基于Q-learning的离散时间多智能体系统一致性

朱志斌, 王付永, 尹艳辉, 刘忠信†, 陈增强

(南开大学 人工智能学院, 天津 300350; 天津市智能机器人技术重点实验室, 天津 300350)

摘要: 针对模型未知的一类离散时间多智能体系统, 本文提出了一种Q-learning方法实现多智能体系统的一致性控制. 该方法不依赖于系统模型, 能够利用系统数据迭代求解出可使给定目标函数最小的控制律, 使所有智能体的状态实现一致. 通过各个智能体所产生的系统数据, 采用策略迭代的方法实时更新求解得到多智能体系统的控制律, 并对所提Q-learning方法进行了收敛性和稳定性分析. 最后, 论文给出了计算机仿真验证了所提方法的有效性.

关键词: 多智能体系统; 一致性; 离散时间; Q-learning

引用格式: 朱志斌, 王付永, 尹艳辉, 等. 基于Q-learning的离散时间多智能体系统一致性. 控制理论与应用, 2021, 38(7): 997 – 1005

DOI: 10.7641/CTA.2021.00533

Consensus of discrete-time multi-agent system based on Q-learning

ZHU Zhi-bin, WANG Fu-yong, YIN Yan-hui, LIU Zhong-xin†, CHEN Zeng-qiang

(College of Artificial Intelligence, Nankai University, Tianjin 300350, China;
Key Laboratory of Intelligence Robotics of Tianjin, Tianjin 300350, China)

Abstract: For a class of discrete-time multi-agent systems with unknown models, a Q-learning method is proposed in this paper to achieve consensus of multi-agent systems. The proposed method does not depend on the system model, and the optimal control law can be obtained through the iteration of system data. Based on the system data, policy iteration is adopted to calculate the optimal control law of the multi-agent systems. Convergence and stability analysis of the proposed Q-learning method for multi-agent systems is also given in this work. Finally, a simulation example is provided to verify the effectiveness of the proposed method.

Key words: multi-agent systems; consensus; discrete-time; Q-learning

Citation: ZHU Zhibin, WANG Fuyong, YIN Yanhui, et al. Consensus of discrete-time multi-agent system based on Q-learning. *Control Theory & Applications*, 2021, 38(7): 997 – 1005

1 引言

近年来, 随着计算机技术和网络技术的不断发展, 有关多智能体系统协同控制方面的研究越来越多. 多智能体系统具有单个智能体无法比拟的优势, 可以完成单个智能体无法完成的任务, 因而具有更广泛的应用, 例如: 无人飞行器编队控制、卫星姿态控制和移动多机器人等[1–5].

多智能体系统(multi-agent systems, MASs)的一致性是复杂动力学系统中一个具有理论和实践意义的重要问题. 现有的一致性问题主要分为两类: 无领导者多智能体系统的一致性和领导–跟随多智能体系统的一致性. 对于无领导者的多智能体系统, 当所有智能体的状态收敛至同一值时, 则系统达到一致; 对于领导–跟随的多智能体系统, 当系统中所有跟随者的状态趋于领导者的状态, 则系统达到一致. 基于多智能体系统的一致性和协同控制方面的研究[6], 文献[7]利用动态输出反馈的控制方法研究了在固定和切换拓扑下多智能体系统的一致性, 并提出了相应的一致性算法. 为了解决线性异构多智能体系统在切换拓扑下的一致性问题, 文献[8]设计了一种分布式分级控制协议. 对于二阶非线性多智能体系统的一致性问题, 文献[9]推导出固定网络拓扑在部分间歇通信的情况下达到一致的充分条件. 假设系统同时存在网络时延

收稿日期: 2020−08−12; 录用日期: 2021−03−29.

†通信作者. E-mail: lzhx@nankai.edu.cn; Tel.: +86 22-85358276.

本文责任编委: 陈皓勇.

天津市自然科学基金项目(20JCYBJC01060, 20JCQNJC01450), 国家自然科学基金项目(61973175), 南开大学中央高校基本科研业务费专项资金项目(63201196)资助.

Supported by the Tianjin Natural Science Foundation of China (20JCYBJC01060, 20JCQNJC01450), the National Natural Science Foundation of China (61973175) and the Fundamental Research Funds for the Central Universities, Nankai University (63201196).

与状态时延, 文献[10]研究了一类二阶马尔可夫切换多智能体系统的一致性问题. 然而, 上述研究均是在模型参数已知的条件下, 而面对现实中复杂的物理系统, 绝对精确的数学模型是难以得到的. 所以, 为了避免建模难度大或者模型未知的问题, 设计一种基于数据的分布式控制律则更为有效.

强化学习作为机器学习的一种方法[11–12], 以环境反馈作为输入并通过不断的试错来寻找最优的行为策略. 不同于监督学习, 智能体并不知道如何做出正确的行为策略, 但可以通过强化学习的方法对行为策略进行评估, 并根据有效反馈信息对行为策略进行改善. 同时, 强化学习的奖励函数对系统信息的需求更少, 也更容易设计, 因此, 强化学习适合解决复杂的控制问题. 在有关强化学习的基础数学研究取得突破之后, 对强化学习的研究也越来越多[13–14].

强化学习还可以解决最优控制问题, 例如带有约束的最优控制[15–16]、带有时延的最优控制[17–18]、最优跟踪控制[19–20]、最优一致性控制[21]等. 作为强化学习中一种重要的方法, Q-leaning是一种无模型的学习方法. 目前, 已经有很多关于Q-learning的研究, 例如: 跟踪控制[22]、零和博弈[23]、鲁棒控制[24]等.

经过上述讨论可知, 在系统模型已知的前提下, 文献[7–10]可以很好地解决多智能体系统的一致性问题, 但是当系统模型未知时, 解决这一问题将变得困难. 为了避免复杂的机理建模, 本文提出一种Q-learning方法解决离散时间多智能体系统的一致性问题, 主要贡献如下:

1) 将Q-learning方法从单智能体拓展应用到多智能体系统中, 解决了离散时间多智能体系统模型未知时的一致性控制问题;

2) 对于多智能体强化学习, 本文在文献[26]的基础上对值函数的结构进行了优化改进, 提出了一种新的关于误差的值函数.

本文结构安排如下: 第2部分将给出基础代数图理论和问题描述; 第3部分将给出的Q-learning一致性算法, 并对所提算法进行了收敛性和稳定性分析; 第4部分将通过计算机仿真验证所提算法的有效性; 最后, 第5部分将对全文工作进行总结和展望.

2 问题描述

2.1 代数图论

令$\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{A})$表示一个有向加权图. 其中: $\mathcal{V}=\{v_1,v_2,\cdots,v_N\}$表示具有$N$个节点的集合, $\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$表示边集, 节点下标集合为$\mathcal{I}=\{1,2,\cdots,N\}$. 定义矩阵$\mathcal{A}=[a_{ij}]$是图$\mathcal{G}$的非负邻接矩阵, 矩阵元素$a_{ij}\geqslant 0$表示节点$v_i$和$v_j$之间的连接权重. 当节点$v_i$可以收到来自节点$v_j$的信息传递时, 则$a_{ij}\geqslant 0$; 否则$a_{ij}=0$, 本文所讨论的图中, $a_{ii}=0,\forall i,j\in\mathcal{I}$. 如果$a_{ij}=a_{ji}$, 则称图$\mathcal{G}$是无向的, 显然无向图对应的加权邻接矩阵$\mathcal{A}=[a_{ij}]$是对称的. 节点$v_i$的邻居下标集合为$N_i=\{j|v_j\in\mathcal{V}(v_j,v_i)\in\mathcal{E}\}$且邻居的个数表示为$|N_i|$. 定义度矩阵为$\mathcal{D}=\mathrm{diag}\{d_i,i=1,2,\cdots,N\}$, $d_i=\sum\limits_{j=1}^{n}a_{ij}$为矩阵$\mathcal{A}$的第$i$行元素的和, 节点$v_i$的入度和出度分别定义为$d_{\mathrm{in}}(v_i)=\sum\limits_{j=1}^{n}a_{ji}$, $d_{\mathrm{out}}(v_i)=\sum\limits_{j=1}^{n}a_{ij}$. 若图$\mathcal{G}$中每个节点的入度都等于出度, 则称图$\mathcal{G}$是平衡图. 定义图$\mathcal{G}$的Laplace矩阵为$L=\mathcal{D}-\mathcal{A}$. 对于图中的节点$v_i$和$v_j$, 存在有序下标集合$\{k_1,k_2,\cdots,k_l\}$, 若$a_{ik_1}>0$, $a_{k_1k_2}>0,\cdots,a_{k_lj}>0$, 则称节点$v_i$和$v_j$之间存在一条有向连接路径. 如果对于图中任意的两个节点v_i和v_j之间存在至少一条有向连接路径, 则称图$\mathcal{G}$是强连通的.

2.2 问题描述

假设多智能体系统由N个智能体组成, 且智能体间的通信网络拓扑是固定、无向和连通的. 给出下面的离散时间系统的一致性算法:

$$x_i(k+1)=Ax_i(k)+Bu_i(k),i=1,\cdots,N,\quad(1)$$

$$u_i(k)=\varepsilon K\sum_{j\in N_i}a_{ij}(x_j(k)-x_i(k)),\quad(2)$$

其中: $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times m}$和$K\in\mathbb{R}^{m\times n}$, $x_i(k)\in\mathbb{R}^n$是智能体i在$k=1,2,3,\cdots$时刻的状态, $u_i(k)\in\mathbb{R}^m$是其控制律, $\varepsilon\in(0,1/\Delta)$是系统的控制参数, $\Delta=\max\limits_i d_{\mathrm{out}}(v_i)>0$为网络节点的最大出度.

假设 1 假设本文中的系统模型是确定且未知的.

控制律(2)可以用每个智能体自身信息和其邻居智能体的信息求得, 那么每个智能体和其邻居智能体的局部误差定义如下:

$$\delta_i(k)=\sum_{j\in N_i}a_{ij}(x_j(k)-x_i(k)),\quad(3)$$

根据式(3)可以得到全局误差向量

$$\delta(k)=[\delta_1^{\mathrm{T}}(k)\ \ \delta_2^{\mathrm{T}}(k)\ \ \cdots\ \ \delta_N^{\mathrm{T}}(k)]^{\mathrm{T}}\in\mathbb{R}^{nN}.$$

本文目标是求解最优控制律(2), 使得当$k\to\infty$时, $\|\delta(k)\|\to 0$, 所有智能体的状态达到一致, 即

$$\lim_{k\to\infty}\|x_i(k)-x_i^*\|=0,$$

各个智能体的最终状态满足[6]

$$x_1^*=x_2^*=\cdots=x_N^*,$$

同时使后面第2.3节中的目标函数式(4)最小.

2.3 值函数的定义

对于每个智能体定义一个如下所示的目标函数:

$$J_i(\delta_i(k),u_i(k),u_j(k))=\sum_{k=0}^{\infty}\gamma^k U_i(\delta_i(k),u_i(k),u_j(k)),\tag{4}$$

上式中的效用函数$U_i(\delta_i(k),u_i(k),u_j(k))$定义如下:

$$U_i(\delta_i(k),u_i(k),u_j(k))=\delta_i^{\mathrm{T}}(k)Q_{ii}\delta_i(k)+u_i^{\mathrm{T}}(k)R_{ii}u_i(k)+\sum_{j\in N_i}u_j^{\mathrm{T}}(k)R_{ij}u_j(k),\tag{5}$$

其中: $Q_{ii}>0\in\mathbb{R}^{n\times n}$, 而$R_{ii}>0\in\mathbb{R}^{m\times m}$, $R_{ij}>0\in\mathbb{R}^{m\times m}$均为正定对称矩阵, $0<\gamma\leqslant 1$为折扣因子.

对每个智能体及其邻居智能体给定控制律$(u_i(l),u_j(l))$, 每个智能体的值函数定义如下

$$V_i(\delta(k))=\sum_{l=k}^{\infty}\gamma^{l-k}U_i(\delta_i(l),u_i(l),u_j(l)),\tag{6}$$

其中$\delta(k)=[\delta_1^{\mathrm{T}}(k)\ \ \delta_2^{\mathrm{T}}(k)\ \ \cdots\ \ \delta_N^{\mathrm{T}}(k)]^{\mathrm{T}}\in\mathbb{R}^{nN}$.

注 1　目标函数(4)用来评价智能体i的性能, 智能体i的值函数(6)可以收集局部信息.

定义 1(容许控制[25])　如果控制律$u_i(k),\forall i\in\mathcal{I}$不仅可以使系统稳定, 还可以保证目标函数有界, 则称其为容许控制.

在满足容许控制律的条件下, 值函数可以改写成贝尔曼方程的形式

$$V_i(\delta(k))=U_i(\delta_i(k),u_i(k),u_j(k))+\gamma V_i(\delta(k+1)).\tag{7}$$

引理 1(二次型值函数)　在满足容许控制律的条件下, 智能体i的值函数式(6)可以改写为如下二次型的形式

$$V_i(\delta(k))=\delta^{\mathrm{T}}(k)P_i\delta(k).\tag{8}$$

证　此引理的证明可以通过以下步骤完成:

步骤 1　根据式(3), 智能体i和邻居智能体的局部误差动力学方程为

$$\delta_i(k+1)=\sum_{j\in N_i}a_{ij}[x_j(k+1)-x_i(k+1)]=A\delta_i(k)-d_iBu_i(k)+\sum_{j\in N_i}a_{ij}Bu_j(k),\tag{9}$$

根据式(9)可得

$$\delta(k+1)=(I\otimes A)\delta(k)+(L\otimes B)u(k),\tag{10}$$

其中$u(k)=[u_1^{\mathrm{T}}(k)\ \ u_2^{\mathrm{T}}(k)\ \ \cdots\ \ u_N^{\mathrm{T}}(k)]^{\mathrm{T}}\in\mathbb{R}^{mN}$. 根据式(2)和式(3), 可以得到$u_i(k)=\varepsilon K\delta_i(k)$, 整理成紧凑的形式

$$u(k)=(I\otimes\varepsilon K)\delta(k),\tag{11}$$

将式(11)代入式(10), 得到

$$\delta(k+1)=(I\otimes A)\delta(k)+(L\otimes\varepsilon BK)\delta(k)=(I\otimes A+L\otimes\varepsilon BK)\delta(k)=A_c\delta(k).\tag{12}$$

步骤 2　定义两个块对角矩阵

$$\bar{Q}_i=\begin{pmatrix}0&&&\\&\ddots&&\\&&Q_{ii}&\\&&&\ddots\\&&&&0\end{pmatrix},$$

$$\bar{R}_i=\begin{pmatrix}0&&&\\&\ddots&&\\&&R_{ii}&\\&&&\ddots\\&&&&0\end{pmatrix}+\begin{pmatrix}0&&&\\&\ddots&&\\&&R_{ij}&\\&&&\ddots\\&&&&0\end{pmatrix},$$

其中: $\bar{Q}_i\geqslant 0\in\mathbb{R}^{nN\times nN}$为对角矩阵, 矩阵元素$Q_{ii}\geqslant 0\in\mathbb{R}^{n\times n}$所在位置为第$i$行、第$i$列, 其他位置元素均为0; $\bar{R}_i\geqslant 0\in\mathbb{R}^{mN\times mN}$为对角矩阵, 矩阵元素$R_{ii}>0\in\mathbb{R}^{m\times m}$所在位置为第$i$行、第$i$列, 矩阵元素$R_{ij}>0\in\mathbb{R}^{m\times m}$所在位置为第$j$行、第$j$列$j\in N_i$, 剩余元素均为0.

步骤 3　根据式(11)–(12), 式(6)可以改写为

$$\begin{aligned}&V_i(\delta(k))=\\&\sum_{l=k}^{\infty}\gamma^{l-k}U_i(\delta_i(l),u_i(l),u_j(l))=\\&\sum_{l=k}^{\infty}\gamma^{l-k}[\delta_i^{\mathrm{T}}(l)Q_{ii}\delta_i(l)+u_i^{\mathrm{T}}(l)R_{ii}u_i(l)+\\&\sum_{j\in N_i}u_j^{\mathrm{T}}(l)R_{ij}u_j(l)]=\\&\sum_{l=k}^{\infty}\gamma^{l-k}[\delta^{\mathrm{T}}(l)\bar{Q}_i\delta(l)+u^{\mathrm{T}}(l)\overline{R}_iu(l)]=\\&\sum_{l=k}^{\infty}\gamma^{l-k}[\delta^{\mathrm{T}}(l)\bar{Q}_i\delta(l)+\\&\delta^{\mathrm{T}}(l)((I\otimes\varepsilon K)^{\mathrm{T}}\bar{R}_i(I\otimes\varepsilon K))\delta(l)]=\\&\sum_{l=0}^{\infty}\gamma^{l}[\delta^{\mathrm{T}}(k+l)\bar{Q}_i\delta(k+l)+\\&\delta^{\mathrm{T}}(k+l)((I\otimes\varepsilon K)^{\mathrm{T}}\bar{R}_i(I\otimes\varepsilon K))\delta(k+l)]=\\&\delta(k)[\sum_{l=0}^{\infty}\gamma^{l}((A_c^l)^{\mathrm{T}}((I\otimes\varepsilon K)^{\mathrm{T}}\bar{R}_i(I\otimes\varepsilon K)+\\&\bar{Q}_i)A_c^l)]\delta(k)=\\&\delta(k)P_i\delta(k),\end{aligned}\tag{13}$$

其中

$$P_i=\sum_{l=0}^{\infty}\gamma^l[(A_c^l)^{\mathrm{T}}(\bar{Q}_i+(I\otimes\varepsilon K)^{\mathrm{T}}\bar{R}_i(I\otimes\varepsilon K))A_c^l].$$

证毕.

基于贝尔曼最优性原则, 智能体i的最优值函数满足离散时间的Hamilton-Jacobi-Bellman (HJB)方程

$$V_i^*(\delta(k))=\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\gamma V_i^*(\delta(k+1))\},\tag{14}$$

通过求偏导, 即$\dfrac{\partial V_i^*(\delta(k))}{\partial u_i(k)}=0$, 可以得到最优控制律$u_i^*(k)$

$$u_i^*(k)=\arg\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\gamma V_i^*(\delta(k+1))\}.\tag{15}$$

定义 2(纳什均衡解[25]) 如果包含N个控制律的控制序列是N个智能体博弈的全局纳什均衡解, 那么满足

$$V_i^*=V_i(u_1^*,u_2^*,\cdots,u_i^*,\cdots,u_N^*)\leqslant V_i(u_1^*,u_2^*,\cdots,u_i,\cdots,u_N^*).$$

根据定义2, 智能体i的耦合离散时间HJB方程为

$$V_i^*(\delta(k))=U_i(\delta_i(k),u_i^*(k),u_j^*(k))+\gamma V_i^*(\delta(k+1)).\tag{16}$$

3 基于Q-learning的多智能体系统的一致性算法

对于模型未知的多智能体系统, 已有的基于模型的算法是无效的. 尽管可以通过系统信息建立模型, 但是实际系统中的不确定因素导致难以建立准确的系统模型. 由于Q-learning算法不需要构建系统模型, 可以利用多智能体系统产生的数据在线实时更新控制律, 故本文采用此方法解决多智能体系统一致性问题.

3.1 Q-learning算法

基于贝尔曼方程的离散时间Q函数定义如下:

$$Q_i(\delta(k),u_i(k)=U_i(\delta_i(k),u_i(k),u_j(k))+\gamma V_i(\delta(k+1)),\tag{17}$$

其中Q函数由智能体i及其邻居智能体的状态信息和控制律构成. 根据式(17), 可以得到

$$Q_i(\delta(k),u_i(k))=V_i(\delta(k)).\tag{18}$$

构造Q函数的二次型的形式

$$Q_i(\delta(k),u_i(k))=\begin{bmatrix}\bar{\delta}_i(k)\\u_i(k)\end{bmatrix}^{\mathrm{T}}H_i\begin{bmatrix}\bar{\delta}_i(k)\\u_i(k)\end{bmatrix}=Z_i^{\mathrm{T}}(k)H_iZ_i(k),\tag{19}$$

其中: $\bar{\delta}_i(k)=[\delta_i^{\mathrm{T}}(k)\ \delta_{ij1}^{\mathrm{T}}(k)\ \delta_{ij2}^{\mathrm{T}}(k)\ \cdots\ \delta_{ijp}^{\mathrm{T}}(k)]^{\mathrm{T}}\in\mathbb{R}^{n(p+1)}$由全局误差向量$\delta(k)$中不为0的元素构成, $Z_i(k)=[\bar{\delta}_i^{\mathrm{T}}(k)\ u_i^{\mathrm{T}}(k)]^{\mathrm{T}}\in\mathbb{R}^{n(p+1)+m}$; $H_i=H_i^{\mathrm{T}}$; 矩阵H_i表示为

$$H_i=\begin{bmatrix}H_{\delta\delta}&H_{\delta u_i}\\H_{u_i\delta}&H_{u_iu_i}\end{bmatrix}=\begin{bmatrix}H_{i(\delta_i\delta_i)}&H_{i(\delta_i\delta_{j1})}&\cdots&H_{i(\delta_i\delta_{jp})}&H_{i(\delta_iu_i)}\\H_{i(\delta_{j1}\delta_i)}&H_{i(\delta_{j1}\delta_{j1})}&\cdots&H_{i(\delta_{j1}\delta_{jp})}&H_{i(\delta_{j1}u_i)}\\\vdots&\vdots&&\vdots&\\H_{i(\delta_{jp}\delta_i)}&H_{i(\delta_{jp}\delta_{j1})}&\cdots&H_{i(\delta_{jp}\delta_{jp})}&H_{i(\delta_{jp}u_i)}\\H_{i(u_i\delta_i)}&H_{i(u_i\delta_{j1})}&\cdots&H_{i(u_i\delta_{jp})}&H_{i(u_iu_i)}\end{bmatrix},$$

$j1,j2,\cdots,jp\in N_i$, p为智能体i的邻居个数. 为了表达方便, 将$H_{i(u_i(k)\delta_i(k))}$和$H_{i(\delta_i(k)\delta_{jp}(k))}$分别写作$H_{i(u_i\delta_i)}$和$H_{i(\delta_i\delta_{jp})}$的形式.

为了得到能使目标函数最小的线性控制律$u_i(k)$, 对Q函数求关于$u_i(k)$的偏导, 并使$\dfrac{\partial Q_i}{\partial u_i(k)}=0$, 则有如下方程

$$\frac{\partial Q_i}{\partial u_i(k)}=2H_{i(u_iu_i)}u_i(k)+2H_{i(u_i\delta_i)}\delta_i(k)+\sum_{j\in N_i}2H_{i(u_i\delta_j)}\delta_j(k)=0,\tag{20}$$

求解式(20)可以得到

$$u_i(k)=-H_{i(u_iu_i)}^{-1}(H_{i(u_i\delta_i)}\delta_i(k)+\sum_{j\in N_i}H_{i(u_i\delta_j)}\delta_j(k))=-H_{i(u_iu_i)}^{-1}\bar{H}_i\delta(k)=K_i\delta(k),\tag{21}$$

其中: $\bar{H}_i=[H_{i(u_i\delta_i)}\ H_{i(u_i\delta_{j1})}\ \cdots\ H_{i(u_i\delta_{jp})}]$, 而$K_i=-H_{i(u_iu_i)}^{-1}\bar{H}_i$是反馈控制增益矩阵. 根据式(17)–(18), Q函数改写为

$$Q_i(\delta(k),u_i(k))=U_i(\delta_i(k),u_i(k),u_j(k))+\gamma Q_i(\delta(k+1),u_i(k+1)),\tag{22}$$

将式(24)带入到式(27), Q函数还可以写成

$$Z_i^{\mathrm{T}}(k)H_iZ_i(k)=\delta_i^{\mathrm{T}}(k)Q_{ii}\delta_i(k)+u(k)\bar{R}_iu(k)+\gamma Z_i^{\mathrm{T}}(k+1)H_iZ_i(k+1),\tag{23}$$

其中$u(k)=[u_1^{\mathrm{T}}(k)\ u_2^{\mathrm{T}}(k)\ \cdots\ u_N^{\mathrm{T}}(k)]^{\mathrm{T}}\in\mathbb{R}^{mN}$.

由式(21)(23)可知, 在策略评估和策略迭代时不需要任何系统模型信息. 矩阵H_i可以通过最小二乘法实

时迭代求解[26]. 将策略迭代直接应用到所提Q-learning算法中, 给出如下算法:

步骤 1　初始化: 对于每个智能体i, 给定初始容许控制律$u_i^0(k), \forall i=1,2,\cdots,N$; $r=0$, r表示迭代次数, 总的迭代次数为$\mathcal{R}$;

步骤 2　策略评估: 根据式(23)迭代求解每个智能体i的第$r+1$次矩阵H_i

$$\begin{aligned}&Q_i(\delta(k),u_i(k))=\\&\delta_i^{\mathrm{T}}(k)Q_{ii}\delta_i(k)+u^T(k)\bar{R}_iu(k)+\\&\gamma Z_i^{\mathrm{T}}(k+1)H_iZ_i(k+1);\end{aligned}\tag{24}$$

步骤 3　策略迭代: 利用求得的矩阵H_i更新第$r+1$迭代控制律$u_i^{r+1}(k), \forall i$

$$\begin{aligned}u_i^{r+1}(k)=-H_{i(u_iu_i)}^{-1}(&H_{i(u_i\delta_i)}\delta_i(k)+\\&\sum_{j\in N_i}H_{i(u_i\delta_j)}\delta_j(k));\end{aligned}\tag{25}$$

步骤 4　若$r=\mathcal{R}$, 停止迭代; 否则, $r=r+1$并返回步骤2.

注 2　为了确保数据训练过程中对状态空间有更充分的探索, 策略迭代算法一般需要一定的激励条件ξ, ξ是会随着迭代次数的增加而逐渐衰减至0的.

3.2　算法的收敛性分析

为了保证系统稳定, 下面对所提的算法进行收敛性分析.

假设 2　假设智能体之间的通信网络拓扑图$\mathcal{G}$是固定、无向且连通的.

引理 2[27]　对 $i\in\mathcal{I}$ 和 $\forall r\in\mathcal{N}$ 值函数$Q_i^r(\delta(k), u_i^r(k))$和控制律$u_i^r(k)$分别通过式(24)–(25)进行更新. 如果智能体i给定的初始控制律u_i^0是容许的, 那么迭代控制律$u_i^r(k)$也是容许的, 则$u_i^r(k)$不仅可以使系统(1)稳定, 还可以保证目标函数(4)有界.

引理 3　对于$\forall i\in\mathcal{I}$和$\forall r\in\mathcal{N}$, 给定一个初始容许控制律(2), 即$u_i^0(k)$, 分别通过式(24)–(25)计算$Q_i^r(\delta(k),u_i^r(k))$和$u_i^r(k)$, 可以得到函数$Q_i^r(\delta(k), u_i^r(k))$是单调非递增的, 即

$$Q_i^{r+1}(\delta(k),u_i^r(k))\leqslant Q_i^r(\delta(k),u_i^r(k)).$$

证　此引理的证明可以通过以下步骤完成:

步骤 1　对于$\forall i\in\mathcal{I}$和$\forall r\in\mathcal{N}$, 定义一个新的性能指标函数$\psi_i^{r+1}(\delta(k),u_i^{r+1}(k))$

$$\begin{aligned}&\psi_i^{r+1}(\delta(k),u_i^{r+1}(k))=\\&U_i(\delta_i(k),u_i^{r+1}(k),u_j^{r+1}(k))+\\&\gamma Q_i^r(\delta(k+1),u_i^r(k+1)\,)=\\&\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\\&\gamma Q_i^r(\delta(k+1),u_i^r(k+1)\,)\}\leqslant\\&Q_i^r(\delta(k),u_i^r(k)),\end{aligned}\tag{26}$$

根据引理2, 对于$\forall i\in\mathcal{I}$和$\forall r\in\mathcal{N}$, 迭代控制律$u_i^r(k)$总是容许的. 因此, 当$k\to\infty$时, $\delta(k)\to 0$.

步骤 2　令$k=T$当$T\to\infty$可以得到

$$\begin{aligned}&Q_i^{r+1}(\delta(T),u_i^{r+1}(T))=\\&\psi_i^{r+1}(\delta(T),u_i^{r+1}(T))=\\&Q_i^r(\delta(T),u_i^r(T)).\end{aligned}\tag{27}$$

令$k=T-1$, 根据式(24)–(25)(27)可得

$$\begin{aligned}&Q_i^{r+1}(\delta(T-1),u_i^{r+1}(T-1)\,)=\\&U_i(\delta_i(T-1),u_i^{r+1}(T-1),u_j^{r+1}(T-1)\,)+\\&\gamma Q_i^{r+1}(\delta(T),u_i^{r+1}(T))=\\&U_i(\delta_i(T-1),u_i^{r+1}(T-1),u_j^{r+1}(T-1)\,)+\\&\gamma Q_i^r(\delta(T),u_i^r(T))=\\&\min_{u_i(T-1)}\{U_i(\delta_i(T-1),u_i(T-1),u_j(T-1)\,)+\\&\gamma Q_i^r(\delta(T),u_i^r(T))\}\leqslant\\&U_i(\delta_i(T-1),u_i^r(T-1),u_j^r(T-1)\,)+\\&\gamma Q_i^r(\delta(T),u_i^r(T))=\\&Q_i^r(\delta(T-1),u_i^r(T-1)),\end{aligned}\tag{28}$$

根据式(28), 当$k=T-1$时, $Q_i^{r+1}(\delta(k),u_i^{r+1}(k))\leqslant Q_i^r(\delta(k),u_i^r(k))$.

步骤 3　假设对$\forall q=0,1,2,\cdots$, 当$k=q+1$时式(28)成立, 即

$$\begin{aligned}&Q_i^{r+1}(\delta(q+1),u_i^{r+1}(q+1))\leqslant\\&Q_i^r(\delta(q+1),u_i^r(q+1)),\end{aligned}$$

令$k=q$, 则

$$\begin{aligned}&Q_i^{r+1}(\delta(q),u_i^{r+1}(q))=\\&U_i(\delta_i(q),u_i^{r+1}(q),u_j^{r+1}(q))+\\&\gamma Q_i^{r+1}(\delta(q+1),u_i^{r+1}(q+1)\,)\leqslant\\&U_i(\delta_i(q),u_i^{r+1}(q),u_j^{r+1}(q))+\\&\gamma Q_i^r(\delta(q+1),u_i^r(q+1)\,)=\\&\psi_i^{r+1}(\delta(q),u_i^{r+1}(q))\leqslant Q_i^r(\delta(q),u_i^r(q)).\end{aligned}\tag{29}$$

根据式(29), 对于$\forall i$和$\forall r$, $Q_i^{r+1}(\delta(k),u_i^{r+1}(k))\leqslant Q_i^r(\delta(k),u_i^r(k))$, $\forall k=0,1,2,\cdots$成立.

证毕.

定理 1 在假设1–2成立的条件下, 给定系统(1)的任意初始容许控制律, 分别通过式(24)–(25)计算得到矩阵 H_i 和控制律 $u_i^r(k)$. 当 $r\to\infty$ 时, $Q_i^r(\delta(k), u_i^r(k))$将收敛至最优值函数$Q_i^*(\delta(k), u_i^r(k))$, $u_i^r(k)$将收敛至最优控制律$u_i^*(k)$.

证 此定理的证明可以通过以下步骤完成:

步骤 1 定义

$$Q_i^\infty(\delta(k),u_i^\infty(k))=\lim_{r\to\infty}Q_i^r(\delta(k),u_i^r(k)),$$

根据引理3, 值函数$Q_i^r(\delta(k),u_i^r(k))$是单调非递增的. 根据式(26)可得

$$\begin{aligned}&Q_i^\infty(\delta(k),u_i^\infty(k))\leqslant\\&\psi_i^{r+1}(\delta(k),u_i^{r+1}(k))=\\&\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\\&\gamma Q_i^r(\delta(k+1),u_i^r(k+1)\,)\}.\end{aligned}\tag{30}$$

当$r\to\infty$时, 满足

$$\begin{aligned}&Q_i^\infty(\delta(k),u_i^\infty(k))\leqslant\\&\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\\&\gamma Q_i^\infty(\delta(k+1),u_i^\infty(k+1)\,)\}.\end{aligned}\tag{31}$$

步骤 2 因为$Q_i^r(\delta(k),u_i^r(k))$是单调非递增的, 则一定存在一个迭代次数$R$使不等式成立, 则可以得到

$$\begin{aligned}&Q_i^\infty(\delta(k),u_i^\infty(k))\geqslant\\&Q_i^R(\delta(k),u_i^R(k))-a=\\&U_i(\delta_i(k),u_i^R(k),u_i^R(k))+\\&\gamma Q_i^R(\delta(k+1),u_i^R(k+1)\,)-a\geqslant\\&\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\\&\gamma Q_i^\infty(\delta(k+1),u_i^\infty(k+1)\,)\}-a,\end{aligned}\tag{32}$$

其中a是任意正常数, 所以

$$\begin{aligned}&Q_i^\infty(\delta(k),u_i^\infty(k))\geqslant\\&\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\\&\gamma Q_i^\infty(\delta(k+1),u_i^\infty(k+1)\,)\}.\end{aligned}\tag{33}$$

结合式(31)(33), 可得

$$\begin{aligned}&Q_i^\infty(\delta(k),u_i^\infty(k))=\\&\min_{u_i(k)}\{U_i(\delta_i(k),u_i(k),u_j(k))+\\&\gamma Q_i^\infty(\delta(k+1),u_i^\infty(k+1)\,)\}.\end{aligned}\tag{34}$$

步骤 3 定义$\phi_i(\delta(k),\mu_i(k))$为一个新的性能指标函数

$$\begin{aligned}&\phi_i(\delta(k),\mu_i(k))=\\&U_i(\delta_i(k),\mu_i(k),\mu_j(k))+\\&\gamma Q_i(\delta(k+1),\mu_i(k+1)\,),\end{aligned}\tag{35}$$

其中$\mu_i(k)$和$\mu_j(k)$为任意容许控制律.

令$k=p, p\to\infty$, 由引理2可知$\delta(k)\to 0$, 从而

$$Q_i^\infty(\delta(p),u_i^\infty(p))=\phi_i(\delta(p),\mu_i(p))=0.$$

令$k=p-1$可得

$$\begin{aligned}&\phi_i(\delta(p-1),\mu_i(p-1))=\\&U_i(\delta_i(p-1),\mu_i(p-1),\mu_j(p-1))+\\&\gamma\phi_i(\delta(p),\mu_i(p))\geqslant\\&\min_{u_i(p-1)}\{U_i(\delta_i(p-1),u_i(p-1),u_j(p-1))+\\&\gamma\phi_i(\delta(p),u_i(p))\}=\\&\min_{u_i(p-1)}\{U_i(\delta_i(p-1),u_i(p-1),u_j(p-1))+\\&\gamma Q_i^\infty(\delta(p),u_i^\infty(p))\}=\\&Q_i^\infty(\delta(p-1),u_i^\infty(p-1)).\end{aligned}\tag{36}$$

假设根据式(36), 当$k=s+1, \forall s=0,1,2,\cdots$时,

$$\begin{aligned}&\phi_i(\delta(s+1),\mu_i(s+1))\geqslant\\&Q_i^\infty(\delta(s+1),u_i^\infty(s+1))\end{aligned}$$

成立, 则

$$\begin{aligned}&\phi_i(\delta(s),\mu_i(s))=\\&U_i(\delta_i(s),\mu_i(s),\mu_j(s))+\\&\gamma\phi_i(\delta(s+1),\mu_i(s+1))\geqslant\\&U_i(\delta_i(s),\mu_i(s),\mu_j(s))+\\&\gamma Q_i^\infty(\delta(s+1),\mu_i^\infty(s+1))\geqslant\\&\min_{\mu_i(s)}\{U_i(\delta_i(s),u_i(s),u_j(s))+\\&\gamma Q_i^\infty(\delta(s+1),u_i^\infty(s+1))\}=\\&Q_i^\infty(\delta(s),u_i^\infty(s)).\end{aligned}\tag{37}$$

根据式(37), 可得

$$\begin{aligned}&\phi_i(\delta(s),\mu_i(s))\geqslant Q_i^\infty(\delta(s),u_i^\infty(s)),\\&\forall s=0,1,2,\cdots.\end{aligned}$$

步骤 4 令$\mu_i(k)=u_i^*(k), \forall i$, 可得

$$Q_i^\infty(\delta(k),u_i^\infty(k))\leqslant$$

$$\phi_i(\delta(k),\mu_i(k))=Q_i^*(\delta(k),u_i^*(k)). \tag{38}$$

因为$Q_i^*(\delta(k),u_i^*(k))$是最优性能指标值, 可得

$$Q_i^\infty(\delta(k),u_i^\infty(k))=\lim_{r\to\infty}Q_i^r(\delta(k),u_i^r(k))=$$
$$Q_i^*(\delta(k),u_i^*(k)). \tag{39}$$

结合式(38)–(39), 可得

$$Q_i^\infty(\delta(k),u_i^\infty(k))=\lim_{r\to\infty}Q_i^r(\delta(k),u_i^r(k))=$$
$$Q_i^*(\delta(k),u_i^*(k)).$$

当$r\to\infty$时, $Q_i^r(\delta(k),u_i^r(k))$收敛至最优性能指标值$Q_i^*(\delta(k),u_i^*(k))$, 此时

$$\lim_{r\to\infty}u_i^r(k)=u_i^*(k).$$

证毕.

3.3 稳定性分析

定理 2　在假设1和假设2成立的条件下, 系统(1)的每个智能体i的$V_i^*(\delta(k))$和$u_i^*(k)$分别满足耦合离散时间HJB方程(16)和最优控制律(15). 智能体i和其邻居的局部误差$\delta_i(k)$是渐近稳定的, 并且当$k\to\infty$时, $\delta_i(k)\to 0$, 此时所有智能体的状态达到一致.

证　根据式(21)可以得到

$$V_i^*(\delta(k))-\gamma V_i^*(\delta(k+1))=$$
$$U_i(\delta_i(k),u_i^*(k),u_j^*(k)). \tag{40}$$

定义Lyapunov函数的差分

$$\Delta(\gamma^k V_i^*(\delta(k)))=$$
$$\gamma^{k+1}V_i^*(\delta(k+1))-\gamma^k V_i^*(\delta(k)), \tag{41}$$

根据式(40), 式(41)可改写为

$$\Delta(\gamma^k V_i^*(\delta(k)))=$$
$$-\gamma^k U_i(\delta_i(k),u_i^*(k),u_j^*(k))\leqslant 0, \tag{42}$$

上述式(42)表明智能体i和其邻居的误差$\delta_i(k)$是渐近稳定的, 即当$k\to\infty$, $\delta_i(k)\to 0$, $i=1,2,\cdots,N$. 最终, 所有智能体的状态将趋于一致.

证毕.

4 仿真

本节用MATLAB仿真验证所提算法的有效性. 假设多智能体系统的通信网络拓扑如图1所示, 该网络是一个固定无向图, 且系统模型未知. 5个顶点分别代表5个智能体, 智能体之间的连接权值均为1. 每个智能体的动力学方程满足式(1), 令

$$A=\begin{bmatrix}0.985 & -0.09887\\ 0.09887 & 0.985\end{bmatrix},\ B=\begin{bmatrix}0.8 & -1\\ 0 & 0.9\end{bmatrix}.$$

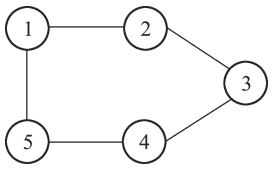


图 1　多智能体系统的通信拓扑

Fig. 1　The communication topology of multi-agent systems

给定初始状态为

$$x_1^0=\begin{bmatrix}17\\-4\end{bmatrix},\ x_2^0=\begin{bmatrix}22\\9\end{bmatrix},\ x_3^0=\begin{bmatrix}13\\5\end{bmatrix},$$
$$x_4^0=\begin{bmatrix}11\\12\end{bmatrix},\ x_5^0=\begin{bmatrix}16\\-3\end{bmatrix}.$$

初始化矩阵H_i^0为

$$H_i^0=\begin{bmatrix}0.1&0&0&0&0&0&0.1&0.1\\0&0.1&0&0&0&0&0.1&0.1\\0&0&0.1&0&0&0&0.1&0.1\\0&0&0&0.1&0&0&0.1&0.1\\0&0&0&0&0.1&0&0.1&0.1\\0&0&0&0&0&0.1&0.1&0.1\\0.1&0.1&0.1&0.1&0.1&0.1&1&0\\0.1&0.1&0.1&0.1&0.1&0.1&0&1\end{bmatrix},$$

相应的初始容许控制律为

$$u_1^0=\begin{bmatrix}1.650\\1.650\end{bmatrix},\ u_2^0=\begin{bmatrix}1.550\\1.550\end{bmatrix},\ u_3^0=\begin{bmatrix}3.200\\3.200\end{bmatrix},$$
$$u_4^0=\begin{bmatrix}-0.950\\-0.950\end{bmatrix},\ u_5^0=\begin{bmatrix}-2.350\\-2.350\end{bmatrix}.$$

令折扣因子$\gamma=0.8$, 迭代总次数$\mathcal{R}=600$, 效用函数式(5)的权值矩阵分别为

$$Q_{11}=Q_{22}=Q_{33}=Q_{44}=Q_{55}=I_{2\times 2},$$
$$R_{11}=R_{22}=R_{33}=R_{12}=R_{21}=R_{23}=R_{32}=1,$$
$$R_{31}=R_{13}=0.$$

下面的图2给出了多智能体系统的一致性误差动态曲线, 可以看出在前1000次的迭代过程中, 由于缺少模型信息, 5个智能体之间的一致性误差较大; 在迭代至1000次至2000次时, 随着不断的获取邻居智能体的信息并学习过往积累的数据信息, 智能体之间的一致性误差开始减小. 迭代至2000次之后, 智能体之间的一致性误差逐渐趋于0.

下面的图3给出了关于控制律的动态曲线, 可以看出在前1200次迭代过程中, 由于迭代次数少还无法学

习到较好的控制策略, 智能体的控制律还不理想. 但值得注意的是, 因为选取了初始容许控制律, 所以控制律的更新始终是有界的. 从1500次迭代开始, 根据过往积累的数据信息和邻居智能体的信息, 智能体开始逐渐学习并优化控制律, 但智能体之间的状态误差是非0的, 所以智能体的控制律需要不断更新且不为0.

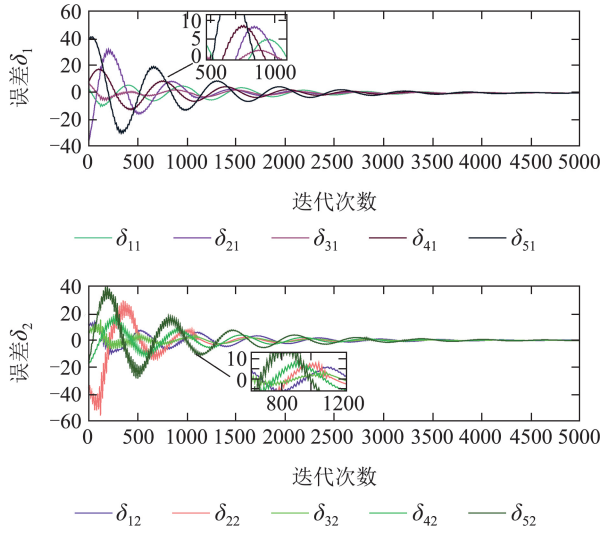


图 2 智能体的一致性误差

Fig. 2 The consensus error of the MASs

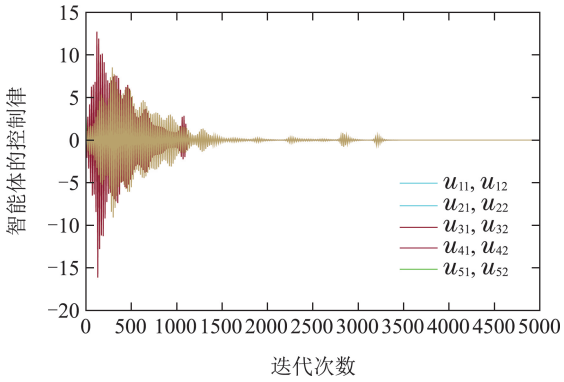
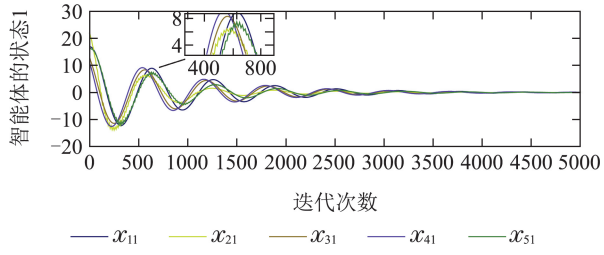


图 3 智能体的控制律

Fig. 3 The controller of the MASs

下面的图4给出了每个智能体状态的动态变化曲线, 可以看出在迭代至0∼1500次时, 由于缺少模型信息, 智能体的状态无法达到一致; 从1500次开始, 经过学习过往积累的数据信息不断更新控制律$u_i(k)$, 每个智能体的状态开始逐渐趋于一致. 图5是智能体状态的三维坐标图.



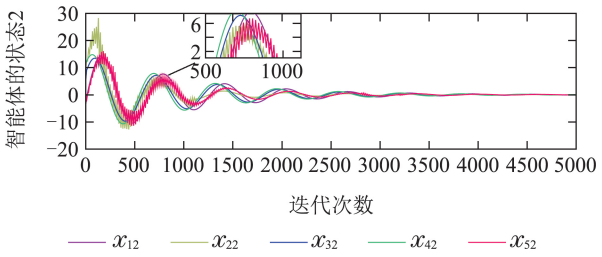


图 4 智能体的状态

Fig. 4 The dynamics of the MASs

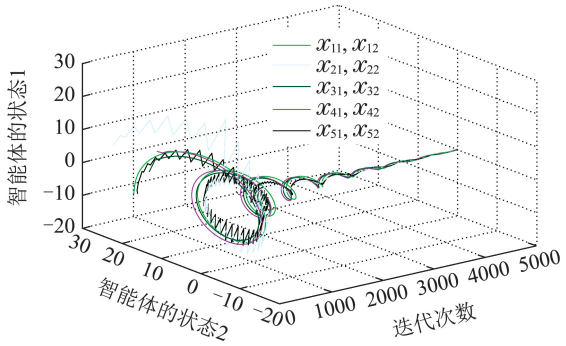


图 5 智能体状态三维坐标图

Fig. 5 3-D phase plane plot of the dynamics of the MASs

5 结论

本文研究在系统模型未知的情况下一类离散时间多智能体系统的一致性问题. 论文提出了一种Q-learning的方法, 这种方法避免了复杂的系统建模与HJB方程求解, 利用系统数据进行策略迭代得到理想控制律, 实现多智能体系统的状态一致. 为了保证系统稳定, 论文给出了所提算法的收敛性和稳定性分析. 最后, 通过MATLAB仿真验证了所提方法的有效性. 未来研究工作可以针对如何利用强化学习方法解决一类模型随机的多智能体系统的一致性问题进行研究.

参考文献:

[1] LAI Yunhui, LI Rui, SHI Yingjing, et al. On the study of a multi-quadrotor formation control with triangular structure based on Graph theory. *Control Theory & Applications*, 2018, 35(10): 1530 – 1537.
(赖云晖, 李瑞, 史莹晶, 等. 基于图论法的四旋翼三角形结构编队控制. 控制理论与应用, 2018, 35(10): 1530 – 1537.)

[2] JIANG Wanyue, WANG Daobo, WANG Yin, et al. A vector field based method for multi-UAV simultaneous arrival. *Control Theory & Applications*, 2018, 35(9): 1215 – 1228.
(蒋婉玥, 王道波, 王寅, 等. 基于时变向量场的多无人机编队集结控制方法. 控制理论与应用, 2018, 35(9): 1215 – 1228.)

[3] PENG Tao, LU Qun, SU Chunyi. Adaptive control of multiple mobile robot formation under slip condition. *Control Theory & Applications*, 2020, 37(2): 439 – 445.
(彭滔, 陆群, 苏春翌. 打滑状态下的多移动机器人编队自适应控制. 控制理论与应用, 2020, 37(2): 439 – 445.)

[4] JIANG Yutao, LIU Zhongxin, CHEN Zengqiang. Distributed finite-time consensus algorithm for multiple nonholonomic mobile robots with disturbances. *Control Theory & Applications*, 2014, 31(4): 531 – 537.
(姜玉涛, 刘忠信, 陈增强. 带扰动的多非完整移动机器人分布式有限时间一致性控制. 控制理论与应用, 2019, 36(5): 737 – 745.)

[5] ZHOU Yuan, HU Hesuan, LIU Yang, et al. Distributed approaches to motion control of multiple robots via discrete event systems. *Control Theory & Applications*, 2018, 35(1): 110 – 120.
(周远, 胡核算, 刘杨, 等. 分布式多机器人运动控制的离散事件系统方法. 控制理论与应用, 2018, 35(1): 110 – 120.)

[6] OLFATISABER R, MURRAY R M. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 2004, 49(9): 1520 – 1533.

[7] XU J, XIE L, LI T, et al. Consensus of multi-agent systems with general linear dynamics via dynamic output feedback control. *IET Control Theory & Applications*, 2013, 7(1): 108 – 115.

[8] LIU X K, WANG Y W, XIAO J W, et al. Distributed hierarchical control design of coupled heterogeneous linear systems under switching networks. *Int. J. Robust. Nonlinear Control*, 2017, 27: 1242 – 1259.

[9] HUANG N, DUAN Z, ZHAO Y, et al. Leader-following consensus of second-order non-linear multi-agent systems with directed intermittent communication. *IET Control Theory & Applications*, 2014, 8(10): 782 – 795.

[10] YI J W, WANG Y W, XIAO J W, et al. Consensus in second-order Markovian jump multi-agent systems via impulsive control using sampled information with heterogenous delays. *Asian Journal of Control*, 2016, 18: 1940 – 1949.

[11] GRONDMAN I, BUSONIU L, LOPES G A, et al. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics*, 2012, 42(6): 1291 – 1307.

[12] STULP F, BUCHLI J, ELLMER A, et al. Model-free reinforcement learning of impedance control in stochastic environments. *IEEE Transactions on Autonomous Mental Development*, 2012, 4(4): 330 – 341.

[13] ZHONG C, LU Z, GURSOY M C, et al. A deep actor-critic reinforcement learning framework for dynamic multichannel access. *IEEE Transactions on Cognitive Communications and Networking*, 2019, 5(4): 1125 – 1139.

[14] HE P, JAGANNATHAN S. Reinforcement learning-based output feedback control of nonlinear systems with input constraints. *IEEE Transactions on Systems, Man, and Cybernetics*, 2005, 35(1): 150 – 154.

[15] MODARES H, LEWIS F L, NAGHIBI-SISTANI MB. Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2013, 24(10): 1513 – 1525.

[16] ABUKHALAF M, LEWIS F L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, 2005, 41(5): 779 – 791.

[17] WEI Q, ZHANG H, LIU D, et al. An optimal control scheme for a class of discrete-time nonlinear systems with time delays using adaptive dynamic programming. *Acta Automatica Sinica*, 2010, 36(1): 121 – 129.

[18] WANG B, ZHAO D, ALIPPI C, et al. Dual Heuristic dynamic programming for nonlinear discrete-time uncertain systems with state delay. *Neurocomputing*, 2014, 134(25): 222 – 229.

[19] KIUMARSI B, LEWIS F L. Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems. *IEEE Transactions on Neural Networks*, 2015, 26(1): 140 – 151.

[20] 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.

[21] 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.

[22] JIANG Y, FAN J, CHAI T, et al. Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout. *IEEE Transactions on Neural Networks*, 2018, 29(10): 4607 – 4620.

[23] AI-TAMIMI A, LEWIS F L, ABU-KHALAF M. Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infifinity control, *Automatica*, 2007, 43(3): 473 – 481.

[24] FU Y, CHAI T, FAN J, et al. Robust adaptive quadratic tracking control of continuous-time linear systems with unknown dynamics. *2015 American Control Conference (ACC)*. Chicago, IL, 2015: 2230 – 2235.

[25] 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.

[26] MU C, ZHAO Q, GAO Z, et al. Q-learning solution for optimal consensus control of discrete-time multiagent systems using reinforcement learning. *Journal of The Franklin Institute-Engineering and Applied Mathematics*, 2019, 356(13): 6946 – 6967.

[27] ABOUHEAF M, LEWIS F L, MAHMOUD M S, et al. Discrete-time dynamic graphical games: Model-free reinforcement learning solution. *Control Theory and Technology*, 2015, 13(1): 55 – 69.

作者简介:

朱志斌　硕士研究生, 目前研究方向为多智能体系统控制, E-mail: 2120190406@nankai.edu.cn;

王付永　讲师, 硕士生导师, 中国人工智能学会智能空天系统专业委员会会员, 主要研究方向为集群智能与协同控制、强化学习与智能博弈、多智能体系统等, E-mail: wangfy@nankai.edu.cn;

尹艳辉　博士研究生, 目前研究方向为分布式系统容错控制, E-mail: yinyanhui2013@163.com;

刘忠信　教授, 博士生导师, 中国人工智能学会智能空天系统专业委员会委员, 中国智能物联网系统建模与仿真专业委员会委员, 主要研究方向为多智能体系统、复杂动态网络、计算机控制与管理, E-mail: lzhx@nankai.edu.cn;

陈增强　教授, 博士生导师, 主要研究方向为智能预测控制、多智能体系统、混沌系统与复杂动态网络, E-mail: chenzq@nankai.edu.cn.