Policy Search in Continuous Action Domains: an Overview

Olivier Sigaud

INRIA Bordeaux Sud-Ouest, équipe FLOWERS
Sorbonne Université, CNRS UMR 7222,
Institut des Systèmes Intelligents et de Robotique, F-75005 Paris, France
olivier.sigaud@isir.upmc.fr +33 (0) 1 44 27 88 53

Freek Stulp

German Aerospace Center (DLR), Institute of Robotics and Mechatronics, Wessling, Germany
freek.stulp@dlr.de

【强化学习 55】连续控制策略搜索

张楚珩
清华大学 交叉信息院博士在读

一篇综述，把里面有意思的东西摘录一下。

原文传送门

Sigaud, Olivier, and Freek Stulp. "Policy search in continuous action domains: an overview." Neural Networks (2019).

特色

连续控制方向的策略搜索（policy search）综述，该综述比较新，不仅涵盖了一些比较古老的方法而且也包括了最新的很多强化学习方法；同时也提出了该领域的一些重点问题。连续控制问题可以看做是半个强化学习领域了，即它不包括一些动作空间离散的情况，即DQN一类的方法不在此讨论范围内。

过程

一、总览

文章对于所有的相关方法进行了一个分类。

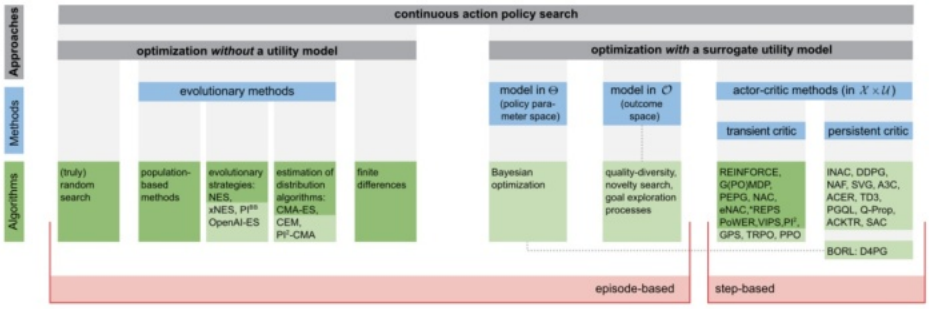

Figure 2: Simplified classification of the algorithms covered in the paper. Θ is the space of policy parameters (see Section 2), O is an outcome space (see Section 4) and $\mathcal{X} \times \mathcal{U}$ is the state and action space (see Section 5). Algorithms not covered in (Deisenroth et al., 2013) have a lighter (green) background. References to the main paper for each of these algorithms is given in a table in the end of each section. From the left to the right, algorithms are grossly ranked in order of increasing [illegible] reuse, but methods using a utility model in Θ and O show better sample choice, resulting in competitive sample efficiency.

策略搜索问题中有带参数 θ 的策略，以及策略在环境中的表现（utility） $J(\theta)$ 。首先分为两类，一类是without utility model，即不需要对 $J(\theta)$ 进行建模，只是不断尝试不同的参数，然后找到能够获得更大utility数值的参数；另一类需要对这个函数进行建模。

二、随机搜索、演化算法和有限差分

这三类都是不需要对于utility函数进行建模的方法。这类方法需要对于整个轨迹进行采样然后仅仅获取一个utility数值，这种方式不能有效利用数据，数据利用率较低；但是其优点在于能够高度并行地进行计算，在计算资源充足的情况下更有优势。

随机搜索很直接，就是找不同的参数，然后去做rollout，找出utility数值最高的参数。

演化算法包括population based、evolutionary strategy （ES）和 estimation of density algorithm（EDA）。其主要的区别一幅图就能说清楚。

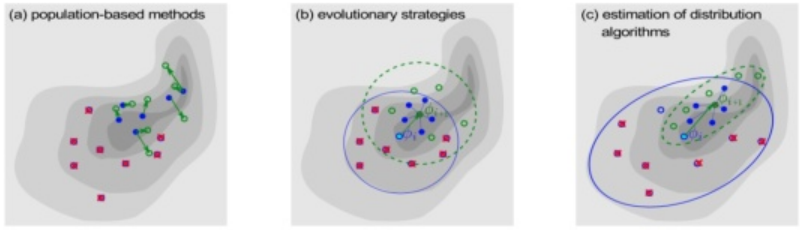

Figure 3: One iteration of evolutionary methods. (a) Population-based methods (b) Evolutionary Strategies (c) EDAs. Blue: current generation and sampling domain. Full blue dots: samples with a good evaluation. Dots with a red cross: samples with a poor evaluation. Green: new generation and sampling domain, empty dots are not evaluated yet. Red dots: optimum guess. In population-based methods, the next generation are offspring from several elite individuals of the previous generation. In ES, it is obtained from an optimum guess and sampling from fixed Gaussian noise. In EDAs, Gaussian noise is tuned using Covariance Matrix Adaptation.

有限差分方法也是在参数空间采样多个点，然后做rollout，并得到相应的utility数值；与前面不同的是，该方法会基于相近的多个点，计算出这些参数附近的有限差分来近似梯度方向，顺着梯度方向来进行搜索。

三、Bayesian optimization 和 directed exploration method

贝叶斯优化（Bayesian optimization）的方法是在参数空间根据Bayesian的原理确定一个值得尝试的参数，然后做rollout获得utility值。由于它能够主动地去尝试新的参数，其采样效率更高一些。不过该方法scale up的可能性还是比较低。

有向探索方法（directed exploration method）分为两步，先在策略参数空间找出参数使其产生的结果能够覆盖所定义的一个outcome space o （比如先找到一堆策略能让一个机器人处于不同的姿态），由于utility的定义与outcome space是密切相关的，因此下一步就比较容易找到好的策略参数。这类方法我读的比较少，但感觉还挺有意思的，它不仅能够更好地解决探索的问题，而且应该可以用来做meta-learning。

四、Actor-critic方法

这是目前模型policy search方法里面效果最好的一类方法，也是近年来研究十分多的方法。之前的方法都是基于MDP的一整个轨迹（episode-based）来对utility建模或者获得估计值的，而这里把轨迹拆成了一步步，基于每一步（step-based）来对utility做估计（即RL里面讲的价值函数）。这样的做法对于数据的利用效率更高。

文章把这一类方法分为了两大类，一类是transient critic，即学习到的critic每次迭代之后都会重新学习，这类方法一般直接使用Monte Carlo样本或者用Monte Carlo样本来做目标训练critic网络；另一类是persistent critic，这类方法每次迭代学习到的critic网络都基于之前的critic网络，即使用TD-learning（bootstrap）方式类学习。后者效率更高，前者更稳定并且无偏。

下表可以作为强化学习中近年来比较优秀算法的一个列表。

Algorithm	Main paper
REINFORCE	(Williams, 1992)
G(PO)MDP	(Baxter and Bartlett, 2001)
NAC	(Peters and Schaal, 2008a)
eNAC	(Peters and Schaal, 2008a)
POWER	(Kober and Peters, 2009)
PI^2	(Theodorou et al., 2010)
REPS	(Peters et al., 2010)
PEPG	(Sehnke et al., 2010)
VIPS	(Neumann, 2011)
iNAC	(Bhatnagar et al., 2007)
GPS	(Levine and Koltun, 2013)
TRPO	(Schulman et al., 2015)
DDPG	(Lillicrap et al., 2015)
A3C	(Mnih et al., 2016)
NAF	(Gu et al., 2016b)
ACER	(Wang et al., 2016b)
Q-PROP	(Gu et al., 2016a)
PGQL	(O'Donoghue et al., 2016)
PPO	(Schulman et al., 2017)
ACKTR	(Wu et al., 2017)
SAC	(Haarnoja et al., 2018)
TD3	(Fujimoto et al., 2018)
D4PG	(Barth-maron et al., 2018)

Table 4: Main reinforcement learning algorithms. Algorithms below the line have not yet been covered in (Deisenroth et al., 2013).

五、未来方向

文章提出该方向最终关心的问题主要是1）算法能够得到的最终性能（performance）；2）算法的稳定性（stability）；3）算法的采样效率（sample efficiency）。文章提出

- 目前对于性能的比较比较多，但是也应该更多的比较稳定性和采样效率，并且仔细分析算法的自身性质；
- 对于不同种类方法的组合可能产生更有效的算法；
- lifelong learning、continual learning、open-ended learning、multitask learning、hierarchical RL、meta-learning、representation learning和transfer learning等方向。

发布于 2019-04-22

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