doi:10.3969/j.issn.1672-1292.2022.01.013

小样本场景下的强化学习研究综述

王哲超[1,2,3],傅启明[1,2,3],陈建平[2,3],胡伏原[1,2,3],陆悠[1,2,3],吴宏杰[1,2,3]

(1.苏州科技大学电子与信息工程学院,江苏 苏州 215009)
(2.苏州科技大学江苏省建筑智慧节能重点实验室,江苏 苏州 215009)
(3.苏州科技大学苏州市移动网络技术与应用重点实验室,江苏 苏州 215009)

[摘要] 根据小样本问题背景,将小样本场景分成两类,第一类场景追求更专业的性能,第二类场景追求更通用的性能. 一般在知识泛化过程中,不同的场景对知识载体的需求有着明显的倾向性. 针对小样本学习方法,以知识载体的角度,将其分为使用过程性知识的方法和使用陈述性知识的方法,再讨论该分类下的小样本强化学习算法. 最后,从理论和应用等方面提出了可能的发展方向,以期为后续研究提供参考.

[关键词] 强化学习,小样本学习,元学习,迁移学习,终身学习,知识泛化
[中图分类号]TP181　[文献标志码]A　[文章编号]1672-1292(2022)01-0086-07

Review of Research on Reinforcement Learning in Few-Shot Scenes

Wang Zhechao[1,2,3],Fu Qiming[1,2,3],Chen Jianping[2,3],Hu Fuyuan[1,2,3],Lu You[1,2,3],Wu Hongjie[1,2,3]

(1.School of Electronic and Information Engineering,Suzhou University of Science and Technology,Suzhou 215009,China)
(2.Jiangsu Provincial Key Laboratory of Building Intelligence and Energy Saving,Suzhou University of Science and Technology,Suzhou 215009,China)
(3.Suzhou Key Laboratory of Mobile Networking and Applied Technologies,Suzhou University of Science and Technology,Suzhou 215009,China)

Abstract:According to the background of the few-shot problem,this paper divides few-shot scenes into two types. The first type of scenes pursues more professional performance,while the other pursues more general performance. In the process of knowledge generalization,different scenes have obvious tendency to the requirement of knowledge carrier. Because of the discovery,the FSL is divided into two types in terms of knowledge carrier,where one type uses procedural knowledge and the other uses declarative knowledge. Then FS-RL algorithms under this classification are discussed. Finally,the possible development direction is proposed from the theory and the application,hoping to provide insights to following research.

Key words: reinforcement learning, few-shot learning, meta-learning, transfer learning, lifelong learning, knowledge generalization

数据驱动下的人工智能技术已经得到了迅猛的发展,特别是对于存在大量样本数据背景下的机器学习方法,其训练结果都显示出了较高的水平,甚至达到超人类水平(如围棋、竞技游戏). 但需要注意的是,这些成果都只是某一智能体于某项特定任务的表现. 特别是对高维数据[1],智能体表现出的对数据的贪婪性和对任务间较差的泛化性,是现今人工智能技术发展的主要瓶颈.

人在处理未知环境中的新任务时,总是基于两个前提:一是对新任务的少量观测,二是经验中的旧知识. 小样本学习(few-shot learning,FSL)正是如此,在大量训练样本中提取一般性的知识,在小量测试样本上泛化这些知识,以适应新的、不同需求的任务[2]. 这也阐明了 FSL 所需要的两个过程,即知识归纳和知识迁移.

强化学习作为机器学习中处理序贯决策问题的重要方法,在广泛的实际应用中也出现了很多小样本场景. 在学习过程中,强化学习虽然不像监督学习一样依赖于大量人工标签,但其同样需要在与环境不断

收稿日期:2021-08-31.
基金项目:国家重点研发计划项目(2020YFC2006602)、国家自然科学基金项目(62072324、61876217、61876121、61772357、62073231、61902272)、江苏省重点研发计划项目(BE2017663).
通讯作者:傅启明,博士,副教授,研究方向:强化学习、深度学习、智能信息处理等. E-mail:fqm_1@126.com

交互中收集样本,而有些场景下(如智能驾驶),因为损耗、响应时延等问题使得智能体能够收集的样本数量是有限的. 利用知识复用解决小样本问题,必然使智能体处于样本分布不一的任务环境中. 智能体必须利用这些由样本体现出来的任务间的相似性,合理复用先验知识,获得在新任务下的较高水平表现.

目前,很多研究者针对强化学习中的小样本问题做了大量工作. 迁移学习,要求智能体通过研究不同任务下变量(如状态、动作等)之间的映射(相似)关系来完成知识(如样本、参数等)的迁移;学会学习,强调智能体能学会学习的方法而不是通过人工设置参数进行学习. 终身学习,要求智能体始终保有并利用历史知识;元学习,强调智能体提炼任务间不变的元知识并复用. 因此,本文以小样本场景下的强化学习方法为对象,针对其前沿研究现状进行综述.

1 小样本强化学习框架

1.1 强化学习

强化学习旨在强化个体或群体在某一特定环境下的适应能力,这样的智能体被称为 agent. 具体地讲,强化学习要求 agent 可以在特定环境状态下,通过与环境的交互,求解能够带来最大期望累计奖赏的行为策略[3]. 其中,奖赏信号是一个标量,它是指导 agent 做出更好行为决策的评估信号. 在强化学习中,通常利用马尔可夫决策过程(Markov decision processes,MDPs)对问题进行建模. MDPs 需要由一个四元组 S,A,T,R 来表示,其中 S 表示状态空间,A 表示动作空间,T 和 R 分别表示状态转移函数和奖赏函数,

$$T_{ss'}^{a}=T[S_{t+1}=s'\mid S_t=s,A_t=a], \tag{1}$$

$$R_{ss'}^{a}=E[R_{t+1}\mid S_t=s,A_t=a,S_{t+1}=s']. \tag{2}$$

式中,$T_{ss'}^{a}$ 表示在 t 时刻状态 s 经动作 a 转移至状态 s' 的概率,$R_{ss'}^{a}$ 表示在 t 时刻状态 s 经动作 a 转移至状态 s' 后获得的期望奖赏.

1.2 小样本强化学习的问题定义

定义任务 M 为一个 MDPs,与其对应的四元组 S_m,A_m,T_m,R_m 规定了该任务的细节,其中 S_m、A_m、T_m 和 R_m 分别对应了该任务的状态空间、动作空间、转移模型和奖赏函数. U 是任务空间,即 $U=\{M\}$;Ω 是 M 在 U 上的概率分布,即 $M\sim\Omega$;环境 ε 负责实例化 M,其包含了 U 和 Ω,也由 U、Ω 两者定义,即 $\varepsilon=\langle U,\Omega\rangle$.

一般小样本强化学习要求满足以下的问题设定:定义两个 MDPs. 其中,一个或者多个旧任务 $M_o=\langle S_o,A_o,T_o,R_o\rangle$ 和一个新任务 $M_n=\langle S_n,A_n,T_n,R_n\rangle$,定义一个环境 ε 包含任务空间 $U=\langle M_o,M_n\rangle$ 和任务概率分布 Ω. 当有多个旧任务时,旧任务之间的 MDPs 可以不同,但为了简化这里都记作 M_o. 因为 FSL 是机器学习中的一个分支,所以这里根据 Mitchell[4] 提出的机器学习定义,给出小样本强化学习的定义:

定义 1 小样本强化学习. 计算机程序在旧任务 M_o 中通过强化学习在知识空间 K_o 中学习后,又在新任务 M_n 的知识空间 K_n 中学习,agent 使用归纳算法 A_{conclude} 和迁移算法 A_{transfer} 将 K_o 与来自新任务 M_n、含有少量知识的 K_n 依次转变为归纳知识 K_{conclude} 和迁移知识 K_{transfer},学习算法 A_{learn} 根据 K_{transfer} 在假设空间 H 学习一个假设 h.

2 小样本强化学习的分类

小样本问题极为普遍,以致于很难提出一种分类方法可以涉及所有已提出的学习算法. 本文主要对一些主流方法进行分类. 虽然以往较多的综述工作都涉及了小样本问题,但都在迁移学习、元学习等这些学习算法各自领域里进行讨论,缺少对小样本这个问题本身的探究. 而本文直接以小样本场景为入口,以不同的场景需求为启发,提出分类方法.

2.1 小样本场景

一般有两种视角定义小样本场景,一种是知识从模拟世界到真实世界泛化的角度,另一种是类人智能发展的角度,这两个都是小样本问题中最常见的视角,因为这两者都涉及到知识在任务间泛化的问题.

一般常用“现实差距”来描述学习方法由模拟世界应用至真实世界时产生的不好结果. Agent 在模拟世界中,可以比物理机器人更快地收集扩展性更强、成本更低的数据,以此加速其学习. 但当 agent 从模拟世界向真实世界迁移时,会因两者的差异,而影响智能体在真实世界中的表现,一般这是 agent 对模拟世界过拟合导致的,称为“sim2real”问题. 这一现象是强化学习应用中经常遇到的问题,如机械手臂的物品

抓取任务[5]. 本文将其称为第一类小样本场景. 在该场景下,FSL 需要做的是充分利用先验知识,找到这种场景下知识复用的最佳方式,使得在真实世界中 agent 要有较好的初始在线性能力[6].

小样本场景的另一种视角是发展类人智能. 虽然专门的人工智能对专门任务有更好适应性,但是人们期望人工智能可以和人一样拥有通用的学习算法. 通用学习算法的核心思想就是提炼通用知识. 本文将其称为第二类小样本场景. FSL 要求智能体在平时的任务中归纳通用的知识,在面对新任务时结合较少的样本将通用的旧知识应用至该任务中. 这里的通用知识不单只是用来适应当前面临的新任务,而是用来解决将来的一类任务. 这与第一类小样本场景需求的区别是,第一类以在单一新任务中产生最佳表现为目的,而第二类则以快速适应多个任务为目的. 第一类希望 agent 成为专家,而第二类则希望 agent 变得更博学.

2.2 小样本强化学习方法

一些心理学研究表明,人类可以通过在相似的任务中迁移知识,以达到更快的学习效果. 美国心理学家 Anderson 提出的思维适应性控制(adaptive control of thought,ACT)理论[7],将人的认知分为过程性认知和陈述性认知. 在王皓等[8]的基础上,本文加入知识归纳作为区分,即,将知识迁移分为未作知识归纳、注重挖掘并利用任务相似性的直接迁移和已作知识归纳、注重抽象并理解任务复杂过程的间接迁移.

“直接”和“间接”的区别是,迁移是否经过知识归纳,即是否产生陈述性认知. 对单个任务来说,agent 需要输出的是一个针对该任务的策略(h). 若知识的主要内容是在该策略下产生的样本($\{s\times a\times s\times r\}^n$),那么这就是直接迁移. 因为这些样本代表的是过程性认知. 相反,若将环境交互中归纳出来的知识作为迁移内容时,即为间接迁移.

3 直接迁移

3.1 从演示中学习

从演示中学习(learning from demonstrations,LFD),是为了解决标准强化学习中的随机探索问题,因为这耗费了大量的时间. 其中具体涉及两个问题,一是如何获得更好的演示样本,二是在当前的演示样本下如何更高效的学习,因为后者更接近一个学习问题,所以本节将重点放在后者上. 一般地,演示样本来自于专家.

Kim 等[9]在近似策略迭代(approximate policy iteration,API)[10]中使用专家示例以降低学习所需的样本量. 在 API 中,给出第 k+1 次迭代的待评估策略 π_k,通过 Bellman 算符 T^{π_k}进行近似估计得到 $\hat{Q}_k$,其为近似固定点,即 $T_\pi:\hat{Q}_k\approx T^{\pi_k}\hat{Q}_k$,最后基于估计使用贪心法计算出新策略 π_{k+1}. LFD 可以被看作一个组合凸优化问题,其包括标准强化学习的优化问题和演示优化问题. Piot 等[11]认为用 API 来解决 LFD 所带来的一个问题,即在策略迭代中每次评估的策略并非最优策略,而专家示例可以看作最优策略的演示,把这两种优化整合,可能降低学习的速率. 因此,该工作中使用值迭代的方式来进行标准强化学习,两部分优化都同时以逼近最优策略为目标,从而避免了不必要的策略评估.

从上述对比可见,这种方法虽然对前人工作进行了补充,但伴随而来的是要解决一个非凸的、不可微和有偏的优化. Chemali 等[12]对此提出了一种基于贝叶斯模型的方法,该方法使用贝叶斯方法从专家示例中整合模型知识.

3.2 样本加权

一般地,来自相似任务的样本分布是不完全一致的. Lazaric 等[13]认为需要从尽可能相似的旧任务中选取样本给新任务学习. Cortes 等[14]提出用重要性权重来对 loss 进行加权,以强调(或减少)某些样本的误差,从而解决分部之间不匹配的问题. 重要性权重的定义是,对于样本点 $X=(S\times A\times S\times R)$,有 $w(X)=P(X)/Q(X)$,其中 P、Q 分别指新、旧任务下该样本点的分布. 但是这种方法存在两个明显的挑战:(1)样本分布很难获得,即使能够获得也可能很难描述;(2)在某些情况中,新、旧任务的环境动态可能很相似,但是由于奖赏分布极为不同,那么权重可能逼近于 0,导致知识迁移无法进行. 针对第一个挑战,大多数做法都是通过采样解决. 在第二个挑战中,Laroche 等[15]在任务间环境动态相同的情况下,提出用监督学习方法学习一个奖赏估计器. 而对不同的环境动态,Tirinzoni 等[16]提出在拟合 Q 迭代(fitted Q-iteration,FQI)[17]的基础上进行重要性样本迁移. 因为第二个挑战需要解决奖赏差异带来的问题,所以第一步需要

训练一个可用的奖赏模型 $\hat{R}$,

$$\hat{R}=\arg\inf_{h\in H^Q}\frac{1}{Z_r}\sum_{j=0}^{m}\sum_{i=0}^{N_j}w_{r,i}^{(j)}\left|h(X_i^{(j)})-r_i^{(j)}\right|^2. \tag{3}$$

式中,H^Q 是状态-动作值函数的假设空间,$Z_r=\sum_{i,j}w_{r,i}^{(j)}$,上标 j 表示样本来自第 j 个任务,权重计算方式如下:

$$w_{r,i}^{(j)}=\frac{R_0(r_i^{(j)}\mid X_i^{(j)})}{R_j(r_i^{(j)}\mid X_i^{(j)})}. \tag{4}$$

3.3 奖赏塑造

奖赏塑造(reward shaping,RS)通过塑造奖赏信号,鼓励(或阻止)特定的行为,进而减少 agent 不必要的探索,降低样本需求量. RS 将原样本($S\times A\times S\times R$)转变成($S\times A\times S\times R'$). 新的奖赏信号 $R'=R+F$,其中 $F:S\times A\times S'\to\mathbf{R}$,是一个有界实数函数,称为塑造函数.

Ng 等[18]在 RS 领域提出的基于势的奖赏塑造(potential-based reward shaping,PBRS)为不同状态定义了势函数 $\Phi(s)$. 当 agent 访问到不同状态时形成势差 $\gamma\Phi(s')-\Phi(s)$,并以此作为塑造奖赏的值,同时文中也证明了奖赏塑造后获得最优策略和奖赏塑造前一致. Wiewiora 等[19]提出的基于势的建议信息(potential-based advice,PBA)将动作加入势函数,并给出势函数的前瞻建议和回溯建议的定义,前者根据后一状态的势来进行奖赏塑造,而后者是根据前一个状态的势来进行奖赏塑造. 此外,PBA 还提出加入动作的势函数在回溯建议中更具优势. Devlin 等[20]提出了基于动态势(dynamic potential-based,DPB)的奖赏塑造,来解决势函数在学习过程中动态变化的问题,特别是当 agent 被要求自动学习势函数时. 同时 DPB 也扩展了其在多 agent 任务方面的研究,并证明了该方法可以获得纳什均衡. Harutyunyan 等[21]提出的(dynamic potential-based advice,DPBA),要求 agent 在学习策略的同时,也要学习关于状态-动作的势函数,同样,逐步学习的势函数能促进 agent 更快地找到最优策略.

4 间接迁移

4.1 元学习

元学习主要使用在多任务中,它要求 agent 在学习过程中归纳元知识,并在新任务中利用元知识. 元学习的目标是让 agent 学到多个任务中的共性知识,而不是和某些任务的基线算法去比较性能. Finn 等[22]提出模型不可知的元学习(model-agnostic meta-learning,MAML)是元强化学习的重要方法. 该方法使用一个策略参数 θ 来适应多个任务 $M_i\sim\Omega$,这使得拥有该策略 h_θ 的 agent 不再只适用于一项特定的任务.

MAML 算法下 agent 的学习过程如图 1 所示,agent 分别在各个任务 M_i 中交互学习,在元参数 θ 的基础上根据学习目标 $L_{M_i}(h_\theta)$ 得到更新后的参数 θ_i',再在策略 $h_{\theta_i'}$ 下产生交互样本集 D_i',最后根据综合目标 $\sum_{M_i\sim\Omega}L_{M_i}(h_{\theta_i'})$ 更新元参数 θ,以此往复. 但是并不是所有适应阶段都会对新任务产生更好的策略,如果出现不好的策略,那这个过程被称为负适应. Deleu 等[23]称这种过度专门化是产生负适应的根本原因,即在元学习中的策略参数对任务空间中的某一个(或某些)任务产生了专门能力,这使得再在这个(或这些)任务上适应时可能出现比初试策略差的策略.

4.2 策略蒸馏

DeepMind 提出的策略蒸馏[24]通过预训练的方法,将教师 agent 的能力“传授”给学生 agent,这体现在通过策略蒸馏将教师样本(过程性知识)转化为学生的策略(陈述性知识). 与一般预训练方法不同的是,为了从教师 agent 那产生的专家样本包含更多知识,策略蒸馏在网络最后产生 Q 值的 Softmax 中设置了较高的温度参数,这使得同一输入下的不同 Q 值较为相近,而学生需要学习的目标 Q 值则通过降低温度参数获得,这使得最优动作的 Q 值与其他 Q 值差距增大,对于学生来说最优动作更为明显. 蒸馏时一般使用 KL 距离作为损失函数. 多任务策略蒸馏[24]与元学习的理念非常相似,它们都将应对不同任务的能力凝练到一个 agent 的策略上. 而且同样地,多任务策略蒸馏的首要目的并不是复制专家的能力,而是使学习者更加“博学”. 图 2 展示了整个策略蒸馏的过程,每个专家都在自己的领域中在线地生产专家示例并放入对应的回放单元,以供博学者来学习. 博学者在这些回放单元中抽取训练样本进行有监督的学习. 需要说

明的是,如果把博学者看成是最终的策略(h),那么这就是一种直接迁移,但在小样本问题下,策略蒸馏并没有完成全部任务,博学者还需要结合特定任务的知识来完成适应阶段,所以这时候的博学者更应被看成是间接迁移中的归纳知识(K_{conclude}).

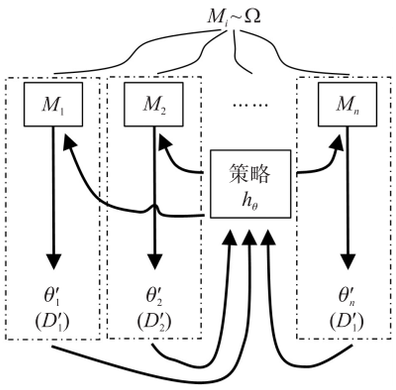

图 1　MAML 算法下 agent 的学习过程

Fig. 1　Learning process of agent in MAML

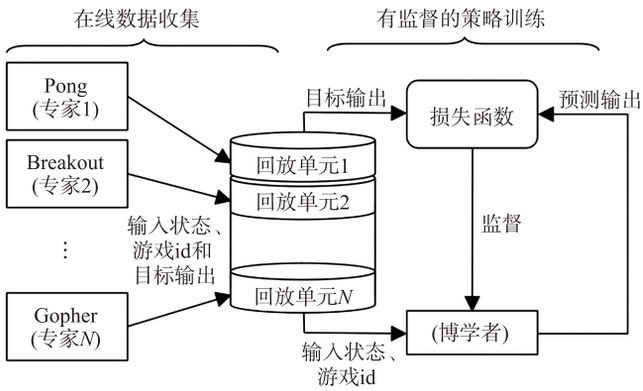

图 2　多任务策略蒸馏过程

Fig. 2　Multi-task policy distillation process

4.3　状态抽象

在机器学习中,抽象是十分重要的理念之一,Abel[25]认为好的抽象可以帮助强化学习快速发现和有效学习高价值的策略. 特别地,当 MDP 中拥有连续的、巨大的状态空间时,状态抽象尤为重要. 状态抽象使 agent 凝练关于状态的知识,简化问题的复杂性,降低对非重要特征的关注度.

几乎任何类型的抽象都会丢掉一定数量的信息. 然而,在强化学习中可取的方法是,抽象保留足够的相关信息以使 agent 最终学会解决感兴趣的问题. 一旦压缩状态,agent 处理的信息量减少,如图 3 所示.

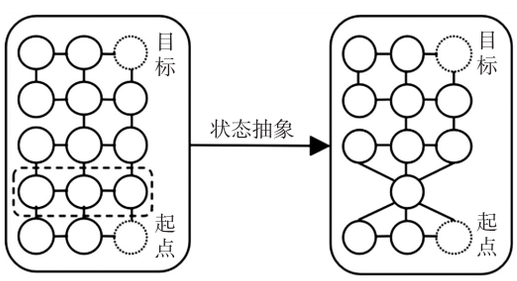

图 3　状态抽象示意图

Fig. 3　State abstraction diagram

首先状态抽象需要定义一个关系 P,

$$P:S\times S\to\{0,1\}. \tag{5}$$

状态抽象方法根据一个固定的映射 $\varphi:S\to S_\varphi$,将状态分入不同状态簇中,当 φ 将不同的状态分入相同的簇时,P 为 1,否则为 0,即

$$\varphi(s_1)=\varphi(s_2)\Rightarrow P(s_1,s_2). \tag{6}$$

Abel 等[26]提出的状态抽象方法将状态根据对应最优 Q 值依间隔值 d 进行分类,其定义如下:

$$P^d(s_1,s_2)\triangleq\forall_a:\frac{Q^*(s_1,a)}{d}=\frac{Q^*(s_2,a)}{d}. \tag{7}$$

式中,$d\in[0,V_{\max}]$,$V_{\max}$ 收敛于 $R_{\max}/(1-\gamma)$,$R_{\max}$ 为奖赏的最大值. 这种方法使得状态抽象具有传递性,即 $[P(s_1,s_2)\wedge P(s_2,s_3)]\Rightarrow P(s_1,s_3)$. 传递性对状态抽象非常重要,这大大减少了计算量.

在终身学习下的状态抽象问题,需要在某一个任务分布下训练,因此算法要保证高概率的适用性. 那么,如果一个状态抽象满足可能近似正确[27],就称为这个抽象属于 PAC-状态抽象. 形式化表达为,对任意一组状态(s_1,s_2),当且仅当 P 在任务分布上为真的概率是 $1-x$,那么 $\rho_x^P(s_1,s_2)$ 为真,其表达式如下:

$$\rho_x^P(s_1,s_2)\triangleq\underset{M\sim\Omega}{Pr}\{P_M(s_1,s_2)=1\}\geqslant 1-x. \tag{8}$$

5　理论与应用

FSL 对现有学习算法在样本数量方面提出更高的要求. 无论是在深度学习上[28-29]还是在强化学习上,知识复用一直以来是小样本问题的主要解决手段,所以这些方法自然而然的发展成了迁移学习、元学习、终身学习等. 在早期的迁移学习和一些域适应方法中,一般更关注知识迁移的方法,并未将样本数量的限制作为问题背景,且任务间知识泛化是一对一的,但其中大部分算法在小样本场景下的具有突出能力. 近来,小样本问题成了机器学习中的热点话题,迁移学习方法与其他方法一样致力于解决小样本问

题,且知识泛化在一批任务中进行[30]. MAML 这种参数微调方法,一直是元学习中许多工作开展的基础. 但面对一批批的任务,如何高效的学习仍是一个关键问题,Mehta 等[31]的思路是应用课程学习[32]的思想调整学习任务的次序. 在严格的终身学习条件下对可利用任务要求更加苛刻,久远的任务将无法被采样,所以其目标是解决知识保有的问题,即在遗忘中尽可能保留重要的信息.

一般地,自动驾驶[33]、协同对抗[34]、游戏竞技[35]很难使用深度强化学习算法. 因为通常算法要获得一个良好的表现需要训练上百万步,而且需要一个非常完美的模拟器,但在真实世界中很难实现. 在真实世界中 agent 要对它的行为负责,它需要从一开始就有一个良好的在线性能[6]. 因此,如何通过 FSL 预训练一个策略并在真实世界应用,是 FSL 应用未来需要不断深入的研究.

[参考文献](References)

[1] 吉珊珊. 基于神经网络树和人工蜂群优化的数据聚类[J]. 南京师大学报(自然科学版),2021,44(1):119-127.

[2] LI F F, FERGUS R, PERSON P. A bayesian approach to unsupervised one-shot learning of object categories[C]//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE, 2003: 1134-1141.

[3] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. London: MIT Press, 2018.

[4] MITCHELL M T. Machine learning[M]. New York: McGraw-Hill, 1997.

[5] TOBIN J, FONG R, RAY A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[J]. arXiv Preprint arXiv: 1703.06907, 2020.

[6] HESTER T, VECERIK M, PIETQUIM O, et al. Deep Q-learning from demonstrations[C]//The 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 3223-3230.

[7] ANDERSON J R. Cognitive psychology and its applications[M]. 3rd ed. New York: Freeman, 1990.

[8] 王皓,高阳,陈兴国. 强化学习中的迁移:方法和进展[J]. 电子学报,2008,36(Suppl 1):39-43.

[9] KIM B, FARAHMAND A, PINEAU J, et al. Approximate policy iteration with demonstration data[C]//The 1st Multi-disciplinary Conference on Reinforcement Learning and Decision Making. Princeton, USA, 2013: 168-172.

[10] BERTSEKAS D P. Approximate policy iteration: a survey and some new methods[J]. Journal of Control Theory and Applications, 2011, 9(3): 310-335.

[11] PIOT B, GEIST M, PIETQUIN O. Boosted bellman residual minimization handling expert demonstrations[C]//The 25th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Nancy, France, 2014: 549-564.

[12] CHEMALI J, LAZARIC A. Direct policy iteration with demonstrations[C]//The 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3380-3386.

[13] LAZARIC A, RESTELI M, ANDREA B. Transfer of samples in batch reinforcement learning[C]//The 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 544-551.

[14] CORTES C, MOHRI M, RILEY M, et al. Sample selection bias correction theory[C]//The 19th International Conference on Algorithmic Learning Theory. Budapest, Hungary, 2008: 38-53.

[15] LAROCHE R, BARLIER M. Transfer reinforcement learning with shared dynamics[C]//The 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 2147-2153.

[16] TIRINZONI A, SESSA A, MATTEO P, et al. Importance weighted transfer of samples in reinforcement learning[C]//The 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 4943-4952.

[17] ERNST D, GEURTS P, WEHENKEL L. Tree-based batch mode reinforcement learning[J]. Journal of Machine Learning Research, 2005, 6(4): 503-556.

[18] NG A Y, HARADA D, RUSSELL S J. Policy invariance under reward transformations: Theory and application to reward shaping[C]//The 16th International Conference on Machine Learning. Bled, Slovenia, 1999: 278-287.

[19] WIEWIORA E, COTTREL G W, ELKAN C. Principled methods for advising reinforcement learning agents[C]//The 20th International Conference on Machine Learning. Washington DC, USA, 2003: 792-799.

[20] DEVLIN S, KUDENKO D. Dynamic potential-based reward shaping[C]//The 11th International Conference on Autonomous Agents and Multiagent Systems. Valencia, Spain, 2012: 433-440.

[21] HARUTYUNYAN A, DEVLIN S, VRANCX P, et al. Expressing arbitrary reward functions as potential-based advice[C]//The 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2652-2658.

[22] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//The 34th Interna-

tional Conference on Machine Learning. Sydney, Australia, 2017: 1126-1135.
[23] DELEU T, BENGIO Y. The effects of negative adaptation in Model-Agnostic Meta-Learning[J]. arXiv Preprint arXiv: 1812.02159, 2018.
[24] RUSU A A, COLMENAREJO S G, GÜLCEHRE C, et al. Policy distillation[C]//arXiv Preprint arXiv: 1511.06295, 2016.
[25] ABEL D. A theory of state abstraction for reinforcement learning[C]//The 31st Innovative Applications of Artificial Intelligence Conference. Honolulu, USA, 2019: 9876-9877.
[26] ABEL D, HERSHKOWITZ D E, LITTMAN M L. Near optimal behavior via approximate state abstraction[C]//International Conference on Machine Learning. New York, USA, 2016: 2915-2923.
[27] VALIANT L G. A theory of the learnable[J]. Communications of the Association for Computing Machinery. 1984, 27(11): 1134-1142.
[28] YAO H, ZHANG C, WEI Y, et al. Graph few-shot learning via knowledge transfer[C]//The 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 6656-6663.
[29] ZHANG C, YAO H, HUANG C, et al. Few-shot knowledge graph completion[C]//The 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3041-3048.
[30] PARISOTTO E, BA J L, SALAKHUTDINOV R. Actor-mimic: deep multitask and transfer reinforcement learning[C]//The 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 156-171.
[31] MEHTA B, DELEU T, RAPARTHY S C, et al. Curriculum in gradient-based meta-reinforcement learning[J]. arXiv Preprint arXiv: 2002.07956, 2020.
[32] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]//The 26th Annual International Conference on Machine Learning. New York, USA, 2009: 41-48.
[33] HESTER T, STONE P. Texplore: real-time sample-efficient reinforcement learning for robots[J]. Machine Learning, 2013, 90(3): 385-429.
[34] 施伟, 冯旸赫, 程光权, 等. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报, 2021, 47(7): 1610-1623.
[35] 孟琭, 沈凝, 祁殷俏, 等. 基于强化学习的三维游戏控制算法[J]. 东北大学学报(自然科学版), 2021, 42(4): 478-482, 493.

[责任编辑: 陈　庆]

《南京师范大学学报（工程技术版）》

版权声明

为适应我国信息化建设发展的需要，扩展广大作者的学术交流渠道，本刊已入编台湾中文电子期刊服务——思博网（CEPS），并成为《中国学术期刊（光盘版）》、《万方数据—数字化期刊群》、《中国期刊网》、《中国核心期刊（遴选）数据库》等全文收录期刊。

凡向本刊所投稿件，均视为愿意供上述各文献数据库收录、转载并上网发行。其作者文章著作权使用费已包含在本刊支付的稿酬中，本刊不再另付其他稿酬。如作者不同意文章被收录，请在投稿时向本刊声明，本刊将做适当处理。

关于抵制学术不端行为的声明

近年来，少数作者在投稿中有抄袭剽窃、伪造注释、一稿多投等学术不端行为。这些无视学术道德与操守的不端行为，不仅违反了国家的有关法律、法规，也给编辑工作造成了极大困扰。为倡导优良学风，规范学术行为，净化学术空气，根据教育部《关于严肃处理高等学校学术不端行为的通知》（教社科 [2009]3 号），本刊发表如下声明：

一、从本声明公布之日起，凡向我校学报投稿，作者均须严格遵守《中华人民共和国著作权法》等国家有关法律、法规。

二、对由于抄袭剽窃、一稿多投等不端行为造成严重后果者，学报将在刊物上对其不端行为进行公开曝光，同时通知当事人所在单位，并保留对其追责的权利。

三、凡被发现有抄袭剽窃、一稿多投等学术不端行为者，本刊将拒绝刊发其任何文章。

本刊愿与广大作者一起，为共同构建规范健康的学术平台而不懈努力。

《南京师范大学学报（工程技术版）》

编 辑 部

南京师范大学学报（工程技术版）
Nanjing Shifan Daxue Xuebao (Gongcheng Jishu Ban)

JOURNAL OF NANJING NORMAL UNIVERSITY
ENGINEERING AND TECHNOLOGY EDITION

季刊
2001年创刊
第22卷第1期（总第85期）
2022年3月20日出版

Quarterly
Started in 2001
Vol.22,No.1 (Sum No.85)
Published Date: 20 Mar,2022

收录本刊的主要数据库或文摘刊物

《中国期刊全文数据库》
《中国科技期刊引证报告》
《万方数据——数字化期刊群》
《中文学术期刊综合评价数据库》
《中国核心期刊（遴选）数据库》
龙源国际期刊网
华艺思博网（CEPS）
《电子科技文摘》
《中国数学文摘》
德国《数学文摘》
美国《化学文摘》（CA）
英国《科学文摘》（SA,INSPEC）
美国《剑桥科学文摘：材料信息》
俄罗斯《文摘杂志》（AJ,VINITI）
美国《乌利希国际期刊指南》

主管单位：江苏省教育厅
主办单位：南京师范大学
出版单位：南京师范大学学报编辑部
（210097，南京市宁海路122号）
主　　编：胡敏强
执行主编：莫祥银
网　　址：http://xuebao.njnu.edu.cn
电子信箱：gkxb@njnu.edu.cn
电　　话：025-83598631
印刷单位：南京凯德印刷有限公司
国内发行：中国版本图书馆发行部
（100005，北京市先晓胡同10号）
国外发行：龙源国际期刊网（www.qikan.com）
发行范围：公开
广告经营许可证号：3200004060748
封面设计：吴振韩

Administrated by：Department of Education of Jiangsu Province
Sponsored by：Nanjing Normal University
Published by：Editorial Board of JNNU
Address：122 Ninghai Road，Nanjing 210097，China
Editor-in-Chief ：HU Minqiang
Executive Editor-in-Chief ：MO Xiangyin
Website：http://xuebao.njnu.edu.cn
E-mail：gkxb@njnu.edu.cn
Tel：86-25-83598631
Printed by：Kaide Typographic Co.,Ltd
Overseas Distributor：Dragonsource.com Inc

ISSN 1672-1292
03
9 771672 129221

国际标准连续出版物号 ISSN 1672-1292
国内统一刊号 CN 32-1684/T

国内定价：6.00元
国外定价：6.00美元