The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

多段・多目的最適化における解の網羅的発見

Finding compromise solutions for multi-stage and multi-objective optimization problem

山本慶佑[*1]　荒井幸代[*1]
Keisuke Yamamoto　Sachiyo Arai

[*1]千葉大学大学院融合理工学府都市環境システム
Department of Urban Environment Systems, Graduate School of Science and Engineering, Chiba University

Many problems in real world could be formalized as the multi-stage and multi-objective optimization problems (MOP) where there exist mutual conflicts among the objective functions. It is said that people hard to find a compromise solution without a sufficient number of solutions as the candidates. Thus, we propose a method to find the pareto-optimal exhaustively. Our approach is based on multi-objective reinforcement learning (MORL) because the real-world problem requires multiple action-sequences until getting the reward. We evaluate our proposed method by applying it to "waste collection problem" where there are two conflicting objective functions, "capacity of collecting vehicles" and "time".

1. はじめに

実世界の問題の多くは複数の競合する目的からなる多目的最適化問題である．多目的最適化問題は様々な制約が存在し，これらの制約をすべて満たす解が実行可能解となる．従来，複数の制約を満たす実行可能解の中で一つの目的関数だけに着目した解が選択されてきた．すなわち，一つの目的関数が他の制約の充足性に与える影響を考慮してこなかった．さらに，多目的最適化問題は複数の制約が複雑に構成されているため，制約を完全に満たす領域に解が存在しない場合がある．また，目的関数があらかじめ明示的に与えられている状況に対しては既に優れた方法が提案されているが，実世界において，すべての状況に対してあらかじめ評価値を与えることは困難である．

そこで本論文では，多目的最適化問題の例としてごみの巡回収集を対象とし，複数の解から意思決定者の嗜好に合う制約を設定し，意思決定者にとって重要度の高い解を選択する手法を提案する．また，ごみ収集過程における不確実性にも着目し，環境の確率的変動に頑健であり，複数の行動後の評価値だけを手がかりに最適解が得られる多目的強化学習を導入する．学習の結果得られたパレート最適解から，獲得可能なすべての解を提示することによって，制約充足を損なわない意思決定の実現を目指す．

2. 準備

2.1 強化学習

強化学習は，未知の環境において最適な制御を試行錯誤的な探索を通して獲得する手法である [Sutton 00]．意思決定主体であるエージェントには，状態入力に対する正しい出力を明示する教師信号が存在せず，報酬と呼ばれるスカラーの情報だけが与えられる．エージェントは報酬の期待総和を最大化する制御を行うことを目的に，制御方策を学習する．

強化学習における環境モデルを $\langle \mathcal{S}, \mathcal{A}, \mathcal{R} \rangle$ と定義する．$\mathcal{S}$ は状態集合，$\mathcal{A}$ は行動集合，$\mathcal{R}$ は報酬関数を表す．エージェントは時刻 t において，状態 $s_t \in \mathcal{S}$ を観測し，自身の方策 π_t に基づいて行動 $a_t \in \mathcal{A}$ を選択する．その後，時刻 $t+1$ では s_t，a_t によって確率的に次状態 s_{t+1} に遷移し，報酬 $r_{t+1} \in \mathcal{R}$ を得る．エージェントは獲得した報酬から価値関数を生成し，この価値関数を用いて方策 π を更新する．

連絡先: 山本慶佑，千葉大学大学院融合理工学府都市環境システム，千葉市稲毛区弥生町　1-33

Algorithm 1 Value Iteration

Initialize array Q arbitrarily (e.g., $Q(s,a) = 0$ for all s, a)
while $\Delta > \theta$ (a small positive number) **do**
  $\Delta \leftarrow 0$
  for $s \in S, a \in A$ **do**
    $q \leftarrow Q(s,a)$
    $Q(s,a) \leftarrow \max_a \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma Q(s', a')]$
    $\Delta \leftarrow \max(\Delta, |q - Q(s,a)|)$
  end for
end while
return $\pi(s,a) = \arg\max_a Q(s,a)$

2.1.1 マルコフ決定過程

マルコフ決定過程（MDP : Markov Decision Process）は，状態遷移にマルコフ性を持つ確率システムの動的最適化のための数学モデルである．マルコフ性とは，次状態 $s_{t+1} \in \mathcal{S}$ が現在の観測状態 s_t と行動 a_t だけによって定まる性質のことをいう．MDP において，強化学習は最適な方策の獲得が保証されており，一般的に強化学習では MDP を仮定する．

MDP では，任意の状態 s と行動 a が与えられたとき，次に可能な各状態 s' の確率 $P^a_{ss'}$ を式 (1) で表す．

$$P^a_{ss'} = \Pr\Big\{ s_{t+1} = s' \mid s_t = s, a_t = a \Big\} \tag{1}$$

$P^a_{ss'}$ は遷移確率と呼ばれる．同様に，式 (2) に現在の任意の状態 s と行動 a が与えられたときの次の報酬の期待値を示す．

$$R^a_{ss'} = E\Big\{ r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s' \Big\} \tag{2}$$

2.1.2 価値反復

価値反復は環境の状態遷移確率を用いて最適な状態価値 V^* あるいは行動価値 Q^* を求める方法である．Q^* は式 (3) を用いて求める．

$$Q^*(s,a) = \max_a \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma \max_{a'} Q^*(s', a')] \tag{3}$$

価値反復は行動価値の推定と方策の改善を交互に繰り返すことで最適方策を求める．価値反復は価値関数の変化量が十分に小さくなった時点で学習を終了する．Algorithm1 に価値反復のアルゴリズムを示す．

2.2 多目的最適化問題

多目的最適化問題とは，「複数個の互いに競合する目的関数を，与えられた制約条件の中で何らかの価値基準に照らし合わせて最大化 (あるいは最小化) する問題」と定義されている [中山 07]．本論文では，最大化問題として論を進める．

一般に多目的最適化問題は，n 個の設計変数を扱う，m 個の互いに競合する目的関数

$$f_i(x_1, x_2, \ldots, x_n) \qquad (i = 1, 2, \ldots, m)$$

を，l 個の不等式制約条件

$$g_j(x_1, x_2, \ldots, x_n) \geq 0 \qquad (j = 1, 2, \ldots, l)$$

のもとで最大化する問題として定式化される．

多目的最適化問題では，目的関数間にトレードオフの関係が存在するため，すべての目的関数 $f_i(x)$ を同時に最大化することはできない．そのため，多目的最適化問題ではすべての目的において最大な値をとる最適解は一般には存在しない．そこで，多目的最適化問題では最適解の代わりに新たな解の概念として，パレート最適解を用いる．

2.2.1 パレート最適解

パレート最適解は，多目的最適化問題における解の優越関係により定義される．まず，ベクトルに対する不等号の使い方について定義し，次に多目的最適化問題における解の優越関係の定義を以下に示す．

定義 2.1 (ベクトル不等式)：$\boldsymbol{y^1}, \boldsymbol{y^2} \in \boldsymbol{R^m}$ に対し

$$\begin{aligned} \boldsymbol{y^1} < \boldsymbol{y^2} \quad &\Leftrightarrow \quad y_i^1 < y_i^2, \quad \forall i = 1, \ldots, m \\ \boldsymbol{y^1} \leqq \boldsymbol{y^2} \quad &\Leftrightarrow \quad y_i^1 < y_i^2, \quad \forall i = 1, \ldots, m \\ \boldsymbol{y^1} \leq \boldsymbol{y^2} \quad &\Leftrightarrow \quad \boldsymbol{y^1} \leqq \boldsymbol{y^2}, \quad \boldsymbol{y^1} \neq \boldsymbol{y^2} \end{aligned}$$

定義 2.2(パレート最適解)：$\boldsymbol{f(\hat{x})} \leq \boldsymbol{f(x)}$ となる $\boldsymbol{x} \in \boldsymbol{X}$ が存在しないとき，$\boldsymbol{\hat{x}}$ をパレート最適解とよぶ．

本論文では，多目的強化学習を扱うため，パレート最適解をパレート最適方策に言い換える．また，パレート最適方策の集合をパレートフロントとよぶ．

2.2.2 多目的マルコフ決定過程

多目的マルコフ決定過程 (MOMDP:Multi-Objective MDP) は通常の MDP を多目的問題に拡張した環境である．MDP では報酬関数はスカラー量として扱っていたが，MOMDP では報酬関数はベクトル値として扱う．エージェントが状態 $s \in \mathcal{S}$ で行動 $a \in \mathcal{A}$ を取って，次状態 $s' \in \mathcal{S}$ へ遷移した際に，報酬ベクトル $\boldsymbol{r}(s, a) = [r_1(s, a), r_2(s, a), \ldots, r_m(s, a)]$ が与えられる．ここで m は目的の個数である．

2.2.3 多目的強化学習

多目的強化学習は一般に MOMDP を仮定したアプローチである．MOMDP においては，報酬ベクトルの各要素は各目的に関わり，目的間に競合が生じることが多い．一つの目的に対して有効な方策は他の目的に有効にならないことが多いため，パレート最適方策が複数存在する．したがって，目的間におけるトレードオフの関係が成立するパレート最適方策を獲得する必要がある．多目的強化学習は複数存在するパレート最適方策をすべて獲得することを目的とするアプローチである．

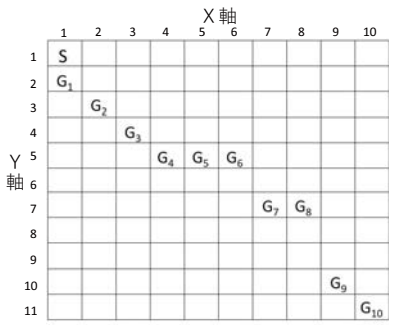

図 1: ごみ収集環境

3. 問題設定

3.1 ごみ収集問題

「収集車のごみ巡回収集における最適経路決定問題」を「ごみ収集問題」とよぶことにする．ごみ収集問題が対象とする環境は，$M \times N$ のグリッドワールドを設定し，ごみステーションを L 個配置したものとする．M, N, L は自然数である．

本研究では，図 1 に示した $M = 10, N = 11, L = 10$ の環境を扱う．S は処理施設，$\mathrm{G}_i (i = 1, 2, \ldots, 10)$ はごみステーションを表す．収集車は処理施設を出発し，任意のごみステーションを巡回した後，処理施設へと戻りごみを搬出する．この工程を 1 回の収集とし，これを最適化問題として定式化する．収集車は 1 ステップごとに { 上，下，左，右 } の行動を一つ選択する．収集車の状態は自己の位置座標，各ごみステーションの巡回の有無からなる 12 次元の状態空間である．したがって全状態数は，収集車の座標で 10×11 ，各ごみステーションの巡回の有無を表した 2^{10} の積，$112,640$ となる．

3.2 多目的最適化問題としての定式化

ごみ収集問題は収集車の容量や台数，収集時間帯など複雑な制約が複数存在する．本研究では，ごみ収集問題の制約として「収集車の容量」と「時間ステップ」を取り上げ，それぞれの目的関数を考える．前者に対しては，収集量の最大化，後者に対しては時間ステップの最小化の 2 目的最適化問題である．また，報酬関数は二つの目的の遂行度を評価する二つの要素からなるベクトルとして定義する．式 (4) に収集量に対する報酬 r_g，式 (5) に時間ステップに対する報酬 r_t を示す．

収集量に対する報酬

$$r_g = \begin{cases} \sum_{i=1}^{10} R_i W_i & \text{if } (x, y) = (1, 1) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

時間ステップに対する報酬

$$r_t = -1 \tag{5}$$

収集量に対する報酬 r_g は，エージェントが処理施設へ戻ったとき，訪れたごみステーションの数に応じて正の報酬を与え，ごみステーションを訪れていない場合は 0 を与える．R_i は各ごみステーションのごみの発生量に対応した正の報酬である．W_i はごみステーションの巡回状態を $\{0, 1\}$ で表す．0 は巡回していない状態，1 は巡回した状態を表す．よって，エージェントが処理施設へ戻ったとき，巡回したごみステーションの収集量に対応した正の報酬の総和を受け取る．時間に対する報酬 r_t は，1 ステップごとに -1 を与える．

Algorithm 2 Pareto Front Value Iteration

Initialize $\hat{Q}(s,a)$ arbitrarily $\forall s,a$
while not converged **do**
  for $s \in S, a \in A$ **do**
    $\hat{Q}(s,a) \leftarrow \mathrm{E}[\boldsymbol{r}(s,a) + \gamma \mathrm{Pareto} \bigcup_{a'} \hat{Q}(s',a') | s,a]$
  end for
end while
return $\hat{Q}$

3.3 関連研究

ごみの巡回収集を解くアプローチとして運搬経路問題 (VRP: vehicle routing problem) がある [久保 02]. VRP とは，デポ (depot) とよばれる特定の地点を出発した運搬車による，顧客の需要を運搬するコスト最小のルートを求める問題の総称である．VRP では，点集合と枝集合で構成されるグラフと移動コストが与えられることを前提としている．しかし，すべての状況に対してあらかじめ評価値を与えることは難しいという問題点がある．また，関連研究 [川中子 02] は制約を絶対条件とするハード制約を対象としているため，多少の時間の遅れや収集車の容量を多少超えてしまう巡回経路を実行不可能な解として扱う．しかし，制約を緩和することで，実行可能解が広がり，関連研究の解より意思決定者にとって重要度の高い解候補を獲得できる可能性がある．

そこで本研究では，制約を緩和し，獲得できるすべての解を提示することによって，意思決定者の嗜好に合う代替案を提案する．そのために，あらかじめ移動コストなどの評価値を与える必要のない多目的強化学習を用いて複数のパレート最適方策を求める．

4. 提案手法

本章ではまず，本研究で用いる多目的強化学習のアルゴリズムについて記述し，次に提案手法について説明する．

4.1 Pareto Front Value Iteration

Pareto Front Value Iteration(PFVI)[齋竹 17] は二つのステップによりパレート最適方策を求める．

Step 1：報酬ベクトルから各状態行動対に対して非劣な Q ベクトル集合を学習

Step 2：学習後の非劣な Q ベクトル集合に対してチェビシェフスカラー化関数を用いてスカラー化を行い，Q ベクトルから Q 値を抽出

PFVI ではまず複数の Q ベクトルを学習する．次に，学習した複数の Q ベクトルをスカラー化しパレート最適方策を獲得する．本節では，まず PFVI の学習アルゴリズムを説明し，次にスカラー化関数について示す．

学習

Algorithm2 に PFVI の学習アルゴリズムを示す．非劣な Q ベクトル集合は各状態行動対のパレートフロントを表す．ここで，非劣な Q ベクトルとは $\boldsymbol{q^1} \leq \boldsymbol{q^2}$ となる$\boldsymbol{q^1}$が存在しないときの$\boldsymbol{q^2}$である．$\hat{Q}(s,a)$ は状態 s で行動 a をとるときの非劣な Q ベクトル集合を表す．$\hat{Q}(s,a)$ の更新は式 (6) を用いる．

$$\hat{Q}(s,a) = \mathrm{E}[\boldsymbol{r}(s,a) + \gamma \mathrm{Pareto} \bigcup_{a'} \hat{Q}(s',a') | s,a] \qquad (6)$$

Pareto は点集合の非劣な Q ベクトル集合を求める操作である．$\bigcup_{a'} \hat{Q}(s',a')$ は，次状態 s' における全行動 a' に対する $\hat{Q}(s,a)$ の非劣な Q ベクトル集合をとっている．

スカラー化

単目的強化学習の枠組みでは報酬をスカラー量として扱うことを前提としている．PFVI ではスカラー化関数を用いて Q ベクトルをスカラー化し方策を獲得する．式 (7) を用いて Q 値を抽出する．

$$TQ(s,a) = \min_{\boldsymbol{q} \in \hat{Q}(s,a)} \max_{i=1,\ldots,m} w_i \cdot |q_i - z_i^*| \qquad (7)$$

z_i^* は学習した q_i の中で最もよい値とする．式 (7) を用いて $\hat{Q}(s,a)$ 中の Q ベクトルと重み $w_i \in [0,1]$ を用いてチェビシェフスカラー化を行い，その中で最小の Q 値を抽出し，パレート最適方策を獲得する．重み w_i は $\sum_{i=1}^{m} w_i = 1$ を満たす．

以上のアルゴリズムを用いてごみの巡回経路を求める．ここで，ごみ収集問題のパレート最適経路とはごみ収集問題におけるパレート最適方策で獲得できる最適経路である．

4.2 提案：パレート最適経路の発見法

関連研究の多くは，二つのステップにより最適経路を求める．

Step 1： 事前に制約の値を設定

Step 2： 制約条件を満たす最適解を一つ求める

関連研究では，事前に複数の制約を設定し，最適解を求めているため，これらの制約条件が適切な値であるか不明であった．そこで，制約の緩和を考慮に入れ，意思決定者の嗜好に合う制約を設定し，意思決定者にとって重要度の高い解を選択する手法を提案する．ここで,「意思決定者にとって重要度の高い解」を「妥協可能な解」とよぶことにする．また，解候補とは任意の制約のもと実行可能なパレート最適方策と定義する．

提案手法は三つのステップによりパレート最適経路を求める．

Step 1：制約を考慮せず，複数の解候補を探索

Step 2：パレート最適方策から制約条件を設定

Step 3：妥協可能な解を選択

Step 1 では，制約を考慮せずに多目的最適化問題として定式化し，PFVI を用いてパレート最適方策を求める．Step 2 以降は意思決定者の意思決定の手順について示す．意思決定者は求めたパレート最適方策を参考に制約の緩和限界を検討し，新しく設定した制約の実行可能解の中から妥協可能な解を選択する．

5. 計算機実験

5.1 実験設定

図 1 に示す 10×11 のグリッドワールドにごみステーションが 10 箇所配置されている環境で実験を行う．エージェントの目的は収集量の最大化と時間ステップの最小化である．割引率 $\gamma = 0.99$，ごみの収集量に対応した報酬は $R_i = 1$ $(i = 1, 2, \ldots, 10)$ とする．$\hat{Q}(s,a)$ の変化量が 1.0×10^{-5} 以下になるとき学習を終了とする．また，Q 値の抽出は重みを 1.0×10^{-2} 刻みで変化させて行う．

5.2 実験結果

図 2 に収集量と時間ステップに関するパレートフロントと重みの関係性を示し，図 3 にごみ収集環境におけるパレート最適経路を示す．縦軸は収集量を，横軸は時間ステップを表し，w_1 は収集量に対する重みである．ごみ収集問題では，10 個のパレート最適方策が存在する．これらのパレート最適方策は最短経路でごみを収集する行動系列と一致する．PFVI を用いてごみ収集問題におけるすべてのパレート最適経路を獲得した．

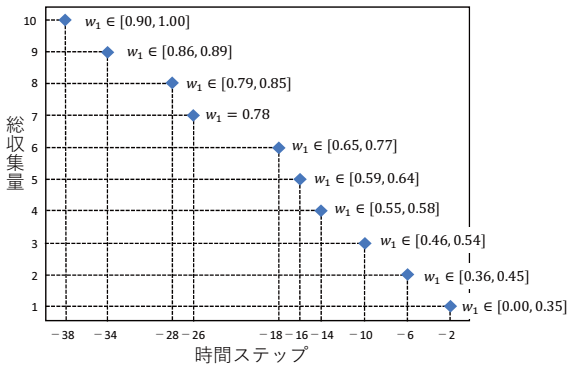

図 2: ごみ収集問題のパレートフロントと重みの関係

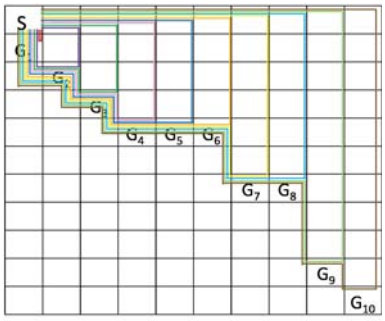

図 3: ごみ収集環境におけるパレート最適経路

ごみの発生量がすべて等しい環境では，パレート最適方策をすべて獲得できることがわかった．

5.3 提案手法の適用例

5.3.1 時間に関する制約

図 2 のパレートフロントを参考に制約の設定を行う．まず，時間に関する制約ついて考える．今回の例では，事前に設定されていた制約を 25 ステップとする．意思決定者は獲得したパレート最適方策を参考に新しい制約を設定する．この例では，収集時間を 5 ステップ延長することによって解候補が 2 個増加し，意思決定者は 8 個の解候補の中から妥協可能な解を選択する．このように意思決定者は，得られた解候補に基づいて制約の緩和限界を検討することができる．

5.3.2 収集車の容量に関する制約

新しく設定した時間の制約を考慮した収集車の容量に関する制約について考える．図 4 に時間と収集量を考慮に入れたパレートフロントを示す．意思決定者は容量の異なる 2 台の収集車を所持しているとする．(a)1 台の収集車の容量だけを満たす解候補，(b) どちらの収集車の容量も満たす解候補，で収集車の選択方法が異なる．(a) の場合，意思決定者は容量を満たす収集車 1 を選択する．(b) の場合，2 台とも容量を満たしているため，意思決定者の嗜好に合う収集車が選択される．図 4 はどの容量の収集車を所持すべきかを検討する指標になると考えられる．

6. まとめ

本研究では，多目的最適化問題における数多くの制約充足を損なわずにパレート最適解を選択するための意思決定実現を目的とし，複雑で多くの制約をもつ多目的最適化問題の例としてごみ収集問題を扱った．そして，ごみ収集問題を多目的最適

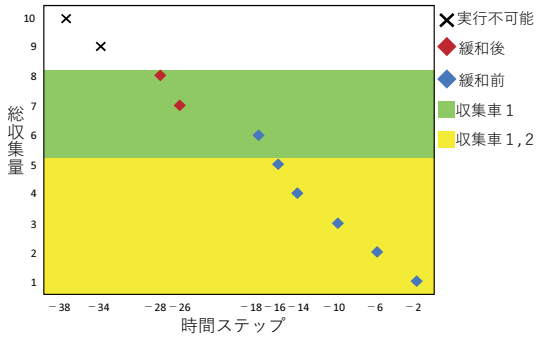

図 4: 時間と収集量に関する制約の解候補

化問題として定式化することによって，パレート最適経路の発見法を適用し，意思決定者が目的の重要度に応じて制約を緩和する手順を示した．本研究の実験環境では，効用の遅れがあるごみ収集環境において多目的強化学習の適用が有用であることを示した．関連研究では，ごみ収集問題を単目的最適化問題として扱っており，また制約を事前に設定するため，解候補が一つだけとなり，制約の設定値が適切であるか不明であった．しかし，提案手法では，複数の解候補をもとに収集車の容量または，収集時間に対する制約を緩和することによって，妥協可能な解を選択可能であることを示した．

今後の課題として，実世界のごみの巡回収集に近い環境においても提案手法が適用可能か確認するために他の環境で実験を行う必要がある．具体的には，ごみの排出量の不確実性を考慮した環境に提案手法を適用させることを考えている．強化学習は環境の確率的変動に頑健であるため，不確実性のごみ収集環境への適用が期待できる．また，提案手法の詳細な評価とともに制約最適化問題と制約充足問題の関連研究を整理し，更なる提案手法の洗練を行う必要がある．

参考文献

[Sutton 00] Richard S. Sutton, Andrew G. Barto(三上貞芳, 皆川雅章 (訳)): Reinforcement learning , 森北出版, 2000.

[中山 07] 中山弘隆，岡部達哉，荒川雅生，尹禮分：多目的最適化と工学設計 -しなやかシステム工学アプローチ-，現代図書, 2007.

[久保 02] 久保 幹雄，田村 明久，松井 知己：応用数理計画ハンドブック，朝倉書店, 2002.

[川中子 02] 川中子 敬至：ゴミ・ステーションを巡回する収集車の経路問題, オペレーションズ・リサーチ, vol.47, No.11, pp.737-742, 2002.

[齋竹 17] 齋竹 良介，荒井 幸代：期待報酬ベクトルの非線形スカラー化による多目的強化学習アルゴリズム，千葉大学院工学研究科 修士論文，2017.