F-018

マルチエージェント強化学習による多主体調停技術
Multi-agent Mediation Technology Based on Multi-agent Reinforcement Learning

犬塚 翔太† 寺本 やえみ†
Shota Inuzuka Yaemi Teramoto

1. はじめに

企業や人間の社会活動においては、社会という同一環境内で各々の目的に従って互いに意思決定を実行し合っている。企業や人間などの各主体はそれぞれ合理的な判断のもと、すなわち各々の目的に応じた指標を最適化する意思決定を実行する。この各意思決定というのは、社会に存在する状態の遷移と各指標に作用するため、自身の意思決定は他者の指標に影響を与え、また逆に他者の意思決定は自身の指標に影響を与える関係性を持つ。したがって、企業や人間というのは状況によって協力、連携することで互いの指標を向上し合う関係を築くことや、逆にライバル関係にある場合には、自身の意思決定によって相手を出し抜くということが起こりえる。

近年、環境問題やカーボンニュートラルなどの環境価値や、社会全体にとって有益な社会価値など、全主体に共通の指標(=全体指標)への貢献が重要視されている。この全体指標の向上という課題は、各々の主体単独では解決しきれず、主体間の協力、連携が必要である。しかしながら、この全体指標の向上に伴って、各主体に特有の指標(=個別指標)がトレードオフの関係となって悪化するケースが存在する。そのような場合、単に主体間で協力、連携して全体指標を向上させようとしても、個別指標のコンフリクトによって連携を阻害されてしまい、互いに合意の取れた意思決定を選択することが非常に難しくなる。また各主体には、相手に秘匿したい情報も存在するため、開示できる情報の中から最適な意思決定を選択しなければならず、これも合意に至る過程を難しくしている要因である。

本論文では、全体指標と個別指標のトレードオフの関係によって、社会的ジレンマに陥っている各主体を調停する多主体調停技術について述べる。2 章では社会的ジレンマについて述べ、続く 3 章ではマルチエージェント強化学習による多主体調停技術を述べる。4 章では社会的ジレンマのユースケースとしてサプライチェーンにおける複社購買問題を取り上げ、5 章で提案手法によるシミュレーション結果を示す。6 章において、本研究の結論を示す。

2. 社会的ジレンマ

全主体にとって共通の全体指標と、各主体に特有の個別指標に応じて各主体が意思決定を行う場合、各主体は全体指標と個別指標を重み付け和した多目的最適化問題を解き、意思決定を選択することとなる。ここで、1) 全体指標と個別指標はほとんどの場合トレードオフの関係、2) 各主体は全体指標に個別指標に対して相対的に大きすぎる重み設定をしていない、3) 各主体は互いに意思疎通を図れない、ということを仮定する。このとき、各主体は社会全体に有益な意思決定を選択するか、自身に都合の良い意思決定を選択するかの社会的ジレンマに陥る。このような場合、ゲーム理論の観点から問題を整理すると、「任意の主体にとって個別指標を最適化するのが支配戦略となり、互いに個別指標を最適化し合う」、または「一部の主体が全体指標を最適化し、主体間で指標の重みづけ和に不公平が生じる」のいずれかがナッシュ均衡となり、十分に全体指標を最適化することは非常に困難となる。

3. マルチエージェント強化学習による手法

本研究では、社会的ジレンマに陥っている各主体に対し、各主体の情報の秘匿性と主体間の公平性を担保しつつ、全体指標、個別指標の向上を目的としたマルチエージェント強化学習による多主体調停技術を開発した。多主体調停技術では、各主体ごとに代理AIを設け、各主体は開示可能な情報を代理 AI に開示し、各代理 AI は開示された情報をもとに、代理AI間で交渉を行い、最適な各意思決定の策定を行う。ここで、各主体 i から開示される情報をエージェントの観測$\boldsymbol{o_i}$とし、各エージェント i は方策$\pi^i(\boldsymbol{a_i}|\boldsymbol{o_i})$によって意思決定$\boldsymbol{a_i}$を実行する。実行された各意思決定$\boldsymbol{a_i}$に基づき、環境の状態$\boldsymbol{s}(t)$は状態遷移確率$P\left(\boldsymbol{s}(t+1)\middle|\boldsymbol{s}(t),\boldsymbol{a}_1(t),\dots,\boldsymbol{a}_{|I|}(t)\right)$によって遷移していく。ここで、$|I|$はエージェントの集合$I$の要素数を表す。これら状態遷移と各意思決定によって各エージェント i の報酬r^iが定まる。本研究では報酬r^iを以下のように定義する。

$$r^i = c_{global}r_{global} + c_{local}r^i_{local} + c_{fair}r^i_{fair} \qquad (1)$$

ここで、r_{global}は全体指標、r^i_{local}は主体 i の個別指標、r^i_{fair}は主体 i の公平性に関する指標、c_{global}、c_{local}、c_{fair}はそれぞれに対する重み係数である。公平性に関する指標r^i_{fair}は対象の問題に応じて公平性を定義し、設定する。各エージェント i は割引率$\gamma \in [0,1]$によるr^iの割引報酬和の期待値を最大化するように方策$\pi^i(\boldsymbol{a_i}|\boldsymbol{o_i})$を更新する。

本研究では、方策を深層学習によってモデル化し、勾配法によってパラメータ更新を行った。本研究の問題設定では、各エージェントは各主体が開示可能な情報のみから意思決定を策定しなければならない部分観測マルコフ決定過程となっている。これに対処するために、本研究では[1]に代表されるように方策を RNN(Recurrent Neural Network)によってモデル化を行った。また、代理 AI 間の交渉を TarMAC(Targeted Multi-Agent Communication)[2]によって実装し、エージェント間でコミュニケーションを取らせながら、協調的な意思決定が策定できる構成とした。

4. サプライチェーンにおける複社購買問題

本研究では、社会的ジレンマのユースケースとして図 1 のようなサプライチェーンにおける複社購買問題を取り上げる。複社購買問題とは複数の調達先から何割ずつ調達す

†株式会社 日立製作所 Hitachi, Ltd.

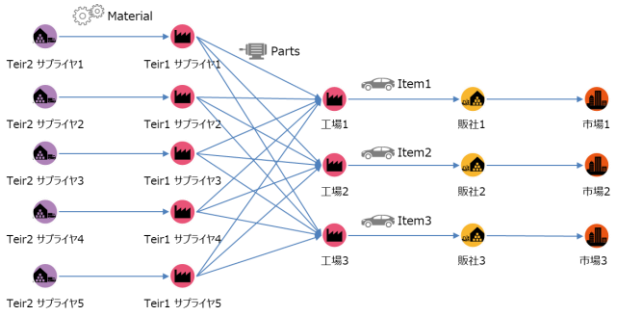

図 1　対象のサプライチェーンの構造

るかを決める問題であり、図 1 の 3 つの工場における Tier1 サプライヤに対する複社購買問題を考える。全体指標は 3 市場に対する需要充足率(OFR: Order Full Rate)の平均と Tier1 サプライヤで発生する製造過程における CO2 排出量の 2 種類を考える。個別指標は各工場における Tier1 サプライヤからの Parts の調達コストを考える。本研究では Tier1 サプライヤの生産能力はすべて異なる場合を想定し，生産能力が大きいほどエネルギー効率が良く，Parts1 個あたりの CO2 排出量が少ないが，Parts の価格は高価な場合を考える。このような設定においては，各工場が個別指標である調達コストを抑えるために，価格が安価な Tier1 サプライヤから Parts を調達しようとすると発注が 1 ヶ所に集中し，市場需要が大きい場面では Tier1 サプライヤの生産能力の観点から各工場十分な Parts の調達が行えず，市場に対する OFR を低下させてしまう。また，Parts1 個あたりの CO2 排出量も多いため，製造過程における CO2 排出量も増加してしまう。これら全体指標を向上させるためには各工場を調停する必要がある。

この複社購買問題に対して、本研究では各個別指標を各工場ごとに正規化し、正規化した個別指標の累積値とその全工場の平均値との偏差が任意の時刻であるδ以下になっている状態を公平性として定義した。この公平性を担保するために、tanh に平均値との偏差を適用した数値の符号を反転させ、公平性に関する指標r^i_{fair}として定義し、DDPG(Deep Deterministic Policy Gradient)[3]によって各工場の調達比率を行動として学習した。また、1 ステップ前の全ての市場における実測需要合計値, 1 ステップ先の全ての市場における予測需要合計値, 2 ステップ先の全ての市場における予測需要合計値, 工場における各 Tier1 サプライヤからの各調達合計値，工場における販社への出荷合計値，工場における OFR，各 Tier1 サプライヤの OFR，各販社の OFR，時刻$t = 0$から$t = t'$までの正規化された個別指標の累積値と全工場の平均値との偏差をエージェントが観測できる情報として定義し、各指標に影響を与える各工場の在庫や注文残高は秘匿性の高い不可観測な情報として扱った。

5. シミュレーション

4 章で導入した複社購買問題に対し、提案手法のシミュレーション検証を行った。シミュレーションでは[4]で述べられているシミュレータを用いた。シミュレータではシミュレーション実行前に，シミュレーション期間，市場需要の実測値・予測値，生産リードタイムや輸送リードタイム，生産能力など様々な値を設定でき，シミュレーションを実行すると設定値とステップごとに変更される各調達比率に

表 2　調停技術有無による比較結果

	市場に対するOFR (1エピソード間平均)	CO2排出量 (1エピソード間平均)
調停技術無し	46 [%]	119892.2
調停技術有り(DDPG)	66 [%]	59640.1
差	+20 [%] (43.5% 改善)	−60252.1 (50.3% 改善)

	個別指標(調達コスト) (1エピソード間平均)	公平性(最大 δ 値)
調停技術無し	−0.37 [−]	9.02 [−]
調停技術有り(DDPG)	−0.65 [−]	0.35 [−]
差	−0.28 [−] (75.7% 悪化)	−8.67 [−] (最大96.1% 改善)

基づいて，入庫実績，出庫履歴，注文残高，在庫数，各拠点の OFR，CO2 排出量などを算出し出力することができる。市場需要のデータは手動で設定した値を予測値とし、それにランダムノイズを加算したものを実測値としてシミュレーションを行った。

また，本研究による調停技術の定量的な有効性を確認するために，調停技術が存在せず，各工場が独立に意思決定を実行している場合と比較検討を行った。各工場が独立して意思決定を実行し合っている状態をシミュレーションで模擬するために，独立型 DDPG の実装を行った。独立型 DDPG では各工場が観測できる情報であればエージェントの観測として扱ってよいと考えられるため，各在庫情報や注文残高を観測できる形で学習を行った。

調停技術有無による比較結果は表 1 のようになった。調停技術が無い場合は、一部の工場が全体指標である市場に対する OFR と CO2 排出量を最適化し，それ以外の工場は個別指標を最適化するような，工場間で不公平が生じているナッシュ均衡に収束したため、最大δ値が大きくなっている。一方で提案手法では、各工場の在庫や注文残高を不可観測な情報として扱っているにもかかわらず、全体指標とほとんどの場合トレードオフの関係になっている個別指標の悪化を抑制しつつ，全体指標を大幅に改善し，かつ工場間の公平性も大幅に改善できている。

6. おわりに

本研究では、社会的ジレンマに陥っている複数の主体に対して、各主体の情報の秘匿性と主体間の公平性を担保しながら、全体指標、個別指標を向上させるような意思決定の提案ができる多主体調停技術に関して述べた。本手法は数理最適化やシングルエージェント強化学習でも実装可能であるが、個々の意思決定をマルチエージェント強化学習によって分散化することで、1 つ 1 つの方策の学習のスケールを低減することができ、調停する主体の数が大きくなった場合でも、ある程度性能を担保することが期待できるのではないかと考えている。

参考文献

[1] Kapturowski, Steven, et al. "Recurrent experience replay in distributed reinforcement learning." International conference on learning representations. 2019.
[2] Das, Abhishek, et al. "Tarmac: Targeted multi-agent communication." International Conference on Machine Learning. PMLR, 2019.
[3] Lillicrap, Timothy P., et al. "Continuous control with deep reinforcement learning." arXiv preprint arXiv:1509.02971 (2015).
[4] Kiuchi, Atsuki, et al. "Bayesian optimization algorithm with agent-based supply chain simulator for multi-echelon inventory management." *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2020.