4H1-4

繰り返しゲームでの強化学習アルゴリズムの組み合わせによる協調行動の学習

Learning to coordinate in repeated games using ensemble reinforcement learning

藤田 渉 *1 Wataru Fujita　森山 甲一 *2 Koichi Moriyama　福井 健一 *2 Ken-ichi Fukui　沼尾 正行 *2 Masayuki Numao

*1大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

*2大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

The interaction of people is modeled as games. Many researchers have proposed reinforcement learning algorithms to obtain most suitable strategies in games. However, state-of-the-art algorithms perform well in some games but poorly in others. We construct an algorithm that maximizes payoffs in various games by combining two reinforcement algorithms with complementary properties. Our proposed algorithm combines M-Qubed and Satisficing algorithm using Boltzmann multiplication ensemble method. An experimental study in which M-Qubed agents, Satisficing algorithm agents, and the proposed agents played ten games, shows that the proposed agent performed well in nine games.

1. 序論

我々人間は，日常生活において様々な意思決定を行っている．ある一人の人の意思決定は他の人々の意思決定に影響を及ぼし，同様に，他の人々の意思決定はある一人の人の意思決定に影響を及ぼす．社会的状況では，人々の間には利害の対立や競争，さらに協力などの複雑で多様な相互関係が混在する．このような人々の関係をモデル化したものをゲームと定義し，合理的な意思決定を分析するゲーム理論が広く研究されている．

また，人はその様々な状況や局面に応じて試行錯誤を行い，どのように行動するかを決定する．人間は本能的に欲求を持っており，その欲求を満たそうと行動を決定する．欲求が満たされない不快な状態を避け，欲求が満たされる快い状態を求める．このような行動をとるのは学習機構が人間の脳に存在するからである．そのモデルとして強化学習を導入する．

「相互作用の関係にある個々」が「自己の利益となる行動を学習する」状況をモデル化するため，人工的に構築した意思決定主体（エージェント）が強化学習アルゴリズムを搭載しゲームを行う問題を考える．ところが，既存の強化学習アルゴリズムは，ゲームの種類によって得意不得意があるという問題点を持つ．Crandall と Goodrich の研究 [1] の結果においても，アルゴリズムごとに高い利得を獲得できるゲームが異なっていることがわかる．

現実世界には多様で複雑な状況が存在するため，どのようなゲームにおいても利用範囲が広く合理的な判断を下せるアルゴリズムが必要である．したがって本研究では，様々なゲームで自己の利益を最大化する行動を学習する強化学習アルゴリズムを構築することを目的とする．

連絡先: 藤田 渉，大阪大学産業科学研究所沼尾研究室，
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1，
Tel: 06-6879-8426，Fax: 06-6879-8428，
E-mail: fujita@ai.sanken.osaka-u.ac.jp

2. 関連研究

人々のやり取りや関係をモデル化したゲームならびに，試行錯誤を通じて環境に適応する学習の枠組みである強化学習について紹介する．

2.1 ゲーム

人間は社会活動を行う中で，それぞれの人が目的を達成するためにどのような行動を取ればよいか意思決定を行う．人がその目的を達成できるかどうかは自分の意思決定だけでなく他の人の意思決定にも依存する．社会における様々な意思決定の相互関係を数理的に解析する理論をゲーム理論 [6] と言う．

ゲーム理論は，複数の意思決定主体がそれぞれの目的を達成するために相互に依存しあっている状況をゲームと定義し，これを分析する．ゲームにおいてプレイヤーは様々な状況や局面に応じて行動を選択する．そして，プレイヤーは自己の戦略に則って行動を決定する．プレイヤー同士の行動の組み合わせが決定するとプレイヤーに利得が与えられる．あるプレイヤーの利得はそのプレイヤー自身の行動だけでなく他のプレイヤーの行動にも影響されるので，自己の利得を最大化するためには他のプレイヤーの行動を十分に鑑みながら自分の行動を決定する必要がある．

2.2 強化学習

強化学習 [4] とは，環境との相互作用から学習して目標を達成する問題の枠組みである．意思決定主体をエージェントと呼び，エージェントの外部の全てから構成されエージェントが相互作用を行う対象を環境と呼ぶ．エージェントと環境は離散的な時間ステップ $t = 0, 1, 2, 3, ...$ において相互作用を行う．各時間ステップ t において，エージェントは環境の状態 $s^t \in S$（S は可能な状態の集合）を受け取り，これに基づいて行動 $a^t \in A(s^t)$ を選択する（$A(s^t)$ は状態 s^t において選択可能な行動の集合）．1 ステップ後に，エージェントはその行動の結果として数値化された報酬 $r^{t+1} \in \Re$ を受け取り，新しい状態 s^{t+1} にいることを観測する．各時間ステップにおいて，状態から可能な行動を選択する確率をエージェントの方策 π^t と呼ぶ．$\pi^t(s, a)$ はもし $s^t = s$ なら $a^t = a$ となる確率である．

3. 提案手法

様々なゲームにおいて最適戦略を学習するアルゴリズムを構築する手法について述べる．初めに，既存の優秀な強化学習アルゴリズムについて紹介し，それらが持つ問題点について述べる．それから，複数の強化学習アルゴリズムを組み合わせることでその問題点を解決する手法を提案する．

3.1 強化学習アルゴリズム

提案手法の基となる既存の 2 つの優秀な強化学習アルゴリズム M-Qubed [1] と Satisficing algorithm [3] を導入する．

3.1.1 M-Qubed

M-Qubed [1] は，状態における行動の価値を算出し，自己の最低限の報酬を確保しつつ，高い報酬を獲得するために相手プレイヤーと協調することを学習するアルゴリズムであり，多くのゲームにおいて優れた戦略を獲得することができる．M-Qubed では自己と他者の行動の組み合わせを状態 s と定義し，状態 s における行動 a の価値 $Q(s,a)$（Q 値）を学習するため Sarsa [2] が使用されている．Sarsa の更新則は以下の式で表される．

$$Q^{t+1}(s^t,a^t) = Q^t(s^t,a^t)+\alpha[r^t+\gamma V^t(s^{t+1})-Q^t(s^t,a^t) \quad (1)$$

$$V^t(s) = \sum_{a\in A} \pi^t(s,a)Q^t(s,a) \quad (2)$$

ここで s^t，a^t は時刻 t における状態と行動，r^{t+1} は a^t によりもたらされる正規化された報酬，α は学習率，γ は割引率，$\pi^t(s,a)$ は時間 t でプレイヤーが状態 s でとる行動 a の確率である．M-Qubed はこの更新則によって Q 値を求めた後，以下の「利益追求」,「損失回避」,「積極的探索」を目的とした 3 つの戦略によってエージェントの方策を決定する．また，相手プレイヤーが自分にとって最も報酬が少なくなるような行動をとった時に自分が最大の報酬を獲得するような戦略をマキシミニ戦略と定義し，この戦略をとった時に獲得できる報酬をマキシミニ値とする．

利益追求

M-Qubed はマキシミニ戦略 π^t_{MM} を取るかどうかを考慮しながら，現在の状態における最大の Q 値を持つ行動を選択する純粋戦略をとる．

$$\overline{\pi}^*(s) \longleftarrow \begin{cases} \arg\max_{a\in A} Q^t(s,a) \\ \quad \text{if} \quad \max_{a\in A} Q^t(s,a) > \frac{v^{MM}}{1-\gamma}, \\ \pi^t_{MM} \quad \text{otherwise.} \end{cases} \quad (3)$$

損失回避

$|A(s^t)|$ を取り得る行動の数，$H(\omega)$ を過去 ω 回分の全エージェントの行動の集合と定義し，許容可能な損失を $L^{tol} = 500\times|A(s_t)|\times H(\omega)$ と定義する．マキシミニ戦略をとることによって獲得できる報酬を v^{MM}，エージェントがこれまでに獲得した平均報酬を v^{avg} とし，時刻 t におけるエージェントの被損失を $L^{accum}(t) = \max(0, t[v^{MM} - v^{avg}(t)])$ と定義する．この時，被損失が許容可能な損失を超過するまで，その状態における最大の Q 値を持つ行動を選択し，超過してからはマキシミニ戦略を選択する．

$$\underline{\pi^*}(s) \longleftarrow \begin{cases} \arg\max_{a\in A} Q^t(s,a) \\ \quad \text{if} \quad \forall\tau \le t, L^{accum}(\tau) < L^{tol}, \\ \pi^t_{MM} \quad \text{otherwise.} \end{cases} \quad (4)$$

積極的探索

利益追求戦略はその瞬間において高い報酬を獲得できるが，より高い未来の報酬を考慮しない近視眼的な戦略に陥りがちである．また，損失回避戦略は高い報酬をもたらす相手プレイヤーとの協調行動を導くことができない．この問題を解決するために Q 値の初期値を最大の可能割引報酬 $1/(1-\gamma)$ に設定することで，M-Qubed は大局的な戦略を学習する．

M-Qubed はこれら 3 つの戦略を組み合わせて最終的な戦略を生成する．まず，損失を基にして利益追求戦略と損失回避戦略の組み合わせを行う．時刻 t における損失回避戦略 $\underline{\pi}^*(s)$ を選択する割合を β^t とし，利益追求戦略 $\overline{\pi}^*(s)$ を選択する割合を $1-\beta^t$ とすると，状態 s における M-Qubed の戦略は

$$\pi^*(s) = \beta^t\underline{\pi}^*(s) + (1-\beta^t)\overline{\pi}^*(s) \quad (5)$$

となる．ここで配分率 β^t を被損失 L^{accum} と許容可能損失 L^{tol} の比率に基づき，

$$\beta^t \longleftarrow \begin{cases} 1 \quad \text{if} \quad \exists\tau \le t \text{ such that } L^{accum}(\tau) \ge L^{tol}, \\ (\frac{L^{accum}(t)}{L^{tol}})^{10} \quad \text{otherwise} \end{cases} \quad (6)$$

とする．次に，M-Qubed 全体の戦略を構築する．S^* を全ての状態 s における最も高い Q 値の近傍の値を持つ行動の組み合わせの集合と定義し，S^{prev} をゲームにおける直近の時間ステップの間に M-Qubed が訪れた状態の集合と定義する．すなわち，$S^* \cap S^{prev}$ は M-Qubed が直近に訪れた Q 値の高い行動の組み合わせの集合となる．この集合が空の時，相手プレイヤーの戦略と M-Qubed の戦略が局所解に留まっていることを意味する．そのような状況で，潜在的に高い報酬を持つ解に到達するために探索をしなければならない．従って状態 s における M-Qubed の戦略を

$$\pi^t(s) \longleftarrow \begin{cases} \pi^*(s) \quad \text{if} \quad \beta^t = 1 \; or \; S^* \cap S^{prev} \neq \emptyset, \\ [1-\eta(t)]\,\pi^*(s) + \eta(t)\chi \quad \text{otherwise} \end{cases} \quad (7)$$

とする．ここで $\chi = 1/|A(s_t)|$，探索率 $\eta(t) \in [0,1)$ を $t\to\infty$ につれて $\eta(t)\to 0$ となるように設定する．

学習率 α は M-Qubed の最適戦略を求める性能に大きく寄与している．エージェントが潜在的な最適解を探索するために学習の速度を十分に遅くし，かつ探索を行っている間に被る潜在的な損失を小さくするために十分に速くしなければならない．従って学習率を

$$\alpha = \frac{1-(\frac{\zeta}{1-v^{MM}+\zeta})^{\frac{\zeta|A||H(\omega)|}{L^{tol}}}}{1-\gamma} \quad (8)$$

とする．ζ はパラメータである．

3.2 Satisficing algorithm

Satisficing algorithm (S-alg) [3] は，自分の満足度 (aspiration level) を計算し，満足度を超える報酬が得られる状態に留まろうとする強化学習アルゴリズムである．相手プレイヤーと高い確率で協調行動を示すことが評価されている．S-alg のアルゴリズムは以下の通りである．

1. 満足度の初期値 (α^0) を報酬の最大値 R_{max} から $2R_{max}$ の間の値にランダムに設定する

2. 次のことを繰り返す
   (a) 行動 a^t を選択

$$a^t \longleftarrow \begin{cases} a^{t-1} & \text{if} \quad (r^{t-1} \geq a^{t-1}), \\ \text{ランダム行動} & \text{otherwise.} \end{cases} \tag{9}$$

   (b) 報酬 r^t を受け取り，満足度を更新

$$\alpha^{t+1} \longleftarrow \lambda\alpha^t + (1-\lambda)r^t. \tag{10}$$

ここで $\lambda \in (0,1)$ はエージェントの学習率である．S-alg はシンプルなアルゴリズムだが最低でもマキシミニ値以上の解に収束することが示されている [3].

3.3 M-Qubed と Satisficing algorithm の問題点

紹介した 2 つの強化学習アルゴリズムは優秀だが，問題点がある．M-Qubed の問題点として，

- 複数の戦略を持つので戦略の配分の決定や学習に時間が掛かり，相手プレイヤーとの協調行動により唯一つの状態が最適解となるゲームにおいて，平均獲得報酬が S-alg よりも少なくなる

ことが挙げられる．一方，S-alg の問題点として，

- 相手プレイヤーがグリーディな戦略を行うアルゴリズムだった場合，アルゴリズム内の満足度が減少し，低い報酬に満足してしまい一方的に搾取されてしまうことや，
- 状態の探索が足りず，次善の報酬に満足し最適戦略を学習しない場合も存在する

ことが挙げられる．すなわち，全てのゲームにおいて高い報酬を獲得するオールマイティなアルゴリズムではない．この問題を解決するため，損失回避戦略により相手プレイヤーに搾取されることを回避し，積極的探索戦略により最適な状態を探索することができる M-Qubed によって S-alg の欠点を補い，協調行動を繰り返しの早い段階で学習することができる S-alg によって M-Qubed の欠点を補うことにより，様々なゲームで最適戦略を学習する強化学習アルゴリズムを構築する．

3.4 強化学習アルゴリズムの組み合わせ

本研究では，得手不得手が相補的な M-Qubed と S-alg という異なる強化学習アルゴリズムの戦略を組み合わせることで最終的な利得を最大化する戦略を学習することを提案する．提案手法により構築したエージェントは複数のアルゴリズムを保持し，複数のアルゴリズムがもたらす戦略を組み合わせて新たな戦略を作り出す．エージェントが選択した行動に基づいて，エージェント内の全てのアルゴリズムが同時に行動の価値や満足度の更新を行う．組み合わせの方法として Boltzmann multiplication (BM) [5] を用いる．BM は，構成するアルゴリズム j の方策 π_j^t に基づいて，各行動の選択確率を乗算して，ボルツマン分布によりエージェントの方策を決定する．各行動の選好度は

$$p^t(s^t, a[i]) = \prod_j \pi_j^t(s^t, a[i]) \tag{11}$$

で計算される．エージェントのアンサンブル戦略は

$$\pi^t(s^t, a[i]) = \frac{p^t(s^t, a[i])^{\frac{1}{\tau}}}{\sum_k p^t(s^t, a[k])^{\frac{1}{\tau}}} \tag{12}$$

で計算される．戦略を計算した後，行動を選択する．そして，エージェントを構成する全ての強化学習アルゴリズムは選択された行動とその実行結果に基づいて学習を行う．

4. 実験

提案手法により構築したアルゴリズムの性能を確かめるため，Crandall と Goodrich の論文 [1] より引用した 10 種のゲームを用いて行った実験について述べる．

M-Qubed の割引率 $\gamma = 0.95$，探索率 $\eta(t) = (0.04 \times 1000)/(1000 + \max_s \kappa^t(s))$ とし，$\max_s \kappa^t(s)$ をエージェントが訪れた各状態の回数のうち最も大きい数値とおいた．M-Qubed の学習率 α を式 (8) のように設定した．ただし $\zeta \in [0.05, 0.1]$ である．また，状態として用いる過去の行動の数 $\omega = 1$ とおいた．S-alg の学習率 $\lambda = 0.99$ に設定した．表 1 は Crandall と Goodrich の論文 [1] で用いられている 10 種の 2 人 2 行動非零和行列ゲームである．斜体はプレイヤーの利得和を最大にする解を表している．

	c	d
a	*1.0,1.0*	0.0,0.0
b	0.0,0.0	0.5,0.5

(a) Common interest game (CIG)

	c	d
a	*1.0,0.5*	0.0,0.0
b	0.0,0.0	*0.5,1.0*

(b) Coordination game (CG)

	c	d
a	*1.0,1.0*	0.0,0.75
b	0.75,0.0	0.5,0.5

(c) Stag hunt (SH)

	c	d
a	0.0,1.0	*1.0,0.67*
b	0.33,0.0	0.67,0.33

(d) Tricky game (TG)

	c	d
a	*0.6,0.6*	0.0,1.0
b	1.0,0.0	0.2,0.2

(e) Prisoner's dilemma (PD)

	c	d
a	0.0,0.0	*0.67,1.0*
b	*1.0,0.67*	0.33,0.33

(f) Battle of the sexes (BS)

	c	d
a	*0.84,0.84*	0.33,1.0
b	1.0,0.33	0.0,0.0

(g) Chicken (Ch)

	c	d
a	0.84,0.33	0.84,0.0
b	0.0,1.0	*1.0,0.67*

(h) Security game (SG)

	c	d
a	0.0,0.0	*0.0,1.0*
b	*1.0,0.0*	0.0,0.0

(i) Offset game (OG)

	c	d
a	*1.0,0.0*	*0.0,1.0*
b	*0.0,1.0*	*1.0,0.0*

(j) Matching pennies (MP)

表 1: 10 種の 2 人 2 行動非零和行列ゲーム

4.1 実験 1

同一のアルゴリズムを持つエージェント同士が対戦した場合に，提案手法を搭載したエージェントが最適戦略を学習するかどうか，また内部のアルゴリズムが互いの不得手なゲームにおいて補い合っているかどうかを確認するため，表 1 の 10 種の 2 人 2 行動非零和行列ゲームを 30 万回繰り返す実験を 50 回行い，平均獲得報酬を比較した．表 2 に 10 種のゲームにおける平均獲得報酬を各ゲームの最大利得和で割り正規化した値を示す．値が 1 に近いほどそのゲームにおいて最適戦略をとれていることを示す．表の太字は各ゲームにおける最も高い値を表している．

表 2: 各アルゴリズムを搭載したエージェント同士でゲームを行った時の平均獲得報酬を最大利得和で割り正規化した値

ゲーム	M-Qubed	S-alg	提案手法
CIG	0.998956	0.999868	**0.999880**
CG	0.974634	**0.999653**	0.977571
SH	0.999241	**0.999876**	0.999836
TG	0.966469	**0.999718**	0.973407
PD	0.917534	**0.999774**	0.926097
BS	0.984214	**0.999763**	0.985918
Ch	0.988794	**0.999838**	0.988785
SG	0.983062	0.700552	**0.984236**
OG	0.493440	0.499923	**0.609852**
MP	1.000000	1.000000	1.000000

Prisoner's dilemma (PD) では S-alg の平均獲得報酬が M-Qubed よりも高いことがわかる．これは，相手プレイヤーとの協調行動を学習するために M-Qubed が多くの探索を行うことにより，報酬の少ない状態にも訪れるため平均獲得報酬が少なくなっているためである．一方，Security game (SG) で M-Qubed がほぼ最適戦略を学習していることがわかるが，S-alg はできていない．これは，S-alg の探索が不十分で，次善の報酬に満足するほどアルゴリズム内の満足度が減少したためである．また，Offset game (OG) で提案手法は M-Qubed と S-alg よりも高い報酬を獲得していることがわかる．これらのゲームで提案手法では，内部のアルゴリズムが欠点を補い合うように戦略を出しあい，不得手とするアルゴリズムよりも高い平均報酬を獲得した．提案手法は Offset game (OG) を除く 9 種においてほぼ最適戦略を学習し，よりオールマイティなアルゴリズムとなった．

4.2 実験 2

10 種のゲームにおいて提案手法と既存のアルゴリズムで総当り戦を行った場合の平均獲得報酬より，提案手法と既存のアルゴリズムと比較を行い，提案手法の性能を確認する．表 1 の 10 種の 2 人 2 行動非零和行列ゲームを 30 万回繰り返す実験を 50 回行った．ゲームのプレイヤーによって非対称なゲームが存在するので，行プレイヤーと列プレイヤーを入れ替えて再度ゲームを 30 万回繰り返す実験を 50 回行い，平均獲得報酬を比較した．表 3 に 10 種のゲームでのそれぞれのアルゴリズムを搭載したエージェントが総当り戦を行った時の平均獲得報酬を各ゲームの最大利得和で割り正規化した値を示す．1 を超えた場合は，他のプレイヤーを搾取して各ゲームの最大利得和よりも高い報酬を獲得したことを表す．表の太字は各ゲームにおける最も高い値を表している．

S-alg は相手と協調することが最適戦略となりただ一つの状態が最適解となる Tricky game (TG)，Prisoner's dilemma (PD)，Chicken (Ch) において他のアルゴリズムと比べ高い報酬を獲得している．しかし，Coordination game (CG)，Battle of the sexes (BS)，Security game (SG)，Offset game (OG)，Matching pennies (MP) において相手プレイヤーと望む状態に背反が起き，相手プレイヤーがグリーディな戦略をとった場合に，アルゴリズム内の満足度が減少し低い報酬に満足したので相手プレイヤーに搾取された．提案手法は相手プレイヤーに搾取されず，また M-Qubed と S-alg が苦手なゲームにおいて最適戦略を学習することができた．

表 3: 各アルゴリズムを搭載したエージェントが総当り戦を行った時の平均獲得報酬を最大利得和で割り正規化した値

ゲーム	M-Qubed	S-alg	提案手法
CIG	0.999266	**0.999846**	0.999734
CG	1.064534	0.833159	**1.070472**
SH	0.999394	**0.999860**	0.999688
TG	0.974606	**0.986435**	0.975419
PD	0.936898	**0.966992**	0.942708
BS	1.033065	0.901039	**1.042997**
Ch	0.988841	**0.994097**	0.988638
SG	0.935408	0.802276	**0.935854**
OG	0.565524	0.265818	**0.623308**
MP	**1.079259**	0.838132	1.077609

5. 結論

人々が相互に影響しあう関係をモデル化した「ゲーム」において，個々が学習を行うことで自己の利益を最大にする戦略を獲得する強化学習アルゴリズムが研究されている．しかし，既存の強化学習アルゴリズムにはゲームの状況によって得意なゲームと不得意なゲームがあるという問題点があった．本研究では，得手不得手が相補的な関係にある強化学習アルゴリズムを組み合わせて，より高い利得を獲得するアルゴリズムを構築した．2 つの強化学習アルゴリズムを組み合わせて構築したエージェント同士が 10 種の 2 人 2 行動非零和行列ゲームを行った場合，9 種でほぼ最適戦略を学習することができた．また，得手不得手が相補的な強化学習アルゴリズムを組み合わせることにより，互いの弱点を補いあい，かつより高い報酬を獲得することを確認した．

参考文献

[1] J.W. Crandall and M.A. Goodrich, "Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning.", Mach Learn, 82: 281–314, 2011.

[2] G.A. Rummery and M. Niranjan."On-line Q-learning using connectionist systems", Technical Report TR166, Cambridge University Engineering Department, 1994.

[3] J.L. Stimpson and M.A. Goodrich, "Learning To Cooperate in a Social Dilemma: A Satisficing Approach to Bargaining", Proc. ICML, 728–735, 2003.

[4] R.S. Sutton and A.G. Barto, 三上貞芳・皆川雅章訳『強化学習』森北出版, 1998.

[5] M.A. Wiering and H. van Hasselt, "Ensemble Algorithms in Reinforcement Learning", IEEE Trans Syst Man Cybern B, 38: 930–936, 2008.

[6] 岡田章『ゲーム理論 新版』有斐閣, 2011.