1F3-5in

因果強度を用いた強化学習における価値配分手法

Value deployment method with causal heuristics in reinforcement learning

小川絢加 *1 Ayaka Ogawa　澤山熱気 *1 Atsuki Sawayama　甲野佑 *2 Yu Kohno　高橋達二 *1 Tatsuji Takahashi

*1東京電機大学 Tokyo Denki University　*2東京電機大学大学院 Graduate School of Tokyo Denki University

One of the biggest issues in reinforcement learning is how an agent should interpret delayed rewards that are considered given to a series of previous state-actions. It takes a causal inference in which one goes from the effect in focus (high reward) to candidate causes (series of actions at the states) to assign the value for the rewards to the actions at the states that generated them. One of the standard technique for the issue is TD(λ). We propose a method that efficiently assigns and propagates using causal heuristics that human being use in causal induction.

1. はじめに

人間は報酬を得たとき，その報酬を得たという結果が何に起因するかを考える．そして結果と原因の関係を把握する事で，報酬を得るためには原因としてどのような行動を取れば良いかを理解して，効率的に報酬を得る事が可能となる．試行錯誤から，環境から得られる報酬を最大化する行動を見つけ出す事を目的とする強化学習という枠組みでは，前述したような報酬と行動系列の関係を把握して，その行動系列に対して適切に価値付ける事が，良い行動を学習する上で重要となる．環境の構造に依存しない学習手法であるモデルフリーな強化学習では，価値を行動系列上の過去に向かって伝播させる事が，環境にマルコフ性を想定する事で，行動系列に対する適切な価値付けとして成り立つ [Sutton 00].

しかしながら，この仮定は環境側のマルコフ性が弱い場合に成立しない．その場合，環境の構造を獲得したり，隠れマルコフを想定する等して，価値付けの方法を大きく変更する必要がある．本研究では前述した手法とは異なるアプローチとして，行動系列から報酬と原因を推定し，その因果関係の強さを学習に応用する事を提案する．従来，正確な因果関係の強さを強化学習で扱う行動系列の要素である “状態行動対” に対して定義するのは困難であった．我々はこの問題を解決するため，服部 [服部 01] や高橋ら [甲野 10, 高橋 14] が考案した，人間の因果関係の強さの推定と高い相関を持つ因果強度モデルに着目した．これは人間の因果強度モデルが簡潔な定義を持つため，比較的応用が容易であると考えたためである．本研究ではこのような人間の因果強度モデルと適格度トレースを組み合わせて，因果関係の強さに応じて行動系列に価値を配分する手法を考案し，既存アルゴリズムや因果強度モデル間の比較を強化学習課題シミュレーション上で行う事で，因果関係の推定が強化学習にどのような影響を与えるか考察した．

2. 強化学習

強化学習とは目的を達成する行動を試行錯誤的に学習する機械学習の一分野である．強化学習において学習エージェントは環境のある状態 ($s_i \in S$) にいる時，取り得る行動 ($a_j \in A_i$) から選択を行い，それによる状態の変化と報酬から目的を達成する行動系列を学習していく．ここで言う目的とは環境から得られる報酬を最大化する事に他ならず，すなわち環境から与えられる報酬という信号は獲得したい行動をどの程度達成しているかの指標であると捉えられる．

2.1 マルコフ性と遅延報酬

報酬は行動に対して即時的に与えられるとは限らず，ある行動系列の最後に与えられる場合が多い．このような報酬を遅延報酬と呼ぶ．強化学習では，得られた報酬を行動系列の各要素に対して，得られた報酬への貢献度を考慮して適切に利益配分する必要がある．通常のモデルフリーな強化学習アルゴリズムでは環境にマルコフ性を想定する事でこの問題を解決している．マルコフ性とは現在の環境の状態量が決まったとき，その後の状態変化が確率的に決定する事を意味する．そのため，報酬を価値として行動系列に対して時間的な過去に価値を伝播していく事で，報酬の配分を適切に行っている事になる．この性質を利用したアルゴリズムがモデルフリー強化学習の代表的な一つである TD 学習である [Sutton 00]．しかしながら，状態量の観測が不完全である等の理由で，エージェントが観測する環境のマルコフ性が弱まったり失われた場合，前述した手法では対応できない．その場合，環境側の行動をエージェント内でモデル化する等の方法で適切な価値配分を行えるような工夫が必要となる．

2.2 マルコフ性とモデルフリー学習

強化学習アルゴリズムはモデルベース学習とモデルフリー学習という二つに大別される．代表的なモデルベース学習として TD 学習に分類される Q 学習，Q(λ)，Sarsa，Sarsa(λ) や，Dyleyed Q-Learning 等がある．代表的なモデルフリー学習としては Dyna-Q，E3，R-MAX 等がある．モデルベース学習の特徴として過去の経験から内部環境モデルを構築し，先に起きることをシミュレーションして現在の状態遷移を決める事で間接的に価値関数を求めることが挙げられる．マルコフ性が弱い複雑な課題では，このようなモデルベース的なアルゴリズムの重要性が増してくる．しかしながら問題への実装の簡便さ等の扱い易さの面でモデルフリー学習の方が汎用的である．そこで我々はモデルフリーをベースにして，モデルベース的な要素を取り入れられないかと考え，因果関係の強さに応じた価値配分を行う学習手法を考案した．

連絡先: 高橋達二，　東京電機大学，　350-0934 埼玉県比企郡鳩山町石坂，　049-296-5416, tatsujit[at]mail.dendai.ac.jp

3. 因果トレース学習

上述したように，強化学習において価値の配分をどう行うかは非常に重要な課題である．我々はその価値配分に，事象 (状態行動対) の変化に対する介入的な観測から推定される因果関係の強さを用いた手法を考案し，因果トレース学習と名付けた．応用前のアルゴリズムのベースには適格度トレースによる価値の伝播を行う TD(λ) 学習の一種である Sarsa(λ) を使用した．因果トレース学習アルゴリズムでは，任意の状態行動対の価値を意味する Q 値の更新の際に，Q 値に対する修正量である TD 差分の反映度合いに因果関係の強さ (因果トレース値) を用いて更新する．因果トレース値の計算にはある状態行動対の発生からの経過時間に相当する e 値 (適格度トレース値と同様)，状態行動対 (s_i, a_j) から (s_k, a_l) に対する適格度トレースの累積量である χ 値，任意の χ 値の更新の発生回数を意味する f 値を使用する．e 値，χ 値，f 値の更新はそれぞれ式 1，式 2，式 3 によって毎ステップ全ての値に対して行う．ここで用いられている β は忘却率という過去の情報の重みを減衰させるためのパラメータである．忘却率が $\beta = 1$ だと過去の情報の重みを完全に残して累積し，0 に近いほど，減衰していく．因果トレース学習ではこれらの数値を基に，事象の共起頻度から因果関係の強さを帰納的に推定する因果強度モデルの中から応用可能な既存のモデルの値を計算して，TD 差分の反映度に用いる事で，因果関係の強さを考慮した価値配分を行う．

表 1: 2 状態行動対間の e 値の累積表

	(s_1, a_1)	(s_1, a_2)	$\cdots$	(s_n, a_m)
(s_1, a_1)	$\chi^1_1(s_1, a_1)$	$\chi^1_1(s_1, a_2)$	$\cdots$	$\chi^1_1(s_n, a_m)$
(s_1, a_2)	$\chi^1_2(s_1, a_1)$	$\chi^1_2(s_1, a_2)$	$\cdots$	$\chi^1_2(s_n, a_m)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
(s_n, a_m)	$\chi^n_m(s_1, a_1)$	$\chi^n_m(s_1, a_2)$	$\cdots$	$\chi^n_m(s_n, a_m)$

$$e = \begin{cases} 1 & (s_i = s_t) \wedge (a_j = a_t) \\ 0 & (s_i = s_t) \wedge (a_j \neq a_t) \\ \lambda\gamma\beta e(s_t, a_t) & otherwise \end{cases} \quad (1)$$

$$\chi^i_j(s_{t+1}, a_{t+1}) = \begin{cases} \beta\chi^t_t(s_{t+1}, a_{t+1}) + e(s_i, a_j) & (s_i = s_t) \wedge (a_j = a_t) \\ \beta\chi^t_t(s_{t+1}, a_{t+1}) & otherwise \end{cases} \quad (2)$$

$$f(s_{t+1}, a_{t+1}) = \begin{cases} \beta f(s_{t+1}, a_{t+1}) + 1 & (s_i = s_t) \wedge (a_j = a_t) \\ \beta f(s_{t+1}, a_{t+1}) & otherwise \end{cases} \quad (3)$$

3.1 因果強度モデル

事象の共起頻度から因果関係の強さを帰納的に推定する因果強度モデルは複数存在する．しかしながら，それは飽くまでも共起頻度を基に計算される量的な指標であり，共起に時間的なズレを含む χ 値に対して定義されているものではない．また，ほとんどの因果強度モデルは原因結果の間に意図的な操作を伴わない非介入な観測を前提にしているため，応用できるモデルが極めて限られる．本研究では人間の因果関係の強さと高い相関を持つモデルである，DH[服部 01]，HS[甲野 10]，pARIs[高橋 14] を応用する．これらのモデルは人間の認知特性の一種である対称性バイアス [服部 08] をもったモデルであるとされる．対称性バイアスとは，$p \rightarrow q$ が真であるなら $q \rightarrow p$ も真であると思い込んでしまう認知傾向を意味する．そのような認知傾向を持つだけでなく，因果強度の値が $P(q|p)$，$P(p|q)$ という双方向の条件付き確率から計算されるという簡便さから，応用し易いと考えて因果トレース値として組み込んだ．

$$CP(s_i, a_j, s_k, a_l) = \frac{\chi^i_j(s_k, a_l)}{f(s_k, a_l)} \quad (4)$$

$$MP(s_k, a_l) = \sum_{(s_x \in S)} \sum_{(a_y \in A_x)} \chi^x_y(s_k, a_l) \quad (5)$$

$$RP(s_i, a_j, s_k, a_l) = \frac{\chi^i_j(s_k, a_l)}{MP(s_k, a_l)} \quad (6)$$

因果トレースでは，原因事象 p から結果事象 q が起こった割合である $P(q|p)$ を式 4 で定義される CP 値を用いる．CP 値は，e 値の累積量である χ 値を，χ 値の更新回数である f 値で除算するため，訪問回数に対する e 値の平均を意味する．また，結果事象 q が起こった際に原因事象 p が起こっていた割合である $P(p|q)$ を式 6 で定義される RP 値を用いる．RP 値は χ 値を用いて，結果として表れる状態行動対 (s_k, s_l) の以前に任意の状態行動対 (s_i, s_j) が他の状態行動対に比べてどの程度発生していたかを表している．そして状態行動対 (s_i, a_j) と (s_k, a_l) の間の共起頻度には $\chi^i_j(s_k, a_l)$ を用いる．すると DH，HS，pARIs の評価値はそれぞれ I_{DH}(式 7)，I_{HS}(式 8)，I_{pARIs}(式 9) となる．

$$I_{DH}(s_i, a_j, s_k, a_l) = \sqrt{CP(s_i, a_j, s_k, a_l) RP(s_i, a_j, s_k, a_l)} \quad (7)$$

$$I_{HS}(s_i, a_j, s_k, a_l) = \frac{2}{\frac{1}{CP(s_i, a_j, s_k, a_l)} + \frac{1}{RP(s_i, a_j, s_k, a_l)}} \quad (8)$$

$$I_{pARIs}(s_i, a_j, s_k, a_l) = \frac{\chi^i_j(s_k, a_l)}{f(s_k, a_l) + MP(s_k, a_l) - \chi^i_j(s_k, a_l)} \quad (9)$$

3.2 因果トレースにおける Q 値の学習

因果トレースアルゴリズムにおける Q 値の更新には，ベースとなる Sarsa(λ) と同じく現在の状態行動対 (s_t, a_t) が持つ価値 $Q(s_t, a_t)$ と，次に表れる状態行動対 (s_{t+1}, a_{t+1}) が持つ価値 $Q(s_{t+1}, a_{t+1})$ との TD 差分 δ (式 10) を用いる．TD 差分 δ を全ての状態行動対に対して計算された因果強度モデルの値 I_{model} に応じた更新を行う．これは本アルゴリズムが TD(λ) 学習における適格度トレース (e 値) を因果強度 I_{model} に置き換えて行っている事を意味している．I_{model} の値は因果強度の計算に用いるモデル (DH，HS，pARis) によって異なる．

$$\delta = r_t - \gamma Q(s_{t+1}, a_{t+1}) - Q(s_i, a_j) \quad (10)$$

$$Q(s_i, a_j) = Q(s_i, a_j) + \alpha I_{model}(s_i, a_j, s_{t+1}, a_{t+1})\delta \quad (11)$$

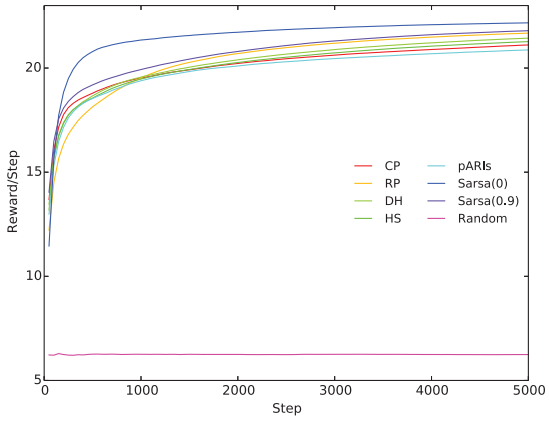

図 1: 獲得した平均報酬の推移

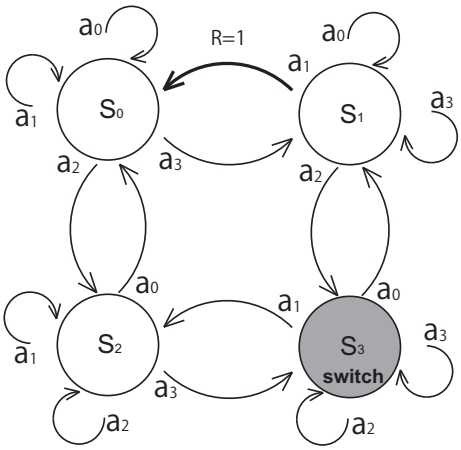

図 2: 状態の推移と報酬のスイッチ

4. シミュレーション 1 -基本格子タスク-

まず因果トレース学習アルゴリズムの基本的動作を検証するため，通常の単純なマルコフ決定過程課題に対するシミュレーションを行った．現時点の原始的な因果トレース学習アルゴリズムでは χ 値の行列の要素数が全状態行動対の平方数となる事から，まず比較的小さい状態数と行動数を持つ課題での検証を行った．環境は縦横 2 マスの格子空間を想定しており，状態はマス数に対応した 4 状態を取る ($S = \{s_0, s_1, s_2, s_3\}$)．エージェントは各状態毎に上下左右への移動という 4 つの行動 ($A_i = \{a_0, a_1, a_2, a_3\}$) をできる．格子空間の外に出てしまう時は壁に阻まれるように状態が遷移できない (例えば左上状態 s_0 で上移動行動 a_0 を選択すると，格子空間上で左上状態 s_0 に状態が存在しないため，遷移先の状態は現在状態と同じ左上状態 s_0 になる)．本シミュレーションでは，右上状態 s_1 において左移動行動 a_1 を選択し，左上状態 s_0 に遷移した時に報酬 $r = 1$ がエージェントに与えられる．このシミュレーションの目的は因果トレース学習が従来のモデルフリーな学習と同等な水準で学習可能かを確かめる事である．シミュレーションでは一回の行動選択と状態遷移を 1 step として 5000 step 行い，その間に得られた報酬の累積値 (収益) を成績として 1000 回のシミュレーションの平均値を比較した．比較には因果トレース学習アルゴリズムの元になった Sarsa(λ) を用い，$\lambda = 0.0$，$\lambda = 0.9$ の場合をそれぞれ示す．学習エージェントのパラメータは学習率 $\alpha = 0.2$，割引率 $\gamma = 0.9$，行動選択に用いる方策 ϵ-greedy のランダム選択割合 $\epsilon = 0.1$ という，経験的によく用いられている値にした．. 因果トレースのパラメータは $\lambda = 0.9$，忘却率 $\beta = 0.9$ とした．また，そもそも学習が促進しているかの水準として完全にランダムな行動を選択するエージェント (Random) の成績も示す．

4.1 結果及び考察

シミュレーションの結果を図 1 に示す．獲得報酬の累積であると学習の進度が解り難いため，縦軸を報酬累積に対する step 数あたりの平均報酬とした．シンプルな課題であるためか，比較アルゴリズムの中で最も単純な学習アルゴリズム Sarsa(0) が最も速く，最も良い成績を有している．また，Sarsa(0) と完全ランダムな行動選択をするエージェント (Random) 以外の学習アルゴリズムには大きな差が見られず，また DH，HS，pARIs，それぞれの因果強度モデルを用いた因果トレース学習アルゴリズムが，いずれも時間と共に報酬をより多く得られる行動系列を学習出来ている事が確認できた．これにより，因果トレース学習アルゴリズムが単純な強化学習課題を Sarsa と同等の水準で学習可能な事がわかった．

5. シミュレーション 2 -スイッチ格子タスク-

次にマルコフ性が弱い課題における因果トレース学習アルゴリズムの成績を示すためシミュレーション 1 より複雑な課題でのシミュレーションを行う．マルコフ性が弱い学習課題として，原因状態を訪れた後に任意の時間内に結果状態を訪れたときのみ報酬が得られる課題を想定した．課題環境はシミュレーション 1 と同じ状態と行動選択肢を持つ 2×2 格子空間を用いる．唯一異なるのは，報酬を得られるのが，右下状態 s_3 を訪れた後 3 step 以内に，右上状態 s_1 において左移動行動 a_1 を選択し，左上状態 s_0 に遷移した時に報酬 r がエージェントに与えられる点のみである (図 2)．即ち，報酬を得る原因として，s_3 への訪問が重要であり，その後 (s_1, a_1) を訪れるまでの系列には固定された経路としての価値が無い．以上の点で本シミュレーション課題はマルコフ性が弱い学習課題であると言える．比較に用いるアルゴリズムやシミュレーション回数についてもシミュレーション 1 と同様であり，step 数のみシミュレーション 1 の 10 倍である 50,000 step 行った．

5.1 結果及び考察

シミュレーションの結果を図 3 に示す．シミュレーション 1 と同様に縦軸が獲得報酬の累積であると学習の進度が解り難いため，step 数あたりの平均報酬とした．全体の傾向としては，多くのエージェントがある程度学習した後，単位 step 当たりの獲得報酬が減少している点が挙げられる．特に Sarsa($\lambda = 0.9$) においてその傾向は顕著に表れている．対して pARIs は学習進度の発展は他の学習アルゴリズムや因果強度よりもやや遅いものの，最終的には最も高い報酬を得られている．また，Sarsa(0) がランダム選択エージェントより低い成績であるという点で，この学習課題の環境のマルコフ性が弱い事も示されている．このような環境では，一つの状態遷移に対するマルコフ性のみに着目した仕組みでは学習できず，適格度トレース等による状態系列全体に対する価値配分が不可欠になる．しかしながら，Sarsa(0.9) は獲得平均報酬が一度上昇した後に減少し

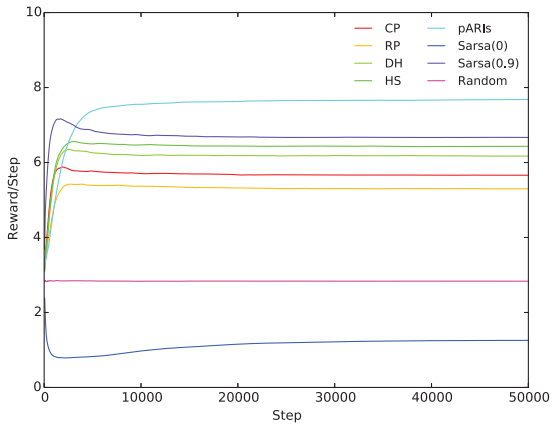

図 3: 獲得した平均報酬の推移

て，その後再び学習が進まないという傾向が見られた．これは適格度トレースもマルコフ性が弱過ぎるタスクには適応できない事を示しており，適格度トレースに優っている因果強度モデルは pARIs のみではあるが，それによって因果トレースの有用性が断片的ながら示されたと言える．

6. 結論

本研究では，強化学習に因果関係の推定がどのような貢献をもたらすかを知るため，状態系列に対する価値配分に人間の推定する因果関係の強さを応用する事を試みた．本研究で使用した因果強度モデルの計算には，共起頻度ではなく要素間の時間的な影響度を意味する適格度トレース (e 値) の累積である χ 値等を使用した．これは事象の発生に対する時間的隔たりを扱うために行った改変であり，その変更が因果強度モデルにどのような性質の変化をもたらすか，本研究では考慮していない．また，因果強度の計算のために全ての状態行動対の数の平方数だけの記憶が必要となるため，状態行動対の数が膨大である事を考慮できていない．しかしながら，シミュレーションの結果からマルコフ性の弱い課題環境下における因果トレースによる価値配分の有効性を断片的ながら表す事が出来た．また前述の問題点についても，まだ認知的な解釈と数学的な解析が進んでいない現時点においては，それらが因果トレース学習の本質的な問題であるとは言えない．因果トレース学習はモデルフリーな学習をベースにしている事から，本研究の結果はモデルフリーでありながら因果関係の強さを導入する事のみで，課題環境にマルコフ性を想定せずに強化学習が可能である事を断片的ながら示す事が出来たと言える．

参考文献

[服部 01] 服部 雅史: 因果帰納の二要因ヒューリスティクス・モデル, 認知科学, 8(4), 444–453 (2001).

[Hattori 07] Hattori, M. and Oaksford, M.: Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis, *Cognitive Science*, 31(5), 765–814 (2007).

[服部 08] 服部 雅史: 推論と判断の等確率性仮説: 志向の対称性とその適応的意味, 認知科学, 15(3), 408-427 (2008).

[服部 08] 服部 雅史, 山崎由美子: 対称性と双方向性の認知科学：特集「対称性」の編集にあたって, Cognitive Studies, 15(3), 315–321.(2008).

[甲野 10] 甲野 佑, 高橋 達二: 因果帰納の調和対象ヒューリスティクス, 日本認知科学会第 27 回大会 (JCSS2010) 発表論文集, 43–46(2010).

[牧野 14] 牧野 貴樹: 実用化する強化学習研究, 生産研究, 66, 3, 305–308 (2014).

[中野 14] 中野 太智, 前田 新一, 石井 信: 状態非依存の方策を用いた新しい強化学習の提案, システム制御情報学会論文誌, 27, 8, pp.327–332 (2014).

[高橋 14] 高橋 達二, 大用 庫智 : 対称性推論モデルとしての「双条件付き確率」と少数サンプルからの因果帰納推論, 日本認知科学会 31 回大会 (2014).

[斎藤 12] 斎藤 淳哉: 実環境における不確実性や遅延を考慮した学習に関する研究, 東北大学 大学院情報科学研究科修士学位論文 (2012).

[澁谷 14] 澁谷 長史, 安信 誠二: 報酬が周期的に変化する環境のための強化学習, 電気学会論文誌 C(電子・情報。システム部門誌), 134, 9, 1325-1332 (2014).

[Sutton 00] Sutton, R. S., Barto, A. G., 強化学習, 森北出版, (三上, 皆川 訳) (2000).