3P-04

任意のパレート解を得るための多目的強化学習のパラメータ推定

斉竹　良介†　　荒井　幸代†
千葉大学大学院工学研究科†

1. はじめに

実社会では多くの競合する目的の最適化が求められる場面が多い。最適化問題の一解法である強化学習研究においても二つ以上の目的を最適化する多目的強化学習が提案されている。

図 1 に示すように，多目的強化学習の最適解は複数存在する。縦軸と横軸はそれぞれの目的関数値を表し，最適解であるパレート解が黒い点，それ以外の支配される解が灰色の点で表されている。以下ではこの最適解群の中からエージェントに任意の解をとる行動を学習させることを考える。

様々なパラメータを組み合わせて最適解群全ての解を得たうえで，その中から設計者が最も好ましい解を選択するのが最良であるが，問題が大規模であり，計算コストが大きいなどの理由で，パラメータの組み合わせを試すことが難しい場合がある。そこで，任意の解をとるエキスパートから，学習のためのパラメータを推定し，エージェントに任意の解をとる行動を学習させることを考える。

多目的強化学習では各目的に対して設計者が重みというパラメータを設定し，好ましい解の方向を定める方法が一般的である。そのため，本研究では任意の解をとるエキスパートの行動軌跡上のQ値から，各目的に対する重みの推定法を提案する。強化学習の歴史における大きな流れの１つは動物の心理学であるため，心理学との関わりがある。重みを推定することは，行動者がどの目的に重きを置いているかを知ることにつながるため，暗黙知の抽出などに役立つと考えられる。多目的強化学習のベンチマークである Deep Sea Treasure 環境[2]を用いて実験を行い，提案法を評価した。

2. 問題設定

強化学習における逆問題として逆強化学習が研究されている[3][4]。逆問題とは，入力から出力を求める問題を順問題と呼ぶのに対し，出力から入力を求める問題のことである。表 1 に逆強化学習と重み推定の入力と出力を整理する。

一般に強化学習は状態，行動，状態遷移確率，報酬を入力として，問題に対して最適な行動系列を出力する。その逆に，逆強化学習では状態，行動，状態遷移確率，そして強化学習では出力であった行動軌跡から，入力であった報酬を求

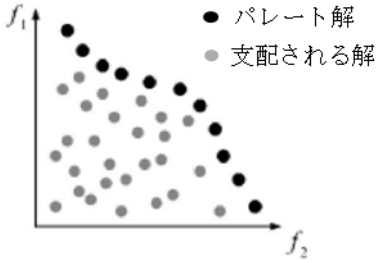

図 1 多目的強化学習の最適解群[1]

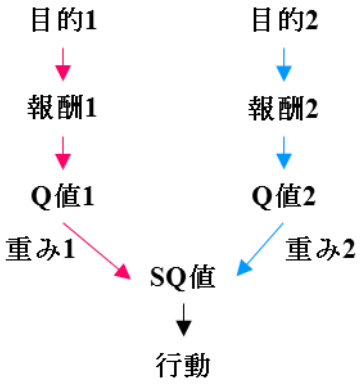

図 2 多目的強化学習の概略

表 1 逆強化学習と重み推定

	逆強化学習	重み推定
逆問題対象	強化学習	多目的強化学習
入力	● 状態 ● 行動 ● 状態遷移確率 ● 行動軌跡	● 状態 ● 行動 ● 状態遷移確率 ● 行動軌跡上のQ値 ● 報酬
出力	● 報酬	● 重み

Parameter Estimation of Multi-Objective Reinforcement Learning to Reach Arbitrary Pareto Solution
†Ryosuke SAITAKE †Sachiyo ARAI
†Graduate School of Engineering, Chiba University

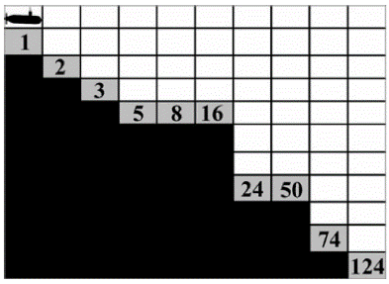

図 3 Deep Sea Treasure[2]

める。

これに対し，本研究では多目的強化学習の逆問題として，重みの推定問題を対象とする。多目的強化学習では図 2 に示すように，各目的に対して報酬を設定し，目的ごとにQ値を計算する。本研究で用いる weighted アプローチでは重みwを用いて Q 値を結合し，その結合した Q 値(SQ 値)に基づいて行動する。多目的強化学習は強化学習と同様，状態，行動，状態遷移確率，報酬，そして重みを入力とし，最適な行動系列を出力する。逆問題である重みの推定では，状態，行動，状態遷移確率，報酬，行動軌跡上のQ値から，各目的に対する重みを求める。

3. 多目的強化学習の重み推定法

任意の解への最適行動をとるエキスパートの行動軌跡上のQ値，状態，行動，報酬を所与として，エキスパートの重みを推定する。多目的強化学習のアルゴリズムは非線形手法であり性能が高い Chebyshev スカラー化関数[4]を用いる。

まず，重みの初期値を設定する。次に，設定した重みを用いて行動軌跡上の Q 値を結合し，SQ値を求める。求めたSQ値に従い行動するエージェントの軌跡がエキスパートの軌跡と合致するまで重みを小さな定数αだけ変化させる。具体的な流れを以下に示す。

1. 重みの初期値を決定
2. 重みを用いてQ値を結合し，SQ値を算出
3. SQ値に従うエージェントのグリーディ行動とエキスパートの行動軌跡が一致する場合，重みを出力し終了
4. 一致しない場合，重みをαだけ変化させ，2.から繰り返し

4. 実験

図 3 に示す Deep Sea Treasure 環境[4]を用いて Chebyshev スカラー化関数の重みを推定する実験を行った。Deep Sea Treasure 環境は縦 10×横 11 のグリッドからなっており，左上のスタート地点から数字の書かれたグリッドのいずれかにたどり着いたときゴールとする。この環境では二つの目的があり，一つは少ない時間でゴールにたどり着くこと，もう一つは価値の高い宝物を得ることである。行動は上下左右に移動する四つ，報酬は 2 要素のベクトルからなる。一つ目の要素はタイムペナルティで，毎ステップ-1 を与える。二つ目の要素は宝物の価値であり，エージェントが宝物の置いてある場所にたどり着いたとき，その宝物分の報酬を得る。

Chebyshev スカラー化関数では重みを 0.91, 0.01 としたとき時間 17,宝物 74 を得る。逆問題である重みの推定では，時間 17，宝物 74 をとる行動軌跡上の Q 値からの重み 0.91, 0.01 を推定する。また，初期重みを時間と宝物それぞれ 0.0, 1.0 とし，$\alpha = 0.01$として実験を行った。

この設定において提案手法を用いて実験を行ったところ，正確に重みを推定することができた。

5. 結論

エキスパートの行動軌跡上のQ値を所与として，各目的の重みを変化させる推定法を提案した。これにより，効率的に任意のパレート解をとる行動を学習させることができる。

今回の実験環境は二目的であり，重みを変化させることが容易であったため，今後はさらに目的が多い環境で提案法を評価する必要がある。また，行動軌跡上のQ値を所与とすることは現実には難しいと考えられるため，エキスパートの行動軌跡だけから重みを推定する方法を考案する予定である。

参考文献

[1] Liu, Chunming, Xin Xu, and Dewen Hu. "Multiobjective reinforcement learning: A comprehensive overview." *Systems, Man, and Cybernetics: Systems, IEEE Transactions on* 45.3 pp.385-398 (2015)

[2] Van Moffaert, Kristof, Madalina M. Drugan, and Ann Nowé. "Scalarized multi-objective reinforcement learning: Novel design techniques." *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), IEEE Symposium on*. IEEE, pp.191-199 (2013)

[3] Abbeel, Pieter and Andrew Y. Ng. "Apprenticeship learning via inverse reinforcement learning." *Proceedings of the 21th international conference on Machine learning*. ACM (2004)

[4] Ng, Andrew Y. and Stuart J. Russell. "Algorithms for inverse reinforcement learning." In proceedings of ICML, pp.663-670 (2000)