基于强化学习的离散事件系统最优定向监控

胡瑜洪[1],王德光[1*],杨 明[1],王 玺[2]
(1. 贵州大学电气工程学院,贵州贵阳 550025;2. 西安电子科技大学机电工程学院,陕西西安 710071)

摘 要: 对于多个可控事件(控制指令)允许同时执行的情形,离散事件系统的监控器进行随机选择. 然而,在实际应用中,如交通调度、机器人路径规划,可控事件的定向选择和数值优化是必须要考虑和解决的两个问题. 对此,引入一种优化机制量化控制成本,将监督控制理论与强化学习结合,提出一种基于强化学习的离散事件系统最优定向监控器求解方法,使被控系统实现以下三个目标:(1)遵循安全性和活性控制规范;(2)每个状态下至多允许一个可控事件执行;(3)从初始状态到标记状态事件执行累计成本最小. 首先,建立系统和控制规范的自动机模型,做同步积运算后可得到目标模型,通过定义的成本函数为目标模型中每个事件的执行赋予成本. 其次,利用监督控制理论求解无阻塞且行为最大许可的监控器. 最后,将监控器转化为马尔可夫决策过程,并利用Q学习算法求解出最优定向监控器. 使用单向列车导轨控制案例和多轨道列车控制案例验证所提方法的有效性和正确性. 仿真结果表明,所提出方法能够实现系统的无阻塞定向控制,并且使得定向监控器的数值成本最小.

关键词: 离散事件系统;定向监控器;强化学习;最优控制;数值优化;交通系统

基金项目: 国家自然科学基金(No.52265066,No.62203132);贵州省省级科技计划资助项目(No.黔科合基础-ZK[2022]一般103);贵州省教育厅青年科技人才成长项目(No.黔教合KY字[2022]138号);贵州大学科研基金资助项目(No.贵大特岗合字[2021]04号);贵州省教育厅创新群体(No.黔科合支撑[2021]012)

中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112(2024)09-3172-13

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221267

Optimal Directed Control of Discrete Event Systems Based on Reinforcement Learning

HU Yu-hong[1], WANG De-guang[1*], YANG Ming[1], WANG Xi[2]
(1. *The Electrical Engineering College*, *Guizhou University*, *Guiyang*, *Guizhou* 550025, *China*;
2. *School of Eletro-Mechanical Engineering*, *Xidian University*, *Xi' an*, *Shaanxi* 710071, *China*)

Abstract: In the case that several controllable events (control commands) are allowed to execute simultaneously, the supervisor in the framework of discrete event systems (DESs) selects one randomly. However, in practical applications, such as traffic scheduling and robot path planning, the problems of directed control and numerical optimization should be considered. This paper introduces an optimization mechanism to quantify the control cost and combines supervisory control theory (SCT) with reinforcement learning. A systematic procedure is proposed to synthesize the optimal directed supervisor of a DES based on reinforcement learning, which makes the controlled system achieve the following three goals: (1) the control specifications relevant to security and liveness are not violated; (2) at most one controllable event can be executed at each state; (3) the cumulative cost of event execution from the initial state to a mark state is minimal. First, given the automaton models of the plant and specifications, the target automaton model is obtained by the synchronous operation of these two models; a cost function is defined and assigns the execution cost for each event in the target model. Second, the nonblocking and maximally permissive supervisor is synthesized by SCT. Finally, the supervisor is transformed into a Markov decision process and then the Q-learning algorithm is utilized to compute the optimal directed supervisor. Two applications are used to verify the effectiveness and correctness of the proposed method. The simulation results show that the proposed method can realize the directed control of the system, and the numerical cost of the directed supervisor is minimized.

Key words: discrete-event system; directed supervisor; reinforcement learning; optimal control; numerical optimiza-

收稿日期:2022-11-03;修回日期:2023-03-01;责任编辑:梅志强
*通讯作者:王德光

tion; traffic systems

Foundation Item(s): National Natural Science Foundation of China (No.52265066, No.62203132); Guizhou Provincial Science and Technology Projects (No.QianKeHeJiChu-ZK[2022]YiBan 103); Youth Science and Technology Talents Development Project of Guizhou Education Department (No.QianJiaoHe KY [2022]138); Scientific Research Fund of GuiZhou University (No.GuiDaTeGangHeZi[2021] 04); Guizhou Provincial Education Innovation Group Projects (No.QianKeHeZhiCheng[2021]012)

1 引言

离散事件系统监督控制理论最初由 Ramadge 和 Wonham 提出,旨在提供可普遍适用于计算机集成的人造系统(如通信与计算机网络、交通系统、生产制造系统、物流系统等)分析与控制的一般原则和方法[1,2]. 该理论求解的监控器以极小约束的方式禁止可控事件,从而确保系统的安全性(即系统永远都不会进入非法状态)以及活性(即系统无阻塞)[3~5]. 在实际应用中,一个控制器通常需要从多个可同时允许发生的控制指令中至多选择一个发生[6~8],并且通过设定的优化机制[9](如最小控制成本、最大经济效益等)驱动系统达到标记状态. 例如,在运输系统[10]中,监督控制的行为只是为车辆指定一个最大的可行路线集. 然而,更合理的方式是定向控制车辆使其遵循特定路线行驶,并且选择的路线是最优的,如路径最短,耗时最短等.

对于离散事件系统的诸多实际应用,给定系统的模型、行为规范的模型以及优化的目标,存在各种方法求解最优控制器. 当所有事件均是可控时,任意事件序列的选择可通过禁用其他事件来实现,也可通过启发式搜索算法,如 A*算法[11],寻找最优解. 然而,系统中存在不可控的事件时,启发式搜索算法将不再适用. 文献[12]中设计一个天线转子控制系统,控制器执行给定的安全性、活性和实时控制规范,在系统的每个状态下至多选择一个可控事件执行. 然而,这种选择是在特定的基础上进行的,尽管初始监控器是无阻塞的,但这种特定的事件选择方式可能导致求解的最优控制器是阻塞的. 文献[13]提出了一种求解离散事件系统最优定向控制器的方法,然而该方法只适用于无环路情形. 文献[14]解决了有环路和无环路情形下最优定向控制器求解问题,但其求解过程十分复杂,需通过两个算法多次迭代计算. 此外,求解的最优控制器忽略了某些状态下可控事件引起的到达标记状态的路径.

强化学习用于描述和解决智能体在与环境的交互过程中通过学习策略以达到回报最大化或实现特定目标的问题. 强化学习需要训练智能体不停地和环境进行交互,通过试错的方式总结每一步的最佳行为决策,即基于环境反馈训练智能体调整行为决策,其优势是不需要先验知识便可完成自主学习. 强化学习适用于处理动态环境,但缺乏对安全性的保障. 离散事件系统的监督控制理论侧重于系统安全性,但缺乏灵活性,且未考虑数值优化问题. 监督控制理论和强化学习均具有特定的优势,但其理论框架并不相同,强化学习的理论框架中并不考虑事件的可控性;离散事件系统监督控制理论框架下使用的有限自动机模型是完全确定的,即如果知道当前的状态和将要触发的事件,就可以预测下一个状态,而强化学习中的马尔可夫决策过程是由转移概率矩阵引起的给定状态的转移结果具有随机性. 因此使用强化学习算法求解离散事件系统最优定向监控问题时,需要综合考虑事件的可控性问题和马尔可夫决策过程的随机性问题.

综上,针对已有方法存在的不足,本文提出一种基于强化学习的离散事件系统最优定向监控器求解方法,可适用于有环路和无环路情形. 以事件执行累计成本最小为优化目标,同时解决了系统安全性和活性控制、事件选择引发的系统阻塞以及数值最优问题. 在使用 Q 学习算法求解最优定向监控器时,通过奖惩因子以及事件执行成本函数构建奖励函数,对选择进入不可控环路和在同时允许可控事件与不可控事件发生的状态下选择可控事件的动作进行惩罚. 经过多次迭代后,根据 Q 值表选择对应的状态和事件构成最优定向监控器,从而为受控系统提供一条最优的可行路径. 最后,本研究将所提算法应用到单向列车导轨和多轨道列车控制两个案例中,分别实现了最优无阻塞定向控制. 与 A*算法、改进蚁群算法、文献[13,14]中方法得到的结果进行对比分析,验证了本研究所提方法的正确性和有效性.

2 离散事件系统监督控制理论

离散事件系统是一类状态空间离散,且通过事件驱动状态改变的动态系统. 通常,有限自动机用来建立系统的数学模型,形式化语言描述系统的行为. 自动机是一个五元组 $G=(H, \Sigma, \delta, h_0, H_m)$,其中,$H$ 表示有限状态集合;Σ 表示有限事件集合;$\delta: H\times\Sigma\to H$ 表示状态转移函数;$h_0\in H$ 表示初始状态;$H_m\subseteq H$ 表示标记状态集合. 状态转移函数可扩展为 $\delta: H\times\Sigma^*\to H$,其中,$\Sigma^*$ 表示包含空字符串和所有有限长度的事件序列构成的集合. 系统 G 的生成语言和标记语言分别表示为 $L(G)=\{t\in\Sigma^*|\delta(h_0,t)\in H\}$ 和 $L_m(G)=\{t\in L(G)|\delta(h_0,t)\in H_m\}$. 用

$\overline{L_m(G)}$表示标记语言的前缀闭包. 当$L(G)=\overline{L_m(G)}$时，系统G是无阻塞的. 若一个状态可由初始状态经过有限事件序列到达，则称该状态是可达的. 用Ac(G)表示删除系统G中不可达的状态以及对应的变迁后的系统模型.

定义1 将多个独立运行的模块自动机模型合成为一个完整系统模型的操作称为同步积[2].

令 $G_i=(H_i,\Sigma_i,\delta_i,h_{i0},H_{im})$，其中，$i$=1，2. $G_1\|G_2=(H_1\times H_2,\Sigma_1\cup\Sigma_2,\delta,(h_{10},h_{20}),H_{1m}\times H_{2m})$. 令$h_1\in H_1,h_2\in H_2,\sigma\in\Sigma$，

$$\delta((h_1,h_2),\sigma)=\begin{cases}(\delta_1(h_1,\sigma),\delta_2(h_2,\sigma)), & 若\sigma\in\Sigma_1\cap\Sigma_2\\(\delta_1(h_1,\sigma),h_2), & 若\sigma\in\Sigma_1-\Sigma_2\\(h_1,\delta_2(h_2,\sigma)), & 若\sigma\in\Sigma_2-\Sigma_1\end{cases}\tag{1}$$

事件集Σ划分为可控事件集Σ_c和不可控事件集Σ_{uc}，且$\Sigma=\Sigma_c\cup\Sigma_{uc}$. 可控事件一般是指控制器发出的指令，而不可控事件一般是指传感器的反馈信号或扰动信号(如发生故障). 用$\Sigma(h)$表示在状态h下允许发生的事件集合，可划分为可控事件集$\Sigma_c(h)$和不可控事件集$\Sigma_{uc}(h)$.

满足$\Sigma_{uc}\subseteq\gamma\subseteq\Sigma$的事件子集$\gamma$称为控制模式. 若$\sigma\in\gamma$，则事件$\sigma$允许发生；否则，事件$\sigma$禁止发生. 将所有控制模式构成的集合记为$\Gamma$：

$$\Gamma=\{\gamma\in 2^{\Sigma}|\Sigma_{uc}\subseteq\gamma\}\tag{2}$$

监控器$V=(G,\psi)$，其中，$\psi:L(G)\to\Gamma$为监督控制映射. (G,ψ)可写为ψ/G，表示G在ψ的监控下. 离散事件系统的监督控制框图如图1所示.

若$L(\psi/G)=\overline{L_m(\psi/G)}$，则称$V$是无阻塞的. 令$K\subseteq L(G)$，如果满足$\bar{K}\Sigma_{uc}\cap L(G)\subseteq\bar{K}$，称$K$关于$G$和$\Sigma_c$可控. 如果$K=\bar{K}\cap L_m(G)$，称$K$是$L_m(G)$-闭.

定理1 令$K\subseteq L(G)$，存在一个无阻塞的监控器V使得$L_m(\psi/G)=K$，当且仅当K关于G和Σ_c可控以及K是$L_m(G)$-闭[15].

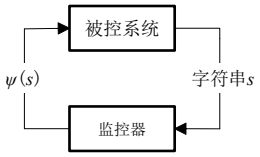


图1 离散事件系统监督控制框图

在本次研究中，使用TCT软件[16]进行自动机模型搭建和监控器求解. 求解监控器的步骤如下：

(1) 建立系统各个模块的自动机模型G_i，i=1，…，N.

(2) 建立控制规范自动机模型E_j，j=1，…，M.

(3) 分别对所有G_i和E_j做同步积运算，可得系统模型$G=G_1\|G_2\|\cdots\|G_N$和控制规范模型$E=E_1\|E_2\|\cdots\|E_M$.

(4) 对G和E做同步积运算，可得目标自动机模型$T=G\|E$.

(5) 求解无阻塞且行为最大许可的监控器V=SupC(G, T).

通过SupC(G, T)可以得到目标语言$L(T)$在控制规范E中关于G和$\sum_c$上的最大可控子语言$L(V)$，且$L(V)$是$L_m(G)$-闭.

3 基于强化学习的最优定向监控

本节定义了成本函数，引入定向监控器、最优定向监控器的概念，介绍强化学习模型和基于值函数的强化学习算法，提出一种将监控器转化为马尔科夫决策过程的方法，给出了基于强化学习算法求解最优定向监控器的步骤，并证明了算法求解的定向监控器是最优的且无阻塞的. 本研究中最优定向监控器的求解过程框图如图2所示.

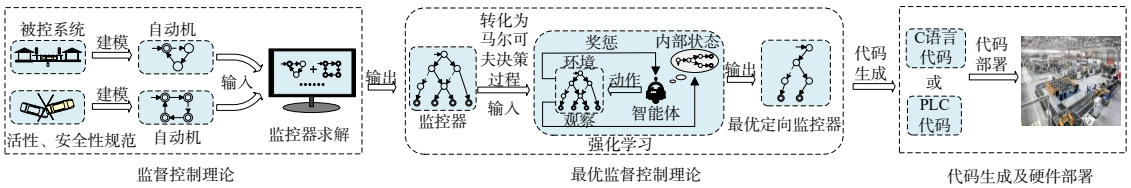


图2 最优定向监控器的求解框图

3.1 成本函数

在离散事件系统中，状态的转移是通过事件驱动. 在本研究中，事件的执行成本由事件的发生成本和控制成本构成. 事件的发生成本具有实际的物理意义. 例如，在柔性汽车制造系统中，事件的发生成本可以表示机械的耗电、磨损、维修等成本；在交通系统调度的应用中，事件的发生成本可以表示为距离、时间等因素. 事件的控制成本为禁止该状态下其他可控事件的成本.

系统模型G和控制规范模型E做同步积运算可得到目标模型$T=G\|E=(Z,\Sigma,\tau,z_0,Z_m)$. 定义成本函数$c$：$Z\times\Sigma\to\mathbb{R}^+$，其中，$\mathbb{R}^+$表示正实数集，满足

$$c(x,\sigma)=\begin{cases}0, & 若\sigma\notin\Sigma(x)\\R_\sigma, & 若\sigma\in\Sigma(x)\end{cases}\tag{3}$$

其中，$c(x,\sigma)$表示在状态x下执行事件σ的成本，它是事件σ的发生成本和禁止状态x下其他允许发生的可控

事件的成本之和. 成本函数可以扩展为 $c: Z\times\Sigma^*\to\mathbb{R}^+$. 令 $x\in Z, s\in\Sigma^*$, $c(x, s)$ 表示在状态 x 下执行事件序列 s 的成本.

对于路径 $z=x_0\xrightarrow{\sigma_1}x_1\xrightarrow{\sigma_2}x_2\cdots x_{n-1}\xrightarrow{\sigma_n}x_n$. 用 $\mathcal{C}(z)$ 表示该路径的成本,其计算公式如下:

$$\mathcal{C}(z)=\sum_{i=0}^{n-1}c(x_i,\sigma_{i+1}) \tag{4}$$

3.2 定向监控器

监控器在某些状态下可能同时允许多个可控事件发生,而定向监控器至多允许每个状态下有一个可控事件发生.

定义 2 令监控器 $V=(X, \Sigma, \varphi, x_0, X_m)$,定向监控定义为 $D: X\to 2^{\Sigma_c}$,满足:

$$\forall x\in X: |D(x)|\leqslant 1 \tag{5}$$

定义 2 表示在定向监控的作用下,状态 x 处至多允许一个可控事件发生. 定向监控器 $V^D=(V, D)$,其并没有针对不可控事件定义控制决策,所以不可控事件一直允许发生. 因此,定向监控器在状态 x 时,允许发生的事件集合为 $D(x)\cup\Sigma_{uc}(x)$. 定向监控器的生成语言和标记语言分别用 $L(V^D)$, $L_m(V^D)$ 表示.

例 1 考虑如图 3 所示监控器 V,其中,状态用圆圈表示,初始状态用未带标签的箭头指向表示,标记状态用双圆圈表示,字母表示事件,括号内的数字表示事件的执行成本,实线表示可控事件,虚线表示不可控事件. 监控器 V 的标记语言为 $L_m(V)=\{\text{hd}, \text{hk}, \text{bf}, \text{bg}\}$. 由图 3 可知,监控器 V 有四个定向监控器 V^{D_1}, V^{D_2}, V^{D_3}, V^{D_4}. 其中, V^{D_1}, V^{D_2} 在初始状态选择事件 h,而 V^{D_3}, V^{D_4} 选择事件 b. 在事件 h 发生后, V^{D_1} 在状态 1 下允许可控事件 d 发生,而 V^{D_2} 禁止可控事件 d 的发生. 同理, V^{D_3} 在发生事件 b 后允许可控事件 f 发生,而 V^{D_4} 禁止可控事件 f 发生. 监控器 V 的四个定向监控器的标记语言分别为 $L_m(V^{D1})=\{\text{hd}, \text{hk}\}$, $L_m(V^{D2})=\{\text{hk}\}$, $L_m(V^{D3})=\{\text{bf}, \text{bg}\}$, $L_m(V^{D4})=\{\text{bg}\}$.

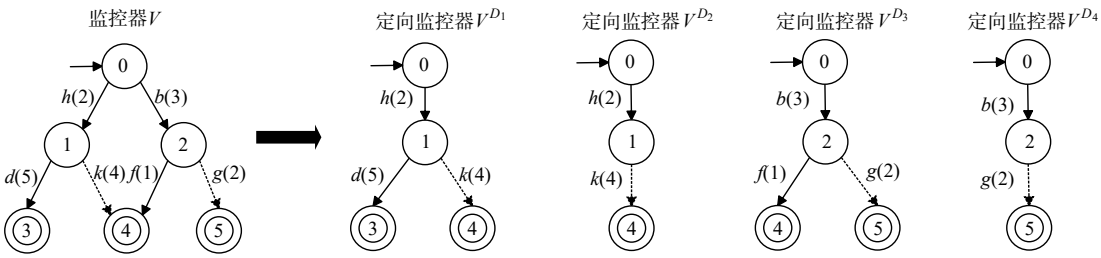


图 3　例 1 中的监控器和定向监控器模型

3.3 最优定向监控器

一个监控器可能具有多个定向监控器,如何从中选择到达标记状态的事件执行累计成本最低的定向监控器,是本研究要解决的问题.

令监控器 $V=(X, \Sigma, \varphi, x_0, X_m)$,状态 $x\in X$. 从状态 x 到达标记状态的路径集合记为 $\Omega_V^z(x)$. 用 $|\Omega_V^z(x)|$ 表示路径集合中路径的个数. 用 $d_V(x)$ 表示从状态 x 经过路径 $z\in\Omega_V^z(x)$ 到达标记状态的最大成本,即

$$d_V(x)=\begin{cases}\max\{\mathcal{C}(z)|z\in\Omega_V^z(x)\}, & \text{若} 0<|\Omega_V^z(x)|<\infty\\ 0, & \text{若 } \Omega_V^z(x)=\varnothing \text{且} x\in X_m\end{cases} \tag{6}$$

当 $\Omega_V^z(x)\neq\varnothing$ 时,

$$\begin{aligned}d_V(x)&=\max\{d_V(x')+c(x,\sigma)|x'=\varphi(x,\sigma)\in X\}\\&=\max\{d_1, d_2\}\end{aligned} \tag{7}$$

其中,

$$d_1=\max\{d_V(x')+c(x,\sigma)|x'=\varphi(x, \sigma)\in X, x'\notin X_m\} \tag{8}$$

$$d_2=\max\{0+c(x,\sigma)|x'=\varphi(x, \sigma)\in X, x'\in X_m\} \tag{9}$$

综上,定向监控器 $V^D=(X^D, \Sigma^D, \varphi^D, x_0, X_m^D)$ 的成本为:

$$P(V^D)=\max_{x\in X^D}d_{V^D}(x) \tag{10}$$

所以,监控器 V 的最优定向监控器为 V^{D^*},

$$V^{D^*}\in\arg\left\{\min_{V^D}P(V^D)\right\}=\arg\left\{\min_{V^D}\max_{x\in X^D}d_{V^D}(x)\right\} \tag{11}$$

例 2 在例 1 中,由式(9)可得定向监控器 V^{D_1}, V^{D_2}, V^{D_3} 和 V^{D_4} 的成本分别为 $P(V^{D_1})=\max d_{V^{D_1}}(x_0)=c(x_0, \text{ad})=7$, $P(V^{D_2})=\max d_{V^{D_2}}(x_0)=c(x_0, \text{ae})=6$, $P(V^{D_3})=\max d_{V^{D_3}}(x_0)=c(x_0, \text{bg})=5$, $P(V^{D_4})=\max d_{V^{D_4}}(x_0)=c(x_0, \text{bg})=5$. 由式(10)可知,监控器 V 的最优定向监控器为 V^{D_3} 和 V^{D_4}.

定义 3 令状态 $x, x'\in X$, 事件序列 $e\in\Sigma^*$. 若 $\varphi(x, e)=x'$,则称状态 x 为状态 x' 的关联状态.

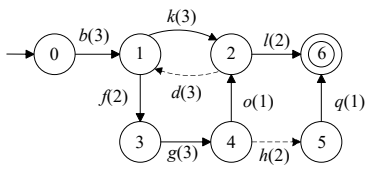
定义 4 令状态 $x, x'\in X$. 若 x 为 x' 的关联状态,且存在事件 $\sigma\in\Sigma_{uc}$ 使得 $\varphi(x', \sigma)=x$,则称状态 x, x' 构成不可控回路.

3.4 强化学习求解最优定向监控器

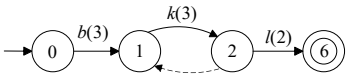
对于求解最优定向监控器问题,可以将其视为在有向图中寻找最优路径问题. 在每个状态下至多有 $|\Sigma|$ 个事件,因此,无环路系统中,从初始状态到标记状态

的路径数量至多为$|\Sigma|^{|X|}$. 若系统中存在环路,则路径数量可以为无穷. 随着系统组件的增加,监控器的状态数量呈指数增长,在求解最优路径时,其搜索空间也会变大. 在求解最优定向监控器时,目前用于寻找最优路径的算法不考虑事件可控性,故最终计算结果可能为次最优. 在图4所示的监控器中,从初始状态到标记状态事件执行累计成本最小的事件序列为s=bkl,其成本为8. 若在状态2下允许不可控事件d发生,由于不可控回路的存在,在最坏情况下由状态0、1、2、6构成的定向监控器为成本为∞. 因此,在求解最优定向监控器时,不仅需要考虑每个状态下不可控事件引起的到达标记状态的每一条路径成本,同时需要考虑监控器中存在的环路情况.

由例2可知,在求解无环路监控器的最优定向监控器时,从某一状态到标记状态的成本总是由该状态下的不可控事件决定的,故在计算定向监控器的成本时可忽略该状态下可控事件引起的变迁. 然而,特定的事件选择方式可能导致求解的最优控制器是阻塞的,如对于有环路的监控器,忽略可控事件引起的变迁可能会导致计算最优定向监控器时出现阻塞. 在图4中,状态2下可控事件l和不可控事件d同时允许发生,在计算定向监控器成本时,若忽略可控事件l,则会造成阻塞. 本研究提出一种基于强化学习的最优无阻塞定向监控器计算方法,在使用Q学习算法求解最优定向监控器时,在奖励函数中加入奖惩因子,对选择进入不可控环路和在同时允许可控事件与不可控事件发生的状态下选择可控事件的动作进行惩罚,对到达标记状态的动作进行奖励. 经过多次迭代运算后,根据Q值表选择对应的状态和事件构成最优定向监控器. 本研究方法可同时适用于有环路和无环路的情形.



(a) 监控器V



(b) 次最优定向监控器V^{D}

图4 有环路监控器及其次最优定向监控器

强化学习主要用于解决序贯决策问题[17],强化学习问题可用有限马尔可夫决策过程(Markov Decision Process, MDP)来表示. 在具有马尔可夫性质的系统状态环境中,它能够模拟智能体实现随机策略与回报奖励的描述[18]. 为了使用强化学习算法求解最优定向监控器,必须先将监控器转换为MDP. 确定性有限自动机是完全确定的,即如果知道当前的状态和将要触发的事件,就可以预测下一个状态. 然而,MDP是由转移概率矩阵引起的给定状态的结果具有随机性. 本研究提出将确定性有限自动机转换为确定性MDP的方法. 首先,将随机性MDP转换为确定性MDP. 其次,将自动机转换为等效的确定性MDP. 最后,定义奖励函数.

如果MDP中每个动作可唯一确定转移过程的下一个状态,则称它是确定性(或非随机性)MDP[19]. 确定性MDP是一个五元组(S,A,T,R,γ),其中,S为状态空间;A为动作空间;$T:S\times A\to S$为确定性转移函数;$R:S\times A\to\mathbb{R}$为奖励函数,$\mathbb{R}$为实数集;$\gamma\in[0,1]$为折扣因子,表示对未来奖励的考虑情况,γ越大表示越注重未来动作的奖励. 对于自动机$G=(H,\Sigma,\delta,h_0,H_m)$,其被转换成等效的确定性MDP的过程如下. 令$S\equiv H;A\equiv\Sigma;T\equiv\delta$;因为$\Sigma$分为可控事件集合和不可控事件集合,故动作集合$A$划分为可控动作集合$A_{\mathrm{c}}\equiv\Sigma_{\mathrm{c}}$和不可控动作集合$A_{\mathrm{uc}}\equiv\Sigma_{\mathrm{uc}}$,且$A=A_{\mathrm{c}}\cup A_{\mathrm{uc}}$. 本研究的目标为求解事件执行累计成本最低的定向监控器,故智能体执行动作后获得的奖励应与执行事件成本有关,奖励函数定义如下:

$$R(s,a)=-c(s,a)+p \tag{12}$$

其中,$c(s,a)$为成本函数,表示监控器V中状态$s\in X$执行事件$a\in\Sigma$的成本,$p\in\mathbb{R}$为奖惩因子. 当p为负数时,表示对动作的惩罚,用于以下两种情况:(1)对进入不可控回路动作的惩罚;(2)在$A_{\mathrm{c}}(s)\neq\varnothing$且$A_{\mathrm{uc}}(s)\neq\varnothing$时,选择可控动作的惩罚. 当$p$为正数时,表示对到达标记状态的奖励.

强化学习以累计奖励最大化为优化目标,常用状态-动作对值函数衡量累计奖励:

$$Q^{\pi}(s,a)=E\left(\sum_{k=0}^{\infty}\gamma^{k}r_{t+1+k}\middle|s_t=s,a_t=a\right) \tag{13}$$

其中,γ为折扣因子,r为执行动作后获得的奖励. $Q^{\pi}(s,a)$为状态-动作值函数,表示t时刻,智能体在环境状态$s_t=s$,执行动作$a_t=a$时累计奖励的数学期望. 求解强化学习问题实际上就是寻求最优策略π^{*},使得智能体与环境交互获得的累积奖励最大化,即

$$\pi^{*}=\arg\max_{a}Q^{\pi}(s,a) \tag{14}$$

Q学习是经典的强化学习算法[20,21],其通过定义Q函数,即状态-动作值函数,评估策略的优劣. 在每个时间步最大化动作值函数$Q(s,a)$选择最优动作. Q学习以MDP为基础,在执行完动作后,通过值迭代方法更新Q函数,其公式为:

$$Q(s,a)=Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')-Q(s,a)] \tag{15}$$

其中,$Q(s,a)$表示状态-动作值函数,α表示学习率,r表

示执行动作a获得的奖励,γ为折扣因子. 持续迭代之后,Q函数收敛进而获得最优策略π^*. Q学习一般通过ε贪婪策略选择动作,强制智能体搜索Q值表中所有的状态-动作对. ε贪婪策略如算法1所示.

算法1　ε贪婪策略

输入:Q值表,状态s,动作集合A,ε.

输出:状态s下的动作a.

1. 创建随机数$n \in [0, 1]$.
2. 若$n > \varepsilon$,则执行步骤3;否则,执行步骤4.
3. 在$Q(s)$中选择Q值最高的动作a,执行步骤5.
4. 在$Q(s)$中随机选择一个动作a,执行步骤5.
5. 返回动作a.

基于Q学习算法求解最优定向监控器的算法如算法2所示.

算法2　Q学习算法求解最优定向监控器

输入:由监控器转换得到的确定性MDP,标记状态集合L_s-X_m,学习率α,折扣因子γ.

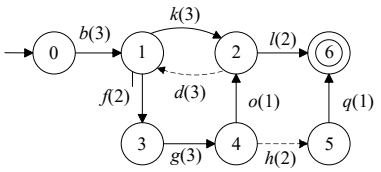
输出:最优定向监控器,且成本为costValue.

初始化:将Q值表初始化为0,设置最大训练次数N,迭代次数$n = 0$,costValue = 0,状态集合D^*=[].

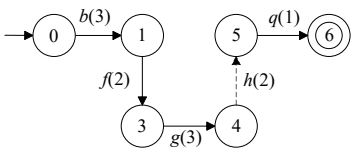
1. 若$n < N$,则循环步骤2 ~ 8;否则,执行步骤9.
2. 初始化状态$s = x_0$.
3. 若$s \in L_s$,则循环步骤4 ~ 7;否则执行步骤8.
4. 根据算法1选择状态s下的一个动作a.
5. 执行动作a,观察奖励r和下一个状态s'.
6. 通过式(15)更新$Q(s, a)$.
7. 令$s = s'$,返回步骤3.
8. $n = n+1$,返回步骤1.
9. 从初始状态s开始,若该状态下没有不可控动作,则选择$Q(s, a)$中Q值最大的动作,costValue = costValue + $c(s, a)$;若该状态下有不可控动作,则选择所有不可控动作并标记Q值最小的不可控动作a,costValue = costValue + $c(s, a)$,并将该状态存入到D^*中;根据选择的动作到达下一个状态s',进行同样的选择,直到$s' \in L_s$,执行步骤10.
10. 根据Q值表和步骤9的动作选择方式,选择D^*中每个状态下的可控动作到达标记状态的状态转移路径,根据式(5)计算出该路径的成本$d_v(x_0)$,若$d_v(x_0) \leqslant$ costValue,则将该路径加入到最优定向监控器中,整合步骤9和步骤10中的状态和动作,可得到监控器V中的最优定向监控器V^{D*},且事件执行累计的成本为costValue.

在监控器的每个状态下,至多有$|\Sigma_c|+1$个动作选择. 因此,从监控器中可找到的定向监控器数量不超过$(|\Sigma_c|+1)^{|X|}$. 采用枚举的方式求解最优定向监控器的计算复杂度是关于监控器状态数量的指数关系. 在算法2中,步骤1的执行次数为N. 为了到达标记状态,在最坏的情况下,需要遍历所有状态,即步骤3需要执行$|S\times A|$次. 步骤4需要遍历状态s下的所有动作,至多执行$|A|$次. 步骤9需遍历所有状态以及该状态下的动作,故要执行$|S|\times|A|$次. 步骤10遍历列表D^*和D^*中状态下的事件,最坏情况下$|D^*|=|S|$,执行$|S|\times|A|$次. 其中,$|S|=|X|$和$|A|=|\Sigma|$分别为监控器的状态和事件数量. 算法2的计算复杂度为$O([|S\times A|\times(|A|+3)+2]\times N+2\times|S|\times|A|)$. 因为$|S\times A|\leqslant|S|\times|A|$,故$[|S\times A|\times(|A|+3)+2]\times N+2\times|S|\times|A|\leqslant(N\times|A|^2+3\times N\times|A|+2\times|A|)\times|S|+2\times N$,且$N >> |A|$,所以可渐近为$O(\vartheta\times N\times|S|)$. 综上,算法2的计算复杂度与监控器状态个数呈多项式关系.

例3　考虑图5(a)所示监控器V,状态0和状态1可分别通过事件序列bk和k到达状态2,故状态0和1为状态2的关联状态. 状态2经过不可控事件d到达状态1,故状态1和状态2构成不可控回路. 同理,状态1、2、3和4也构成不可控回路. 对选择事件k和事件o的动作进行惩罚. 选择事件l和q的动作将到达标记状态,对选择这两个事件的动作进行奖励. 状态4允许可控事件o和不可控事件h发生,在状态4下对选择可控事件o的动作进行惩罚. 将惩罚设置为−50,奖励设置为100. 通过算法2可以得到图5(a)的最优定向监控器,如图5(b)所示,其成本为11. 在图5(a)中,路径$0 \xrightarrow{b(3)} 1 \xrightarrow{k(3)} 2 \xrightarrow{l(2)} 6$的成本为8,但在状态2下存在不可控事件到达状态1,由于监控器并不能禁止不可控事件的发生,故状态1和状态2的循环次数为未知数. 考虑到最坏情况,将具有不可控环路的监控器成本设置为无穷大. 在状态4下,可控事件o到达标记状态的转移路径的成本大于该状态下不可控事件的转移路径成本,故禁止该事件的发生.



(a) 监控器V



(b) 最优定向监控器V^D

图5　例3中监控器合最优定向监控器模型

定理2　算法2求解的最优定向监控器是最优的.

证明　对于监控器V,假设在状态x下,发生不可控事件w和任意事件序列e_1后到达标记状态的定向监控器为V_1^D,其执行事件序列we_1的累计成本为P_1;在该状态下,通过可控事件w'和任意事件序列e_2到达标记状态的定向监控器为V_2^D,其执行事件序列$w'e_2$的累计成本为P_2.

当 $P_1>P_2$ 时，定向监控器 V^D 的成本为 $P(V^D)=\max\limits_{x\in X^D} d_{V^D}(x)=P_1$，最优定向监控器为

$$V^{D^*}\in \arg\left\{\min_{V^D} P(V^D)\right\}=\arg\left\{\min_{V^D} P_1\right\}=V_1^D \quad (16)$$

当 $P_1<P_2$ 时，考虑以下两种情况：可控事件允许发生时，定向监控器 V^D 的成本为 $P(V^D)=\max\limits_{x\in X^D} d_{V^D}(x)=P_2$；可控事件禁止发生时，定向监控器 V^D 的成本为 $P(V^D)=\max\limits_{x\in X^D} d_{V^D}(x)=P_1$. 此时，最优定向监控器为

$$V^{D^*}\in \arg\left\{\min_{V^D} P(V^D)\right\}=\arg\left\{\min_{V^D} (P_1,P_2)\right\}=V_1^D \quad (17)$$

当 $P_1=P_2$ 时，最优定向监控器 $V^{D*}\in\arg\{V_1^D,V_2^D\}$.

综上，在求解最优定向监控器时，在每个状态下到达标记状态的成本是由不可控事件引起的状态转移路径决定的.

使用Q学习算法求解最优定向监控器：当 $P_1>P_2$，在对选择可控事件的动作进行惩罚时，为了使智能体在训练过程中认为选择可控事件得到的定向监控器为次最优，应存在惩罚 p，使得

$$P_1<P_2+|p| \quad (18)$$

在训练完成后，可得到收敛的 Q 值表. Q 值表间接表示了当前状态到达标记状态的成本大小，即 Q 值越大，成本越小. 对于任意状态 x，若 $\Sigma_c(x)\neq\varnothing$，$\Sigma_{uc}(x)=\varnothing$，则最优定向监控器为 $V^{D^*}\in\arg\{\min\limits_{V^D} P(V^D)\}$，即应选择 Q 值最大的动作. 若 $\Sigma_c(x)=\varnothing$，$\Sigma_{uc}(x)\neq\varnothing$，当 $|\Sigma_{uc}(x)|>1$ 时，在最坏情况下，最优定向监控器的成本为不可控事件到达标记状态的路径中成本最大的路径，即 $\max\limits_{x\in X^D} d_{V^D}(x)$，所以应在 Q 值表中选择不可控动作中 Q 值最小的. 若 $\Sigma_c(x)\neq\varnothing$，$\Sigma_{uc}(x)\neq\varnothing$，当可控事件引起的转移路径成本大于不可控事件引起的转移路径中的最大成本时，根据式(17)可知，禁止该可控事件后，得到的监控器为最优定向监控器.

综上，算法2求解的定向监控器为最优的.

证毕.

4 案例仿真与结果分析

在本节中，适用单向列车导轨案例和多轨道列车控制案例验证所提方法的有效性和正确性，并与 A^* 算法、改进蚁群算法、文献[13, 14]中的方法进行对比分析.

4.1 系统建模

图6为单向列车导轨案例示意图，在站台A和站台B之间有一个单向轨道，该轨道由区域1、2、3和4组成. 在站台A和轨道区域1、轨道区域1和轨道区域2、轨道区域3和轨道区域4之间分别通过交通信号灯控制列车的驶入和驶出. 此外，在轨道的五个不同位置分别安装信号探测器，用来检测列车是否到达该位置. 假设列车C1和C2同时使用该轨道，并且初始位置均在站台A. 为防止碰撞事故发生，安全规范为确保列车C1和C2不会同时处于同一个轨道区域.

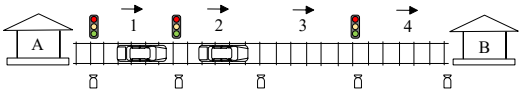


图6 单向列车导轨案例示意图

列车C1和C2的自动机模型如图7所示，其中，初始状态0表示站台A，标记状态5表示站台B，状态1~4分别表示轨道区域1~4，奇数表示可控事件，偶数表示不可控事件. 每当有一个事件发生，其意义表示对应的列车从当前轨道区域行驶到下一轨道区域或站台B内.

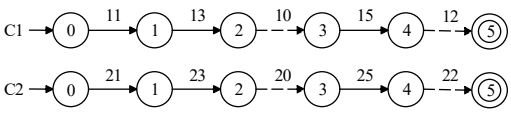


图7 C1与C2的自动机模型

对C1和C2做同步积运算后，得到单向列车导轨系统模型 C = C1||C2，如图8所示. 有序对 (m, n) 表示C中的状态，其中 m、n 分别表示C1、C2当前所处的状态. 自动机模型C中状态(1, 1)、(2, 2)、(3, 3)和(4, 4)表明C1、C2位于同一轨道区域. 此时，列车C1和C2会发生碰撞. 因此，它们是不满足系统安全性规范的状态.

4.2 仿真结果与分析

根据监督控制理论，可以求得单向列车导轨系统的监控器，如图9所示. 监控器提供的事件序列能够避免系统进入阻塞和违背安全性和活性控制规范.

对于图9的监控器，将惩罚设置为−50，奖励为100，学习率为0.8，折扣因子为0.9，ε 为0.2，迭代次数为200次. 通过算法2可求得单向列车导轨系统的最优定向监控器，如图10所示. 在训练过程中，求解的定向监控器的成本如图11(a)所示，其中，横坐标为迭代次数，纵坐标为每次迭代的最优定向监控器成本. 在训练过程中，Q 值表收敛过程如图11(b)所示. 奖励和惩罚的取值大小会对动作的选择产生影响，图12所示为设置不同奖励和惩罚下得到的最优定向监控器成本. 图12(a)表明当惩罚值满足式(18)时，求解的定向监控器为最优，并且惩罚值越大，算法收敛速度越快. 图12(b)所示为在惩罚值为10时，不同奖励值对最优定向监控器的求解影响. 对选择到达标记状态的动作进行奖励，是避免在求解的过程中陷入局部循环. 从图中可知，设置

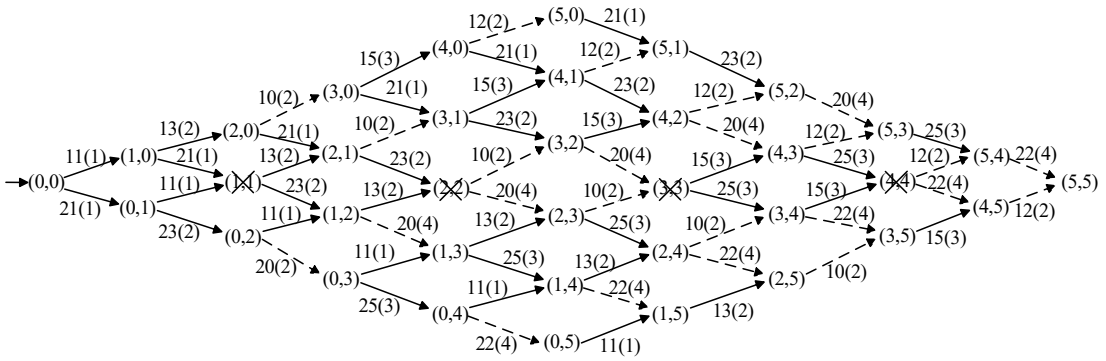


图 8　单向列车导轨案例自动机模型

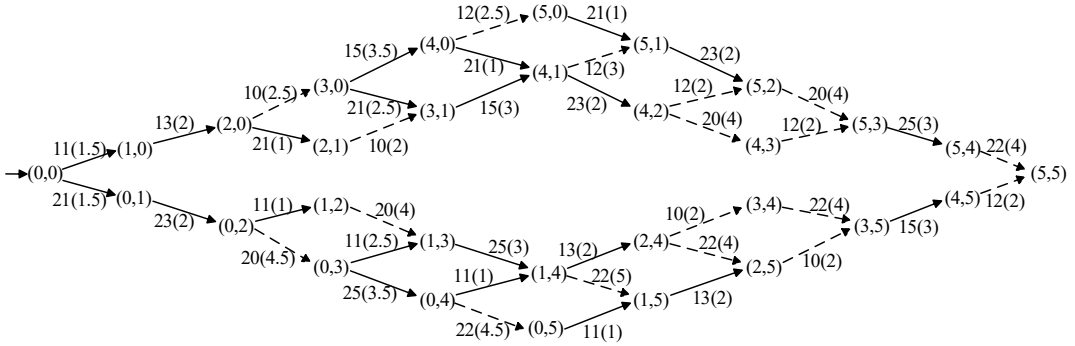


图 9　单向列车导轨案例监控器模型

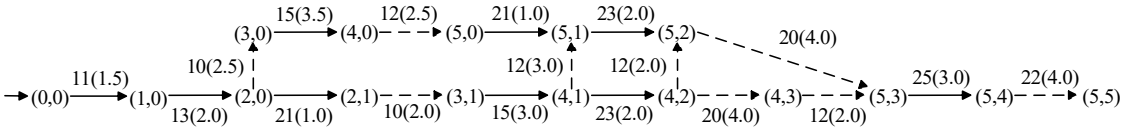
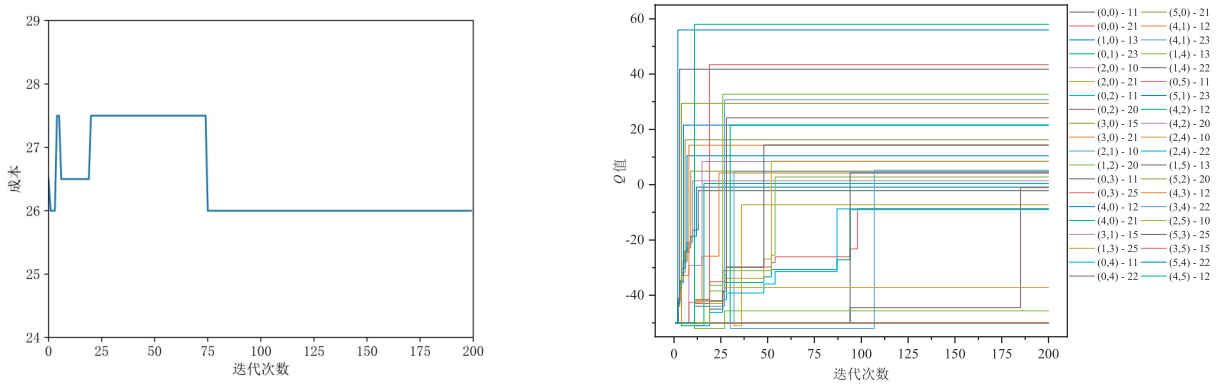


图 10　本研究方法所求解的最优定向监控器

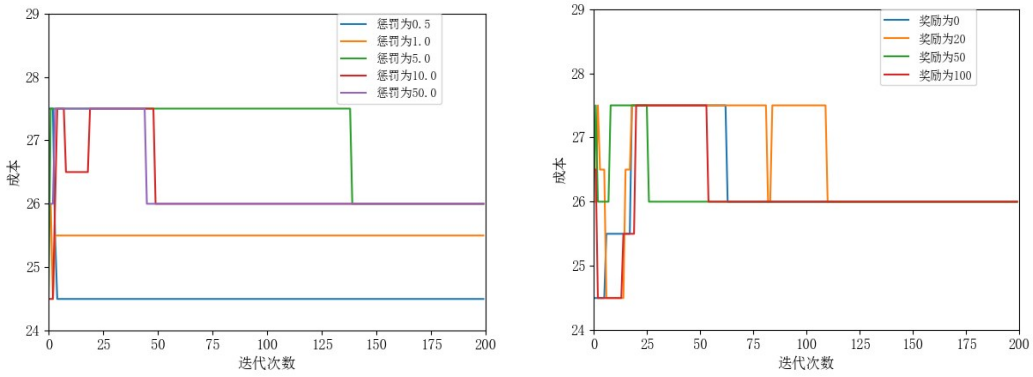


(*a*) 定向监控器成本　　(*b*) 监控器中动作的 Q 值收敛过程

图 11　单向导轨案例训练结果

不同的奖励均能求得最优定向监控器,且奖励的大小不影响算法收敛速度.

目前,存在很多成熟的搜索最优路径(最优解)算法,比如,A*算法和Dijkstra算法等图搜索类算法[22]、遗传算法[23]和蚁群算法[24,25]等智能优化算法以及强化学习[26~28]等机器学习算法. 在图8中,上述算法求解最优定向监控器时,可将其视为求解有向图的最短路径,将不符合安全性规范的状态视为障碍,从而求解出包含所有符合安全性规范状态的路径. 然而,由于上述算法不考虑事件的可控性,所求解的定向监控器可能不是

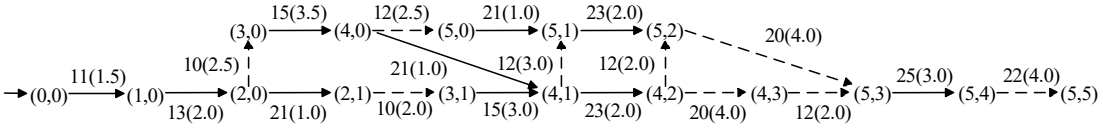


(*a*) 不同惩罚值对最优定向监控器求解的影响 (*b*) 惩罚为10时,不同奖励值对最优定向监控器求解的影响

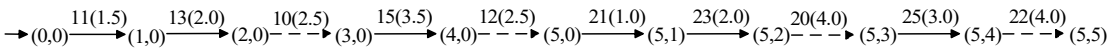
图 12 不同奖惩值对最优定向监控器求解的影响

最优的. 特别地,文献[26]结合监督控制理论和强化学习算法求解最大化柔性汽车制造系统回报问题,其奖励函数由系统中事件发生的成本映射. 例如,表示产品合格的事件返回高的正利润,而表示产品不合格的事件返回高的负利润,与机器动作相关的事件返回较低的负利润. 由于文献[26]未考虑同时允许可控事件和不可控事件发生的状态下不可控事件到达标记状态的路径成本大于可控事件到达标记状态的路径成本的情况,导致求解的定向监控器可能为次最优.

利用A*算法、改进蚁群算法和文献[26]中强化学习方法求解的最优控制器如图13(*a*)所示,其成本为26.5. 如图10所示,本研究在求解过程中禁止了状态(4,0)下的可控事件21,使得所求解的定向监控器为最优. 在图13(*a*)中,状态(4,0)到标记状态(5,5)的最短路径为(4,0)→(4,1)→(4,2)→(4,3)→(5,3)→(5,4)→(5,5),其路径成本为16. 然而,监控器在状态(4,0)、(4,1)下不能禁止不可控事件12发生,所以状态(4,0)构成的定向监控器的成本由路径(4,0)→(4,1)→(4,2)→(4,3)→(5,3)→(5,4)→(5,5)决定,其成本为17. 在将(4,0)下的可控事件21禁止后,状态(4,0)构成的定向监控器的成本由路径(4,0)→(5,0)→(5,1)→(5,2)→(5,3)→(5,4)→(5,5)决定,其成本为16.5,为最优定向监控器. 此外,所求解的路径可能会导致系统进入到不符合安全性规范的状态,如图8中,状态(3,2)处存在不可控事件使得系统进入违背安全性规范的状态(3,3),将导致系统违背安全性规范.



(*a*) 采用A*算法、改进蚁群算法和文献[26]求解的最优定向监控器



(*b*) 文献[13,14]求解的最优定向监控器

图 13 相关文献求解的最优定向监控器结果

在求解最优定向监控器时,由于不能禁止不可控事件,在某些状态下具有多个不可控事件到达标记状态的转移路径,需要对每一条路径都进行成本判断. 文献[13]基于动态规划思想提出求解无环路系统的最优定向控制器的方法,该方法通过从标记状态开始向前迭代计算求解最优定向监控器. 在计算过程中,对于同时具有可控事件和不可控事件(非扰动事件)的状态,忽略了该状态下的可控事件. 如图9所示,状态(4,1)到标记状态共有3条路径,其中由两条路径是由可控事件23引起的,由于状态(4,1)同时允许可控事件和不可控事件发生,在考虑最坏情况下,状态(4,1)到标记状态的成本由不可控事件12到标记状态的路径决定,文献[13]中忽略可控事件23到标记状态的路径,从而计算出状态(4,1)到标记状态的成本. 文献[14]提出了求解有环路和无环路系统的最优定向控制方法. 然而,其求解过程十分复

杂,通过两个算法进行迭代计算,并且其求解的最优控制器同样忽略了某些状态下的可控事件引起的到达标记状态的路径. 文献[13,14]中方法求解的最优定向监控器如图13(*b*)所示,其通过忽略具有不可控事件使能的状态下的可控事件到达标记状态的路径,减少搜索空间. 然而,所忽略的状态转移路径可能具有更小的路径成本. 图12(*a*)中,在状态(2,0)和状态(4,1)处分别选择可控事件21和23时,路径(0,0)→(1,0)→(2,0)→(2,1)→(3,1)→(4,1)→(4,2)→(4,3)→(5,3)→(5,4)→(5,5)的成本为24.5,小于图13(*b*)的路径成本. 因此,在实际应用中,最优定向监控器中应该保留该路径. 为此,本研究对文献[13,14]的定向映射条件进行改变. 文献[13,14]在以下条件下允许可控事件发生:条件(1),若一个状态下只有可控事件可以发生,此时保留该状态到标记状态成本最小的事件;条件(2),若一个状态下的不可控事件为扰动事件(故障等),保留该状态到标记状态下成本最小的可控事件,若该不可控事件不为扰动事件,则禁止所有可控事件. 在本研究中,保持条件(1)不变,将条件(2)改为若该状态允许不可控事件发生,则判断该状态下所有可控事件到标记状态的成本是否小于该状态下不可控事件到标记状态的成本中的最小值. 若结果为是,则允许到标记状态成本最小的可控事件发生,禁止其他可控事件. 此外,文献[13,14]是以整个系统作为输入,通过删除系统自动机模型中违背安全性、活性控制规范的状态和变迁来满足系统的控制规范,并使用trim()操作[2]保证系统是无阻塞的. 然而,trim()操作后并不能保证系统是可控的,并且当系统规模较大且控制规范为自动机模型时,无法直接从系统自动机模型中删除违背控制规范的状态和变迁. 为解决这一问题,本研究通过求解出系统的监控器模型后,将其作为解的搜索空间,使得所求解的最优定向监控器满足控制规范.

为进一步体现本研究方法的适用性,增加了图14所示的多轨道列车控制案例. 两辆列车C3和C4分别位于轨道区域1和轨道区域2,列车C3可在轨道区域1~8之间行驶,列车C4可在轨道区域2,4~8之间行驶. 通过5个交通信号灯分别控制列车驶入驶出轨道. 为防止碰撞的事故发生,多轨道列车控制案例的安全规范为确保列车C3和C4不会同时处于同一个轨道上.

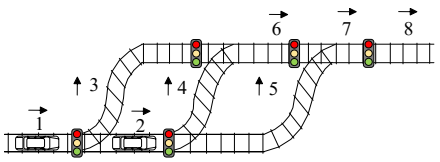


图14　多轨道列车控制案例示意图

在表1中,$|Z|$、$|\Sigma|$分别表示目标模型的状态和事件数量;y表示蚂蚁数量,T表示迭代次数;$|Z_m-Z_t|$表示非终结的标记状态个数;ϑ为常数,N为迭代次数,E为事件具有的步骤数,n为步长. 由表1可知,相较于A*算法和改进蚁群算法,本研究采用的方法保证了所求解的定向监控器是最优的. 与文献[13]中所采用的动态规划方法相比,本研究所提方法不仅能够求解出无环路情形下的最优定向监控器,同时也适用于求解有环路的情形. 与文献[14]中的方法比较,研究中的方法同时保证定向监控器为最优,并且保留了由可控事件引起的到达标记状态事件执行累计成本最低的状态转移路径.

在表2中,$|X|$、$|\varphi|$分别表示监控器中状态数量和状态变迁数量;$|X^D|$、$|\varphi^D|$分别表示计算得到的定向监控器V^D中状态数量和状态变迁数量;$P(V^D)$表示V^D的成本;$c(z_{\text{best}})$为最优定向监控器中从初始状态到标记状态的最小路径成本.

表1　相关方法对比

方法	适用范围	是否最优	算法复杂度	是否为定向控制	是否满足安全规范
A*算法	有环路和无环路情形	否	$O(\|Z\|\times(\log\|Z\|^{\|Z\|}+\|\Sigma\|))$	是	否
改进蚁群算法	有环路和无环路情形	否	$O((y\times T\times\|Z\|^2)/2)$	是	否
文献[13]中的方法	无环路情形	是	$O(\|Z\|\times\|\Sigma\|)$	是	是
文献[14]中的方法	有环路和无环路情形	是	$O(\|Z\|\times\|\Sigma\|\times(\|Z_m-Z_t\|+1))$	是	是
文献[1]中的方法	有环路和无环路情形	否	$O(\|Z\|\times\|\Sigma\|)$	否	是
文献[26]中的方法	有环路和无环路情形	否	$O(T\times\|\Sigma\|)/O(T\times\|E\|\times n)$	是	是
本研究中的方法	有环路和无环路情形	是	$O(\vartheta\times N\times\|Z\|)$	是	是

表2 相关方法应用到两个案例的仿真结果

案例	$\|X\|$	$\|\varphi\|$	方法	$\|X^D\|$	$\|\varphi^D\|$	$P(V^D)$	$c(z_{\text{best}})$
单向列车导轨控制案例	30	40	A*算法	16	19	26.5	24.5
			改进蚁群算法	16	19	26.5	24.5
			文献[13]中的方法	11	10	26.0	26.0
			文献[14]中的方法	11	10	26.0	26.0
			文献[26]中的方法	16	19	26.5	24.5
			本研究所提的方法	16	18	26.0	24.5
多轨道列车控制案例	17	23	A*算法	7	6	14.0	14.0
			改进蚁群算法	7	6	14.0	14.0
			文献[13]中的方法	7	6	14.0	14.0
			文献[14]中的方法	7	6	14.0	14.0
			文献[26]中的方法	7	6	14.0	14.0
			本研究所提的方法	7	6	14.0	14.0

5 结论

针对经典监督控制理论无法实现定向控制和处理数值优化问题,本研究引入一种最优机制量化控制成本,提出将自动机转换为马尔科夫决策过程的策略,并利用强化学习中的Q学习算法求解最优定向监控器,可适用于有环路和无环路情形. 本研究的方法计算所得最优定向监控器不仅可使受控系统满足安全性和活性控制规范,同时实现了定向控制和累计数值成本最低. 证明了所提出的算法得到的结果是最优的.

在实际问题中,由于传感器的数量不足或量程有限,导致部分事件无法被观测,未来的工作中将致力于提出部分可观离散事件系统的最优定向控制方法.

参考文献

[1] RAMADGE P J G, WONHAM W M. The control of discrete event systems[J]. Proceedings of the IEEE, 1989, 77(1): 81-98.

[2] CAI K, WONHAM W M. Supervisory control of discrete-event systems[M]//Encyclopedia of Systems and Control. Cham: Springer, 2021: 2245-2253.

[3] 辛宗洋. 基于RW框架的离散事件系统监督控制理论计算和仿真平台[J]. 工业控制计算机, 2009, 22(12): 39-41, 43.
XIN Z Y. Computation and simulation platform for supervisory control of discrete event system based on RW structure[J]. Industrial Control Computer, 2009, 22(12): 39-41, 43. (in Chinese)

[4] 焦亭, 刘振国. 组件多工作模式下的离散事件系统监督控制方法研究[J]. 控制理论与应用, 2020, 37(3): 534-539.
JIAO T, LIU Z G. Study on the supervisory control of discrete-event systems incorporating components with several working modes[J]. Control Theory & Applications, 2020, 37(3): 534-539. (in Chinese)

[5] 史晶锌, 舒少龙, 林峰, 等. 基于监督控制理论的家庭住宅用电安全性控制研究[J]. 现代建筑电气, 2014, 5(1): 9-15.
SHI J X, SHU S L, LIN F, et al. Control for safety of home electric usage based on supervisory control theory[J]. Modern Architecture Electric, 2014, 5(1): 9-15. (in Chinese)

[6] 荣胜波, 朱军, 史勃, 等. 监督控制理论在多任务机床控制系统中的应用[J]. 机床与液压, 2011, 39(22): 85-87.
RONG S B, ZHU J, SHI B, et al. The application of supervisory control theory to control system of multiple-task machine tool[J]. Machine Tool & Hydraulics, 2011, 39(22): 85-87. (in Chinese)

[7] TATSUMOTO Y, SHIRAISHI M, CAI K, et al. Application of online supervisory control of discrete-event systems to multi-robot warehouse automation[J]. Control Engineering Practice, 2018, 81: 97-104.

[8] GONZALEZ A G C, ALVES M V S, VIANA G S, et al. Supervisory control-based navigation architecture: A new framework for autonomous robots in industry 4.0 environments[J]. IEEE Transactions on Industrial Informatics, 2018, 14(4): 1732-1743.

[9] UMEMOTO H, YAMASAKI T. Optimal LLP supervisor for discrete event systems based on reinforcement learning [C]//2015 IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE, 2015: 545-550.

[10] KAYMAKCI O, ANIK V G, USTOGLU I. A local modular supervisory controller for a real railway station[C]// 5th IET International Conference on System Safety 2010. Manchester: IET, 2010: 1-6.

[11] PASSINO K M, ANTSAKLIS P J. On the optimal control of discrete event systems[C]//Proceedings of the 28th IEEE Conference on Decision and Control. Piscataway: IEEE, 1989: 2713-2718.

[12] BARBEAU M, FRAPPIER M, KABANZA F. A supervisory control synthesis case study: The antenna control system[C]//Allerton Conference on Communication, Control, and Computing. Nederlanden: Elsevier, 1997: 533-542.

[13] HUANG J, KUMAR R. An optimal directed control framework for discrete event systems[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2007, 37(5): 780-791.

[14] HUANG J, KUMAR R. Optimal nonblocking directed control of discrete event systems[C]//2007 American Control Conference. Piscataway: IEEE, 2007: 4285-4290.

[15] 王飞跃, 陈俊龙. 智能控制方法与应用[M]. 北京: 中国科学技术出版社, 2020: 65-80.
WANG F Y, CHEN J L. Intelligent Control Method and Application[M]. Beijing: China Science and Technology Press, 2020: 65-80. (in Chinese)

[16] 王玺. 实时系统的基于优先级的实时重构和不基于优先级的条件剥夺调度算法[D]. 西安: 西安电子科技大学, 2016.
WANG X. Priority-based Reconfiguration and Priority-free Conditionally-preemptive Scheduling of Real-time Systems[D].Xi'an: Xidian University, 2016. (in Chinese)

[17] 张涛, 张文涛, 代凌, 等. 基于序贯博弈多智能体强化学习的综合模块化航空电子系统重构方法[J]. 电子学报, 2022, 50(4): 954-966.
ZHANG T, ZHANG W T, DAI L, et al. Integrated modular avionics system reconstruction method based on dequential game multi-agent teinforcement learning [J]. Acta Electronica Sinica, 2022, 50(4): 954-966. (in Chinese)

[18] 袁都佳. 基于Q学习的钢琴指法自动生成方法研究[D]. 广州: 华南理工大学, 2020.
YUAN D J. Research on Automatic Generation Method of Piano Fingering Based on Q-learning[D]. Guangzhou: South China University of Technology, 2020. (in Chinese)

[19] MADANI O, THORUP M, ZWICK U. Discounted deterministic Markov decision processes and discounted all-pairs shortest paths[J]. ACM Transactions on Algorithms, 2010, 6(2):1-25.

[20] SHUAI B, LI Y F, ZHOU Q, et al. Supervisory control of the hybrid off-highway vehicle for fuel economy improvement using predictive double Q-learning with backup models[J]. Journal of Central South University, 2022, 29 (7): 2266-2278.

[21] 张德干, 葛辉, 刘晓欢, 等. 一种基于Q-Learning策略的自适应移动物联网路由新算法[J]. 电子学报, 2018, 46 (10): 2325-2332.
ZHANG D G, GE H, LIU X H, et al. A kind of new routing algorithm with adaptivity for mobile IOT based on Q-Learning [J]. Acta Electronica Sinica, 2018, 46(10): 2325-2332. (in Chinese)

[22] 俞凯. 基于图割/图搜索的医学影像分割算法研究及应用[D]. 苏州: 苏州大学, 2020.
YU K. Study of Graph Cut/Search Based Medical Image Segmentation Algorithms and Applications[D]. Suzhou: Soochow University, 2020. (in Chinese)

[23] 徐超群, 易忠, 孟立飞, 等. 基于遗传算法的卫星磁模型研究[J]. 电子学报, 2020, 48(6): 1108-1112.
XU C Q, YI Z, MENG L F, et al. Research on satellite magnetic model based on genetic algorithm[J]. Acta Electronica Sinica, 2020, 48(6): 1108-1112. (in Chinese)

[24] 廖伟志, 夏小云, 贾小军. 基于蚁群算法的多路径覆盖测试数据生成[J]. 电子学报, 2020, 48(7): 1330-1342.
LIAO W Z, XIA X Y, JIA X J. Test data generation for multiple paths coverage based on ant colony algorithm[J]. Acta Electronica Sinica, 2020, 48(7): 1330-1342. (in Chinese)

[25] 许凯波, 鲁海燕, 黄洋, 等. 基于双层蚁群算法和动态环境的机器人路径规划方法[J]. 电子学报, 2019, 47(10): 2166-2176.
XU K B, LU H Y, HUANG Y, et al. Robot path planning based on double-layer ant colony optimization algorithm and dynamic environment [J]. Acta Electronica Sinica,

2019, 47(10): 2166-2176. (in Chinese)
[26] ZIELINSKI K M C, HENDGES L V, FLORINDO J B, et al. Flexible control of discrete event systems using environment simulation and reinforcement learning[J]. Applied Soft Computing, 2021, 111: 107714.
[27] SHI H R, LIU G J, ZHANG K W, et al. MARL Sim2real transfer: Merging physical reality with digital virtuality in metaverse[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 53(4): 2107-2117.
[28] ZHOU Z Y, LIU G J. RoMFAC: A robust mean-field actor-critic reinforcement learning against adversarial perturbations on states[EB/OL]. (2022-05-15) [2022-08-02]. http://arxiv.org/abs/2205.07229.

作者简介

胡瑜洪 男,1999 年 4 月出生于广东省揭阳市. 现为贵州大学电子信息硕士研究生. 主要研究方向为监督控制理论、机器人路径规划、调度等.
E-mail: Yhong@163.com

王德光 男,1991 年 6 月出生于山西省侯马市. 2019 年毕业于西安电子科技大学机电工程学院. 现为贵州大学电气工程学院讲师(特岗教授). 从事监督控制理论、离散事件系统故障诊断、复杂系统建模与分析等方面的研究工作.
E-mail: dgwang@gzu.edu.cn

杨 明 男,1990 年 12 月出生于贵州省铜仁市. 2020 年于北京化工大学获工学博士学位. 现为贵州大学电气工程学院讲师(特岗教授). 主要研究方向为振动测试与计量、机器视觉检测、激光测量等.
E-mail: myang23@gzu.edu.cn

王 玺 男,1986 年 1 月出生于河北省邯郸市. 现为西安电子科技大学机电工程学院讲师. 主要研究方向为离散事件系统监督控制理论、实时系统调度及重构的理论与应用研究.
E-mail: wangxi@xidian.edu.cn