The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

他者の行動による目的移行のメカニズム ～動的選好の導入による多目的意思決定理論の拡張

Mechanism of objective transition by behavior of others
～Expansion of multi-objective decision theory by introducing dynamic preference～

西 孝介　　荒井 幸代
Nishi Kousuke　　Arai Sachiyo

*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 of the decisions in the real world are multi objective decision-making. Since there is generally no choice to optimize all objectives, the decision maker selects one from multiple alternatives with different importance for each objective based on their own preferences. However, in situations where there are other decision makers , the choice you choose may change. At this time, I focus on that a latent "objective for considering the selection of others" appears and decision maker shifts the objective. Therefore, in this research, I propose a modeling of decision-making to shift objectives due to the existence of others in the multi-agent system. I also control the behavior of the whole decision-maker using the objective transition mechanism. As a result of the experiment, I could model decisions considering the selection of others, and succeeded to control the whole decision-maker to the ideal behavior.

1. はじめに

実社会における多くの問題は，効用が相反する複数の目的をもつ多目的意思決定問題であり，一般にすべての目的を最適化する解は存在しない．そこで意思決定者は各々がもつ固有の選好に基づき，各目的への重要度に応じた複数の解の中から一つを選択する．しかし，自身のほかに意思決定者が存在する状況では，他者の影響を受けて選択する解が変わることがある．例えば，多くの意思決定者が集中する解を避ける状況があげられる．ここでは，意思決定者の中に潜在していた" 他者を考慮する目的" が発生し，目的を移行したことにより選択を変更している．このように、実社会では意思決定者が複数存在する環境で，他者の影響を受けて意思決定をすることが多い．また、意思決定を一度ではなく，何度も繰り返す逐次的意思決定が多い．

そこで本研究では，意思決定者が複数存在するマルチエージェント系の下，多目的逐次意思決定における目的移行のモデル化を行う．多目的逐次意思決定は，最適な意思決定系列を獲得する手法である強化学習を，多目的最適化問題に拡張した多目的強化学習によりモデル化されている．多目的強化学習によるモデル化は今までシングルエージェント系を対象としていた．本研究では，他者の意思決定により，選択肢の価値が変化することに着目し，動的な報酬設計を加えることで，今までの多目的強化学習のモデル化をマルチエージェント系に拡張する．

また意思決定者の目的移行のメカニズムが明らかになれば，意思決定者の選択を誘導したい場面において，最適な制御則の提案に活用できる．そこでマルチエージェント系において，意思決定者の目的を網羅的に移行させることにより，系全体の選択の挙動を制御できることを計算機実験によって確認する．

以下，2 章では準備として，多目的最適化問題と強化学習の説明を行う．3 章では対象問題とする，マルチエージェント系における多目的逐次意思決定問題について述べる．4 章ではアプローチとして，多目的意思決定における目的移行のモデル化と，マルチエージェント系全体の挙動の制御の二つを説明し，5 章で二つのアプローチを用いた計算機実験について述べる．最後に 6 章で，まとめと今後の課題について述べる．

連絡先: 西孝介，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, sachiyo@faculty.chiba-u.jp

2. 準備

2.1 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意味で最大化 (最小化) する問題」と定義されている．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

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

を，l 個の不等式制約条件

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

のもとで最大化する問題として定式化される [中山 07].

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，全ての目的関数 $f_i(x)$ 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、最適解に代わる新たな解の概念として，パレート最適解を用いる．

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．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

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

定義 2.2 (パレート最適解) : $\boldsymbol{f(x)} \leq \boldsymbol{f(\hat{x})}$ 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.

2.2 強化学習

強化学習は，未知の環境において試行錯誤を繰り返しながら最適な制御則を獲得する手法である [Sutton 98]．環境モデルを $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ 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，$\mathcal{P}$ は遷移確率集合，γ は割引率を表す．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} を得る．獲得した報酬から価値関数を生成し，その値を用いて状態 s から可能な行動 a を選択する確率である方策 π を学習する．

最適方策

状態行動対 (s, a) の価値関数 $Q(s, a)$ を行動価値関数あるいは Q 値とよび、式 (3) に示す．Q 値は状態 s で行動 a をとった後，方策 π に従うときの期待報酬を表す．最適行動価値関数 $Q^*(s, a)$ は，状態 s において行動 a を実行した後に，最適方策を取り続ける場合の無限期間の割引報酬の期待値として，式 (4) で定義される．

$$Q^{\pi}(s,a) = E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \;\middle|\; s_t = s, a_t = a \right\} \tag{3}$$

$$Q^*(s,a) = \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma \max_{a'} Q^*(s', a')] \tag{4}$$

3. 対象問題

意思決定の分類を表 1 に示す．意思決定の種類は，意思決定に考慮する目的の数，意思決定の回数，意思決定者の数の三つの軸で分類できる．本論文では，目的の数，意思決定の回数，意思決定者の数，いずれも複数である，マルチエージェント環境における多目的意思決定問題を扱う．

3.1 多目的逐次意思決定問題

多目的意思決定問題の目的は，パレート最適解集合 $X^* = \{\boldsymbol{x}^*_1, \boldsymbol{x}^*_2, ..., \boldsymbol{x}^*_n\}$ の中から，意思決定者の選好に合致した選好解 $\boldsymbol{x}^*_k$ を獲得することである．ここで選好とは，意思決定時の各目的に対する重要度を要素とする，重要度ベクトル $\boldsymbol{w} = \{w_1, w_2, ..., w_m\}$ と定義する．ここで m は意思決定者に潜在する目的の個数である．本来，意思決定者は潜在する全ての目的を考慮することはなく，m 個のうちの k 個の要素からなる重要度ベクトルにより，選好解を獲得する．

多目的逐次的意思決定問題とは，複数の目的を考慮しながら，状況に応じた適切な意思決定を何度も繰り返す問題である．一般に多目的逐次意思決定問題は，多目的マルコフ決定過程 (MOMDP)，$\langle \mathcal{S}, \mathcal{A}, \boldsymbol{R}, \mathcal{P}, \gamma \rangle$ によりモデル化できる．MOMDP において、$\mathcal{S}$ は状態集合，$\mathcal{A}$ は行動集合，$\boldsymbol{R}$ は報酬ベクトル，$\mathcal{P}$ は遷移確率集合，γ は割引率を表す．状態 $s \in \mathcal{S}$ を観測したエージェントは，方策 $\pi : \mathcal{S} \mapsto \mathcal{A}$ にしたがって行動

表 1: 意思決定の分類

	目的数	意思決定回数	意思決定者数
単目的意思決定	1	1	1
多目的意思決定	複数	1	1
多目的逐次意思決定	複数	複数	1
本研究	複数	複数	複数

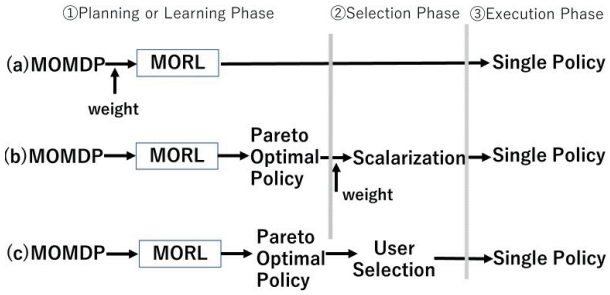

図 1: 多目的強化学習のアプローチ

$a \in \mathcal{A}$ を選択し，次状態 $s' \in \mathcal{S}$ へ遷移した際に，報酬ベクトル $\boldsymbol{R}(s, a, s') = (R_1(s, a, s'), R_2(s, a, s'), \ldots, R_m(s, a, s'))$ が与えられる．

3.2 多目的強化学習

多目的強化学習 (MORL) とは，多目的マルコフ決定過程を仮定した，パレート最適方策を獲得する手法である．多目的強化学習によって，一つの方策を獲得するまでの過程を図 1 に示す．多目的強化学習による手法は，学習により一つの方策を獲得する single-policy approach と，複数の方策を獲得する multiple-policy approach の 2 種類に大別できる [Roijers 13]．single-policy approach は (a) の手法にあたり，スカラー化関数によって多目的から単一目的に変換する手法である．スカラー化手法は，各目的に対する重要度ベクトル $\boldsymbol{w}$ を用いて複数の目的関数 $\boldsymbol{f}(\boldsymbol{x})$ をスカラー化することにより，単一のパレート最適解を獲得する．そのため設計者は，事前に各目的に対する重要度ベクトルを設定する必要がある．一方，multiple-policy approach は (b) と (c) の手法にあたり，事前に各目的に対する重要度ベクトルを設定する必要がない手法である．この手法は，学習によって重要度に応じた複数のパレート最適解からなる，パレート最適解集合を求めることを目的とする．

3.3 マルチエージェント系での多目的逐次意思決定

本研究では，意思決定者が複数存在するマルチエージェント系における多目的逐次意思決定問題を考える．マルチエージェント系において意思決定をする際に，意思決定者は他者の選択を考慮に入れる場合がある．この時，意思決定者に潜在していた” 他の意思決定者の選択を考慮した目的” が現れ，目的を移行している．すなわち，意思決定者に潜在する m 個の目的のうち，k 個の目的を考慮してた時に，他者が存在することで，$k+1$ 個目の目的に移行する．意思決定者の集合を $Agent = \{agent_1, agent_2, ...agent_n\}$ とし，$agent_i$ が獲得した選好解を x^{agent_i} と定義する．式 (5) に他者の選択を考慮する目的に対する重要度 w_{k+1} を定式化する．

$$w^{agent_i}_{k+1} = f(\boldsymbol{x}^{agent_1}, \boldsymbol{x}^{agent_2}, ..., \boldsymbol{x}^{agent_{i-1}}) \tag{5}$$

$agent_i$ の重要度 w_{k+1} は $agent_{i-1}$ までの選択解集合 $X = \{\boldsymbol{x}^{agent_1}, \boldsymbol{x}^{agent_2}, ..., \boldsymbol{x}^{agent_{i-1}}\}$ によって決まる．

4. アプローチ

アプローチは，マルチエージェント系の下，多目的意思決定による目的移行のモデル化とマルチエージェント系全体の挙動の制御の二つである．

4.1 多目的意思決定による目的移行のモデル化

多目的強化学習における single-policy approach を用いて，マルチエージェント系の下，多目的意思決定による目的移行

Algorithm 1 モデルアルゴリズム

1: E : The maximum number of episodes
2: K : The number of objectives in single agent system
3: N : The number of agents
4: Initialize $TQ(s,a)$ arbitrary
5: **for** $h = 0$ to E **do**
6: **for** $j = 1$ to N **do**
7: initialize s
8: **repeat**
9: Choose a from s using policy derived from $TQ(s,a)$
10: Take action a and observe state $s' \in S$ and reward vector $\boldsymbol{r} \in \boldsymbol{R}$
11: $greedy_{a'}(s') \leftarrow$ choose a' from s' using policy derived from $TQ(s',a')$
12: **for** $i = 1$ to $k+1$ **do**
13: $Q_i(s,a) \leftarrow Q_i(s,a) + \alpha[r_i(s,a) + \gamma Q_i(s', greedy_{a'}) - Q_i(s,a)]$
14: **end for**
15: Compute $TQ(s,a)$
16: $s \leftarrow s'$
17: **until** s is terminal
18: **return** x^{*agent_j}
19: **end for**
20: $r_{k+1} \leftarrow r'_{k+1}$
21: **end for**

のモデル化を行う．single-policy approach を用いたモデルアルゴリズムを Algorithm1 に示す．従来の手法はシングルエージェント系を前提としていたが，本研究ではこれをマルチエージェント系に拡張する．従来のアルゴリズムと異なる点を赤字で示す．マルチエージェント系では，エージェントが複数存在するためエージェントの数を N と定義する．各エージェントは Q 学習を行うことで，最終的に最適な意思決定をする．多目的強化学習では，報酬がベクトルであるため Q 値が目的の数存在する．ングルエージェント系では k 個の目的を考慮するが，マルチエージェント系では，潜在していた，他エージェントの意思決定を考慮した目的が生まれ，$k+1$ 個の目的に対して Q 値を計算する．12〜14 行目で計算した Q ベクトルを，エージェントの各目的に対する重要度ベクトルによってスカラー化する．スカラー化の式を式 (6) に示す．

$$TQ(s,a) = \sum_{i=1}^{k+1} w_i Q_i(s,a) \tag{6}$$

17 行目において，エージェントが終端状態に着いたら，最終的な意思決定を行い，18 行目においてエージェントは選好解を獲得する．20 行目において，エージェントが意思決定を行うたびに報酬 R_{k+1} が動的に変化する．次に学習を行うエージェントは変化した報酬のもと学習を行う．全エージェントが 1 度ずつ学習をし意思決定するまでを 1 エピソードとし，最大エピソードである E エピソードまで，全エージェントの意思決定を繰り返す．

4.2 マルチエージェント系全体の挙動の制御

制御では，一般にシングルエージェント系での選好は不変だが，マルチエージェント系では選好が変動することに着目する．すなわちシングルエージェント系において k 個の目的を考慮してる場合，重要度ベクトル $\boldsymbol{w} = \{w_1, w_2, ..., w_k\}$ の各要素の値は変えられない．そこで，マルチエージェント系において生まれる目的に対する重要度 w_{k+1} を制御対象とし，目的を移行させることによって意思決定の変更を促す制御アルゴリズムを提案する．制御アルゴリズムを Algorithm2 に示す．全

Algorithm 2 制御アルゴリズム

1: **Given** : MOMDP$\langle S, A, T, \gamma, \boldsymbol{R}\rangle$
 $\boldsymbol{W} = \{\boldsymbol{w}^{agent_1}, \boldsymbol{w}^{agent_2} ... \boldsymbol{w}^{agent_n}\}$
2: **while** not converged **do**
3: **for** $i = 1$ *to* n **do**
4: $agent_i$ が多目的強化学習により選好解を獲得
5: $R_{k+1} \leftarrow R'_{k+1}$
6: **end for**
7: **end while**
8: 意思決定者の選好解の分布を確認
9: 理想の挙動が得られたら 10 へ
 そうでない場合は $w_{k+1} \leftarrow w'_{k+1}$，2 へ
10: **return** $\boldsymbol{W}$

エージェントの各目的に対する重要度を所与として，これを用いて単一目的にスカラー化し全エージェントが多目的強化学習により選好解を一回ずつ獲得する．これを 1 エピソードとする．1 エピソードが終了したら選好解の分布を確認する．選好解の分布から，それが理想の挙動である場合は全エージェントの重要度ベクトルを出力し，そうでない場合は全エージェントの重要度ベクトルのうちの w_{k+1} を網羅的に動かすことで目的を移行させ，理想の挙動を獲得するまで全エージェントの意思決定を繰り返す．

5. 計算機実験

計算機実験はモデル化と制御に分かれる．モデル化では，他者の選択を考慮して選択を変更することを確認する．制御では，意思決定者が均等に選択するように制御できることを確認する．

5.1 実験設定

実験環境は Deep Sea Treasure(DST) 環境 [Vamplew 11] を用い，図 3 に示す．DST 環境は 5 × 5 のグリッドで，黒いセルは海底を，グレーのセルは宝物を表す．潜水艦の行動は上下左右に 1 セル動くこととし，左上のセルからスタートして，宝物のあるセルに着いたらエピソードを終了する．本実験では潜水艦を操作するエージェント数を 30 とする．エージェントの目的は，宝物へのステップ数の最小化と，宝物の価値の最大化に，マルチエージェント系において新たに待ち時間の最小化が加わる．報酬は三つの目的に対してそれぞれ存在し，報酬ベクトル $\boldsymbol{R}$ として定義する．各目的に対する報酬と重要度の対応を表 2 に示す．一つはタイムペナルティーで 1 ステップごとに-1 を与える．二つ目は宝物の価値で，エージェントが宝物のあるセルに着いたら，そのセルの宝物の価値分報酬を与える．三つ目は既存の選択人数によるペナルティーで，自身より前のエージェントが 6 人以上選択した宝物のセルに着いた場合，獲得するまでの待ち時間が生じ，タイムペナルティーの-5 が与えられる．そのほかの宝物のセルに着いた場合はタイムペナルティーは与えられない．割引率 γ は 0.95 とした．

各エージェントが一度ずつ宝物を獲得し，全エージェントが宝物を獲得するまでを 1 エピソードとする．各エージェントは毎エピソード違う順番で宝物を獲得する．そこで各エージェントは，各宝物に対して，表 2 に示す [空，混] の状態を新たに観測する．例えば各宝の選択人数が [0.5,28,52,73,81]=[7,1,3,7,5] である場合 [混, 空, 空, 混, 空] をエージェントは観測する．

表 2: 各目的に対する報酬と重み

目的	報酬	重み
ステップ数の最小化	R_1 :1 ステップ毎に-1	w_1
宝物の価値の最大化	R_2 : 宝物の価値分 or 0	w_2
待ち時間の最小化	R_3 選択人数 0〜5 人 (空) → 0 選択人数 6〜10 人 (混) →-5	w_3

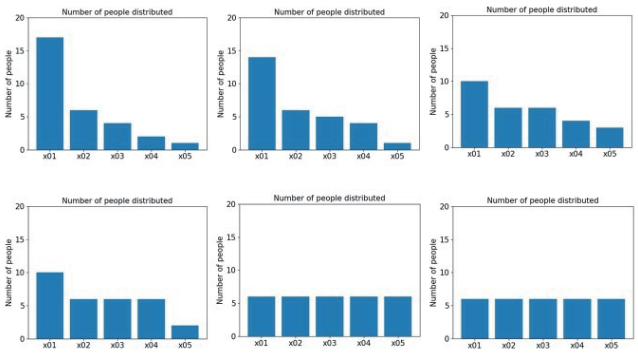

図 2: 制御による結果

5.2 モデル化

モデル化では，(0.95,0.05) の固有の重要度ベクトルを持つ 1 エージェントを対象とし，意思決定を観察する．一つの宝物が [混] の状態をエージェントが観測したときに，待ち人数に対する重み w_3 が生まれ目的を移行する．本実験では，w_3=0.1 と事前に設定し，対象エージェントの意思決定を様々な順番において観察した．図 3〜図 6 に実験結果を示す．

図 3 は 1 番に意思決定を行った時である．全ての宝物の状態が空であったため，2 目的による自身の固有の選好の下，1 番近くの 0.5 の宝物を獲得した．図 4〜図 6 では 6 人以上が既に選択した [混] の宝物があり，待ち人数に対する目的が生まれる．図 4〜図 6 のどれにおいても，エージェントは [混] の宝物を避けて，[空] の宝物を選択したことが確認された．

5.3 制御

制御では，30 の全エージェントを対象とする．現状 5 つの宝物の選択分布が幾何分布であると仮定して，これを一様分布に制御を行うことを目標とする．エージェント全体が考慮する目的を，選択人数を考慮した目的に移行させることによって，制御を行った．選択分布を図 2 に示す．スタートに近い宝物から x01，x02...x05 と表す．w_3 を大きくすることで選択分布は分散し，$w_3 = 0.6$ で理想な挙動を獲得した．$w_3 \geq 0.6$ においても，同じく一様分布であることを確認した．

6. まとめ

本研究では，多目的意思決定問題において，他の意思決定者が存在すると，潜在していた新たな目的が発生し，目的を移行することに着目した．また意思決定の多くは，逐次的意思決定が多いことから，多目的強化学習を用いて，多目的意思決定における目的移行のモデル化を行った．計算機実験では，エージェントが集中する解を避ける行動を獲得し，他エージェントの選択を考慮した意思行動を確認できた．また，意思決定者の目的移行により，意思決定を変更することに着目し，意思決定者全体の挙動の制御を行った．制御では，マルチエージェント系において発生する” 他者の選択を考慮した目的” に移行させることにより，系全体の理想な挙動に制御できたことを確認した．

本研究では，意思決定者の選好を事前に設定しモデル化と制御を行った．しかし実際は，意思決定者の選好が定量的にわかる場面は少ない．また，制御において，マルチエージェント系において発生する目的に対する重要度は，全エージェント一律の値で設定した．意思決定者の重要度ベクトルの設定の仕方は今後考えるべき課題である．

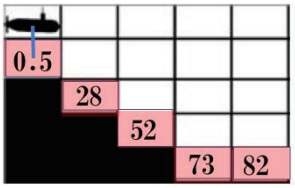

図 3: 実験環境
1 番目の意思決定

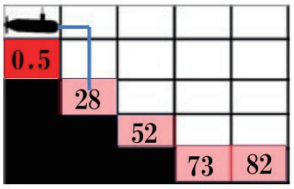

図 4: 12 番目の意思決定

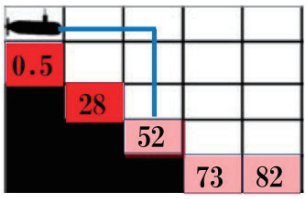

図 5: 18 番目の意思決定

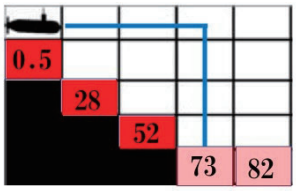

図 6: 23 番目の意思決定

参考文献

[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)

[Sutton 98] Sutton, Richard S., and Andrew G. Barto. Introduction to reinforcement learning. Vol. 135. Cambridge: MIT press, 1998.

[Roijers 13] Roijers, Diederik M., et al. ”A survey of multi-objective sequential decision-making.” Journal of Artificial Intelligence Research 48 (2013): 67-113.

[Vamplew 11] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker: “Empirical evaluation methods for multiobjective reinforcement learning algorithms”, Machine learning, 84(1-2), pp. 51-80 (2011)