of China (62163023, 61873116, 62366031, 62363023), the Gansu Provincial Science and Technology Plan Project of China (25ZYJA040, 25JRRA058), the 2024 Gansu Provincial Key Talent Project of China and the 2023 Gansu Provincial Special Fund for Military-Civilian Integration Development of China

随机有限集(Random Finite Set, RFS)理论能够统一描述目标数量和状态的不确定性，标签多伯努利(Labeled Multi-Bernoulli, LMB)滤波器则能够以带标签的多目标概率密度表示目标存在概率及运动状态。基于此，本文提出一种基于TD3的LMB传感器管理方法。首先，根据LMB预测与更新结果构造多目标信念状态，将移动传感器管理转化为连续航向角决策问题；其次，以预测LMB与伪更新LMB之间的CS散度构造信息增益奖励，使动作评价直接关联整体多目标后验；最后，利用TD3学习连续传感器控制策略，以提高复杂多目标环境中的观测配置能力和跟踪稳定性。

2 问题描述

2.1 标签多伯努利随机有限集

为了追踪目标的轨迹, 将状态$x \in X$加上标签$\ell \in \mathbb{L}$后记为$\mathrm{x} \in \mathbb{X} \times \mathbb{L}$，因此参数集为$\pi = \{r^{(\ell)}, p^{(\ell)}\}_{\ell \in \mathbb{L}}$的标签RFS$\boldsymbol{X}$的概率密度可表示为

$$\pi(\boldsymbol{X}) = \Delta(\boldsymbol{X})\, w(\mathcal{L}(\boldsymbol{X}))\, p^{\boldsymbol{X}} \tag{1}$$

其中$\Delta(\boldsymbol{X}) \triangleq \delta_{|\boldsymbol{X}|}(|\mathcal{L}(\boldsymbol{X})|)$为不同的标签指示，$|\cdot|$表示集合的势，$\mathcal{L}$表示由$\mathbb{X} \times \mathbb{L} \to \mathbb{L}$的映射，即$\mathcal{L}((x,\ell)) = \ell$，且

$$w(L) = \prod_{i \in \mathbb{L} \setminus L} \left(1 - r^{(i)}\right) \prod_{\ell \in L} \frac{1_{\mathbb{L}}(\ell)\, r^{(\ell)}}{1 - r^{(\ell)}} \tag{2}$$

$$p(x,\ell) = p^{(\ell)}(x) \tag{3}$$

2.2 RFS下的多目标跟踪建模

假设k时刻在目标的状态空间$\mathbb{X} \times \mathbb{L}$上存在$N(k)$个目标，其状态可分别表示为$\mathrm{x}_{k,1}, \mathrm{x}_{k,2}, \cdots \mathrm{x}_{k,N(k)}$，同时在观测空间中接收到$M(k)$个量测$z_{k,1}, z_{k,2}, \cdots, z_{k,M(k)}$。故多目标状态与多目标量测的RFS可表示为

$$\boldsymbol{X}_k = \left\{\mathrm{x}_{k,1}, \mathrm{x}_{k,2}, \cdots, \mathrm{x}_{k,N(k)}\right\} \tag{4}$$

$$Z_k = \left\{z_{k,1}, z_{k,2}, \cdots, z_{k,M(k)}\right\} \tag{5}$$

假设在$k-1$时刻，每个目标$\mathrm{x}_{k-1} \in \boldsymbol{X}_{k-1}$都有$p_\nu(\mathrm{x}_{k-1})$的概率继续存活，因此在$k$时刻的多目标RFS可表示为

$$\boldsymbol{X}_k = \left[\bigcup_{\mathrm{x}_{k-1} \in \boldsymbol{X}_{k-1}} S_{k|k-1}(\mathrm{x}_{k-1})\right] \cup \Gamma_k \tag{6}$$

其中$S_{k|k-1}(\cdot)$表示从$k-1$时刻到k时刻存活目标的RFS，Γ_k表示新生目标RFS。同时，状态转移方程可表示为

$$\mathrm{x}_{k+1} = F\mathrm{x}_k + w_k \tag{7}$$

其中F表示状态转移矩阵，w_k表示过程噪声。

当传感器选择动作a_k后，以$P_D(\mathrm{x}_k)$的检测概率检测到目标x_k所对应的量测$z_k \in Z_k$可表示为

$$z_k = h(\mathrm{x}_k, a_k) + v_k \tag{8}$$

其中h为量测方程。考虑到杂波影响，多目标量测集Z_k可表示为

$$Z_k = \left[\bigcup_{\mathrm{x}_k \in \boldsymbol{X}_k} \Theta_k(\mathrm{x}_k)\right] \cup \mathcal{K}_k \tag{9}$$

其中$\Theta_k(\mathrm{x}_k)$表示目标x_k对应的量测RFS，$\mathcal{K}_k$表示k时刻杂波分布。

3 基于TD3的LMB滤波器传感器管理

3.1 基于深度强化学习的多目标贝叶斯滤波器传感器管理框架设计

本文构造的基于深度强化学习的多目标贝叶斯滤波器传感器管理框架如图1所示。

当$k > 1$时，传感器接收的量测受其策略序列$a_0{:}\, a_k$影响，通过历史量测信息$Z_{0:\,k,a_0:\,a_k} = (Z_{0,a_0}, \cdots, Z_{k,a_k})$计算$k$时刻多目标的后验密度$\pi_k(X_{0:\,k}|Z_{0:\,k,a_0:\,a_k})$：

$$\begin{aligned}\pi_k(X_{0:\,k}|Z_{0:\,k,a_0:\,a_k}) \propto\ & g_k(Z_{k,a_k}|X_k) \\ & f_{k|k-1}(X_k|X_{k-1}) \cdot \pi_{k-1}(X_{0:\,k-1}|Z_{0:\,k-1,a_0:\,a_{k-1}})\end{aligned} \tag{10}$$

其中$X_{0:\,k} = (X_0, \cdots, X_k)$，$g_k(\cdot|\cdot)$表示在$k$时刻多目标似然函数，$f_{k|k-1}(\cdot\,|\,\cdot)$为多目标状态转移函数。

假设多目标状态的演变遵循一阶马尔可夫过程，先验与后验多目标概率密度可表示为

$$\pi_{k+1|k}(X_{k+1}) = f_{k+1|k}(X_{k+1}|X_k)\pi_k(X_k|Z_{k,a_k})\delta X_k \tag{11}$$

$$\pi_k(X_k|Z_{k,a_k}) = \frac{g_k(Z_{k,a_k}|X_k)\pi_{k|k-1}(X_k)}{g_k(Z_{k,a_k}|X_k)\pi_{k|k-1}(X_k)\delta X_k} \tag{12}$$

为了在每一个时刻使传感器实时获得最优量测，首先从先验概率密度中提取出预测多目标状态

$$\hat{X}_{k+1|k} = sef(\pi_{k+1|k}(X_{k+1})) \tag{13}$$

其中$sef(\cdot)$表示状态提取操作。

再用每一个预测状态$\hat{x}_{k+1|k} \in \hat{X}_{k+1|k}$与伪动作$a_k^\circ$构造伪量测$z_k^\circ \in Z_k^\circ$

$$z_k^\circ = h\left(\hat{x}_{k+1|k}, a_k^\circ\right) \tag{14}$$

利用式(14)构造的伪量测，可得伪更新的多目标后验概率密度：

$$\pi_k(X_k|Z_k^\circ) = \frac{g_k(Z_k^\circ|X_k)\pi_{k|k-1}(X_k)}{g_k(Z_k^\circ|X_k)\pi_{k|k-1}(X_k)\delta X_k} \tag{15}$$

由式(14)、式(15)可知，伪量测及伪更新结果与传感器动作有关，因此为了优化滤波器性能，动作的选择指标应满足：

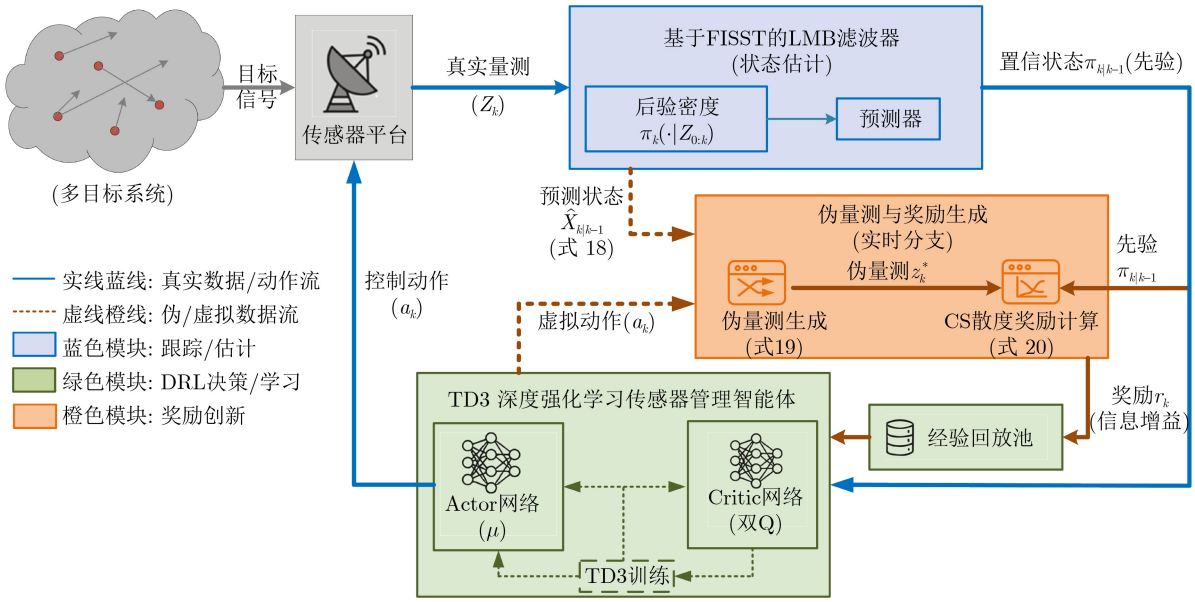

图 1 基于深度强化学习的多目标贝叶斯滤波器传感器管理框架

$$a_k = \underset{a_k \in \mathbb{A}_k}{\arg\max} \mathbb{E}\left[J\left(a_k\right)\right] \tag{16}$$

其中$J\left(a_k\right)$表示滤波器性能的量化指标。

为实现有效的传感器管理，需要构造能够反映当前感知质量的状态信息，并据此对下一时刻每个可能的传感器动作的价值进行量化。目前，已有多种量化方法被提出，如平均风险与信息增益等。综合考虑感知精度与信息价值，本文采用先验概率密度与后验概率密度之间的信息增益作为评估传感器动作价值的指标：

$$S\left(\overset{\circ}{a}_k\right) = div\left(\pi_k\left(X_k|\overset{\circ}{Z}_k\right), \pi_{k|k-1}\left(X_{k-1}\right)\right) \tag{17}$$

其中div (A,B)表示计算A,B之间的散度。基于此，滤波器性能的量化指标可作为深度强化学习中的奖励信号，用于评估各个传感器动作的价值，并构造经验回放机制，通过深度强化学习算法推导出在k时刻使滤波器性能最优的传感器策略$\hat{a}_k$。

3.2 基于TD3的LMB滤波器传感器管理问题建模

传感器管理可建模为部分可观测马尔可夫决策过程。由于真实多目标状态无法直接获得，本文利用LMB滤波器的后验结果构造信念状态，并以传感器航向角作为连续动作。状态转移和观测过程分别由多目标运动模型与量测模型描述，奖励函数则根据第3.3节定义的CS散度构造。

3.3 LMB滤波器的传感器管理方案求解

假设在k时刻，已存在目标与新生目标都是LMB，密度参数为$\left\{\left(r_{k,S}^{(\ell)}, p_{k,S}^{(\ell)}\right)\right\}_{\ell\in\mathbb{L}}$，$\left\{\left(r_{\mathrm{B}}^{(\ell)}, p_B^{(\ell)}\right)\right\}_{\ell\in\mathbb{B}}$：

$$\pi_{k,S}(\boldsymbol{X}) = \Delta(\boldsymbol{X}) w_S(\mathcal{L}(\boldsymbol{X}))[p_S]^{\boldsymbol{X}} \tag{18}$$

$$\pi_{k,B}(\boldsymbol{X}) = \Delta(\boldsymbol{X}) w_B(\mathcal{L}(\boldsymbol{X}))[p_B]^{\boldsymbol{X}} \tag{19}$$

其中，

$$w_S(L) = \prod_{i\in\mathbb{L}}\left(1 - r_{k,S}^{(i)}\right) \prod_{\ell\in L} \frac{1_{\mathbb{L}}(\ell) r_{k,S}^{(\ell)}}{1 - r_{k,S}^{(\ell)}} \tag{20}$$

$$w_B(L) = \prod_{i\in\mathbb{B}}\left(1 - r_B^{(i)}\right) \prod_{\ell\in L} \frac{1_{\mathbb{B}}(\ell) r_B^{(\ell)}}{1 - r_{\mathrm{B}}^{(\ell)}} \tag{21}$$

$$p_S(x,\ell) = p_S^{(\ell)}(x) \tag{22}$$

$$p_B(x,\ell) = p_B^{(\ell)}(x) \tag{23}$$

合并后可得一步预测的概率密度：

$$\pi_{k+1|k} = \left\{\left(r_{k+1|k,S}^{(\ell)}, p_{k+1|k,S}^{(\ell)}\right)\right\}_{\ell\in\mathbb{L}} \cup \left\{\left(r_{\mathrm{B}}^{(\ell)}, p_B^{(\ell)}\right)\right\}_{\ell\in\mathbb{B}} \tag{24}$$

$$r_{k+1|k,S}^{(\ell)} = \eta_S(\ell) r^{(\ell)} \tag{25}$$

$$p_{k+1|k,S}^{(\ell)}(x) = \left\langle p_v(x,\ell) f(x|\cdot,\ell), p_S(x,\ell)\right\rangle / \eta_S(\ell) \tag{26}$$

其中$\langle\cdot,\cdot\rangle$为内积计算符，$p_v\left(x,\ell\right)$为状态相关存活概率，$\eta_S\left(\ell\right)$表示轨迹$\ell$的存活概率，$f\left(x|\cdot,\ell\right)$表示轨迹$\ell$的转移密度，并将预测中的合并标签空间定义为$\mathbb{L}_{k+1|k} = \mathbb{L}_k \cup \mathbb{B}$，记为$\mathbb{L}_+$。

由式(24)～式(26)可提取出一步预测中的状态参数，利用TD3策略网络输出的伪动作生成理想量测集：

$$Z_{k+1}^{\circ} = \underset{\hat{x}_{k+1|k}\in\hat{X}_{k+1|k}}{\cup}\left[h\left(\hat{x}_{k+1|k}, \overset{\circ}{a}_k\right)\right] \tag{27}$$

利用伪量测Z_{k+1}°对预测的概率密度$\pi_{k+1|k}$更新：

$$\pi_{k+1}(\cdot|Z_{k+1}^{\circ})=\{(r_{k+1}^{(\ell)},p_{k+1}^{(\ell)})\}_{\ell\in\mathbb{L}_{+}} \tag{28}$$

$$r_{k+1}^{(\ell)}=\sum_{(I,\theta)\in\mathcal{F}(\mathbb{L}_{+})\times\Theta_{I}}w^{(I,\theta)}(Z_{k+1}^{\circ})1_{I}(\ell) \tag{29}$$

$$p_{k+1}^{(\ell)}(x)=\frac{1}{r_{k+1}^{(\ell)}}\sum_{(I,\theta)\in\mathcal{F}(\mathbb{L}_{+})\times\Theta_{I}}w^{(I,\theta)}(Z_{k+1}^{\circ})\times 1_{I}(\ell)\cdot p^{(\theta)}(x,\ell) \tag{30}$$

其中$\mathcal{F}(\cdot)$表示有限集中所有子集，I表示轨迹标签，Θ_I表示映射空间$\theta\colon I\to\{0,1,\cdots,|Z|\}$，$P_D(x,\ell)$表示$(x,\ell)$的检测概率，对于每一个映射$\theta$有：

$$w^{(I,\theta)}(Z_{k+1}^{\circ})\propto w_{k+1|k}(I)\left[\eta_{Z_{k+1}}^{(\theta)}\right]^{I} \tag{31}$$

$$p^{(\theta)}(x,\ell|Z_{k}^{\circ})=\frac{p_{k+1|k}(x,\ell)\psi_{Z}(x,\ell;\theta)}{\eta_{Z_{k}^{\circ}}^{(\theta)}(\ell)} \tag{32}$$

$$\psi_{Z}(x,\ell;\theta)=\delta_{0}(\theta(\ell))q_{D}(x,\ell)+(1-\delta_{0}(\theta(\ell)))\cdot\frac{p_{D}(x,\ell)g(z_{k,\theta(\ell)}|x,\ell)}{\kappa(z_{k,\theta(\ell)})} \tag{33}$$

其中$q_D(x,\ell)=1-P_D(x,\ell)$，$\kappa(\cdot)$为泊松分布的噪声的强度函数.

传感器动作通过改变传感器与目标之间的相对观测关系，影响伪量测质量及伪更新结果，进而影响信息增益。在短视策略中，最大化信息增益(如数据与目标之间的互信息)可作为替代任务特定奖励函数的通用性能指标。鉴于在SMC-LMB滤波中，预测密度与更新密度间一阶统计距的CS散度具有闭式解，本文采用通过散度设定奖励函数。

多目标状态的预测LMB密度可表示为$\left\{r_{k+1|k}^{(\ell)},p_{k+1|k}^{(\ell)}\colon\ell\in\mathbb{L}_{k+1|k}\right\}$，在LMB滤波器的粒子实现中，每一个$p_{k+1|k}^{(\ell)}$都可由一组粒子近似：

$$p_{k+1|k}^{(\ell)}=\sum_{j=1}^{J_{k+1|k}^{(\ell)}}w_{k+1|k,j}^{(\ell)}\delta\left(x-x_{k+1|k,j}^{(\ell)}\right) \tag{34}$$

多目标状态的后验LMB密度可表示为$\left\{r_{k+1}^{(\ell)},p_{k+1}^{(\ell)}\colon\ell\in\mathbb{L}_{k+1}\right\}$，每一个$p_{k+1}^{(\ell)}$与预测LMB密度的$p_{k+1|k}^{(\ell)}$粒子相同，但具有不同的权重：

$$p_{k+1}^{(\ell)}=\sum_{j=1}^{J_{k+1|k}^{(\ell)}}w_{k+1,j}^{(\ell)}\delta\left(x-x_{k+1|k,j}^{(\ell)}\right) \tag{35}$$

因此预测与后验的LMB密度的一阶统计矩可以表示为：

$$D_{k+1|k}(x)=\sum_{\ell\in\mathbb{L}_{k+1|k}}\sum_{j=1}^{J_{k+1|k}^{(\ell)}}r_{k+1|k}^{(\ell)}w_{k+1|k,j}^{(\ell)}\delta(x-x_{k+1|k,j}^{(\ell)}) \tag{36}$$

$$D_{k+1}(x)=\sum_{\ell\in\mathbb{L}_{k+1|k}}\sum_{j=1}^{J_{k+1|k}^{(\ell)}}r_{k+1}^{(\ell)}w_{k+1,j}^{(\ell)}\delta(x-x_{k+1|k,j}^{(\ell)}) \tag{37}$$

其中，

$$w_{k+1}^{(\ell)}=w_{k+1|k}^{(\ell)}\cdot g_{k}(z_{k}^{\circ}|x_{k+1|k,j}^{(\ell)}) \tag{38}$$

$$r_{k+1}^{(\ell)}=\frac{r_{k+1|k}^{(\ell)}p_{D}\sum\limits_{z_{k}^{\circ}\in Z_{k}^{\circ}}g_{k}(z_{k}^{\circ}|x_{k+1|k,j}^{(\ell)})}{(1-r_{k+1|k}^{(\ell)}p_{D})+r_{k+1|k}^{(\ell)}p_{D}\sum\limits_{z_{k}^{\circ}\in Z_{k}^{\circ}}g_{k}(z_{k}^{\circ}|x_{k+1|k,j}^{(\ell)})} \tag{39}$$

由此可计算出预测与更新的LMB密度一阶统计矩之间的CS散度为：

$$\begin{aligned}D_{CS}\left(a_{k}^{\circ},\pi_{k}(x)\right)&=-\ln\left(\frac{\langle D_{k+1|k},D_{k+1}\rangle}{\sqrt{\langle D_{k+1|k},D_{k+1}\rangle\cdot\langle D_{k+1},D_{k+1}\rangle}}\right)\\&=-\ln\left(\frac{\sum\limits_{\ell,j}\left(r_{k+1|k}^{(\ell)}w_{k+1|k,j}^{(\ell)}\cdot r_{k+1}^{(\ell)}w_{k+1,j}^{(\ell)}\right)}{\sqrt{\left[\sum\limits_{\ell,j}\left(r_{k+1|k}^{(\ell)}w_{k+1|k,j}^{(\ell)}\right)^{2}\right]\cdot\left[\sum\limits_{\ell,j}\left(r_{k+1}^{(\ell)}w_{k+1,j}^{(\ell)}\right)^{2}\right]}}\right)\end{aligned} \tag{40}$$

因此奖励函数设计为：

$$\begin{aligned}r_{k}=R\left(\pi_{k}(\mathrm{x}),a_{k}^{\circ}\right)&=D_{CS}\left(a_{k}^{\circ},\pi_{k}(\mathrm{x})\right)\\&-D_{CS}\left(a_{s}^{\circ},\pi_{k}(\mathrm{x})\right)\end{aligned} \tag{41}$$

其中a_s°表示传感器处于静止不动(静默状态)时的动作。由此构建深度强化学习的经验样本$\left\{\hat{X}_{k+1|k},a_{k}^{\circ},r_{k},\hat{X}_{k+1}\right\}$。

鉴于每一滤波步都需获得最优策略，本文将每一滤波时刻视为一个训练回合，在每回合的结束时重置网络与经验回放池。如图4所示，深度强化学

习模习模块由一个策略网络(参数为θ^{μ})、一个目标策略网络(参数为$\theta^{\mu'}$)，以及两个评论家网络(参数为θ^{Q} $(i=1,2)$)及其对应的目标网络(参数为$\theta^{Q'}$)构成。其中策略网络与目标策略网络均包含输入层，全连接层，输出层，激活层，扩尺函数，评论家网络与目标评论家网络结构为：输入层，全连接层，输出层。由当前状态、传感器动作、奖励及下一时刻状态构成的经验样本被持续填充至经验回放池。训练过程中智能体从中随机抽取样本，输入评论家网络，通过最小化均方误差来更新网络参数，具体损失函数定义如下：

$$L\left(\theta_i^{Q}\right)=\frac{1}{N}\sum_{j=1}^{N}\left(y-Q_i\left(\hat{X}_{k+1|k},\mathring{a}_k\right)\right)^2 \tag{42}$$

$$y=r_{k+1}+\gamma\min_{i=1,2}Q_i'\left(\hat{X}_{k+2|k+1},\mu'\left(\hat{X}_{k+2|k+1}\right)\right) \tag{43}$$

其中Q_i表示两个评论家网络，Q'_i为两个目标评论家网络，$\mu'\left(\hat{X}_{k+1}\right)$表示多目标状态为$\hat{X}_{k+1|k}$时的目标策略网络。评论家网络的参数更新公式为：

$$\theta_i^{Q}\leftarrow\theta_i^{Q}-\alpha\nabla_{\theta_i^{Q}}L\left(\theta_i^{Q}\right),i\in\{1,2\} \tag{44}$$

其中α为评论家网络学习率，$\nabla_{\theta_i^{Q}}$表示对参数θ_i^{Q}计算梯度。策略网络采用梯度上升的算法进行网络参数的更新，其目的是最大化价值：

$$J\left(\theta^{\mu}\right)\approx\frac{1}{N}\sum_{m=1}^{M}Q_i(\hat{X}_{k+1|k},\mu(\hat{X}_{k+1|k})) \tag{45}$$

其中M为样本数。策略网络的参数更新公式为:

$$\theta^{\mu}\leftarrow\theta^{\mu}+\beta\nabla_{\theta^{\mu}}J\left(\theta^{\mu}\right) \tag{46}$$

其中β为策略网络学习率。目标策略网络的与目标评论家网络的更新采用软更新策略：

$$\theta_i^{Q'}\leftarrow\tau\theta_i^{Q}+(1-\tau)\,\theta_i^{Q'} \tag{47}$$

$$\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\,\theta^{\mu'} \tag{48}$$

其中τ为软更新系数。

4　仿真实验

在正方形监控区域内，部署一台搭载搜索雷达的移动感知平台，同时跟踪10个目标。监控区域边长为$a=2\,000$ m。目标状态向量为$\mathrm{x}=[x_p,y_p,\dot{x}_p,\dot{y}_p,\omega]^{\mathrm{T}}$，按恒转速模型运动。假设传感器的检测概率随传感器与目标之间距离的增大而减小：

$$p_D(\boldsymbol{p})=\begin{cases}1, & \text{if } d_k\le R_0\\ \max\{0,1-(d_k-R_0)h\}, & \text{if } d_k>R_0\end{cases} \tag{49}$$

其中$R_0=100$ m，$d_{\mathrm{k}}=||\boldsymbol{p}-\mathrm{x}_{s,k}||$为时刻目标与传感器之间的距离，h为检测概率随距离变化的衰减系数：

$$||\boldsymbol{p}-\mathrm{x}_{s,k}||=\sqrt{(x_p-x_{s,k})^2+(y_p-y_{s,k})^2} \tag{50}$$

图2展示了传感器在100个时间步内的运动轨迹。实验结果表明，与缺乏明确导向性且信息获取效率低下的随机控制策略相比，本文提出的基于TD3的传感器策略在复杂动态环境下表现出更高的平滑性与显著的目标导向性；尽管基于信息论和策略梯度的离散控制方法能够引导传感器向目标区域聚集，但由于受限于局部最优及动作空间的离散性，其轨迹在目标分布变化时常出现不稳定的脉冲式跳跃。相比于上述离散控制方法，DDPG同样采用连续动作空间，能够直接输出连续航向角控制量，因此其传感器运动轨迹整体更加平滑，说明连续动作能够在一定程度上缓解离散动作切换带来的轨迹突变问题。然而，由于DDPG采用单评论家结构，在目标数量和空间分布快速变化时，其动作价值估计容易受到函数逼近误差和价值高估的影响，导致部分时刻的动作调整仍可能出现局部不稳定。相比之下，所提算法利用TD3的连续动作决策机制、双评论家结构和基于CS散度的奖励函数，使智能体能够根据目标的出生、消亡及运动状态实时、平滑地调整航向，主动识别并接近高信息量区域，从而在目标密度变化与运动复杂化背景下保持稳定性能。

10个目标的检测概率对比如图3所示。不同目标对应的检测概率存在差异，说明传感器与各目标之间的相对位置和运动状态会影响量测获取效果。随机控制在多数目标上的检测概率相对较低，PG方法和CS散度方法能够通过动作选择改善部分

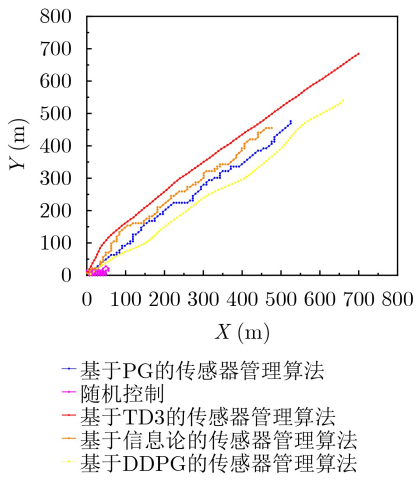

图 2 传感器运动轨迹

目标的观测条件。DDPG-LMB和TD3-LMB采用连续航向角控制，能够对传感器运动方向进行较细致的调整，其中TD3-LMB在多数目标上取得了最高或接近最高的检测概率。部分目标上DDPG-LMB与TD3-LMB的结果较为接近，表明连续动作建模本身有助于改善传感器与目标之间的相对观测几何关系。TD3-LMB在不同目标上的检测概率变化相对均衡，说明其在当前场景下能够兼顾多个目标的观测需求。从不同目标的结果可以看出，单一目标上的最高检测概率并不能完整代表多目标场景的整体性能，因此还需要结合信息增益和OSPA结果评价传感器策略。

500次蒙特卡洛实验下不同算法的CS散度结果如图4所示。各算法的CS散度均随时间发生变化，表明预测密度与伪更新密度之间的信息差异受到目标分布、目标数量以及传感器相对位置变化的影响。随机控制的CS散度整体较低，说明缺少状态反馈的航向选择难以持续获得较大的信息增益。

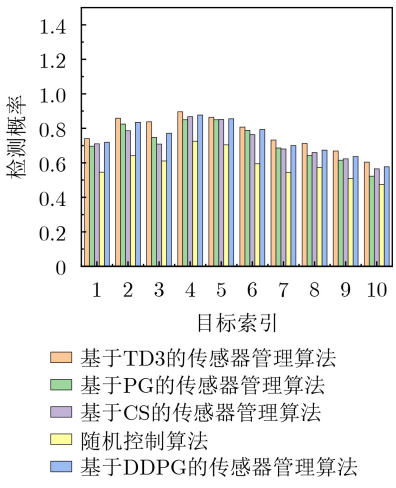

图 3 10个目标的检测概率对比

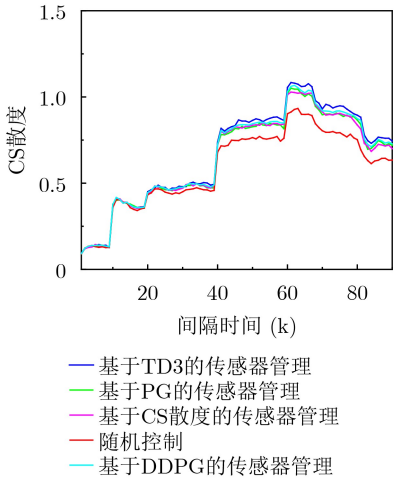

图 4 500次蒙特卡洛实验的CS散度结果

PG方法和CS散度方法在多个时段提高了CS散度，DDPG-LMB在多数时刻又高于两种离散控制基线。TD3-LMB的曲线在多数时刻位于较高位置，表明其所选航向角能够使伪更新后的LMB密度相对于预测密度产生较大的信息变化。曲线在若干时间段出现同步上升或下降，也说明场景状态变化会同时影响不同传感器控制策略的信息获取结果。CS散度评价的是当前候选动作带来的信息变化，其数值较大并不直接等同于最终状态误差最低，但能够反映该动作对后验更新的潜在贡献。因此，CS散度结果可作为理解后续OSPA变化的重要中间指标。

500次蒙特卡洛实验下不同算法的OSPA结果如图5所示。仿真初始阶段各算法的OSPA均较高，随后随着滤波递推逐渐下降；在部分时间步，OSPA出现短时升高，反映出目标状态或目标数量变化会暂时增加多目标估计难度。随机控制的OSPA整体较高，PG方法和Rényi散度方法能够在一定程度上降低多目标跟踪误差，说明基于状态或信息增益的动作选择优于缺少反馈的随机动作。DDPG-LMB采用连续航向角控制，其OSPA在多数时刻低于两种离散控制基线。TD3-LMB的OSPA整体最低，表明其在当前场景下获得了较低的多目标综合跟踪误差。由于OSPA同时反映目标状态定位误差和目标数目估计误差，该结果与图3中的检测概率和图4中的信息增益结果共同说明，较好的观测条件有助于改善LMB滤波器的整体估计结果。

5 结束语

本文的主要工作和创新在于提出了面向多目标跟踪性能优化的 LMB-CS-TD3 移动传感器管理方法。在 LMB 随机有限集滤波框架下，本文利用预测多目标密度构造伪量测并完成候选动作伪更新，

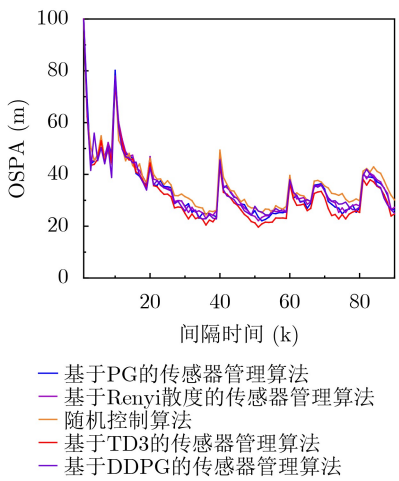

图 5 500次蒙特卡洛实验的OSPA结果

以预测密度与更新密度一阶统计矩之间的 CS 散度作为信息增益奖励，从而将传感器管理建模为信念空间内的连续动作优化问题。未来工作将进一步考虑动作相关量测噪声、速度/功率/驻留时间等多维动作变量以及多传感器协同管理问题。

参 考 文 献

[1] LIU Zhihao, SHANG Yuanyuan, LI Timing, *et al*. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark[J]. *IEEE Transactions on Multimedia*, 2023, 25: 1462–1476. doi: 10.1109/TMM.2023.3234822.

[2] HU Juan, ZUO Lei, VARSHNEY P K, *et al*. Resource allocation for distributed multitarget tracking in radar networks with missing data[J]. *IEEE Transactions on Signal Processing*, 2024, 72: 718–734. doi: 10.1109/TSP.2024.3352915.

[3] COX P B and VAN ROSSUM W L. Split-aperture phased array radar resource management for tracking tasks[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2025, 61(3): 6476–6486. doi: 10.1109/TAES.2025.3531843.

[4] HAYKIN S. Cognitive radar: A way of the future[J]. *IEEE Signal Processing Magazine*, 2006, 23(1): 30–40. doi: 10.1109/MSP.2006.1593335.

[5] ZHANG Peng, YAN Junkun, PU Wenqiang, *et al*. Multi-dimensional resource management scheme for multiple target tracking under dynamic electromagnetic environment[J]. *IEEE Transactions on Signal Processing*, 2024, 72: 2377–2393. doi: 10.1109/TSP.2024.3390119.

[6] CHARLISH A, HOFFMANN F, DEGEN C, *et al*. The development from adaptive to cognitive radar resource management[J]. *IEEE Aerospace and Electronic Systems Magazine*, 2020, 35(6): 8–19. doi: 10.1109/MAES.2019.2957847.

[7] KURNIAWATI H. Partially observable Markov decision processes and robotics[J]. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022, 5(1): 253–277. doi: 10.1146/annurev-control-042920-092451.

[8] MAHLER R. Objective functions for Bayesian control-theoretic sensor management, II: MHC-like approximation[M]. BUTENKO S, MURPHEY R, and PARDALOS P M. Recent Developments in Cooperative Control and Optimization. Boston: Springer, 2004: 273–316. doi: 10.1007/978-1-4613-0219-3_16.

[9] GOSTAR A K, HOSEINNEZHAD R, BAB-HADIASHAR A, *et al*. OSPA-based sensor control[C]. 2015 International Conference on Control, Automation and Information Sciences (ICCAIS), Changshu, China, 2015: 214–218. doi: 10.1109/ICCAIS.2015.7338664.

[10] JONES G, GARCÍA-FERNÁNDEZ Á F, and WONG P W. GOSPA-driven Gaussian Bernoulli sensor management[C]. 2023 26th International Conference on Information Fusion (FUSION), Charleston, USA, 2023: 1–8. doi: 10.23919/FUSION52260.2023.10224220.

[11] GOSTAR A K, HOSEINNEZHAD R, and BAB-HADIASHAR A. Multi-Bernoulli sensor control for multi-target tracking[C]. 2013 IEEE 8th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Melbourne, Australia, 2013: 312–317. doi: 10.1109/ISSNIP.2013.6529808.

[12] RISTIC B, VO B N, and CLARK D. A note on the reward function for PHD filters with sensor control[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2011, 47(2): 1521–1529. doi: 10.1109/TAES.2011.5751278.

[13] GOSTAR A K, HOSEINNEZHAD R, RATHNAYAKE T, *et al*. Constrained sensor control for labeled multi-Bernoulli filter using Cauchy-Schwarz divergence[J]. *IEEE Signal Processing Letters*, 2017, 24(9): 1313–1317. doi: 10.1109/LSP.2017.2723924.

[14] HOFFMANN F, CHARLISH A, RITCHIE M, *et al*. Sensor path planning using reinforcement learning[C]. 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 2020: 1–8. doi: 10.23919/FUSION45008.2020.9190242.

[15] SHI Yuchun, JIU Bo, YAN Junkun, *et al*. Data-driven simultaneous multibeam power allocation: When multiple targets tracking meets deep reinforcement learning[J]. *IEEE Systems Journal*, 2021, 15(1): 1264–1274. doi: 10.1109/JSYST.2020.2984774.

[16] LU Ziyang and GURSOY M C. Resource allocation for multi-target radar tracking via constrained deep reinforcement learning[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2023, 9(6): 1677–1690. doi: 10.1109/TCCN.2023.3304634.

[17] OBOREH-SNAPPS O, SHE Buxin, FAHAD S, *et al*. Virtual synchronous generator control using twin delayed deep deterministic policy gradient method[J]. *IEEE Transactions on Energy Conversion*, 2024, 39(1): 214–228. doi: 10.1109/TEC.2023.3309955.

张新迪：男，博士生，研究方向为多目标跟踪和传感器管理.

陈　辉：男，教授，博士生导师，研究方向为多目标跟踪、数据融合、最优控制等.

张虹芸：女，博士生，研究方向为扩展目标跟踪和雷达资源调度.

连　峰：男，教授，博士生导师，研究方向为多源信息融合、滤波与估计算法.

张光华：男，副教授，博士生导师，研究方向为目标跟踪、信息融合和随机有限集等.

阴志鹏：男，研究方向为目标跟踪、信息融合.

责任编辑：廖海贝

Labeled Multi-Bernoulli Sensor Management Strategy Based on Twin-Delayed Deep Deterministic Policy Gradient Learning Mechanism

ZHANG Xin-di[①] CHEN Hui[①] ZHANG Hong-yun[①] LIAN Feng[②]
ZHANG Guang-hua[②] YIN Zhi-peng[③]

[①](*School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou* 730050, *China*)
[②](*School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an* 710049, *China*)
[③](*Hebei Xintian Kechuang New Energy Technology Co., Ltd., Zhangjiakou* 075000, *China*)

Abstract:

Objective Multi-target tracking requires sensor management to adapt the observation process to clutter, missed detections, target-number variations, and changes in target motion. Conventional methods often search over a finite set of sensor actions, resulting in increasing computational cost and limited control resolution. Moreover, rewards formed by combining several single-target quantities may not adequately represent the joint multi-target posterior. A continuous-action sensor management method integrating the Twin-Delayed Deep Deterministic Policy Gradient algorithm with the Labeled Multi-Bernoulli filter is therefore developed to optimize mobile-sensor heading according to the multi-target belief state.

Methods The LMB posterior, including target existence probabilities and state densities, is used to construct the belief state. At each filtering step, the mobile sensor selects a continuous heading angle that affects the sensor-target geometry, detection probability, and LMB update. Predicted target states and candidate headings are used to generate ideal measurements and obtain pseudo-updated LMB densities. The Cauchy-Schwarz divergence between the predicted and pseudo-updated densities is used to construct the information-gain reward. TD3 employs two critics, target policy smoothing, and delayed actor updates to reduce value-estimation bias. Random control, policy-gradient control, an information-driven discrete method, and DDPG-based continuous control are used for comparison.

Results and Discussions DDPG-LMB and TD3-LMB produce more continuous steering changes than the discrete-action methods (Fig. 2). TD3-LMB obtains the highest or near-highest detection probabilities for most targets (Fig. 3) and gives relatively large Cauchy-Schwarz divergence values during most time steps, while random control remains relatively low (Fig. 4). TD3-LMB also achieves the lowest overall OSPA in the tested scenario, with DDPG-LMB generally outperforming the discrete baselines (Fig. 5). These results show that continuous heading control improves observation quality and overall LMB tracking performance.

Conclusions A TD3-based continuous-action sensor management framework for the LMB filter is presented. Candidate heading actions are evaluated through pseudo-updated LMB densities and Cauchy-Schwarz divergence, linking action selection directly to the joint multi-target posterior. The simulation results show more continuous sensor motion, higher detection probabilities for most targets, larger information gain, and lower OSPA in the tested scenario. Future work will consider higher-dimensional actions and cooperative multi-sensor management.

Key words: Multi-target tracking; Sensor management; Deep reinforcement learning; Label Multi-Bernoulli filter