5U-09

サプライチェーンにおける自動交渉のための エンベディングを用いた強化学習フレームワーク

宮島龍冴 †
東京農工大学 工学部 知能情報システム工学科 †
藤田桂英 ‡
東京農工大学 工学研究院 先端情報科学部門 ‡

1 はじめに

マルチエージェントシステムにおいて，それぞれの選好に従って行動するエージェント同士が交渉により競合を解消する自動交渉という技術が注目されている [1]．自動交渉の応用先の一つとしてサプライチェーンマネジメントがある．サプライチェーンにおける自動交渉では，複数のエージェントと並列に交渉を行う．最近では二者間交渉問題において強化学習により戦略を獲得する研究が行われている一方で，並列交渉に強化学習を導入した例はまだない．そこで本研究では，サプライチェーンにおける自動交渉において，相手の戦略モデルを考慮した強化学習のためのフレームワークを提案する．

2 問題設定

本論文では，エージェント A_0 がエージェント $A_1, A_2, \ldots, A_n$ に対して製品を売るために交渉を行うシナリオについて扱う．交渉の論点は取引する製品の単価と数量である．交渉プロトコルとしては，Alternating Offers Protocol[3] を並列交渉に拡張して用いる．エージェントは，交渉相手から提案を受け取った際に，その提案を受諾するか，拒否して新たな提案を送るかを選択する．

A Reinforcement Learning Framework Using Embedding for Automated Negotiation in Supply Chain
† Ryoga Miyajima, Faculty of Engineering, Tokyo University of Agriculture and Technology
‡ Katsuhide Fujita, Institute of Engineering, Tokyo University of Agriculture and Technology

3 提案手法

3.1 交渉問題の定式化

学習エージェント A_0 とエージェント $A_i(i = 1, 2, ..., n)$ の交渉をマルコフ決定過程 $MDP_i = \langle \mathcal{S}_i, \mathcal{A}_i, r_i, \mathcal{T} \rangle$ として定式化する．以下に状態，行動，報酬と方策について示す．

状態 $s_i \in \mathcal{S}_i$

状態には以下の４要素を用いる．

- 正規化した経過日数 d/D
- 正規化した交渉ラウンド数 r/R
- 前回自身がエージェント A_i に出した提案の効用値 $U(\omega_{r-1}^{\mathrm{for}i})$
- 直前にエージェント A_i から受けた提案の効用値 $U(\omega_{r-1}^{\mathrm{from}i})$

行動 $a_i \in \mathcal{A}_i$

行動は，エージェント A_i に提案する合意案の目標効用値 $u_{target}^{\mathrm{for}i}$ とする．

報酬関数 $r_i : \mathcal{S}_i \times \mathcal{A}_\rangle \to \mathbb{R}$

報酬として，合意形成時にはその効用値 $U(\omega)$, 交渉決裂時には -0.5，交渉継続時には次に相手の提案の効用値が上がれば 0.01，下がれば -0.01，変わらなければ 0 を付与する．

方策 $\pi_\phi : \mathcal{S} \times \mathcal{A} \to [0, 1]$

共通の確率的方策を用い，どの相手エージェントとの交渉においても適切な行動選択を行えるように最適化する．確率的方策にはベータ分布を利用する．

3.2 表現関数

交渉相手の戦略のモデリングに，Grover ら [2] によって提案された表現関数を用いる．表

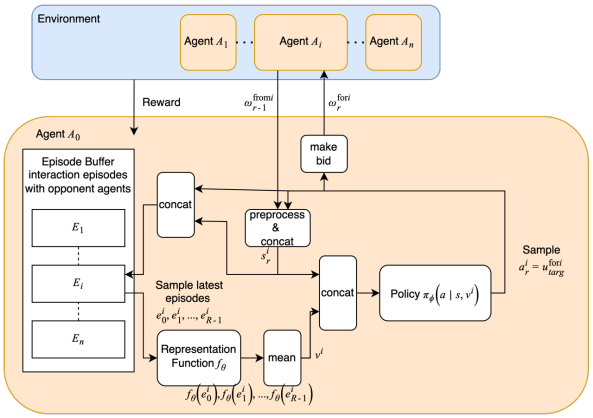

図 1　本稿で提案する学習フレームワーク

現関数 $f_\theta : \mathcal{S} \to \mathbb{R}^d$ は，エージェント A_i と交渉における状態を入力すると実数値ベクトルを出力する関数である．表現関数の学習は，過去の交渉データを用いて生成的表現と識別的表現を目的として教師なし学習を行う．

生成的表現と識別的表現の学習は以下の式 (1) で求められる目的関数 $J(\theta)$ を最大化するように行う．

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{e_+\sim E_i,\\ e_*\sim E_i}} \left[\underbrace{\sum_{\langle o,a\rangle\sim e_+} \log \pi(a|o, f_\theta(e_+))}_{\text{imitation}} - \underbrace{\lambda \sum_{j\neq i} \mathbb{E}_{e_-\sim E_j}\left[d_\theta(e_+, e_-, e_*)\right]}_{\text{agent identification}} \right],$$

$$d_\theta(e_+, e_-, e_*) = \left(1 + \exp\left(||f_\theta(e_*) - f_\theta(e_-)||_2 - ||f_\theta(e_*) - f_\theta(e_+)||_2\right)\right)^{-2} \quad (1)$$

ここで，$e \in E_i$ はエージェント A_i との交渉における 1 日分の状態，行動の組 (s, a) の，交渉終了時までの時系列データである．

3.3　学習フレームワーク

本稿で提案する学習フレームワークを図 1 に示す．学習エージェント A_0 がエージェント A_i から提案を受けると状態 s_r^i に遷移する．前日 1 日分の A_i との交渉データに対する表現関数 f_θ の出力を平均して得られる実数値ベクトル v^i を s_r^i と連結して方策 π_ϕ に入力して目標効用値を得る．目標効用値をもとに提案し，それに対して A_i のとる行動に応じて報酬を得る．

方策 π_ϕ の学習は 1 日の終わりに全ての相手エージェント $A_1, A_2, ..., A_n$ との交渉で得た報酬を用いて PPO により更新し，表現関数 f_θ は数日に一度，全ての 2 体の相手のパターン $A_i, A_j (1 \leq i, j \leq n, i \neq j)$ について式 (1) により更新する．

表 1　エージェントの獲得スコア

	提案手法	表現関数なし	ランダム
獲得スコア	1.32	0.95	0.92

4　実験

実験には交渉相手として，時間依存の戦略をとるエージェント 2 種類，行動依存の戦略をとるエージェント 1 種類，時間と行動の両方に依存する戦略をとるエージェント 1 種類の計 4 種類を用いる．またベースラインとして，ランダム戦略をとるエージェントと表現学習なしで同様の学習を行ったエージェントを用いる．この条件下で実験を行い，エージェントの獲得スコア (平均所持金増加率) を比較する．

実験結果を表 1 に示す．提案手法により学習したエージェントが，ベースラインと比較して高いスコアを獲得することを確認できた．

5　おわりに

本稿ではサプライチェーンにおける自動交渉のための表現関数を導入した強化学習を提案した．実験により提案手法の有効性を示した．

参考文献

[1] 産業競争力懇談会 2017 年度プロジェクト最終報告　人工知能間の交渉・協調・連携. http://www.cocn.jp/report/theme98-L.pdf, 2018.

[2] Aditya Grover, Maruan Al-Shedivat, Jayesh Gupta, Yuri Burda, and Harrison Edwards. Learning policy representations in multiagent systems. In *International conference on machine learning*, pages 1802–1811. PMLR, 2018.

[3] Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109, 1982.