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

満足化基準値共有を用いた社会的強化学習

Social reinforcement learning with shared reference satisficing

其田 憲明 *1 Noriaki Sonota　神谷 匠 *2 Takumi Kamiya　甲野 佑 *3 Yu Kono　高橋 達二 *1 Tatsuji Takahashi

*1東京電機大学理工学部 School of Science and Engineering, Tokyo Denki University

*2東京電機大学大学院 Graduate School of Tokyo Denki Univerity

*3株式会社ディー・エヌ・エー DeNA Co., Ltd.

People and animals learn not only through individual trial-and-error, but also from other individuals. It is known that vertebrates cleverly utilize learning strategies such as copy-when-uncertain and copy-successful-individuals. These strategies can be applied to social reinforcement learning, although their formalizations are yet to be established. We propose a social reinforcement learning algorithm with a very narrow information sharing. The algorithm exploits RS value function that models the satisficing principle for exploration and exploitation.

1. はじめに

強化学習では, エージェントが試行錯誤を通じて未知の環境において収益を最大化する行動系列を学習する. 試行錯誤における環境の探索と知識の活用にはトレードオフが存在し, エージェントは両者のバランスを調整しながら学習を行うが, 実環境のような連続状態空間を扱う場合に適切な探索と知識活用を行うことは非常に困難である.

一方で, 人間はある目的水準を満たすことを目的とした場合に満足化原理 [Simon56] と呼ばれる, 現状が非満足状態であれば探索を行い, 満足状態であれば知識活用を行うように意思決定を行うことで, 探索空間の広大な未知の環境であっても目的とする基準を満たすような行動を可能としている. この満足化原理を反映した意思決定手法として, 満足化価値関数 (reference satisficing: RS) が考案された. RS は多腕バンディット問題をはじめとした様々なタスクにおいて少数の探索で素早く満足する行動系列を獲得でき, 適切な基準値を与えられると最適な行動系列を獲得できることがわかっている [高橋 16]. 適切な基準値が不明な場合 RS は動的に基準値を獲得する必要があるが, 効果的な基準値の決定手法は未だ考案されておらず RS の大きな課題となっている.

また, 人間含む脊椎動物は社会的な学習を巧みに行なっており, 社会的学習の戦略として行為の成否が不確実な場合や, 成功している個体がわかっている場合に他の個体の模倣をすることが知られている [Laland 04]. このような社会的学習の性質を利用した複数のエージェントによる群学習は, 単体のエージェントによる学習よりも素早く学習が進められることがわかっている. 本研究では, 複数のエージェントが探索により獲得した情報のうち, 推定された価値から最も良いものを新しい基準値として共有することにより, 社会的な環境における基準値の動的獲得と社会性の表現が可能であるか検討する.

2. 満足化価値関数と満足化方策

強化学習において満足化価値関数 RS は, 状態行動対 s_i, a_j に対する試行量 $\tau(s_i, a_j)$ と行動価値 $Q(s_i, a_j)$, 満足化基準値 $R(s_i)$ から以下の式のように定義され, RS による評価値を最大化するような行動を選択する方策を満足化方策と呼ぶ.

$$RS(s_i, a_j) = \tau(s_i, a_j)(Q(s_i, a_j) - R(s_i)) \tag{1}$$

$$a^{\text{select}} \leftarrow \arg\max_{a_k} \Big(RS(s_i, a_k)\Big) \tag{2}$$

$\tau(s_i, a_j)$ は以下の式で定義され, γ_τ は試行量割引率, α_τ は試行量学習率を表す.

$$\tau(s_i, a_j) = \tau_{\text{curr}}(s_i, a_j) + \tau_{\text{post}}(s_i, a_j) \tag{3}$$

$$\tau_{\text{curr}}(s_i, a_j) \leftarrow \tau_{\text{curr}}(s_i, a_j) + 1 \tag{4}$$

$$\tau_{\text{post}}(s_t, a_t) \leftarrow \tau_{\text{post}}(s_t, a_t) + \alpha_\tau\Big(\gamma_\tau(s_{t+1}, a_{t+1}) - \tau_{\text{post}}(s_t, a_t)\Big) \tag{5}$$

満足化方策では, 推定価値が満足化基準に達しているか否かによって探索と知識利用を切り替える. 基準値を満たす行動が存在しなければ非満足となり楽観的探索を行い, 基準値を満たす行動が存在すれば満足し知識の悲観的活用を行うことで, 少数の探索によって満足な行動系列を素早く学習できる. しかし, 未知の環境に対する適切な基準値の推定は困難で, 探索によって得た知識から動的に獲得する必要がある.

3. 社会的な満足化基準決定

脊椎動物は不確かなものや, 成功している個体の行動の真似をするといった習性が見られる. 人間社会においてもスポーツなどを始めとした様々な分野において, 新たな世界記録保持者が誕生すると今までの記録が次々と抜かれてしまい, 全体のレベルが押し上げられることがある.

強化学習も同様に, エージェント単体で学習を行う場合よりも複数エージェント間で価値を共有し学習を行うことで単体エージェントに比べ効率的に学習できることがわかっている [飯間 06]. しかし価値の共有のみでは, 他のエージェントと同様に行動できるがさらに良い行動を求めて探索することが難しく, 結果として最適な行動を学習できない可能性がある. そこで本研究では, 価値をそのまま共有するのではなく, より良い行動系列が存在するというメタ情報として共有することで, 意思決定における満足化基準を動的に更新しエージェントに探索を促す社会的な満足化基準の決定手法を提案する.

連絡先: 高橋達二, 東京電機大学理工学部, 350-0394 埼玉県比企郡鳩山町石坂, 049-296-5416, tatsujit@mail.dendai.ac.jp

3.1 行動価値の共有

独立した環境上で駆動するエージェント間の行動価値 Q の共有手法として, K 体のエージェント $(0...K)$ の中の最良行動価値 $Q^{\text{best}}(s,a)$ で各エージェントの行動価値を更新する手法が存在する (式 (6)). 他にも各エージェントの情報を残すように $Q^{best}(s,a)$ と自身の $Q(s,a)$ を平均する手法がある.

$$Q(s,a) \leftarrow Q^{\text{best}}(s,a), (\forall s,a) \tag{6}$$

3.2 メタ情報の共有による満足化基準の動的更新

提案手法では式 (6) 式を参考に, 独立した環境上で駆動する各エージェントが 1 試行で得た結果から各状態 s の最良の行動価値である $\max_a Q^{\text{best}}(s,a)$ を全てのエージェントの基準値 $R(s)$ として共有し, 以下の式のように動的に更新する.

$$R(s) \leftarrow \max_a Q^{\text{best}}(s,a), (\forall s) \tag{7}$$

この基準値 R の動的更新によって社会的な行動学習を行えるか, 決定論的バンディット問題によって評価する.

4. 決定論的多腕バンディット問題

決定論的バンディット問題では, 確率的バンディットとは異なり, マシンを選択すると既定の報酬値を確実に獲得できる. また, 獲得できる報酬が多いマシンほど, 線形的に数が少なくなる. 決定論的多腕バンディット問題を用いた理由として, 得られる報酬が決定論的であるため既に探索した行動と未探索の行動の区別や, 最適報酬と非最適報酬の割合などの報酬分布の操作が容易であることが挙げられる.

4.1 シミュレーション設定

報酬 R と対応するマシンの台数 x の関係 (R,x) がそれぞれ, $(0.0,10),(0.1,9),(0.2,8),...,(0.9,1)$ となるように設定した. 合計 55 台のマシンから方策に従いマシンを選択し, 対応した報酬を受け取る. これを 1 エピソードとして 1000 エピソード行い, 累積報酬が最大となる方策を獲得することを目的とした. 提案手法を用いたエージェントを RS エージェントペアと呼び, 比較対象として, 最適基準値 RS, UCB1-Tuned, ϵ-greedy 単体,ϵ-greedy のペア, そして greedy エージェントを用いた. greedy エージェントを除く各エージェントの行動価値 Q は, 初期値を 0 とし, 獲得報酬の平均値である. RS エージェントペアは, 満足化基準値 R の初期値を 0 として 1 回の試行毎に自身の最良行動価値 $Q^{best}(s)$ を相手に通知し, より高い方を基準値 R として更新した. 最適基準値 RS は基準値を 0 から更新する RS ペアと適切な基準を初期から所持する RS との性能比較をするためであり, このタスクで得られる最大報酬である 0.9 と次に大きい 0.8 の間である 0.85 を基準値とした.

行動決定手法として代表的である ϵ-greedy は, ϵ の初期値を 1.0 に設定し, 単体の場合は 1 エピソード毎に 0.0025 ずつ減衰させ, 400 エピソード終了時点で ϵ を 0 になるように設定した. ペアの場合は, 同じ環境下でエージェントが行動するため行動価値 Q を共有し, ペアで行動するため ϵ を 1 エピソード毎に 2 倍の 0.005 ずつ減衰させ, 200 エピソード終了時点で 0 になるように設定した. greedy エージェントは各行動を 1 度試行した上で最高報酬を選び続けた場合と RS ペアの成績を比較をするため, 各状態の行動価値の初期値を 1.0 に設定した.

4.2 結果と考察

図 1 と図 2 に, 1000 回のシミュレーションを平均した結果を示す. 図 1 は期待損失である, regret の推移を表したグラフである. また, RS ペアに関してはエージェント数で平均したものをグラフに利用している. ϵ-greedy 単体と ϵ-greedy ペアを比較すると, ϵ-greedy ペアは ϵ-greedy 単体の 2 倍の速さの ϵ 減少で regret の上昇をほぼ抑えられている. このことから ϵ-greed ペアは単体よりも 2 倍に近い速さで学習を行うことに成功していると考えられるが, RS ペアはさらに regret を半分程度に抑えることに成功している. そして, RS ペアと greedy エージェントの regret が一致しているため, RS ペアは各腕を 1 回ずつ試行した上で最高報酬となる腕を選択していることがわかる. RS ペアと最適基準値 RS を比較すると, 最適基準値 RS は RS ペアよりも regret を低く抑えられている. これは RS ペアは各行動を最低でも 1 度は試行する必要があるのに対し, 最適基準値 RS は報酬 0.9 を得られる行動を発見した時点で満足するためであると考えられる. そして, RS ペアと最適基準値 RS の regret の上昇が停止するタイミングがほぼ一致しているが, これは最適基準値 RS が 55 回目に 0.9 の腕を引く可能性が存在しているためである.

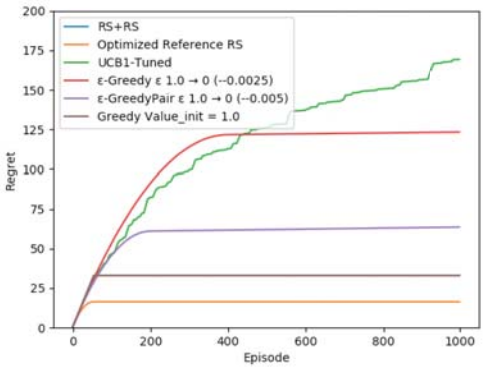

図 1: regret の推移 (バンディット)

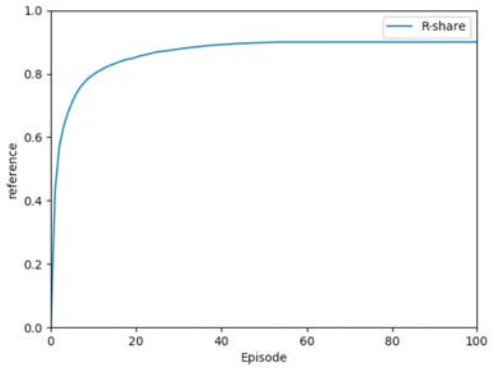

図 2: RSpair 基準値の推移 (バンディット)

図 2 は,RS ペアの共有基準値の推移を表したものだが, 最適基準値 R_{opt} となる範囲である $0.8 < R_{opt} < 0.9$ には 20 エピソード時点で到達しており, 60 エピソードを超えたあたりで最適基準値の 0.9 に到達している. 以上から, RS ペアは最適基準値 RS よりも劣るものの, ϵ-greedy ペアよりも良い成績で最適行動を見つけることに成功しているため, お互いの最良行動価値から基準値を更新する方法は効果的であると考えられる.

次に探索領域の変化と情報共有を行うエージェント数の変化が成績にどのような変化をもたらすかを検証する.

5. 決定論的ツリーバンディット問題

社会的な学習において, 同一環境下で学習を行うエージェントの数や, 探索領域の変化が学習効率に影響を与えると予想される. 決定論的バンディットの性質を維持したまま, エージェント数の変化と探索領域の変化の関係性を観測するため, 満足化方策を用いた先行研究である [牛田 16] で用いられたツリーバンディットと組み合わせた, 決定論的ツリーバンディット問題を行なった.

決定論的ツリーバンディットを用いた理由として, 決定論的多腕バンディットと同様に既に探索した行動と未探索の行動の区別が容易であることに加え, 遅延報酬を考慮した行動系列の複雑さ, 探索領域の変化を層の深さとして容易に管理できることが挙げられる.

5.1 シミュレーション設定

最適累積報酬が得られる行動系列の不確定性, 可変式である層に対応するため, 各ノードのバンディットの設定は, [0.1, 0.2, 0.3, 0.5], [0.1, 0.2, 0.8, 0.9], [0.1, 0.3, 0.4, 0.6], [0.1, 0.3, 0.6, 0.7], [0.3, 0.5, 0.6, 0.8], [0.4, 0.5, 0.6, 0.7] のうちからランダムに一つずつ当てはめる. 始点から終点にたどり着くまでを 1 エピソードとして 5000 エピソード行い, 累積報酬が最大となる方策を獲得することを目的とした. Q 値の更新に用いる学習率 α を 0.1, 割引率 γ を 1.0 に設定した. そして, 信頼度 τ の更新に用いる試行量学習率 α_τ0.1, 試行量割引率 γ_τ を 0.9 に設定した. エージェント数による成績の差を比較するために, RS エージェントに関してはエージェント数を 1, 2, 3, 4, 5 とした 5 通りのグループを用意し, 1 エピソード終了毎に自身の最良行動価値 $Q^{best}(s), (\forall s)$ をグループに通知し, 最も高いものを $R(s), (\forall s)$ として更新した. 探索領域の変化が各グループの成績にどう変化を及ぼすか観測するため, 2 層と 5 層のツリーバンディットを行なった.

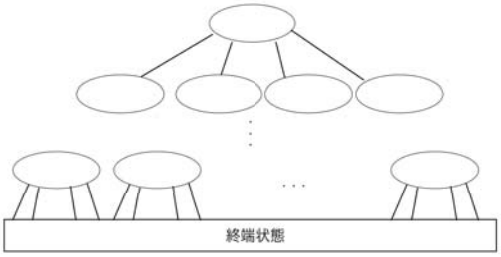

図 3: ツリーバンディットの概要

5.2 結果と考察

図 4 以降に 1000 回の結果を平均した結果を示す. また, 図 6 と図 9 は始点における基準値 R の推移である. そして, RS グループに関してはエージェント数で平均した物をグラフに利用している.

図 4 から 2 層の場合, エージェント数が 2 体以上であると最適累計報酬が得られる最適行動系列を発見することに成功しているが, 1 体の場合は失敗している. これは図 6 と RS アルゴリズムの性質からエージェント数が 1 体のみの場合, 始点の選択肢は全て最低 1 回ずつ選択され, 最初に 2 回以上選択された行動以外は基準値を満たすことができなくなる. そのため, 唯一基準を満たしている行動が非最適行動である場合, 最適行動系列の発見が不可能になる事態に陥ったと考えられる. また, 5 層の図 7 からは, 探索領域に拡大に伴い, エージェント数が多いグループほど最終的に得られる報酬が大きくなるが, 図 7 と図 8 から反対に初期の探索段階での獲得平均報酬は少なくなっているため, regret の上昇が大きくなっている.

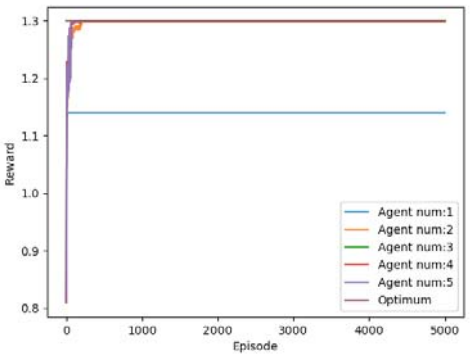

図 4: 獲得平均報酬 (2 層)

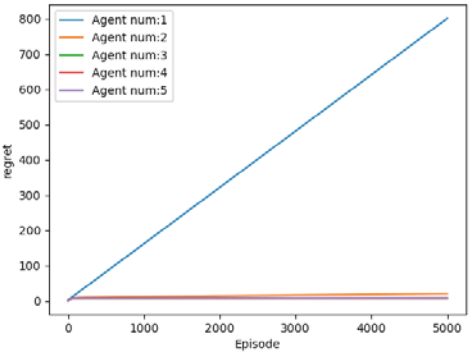

図 5: regret 推移 (2 層)

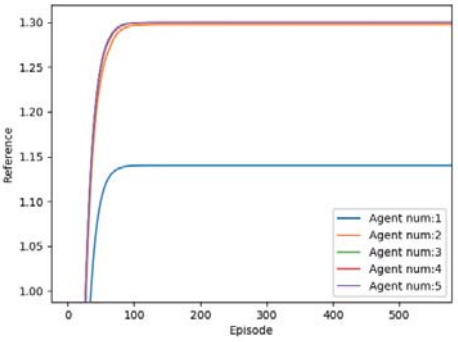

図 6: 基準値 R の推移 (2 層)

一方, 図 6 と図 9 から探索領域が拡大されるに従ってグループのエージェント数が多いほど素早く, より高い基準値に設定されていたことがわかった. これは, エージェント数が多いほど初期段階で得られる報酬が確率的に大きくなるからであると考えられる. これらから, エージェント数が多いほど基準値が高く設定されるため, 図 8 などに見られるように序盤の regret が大きく上昇していると考えられる. このことから, 探索領域が大きくなるに従って同一環境下で探索するエージェント数が多くなるほど, 最終的により良い結果が得られていたため, エージェント数が多いほど社会的学習が効率よく行われることがわかった. しかし, エージェント数が増えるほど序盤の regret が大きく上昇されるため, 低く抑えるための工夫が必要であると考えられる.

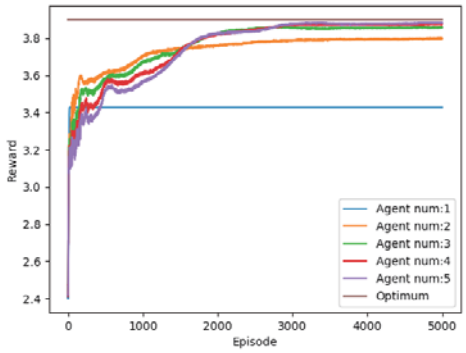

図 7: 獲得平均報酬 (5 層)

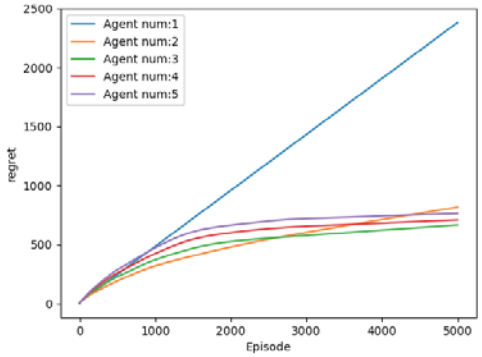

図 8: regret 推移 (5 層)

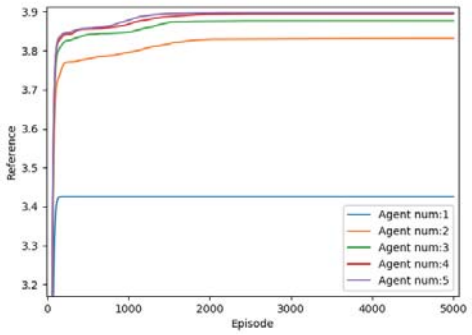

図 9: 基準値 R の推移 (5 層)

6. 総合考察

基準値の動的獲得の面に関して，1 体で学習を行なった場合は早期に基準値の更新が行われなくなった．それに対して 2 体以上のエージェントで学習を行なった場合, 基準値を社会的に更新することに成功しており，エージェント数が増えるほど，より高く適切な基準値に設定することに成功していた．

また, 獲得平均報酬の推移から 1 体で学習を行う場合は基準値の更新が早期に行われなくなるため, より良い行動を発見することが出来ずに局所解で満足することが見られた．それに対して，2 体以上のグループで行う社会的学習において, 他者のより良い成績から自身の既知である行動よりも報酬を得られる行動を探索することで, グループ全体でより高い報酬を得ることに成功している．これら結果を得られた主な要因としては, エージェント数の変化によって一度のエピソードで選ばれる行動が増えることで, 確率的に設定される基準値が高くなること, そしてエージェント数が多いほど基準値がより高い値に素早く設定されるため, より多い報酬を得るための探索に素早く移ることができるためであると考えられる．

以上から，人間社会の競技などにおける新たに良い成績が記録されると, そのほかの競技者の探索によって過去の成績を次々と抜いていき, 全体のレベルを向上させるといった社会的な学習を行うことに成功していると考えられる．

しかし, 今回は決定論的なタスクのみを扱っていること, エージェント数が多くなるに従って序盤の regret の上昇が大きくなっていることから, 確率的にタスクへの適応に向けたより良い基準値共有手法の発見, そして今回情報共有を行っていない試行量と行動価値 Q を共有する方法を模索することで改善できるのではないかと考えられる [柄谷 85].

7. 結論

本研究ではグループ内での最高成績を満足化基準値として共有することで, RS エージェントグループが社会的に基準値の動的獲得を行うことが出来るのか, そしてエージェント数の変化と探索領域の変化が成績にどう変化をもたらすかを検証した．RS エージェントグループはグループ内での最高成績から社会的に基準値を更新し, さらに良い行動を探索をすることで最終的により良い行動を発見することに成功していた．これらの結果により最小限の情報共有 (基準値 R) を利用した, 満足化による社会的な最適行動の探索・学習側面をシミュレーションによって端的に示すことができた．しかし, エージェント数が増えるに従って序盤の regret が大きく上昇する現象が見られたため, 今回情報共有を行っていない試行量と行動価値 Q を利用することで, さらに良い社会的学習手法を模索することが可能であると考える．

参考文献

[Laland 04] Laland, K.N.:Social learning strategies, Learning & Behavior, 2004, 32(1), 4-14 (2004)

[Simon56] Simon, H.A.:Rational choice and the structure of the environment, Psychological Review, Vol. 63, No.2, pp. 129-138 (1956)

[飯間 06] 飯間　等, 黒江康明:エージェント間の情報交換に基づく群強化学習法, 計測自動制御学会論文集, Vol. 42, No. 11, 1224/1251(2006)

[牛田 16] 牛田有哉, 甲野佑, 浦上大輔,　高橋達二:探索割合を自立調整する強化学習手法-満足化基準の動的獲得-, JSAI 2016 (2016 年度人工知能学会全国大会 (第 30 回)) 予稿集, 1M2-3 (2016)

[高橋 16] 高橋達二, 甲野佑, 浦上大輔：認知的満足化-限定合理性の強化学習における効用, 人工知能学会論文誌, Vol. 31, No. 6, pp. 111 (2016)

[柄谷 85] 柄谷 行人：ブタに生れかわる話, 批評とポスト・モダン, pp. 257260. (1985)