真实物理系统中的奖励函数

作者：周新　邮箱：inksci@qq.com

（墨之科技，人工智能自动化）

空间机械臂动力学系统属于多自由度时变非线性系统，通过控制方程的求解相对复杂并且运算量大，因此对于实时性的要求难以满足，并且在逆运动学的求解中可能出现奇异矩阵以及解的不唯一。而使用增强学习实现的控制仅通过定义奖励函数对机械臂完成任务进行策略优化引导，而不需要全程设计机械臂完成任务的流程和动作。机械臂通过动作试错，具有一定程度的策略学习能力，从而使人类在应对多自由度百线性控制任务的工作中变得更轻松。

使用奖励回报函数来告诉机械臂需要做什么，而不关心机械臂具体如何去做，这相对容易。增强学习依赖于马尔可夫决策模型，智能体与环境的交互信息被抽象成状态、动作、奖励 3 个要素。即智能体感知环境的状态并向环境中施加动作，与此同时，接收来自环境中的奖励反馈。智能体的学习即是采取更好的动作以获得更大的奖励。可见，奖励函数对于智能体完成任务有着关键的作用。奖励函数通常是对当前环境的状态与智能体采取的施加于环境的相应动作给出一个评价值，如式(1.1)。

$$r \leftarrow r(s,a) \tag{1.1}$$

状态可以充分地描述机械臂当前所处的环境（不仅包括机械臂所处的外界环境，也包括机械臂自身的构型），如视觉图像、机械臂的关节角度和角速度。动作能够使机械臂在指定的自由度范围内到达所有状态，动作可以是机械臂所有关节的控制力矩、角加速度等。然而相比较而言，奖励函数的设计却不容易实现。不同的奖励函数将引导机械臂完成不同的任务。甚至奖励函数中的参数对机械臂完成任务也有相应的影响。

在空间机械臂的控制任务中，奖励函数中通常需要考虑距离，即被操纵物体到空间机械臂执行器的距离 d 。一种最简形式的奖励函数定义如式(1.2)，奖励 r 为距离的惩罚项，机械臂通过关节角的运动使执行器移动到被操纵物体附近从而减少距离来提高奖励。

$$r = -d \tag{1.2}$$

在训练阶段中，奖励信息是不可缺少的，在每一个控制节拍中，智能体施加给环境动作并接收相应的新的状态和即时奖励。增强学习算法根据环境提供的奖励信息对策略进行优化，不同的奖励函数将引导策略不同的优化方向。因此在训练阶段，需要准确地测量被操纵物体的位置信息。

在真实物理系统中奖励函数的设计是一个难题。因为距离是奖励函数中重要的一项，为了知道距离需要知道被操纵物体的三维位置。在真实的物理环境

中，为了知道物体的三维位置可以使用双目相机进行特征点三维定位，然而相机的标定过程和图像平面上特征点的提取是一个充满噪声和误差的工作。使用标定后的相机对物体进行三维位姿确定是一项复杂的工作，而且精度和稳定性均比较差。例如当摄像机距离被操纵物体较远时，一个特征点像素级的提取误差可能会导致厘米级的测量误差或者更大，这样量级的误差不利于实现空间机械臂对物体进行精细的操作。

尽管增强学习依赖的奖励函数中包含了距离信息的测量，但在增强学习中实现了训练和应用的分离。策略神经网络的使用分为训练阶段和应用（测试）阶段，如图 1，奖励信息仅在训练阶段中使用而在测试阶段不需要环境提供的任何奖励信息。

因此在训练阶段，应当综合借助高精度测量仪器，尽可能充分地了解当前环境所处的状态信息，使设计的奖励函数尽量合理从而引导策略的优化方向。然而在测试（应用）阶段，应当仅获取那些容易测量得到的且满足任务操作需要的信息，从而不依赖于高精度的测量设备，实现策略神经网络在多种应用场景中的普遍可行性，降低控制系统的成本。

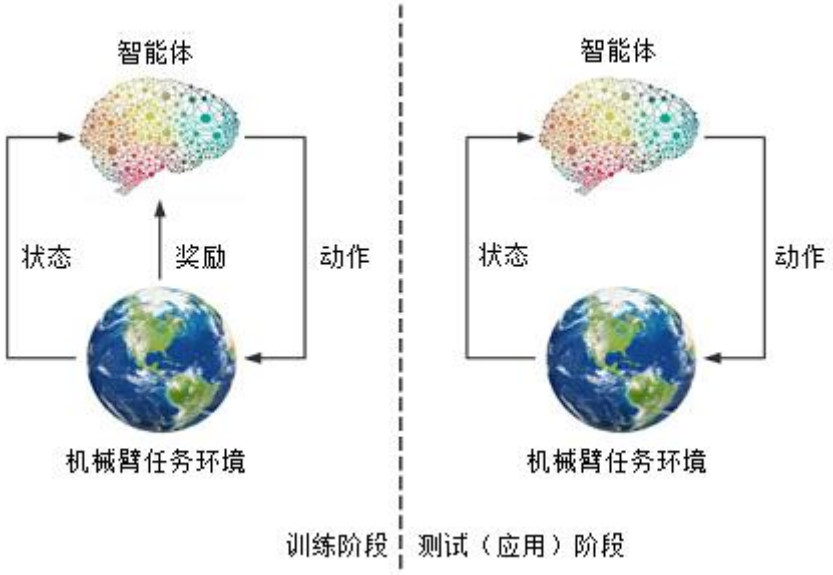

图 1 增强学习中的训练阶段和测试（应用）阶段

为了实现被操纵物体的位置测量，可以将被操纵物体固定在另一个可运动的机械臂上，根据这个机械臂相应的关节角测量值来计算被操纵物体的位置。从而实现了真实物理系统中位置信息以及距离值的精确获取。在文献[1]的训练过程中，通过将被操纵物体装夹于 PR2 机器人的左手中，从而可以方便地知道被操纵物体的相应位置信息。而在应用中，被操纵物体是放置于桌面的任意位置，物体的位置坐标是不可以直接得知的。

[1] Levine S, Finn C, Darrell T, et al. End-to-End Training of Deep Visuomotor Policies[J]. Journal of Machine Learning Research, 2015, 17(1):1334-1373.