1L4-OS-24b-5in

ヒトの適応的認知特性によるモンテカルロ木探索の効率化

Efficiency of the Monte-Carlo Tree Search by adaptive characteristic of human cognition

西村友伸[*1] Tomonobu Nishimura　大用庫智[*1] Kuratomo Oyo　高橋達二[*2] Tatsuji Takahashi

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

事象を相対評価する因果推論モデルとして LSVR が近年研究されている. 本研究では囲碁、リバーシ上の MCTS の評価値として LSVR を適用した. 既存の MCTS の評価値としては損失の限界の保証された UCB1 がよく用いられるが、UCB1 と LSVR の双方を場合により使い分ける事で、単体で用いるより良い結果を得られた。また LSVR の汎用性を示すために AMAF 等のボードゲームの手法との組み合わせについても論じる.

1. 序論

近年, 囲碁や将棋といったボードゲーム AI の成長は著しく, 人間のプロの強さに迫る勢いとなっている. 囲碁 AI の成長についてはモンテカルロ木探索というサンプリングベースの手法が生まれた事が大きい [Kocsis 2006] . モンテカルロ木探索には, 複数のスロットマシンを一定回数プレイしその報酬を最大化する n 本腕バンディット問題が内包されており, この問題への価値関数として, 囲碁 AI では UCB1 と呼ばれるモデルが使われ良い成果を出している [Auer 2002; Gelly 2006] .

本研究では, n 本腕バンディット問題において少ないサンプリング数では UCB1 より良い結果を出す事が分かっている LSVR をモンテカルロ木探索の価値関数として適用した.

LSVR を価値関数とした AI は, リバーシ上でも UCB1 を用いた AI より少ないサンプリング数では強い事がわかっている [西村 2012]. 一方サンプリング数が十分多い時には UCB1 を用いた AI が強い事も分かっている. よって今回は場合により 2 つの価値関数を使い分けるモデルを作成し, その効果を調べた.

また, リバーシで確認された少ないサンプリング数では LSVR が UCB1 に勝ち越すという結果が, 他のボードゲームについても再現出来るかはまだ確認されていない. その為, 本研究では囲碁によりその検証を行ない LSVR の汎用性を確かめる事も行なった. 更に実際の囲碁 AI では, モンテカルロ木探索以外にも様々なヒューリスティクスにより AI の強化が行われる. 今回は数ある手法の中でも代表的な RAVE と LSVR を組み合わせ, その相性の確認も行なった.

2. n 本腕バンディット問題

n 本腕バンディット問題とは, 当たり確率が不明の n 台のスロットマシンを一定回数プレイし報酬の最大化を目指す意思決定の課題である. この問題では最も当たりの確率が高いマシンを探す探索行動と, 現在当たりの確率がもっとも高いであろうマシンをプレイする収穫行動が考えられる. しかし探索行動を起こせば利益の最大化を目指す収穫をする事は出来ず, 収穫行動を起こせばより良い台の探索は行えない.

この 2 つの行動が両立しないトレードオフは探索と知識利用のジレンマとして知られている. n 本腕バンディット問題で良い成績を目指す為には, このジレンマを上手く弱める事が必要となる.

n 本腕バンディット問題は様々な問題に当てはめる事ができ, 本研究で扱うリバーシや囲碁では各盤面での合法手をスロットマシンと考える事で n 本腕バンディットに当てはめる事が出来る.

2.1 UCB1

UCB1 は Auer らにより考案された n 本腕バンディット問題の台の選択アルゴリズムである. このアルゴリズムではまず始めに全ての台を一度プレイする. その後は式(1) で計算される値が最も高い台をプレイする[Auer 2002] .

X_i は台 A_i の報酬の期待値, n_i は台 A_i のプレイ回数, n は全ての台のプレイ回数の和を表す. c は第二項の重みを表し, c が大きいとプレイ回数が少ない台を選ぶ探索の傾向が強まり, 逆に小さいと期待値の反映が大きくなり収穫の傾向が強まる.

$$UCB1(A_i) = X_i + c\sqrt{\frac{2\log n}{n_i}} \quad (1)$$

UCB1 は試行回数が十分に多い時, 後悔の上限が保証されており, 最も良い台を選ぶ事が出来る. しかしスロットマシンの台数が多い時, 初期状態の入手や, 最良の台の発見に時間が掛かる事も知られる.

2.2 LSVR

人間認知の適応特性を利用した因果推論のモデルとして LSVR が近年研究されている[篠原 2007; Kohno 2012] . LSVR はヒトの認知バイアスである, $p \rightarrow q$ から$q \rightarrow p$ を導く対称性バイアスと$p \rightarrow q$ から$\bar{p} \rightarrow \bar{q}$ を導く相互排他性バイアスを緩く含み, 選択肢を相対評価するモデルである.

LSVR では表 1 の$n \times 2$ 分割表データを利用し, 式(6) で計算される最大の値の選択肢を採用する. 表 1 中の a_1 は選択肢 A_1 を選び, 結果が W となった回数であり, b_1 はその結果が$\bar{W}$となった回数である. n 本腕バンディット問題であれば, W は当たり, $\bar{W}$は外れという結果を表す.

表 1: LSVR で用いる分割表

	結果 W	結果$\bar{W}$
選択肢 A_1	a_1	b_1
⋮	⋮	⋮
選択肢 A_n	a_n	b_n

西村 友伸 (Nishimura tomonobu)
E-mail: 12rmj24@ms.dendai.ac.jp

$$S_p = \frac{b_{\max} b_{\min}}{b_{\max} + b_{\min}} \quad (2) \qquad S_n = \frac{a_{\max} a_{\min}}{a_{\max} + a_{\min}} \quad (3)$$

$$\rho_R = \frac{1}{R_t} - 1 \quad (4) \qquad R_{t+1} = \alpha R_t + (1-\alpha) r_t \quad (5)$$

$$LS(W \mid A_i) = \frac{a_i + S_p}{a_i + b_i + \rho_R (S_p + S_n)} \quad (6)$$

ここで a_{max} , b_{max} は選択回数が最も多い選択肢の a_i , b_i であり, a_{min} , b_{min} は選択回数が少ない選択肢の a_i , b_i である. R_t は時刻 t での LSVR の価値の基準であり, LSVR はこの価値基準を満たす選択肢を探索し, 収穫を行う. 初期状態では$R_t = 0.5$ に設定されている. 価値基準は式(5) の漸化式で更新され α は学習率, r はリファレンス報酬を表し, n 本腕バンディット問題では当たりなら 1 , 外れならば 0 となる. ρ_r は判断の基準を式(6) に反映させる為のパラメータである.

LSVR は n 本腕バンディット問題で少ない試行回数では UCB1 よりも良い結果を出している. またスロットマシンの台数が増えた場合であっても結果が変わりづらくロバストなモデルと言える.

3. モンテカルロ木探索

問題に対してランダムなサンプリングを行い, 大数の法則からその近似解を得る手法はモンテカルロ法として知られ, さらにそのサンプリングを工夫し方針を持たせたものはモンテカルロ計画法と呼ばれる. 近年, これらの手法を更に木構造へ拡張したモンテカルロ木探索(Monte Carlo Tree Search, MCTS)が囲碁 AI に用いられ非常に良い成果を出している.

ボードゲームに置けるモンテカルロ木探索は, 現在の盤面を根ノード, 合法手をエッジ, 子ノードは各合法手を打った後の盤面とするゲーム木の探索をする事で実現する.

ボードゲームでの具体的な流れとしては, 根ノードから何らかの価値関数により, 良いと考えられる子ノードへ遷移を繰り返し葉ノードへ移動する. そこでプレイアウトを行い, 勝ちならば 1 , 負けならば 0 の報酬を得てそれを全ての親ノードへ伝播させる. プレイアウトとはある局面から終局までランダムにプレイを進める事をいう. 葉ノードへの訪問回数がある閾値を超えた場合, そのノードから更に木を展開する. ゲーム木を深く探索する事は, ゲームの 2 手 3 手と先まで予想して打つ事に相等する.

4. RAVE

RAVE(Rapid Action Value Estmate) とは囲碁のプレイアウト内で使われる手法であり, AMAF(All Moves As First) もほぼ同じ意味で使われる[Gelly 2007] . これはプレイアウト内で打たれた手については, 直接訪問していないノードであってもデータを更新する手法である. 例を挙げると, 先手側が A と言う手をプレイアウト内で打って勝った場合, ゲーム木内で先手が A という手を打った全てのノードの評価値を更新するといった行為である.

しかし, この方法は通常の更新に比べデータの信用性が落ちる為, データとしては別に保持する必要がある. またプレイアウトの後半程データとしての信用性が落ちる為, 更新量を本来の値より減衰させる. 例えばデータを 1 加算するような場合であれば, 1 − w × d の様に減衰させる. w は減衰の重み, d はプレイアウトの開始から何手目かを表す.

実際に UCB1 と RAVE を用いたモンテカルロ木探索での評価値は次の用に通常の UCB1 の値と, RAVE により保持されたデータを用いて算出された $UCB1_{RAVE}$ の値に重みを付けた和である $UCB1_{UCB1+RAVE}$ となる beta は UCB1 と $UCB1_{RAVE}$ のどちらを重視するか, その割合を決めるパラメータである.

$$UCB1_{RAVE}(A_i) = X_{i\_RAVE} + c\sqrt{\frac{2\log n_{RAVE}}{n_{i\_RAVE}}} \quad (7)$$

$$UCB1_{UCB1+RAVE}(A_i) = (1-beta) \times UCB1(A_i) + beta \times RAEV(A_i) \quad (8)$$

$$beta = \sqrt{\frac{K}{K+M+n}} \quad (9)$$

K, M はシミュレーションで適する定数を見つけ設定する. K, M の値により, RAVE の値の反映の強さを変える事が出来る.

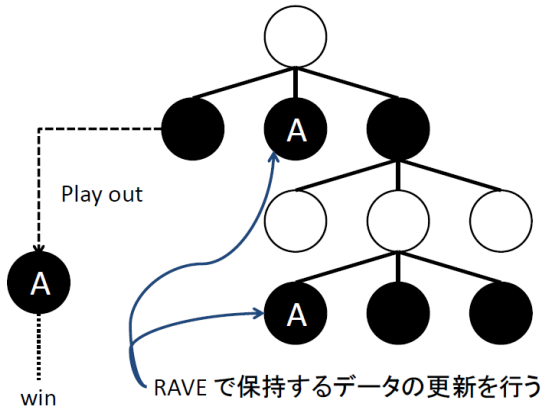

図 1: RAVE で行うデータ更新

5. 提案手法: ハイブリッドモデル

本研究で扱う価値関数, UCB1 は試行回数が十分に多い場合最良の手に辿り着く事が証明されているが, 少ない試行回数ではあまり良い結果は得られない. 一方 LSVR はある程度の試行回数までは UCB1 より良い結果を得られるが, 試行回数が非常に多い時には UCB1 程の結果は得られない.

そこで本研究ではモンテカルロ木探索で行われる各 n 本腕バンディット問題に毎に, サンプリング数が閾値以下の時には LSVR を用い, 閾値より多い時には UCB1 を用いるハイブリッドなモデルを作成し, リバーシにおける通常の UCB1 と比べ勝率がどう変化するか調査した.

6. シミュレーション 1:ハイブリッドモデル

通常の UCB1 と違い, 序盤の探索を LSVR に任せるハイブリッドモデルの性能を調査する為, UCB1, ハイブリッドモデルを実装した AI をリバーシで対局させた.

6.1 シミュレーション 1 設定

UCB1 に対しハイブリッドモデルをリバーシで対局させる. また対照実験として LSVR を実装した AI を UCB1 と対局させる事も行なった. 対局はで先手後手による有利不利を取り除く為, 先手後手入れ替えて計 600 局対戦させる. サンプリング数は 10,50,100,150,300,500,1000,5000,10000 回とし, モデル切り替

えの閾値は 50 回とした. 各モデルのパラメータは, LSVR の学習率 α を 0.8, UCB1 のパラメータ c を 0.1 , 木の成長の閾値は 1 回に設定した.

6.2 シミュレーション 1: 結果

シミュレーション 1 の結果を図 2 に示す. 縦軸は勝率, 横軸はサンプリング回数を表した片対数グラフとなっている.

6.3 シミュレーション 1: 考察

シミュレーションの結果, UCB1 と LSVR をサンプリングの回数により使い分けるハイブリッドモデルは, 通常の LSVR よりも良い結果を残せる事がわかった. また UCB1 に対してサンプリング数が 5000, 10000 回と増えた場合であっても勝率 50%を上回り, UCB1 単体よりも良い結果を残す事が出来た. これは少ないサンプリング数でも良い選択が出来る LSVR の性質により初期のサンプリングが効率化された結果と考えられる.

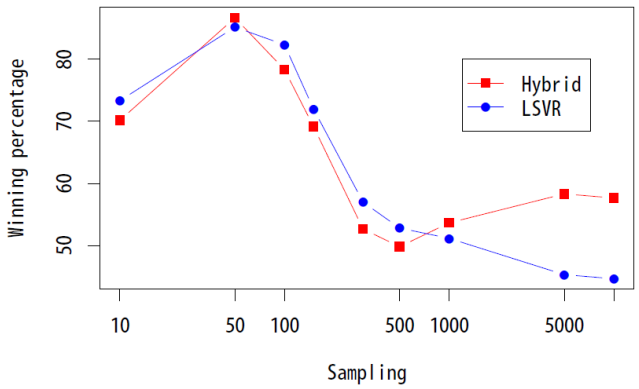

図 2: UCB1 に対するハイブリッドモデル, LSVR の勝率

7. シミュレーション 2: 囲碁に置ける LSVR

囲碁上での LSVR の効果を確認するため, LSVR と UCB1 を実装した AI を 9 路盤の囲碁で対局させた.

7.1 シミュレーション 2: 設定

LSVR と UCB1 を実装した AI を囲碁で先手後手入れ替えて計 200 局対戦させる. サンプリング数は 100,500,1000,5000 回とし, 各モデルのパラメータは, LSVR の学習率 α を 0.8, UCB1 のパラメータ c を 0.31 , 木の成長の閾値は 1 回に設定した. UCB1 のパラメータ設定には囲碁 AI で有名な彩の設定を参考にした [美添 2012] .

7.2 シミュレーション 2: 結果

シミュレーション 2 の結果を表 3 に示す. 表 1 列目はサンプリング数, 2 列目は LSVR の UCB1 に対する勝率を表す.

表 3: 囲碁における LSVR の UCB1 に対する勝率

サンプリング数	勝率
100	46%
500	47.5%
1000	46%
5000	39%

7.3 シミュレーション 2: 考察

シミュレーションの結果, 囲碁で LSVR と UCB1 を対戦させた場合, 少ないサンプリング数では若干 LSVR は負け越し, サンプリング数が増えると勝率が徐々に下がる様な傾向が確認できた. これは UCB1 の, サンプリング数が十分に多い時に最良の手に辿り着くという性質に一致する.

囲碁において LSVR は UCB1 に若干負け越してしまったが, LSVR 側が強化学習での一般的な学習率を用いたのに対し, UCB1 側はしっかりとしたパラメータチューニングを行なっている. この事から, LSVR のチューニングの簡単さと, 囲碁の様な複雑なゲームへの対応力が確認できる.

8. シミュレーション 3: RAVE

LSVR と囲碁における有名手法である RAVE との相性を調べるため, 囲碁上で UCB1 と, UCB1 に RAVE 値を UCB 式により算出し加えた UCB_{RAVE}, UCB1 に RAVE 値を LSVR 式により算出し加えた $UCB1_{LSVR}$ 作成した. そして二つのモデルそれぞれ UCB1 と対局させた.

8.1 シミュレーション 3: 設定

UCB1 と UCB_{RAVE}, $UCB1_{LSVR}$ を実装した AI をリバーシで先手後手入れ替えて計 200 局対戦させる. サンプリング数は 1000 回とし, 各モデルのパラメータは, LSVR の学習率 α を 0.8, UCB1 のパラメータ c を 0.31 , 木の成長の閾値は 1 回に設定した. RAVE のパラメータとしては $w = 0.0015$, $K = 100$, $M=3$, とした, UCB1 のパラメータ設定には囲碁 AI で有名な彩の設定を参考にした .

8.2 シミュレーション 3: 結果

シミュレーション 3 の結果を表 4 に示す. 表 1 列目は AI 中で扱われる価値関数を表し, 2 列目は UCB1 を実装した AI に対する勝率を表す.

表 4:$UCB1_{RAVE}$,$UCB1_{LSVR}$ の UCB1 に対する勝率

モデル	勝率
$UCB1_{RAVE}$	53.5%
$UCB1_{LSVR}$	50.5%

8.3 シミュレーション 3:考察

UCB1 を実装した AI に, UCB1 , LSVR により算出された RAVE 値を追加したモデルを, 通常の UCB1 と対局させた. その結果として大きな差は見られなかったものの, 僅かに $UCB1_{RAVE}$ が上回った. これはプレイアウトから大量のサンプリングを得る RAVE と, サンプリング数の増大と共に良い結果が得られる UCB1 との相性が良い事によると言える.

また $UCB1_{LSVR}$ については, 結果が悪くなることは無かったが, 性能の向上は見られなかった.

9. 終わりに

本研究では, リバーシ上での UCB1, LSVR の両方を使い分けるハイブリッドモデルの検証と囲碁上のモンテカルロ木探索時の LSVR の効果の確認, 囲碁 AI での有名手法である RAVE と LSVR の相性の確認を行いモンテカルロ木探索の効率化を図った.

リバーシ上でのハイブリッドモデルのシミュレーションでは, UCB1 に対して勝ち越し, UCB1 単体で扱うより, UCB1 と LSVR の 2 つのモデルを場合により使い分けた方が良い結果を出せる事を示せた.

囲碁 AI 上での価値関数としての LSVR は, 既存のモデルである UCB1 に勝ち越す事は出来なかった. しかしながら, 緻密

なパラメータチューニングを行った UCB1 に対して, 強化学習での一般的な学習率を設定した LSVR は 45% から 40% 程の勝率を収めた. この事から囲碁のような複雑なゲームであっても, モンテカルロ木探索と LSVR を組み合わせる事で容易にある程度の強さを持った AI を作成出来る事が示せた.

最後に, 囲碁 AI に置ける有名手法である RAVE と LSVR の相性の確認では, 目立った結果を確認する事は出来なかった. これは少ないサンプリング数で有効な LSVR と, プレイアウトから信頼性には欠けるが大量のサンプリングを得る RAVE との相性が良くないからだと考えられる.

今後, LSVR を用いて囲碁上でのモンテカルロ木探索を効率化する場合は, ハイブリッドモデルを応用しての方法が考えられる. その場合には通常の LSVR 使用時の少ないサンプリングでの強さ引き上げる事が課題となる. その為の手法としては, LSVR の判断の価値基準である R_t の学習方法の変更や, 価値基準に下限の設定による探索の促進等の方法が考えられ, 今後の更なる研究が必要となる.

参考文献

[Auer 2002] Peter Auer, Nicolo Casa-Bianchi, Paul Fischer: Finite-time analysis of the multiarmed bandit problem , *Machine Learning* 47 235-256, 2002.

[Gelly 2006] Sylvain Gelly, Yizao Wang, Remi Munos, Olover Teytaud : Modification of UCT with Patterns in Monte-Carlo Go, *INRIA* 6062, 2006.

[Gelly 2007] Sylvian Gelly, David Silver: Combining Online and Offline Knowledge in UCT, ICML'07: Proceedings of the 24th international conference on Machine leraning New York, pp.273-280, 2007.

[Kocsis 2006] Levente Kocsis, Csaba Szepesvari :
Bandit based Monte-Carlo Planning , ECML'06 In: ECML-06, *LNCS*, 4212, pp. 282-293 , 2002.

[Kohno 2012] Yu Kohno, Tatsuji Takahashi: Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off, *The 6th International Conference on Soft Computing and Intelligent Systems, The 13th International Symposium on Advanced Intelligent Systems*. 2012.

[大用 2010] 大用庫智: ヒト認知バイアスのモンテカルロ法への応用, 2009 年度情報科学科卒業文集, 2010.

[篠原 2007] 篠原 修二, 田口 亮, 桂田 浩一, 新田 恒雄: 因果性に基づく信念形成モデルと N 本腕バンディット問題への適用, 人工知能学会論文誌, 22 巻 1 号, pp.58-68 ,2007.

[西村 2012] 西村 友伸, 大用 庫智, 高橋達二: 可変参照型緩対称性推論のモンテカルロ木探索での効果, ゲームプログラミングワークショップ 2012 論文集, 2012(6), pp.191-196, 2012.

[美添 2012] 美添 一樹, 山下 宏: コンピュータ囲碁 – モンテカルロ法の理論と実践, 編者 松原 仁,共立出版, 2012.