数据驱动的多智能体网络鲁棒包容控制

于　镝†

(北京信息科技大学 自动化学院, 北京 100192)

摘要: 针对输入受限的受扰多智能体网络, 提出具有领航层、估计层、控制层和跟随者层的新型鲁棒包容控制方案. 首先, 设计有限时间估值器获得跟随者的期望状态, 然后基于包容误差引入非均方折扣代价函数, 从而将鲁棒包容控制问题转换成受限最优控制问题. 并应用Laypunov拓展原理证明得到的最优控制策略使得网络实现一致最终有界稳定. 在系统动态完全未知的情况下, 采用提出的积分增强学习算法和执行器–评价器结构, 在线得到近似最优控制策略. 仿真结果验证了理论方案的有效性和可行性.

关键词: 数据驱动; 多智能体网络; 积分增强学习; 包容控制

引用格式: 于镝. 数据驱动的多智能体网络鲁棒包容控制. 控制理论与应用, 2020, 37(9): 1963 – 1970

DOI: 10.7641/CTA.2020.90433

Data-driven robust containment control of multi-agent networks

YU Di†

(College of Automation, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: A novel robust containment control scheme is proposed for multi-agent networks with constrained input, including leaders layer, estimation layer, control layer and followers layer. At first, finite time estimators are designed to obtain the desired states of followers. Then a non-quadratic discounted cost function is introduced based on the containment errors, so the robust containment control problem is transformed into a constrained optimal control problem. Moreover, the uniform ultimate bounded stability is verified of whole networks with obtained optimal control policy according to Lyapunov extension theorem. When the dynamics of followers are completely unknown, the approximate optimal control policy is obtained online applying the developed integral reinforcement learning algorithm and actor-critic architecture. Simulation results are provided to demonstrate the effectiveness of the proposed scheme.

Key words: data-driven; multi-agent networks; integral reinforcement learning; containment control

Citation: YU Di. Data-driven robust containment control of multi-agent networks. *Control Theory & Applications*, 2020, 37(9): 1963 – 1970

1　引言

由于多智能体协调控制在众多领域中存在广泛成功的应用, 所以其研究受到广大研究人员的关注. 譬如自组装机器人聚集、无人机火灾救援、卫星姿态调整和智能电网分配等等. 作为典型的协调控制, 包容控制由于在危险物资搬运和火灾救援等军事和民用方面具有潜在的大量应用, 已经吸引了众多学者的研究热情. 在包容控制中, 存在多个领航者, 并且跟随者的运动限定在领航者所围成的最小几何空间中. 迄今为止, 在多智能体网络包容控制研究方面已经涌现出很多优秀的研究成果[1–4].

但上述成果均要求系统动态已知且非最优控制. 在实际应用中, 未知的外界环境可导致系统动态的不确定性变化, 由原有动态得到的控制方法并不准确或奏效. 因此基于数据驱动的控制思想深受研究人员的青睐, 主要依据可测得的网络系统数据信息进行系统监控与故障诊断等行为. 并且在实现包容控制的同时需考虑能量的损耗, 所以需要实现最优控制. 作为非常典型的自适应动态规划方法, 增强学习(reinforcement learning, RL)思想已被研究人员用来解决这个有趣且具有挑战性的问题. RL方法中智能体与周围未知环境进行交互, 从而学习最优控制策略[5–7]. 因此, 对于线性和非线性系统, 文献[8–9]提出了连续时间在线策略迭代算法, 其由策略评估和策略更新两步组成,

收稿日期: 2019−06−09; 录用日期: 2020−03−26.

†通信作者. E-mail: yudizlg@aliyun.com; Tel.: +86 10-82426829.

本文责任编委: 高会军.

北京信息科技大学学科群建设项目(5121911003), 国家自然科学基金项目(61903043)资助.

Supported by the Subject Group Construction Project of Beijing Information Science and Technology University (5121911003) and the National Natural Science Foundation of China (61903043).

并且分别采用评价神经网络和执行神经网络参量化地表示值函数和控制策略. 在系统内动态信息未知的情况下, 得出最优控制解的收敛性. 在文献[10]中, 针对控制输入受限的非线性系统, 在系统转移动态未知的情况下, 拓展积分增强学习(integral reinforcement learning, IRL)方法来解决其最优跟踪控制问题且在保持激励条件下得出系统的收敛性和稳定性. 对于完全未知动态的非线性系统, 基于Nash平衡解和最小–最大优化思想设计跟踪控制器, 并且采用离策略RL算法来学习最优控制策略[11]. 而文献[12]对于动态完全未知的输入受限非线性系统, 合适的选取标称系统的代价函数使得获得的近似最优控制使得系统一致最终有界稳定. 并且提出积分增强学习算法基于系统数据同时更新值函数和控制策略来解决鲁棒自适应调节问题.

以上的成果均针对单个系统, 文献[13–14]将RL算法应用到多智能体系统的最优包容控制. 对于线性异构多智能体系统, 文献[13]基于内模原理并采用全状态反馈和静态输出反馈来研究输出包容问题. 在文献[14]中, 提出离策略增强学习算法来解决部分模型未知的线性多智能体系统的最优包容控制问题. 上述成果均未考虑控制输入受限和网络受扰情况. 然而, 在实际应用中均需限定执行器的幅值来满足物理结构和运行安全的要求, 而且网络个体会受到模型不确定性、随机干扰等非线性摄动的影响. 所以, 在考虑非线性扰动情况下研究输入受限的多智能体网络的鲁棒包容控制具有重要的理论意义和实际价值, 但此方面研究至今无人问津. 本文受文献[10, 12]的启发, 提出了包含有领航层、估计层、控制层和跟随者层的新型控制结构, 设计有限时间估值器以及在线无模型IRL算法实现输入受限的受扰网络的鲁棒包容控制. 本文从以下3个方面对现有成果进行了拓展:1) 与文献[10, 12]相比, 考虑多智能体网络的鲁棒包容控制, 比单个系统的跟踪控制或鲁棒调节要复杂得多; 2) 与文献[13–14]相比, 考虑输入受限的受扰多智能体网络的包容控制, 更具实际意义; 3) 与文献[1, 15]相比, 考虑系统动态未知情况下, 输入受限的多智能体网络的最优鲁棒包容控制, 降低了对系统动态的限制.

本文其余部分组成如下: 第2节介绍了相关定义及引理; 第3节阐述问题; 第4节给出本文控制方案的主要结果, 设计了有限时间估计器和IRL迭代学习算法, 并且证明了多智能体网络的最终一致有界稳定性; 第5节仿真研究验证了本文控制方案和学习算法的有效性; 最后得出结论.

2 预备知识

定义 1 设X是实矢量空间$V \subseteq \mathbb{R}^n$的集合. 用$\mathrm{Co}(X)$表示X的凸包,

$$\mathrm{Co}(X) = \{\sum_{i=1}^{k} \alpha_i x_i | x_i \in X, \alpha_i \in \mathbb{R}, \alpha_i \geqslant 0, \sum_{i=1}^{k} \alpha_i = 1, k = 1, 2, \cdots\}.$$

定义 2[16] 令$\Upsilon : U(V) \subseteq P \to Q$为一给定映射, 其中$P$和$Q$为Banach空间且$U(V)$代表$V$的邻域. 当且仅当存在有界线性算子$\Pi : X \to Y$使得对于满足$\|M\|_{\bar{\Omega}} = 1$和$M \in U(V)$的所有$M$及在零附近有$\lim\limits_{s\to 0} o(s)/s = 0$的所有实数$s$, 有式$\Upsilon(V + sM) - \Upsilon(V) = s\Pi(M) + o(s), s \to 0$成立, 则称映射$\Upsilon$在$V$处是Gâteaux可微的, 且$\Pi$称为$\Upsilon$在$V$处的Gâteaux导数. 则$V$处的Gâteaux导数定义为

$$\Pi(M) = \lim_{s\to 0} \frac{\Upsilon(V + sM) - \Upsilon(V)}{s}.$$

引理 1[16] 如果Gâteaux导数Υ'在V的邻域内存在, 且Gâteaux导数Υ'在V处是连续的, 则$\Pi = \Upsilon'(V)$也是V处的Frechet导数.

3 问题描述

令多智能体网络由智能体$\Sigma_i (i = 1, \cdots, n)$组成, 其对应的有向图为$\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{A})$. 令$F = \{1, \cdots, m\}$和$L = \{m + 1, \cdots, n\}$分别代表跟随者集合和领航者索引集合. 则$\mathcal{V}$由跟随者节点集$\mathcal{V}_{\mathrm{F}} = \{\nu_i, i \in F\}$和领航者节点集合$\mathcal{V}_{\mathrm{L}} = \{\nu_i, i \in L\}$组成. 本文的控制目的是只基于系统数据, 设计合适的近似最优控制策略驱使受扰的跟随者收敛并保持在领航者所构成的动态凸包中.

3.1 网络动态

跟随者动态描述为

$$\dot{x}_i(t) = f(x_i(t)) + g(x_i(t))(u_i(t) + d(x_i(t))), \quad i \in F, \tag{1}$$

其中: $x_i \in \mathbb{R}^p$和$u_i \in \Xi$分别代表第i个跟随者的状态矢量和控制输入矢量, $\Xi = \{u | u \in \mathbb{R}^q, \|u_i(t)\| \leqslant \alpha, i = 1, 2, \cdots, m\}$且$\alpha > 0$. $f(x_i(t)) \in \mathbb{R}^p$和$g(x_i(t)) \in \mathbb{R}^{p\times q}$是状态$x_i(t)$的连续未知函数, 且$d(x_i(t)) \in \mathbb{R}^q$代表干扰. 令初始状态$x_{i0} = x_i(0)$且$f(0) = 0$.

领航者的动态描述为

$$\dot{x}_i(t) = h(x_i(t)), \ i \in L, \tag{2}$$

其中: $x_i \in \mathbb{R}^p$是第i个领航者的状态矢量; $h(x_i(t)) \in \mathbb{R}^p$是状态$x_i(t)$的连续未知函数, 对于$\forall x_i \in \mathbb{R}^p$有$0 < \|h(x_i(t))\| < h_{\mathrm{M}}$且$h(0) = 0$. 令跟随者和领航者状态矢量分别由$x_{\mathrm{F}} = [x_1^{\mathrm{T}} \ \cdots \ x_m^{\mathrm{T}}]^{\mathrm{T}}$和$x_{\mathrm{L}} = [x_{m+1}^{\mathrm{T}} \ \cdots \ x_n^{\mathrm{T}}]^{\mathrm{T}}$代表, $u_{\mathrm{F}} = [u_1 \ \cdots \ u_m]^{\mathrm{T}}$代表跟随者的控制矢量. 为了简便起见, 有如下假设.

假设 1 令$f(x_i) + g(x_i)u_i$在包含原点的紧集$\Omega \in \mathbb{R}^p$内是Lipschitz连续的, 使得对于任何有限初始条件$x \in \mathbb{R}^p$和控制输入$u_i(t) \in \Xi$, 跟随者的轨迹是唯一的. 而且, 跟随者的动态(1)是可控的并且存在正数b_{f}和g_{M}使得$\|f(x)\| \leqslant b_{\mathrm{f}}\|x\|$, $0 < \|g(x)\| \leqslant g_{\mathrm{M}}$,

$\forall x \in \mathbb{R}^p$.

假设 2　干扰有界且$\|d(x)\| \leqslant d_{\rm M}, \forall x \in \mathbb{R}^p$, 其中$d_{\rm M}(x)$是已经有界函数且$d(0)=0,\ d_{\rm M}(0)=0$.

3.2　网络拓扑

在本文中令领航者之间无通信, 且领航者与跟随者之间通信是单向的, 即领航者发送信息. 所以跟随者之间的网络拓扑和领航者与跟随者之间的网络拓扑决定整个网络通信. 由此对Laplacian阵$\mathcal{L}$进行结构划分, 则

$$\mathcal{L}\begin{bmatrix} x_{\rm F} \\ x_{\rm L} \end{bmatrix} = \begin{bmatrix} \mathcal{T} & \mathcal{T}_{\rm d} \\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{bmatrix}\begin{bmatrix} x_{\rm F} \\ x_{\rm L} \end{bmatrix}, \quad (3)$$

其中: $\mathcal{T}=[\mathcal{T}_{ij}] \in \mathbb{R}^{m\times m},\ \mathcal{T}_{\rm d} \in \mathbb{R}^{m\times(n-m)}$.

假设 3　令跟随者之间的拓扑强连通, 并且对于每个跟随者至少存在一个领航者与其通信.

3.3　网络误差

定义误差函数$e_i = \sum\limits_{j=1}^{n} a_{ij}(x_i - x_j),\ i \in F$. 因此整个网络的误差动态可描述为$E = \mathcal{T}x_{\rm F} + \mathcal{T}_{\rm d}x_{\rm L}$, 其中$E=[e_1^{\rm T}\ \cdots\ e_m^{\rm T}]^{\rm T}$. 根据文献[15]中的引理2.3, 可得矩阵$\mathcal{T}$是强对角占优$M$矩阵. 因此$\mathcal{T}^{-1}=[T_{ij}] \in \mathbb{R}^{mp\times mp}$存在且正定, 则$-\mathcal{T}^{-1}\mathcal{T}_{\rm d}$是行和为1的非负矩阵. 因此跟随者的期望状态矢量可表示为$x_{\rm d}=-\mathcal{T}^{-1}\mathcal{T}_{\rm d}x_{\rm L}$, 其中$x_{\rm d}=[x_{\rm d1}^{\rm T}\ \cdots\ x_{\rm dm}^{\rm T}]^{\rm T}$. 令$e_{\rm c}=x_{\rm F}-x_{\rm d}$ 代表包容误差, 其中$e_{\rm c}=[e_{\rm c1}^{\rm T}\ e_{\rm c2}^{\rm T}\ \cdots\ e_{\rm cm}^{\rm T}]^{\rm T}$.

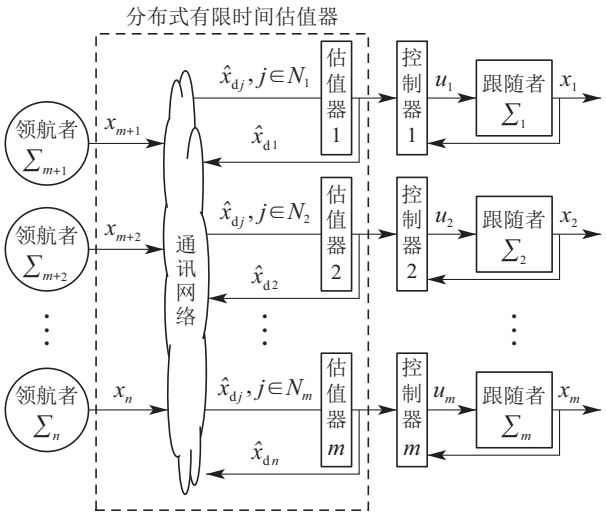

图 1　鲁棒最优包容控制结构示意图

Fig. 1　The diagram of robust optimal containment control

4　主要结果

由于领航者的动态只有部分跟随者已知, 所以需要设计估值器估计出跟随者在领航者所围成凸包中的期望状态. 因此, 本文提出鲁棒包容分布式结构如图1所示, 由领航者层、有限时间估计层、鲁棒最优包容控制层和跟随者层组成. 在估计层中, 有限时间估值器在有限时间内可获得$\hat{x}_{{\rm d}i} \to x_{{\rm d}i},\ i \in F$. 在控制层中, 基于跟随者期望状态的精确估计和后续提出的IRL算法, 跟随者的状态一致最终有界收敛到领航者所围成的凸包中.

4.1　有限时间估计

提出下列估值器:

$$\dot{\hat{x}}_{{\rm d}i} = -\beta\sum_{j=1}^{m} T_{ij}{\rm sgn}[\sum_{k=1}^{n} a_{jk}(\hat{x}_{{\rm d}j} - \hat{x}_{{\rm d}k})],\ i \in F, \quad (4)$$

其中$\beta>0$和$\hat{x}_{{\rm d}i}$代表第i个跟随者的期望状态$x_{{\rm d}i}$的估值, 而且$\hat{x}_{{\rm d}i}=x_i,\ i \in L$.

定理 1　考虑由动态(1)–(2)所描述的有向网络, 若假设3成立, 则当$\beta > h_{\rm M}\|\mathcal{T}_{\rm d}\|_\infty$时, 使用估值器(4)可在有限时间内实现$\hat{x}_{{\rm d}i} \to x_{{\rm d}i},\ i \in F$.

证　令估计误差为$\varepsilon_i = \hat{x}_{{\rm d}i} - x_{{\rm d}i}$和估计误差矢量为$\varepsilon=[\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_m]^{\rm T}$. 则采用与文献[13]中定理4.1类似的方法, 可得到当$t > \dfrac{\sqrt{2V'(0)}}{\beta - h_{\rm M}\|\mathcal{T}_{\rm d}\|_\infty}$时, 可获得跟随者期望状态的精确估计.　　证毕.

4.2　折扣非均方代价函数

针对式(1)的标称系统, 如式(5)所描述:

$$\dot{x}_i(t) = f(x_i(t)) + g(x_i(t))u_i(t),\ i \in F. \quad (5)$$

基于跟随者期望状态的精确估计可得到包容误差动态

$$\dot{e}_c = \dot{x}_{\rm F} - \dot{x}_{\rm d} = f(x_{\rm F}) + g(x_{\rm F})u_{\rm F} + \mathcal{T}^{-1}\mathcal{T}_{\rm d}x_{\rm L}, \quad (6)$$

其中:

$$f(x_{\rm F}) = [f^{\rm T}(x_1)\ f^{\rm T}(x_2)\ \cdots\ f^{\rm T}(x_m)]^{\rm T},$$
$$g(x_{\rm F}) = [g^{\rm T}(x_1)\ g^{\rm T}(x_2)\ \cdots\ g^{\rm T}(x_m)]^{\rm T}.$$

定义第i个跟随者的增广网络状态为$X_i=[e_{{\rm c}i}^{\rm T}\ \ x_{{\rm d}i}^{\rm T}]^{\rm T}$, 并依据式(2)(5)和式(6), 得到与第$i$个跟随者相关增广网络为

$$\dot{X}_i = G(X_i) + H(X_i)u_i, \quad (7)$$

其中: $G(X_i)=\begin{bmatrix} f(X_i) - \dot{x}_{{\rm d}i} \\ \dot{x}_{{\rm d}i} \end{bmatrix},\ H(X_i)=\begin{bmatrix} g(X_i) \\ 0 \end{bmatrix}$.

基于增广网络(7), 引入非均方折扣代价函数,

$$V(X_i) = \int_t^\infty {\rm e}^{-\gamma(\tau-t)}r(X_i,u_i)\,{\rm d}\tau, \quad (8)$$

其中折扣因子$\gamma>0$且$r(X_i,u_i)=X_i^{\rm T}Q'X_i + U(u_i) + \rho d_{\rm M}^2(X_i),\ \rho>0,\ Q'=\begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}$, Q 为正定矩阵, $U(u_i)$为正定函数且定义为

$$U(u_i) = 2\alpha\int_0^{u_i}(\tanh^{-1}(v/\alpha))^{\rm T}R\,{\rm d}v. \quad (9)$$

为了分析简便, 选择$R={\rm diag}\{r_1,r_2,\cdots,r_q\},\ r_i>0$, $i=1,\cdots,q$, 并用$\tanh^{-{\rm T}}(\cdot)$表示$(\tanh^{-1}(\cdot))^{\rm T}$.

4.3 Hamilton-Jacobi-Bellman(HJB)方程及最优解

在本节中, 给出与式(8)中代价函数相关的包容Bellman方程和HJB方程. 沿着增广网络轨迹(7)对$V(X_i)$取微分, 则获得下列包容Bellman方程:

$$\dot{V}(X_i)=\gamma\int_t^{\infty}\mathrm{e}^{-\gamma(\tau-t)}r(X_i,u_i)\,\mathrm{d}\tau-r(X_i,u_i)=\gamma V(X_i)-r(X_i,u_i). \tag{10}$$

可进一步得到

$$X_i^{\mathrm{T}}Q'X_i+U(u_i)+\rho d_{\mathrm{M}}^2(X_i)-\gamma V(X_i)+\dot{V}(X_i)=0. \tag{11}$$

定义哈密尔顿函数为

$$\mathcal{H}(X_i,u,\nabla V_{X_i})=X_i^{\mathrm{T}}Q'X_i+U(u_i)+\rho d_{\mathrm{M}}^2(X_i)-\gamma V(X_i)+(\nabla V_{X_i})^{\mathrm{T}}(G(X_i)+H(X_i)u_i), \tag{12}$$

其中$\nabla V_{X_i}=\dfrac{\partial V}{\partial X_i}\in\mathbb{R}^{2p}$. 令$V^*(X_i)$代表最优代价函数, 则求解HJB方程(13)可计算得到$V^*(X_i)$,

$$\mathcal{H}(X_i,u^*,\nabla V_{X_i}^*)=X_i^{\mathrm{T}}Q'X_i+U(u_i^*)+\rho d_{\mathrm{M}}^2(X_i)-\gamma V^*(X_i)+(\nabla V_{X_i}^*)^{\mathrm{T}}(G(X_i)+H(X_i)u_i^*)=0, \tag{13}$$

并且$V^*(0)=0$. 应用静止条件$\dfrac{\partial\mathcal{H}}{\partial u_i}=0$, 可得到最优控制输入如式(14)所示:

$$u_i^*(X_i)=-\alpha\tanh\Big(\frac{1}{2\alpha}R^{-1}H^{\mathrm{T}}(X_i)\nabla V_{X_i}^*\Big). \tag{14}$$

4.4 多智能体网络的最终一致有界稳定性

定理 2 考虑与代价函数(8)相关的增广网络(7). 若满足假设1和假设2的条件, 当$d^{\mathrm{T}}(X_i)Rd(X_i)\leqslant\rho d_{\mathrm{M}}^2(X_i)$时, 式(14)中的最优控制策略$u_i^*$确保整个网络最终一致有界.

证 令$V^*(X_i)$为包容HJB方程(13)的光滑正定解, u_i^*为式(14)给出的最优控制策略. 因为$V^*(X_i)$是等式(8)的最优值, 因此对于$X_i\neq0$有$V^*(X_i)>0$并且$V^*(0)=0$. 沿着受扰增广网络中第i个跟随者的轨迹对$V^*(X_i)$求导, 得到

$$\dot{V}_{X_i}^*=(\nabla V_{X_i}^*)^{\mathrm{T}}(G(X_i)+H(X_i)u_i^*)+(\nabla V_{X_i}^*)^{\mathrm{T}}H(X_i)d(X_i).$$

因为V^*连续可微, 且$\nabla V_{X_i}^*$有界, 则令$\|\nabla V_{X_i}^*\|\leqslant\mu_{\mathrm{M}}$, 其中$\mu_{\mathrm{M}}$为正常数. 基于式(13)–(14), 采用与文献[10]中定理1类似的方法, 则当$d^{\mathrm{T}}(X_i)Rd(X_i)\leqslant\rho d_{\mathrm{M}}^2(X_i)$时, 可得到

$$\dot{V}^*(X_i)\leqslant-\lambda_{\min}(Q)\|X_i\|^2+\frac{1}{2}\lambda_{\max}(R^{-1})g_{\mathrm{M}}^2\mu_{\mathrm{M}}^2+\gamma V^*(X_i),$$

其中$\lambda_{\min}(Q)$和$\lambda_{\max}(R^{-1})$分别是矩阵Q和R^{-1}的最小和最大特征值. 同时, 得到当$\gamma>0$时, $V^*(X_i)$有界, 并令$\|V^*(X_i)\|\leqslant\tau$. 则可以推断出当状态矢量在紧集

$$\Omega_{X_i}=\{X_i:\|X_i\|\leqslant\sqrt{\frac{\lambda_{\max}(R^{-1})g_{\mathrm{M}}^2\mu_{\mathrm{M}}^2+2\gamma\tau}{2\lambda_{\min}(Q)}}\}$$

之外时, $\dot{V}^*(X_i)<0$. 因此应用Lyapunov拓展原理得到最优控制u_i^*确保网络动态是最终一致有界的. 而且, 通过选择尽可能小的折旧因子γ以及尽可能大的矩阵Q, 可得到期望的包容误差, 从而实现受扰多智能体网络的最优鲁棒包容控制. 证毕.

因此可见, 通过求解HJB方程(13), 可得到$V^*(X_i)$和对应的u_i^*, 从而实现整个多智能体网络的最优鲁棒包容控制. 然而, 式(13)为非线性偏微分方程, 得到其解析解极其困难. 因此, 在下节中采用提出的IRL算法来求解HJB方程.

4.5 IRL迭代算法

在本小节中, 首先引入基于模型的策略迭代算法, 该算法是后面提出的基于数据的IRL迭代算法的基础.

算法 I 基于模型的迭代算法.

算法的步骤如下: 令$V^0\in V_0$为初始的代价函数, 其数值可由文献[17]中的引理5所确定. 因此初始控制策略$u_i^{(0)}=-\alpha\tanh(\dfrac{1}{2\alpha}R^{-1}H^{\mathrm{T}}\nabla V_{X_i}^{(0)})$, 令$k=0$.

Step 1 根据下述式子求解$V^{(k+1)}$:

$$(\nabla V_{X_i}^{(k+1)})^{\mathrm{T}}(G+Hu_i^{(k)})+r(X_i,u_i^{(k)})-\gamma V^{(k+1)}=0. \tag{15}$$

Step 2 由下式更新控制策略:

$$u_i^{(k+1)}=-\alpha\tanh\Big(\frac{1}{2\alpha}R^{-1}H^{\mathrm{T}}\nabla V_{X_i}^{(k+1)}\Big). \tag{16}$$

Step 3 若$\|V^{(k)}-V^{(k-1)}\|\leqslant\varepsilon$, 其中$\varepsilon$为计算精度, 则停止并获得最优代价函数$V^*=V^{(k)}$和最优控制策略$u^*=u^{(k)}$, 否则, 令$k=k+1$, 然后返到Step 1并继续.

下面算法I的收敛性借助牛顿迭代法进行证明. 考虑Banach空间$\Psi\subset V(X,t):\bar{\Omega}\to\mathbb{R}$, 定义映射

$$\Upsilon=r(X_i,u_i)-\gamma V(X_i)+\frac{\partial V^{\mathrm{T}}}{\partial X_i}(G+Hu_i). \tag{17}$$

然后基于定义2和引理1, 可得到以下引理.

引理2 令Υ定义如式(17)所示, 则其在V处的Frechet导数为

$$\Upsilon'(V)M=\Pi(M)=(\frac{\partial M}{\partial X_i})^{\mathrm{T}}(G-\alpha H\tanh D_i-\gamma M), \tag{18}$$

其中$D_i=\dfrac{1}{2\alpha}R^{-1}H^{\mathrm{T}}\dfrac{\partial V}{\partial X_i}$.

证　首先得出Υ在V处的Gâteaux导数, 然后证明其连续性. 基于式(17)中Υ的表达式以及定义2, 可得出Υ在V处的Gâteaux导数

$$\Pi(M)=\lim_{s\to 0}\frac{\Upsilon(V+sM)-\Upsilon(V)}{s}=$$
$$(\frac{\partial M}{\partial X_i})^{\mathrm{T}}(G-\alpha H\tanh D_i)-\gamma M.$$

对于$\forall M_0\in\Psi$, 有下列不等式成立

$$\|\Pi(M)-\Pi(M_0)\|_{\bar{\Omega}}\leqslant$$
$$(\beta\|G\|_{\bar{\Omega}}+\beta\|\alpha H\tanh D_i\|_{\bar{\Omega}}+\gamma)\|M-M_0\|_{\bar{\Omega}},$$

其中$\|\dfrac{\partial(M-M_0)}{\partial X_i}\|_{\bar{\Omega}}\leqslant\beta\|M-M_0\|_{\bar{\Omega}}$. 因此, 对于$\forall\epsilon>0$, 存在$\eta=\epsilon/\chi$使得当$\|M-M_0\|_{\bar{\Omega}}<\eta$时, 有$\|\Pi(M)-\Pi(M_0)\|_{\bar{\Omega}}<\epsilon$, 这表明$\Pi=\Upsilon'(V)$是连续的. 因此根据引理1可证得式(18).　　证毕.

定理 3　基于算法I, 迭代序列$V^{(k+1)}$和$u_i^{(k+1)}$都收敛到它们的最优值, 即当$k\to\infty$, 有$V^{(k+1)}\to V^*$, $u_i^{(k+1)}\to u_i^*$.

证　基于式(17)得到

$$\Upsilon'(V^{(k)})V^{(k)}-\Upsilon(V^{(k)})=$$
$$X_i^{\mathrm{T}}Q'X_i-\rho d_{\mathrm{M}}^2(X_i)+$$
$$\alpha(\frac{\partial V^{(k)}}{\partial X_i})^{\mathrm{T}}H\tanh(D_i^k)-\alpha^2\bar{R}\ln(1-(u_i^{(k)}/\alpha)^2)=$$
$$-r(X_i,u_i^{(k)}),$$

且基于式(18), 可得到

$$\Upsilon'(V^{(k)})V^{(k+1)}=$$
$$(\frac{\partial V^{(k+1)}}{\partial X_i})^{\mathrm{T}}(G-\alpha H\tanh D_i^k)-\gamma V^{(k+1)}=$$
$$-r(X_i,u_i^{(k)}).$$

因此$\Upsilon'(V^{(k)})V^{(k)}-\Upsilon(V^{(k)})=\Upsilon'(V^{(k)})V^{(k+1)}$, 即

$$V^{(k+1)}=V^{(k)}-\frac{\Upsilon(V^{(k)})}{\Upsilon'(V^{(k)})}.\tag{19}$$

则推断出算法I等价于牛顿迭代序列(19), 而且, 根据文献[17]中的引理4和引理5可以得出牛顿迭代序列(19)一定收敛到HJB方程(13)的解.　　证毕.

显而易见, 算法I依赖系统动态信息, 然而, 由于外部环境的复杂性很难获得这些信息. 在此种情况下, 设计无模型迭代算法势在必行.

算法 II　无模型IRL策略迭代算法.

针对数据样本集, 强化学习算法强调在探索新的数据样本和利用已有数据样本之间达到平衡. 鉴于此, 用下式描述与第i个跟随者相关的增广网络的轨迹动态:

$$\dot{X}_i=G(X_i)+H(X_i)u_i^{(k)}+H(X_i)(u_i-u_i^{(k)}),\tag{20}$$

并选取$u_i=u_i^{(k)}+n_{\mathrm{e}}$, 其中$n_{\mathrm{e}}\in\Phi$是探索信号且$\Phi$为有界集. 则沿着此轨迹对$V^*(X)$求导, 应用式(13), 可得

$$\dot{V}^{(k+1)}(X_i)-\gamma V^{(k+1)}(X_i)=$$
$$-r(X_i,u_i^{(k+1)})+(\nabla V_{X_i}^{(k+1)})^{\mathrm{T}}Hn_{\mathrm{e}}.\tag{21}$$

然后在等式(21)两边同时乘以$\mathrm{e}^{-\gamma(\tau-T)}$并对其在$t$和$t+T$之间取积分, 则有

$$\mathrm{e}^{-\gamma T}V^{(k+1)}(X_i(t+T))-V^{(k+1)}(X_i(t))=$$
$$-\int_t^{t+T}\mathrm{e}^{-\gamma(t-T)}r(X_i,u_i^{(k)})\,\mathrm{d}\tau-$$
$$2\alpha\int_t^{t+T}\mathrm{e}^{-\gamma(t-T)}\tanh^{-\mathrm{T}}(u_i^{(k+1)}/\alpha)Rn_{\mathrm{e}}\,\mathrm{d}\tau,\tag{22}$$

其中T是增强采样周期. 则无模型IRL算法如下所示. 初始条件的选取办法和算法I相同. 算法II的流程如图2所示.

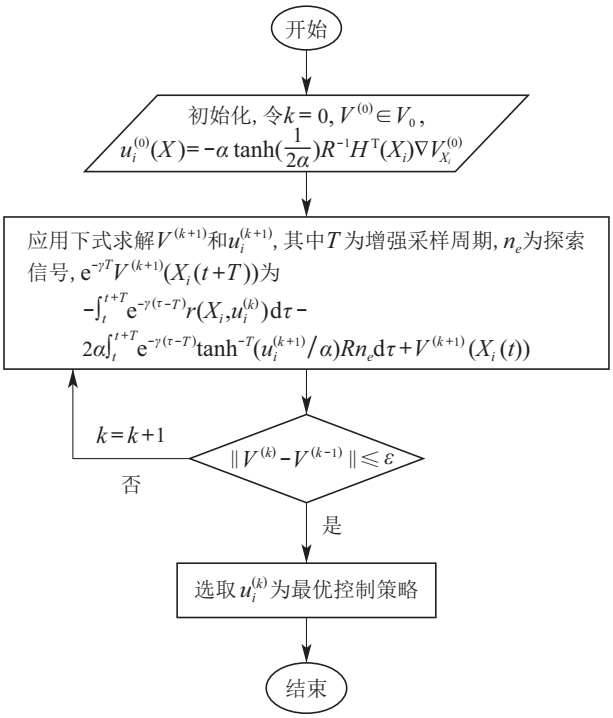

图 2　算法II流程图

Fig. 2　The flowchart of algorithm II

定理 4　采用算法II, 当$k\to\infty$时, 有$V^{(k+1)}\to V^*$和$u_i^{(k+1)}\to u_i^*$.

证　根据算法II的推导过程, 可得到算法I和算法II等价, 则基于定理3, 当$k\to\infty$时, 有$V^{(k+1)}\to V^*$和$u_i^{(k+1)}\to u_i^*$.　　证毕.

4.6　算法II的在线执行

在本节中, 为了实现算法II, 对每个跟随者应用执行器–评价器结构来逼近控制策略$u_i^{(k)}$以及代价函数$V^{(k)}(X_i)$. 方便起见, 定义$\mu_i^{(k)}=\tanh^{-1}(u_i^{(k)}/\alpha)$. 在此种情况下, 评价器NN和执行器NN都具有输入–

隐层–输出三层结构, 并且它们的输出由下式给出:

$$\hat{V}^{(k+1)}(X_i)=\hat{\omega}_i^{\mathrm{T}}\varphi(X_i),\ \hat{u}_i^{(k)}(X_i)=\hat{\varpi}_i^{\mathrm{T}}\phi(X_i),\quad(23)$$

其中: $\varphi=[\varphi_1\ \cdots\ \varphi_{r_1}]\in\mathbb{R}^{r_1}$ 和 $\phi=[\phi_1\ \cdots\ \phi_{r_2}]\in\mathbb{R}^{r_2}$为合适的隐层激励函数矢量, 且$\|\varphi\|\leqslant b_\varphi$, $\|\phi\|\leqslant b_\phi$; $\hat{\omega}_i^{\mathrm{T}}\in\mathbb{R}^{r_1}$和$\hat{\varpi}_i^{\mathrm{T}}\in\mathbb{R}^{r_2}$为常权值矢量. 然后将式(23)代入到式(22)得到

$$\begin{aligned}\delta(t)=&\hat{\omega}_i^{\mathrm{T}}(\mathrm{e}^{-\gamma T}\varphi(X_i(t+T))-\varphi(X_i(t))+\\&\int_t^{t+T}\mathrm{e}^{-\gamma(t-T)}r(X_i,u_i^{(k)})\,\mathrm{d}\tau+\\&2\alpha\sum_{j=1}^{q}r_j\int_t^{t+T}\mathrm{e}^{-\gamma(t-T)}\hat{\varpi}_{i,j}^{\mathrm{T}}\phi(X_i(\tau))n_{\mathrm{e}}\mathrm{d}\tau,\end{aligned}\quad(24)$$

其中$\delta(t)$是逼近误差. 然后重新整理式(24), 可得到

$$z(t)=\delta(t)+\hat{W}^{\mathrm{T}}y(t),\quad(25)$$

其中:

$$z(t)=-\int_t^{t+T}\mathrm{e}^{-\gamma(t-T)}r(X_i,u_i^{(k)})\,\mathrm{d}\tau,$$

$$\hat{W}=[\hat{\omega}_i^{\mathrm{T}}\ \hat{\varpi}_{i,1}^{\mathrm{T}}\ \cdots\ \hat{\varpi}_{i,q}^{\mathrm{T}}]^{\mathrm{T}},$$

$$y(t)=\begin{bmatrix}\mathrm{e}^{-\gamma T}\varphi(X_i(t+T))-\varphi(X_i(t))\\2\alpha r_1\displaystyle\int_t^{t+T}\mathrm{e}^{-\gamma(t-T)}\phi(X_i(\tau))n_{\mathrm{e}}\mathrm{d}\tau\\\vdots\\2\alpha r_q\displaystyle\int_t^{t+T}\mathrm{e}^{-\gamma(t-T)}\phi(X_i(\tau))n_{\mathrm{e}}\mathrm{d}\tau\end{bmatrix}.$$

为了最小化逼近误差, 采用最小二乘法进行计算. 假定从时间t_1到t_K内, 每隔相同的时间间隔T对系统数据进行充分的采样, 共得到$K\geqslant r_1+r_2q$组系统数据, 则得K组数据从而构成$Y=[y^{\mathrm{T}}(t_1)\ \cdots\ y^{\mathrm{T}}(t_K)]$和$Z=[z(t_1)\ \cdots\ z(t_K)]^{\mathrm{T}}$. 式(25)的最小二乘解等于$\hat{W}=(YY^{\mathrm{T}})^{-1}YZ$. 因此得到式(22)中的$V^{(k+1)}$和$u_i^{(k+1)}$的近似值.

5 仿真研究

本节用3组仿真研究验证仿真结果的有效性.

5.1 仿真实验1

考虑由8个智能体组成的多智能体网络. 有向拓扑如图3所示. 第i个跟随者动态由下式所描述:

$$\dot{x}_i=v_i,\dot{v}_i=-x_i^3-0.5v_i+u_i+d_i,\quad(26)$$

其中非线性干扰选为 $d_i=v_i\sin^3x_i\cos(0.5v_i),\ i=1,2,3,4$. 假设控制上界为 $|u_i|\leqslant0.25,\ i=1,2,3,4$. 式(4)中的参数选为$\beta=10$, 式(8)中的参数选为$Q=\begin{bmatrix}20&0\\0&0\end{bmatrix}$, $R=1$, $\rho=3$和$\gamma=0.1$.

对于第i个跟随者, 其评价器NN和执行器NN的激励函数分别选为 $\varphi(X_i)=[e_{ci}^2\ \ e_{ci}\dot{e}_{ci}\ \ \dot{e}_{ci}^2]^{\mathrm{T}}$ 和 $\phi(X_i)=[e_{ci}\ \ \dot{e}_{ci}\ \ e_{ci}\dot{e}_{ci}]^{\mathrm{T}}$. 采样周期选为$T=0.01$且探索信号的选择与文献[12]类似. 网络拓扑满足假设3, 参数R, β和ρ的选取满足定理1和定理2的条件. 跟随者的期望状态的有限时间估计误差变化曲线如图4所示, 可见不到2 s便实现$\hat{x}_{\mathrm{d}i}=x_{\mathrm{d}i}$, $i\in F$. 基于文献[12]中提出的无模型IRL算法和本文得到的上述估值及所提出的无模型IRL算法, 可实现受扰多智能体网络的鲁棒最优包容控制. 智能体的运动轨迹分别如图5和图6所示. 其中: 实心方块代表跟随者的初始位置, 实心圆点代表动态领航者分别在不同时刻的位置. 而且, 4种不同线型的曲线代表跟随者的实际运动轨迹, 黑色方框代表领航者所围成的动态凸包. 由仿真结果可得, 当基于文献[12]中的控制方案时, 跟随者在20 s左右进入到领航者所围成的凸包中. 而采用本文所提出的控制方案时, 跟随者在15 s左右便进入到领航者所围成的凸包中. 可见本文的控制方法能够使得跟随者更加快速地收敛并保持在领航者所围成的凸包中, 在其期望轨迹的微小邻域内运动.

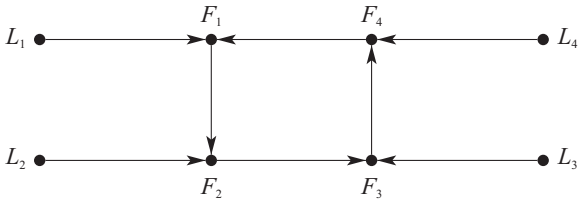

图 3 网络拓扑结构1

Fig. 3 The structure of No.1 network topology

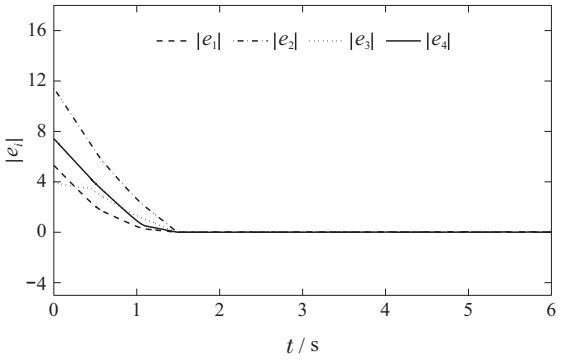

图 4 估值误差变化曲线

Fig. 4 The curves of estimation error

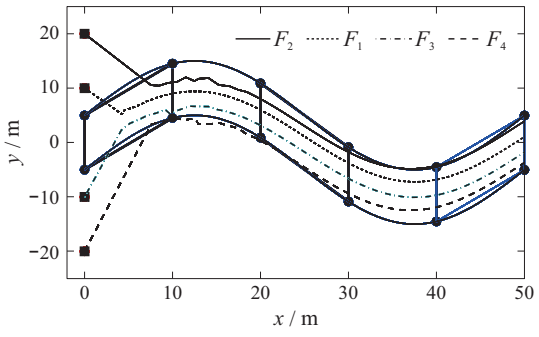

图 5 受扰多智能体网络运动轨迹(基于文献[12]的算法)

Fig. 5 The trajectories of perturbed multi-agent network (based on the algorithm in [12])

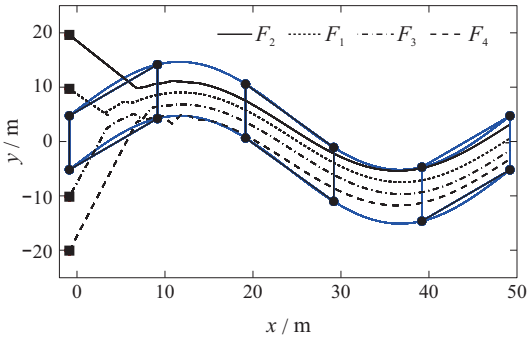

图 6　受扰多智能体网络运动轨迹(基于本文的算法)

Fig. 6　The trajectories of perturbed multi-agent network (based on the proposed algorithm in the paper)

5.2　仿真实验2

本小节考虑当跟随者与多个领航者存在通信时, 由10个智能体组成的多智能体网络. 有向拓扑如图7所示. 网络动态同仿真实验1, 采用本文的控制方案和学习算法, 可实现受扰多智能体网络的鲁棒最优包容控制. 智能体的运动轨迹分别如图8所示. 可见跟随者在10 s内便可以收敛到领航者所围成的凸包中, 与网络拓扑1的仿真结果比较具有快速性. 并且进行了多组实验分析折扣因子对网络控制效果的影响, 得出$\gamma \leqslant 0.05$时跟随者运动轨迹收敛的结论. 可见不同的网络拓扑结构直接影响网络控制参数的选取.

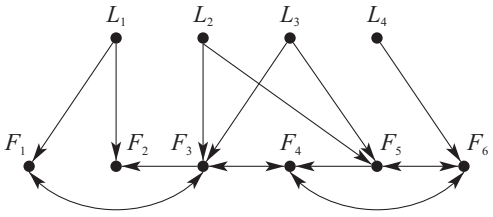

图 7　网络拓扑结构2

Fig. 7　The structure of No.2 network topology

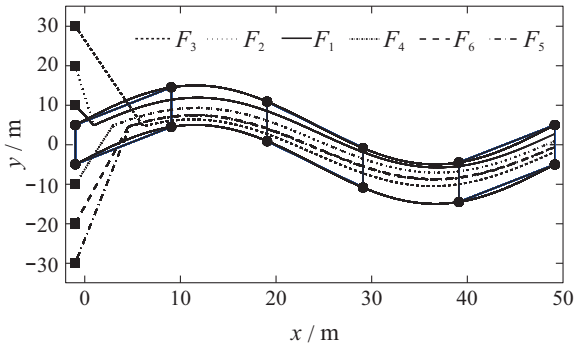

图 8　受扰多智能体网络运动轨迹

Fig. 8　The trajectories of perturbed multi-agent network

5.3　仿真实验3

本小节考虑多AmigoBots机器人[18]网络, 网络拓扑结构如图9所示. 微分驱动轮式机器人模型如图10所示.

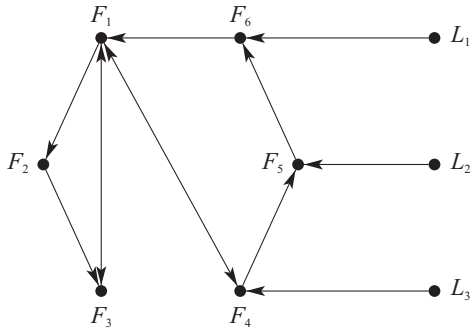

图 9　网络拓扑结构3

Fig. 9　The structure of No.3 network topology

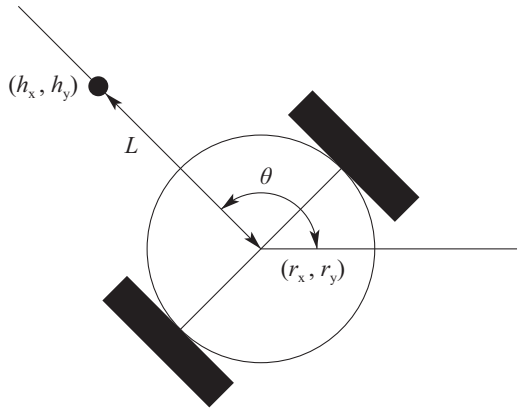

图 10　微分驱动轮式机器人模型

Fig. 10　The model of differentially driven wheeled mobile robot

第i个机器人的位姿位置用$h_i \triangleq [h_{\mathrm{x}i}\ h_{\mathrm{y}i}]^{\mathrm{T}}$表示, 该点位于与轮轴垂直的线上, 并且与轮轴中心交点相距d_i, 轮轴中心点用$r_i \triangleq [r_{\mathrm{x}i}\ r_{\mathrm{y}i}]^{\mathrm{T}}$表示. 令$(r_{\mathrm{x}i}, r_{\mathrm{y}i})$, θ_i, (v_i, ω_i)分别代表第i个机器人的轮轴中心位置、导航角、线速度和角速度. 则第i个机器人的动态方程为

$$\dot{r}_{\mathrm{x}i} = v_i \cos\theta_i,\ \dot{r}_{\mathrm{y}i} = v_i \sin\theta_i,\ \dot{\theta}_i = \omega_i, \tag{27}$$

则可以得出

$$\begin{bmatrix} \dot{h}_{\mathrm{x}i} \\ h_{\mathrm{y}i} \\ \dot{h}_{\mathrm{x}i} \\ \dot{h}_{\mathrm{y}i} \end{bmatrix} = \begin{bmatrix} p_{i1} \\ p_{i2} \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ u_{\mathrm{x}i} \\ u_{\mathrm{y}i} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ q_{i1} \\ q_{i2} \end{bmatrix}, \tag{28}$$

其中: $u_i = [u_{\mathrm{x}i}\ u_{\mathrm{y}i}]^{\mathrm{T}}$代表第$i$个机器人的控制作用,

$$\begin{aligned} p_{i1} &= \cos\theta_i v_i - d_i \omega_i \sin\theta_i, \\ p_{i2} &= \sin\theta_i v_i + d_i \omega_i \cos\theta_i, \\ q_{i1} &= -\sin\theta_i v_i \omega_i - d_i \cos\theta_i \omega_i^2, \\ q_{i2} &= \cos\theta_i v_i \omega_i - d_i \sin\theta_i \omega_i^2. \end{aligned}$$

由此采用本文提出的控制方案对多机器人网络进行仿真研究. 其中$d_i = 0.15$ m, $T = 6$ min, 机器人网络的运动轨迹如图11所示. 其中: 空心方块代表跟随者的初始位置, 实心圆点代表动态领航者分别在 $t = 0$, $\dfrac{T}{4}$, $\dfrac{T}{2}$, $\dfrac{3T}{4}$不同时刻的位置. 而且, 6种不同线型的曲

线代表跟随者的实际运动轨迹, 蓝色方框代表领航者所围成的动态凸包. 仿真结果表明受扰多机器人网络同样可实现鲁棒包容控制.

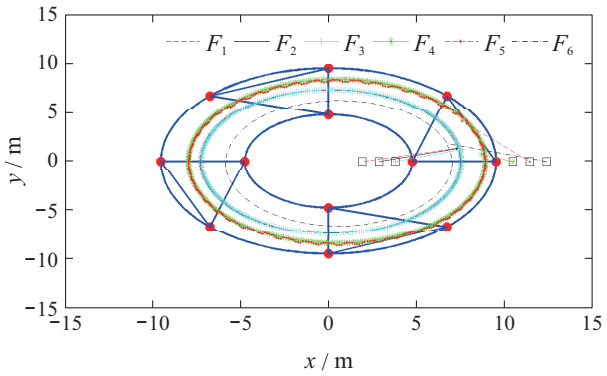

图 11 多机器人网络运动轨迹

Fig. 11 The trajectories of multi-robot network

6 结论

本文提出新的控制方案解决输入受限多智能体网络的鲁棒包容控制问题. 基于包容误差和跟随者在领航者所围成凸包中的期望状态构建增广网络, 并引入非均方折扣代价函数和HJB方程获得最优控制策略. 为了克服系统动态完全未知的困难, 基于执行器–评价器结构和最小二乘法, 基于系统数据在线执行所提出的无模型IRL算法, 得到近似最优控制策略. 并且网络的最终一致有界稳定性和所提IRL算法的收敛性都得以证明. 下一步将针对有限域内的鲁棒包容控制以及避碰问题展开研究.

参考文献:

[1] YU D, JI X Y. Finite-time containment control of perturbed multi-agent systems based on sliding-mode control. *International Journal of Systems Science*, 2018, 49(6): 299 – 311.

[2] WEN S X, YUE D, SUN Z G, et al. Distributed robust finite-time attitude containment control for multiple rigid bodies with uncertainties. *International Journal of Robust and Nonlinear Control*, 2015, 25(15): 2561 – 2581.

[3] WANG H Z, WANG C, XIE G M. Finite-time containment control of multi-agent systems with static or dynamic leaders. *Neurocomputing*, 2017, 226(8): 1 – 6.

[4] DONG X W, SHI Z Y, LU G, et al. Formation-containment analysis and design for high-order linear time-invariant swarm systems. *International Journal of Robust and Nonlinear Control*, 2015, 25(17): 3439 – 3456.

[5] VAMVOUDAKIS K G, LEWIS F L. Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton-Jacobi equations. *Automatica*, 2011, 47(8): 1556 – 1569.

[6] ZHANG H, JIANG H, LUO Y, et al. Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method. *IEEE Transactions on Industrial Electronics*, 2017, 64(5): 4091 – 4100.

[7] JIAO Q, MODARES H, XU S, et al. Multi-agent zero-sum differential graphical games for disturbance rejection in distributed control. *Automatica*, 2016, 69(4): 24 – 34.

[8] VRABIE D, PASTRAVANU O, ABUKHALAF M, et al. Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, 2009, 45(2): 477 – 484.

[9] VRABIE D, LEWIS F L. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, 2009, 22(3): 237 – 246.

[10] MODARES H, LEWIS F L. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica*, 2014, 50(7): 1780 – 1792.

[11] MODARES H, LEWIS F L, JIANG Z P. H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(10): 2550 – 2562.

[12] YANG X, LIU D R, LUO B, et al. Data-based robust adaptive control for a class of unknown nonlinear constrained-input systems via integral reinforcement learning. *Information Science*, 2016, 369(6): 731 – 747.

[13] ZUO S, SONG Y D, LEWIS F L, et al. Output containment control of linear heterogeneous multi-agent systems using internal model principle. *IEEE Transactions on Cybernetics*, 2017, 47(8): 2099 – 2109.

[14] YANG Y L, MODARES H, WUNSCH D C, et al. Optimal containment control of unknown heterogeneous systems with active leaders. *IEEE Transactions on Control Systems Technology*, 2019, 27(3): 1228 – 1236.

[15] YU D, WU Q H. Finite time estimation and containment control of second order perturbed directed networks. *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference*. Orlando, FL, USA: IEEE, 2011: 4126 – 4131.

[16] ZEIDLER E. *Nonlinear Functional Analysis: Fixed Point Theorems*. New York: Springer-Verlag, 1985.

[17] WU H N, LUO B. Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear H_∞ control. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 23(12): 1884 – 1895.

[18] REN W, SORENSEN N. Distributed coordination architecture for multi-robot formation control. *Robotics and Autonomous Systems*, 2008, 56(4): 324 – 333.

作者简介:

于 镝 博士, 目前研究方向为多智能体协调控制、自适应动态规划、计算智能, E-mail: yudizlg@aliyun.com.