N SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL. 45, NO. 3, MARCH 2015

Iultiobjective Reinforcement Learnin A Comprehensive Overview

inming Liu, Xin Xu, *Senior Member, IEEE*, and Dewen Hu, *Senior Member, I*

【强化学习 64】Multiobjective RL

张楚珩
清华大学 交叉信息院博士在读

一篇综述。

原文传送门

Liu, Chunming, Xin Xu, and Dewen Hu. "Multiobjective reinforcement learning: A comprehensive overview." IEEE Transactions on Systems, Man, and Cybernetics: Systems 45.3 (2014): 385-398.

特色

因为最近想了解一下multiobjective RL，就从读一篇综述开始吧。但是这篇综述写的实在不咋地，读起来比较难受，费了一下午。记录一下其中可能有帮助的点吧。

过程

Multiobjective讲的就是RL里面的奖励由之前的一个标量变成一个矢量（即存在多个可能相互冲突的目标）时应该如何去解。这种情况下可能没有一个确定的最优解。解决这类问题主要有两大类方法，一种是学习单个策略，另一种是学习多个策略。

学习单个策略的框架大体如下，看文章中的描述应该是针对于每个奖励的分量 r_i 都去学习其相应的在当前策略下的Q函数，当前的策略通过一个综合的Q函数 TQ 来得到，不同的单策略方法提出不同 TQ 计算方法，以此产生相应的策略（或者直接提出某个策略）。（由此看来，图中红线部分应该替换成SARSA这样的on-policy更新公式，而不是图中的Q-learning更新公式）

Algorithm 3 Naïve Solution of Single-Policy Approaches to MORL

\\K: The maximum number of episodes
\\N: The number of objectives
Initialize $TQ(s, a)$ arbitrarily;
repeat (for each episode j)
  Initialize s;
  repeat (for each step of episode)
  Choose a from s using policy derived from $TQ(s, a)$;
  Take action a, observe $r_1, r_2, \ldots, s'$;
  for $i = 1, 2, \ldots, N$ **do**
  $Q_i(s, a) \leftarrow Q_i(s, a) + \alpha[r_i + \gamma \max_{a'} Q(s', a') - Q_i(s, a)]$;
  end for
  Compute $TQ(s, a)$;
  $s \leftarrow s'$;
  until s is terminal
until $j = K$

Single policy approach for MORL

1. Weighted-sum approach

一种最直接的方式就是直接把各个奖励分量的Q函数加权相加得到 TQ 函数（given加权系数）。

$$TQ(s, a) = \sum_{i=1}^{N} w_i Q_i(s, a).$$

这种方法目测应该等效于直接观察到的就是加权之后的标量奖励，因此这种方法实际上不是一个multiobjective RL，或者说转化成了一个普通的RL问题。

2. W-learning approach

这种方法下取所有奖励分量里面Q函数最大的那个，它能够保证其结果至少对于某一个分量上来说是最优的策略。

$$TQ(s,a) = \max_i Q_i(s,a) \quad 1 \le i \le N. \qquad (13)$$

这种方法的问题是它对于各个reward的数值缩放敏感，即要求各个reward之间绝对数值大小是要可比的。

3. Analytic hierarchy process approach

这里假设不同的奖励分量之间有定性的重要性程度排序关系，假设第 i 个奖励分量相对于第 j 个奖励分量的相对重要性为 $c_{i,j}$ ，可以计算一个重要性因子 I_i 。

$$I_i = \frac{SL_i}{\sum_{j=1}^{N} SL_j}$$

where

$$SL_i = \sum_{j=1, j\neq i}^{N} c_{i,j}$$

文中提到，根据这个重要性因子和两个动作的相对 Q 函数值

$$D_i(a_p, a_q) = Q_i(s, a_p) - Q_i(s, a_q)$$

可以使用一个 fuzzy system 去判断这两个动作的相对好坏。（这里讲的 fuzzy system 我不是太懂，同时我猜这里说的判断 action 的好坏这个环节等价于相对于 TQ 函数的 greedy policy）。

这种方法的劣势在于它需要关于问题的先验知识。

4. Ranking approach

这种方法先对于不同的奖励分量排个序，并且对于每个奖励分量设置一个阈值。为了在每轮迭代的时候，基于现在各个 Q 函数的数值得到一个策略，当遇到一个状态 s 的时候，采取如下方法来得到一个 action。按照顺序去比较各个奖励分量上的 Q 函数，如果不同的 action 在该分量上的 Q 函数小于设定的阈值，就按照它们的大小关系来选择 action。如果都大于这个阈值，或者它们数值相同，那么再比较它们后一个分量上的 Q 函数。即，（我截取了原论文上的算法示意图，这篇文章

里面写的不对）

$$CQ_{s,a,j} \leftarrow \min(Q_{s,a,j}, C_j)$$

In state s, the greedy action a' is selected such that superior$(CQ_{s,a'}, CQ_{s,a}, 1)$ is true $\forall a \in A$ where superior$(CQ_{s,a'}, CQ_{s,a}, i)$ is recursively defined as:

if $CQ_{s,a',i} > CQ_{s,a,i}$
  return true
else if $CQ_{s,a',i} = CQ_{s,a,i}$
  if $i = n$
    return true
  else
    return superior$(CQ_{s,a'}, CQ_{s,a}, i+1)$
else
  return false

5. Geometric approach

文中说该方法认为 π_Q 需要服从某些几何约束，其他的没读懂了。

Fig. 5. Predicted target set (two objectives).

6. Convex hull approach

这是一个多策略的方法，解出不同 linear preference 的 optimal policy。同样也不是很明白。

7. Varying parameter approach

这个比较好理解，就是使用不同的参数（表征不同的对于各个奖励分量的偏好程度）来训练多个策略，把这些策略的结果作为 Pareto 前沿上的解。

总结

TABLE I
REPRESENTATIVE APPROACHES TO MORL

MORL Approaches		Basic Principle
Single-policy approaches	The weighted sum approach	A linear weighted sum of Q-values is computed as the synthetic objective function.
	The W-learning approach	Each objective has its own recommendation for action selection and the final decision is based on the objective with the largest value.
	The AHP approach	The analytic hierarchy process (AHP) is employed to derive a synthetic objective function.
	The ranking approach	"Partial policies" are used as the synthetic objective function.
	The geometric approach	A target set satisfying certain geometric conditions in multi-dimensional objective space is used as the synthetic objective function.
Multiple-policy approaches	The convex hull approach	Learn optimal value functions or policies for all linear preference settings in the objective space.
	The varying parameter approach	Performing any single-policy algorithm for multiple runs with different parameters, objective threshold values and orderings.

最后贴一个关于 Pareto 前沿的东西

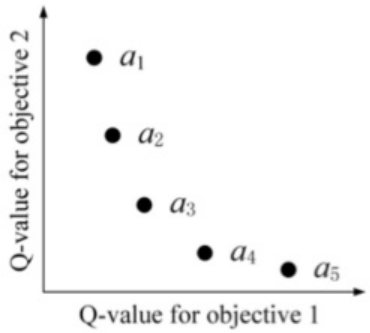

Fig. 4. Concave region of the weighted sum approach.

考虑两个优化目标，图上的五个点都是 Pareto 前沿，但是行程了一个 concave 的形状。如果对于不同的权重，优化两个目标的线性组合，只能得到 a_1 和 a_5 而不能得到中间的点。这个例子告诉我们通过采集不同的线性组合权重是不能够得到整个 Pareto 前沿的。

发布于 2019-05-28

强化学习 (Reinforcement Learning)