鸽子序贯决策中动态强化学习建模与策略演变

李志辉[1,2], 马　莹[1,2], 尚志刚[1,2], 杨莉芳[1,2,3*]

(1.郑州大学 电气与信息工程学院,河南 郑州 450001;2.河南省脑科学与脑机接口技术重点实验室,河南 郑州 450001;3.郑州大学附属脑病医院,河南 驻马店 463000)

摘　要:生物体为了最大化未来回报,在复杂环境中需灵活调整学习策略。为探究生物体在序贯决策学习过程中学习策略的动态演变规律,以鸽子为模式动物,设计了两步序贯决策实验范式,记录了鸽子从初始探索到习得整个阶段的行为学数据,分别构建了基于奖励预测误差驱动的 Model-Free(MF)与状态间关系驱动的 Model-Based(MB)两类动态强化学习(reinforcement learning, RL)模型,利用实验数据对上述模型进行拟合,并系统分析了模型中关键参数—学习率(表征学习新信息的速度)、折扣率(表征对未来奖励的重视程度)和"逆温度"参数(表征决策的确定性)的动态变化特征。结果表明:鸽子在学习的早期阶段主要采用 MB 策略,侧重于掌握状态之间的关系并形成价值表征;随着经验的积累,逐渐转向 MF 策略,更直接地利用已经获得的价值信息。此外,模型参数分析显示,学习过程中学习率逐渐降低,折扣率逐渐升高,"逆温度"参数也逐渐增大,表明鸽子对未来奖励的关注和决策确定性均随经验的增加而显著提升,体现了生物体在序贯决策学习中从探索环境到利用已有经验的自然转变过程。本研究不仅有助于揭示生物体在复杂环境中如何灵活调整学习策略,还为机器强化学习模型中参数的设置提供了有益的启示。

关键词:序贯决策;强化学习;鸽子;学习策略;Model-Based;Model-Free

中图分类号:Q811.211　　**文献标志码**:B　　**doi**:10.13705/j.issn.1671-6833.2025.05.026

决策是人类及动物在环境信息有限的情况下,为了维持生存而采取有利行动的核心能力,这一过程依赖于生物体与环境之间的交互反馈[1-2]。自然界中,生物体在特定时刻所做的选择会对未来可能遇到的情境产生影响,进而影响长期回报[3-4]。这种需要预测未来的可能情境并据此制定行动计划的决策过程,被定义为序贯决策[5-6]。在这一过程中,生物体需要对决策过程中的行为选项进行价值评估,以决定哪些行为能最大化未来的回报[7]。与大多数机器智体的固定学习策略相比,生物体在价值学习过程中展现出较强的适应性,能够根据不同的情境调整学习策略,从而在复杂环境中优化决策行为[8]。生物体学习策略的演变规律成为行为心理学、神经科学和人工智能等跨学科领域的研究热点。

强化学习(reinforcement learning, RL)被认为是与人类及其他动物学习模式最接近的学习机制,不仅为研究生物体如何学习最佳行动以最大化未来奖励方向提供了有利的工具[9],也受到生物体学习策略的启发,推动了关键算法的创新[10]。关于生物体学习策略的研究,早在 1972 年,Rescorlar 和 Wagner[11]提出的 Rescorla-Wagn er(RW)模型已经解释了经典条件反射的学习现象,该模型阐明了刺激和结果之间的关联如何形成和调整,但未能解释时间上下文以来的刺激中的价值传递现象,缺乏时间敏感性。1988 年,Sutton[12]提出的时序差分学习(Time difference,TD)克服了上述局限,指出生物体通过评估不同动作或状态的预期价值来做决策。在机器 RL 范畴内[13],上述模型使用的学习策略被称

收稿日期:2025-06-10;**修订日期**:2025-07-16

基金项目:国家自然科学基金资助项目(62301496);河南省科技攻关项目(232102210072, 252102210008)

作者简介:李志辉(1978—),女,河南郑州人,郑州大学副教授,博士,主要从事认知行为的脑机制方面的研究,E-mail:lizhrain@zzu.edu.cn。

通讯作者:杨莉芳(1992—),女,河南林州人,郑州大学博士后,主要从事生物认知行为建模,强化学习,神经信号检测与处理方向的研究,E-mai: lifang_yang1014@zzu.edu.cn。

为无模型(model-free,MF)策略,其核心是生物体直接与环境互动,通过试错并寻找最大化累积奖励的策略。2017 年,师黎等[14]构建了动态 Q-Learning 模型,并指出鸽子学习过程中的学习率是动态变化的,但其涉及的任务简单,本质上还是基于奖励预测误差驱动的 MF 策略,不涉及策略的动态演变。而 MF 策略对关系知识不敏感,尤其在环境反馈稀疏或状态空间较大时,需要经历大量的试错迭代,学习过程缓慢,难以适应复杂的序贯决策任务。

进一步的研究表明在序贯决策的学习过程中,生物体不仅采用了基于奖励预测误差驱动的 MF 方式,还使用了基于状态转移知识学习的(model-based,MB)策略[3,15]。MB 策略是通过学习和理解环境中的关系知识,构建一个环境的认知模型用来预测未来情境,并据此规划最佳行为,这一过程被视为一种“深思熟虑”的决策方式。2005 年,Daw 等[15]通过两步决策任务首次证实了人类大脑使用了基于规划推理的 MB 策略。2015 年 Doll 等[16]进一步证明了人类在进行两阶段序贯决策任务时,主要依赖基于模型的推理。2020 年,Momennejad 等[17]指出,在涉及复杂的序贯决策时,生物体不仅需要学习环境的结构,还需形成对未来事件的预测性表征。尽管已有研究揭示了生物体在序贯决策中的多样化学习策略,但当前研究大多侧重于探讨生物体采用了哪种策略模式进行学习,缺乏对学习过程中策略演变的动态观察。针对上述现状与不足,本文主要研究的问题是:当生物体面临一个全新的序贯决策任务时,从最初的探索阶段到最终的稳定利用阶段,其学习策略是否会发生适应性的转变?如果发生转变,其学习策略的演变规律又是如何?

针对上述问题,本文以具有良好认知决策能力的鸽子作为模式动物。首先设计两步序贯决策实验范式,获取鸽子整个学习进程的行为学数据。在此基础上,采用分阶段行为计算建模和模型比较的研究思路,分别构建基于 MB 和 MF 学习策略的动态行为强化学习模型,模拟鸽子整个学习过程。结合模型比较与参数解析,阐明鸽子从初始的随机探索到后期的高效利用过程中学习策略的演变规律。本研究不仅能够加深生物体在序贯决策学习机制的理解,也能为机器强化学习算法中策略设计和参数的动态调整提供启发和生物学习机制的依据。

1 研究方法

1.1 实验范式设计

两步序贯决策任务设计的核心思想是动物当前的行为选择会影响下一步的状态,进而影响未来的奖励,需要同时考虑行为选择和后继状态的关系,以及后继状态与奖励的关系。因此,本文设计的序贯决策任务包含两个基本单元:状态转移单元,对应任务中的第一步(Step 1);奖励单元,即第二步(Step 2)。如图 1 所示,实验的具体操作流程阐述如下。

步骤 1 在鸽子的视野范围内,会同时呈现两个选项标志,即 $S1^+$ 和 $S1^-$,即初始状态 S1,分别通过红色方框与绿色方框进行区分。训练鸽子在选项呈现的 2 秒内选择一个选项,并通过啄击屏幕对应选项下方的按键确认其选择。一旦鸽子完成选择确认,系统将以预设的转移概率 $\boldsymbol{T}_{s,s'}$ 进入后继状态 S2。

步骤 2 在鸽子进入后继状态时,其面前的屏幕上会呈现三角形或圆形后继状态标记,记为 $S2^+$ 或 $S2^-$。每个状态标记都与不同的奖励概率相关联。鸽子需在状态呈现的 2 秒内,通过啄击按键确认其已注意到当前的状态。系统根据后继状态以及奖励概率 $\boldsymbol{P}_r$ 呈现食物奖励 R,$P_r(S2^+) = 0.8$,$P_r(S2^-) = 0.2$,奖励时间为 3 s。奖励结束后,实验将进入一个 5 秒的试次间间隔(Inter-trial Interval,ITI)。

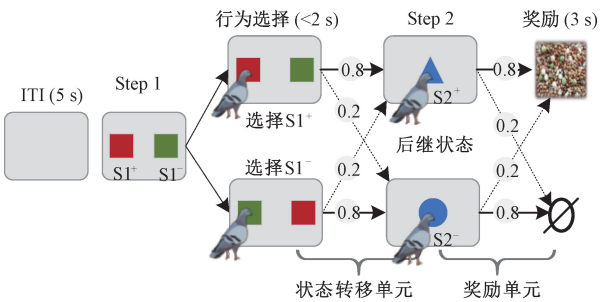

图 1 鸽子序贯决策任务实验范式

Figure 1 The paradigm of the sequential decision task

1.2 实验数据采集

按照上述实验范式,设计了鸽子序贯决策强化学习系统,包括参数设置、数据记录和行为训练箱三部分组成,通过 STM32 单片机为核心的微控制器以及其外围电路实现三者通信。参数设计界面以及数据采集场景如图 2 所示。在此系统下,共训练 4 只鸽子完成了整个序贯决策学习任务。每只被试鸽每天完成 2 个 Session 的实验,每个 Session 预设总试次数为 100。本文所用到的数据信息如表 1 所示。

1.3 动态强化学习模型构建

为了探究鸽子在序贯决策任务中的学习策略演变规律,本文分别基于 MB 和 MF 学习策略,构建了两类动态强化学习模型去模拟鸽子的学习过程:

(a) 实验参数设置界面　　(b)正在参与实验的鸽子

图 2　鸽子序贯决策强化学习训练系统以及训练场景

Figure 2　Reinforcement learning system for sequential decision in pigeons

表 1　所有被试鸽子的数据采集信息统计

Table 1　Statistics of data information for all pigeons

鸽子编号	总 Session	总有效试次数
P090	60	1 517
P025	60	1 374
P014	64	1 875
P021	70	1 987

(1)基于 MF 学习策略。构建了基于奖励预测误差驱动的动态 RW-Q-Learning 模型,其中 RW 模型[18]用来建立奖励与后继状态间的映射,即后继状态 S2 对应的行为价值 $\boldsymbol{Q}_{S2}$,而 Q-Learning 则用于学习 Step 1 中每个选项 S1 的行为价值 $\boldsymbol{Q}_{S1}$。RW 模型中的非条件刺激等同于本实验中的奖励信息,故第 $i+1$ 个试次中步骤 2 中行为价值 $\boldsymbol{Q}_{S2}$ 的更新方式为

$$\boldsymbol{\delta}_{Q_{S2}}(i) = r(i+1) - \boldsymbol{Q}_{S2}(i); \tag{1}$$

$$\boldsymbol{Q}_{S2}(i+1) \leftarrow \boldsymbol{Q}_{S2}(i) + \eta\boldsymbol{\delta}_{Q_{S2}}(i)。 \tag{2}$$

式中:$\boldsymbol{Q}_{S2}(i+1)$ 表示第 $i+1$ 个试次更新后的步骤 2 的行为价值;η 为步骤 2 中的学习率;r 为奖励信息,获得奖励时,$r=1$,反之 $r=0$。S2 到 S1 的价值传递采用 Q-Learning 模型,将步骤 2 的状态值作为反馈,替代了传统贝尔曼方程中“打折后的未来收益 $r+\gamma\max_{a'}\boldsymbol{Q}(s',a')$”部分;第 $i+1$ 个试次中步骤 1 中行为价值更新为

$$\boldsymbol{\delta}_{Q_{S1}}(i) = \boldsymbol{Q}_{S2}(i) + \eta(\boldsymbol{r}(i+1) - \boldsymbol{Q}_{S2}(i)) - \boldsymbol{Q}_{S1}(i),$$

$$\boldsymbol{Q}_{S1}(i+1) \leftarrow \boldsymbol{Q}_{S1}(i) + \alpha\boldsymbol{\delta}_{Q_{S1}}(i)。 \tag{3}$$

式中:α 为 Step 1 中的学习率。上述基于 MF 学习方式计算得到的行为价值记作 $\boldsymbol{Q}_{S1}^{MF}$ 和 $\boldsymbol{Q}_{S2}^{MF}$。

步骤 1 中动作选择采用 SoftMax 策略:

$$\mathbf{P}_{MF}^{j} = \frac{\exp(\beta_{MF}\boldsymbol{Q}_{j}^{MF}(i))}{\sum_{j=1}^{K}\exp(\beta_{MF}\boldsymbol{Q}_{j}^{MF}(i))}, \tag{4}$$

式中:K 为选项个数;j 为选择;i 为试次数;β_{MF} 是控制选择中随机性水平的“逆温度”参数,取值范围为 0(完全随机响应)和 ∞(完全选择价值最高的选项)。

最终的模型参数为:$\theta_{MF} = (\alpha, \beta_{MF}, \eta)$。

(2)基于状态转移知识的 MB 学习策略。借鉴后继表征(successor representation, SR)[19]方法构建了一种基于 MB 学习策略的动态 RL 模型,使用奖励函数和状态转移概率估计价值。模型使用两个结构计算价值:奖励期望 $\boldsymbol{V}_S$ 和状态转移矩阵 $\boldsymbol{T}_{S,S'}$。

状态 s' 对应的奖励期望 $\boldsymbol{V}_{s'}$ 计算公式为:

$$\boldsymbol{V}_{S'} = \boldsymbol{E}[\gamma\boldsymbol{I}(r=1) \mid S0 = S']。 \tag{5}$$

状态转移矩阵 $\boldsymbol{T}_{s,s'}$ 是基于经验频率估计的。具体而言,对于每个 Session 或每个分析窗口,通过计算从步骤 1 的每个选项(状态 S0)转移到步骤 2 的每个后继状态(状态 S1 或 S2)的频率来估计状态转移概率。状态转移矩阵 $\boldsymbol{T}_{s,s'}$ 计算公式为

$$\boldsymbol{T}_{s,s'} = \boldsymbol{E}[\sum_{t=0}^{\infty}\gamma^{t}\boldsymbol{I}(S_t = S') \mid S0 = S]。 \tag{6}$$

式中:$\boldsymbol{E}$ 表示期望值;$\boldsymbol{I}(\cdot)$ 为指示函数,如果其参数为真则为 1,反之为 0;S_t 是在状态 $S0$ 后 t 个时间步遇到的状态;γ 为折扣参数,表示对未来状态转移的打折程度。由于本文的研究内容从初始状态 s 到最终状态 s' 只有一步,因此设置 $t=1$。

然后计算 $\boldsymbol{R}_{s'}$ 和 $\boldsymbol{T}_{S,S'}$ 的内积,得到试次 i 中状态 s 的价值 $\boldsymbol{Q}_S$:

$$\boldsymbol{Q}_S(i) = \sum_{S'}\boldsymbol{T}_{S,S'} \times \boldsymbol{V}_{S'}。 \tag{7}$$

对应于本文,Step 1 的行为价值更新如下式:

$$\boldsymbol{Q}_{S1}(i) = \sum_{S2}\boldsymbol{T}_{S1,S2} \times \boldsymbol{V}_{S2}(i)。 \tag{8}$$

Step 2 的行为价值 $\boldsymbol{Q}_{S2}$ 对应 S2 的奖励期望 $\boldsymbol{V}_{S2}$:

$$\boldsymbol{Q}_{S2}(i) = \boldsymbol{V}_{S2}(i)。 \tag{9}$$

上述基于 MB 学习策略计算得到的行为价值记为 $\boldsymbol{Q}_{S1}^{MB}$ 和 $\boldsymbol{Q}_{S2}^{MB}$。

动作选择同样采用 SoftMax 策略:

$$\mathbf{P}_{MB}^{j} = \frac{\exp(\beta_{MB}\boldsymbol{Q}_{j}^{MB}(i))}{\sum_{j=1}^{K}\exp(\beta_{MB}\boldsymbol{Q}_{j}^{MB}(i))}。 \tag{10}$$

MB 的模型参数:$\theta_{MB} = (\beta_{MB}, \gamma)$。

1.4　模型参数拟合

使用最大后验概率估计 *MAP* 方法估计上述模型的参数。*MAP* 结合了最大似然估计和先验概率的信息来估计模型参数。假设有观测数据集合 $D = \{d_1, d_2, \cdots, d_n\}$,通过调整参数 θ,使得在给定数据集合 D 条件下,观测数据集合 D 的概率 $P(\theta \mid D)$ 最大。计算公式如下:

$$MAP = -\log(P(\theta_M \mid D, M))\alpha - \log(P(D \mid M, \theta_M) - \log(P(\theta_M \mid M))。 \tag{11}$$

式中:$P(\theta_M|D,M)$ 是给定行为数据 D 和时参数 θ_M 的可能性,$P(D|M,\theta_M)$ 是给定当前模型 M 及其参数 θ_M 时行为数据 D(即选择序列)的可能性,$P(\theta_M|M)$ 是参数的先验概率。

为确保广泛的先验,同时防止极端参数估计,本文对待估计参数的先验分布设置如下:

$$\theta(\alpha_{\text{prior}}) \sim Beta(1.1,1.1);$$

$$\theta(\beta_{\text{prior}}) \sim Gamma(1.2,5.0);$$

$$\theta(\gamma_{\text{prior}}) \sim Beta(1.1,1.1);$$

$$\theta(\omega_{\text{prior}}) \sim Beta(1.1,1.1)。$$

1.5 模型比较

使用基于贝叶斯信息准则 *BIC* 对模型进行比较。*BIC* 同时考虑模型拟合优度和复杂度,通过最大化数据的似然函数,同时惩罚模型的参数数量来平衡模型的拟合能力和复杂度。*BIC* 计算公式为

$$BIC = -2\ln L + k\ln n。 \quad (12)$$

式中:L 为模型的似然函数值,k 是模型的参数数量,n 是数据点的数量。公式中的 $k\ln n$ 表示对参数数量进行惩罚,防止过度拟合,BIC 的值越小,表示模型的拟合性能越好,对应于本文研究,*BIC* 值越小的模型越能够反映鸽子真实的学习策略。

2 结果分析

2.1 行为结果

共采集了 4 只鸽子的行为数据,计算了每个 Session 中鸽子的最优动作选择率和获奖率。行为表现如图 3 所示。

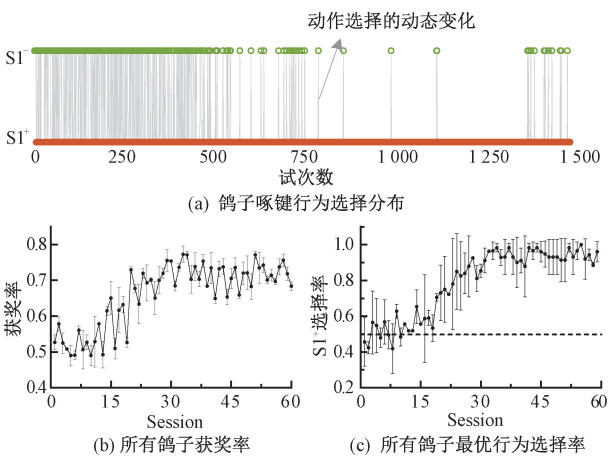

图 3 序贯决策学习进程中受试鸽的行为表现.

Figure 3 Behavioral performance of subject pigeons during the learning process of sequential decision making

从图 3(a)中可以看出,学习前期鸽子在 S1+和 S1-两个选项间频繁切换,学习后期则在绝大多数试次中选择 S1+,表明鸽子能够通过不断的试错探索,习得最优的动作选择。进一步统计了所有受试鸽 S1+选择率与获奖率,结果图 3(b)和(c),横坐标为 Session。从图中可以看出,早经过约 20 个 Session 后,鸽子最优动作选择率和获奖率逐步增加,第 30 个 Session 时表现稳定,获奖率达预设水平。综上,鸽子经过学习掌握了最优行为策略。

2.2 模型比较与学习策略演变分析

为系统揭示鸽子在序贯决策任务中学习策略的动态演变规律,本文将鸽子的整个学习进程划分为若干连续的窗口,每个窗口包含 200 个试次,相邻窗口间重叠率设为 50%。基于前文建立的 MF 和 MB 强化学习模型框架,对每个时间窗口内鸽子的行为数据进行了单独的参数拟合和模型拟合优度分析,通过比较两个模型在每个窗口中的 BIC 值,分析鸽子在不同学习阶段更倾向于采用哪种学习策略。所有鸽子在不同学习阶段的模型比较结果如图 4 所示。

从图 4 可以看出,所有鸽子在学习初期均呈现出 MB 策略的拟合效果显著优于 MF 策略的趋势。这一现象表明,鸽子在初始探索环境阶段更倾向于通过显式地学习和掌握环境中各状态之间的关系,从而建立起环境状态转移的内部认知模型,以此形成更准确的价值预期。随着经验的逐渐积累,模型比较的趋势逐渐发生逆转,在学习后期阶段,所有鸽子的 MF 策略表现明显优于 MB 策略。表明鸽子开始更为直接地依赖已经形成的价值预测,以快速且高效地做出决策。

根据上述 BIC 比较结果,选择每个时间窗口内最优模型,比较了鸽子在学习过程中真实行为选择和模型估计的选择,结果如图 5 所示。图 5 中虚线为 50 次运行结果的平均值,阴影为标准差。横坐标是按照学习时间划分的时间窗口,每个窗口内含 200 个试次。从图 5 中可以看出,动态强化学习模型能够很好地捕获所有鸽子的真实选择行为,且即使鸽子的行为表现出现突然波动的情况下,RL 模型依然能够很好的跟随鸽子行为的变化。

2.3 学习参数变化规律分析

为精细探究鸽子在学习过程中反应其学习特征的学习率 α、“逆温度”参数 β 和折扣率 γ 的变化规律,本小节首先使用一元线性回归模型分别分析了 MF 策略下学习率 α、“逆温度”参数 β 与鸽子优势行为选择率的相关性,结果如图 6 所示。

从图 6 可以看出,所有鸽子的学习率与其行为选择的正确率呈现出显著的负相关性,并且逐渐趋

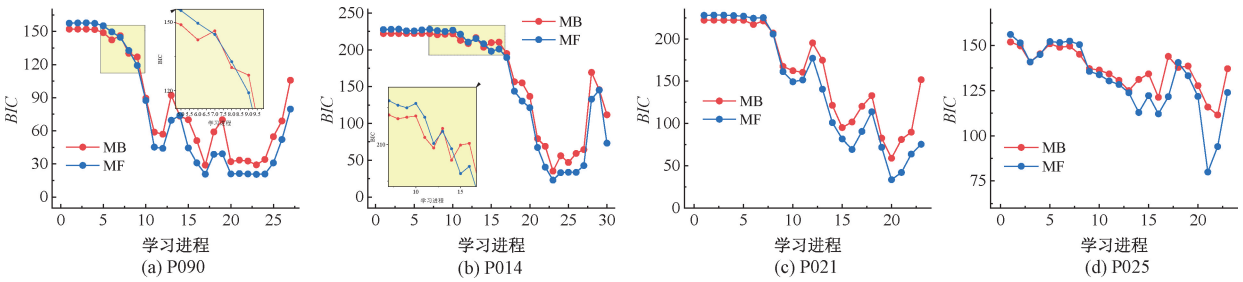

图 4　序贯决策学习进程中两种学习模型的 *BIC* 值比较

Figure 4　Comparison of BIC values of two RL models in sequential decision learning process

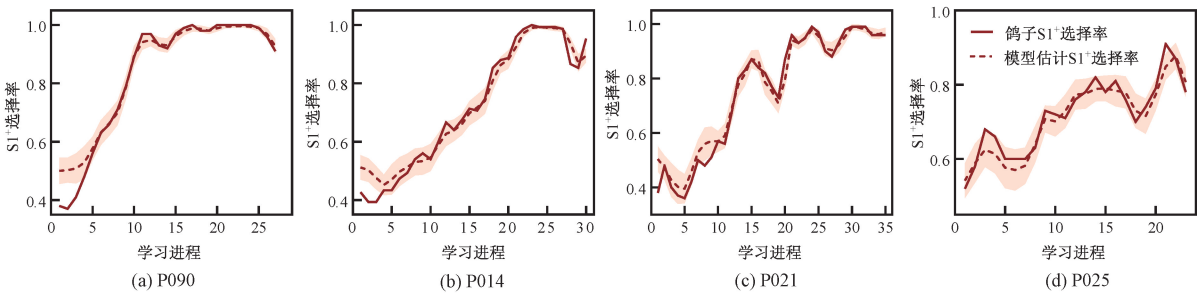

图 5　鸽子在学习过程中真实行为选择和模型估计结果

Figure 5　Behavioral choices of pigeons and RL model estimation results during learning process

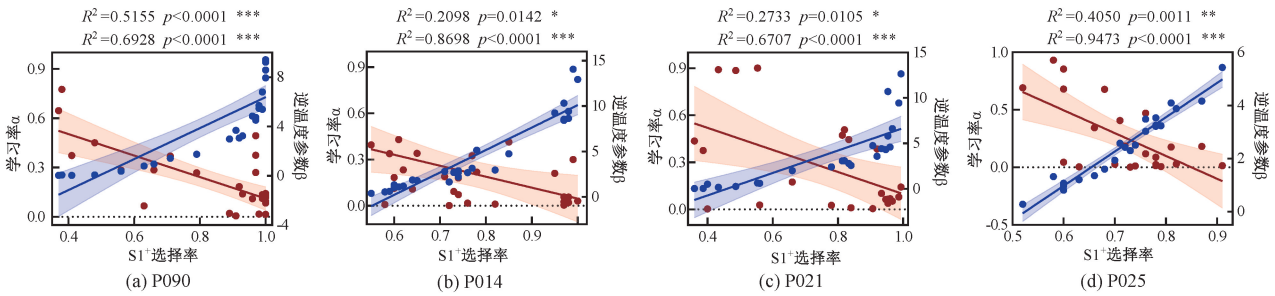

图 6　优势动作选择率与 MF 学习模型参数回归结果

Figure 6　Regression results of pigeon dominant action selection rate with MF learning model parameters

于零。这表明在学习早期阶段,鸽子更倾向于利用奖励预测误差信息来更新其价值判断;而在学习的后期,对预测误差的依赖明显减少。同时控制鸽子行为选择水平的“逆温度”参数 β 与行为选择的正确率之间表现出显著的正相关,表明鸽子逐渐从探索未知转向利用已知信息的行为模式。

进一步分析了鸽子基于 MB 策略时表征对未来奖励重视程度的折扣率 γ、“逆温度” 参数 β 与最优动作选择率的相关性,结果如图 7 所示。

从图中可以看出,4 只受试鸽的行为表现与逆温度参数均呈正相关,表明鸽子行为逐渐发生了从探索到利用的转变。MB 学习策略中的折扣率 γ 与最优动作选择率表现出正相关,且随着鸽子逐渐习得最优策略,γ 逐渐趋于 1,结果表明鸽子在做选择时更多地考虑了长远的回报,而不仅仅是短期的奖励。上述结果凸显了鸽子在学习过程中学习策略变化的灵活性。这种策略的调整有助于最大化长期利益,而不仅仅是追求短期回报。

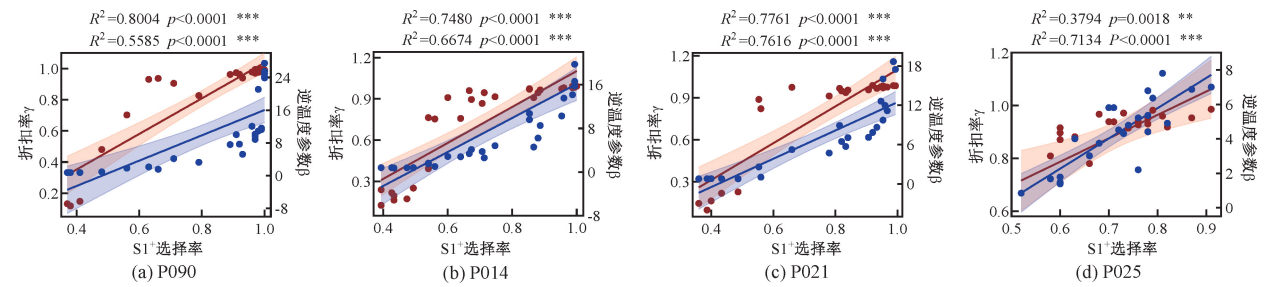

图 7　优势动作选择率与 MB 学习模型参数回归结果

Figure 7　Regression results of dominant action selection rate of pigeons and MB learning model parameters

3 结论

为探究鸽子在序贯决策学习中学习策略的演变规律,本文分别构建了基于 MB 和 MF 学习策略的动态强化学习模型,并对鸽子的学习过程进行了模拟。研究结果表明,在学习初期,鸽子主要依赖 MB 策略,通过学习状态之间的关系形成价值表征;随着学习的深入和经验的积累,逐渐转向 MF 策略,更多地依赖奖励预测误差来指导行为,表现出对已学价值的利用。进一步的参数分析还显示,反映鸽子对新信息敏感度的学习率逐渐减小、反映对未来长期回报重视程度的折扣率则逐渐升高。这些结果反映了鸽子能够根据经验和环境反馈灵活地调整其学习策略。

与传统多侧重于探讨生物体在特定阶段主要采用哪种策略的研究相比,本研究的创新之处在于,以动态视角观察鸽子从探索到利用的完整学习过程。通过将学习历程分段并结合滑窗拟合策略参数,得到了鸽子在不同阶段对 MB 与 MF 策略的依赖度的变化规律,从而更准确地刻画了其学习策略的动态演变过程。这一发现与 Daw 等[20]近期关于人类在两步决策任务中学习策略随时间动态演变的研究结果高度吻合,支持了“生物体的学习策略并非固定不变,而是能够根据学习进程和经验积累进行动态调整”这一观点。研究结果不仅深化了对生物体在序贯决策中如何灵活调整策略的认识,也为在机器强化学习中引入动态策略与自适应参数设定提供理论支持与生物学习依据。

参考文献:

[1] DAYAN P, DAW N D. Decision theory, reinforcement learning, and the brain[J]. Cognitive, Affective, & Behavioral Neuroscience, 2008, 8(4): 429~453.

[2] GUPTA N, AHIRWAL M K, ATULKAR M. Development of human decision making model with consideration of human factors through reinforcement learning and prospect utility theory[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2024, 36(7): 1003-1019.

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

[4] MILLER K J, VENDITTO S J C. Multi-step planning in the brain[J]. Current Opinion in Behavioral Sciences, 2021, 38: 29-39.

[5] DEHAENE S, SIGMAN M. From a single decision to a multi-step algorithm[J]. Current Opinion in Neurobiology, 2012, 22(6): 937-945.

[6] 张倩倩. 面向人机序贯决策的混合智能方法研究[D]. 合肥: 中国科学技术大学, 2021.
ZHANG Q Q. Research on hybrid intelligent method forman-machine sequential decision-making[D]. Hefei: University of Science and Technology of China, 2021.

[7] MATTAR M G, THOMPSON-SCHILL S L, BASSETT D S. The network architecture of value learning[J]. Network Neuroscience, 2018, 2(2): 128-149.

[8] 王东署, 杨凯. 基于状态转移学习的机器人行为决策认知模型[J]. 郑州大学学报(工学版), 2021,42(6): 7-13.
WANG D S, YANG K. Behavior decision-making cognitive model of mobile robot based on state transfer learning [J]. Journal of Zhengzhou University (Engineering Science), 2021,42(6): 7-13.

[9] 蒲慕明. 跨学科开启头脑风暴 促进脑科学交叉与融合[J]. 科学通报, 2023, 68(35): 4749-4750.
PU M M. Initiate interdisciplinary brainstorming, promotecross-disciplinary integration in neuroscience[J]. Chinese Science Bulletin, 2023, 68(35): 4749-4750.

[10] Huang J, Zhang Z, Ruan X. An improved dyna-Q algorithm inspired by the forward prediction mechanism in the rat brain for mobile robot path planning[J]. Biomimetics, 2024, 9(6): 315.

[11] Rescorla R A. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement[J]. Current Research & Theory, 1972, 64-99.

[12] SUTTON R S. Learning to predict by the methods of temporal differences[J]. Machine Learning, 1988, 3(1): 9-44.

[13] 李琳, 李玉泽, 张钰嘉, 等. 基于多估计器平均值的深度确定性策略梯度算法[J]. 郑州大学学报(工学版), 2022, 43(2): 15-21.
LI L, LI Y Z, ZHANG Y J, et al. Deep deterministic policy gradient algorithm based on mean of multiple estimators[J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(2): 15-21.

[14] 师黎, 陶梦妍, 李志辉. 鸽子强化学习过程中内部学习状态的动态建模研究[J]. 科学技术与工程, 2017, 17(13): 120-125.
SHI L, TAO M Y, LI Z H. Dynamic modeling of internal cognitive status of pigeon in the process of reinforcement learning [J]. Science Technology and Engineering, 2017, 17(13): 120-125.

[15] DAW N D, NIV Y, DAYAN P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control[J]. Nature Neuroscience, 2005, 8(12): 1704-1711.

[16] DOLL B B, DUNCAN K D, SIMON D A, et al. Model-based choices involve prospective neural activity[J]. Nature Neuroscience, 2015, 18(5): 767-772.
[17] MOMENNEJAD I. Learning structures: predictive representations, replay, and generalization[J]. Current Opinion in Behavioral Sciences, 2020, 32: 155-166.
[18] ESBER G R, SCHOENBAUM G, IORDANOVA M D. The rescorla-Wagner model: it is not what you think it is [J]. Neurobiology of Learning and Memory, 2025, 217: 108021.
[19] YANG L F, JIN F L, YANG L, et al. The hippocampus in pigeons contributes to the model-based valuation and the relationship between temporal context states[J]. Animals, 2024, 14(3): 431.
[20] VENDITTO S J C, MILLER K J, BRODY C D, et al. Dynamic reinforcement learning reveals time-dependent shifts in strategy during reward learning[J]. bioRxiv, 2024: 2024.02.28.582617.

Dynamic Reinforcement Learning Modeling and Strategy Evolution in Pigeon Sequential Decision-making

LI Zhihui[1,2], MA Ying[1,2], SHANG Zhigang[1,2], YANG Lifang[1,2,3*]

(1. School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. Henan Key Laboratory of Brain Science and Brain Computer Interface Technology, Zhengzhou 450001, China; 3. The Affiliated Encephalopathy Hospital of Zhengzhou University, Zhumadian 463000, China)

Abstract: To maximize future rewards, organisms must flexibly adjust their learning strategies within complex environments. To investigate how learning strategies dynamically evolve during sequential decision-making, we used pigeons—a model species with robust cognitive capabilities—in a two-step sequential decision-making task. Behavioral data were collected throughout the entire learning process, from initial exploration to proficient performance. We developed two dynamic reinforcement learning (RL) models: a reward prediction error-driven Model-Free (MF) model and a state-transition relationship-driven Model-Based (MB) model. Using experimental data, we fitted these models and systematically analyzed the dynamic changes in key learning parameters, including learning rate (reflecting the speed of new information acquisition), discount factor (indicating the valuation of future rewards), and the inverse temperature parameter (representing choice certainty). Model comparisons revealed that pigeons predominantly utilized an MB strategy in early learning stages, focusing on acquiring relationships between states to form accurate value representations. With accumulated experience, pigeons progressively shifted toward the MF strategy, directly utilizing established value predictions for decision-making. Furthermore, analysis of model parameters showed that the learning rate gradually decreased, while both discount factor and inverse temperature increased over the learning period. These changes indicate that pigeons progressively place greater emphasis on future rewards and decision certainty, illustrating a natural shift from environmental exploration to exploitation of acquired knowledge. This study not only elucidates the mechanisms underlying adaptive strategy adjustments in biological systems during sequential decision-making but also provides valuable biological insights for parameter optimization in artificial reinforcement learning models.

Keywords: sequential decision-making; reinforcement learning; pigeons; learning strategies; Model-Based; Model-Free