Exemplar-Based Direct Policy Search with Evolutiona Optimization

Kokolo IKEDA
Academic Center for Computing and Media Studies, Kyoto University
Yoshida-nihonmatsu-cho, Sakyo-ku, Kyoto-city, Japan
kokolo@media.kyoto-u.ac.jp

【强化学习 79】Exemplar

张楚珩
清华大学 交叉信息院博士在读

比较老的一篇 Exemplar-based，结合了演化算法。

原文传送门

Ikeda, Kokolo. "Exemplar-based direct policy search with evolutionary optimization." 2005 IEEE Congress on Evolutionary Computation. Vol. 3. IEEE, 2005.

特色

使用一个 exemplar 的集合来表示 policy，并且使用遗传算法来更新策略。

过程

1. 优势

Exemplar-based 的优势在于，策略的表示能力较强，同时对于迁移学习、模仿学习比较友好。

2. 做法

一个 exemplar set E 再加上一个 case-based action selector D 就能够确定一个策略，每当遇到一个状态时，就使用 D 基于 E 来选择一个行动。算法维护多个这样的策略，并且对这些策略做遗传算法挑选出较好的策略。整体算法如下图所示。

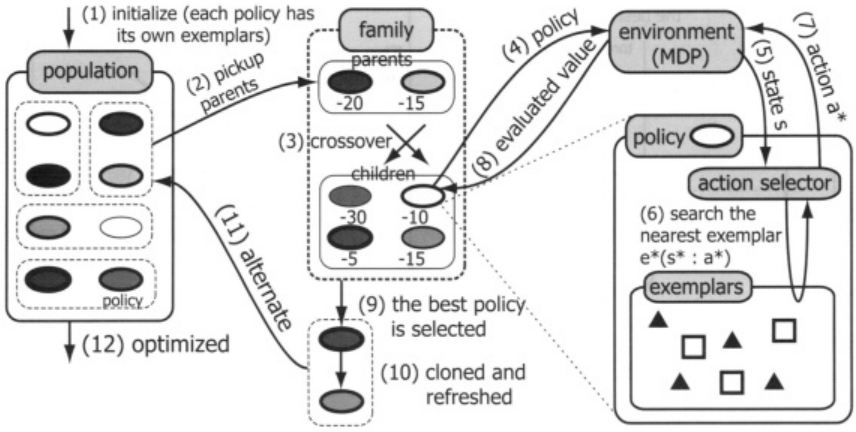

Figure 1: Basic procedure of the SAP-GA method. (1) Given a problem, the population is initialized. (2) In a generation, two solutions (parents) are picked for alternation. (3) Several children are produced by the crossover operator. (4) A policy, a solution in the family, is evaluated in the environment of the target problem. (5) For each step, the observation of state s is sent to the action selector of the policy. (6) The nearest exemplar (s^*, a^*) to s is selected. (7) The action a^* is taken. (8) At the end of the MDP, the evaluated value of the policy is calculated and returned. (9) The best policy in the family is selected. (10) The best policy is cloned and the refresh operator is applied. (11) Two solutions, the best solution and the refreshed solution, are set back to the population instead of the parents. (12) The policy is optimized.

其中，种群中的策略都是成对的，每次取一对策略 (2)，对其进行交叉 (3)，交叉的过程很简单，就是从两个策略的 exemplar set 中各选取一定的 exemplar 来组成新的 exemplar set，其中需要注意不要挑选到重复的 exemplar。这样就能够产生若干个子策略，接下来就需要对这些子策略进行 rollout 并测试其性能。rollout 中比较关键的是如何确定 exemplar 的形式和 action selector $_D$ (5-7)。文章中使用 state-action pair 来作为 exemplar，action selector 就是找到 exemplar 状态和当前状态最为相近的状态，然后选中其中的 action 来作为将要采取的行动。即

Nearest-neighbor action selector

1. The current state s and exemplars $\{e_j\}$ (i.e., state-action pairs $\{(s_j, a_j)\}$) are given.
2. For each s_j, the distance $d(s, s_j)$ is measured.
3. The nearest state s_{j*} is selected.
4. The action a_{j*} is selected and taken.
5. The referred exemplar e_{j*} is marked for the refresh operator.

策略评判 (8) 的标准与标准强化学习类似，比较 discounted return 即可。在所产生的子策略中，选取最好的策略作为策略对的一个策略 (9)，另一个策略为该策略的复制然后在进行一定的变异 (10)，即图中注明的 refresh。在该操作中，以一定概率选中测试时没有使用到的 exemplar，然后随机重新初始化这些 exemplar，以起到变异的效果。

3. 实验

做了两个比较简单的基于物理系统的实验，一个是 acrobat，一个是 parallel double inverted pendulum。效果与 Q-learing 对比。

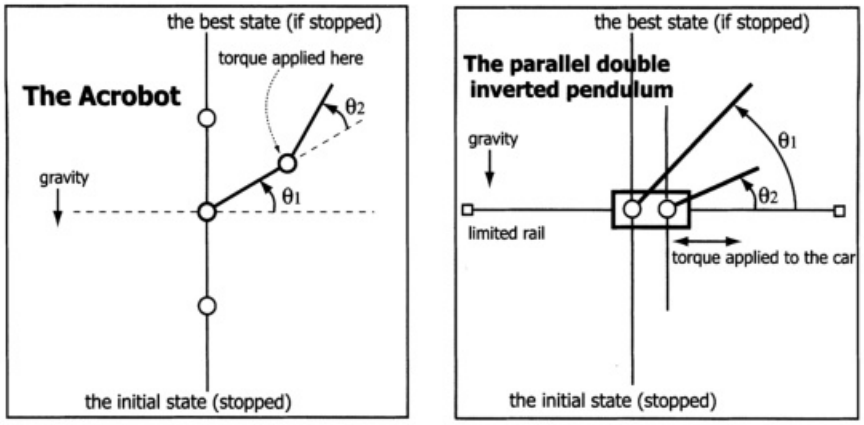

Figure 2: The Acrobot (left) and the PDIP (right)

发布于 2019-07-10

强化学习 (Reinforcement Learning)

文章被以下专栏收录

强化学习前沿

读呀读paper

进入专栏