中国科学: 技术科学 2023 年 第 53 卷 第 2 期: 258 ~ 267
SCIENTIA SINICA Technologica techcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
论 文 群体智能激发汇聚及应用专辑

面向群体共识机制的逆强化学习辨识方法

于鑫[1], 吴文峻[2,3*], 罗杰[1,3], 李未[1,3]

1. 北京航空航天大学计算机学院, 北京 100191;
2. 北京航空航天大学人工智能研究院, 北京 100191;
3. 软件开发环境国家重点实验室, 北京 100191
* E-mail: wwj09315@buaa.edu.cn

收稿日期: 2021-08-13; 接受日期: 2021-11-26; 网络版发表日期: 2022-08-23
科技创新2030-"新一代人工智能"重大项目(编号: 2018AAA0102300)资助

摘要 作为新一代人工智能的重要研究领域, 群体智能是解决开放不确定环境中大规模复杂问题的必由途径, 对人工智能的其他研究领域有着基础性和支撑性的作用. 群体智能系统中, 智能体遵循共识机制进行交互演化产生群体共识, 辨识共识机制是构建和理解群体智能系统的关键. 传统的共识机制建模方法需要做过多简化假设, 难以面对复杂多样的群体智能系统, 应建立数据驱动的共识机制辨识方法. 本文将共识机制的辨识问题转化为群体智能系统的逆强化学习问题, 提出面向群体共识机制的逆强化学习辨识方法, 并将上述辨识方法应用于集群系统, 在多个场景中验证了对群体智能系统的辨识能力, 实现了对群体智能系统的共识机制的反演.

关键词 系统主义, 群体智能, 逆强化学习

1 引言

群体智能最早源于对自然界中蚂蚁、蜜蜂等社会性昆虫群体行为的研究, 这些昆虫群体有一定的结构与组织, 能够通过简单规则涌现出群体性的智慧, 同时具有一定的学习能力来适应环境的变化[1]. 其他类型的生物也有类似的群体智能行为, 例如鱼群集体游动以减少阻力, 大型食草动物集聚在一起躲避天敌, 甚至连细菌都具备一定的集体决策能力. 在人类社会中大规模复杂群体行为, 如: 开源社区的软件创新、基于众包众享的共享经济、各类市场中的群体商业金融博弈等, 都是通过社群化的组织结构来管理、协调和运行, 以竞争、合作、对抗等多种自主协同方式来共同完成挑战性任务, 涌现出超越个体能力的群体智能[2,3]. 群体智能系统的本质是动态认知复杂网络, 涌现强弱决定网络演化的复杂程度. 自然界和人类社会中的群体智能虽然各具形态, 其蕴含的核心概念却是相同的, 即复杂认知网络的群体性、涌现性、共识性、演化性. 共识性是群体智能系统最重要的性质, 指智能体在局部交互中, 按照一定规则形成全局共识, 驱动群智涌现.

智能体在局部交互中所遵循的规则即为共识机制. 研究表明, 遵循局部共识机制的个体能够在系统的全局层面产生复杂的群体行为, 辨识个体共识机制是

引用格式: 于鑫, 吴文峻, 罗杰, 等. 面向群体共识机制的逆强化学习辨识方法. 中国科学: 技术科学, 2023, 53: 258–267
Yu X, Wu W J, Luo J, et al. Identification method for collective consensus mechanism based on inverse reinforcement learning (in Chinese). Sci Sin Tech, 2023, 53: 258–267, doi: 10.1360/SST-2021-0370

构建和理解群体智能系统的关键[4]. 现有对共识机制的辨识方法主要有两种, 包括基于代理的建模方法(agent based model, ABM)和基于机器学习的建模方法. ABM被广泛应用于对复杂群体行为的研究, 经典的群体运动模型有Vicsek模型、Cuker-Smale模型等. Vicsek模型将所有个体的共识机制设计为与邻居对齐运动方向[5]. 作为Vicsek模型的扩展, Cucker-Smale模型指定个体通过计算其邻居速度大小的加权平均值来确定其下一时刻速度大小[6]. 然而, 不同的群体系统具有不同的行为模式, 很难为所有群体运动构建统一模型[7]. 此外, 构建上述模型需要大量领域知识以及对模型参数的精心调整, 增加了模型的构建难度. 为了应对这些挑战, 机器学习被用来以数据驱动的方式对群体行为进行建模. 基于深度神经网络预测生物群体的未来运动方向取得了比以前基于代理的模型更高的准确度[8]. 尽管这些模型在行为预测方面表现出很高的准确性, 但它们并没有揭示具体的共识机制以及产生共识机制的内在激励.

共识机制的辨识本质上是对智能体策略函数的辨识, 而对智能体策略函数的辨识已经发展出了大量方法和范式. 行为克隆(behavior cloning, BC)以监督学习来学习策略, 会造成累积误差和分布漂移的问题[9]. 逆强化学习方法包括学徒学习方法以及最大熵逆强化学习方法等, 最大熵逆强化学习解决了最优策略不唯一的问题[10,11]. 相比于行为克隆, 逆强化学习能够考虑策略的长期回报, 而不像行为克隆只考虑状态到动作的单步映射, 具有更好的鲁棒性及泛化能力. 生成对抗模仿学习(generative adversarial imitation learning, GAIL)和生成对抗逆强化学习(adversarial inverse reinforcement learning, AIRL)利用生成对抗网络和最大熵逆强化学习之间的联系, 提出了基于生成对抗架构的模仿学习, 将逆强化学习和深度神经网络结合起来, 拓展了模仿学习的应用场景. 围绕多智能体场景的逆强化学习研究较少, 大多数工作假设奖励函数由人工特征的线性组合构成, 可扩展性差且无法应用于高维任务[12,13]. Song等人[14]提出了多智能体生成对抗模仿学习(multi-agent GAIL, MA-GAIL), 是GAIL在多智能体场景的扩展, 旨在辨识专家行为策略, 适用于一般马尔可夫博弈. 多智能体逆强化学习(multiagent AIRL, MA-AIRL)是AIRL在多智能体场景的扩展, 能够从最优判别器中计算奖励函数[15]. 然而MA-GAIL和MA-AIRL没有考虑群体系统的特点, 不适合用来辨识大规模同构群体系统.

针对群体智能的复杂系统性质, 需要从其非线性、随机和动态的特征, 遵循系统主义的精准智能研究范式, 构建面向群体共识机制的辨识方法[16]. 本文主要围绕群体智能系统的建模分析问题, 研究基于逆强化学习的系统辨识方法. 具体贡献如下: 分析了群体智能系统的基本性质, 并在此基础上提出面向群体共识机制的逆强化学习辨识方法; 将对共识规则的辨识问题转化为对智能体策略函数和奖励函数的建模问题, 提出了面向集群系统的群体逆强化学习算法; 以群体运动为例, 将上述群体智能理论的模型和方法应用于集群系统, 通过实验验证了该方法对群体智能系统的辨识能力, 实现了对群体智能系统共识机制的反演.

2　群体智能的共识机制

群体智能的共识机制定义了智能体之间的交互模式, 使得智能体通过分布式认知网络完成信息交互, 推动网络中所有智能体的局部状态达成共识, 从而涌现群体智能. 本节首先分析了群体智能的两大类主要的共识形态, 随后介绍了经典的群体共识建模分析方法.

2.1　群体共识机制分类

自然界和人类社会存在两大类型的群体智能共识机制, 一类是以蜂群、蚁群为代表的生物群体智能, 另一类是人类高级群体智能. 表1给出两类群体智能共识机制的比较.

蜜蜂和蚂蚁这类低级生物虽然感知和认知能力很弱, 不具备记忆能力、自我意识以及对同伴的相互感知意识, 更没有自主的任务分配和相互协同的能力, 但是蜂群和蚁群在完成觅食、筑巢这些活动时, 从整体上看是以一种有组织、有协调的方式在运转的. 这类活动的本质是间接式的协同机制, 通过环境和智能体之间的交互而实现. 其基本原则是, 智能体的动作会在环境中留下轨迹(或者信息), 信息量的累积和汇聚将影响其他智能体的后续行动. 所以智能体与环境交互的结果实际产生了强化的效果, 使得群体的行为逐渐涌现出整体性、趋同化的动作模式[17]. 研究生物群体智能系统共识机制, 对解决人工集群协同决策问题具有重要借鉴意义[18].

表 1 生物群体智能与人类群体智能的比较

Table 1 Comparison of collective intelligence between human and animal

群智类型	智能体性质	共识机制
蜂群、蚁群等生物群智	受限智能体: 低级智能体简单行为模式, 只具备较简单的决策适应能力	服从性共识: 个体受周边的生物激素、同伴运动的激励, 被动地模仿和趋同
人类群智	自由意志智能体: 高级智能体, 复杂行为模式, 具备自主决策能力	自主性共识: 根据内在的需要, 结合外在激励, 自主选择策略, 通过信息交互形成共识

2.2 群体共识机制建模

传统的群体行为研究聚焦某一类动物如何完成群体任务, 例如, 鱼群如何在有天敌追逐的情况下集体调整移动方向, 蚁群如何群体地觅食等. 研究结果最终都归结为通过针对特定任务的共识算法, 来解释局部的交互如何形成全局共识, 产生有利于动物群体生存和发展的结果. 在生物科学和物理科学领域中, 对自然中的蚁群、蜂群、鸟群等集群运动进行了大量的理论和实验研究, 提出了针对不同生物集群运动的共识规则和运动方程. 本节以鱼群运动为例, 加以说明.

(1) 鱼群的集群共识规则

鱼群通常包括三种主要的行为模式: 避碰性、同步性、内聚性[19]. 其共识机制可由若干条简单的个体行为规则来刻画, 这些规则规范了单条鱼在根据自身的形状和运动方式特点的前提下, 所采取的运动决策模式. 在这些针对个体的局部规则的集合作用下, 会产生鱼群的集群智能行为.

排斥规则: 实现避免碰撞, 过于靠近的个体需要对运动速度和方向加以控制, 防止互相碰撞的发生.

对齐规则: 实现与邻居个体的运动同步, 这个规则也被称为“互效行为”, 每个个体都试图调节自己运动的速度和方向, 以匹配其邻居的运动矢量.

吸引规则: 邻居选择, 根据运动过程中的注意力选择机制, 来动态选择自己的邻居.

上述鱼群的三类规则中对齐规则发挥着群体共识形成的核心作用, 通常以Vicsek运动模型来刻画和分析.

(2) 生物群体Vicsek运动模型

Vicsek模型最初用于研究大量群体构成的群体系统行为, 借助该模型能够便捷而有效地仿真大规模群体运动的同步行为[5]. 在Vicsek模型中, 每个个体可全方位感知到所有位于自身感知范围内邻居个体, 每个个体的运动方向由其邻居运动角度的矢量平均来更新, 且更新方向过程中受到噪声干扰.

所有个体都具有相同的速率v_0, 个体i在t时刻的运动方向为$\theta_i(t)$, 则其速度矢量为$\mathbf{v}_i(t)=\left[v_0\cos\theta_i(t), v_0\sin\theta_i(t)\right]^{\mathrm{T}}$. 个体$i$在$t+1$时刻的位置按照式(1)进行更新:

$$\mathbf{x}_i(t+1)=\mathbf{x}_i(t)+\mathbf{v}_i(t). \tag{1}$$

每个个体的运动方向按照下式进行更新:

$$\theta_i(t+1)=\langle\theta_i(t)\rangle_{\Gamma}+\xi_i(t), \tag{2}$$

其中$\langle\theta_i(t)\rangle_{\Gamma}$表示个体$i$包含自身在内的所有邻居个体的平均运动方向, 可由下式进行计算:

$$\langle\theta_i(t)\rangle_{\Gamma}=\arctan\frac{\sum\limits_{j\in\Gamma_i(t)}\sin\theta_j(t)}{\sum\limits_{j\in\Gamma_i(t)}\cos\theta_j(t)}. \tag{3}$$

Vicsek模型虽然简单, 却可以进行大规模群体运动仿真, 是研究群体动力学的有力工具. 理论研究结果表明: Vicsek这类分布式共识算法通过耦合交互网络和控制方程, 可以建模为多输入多输出系统[20]. 分布式共识算法通过局部智能体之间的交互, 会确定收敛到一个集体的决策状态, 而智能体之间的网络特征决定分布式共识算法的稳定性和收敛性[21].

传统的共识机制设计方法的局限在于它需要手工设计控制规则和面向具体问题的特征, 这就需要大量的领域知识和工作量, 很难扩展到通用的领域. 要解决这一问题, 需要采用基于深度神经网络的建模方法, 直接从行为数据中学习模型, 而不再依靠单纯的手工特征和启发式规则.

3 群体共识机制辨识

智能群体通过不断交互信息, 形成复杂认知网络, 实施行为演化和智能涌现, 适应动态变化的环境. 通过

将群体智能系统建模为马尔可夫博弈，能够将对共识机制的辨识转化为反演智能体的策略函数和奖励函数, 并进一步基于逆强化学习算法进行求解. 本节首先介绍马尔可夫博弈框架，随后阐述系统辨识与逆强化学习的关系，最后提出面向群体共识机制的逆强化学习辨识方法.

3.1　随机马尔可夫博弈框架

对于上文所述群体智能系统，可以采用部分可观察随机马尔可夫博弈框架(partially observable stochastic game, POSG)来统一描述[22]. 对于合作场景, 通常使用分布式部分可观察马尔可夫决策过程(decentralized partially observable Markov decision process, Dec-POMDP)表述. 而POSG相比于Dec-POMDP是更通用的博弈框架，适用于更广泛的场景，能够表述更多样化的群体智能系统.

部分可观察随机马尔可夫博弈框架:

POSG: N个智能体的马尔可夫博弈定义为元组 $N, S, \{A_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \{O_i\}_{i=1}^n, T, \xi$
N: 智能体数量
S: 状态空间
A_i: 智能体i的动作空间
$A = A_1 \times A_2 \times \cdots \times A_n$ 联合动作空间
$R_i: S \times A \to R$ 智能体i的奖励函数
$T: S \times A \times S \to [0,1]$ 状态转移函数
O_i智能体i的观测空间
$O = O_1 \times O_2 \times \cdots \times O_n$联合观测空间
$\xi: S^N \to O^N$ 系统的观测函数
$\pi_i: S \times A_i \to [0,1]$ 智能体i的策略函数

马尔可夫博弈中每个智能体试图最大化自身的累积折扣奖励$E\left\{\sum_{k=0}^{\infty}\gamma^k r_{i,t+k}\right\}$，其中$r_{i,t+k}$是$k$步之后智能体所收集到的累积奖励, γ是折扣因子. 智能体i的策略记作 $\pi_i: S \times A_i \to [0,1]$. π_i为智能体i的策略函数，基于智能体状态和观测进行决策，将智能体的观测映射到智能体动作. T为系统状态转移函数, 本质上等同于生物群体智能的动力学方程, 但是由于现实系统的复杂性, 通常无法直接得到函数的解析表示，可以通过内嵌系统的深度动力学模型来刻画. ξ为观测函数，由于每个智能体的观察视野往往局限在其邻域范围，并对其决策产生直接影响，所以需要通过观测函数来刻画. R_i为智能体i的奖励函数，通过奖励函数能够定义智能体所要完成的任务. 在POSG框架下，协同合作的群智系统与竞争博弈群智系统的主要区别在于奖励函数的定义.

POSG能够充分表征群体智能系统, 其中系统的观测函数刻画了共识性产生的物理边界及智能体的信息感知能力, 奖励函数刻画了系统共识性的目标倾向, 策略函数刻画了个体的共识决策规则，状态转移函数刻画了群体智能系统的状态变化.

3.2　系统辨识与逆强化学习

群体智能本质上是具有复杂系统属性的智能形态. 系统要素之间的耦合关联构成了群体智能系统的结构与组织，系统行为难以通过单一自治系统特征的简单叠加来刻画体现了群体智能的涌现性，而复杂系统在不断适应组织和环境变化过程中所需的逐步成长体现了群体智能系统的学习能力.

群智系统的复杂性对于分析和构造智能提出了挑战. 系统辨识是利用统计学, 从测量到的数据出发构建动力系统数学模型的方法[23]. 系统辨识首先需要确定对系统模型的刻画方法. 传统的理论建模主要依赖物理原理，实现对系统模型的建立. 但这种方式往往需要做很多对系统的简化假设，才能简化模型方程，实现求解和分析的目的. 所以对于复杂的非线性动力系统，这种完全基于方程的模型具有相当的领域局限性, 无法适用于普遍的情况，并且简化后的方程与实际系统的机理和行为存在一定的偏差. 数据驱动的系统辨识无需对系统模型的方程做过多假设，能够采用神经网络这类黑盒模型，根据系统行为轨迹的测量数据, 找到符合系统行为规律的最优黑盒模型. 数据驱动的辨识方式去除了对系统的简化假设，能够逼近系统的行为特征, 可以应用于各种不同的复杂系统. 当前, 随着算法的发展和算力的提升，数据驱动的系统辨识方法被广泛应用.

系统辨识提供了理解群体智能系统的理论框架, 而非线性动力学模型中的参数估计仍是一个非常具有挑战性的问题. 逆强化学习能够从已有的自然群体智

能的状态轨迹中，反演群智系统的奖励函数和最优策略，揭示其内在的共识机理，完成对群体共识机制的建模. 逆强化学习是基于给定的专家示例数据推断未知的奖励函数和专家策略的过程，其基本思想是寻找奖励函数R，使专家示例数据在使用R进行评估时所能获得的奖励大于等于其他策略所获得的奖励[24]. 逆强化学习的一次完整学习过程可以表示如式(4)所示:

$$\begin{aligned}&\mathrm{RL}\odot\mathrm{IRL}(\pi_E):\\&\max_{\pi}\min_{R}\mathbb{E}_{\pi}[R(s,a)]-\mathbb{E}_{\pi_E}[R(s,a)],\end{aligned}\tag{4}$$

其中s和a为智能体的状态和动作，π_E为专家策略，R为奖励函数.

3.3 群体共识机制辨识方法

图1为面向群体共识机制的逆强化学习辨识方法框架图. 采集群体智能系统所产生的轨迹数据, 随后将采集到的数据作为专家数据，基于逆强化学习算法对马尔可夫博弈中的奖励函数以及策略函数进行辨识.

将群体智能系统建模为马尔可夫博弈，从而将对共识机制的辨识问题转化为还原马尔可夫博弈中智能体的奖励函数和策略函数的问题，其中策略函数是共识机制的外在表现，奖励函数是产生共识机制的内在激励. 通过对共识机制外在表现和内在激励的建模, 能够更全面地分析群体智能的共识机制.

4 逆强化学习辨识方法

4.1 集群马尔可夫博弈框架

对于群体智能系统，可以采用POSG来统一描述. 尽管POSG有较强的描述能力，但很难对该框架进行精确辨识. 集群系统例如鱼群和鸟群具有同构性和局部性的特点，系统中的所有个体具有相同的结构并且个体仅能感知一定范围内的系统状态. 考虑集群系统的特点, 做出式(5)假设:

$$\begin{aligned}&O_i=O_j,\quad\forall i,j,\\&A_i=A_j,\quad\forall i,j,\\&R_i=R_j,\quad\forall i,j,\\&R_i(s,a_1,\ldots,a_i,\ldots,a_k)=R_i(\xi(s)).\end{aligned}\tag{5}$$

对于任意智能体i,j，有相同的观测空间和动作空间、观测函数、策略函数；所有智能体的奖励函数相同；每个智能体的奖励取决于各自感知状态有关. 对于鱼群、鸟群以及无人机集群等系统，式(5)的假设均能够满足.

考虑上述特点, 将POSG框架修改为集群马尔可夫博弈框架(swarm-Markov game, SWARM-MG). 集群马尔可夫博弈如下所示:

集群马尔可夫博弈框架:

SWARM-MG, N个智能体的马尔可夫博弈定义为元组 $N,S,\{A_i\}_{i=1}^n,R,\{O_i\}_{i=1}^n,T,\xi$

N: 智能体数量

S: 状态空间

A: 智能体动作空间

R: $S\times A\to R$: 智能体的奖励函数

T: $S\times A\times S\to[0,1]$: 状态转移函数

O: 智能体i的观测空间

ξ: $S^N\to O^N$: 系统的观察函数

$\pi(\xi(s_i))$: 仅和局部观察有关的策略

其中，$\xi:S^N\to O^N$为观察函数，将个体的状态S映射到局部状态信息O. $R:S\times A\to R$为智能体奖励函数，代表状态S所能获取的奖励. $\pi(O)$: $O^N\to$A为智能体策略函数，观察为O时，所采取的决策. 个体通过调整策略π，最大化累积折扣奖励.

4.2 群体逆强化学习

强化学习在复杂高维环境中的应用越来越广泛. 为了获得理想策略，训练智能体时通常需要人为设计

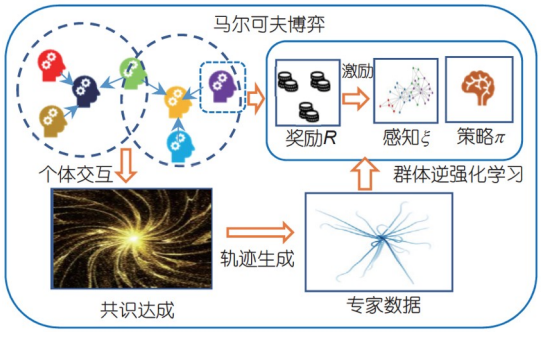

图 1 群体共识机制辨识方法

Figure 1 Identification method for collective consensus.

奖励函数. 对于简单任务, 设计奖励函数较为容易, 但对于复杂任务环境, 设置合理的奖励函数十分困难, 而群体场景使得奖励函数的设计更加困难. 群体逆强化学习能够在不依赖奖励函数的情况下, 从已有的智能体状态轨迹中, 反演群体智能系统的策略和奖励函数, 从而揭示其内在的共识机理. 当集群系统中个体是同构时, 利用参数共享能够更高效地训练策略. 参数共享的智能体共享同一个策略网络, 利用所有智能体的经验同时训练, 由于不同智能体对环境的观测不同, 基于相同策略能够产生不同的动作. 基于参数共享的方法具有良好的可扩展性, 适合对集群系统进行建模[25].

本文将单智能体的生成对抗逆强化学习方法扩展为群体场景的逆强化学习方法(PS-AIRL). 算法结构如图2所示, 基于生成对抗框架, 迭代训练生成器和判别器. 以系统所产生的轨迹为专家轨迹, 基于策略优化算法训练生成器并与环境交互采样轨迹, 训练判别器区分专家轨迹和策略采样轨迹. 判别器$D_{\theta,\phi}$需要对输入的专家轨迹τ_E和生成轨迹τ_π进行判别, 输出该样本为专家样本的概率, 最后根据判别结果的误差更新网络参数. 策略π将判别器$D_{\theta,\phi}$的输出以式(6)计算后作为奖励函数对自身进行评估和更新.

$$r_{\theta,\phi}=\log D_{\theta,\phi}-\log\left(1-D_{\theta,\phi}\right). \tag{6}$$

在交替训练中, $D_{\theta,\phi}$需要尽可能地区分专家轨迹和策略生成轨迹, 使得$D_{\theta,\phi}$在下一轮迭代中能更加准确地进行判别. 而生成器π需要不断逼近专家策略π_E, 从而消除生成轨迹和专家轨迹之间的差异, 生成能够以假乱真的样本, 使$D_{\theta,\phi}$无法对样本进行准确判别. 随着训练的进行, 便得到了策略函数和奖励函数.

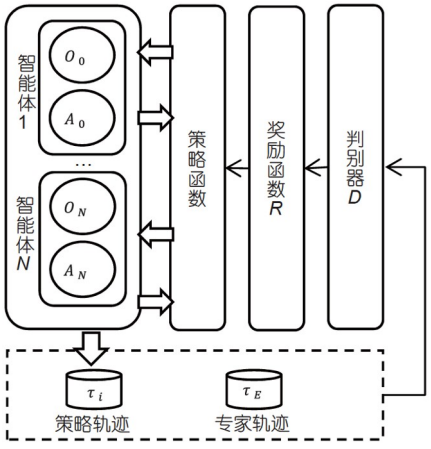

图 2　PS-AIRL算法架构图

Figure 2　Framework of PS-AIRL.

算法1描述了参数共享的对抗逆强化学习算法, 在算法的每次迭代中, 所有智能体共享同一个策略网络, 并以该网络为生成器与环境交互采样, 为每个智能体分布式地生成样本, 将所有智能体的样本作为一批数据, 更新网络参数. 奖励由判别器为每个状态计算, 由算法2更新生成器参数, 并在下一次迭代中更新判别器. 特别地, PS-AIRL算法中所有智能体共享同一个判别器网络参数产生奖励.

生成器采用参数共享的近端策略优化算法(proximal policy optimization, PPO)进行优化. 如算法2 PS-PPO所示. 在每次迭代, 每个智能体基于分布式策略采集样本, 将所有智能体的样本作为一批数据, 用以计算优势函数并更新网络参数.

5　实验与评估

本节在多个集群运动任务场景中评估本文所提出的共识机制辨识方法. 首先介绍仿真环境中智能体的基本属性, 随后分析和评估PS-AIRL算法的有效性.

算法1: PS-AIRL

输入: 专家轨迹$\tau_E \sim \pi_E$

初始化策略π和判别器$D_{\theta,\phi}$

For I = 1, 2, 3, ⋯, N

　执行策略π, 收集轨迹$\tau_i^{\,j}=(s_0,a_0,\ldots,s_T,a_T)$

　整理轨迹数据

　训练判别器$D_{\theta,\phi}$, 区分专家轨迹和生成轨迹

$$D_{\theta,\phi}(s,a,s')=\frac{\exp\left\{f_{\theta,\phi}(s,a,s')\right\}}{\exp\left\{f_{\theta,\phi}(s,a,s')\right\}+\pi(a\mid s)}$$

$$f_{\theta,\phi}(s,a,s')=g_\theta(s,a)+\gamma h_\phi(s')-h_\phi(s)$$

　更新奖励函数

$$R_{\theta,\phi}\leftarrow \log D_{\theta,\phi}-\log\left(1-D_{\theta,\phi}\right)$$

　使用算法2PS-PPO更新关于奖励$R_{\theta,\phi}$的策略π

End for

算法2: PS-PPO

输入: 初始化策略网络参数θ_0, 初始化值函数参数ϕ_0

For k = 0, 1, 2, ⋯ do

智能体j = 0, 1, 2, 3, ⋯, M 执行策略$\pi_{\theta_k}(o_j)$, 收集轨迹$\mathcal{D}_k = \{\tau_i\}^j$

轨迹数据归一化

计算累积折扣奖励reward-to-go $\hat{R}_t$

基于当前值函数V_{ϕ_k}, 估计优势函数$\hat{A}_t$

更新策略

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau\in\mathcal{D}_k} \sum_{t=0}^{T} \min \left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_k}(a_t \mid s_t)} A^{\pi_{\theta_k}}(s_t, a_t), g\left(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)\right) \right)$$

利用均方误差拟合值函数

$$\phi_{k+1} = \underset{\phi}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau\in\mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t \right)^2$$

End for

智能体的运动遵循一阶运动模型, 如式(7)所示:

$$\begin{aligned} \dot{x} &= v\cos\phi, \\ \dot{y} &= v\sin\phi, \\ \dot{\phi} &= \omega, \end{aligned} \tag{7}$$

式中x和y为智能体横坐标和纵坐标, v为智能体速度, ϕ为智能体运动方向. 每个智能体的感知模型如图3所示, 其中$d^{i,j}$为智能体i和j之间的距离, $\phi^{i,j}$为智能体i和j之间的速度位置夹角. 智能体能够感知到和邻居的距离以及速度位置夹角, 智能体i对智能体j的观察为$o^{i,j} = \{d^{i,j}, \phi^{i,j}\}$.

智能体通过感知函数$\xi(s_i)$对周围状态进行感知, 进而将所感知到的信息输入策略函数, 由策略函数输出动作. 感知函数定义为式(8), 使用神经网络作为特征映射函数, 其参数由强化学习算法确定.

$$\xi(s_i) = \frac{1}{n} \sum_{o^{i,j}\in O^i} h(Wo^{i,j} + b), \tag{8}$$

$\xi(s_i)$为智能体i对状态的观察, 其中n为智能体i邻居数量, h为全连接层和RELU激活函数, b为偏差. 智能体能够控制自身的线速度和角速度, 即$a = \{v, \omega\}$. 智能体i执行动作$a_i = \pi(\xi(s_i))$与其他智能体以及环境进行交互. 根据问题的规模不同, 可选择不同复杂程度的神经网络结构, 本文使用64个单元的全连接网络作为策略网络, 使用32个单元的全连接网络作为判别器网络. 本节实验均在E5-1650V3 CPU, 32G内存的硬件条件下进行, PS-AIRL算法经过约150000步训练可完成共识机制辨识, 在两个场景中均耗时120 min.

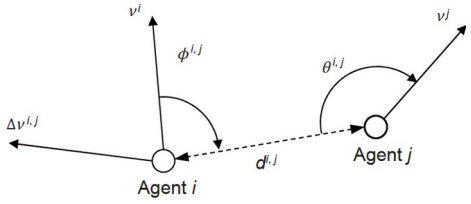

图 3 智能体状态定义

Figure 3 Definition of agent state.

5.1 聚集任务

本节以神经网络智能体为辨识对象, 在聚集任务中基于逆强化学习辨识其策略函数和奖励函数. 聚集任务是智能体在仅具有局部观测能力时, 分布式决策并汇聚到某一地点. 智能体的目标是最小化所有智能体之间的距离. 以式(9)为奖励函数, 其中d_c为智能体的感知距离, 所有智能体与邻居智能体的距离越小奖励值越大.

$$R(\mathbf{s}) = -\sum_{i=1}^{N} \sum_{j=i+1}^{N} \min\left(d^{i,j}, d_c\right). \tag{9}$$

基于参数共享的PPO算法在聚集任务中训练智能体, 使用训练完成的策略网络作为专家策略π_E, 与仿真环境交互, 收集轨迹作为专家样本τ_E. 辨识生成专家样本τ_E的策略和奖励函数, 以智能体的累积奖励情况评估算法.

实验比较PS-AIRL和BC对专家策略π_E和奖励函数R_E的辨识能力. 图4为随智能体数量变化和专家轨迹数量变化时, 各个算法对专家策略的辨识情况, 其中Exp为专家轨迹在不同智能体数量下的累积奖励变化, Airl为本文所提算法的累积奖励情况. 从结果可以看出, 本文所提算法辨识出的策略能够取得接近专家轨迹的效果, 且该算法对智能体数量的增加不敏感. 行为克隆算法BC的累积奖励不稳定, 智能体在环境中所能获取的奖励较低. 实际上, 由于所辨识的系统满足同

构性和局部性假设，基于参数共享进行逆强化学习不受智能体数量变化影响，而行为克隆受到分布漂移的影响, 不能取得好的辨识效果.

在动作空间中等间隔采样动作, 获取动作集合, 对特定状态下智能体执行动作集合中所有动作，可视化其所能获得的奖励. 图5为10个智能体聚集场景的奖励函数可视化情况，其中(a)为目标智能体在执行动作集合时所获取的奖励函数可视化情况，横纵坐标分别为智能体线速度与角速度，颜色越深代表奖励函数值越小. (b)中, 目标智能体距离其他智能体距离较远, 运动方向为远离其他智能体. (a)表明，在当前状态下线速度对奖励函数的影响较小，智能体以顺时针或逆时针转向其他智能体时奖励值较大. 实验结果表明基于PS-AIRL的群体逆强化学习算法能够辨识合理的奖励函数.

5.2　Vicsek模型

基于Vicsek模型生成群体专家数据, 使用PS-AIRL算法辨识其专家策略和奖励函数. 由N个智能体组成离散时间系统，个体的初始位置和运动方向随机分布，在$L\times L$的周期性边界条件的区域内自由移动，个体运动遵循Vicsek模型控制规则，运动方向依照其所有邻居运动角度的平均进行更新，并且会受到均值为零的噪声信号的干扰. 使用序参量衡量辨识效果，当所有智能体速度方向一致时，序参量最大. 序参量如下式定义:

$$V_a(t)=\frac{1}{Nv_0}\left\|\sum_{i=1}^{N}\mathbf{v}_i(t)\right\|, \tag{10}$$

式中N为智能体数量，v_0为智能体线速度，$\mathbf{v}_i(t)$为智能体速度向量.

实验比较PS-AIRL，BC对专家策略π_E和奖励函数R_E的辨识能力. 图6为随智能体数量以及专家轨迹数量变化时，各算法对专家策略的辨识情况. 其中Exp为专家轨迹在不同智能体数量下的累积序参量变化，Airl为本文所提算法的累积序参量情况. 从结果可以看出, 在对Vicsek模型进行辨识时, 我们所提方法对智能体数量的变化不敏感，能够达到接近专家轨迹的效果.

如图7可视化了N=10时辨识出的Vicsek模型奖励函数，(a)图横纵坐标为智能体线速度与角速度. (b)中

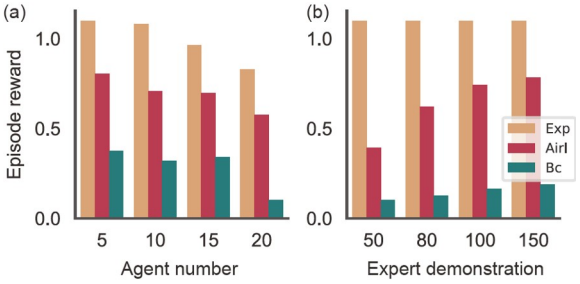

图 4　聚集任务策略辨识

Figure 4　Policy identification of the rendezvous task.

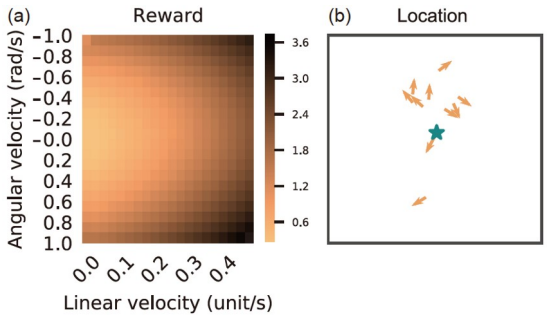

图 5　聚集任务奖励辨识

Figure 5　Reward identification of the rendezvous task.

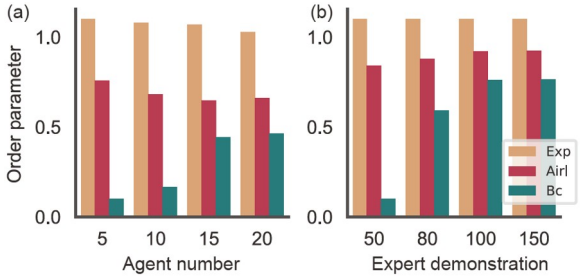

图 6　Vicsek模型策略辨识

Figure 6　Policy identification of the Vicsek model.

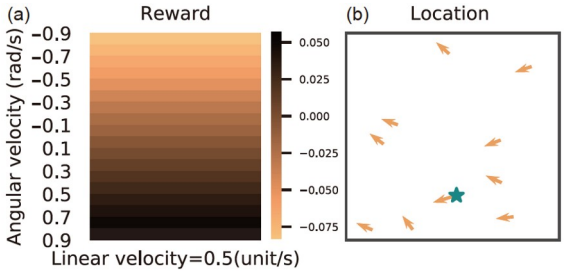

图 7　Vicsek模型奖励辨识

Figure 7　Reward identification of the Vicsek model.

智能体平均运动方向如红色箭头所示，目标智能体与平均运动方向有一定差距，需要顺时针旋转才能够满足Vicsek模型运动规则. (a)图表明智能体顺时针旋转一定角度时会获取更大奖励值. 实验结果表明PS-AIRL能够合理辨识Vicsek模型奖励函数.

6 结论与讨论

本文围绕共识机制辨识问题，从系统主义的角度出发，提出面向群体共识机制的逆强化学习辨识方法，并将上述辨识方法应用于集群系统，在多个场景中验证了对群体智能系统的辨识能力. 首先分析了群体智能系统的基本性质，并进一步将共识规则的辨识问题转化为对智能体策略函数和奖励函数的建模问题，随后考虑同构集群系统特点，设计了基于参数共享的群体逆强化学习算法，最后以群体运动为例，将上述群体智能理论的模型和方法应用于集群系统，通过实验验证了该方法对群体智能系统的辨识能力. 实验结果表明: 基于本文所提出的群体逆强化学习算法，能够从大规模同构集群系统的示例数据中辨识出策略函数, 实现对共识机制的反演; 同时, 所学习的奖励函数对于智能体的行为有一定的解释能力. 未来将基于所提出的群体逆强化学习算法, 在更多的集群场景中辨识共识机制, 并对所学习到的奖励函数进行更完善的分析和解释，以研究个体互动与群体行为之间的关系.

参考文献

1 Penn A, Turner J S. Can we identify general architectural principles that impact the collective behavior of both human and animal systems? Philos Trans R Soc Lond B Biol Sci, 2018, 373: 20180253

2 Li W, Wu W, Wang H, et al. Crowd intelligence in AI 2.0 era. Front Inf Technol Electron Eng, 2017, 18: 15–43

3 Krafft P M. A simple computational theory of general collective intelligence. Top Cogn Sci, 2019, 11: 374–392

4 Couzin I D, Krause J, James R, et al. Collective memory and spatial sorting in animal groups. J Theor Biol, 2002, 218: 1–11

5 Vicsek T, Czirók A, Ben-Jacob E, et al. Novel type of phase transition in a system of self-driven particles. Phys Rev Lett, 2006, 75: 1226–1229

6 Cucker F, Smale S. Emergent behavior in flocks. IEEE Trans Automat Contr, 2007, 52: 852–862

7 Sumpter D J T, Mann R P, Perna A. The modelling cycle for collective animal behaviour. Interface Focus, 2012, 2: 764–773

8 Heras F J H, Romero-Ferrero F, Hinz R C, et al. Deep attention networks reveal the rules of collective motion in zebrafish. PLoS Comput Biol, 2019, 15: e1007354

9 Pomerleau D A. Efficient training of artificial neural networks for autonomous navigation. Neural Comput, 1991, 3: 88–97

10 Abbeel P, Ng A Y. Apprenticeship learning *via* inverse reinforcement learning. In: Proceedings of the Twenty-First International Conference on Machine Learning. Banff Alberta, 2004. 1–8

11 Ziebart B, Maas A, Bagnell A, et al. Maximum entropy inverse reinforcement learning. In: Proceedings of the 23rd International Conference on Artificial Intelligence. Chicago, 2008. 1433–1438

12 Šošić A, KhudaBukhsh W R, Zoubir A M, et al. Inverse reinforcement learning in swarm systems. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems. São Paulo, 2017. 1413–1421

13 Reddy T S, Gopikrishna V, Zaruba G, et al. Inverse reinforcement learning for decentralized non-cooperative multiagent systems. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Seoul, 2012. 1930–1935

14 Song J, Ren H, Sadigh D, et al. Multi-agent generative adversarial imitation learning. In: Proceedings of the Annual Conference on Neural Information Processing Systems 2018. Montreal, 2018. 7472–7483

15 Yu L, Song J, Ermon S, Multi-agent adversarial inverse reinforcement learning. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, 2019. 7194–7201.

16 Zheng Z M, Lü J H, Wei W, et al. Refined intelligence theory: Artificial intelligence regarding complex dynamic objects (in Chinese). Sci Sin Inf, 2021, 51: 678–690 [郑志明, 吕金虎, 韦卫, 等. 精准智能理论: 面向复杂动态对象的人工智能. 中国科学: 信息科学, 2021, 51: 678–690]

17 Heylighen F. Stigmergy as a universal coordination mechanism I: Definition and components. Cogn Syst Res, 2016, 38: 4–13

18 Duan H B, Zhang D F, Fan Y M, et al. From wolf pack intelligence to UAV swarm cooperative decision-making (in Chinese). Sci Sin Inf, 2019, 49: 112–118 [段海滨, 张岱峰, 范彦铭, 等. 从狼群智能到无人机集群协同决策. 中国科学: 信息科学, 2019, 49: 112–118]

19 Lopez U, Gautrais J, Couzin I D, et al. From behavioural analyses to models of collective motion in fish schools. Interface Focus, 2012, 2: 693–707
20 Olfati-Saber R, Fax J A, Murray R M. Consensus and cooperation in networked multi-agent systems. Proc IEEE, 2007, 95: 215–233
21 Olfati-Saber R, Murray R M. Consensus problems in networks of agents with switching topology and time-delays. IEEE Trans Automat Contr, 2004, 49: 1520–1533
22 Hansen E A, Bernstein D S, Zilberstein S. Dynamic programming for partially observable stochastic games. In: Proceedings of the 19th National Conference on Artifical Intelligence. San Jose, 2004. 709–715
23 Åström K J, Eykhoff P. System identification—A survey. Automatica, 1971, 7: 123–162
24 Zhang K F, Yu Y. Methodologies for imitation learning via inverse reinforcement learning: A review (in Chinese). J Comp Res Devel, 2019, 56: 254–261 [张凯峰, 俞扬. 基于逆强化学习的示教学习方法综述. 计算机研究与发展, 2019, 56: 254–261]
25 Gupta J K, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Sao Paulo, 2017. 66–83

Identification method for collective consensus mechanism based on inverse reinforcement learning

YU Xin[1], WU WenJun[2,3], LUO Jie[1,3] & LI Wei[1,3]

[1] *School of Computer Science and Engineering, Beihang University, Beijing 100191, China;*
[2] *Institute of Artificial Intelligence, Beihang University, Beijing 100191, China;*
[3] *State Key Laboratory of Software Development Environment, Beijing 100191, China*

Collective intelligence, an important topic of the new generation of artificial intelligence, is a significant way to solve large-scale and complex problems in an open environment. It is crucial in other branches of artificial intelligence. Agents interact and evolve in response to the consensus mechanism to reach a group consensus. Identifying the consensus mechanism is essential for building and comprehending the collective intelligence system. Traditional consensus mechanism modeling methods require many simplified assumptions, and meeting the challenge of complex collective intelligence systems is difficult. A method for identifying consensus mechanisms based on data should be developed. In this paper, the consensus mechanism's identification problem is transformed into an inverse reinforcement learning problem for the collective intelligence system. We proposed inverse reinforcement learning methods for collective systems and evaluated them in two tasks. The results indicate that the proposed methods can identify the policy function and the reward function of the collective system.

systemism, collective intelligence, inverse reinforcement learning

doi: 10.1360/SST-2021-0370