Vol.2010-GI-24 No.4
2010/6/25

モンテカルロ＋UCTにおける探索木のだまし構造

池田　心[†1]　　橋本隼一[†1]　　土井佑紀[†1]

マルチアームバンディット問題 (MAB) を対象に発展した　B 戦略は，探索と収穫のジレンマに対する一つの有力な回答である．近年，　B 戦略をゲームの木探索に応用した　が盛んに研究され，モンテカルロ法と組み合わせた囲碁プログラムが登場している．しかし，報酬を得られる確率が静的な MAB と異なり，　in　ax 原理が支配する二人ゲームでは　B 値に基づく探索が非効率的である可能性もある．本稿では，囲碁において　が好ましくない挙動をしうることを指摘し，その本質を抜き出してベンチマーク化する．

Deceptive Structure of Search Tree for UCT with Monte-Carlo

Ikeda Kokolo,[†1] Hashimoto Junichi[†1]
and Doi Yuuki[†1]

B strate　which has　een develo　ed for solvin　ulti-ar　ed　andit　ro -le　(MAB) is a　ajor solution for the ex　loitation-ex　loration dile　a.　is the a　lication of　B for　a　e tree search, and is intensivel　studied es e-ciall　for　o-　a　e　ro ra　with　onte carlo　ethod. However, in contrast to the MAB, the ratio for rewards is not static　ut d　na　ic in two-　la ers　a　e such as　o. In this　a　er, we　resent a failure case of　for　o,　eneralize it and　uild a　ench　ark tree search　ro　le　with dece　tive structure.

1. はじめに

マルコフ決定過程 (MDP) などの繰り返し環境にエージェントが置かれ，不完全情報下でできるだけ多くの報酬を得ようとする場合，現在分かっている範囲で高い報酬を得る “収穫 (exploitation)” とより高い報酬が得られる可能性を模索する “探索 (exploration)” はしばしば両立せず，探索と収穫のジレンマと呼ばれる．このジレンマを簡潔に表したモデルにマルチアームドバンディット問題 (MAB) がある．これは 1 状態（局面）複数行動の MDP であり，各行動 a_i 後に一定の確率 p_i で報酬 1 を得られるというものである．

UCB(Upper Confidence Bound)[1] とは，行動（指手）選択回数 n，行動 a_i の選択回数 n_i と報酬（勝率）の期待値 e_i に対して，$e_i + C\sqrt{\frac{\ln(n)}{n_i}}$ で定義される値であり（C は設計パラメータ），n に対して n_i が相対的に小さいときに報酬の期待値が “上がりうる幅” を組み込んだものである．UCB 値が最大となる行動を選択していくことで，期待値の高い行動の exploitation と発展の見込みのある行動の exploration を制御することを可能にしており，前述の MAB などで良い性能を示している．

UCT(Upper Confidence bound for Tree search)[4] は，ゲームなどの木探索 (Tree Search) に UCB の考え方を導入したものである．例えば囲碁や将棋などの多くのゲームでは，状態分岐数（取り得る着手の数）や木の深さ（終局までの手数）が非常に大きく，厳密解法を用いることは事実上不可能である．そのため，良さそうな着手を辿りながら深くまで探索し (exploitation)，かつ可能性のあるそれ以外の分岐にも探りを入れる (exploration) ことが必要になる．UCT では，探索したノードそれぞれに訪問回数と報酬の平均値を保持し，子ノードの UCB 値が最も高くなるものを選択することでこのバランスを取る．

多くの場合，UCT は優れた探索性能を示す，すなわち，深く読むべきところに多くの探索資源を割きつつ，二番手三番手の着手にも現在の報酬平均に応じた資源を投じる．盤面の評価をランダムゲームによって求めるモンテカルロ法との組み合わせは，多くの囲碁プログラムで用いられているところである[3)5)]．しかしながら，UCB 戦略が開発された MAB とは異なり，minmax 原理に支配された二人ゲームでは各行動を取った場合の報酬の期待値は静的ではない．UCB 値における補正項 $C\sqrt{\frac{\ln(n)}{n_i}}$ は報酬の期待値が静的であることを前提としており，動的に変化する場合には不十分である可能性がある．

本稿では，囲碁で UCT が好ましくない挙動を示す場合を発見したことを報告したうえで，その本質を抜き出してベンチマーク化する．木探索を学問として研究していくためには，具体的なゲームを題材に勝率を云々すると同時に，容易に共有・調整・実験ができるベンチマーク問題の整備が必須であると考えている．

†1 北陸先端科学技術大学院大学
JAIST Hokuriku

2. 囲碁における問題点の発見

2.1 nomitan プロジェクト

我々のチームでは，2 年前に野口・松井両氏によって開発された囲碁プログラム nomitan をベースに囲碁プログラムの研究を進めている．nomitan は UCT に基づく探索と，ランダムゲームによるモンテカルロ評価を組み合わせたプログラムであり，ランダムゲームでは勾配法で棋譜学習した係数による着手の選択確率計算，UCT では UCB1-Tuned 値による探索ノード選択[1] を行っている．探索終了後は，現局面で最も訪問回数の高い着手（子ノード）を打つが，これは最も有望な着手を最も丹念に読んでいるはずだという信念に基づく[2]．

2.2 次の一手問題

囲碁プログラムの強さは，通常，プログラム同士の対戦によって計測する．しかし，より詳細にその挙動を知りたい場合には，局面を与えて次の一手を探索させることが有効であるし，また低コストである．

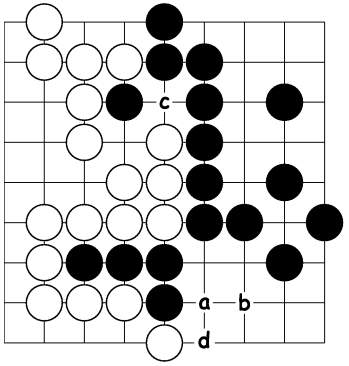

図 1 次の一手問題の例：コミ 3 目半．b(7,8) が正解

図 1 は次の一手問題の例である．この問題では a(6,8) が普通の手に見えるが，後手を引いて c(5,3) に回られてしまうので失敗であり，b(7,8) が c,d を見合いにした正解である．なお初手 d は a に切られて失敗である．この問題は人間にとっては 5 級程度の問題であり，決して難しい部類のものではない．

2.3 好ましくない探索結果

nomitan に 80 万回のプレイアウト（ランダムゲーム）を与えて探索をさせたところ，有力な手 a(6,8),b(7,8) について，その勝率の期待値と訪問回数は図 2 のように変化した．なお，nomitan の場合 80 万プレイアウトの探索には通常の PC でおおよそ 1 分を要する．

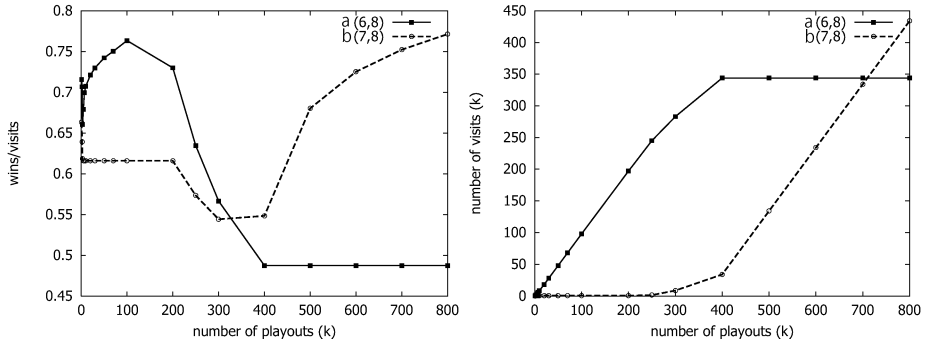

図 2 nomitan による探索，着手 a(6,8),b(7,8) の勝率の期待値（左）と訪問回数（右）

探索の序盤 20 万プレイアウトくらいまでは，a の勝率が b よりかなり高く（左図），殆どの探索資源が a に割かれている（右図）．中盤 20 万〜40 万プレイアウトで a の勝率は大きく下がり，b にも探索資源が割かれるようになる．終盤 40 万〜70 万プレイアウトでは逆に b の勝率が上昇し，a には探索資源が割かれない．

このような挙動は別段奇妙なものではない．有望そうな手を読んでいるうちに相手側の好手に気づくこと，逆に駄目そうな手を読み返して自分側の好手に気づくこと，これらは囲碁に限らず二人ゲームには頻繁に生じる現象である．この探索について問題を挙げるとすれば，(1) 序盤 20 万プレイアウトまで b は殆ど探索されないこと，(2) 勝率が逆転したあとも「訪問回数最大の手を打つ」というルールに従えば 50 万〜70 万プレイアウトのような明らかに b を選ぶべき場合でも a を選んでしまうこと，である．これらの問題点を整理すべく，次章ではより簡潔なベンチマーク問題を提案する．

3. だまし構造のあるベンチマーク探索木

確率的最適化の分野では，“一見悪そうな領域に最適解があり，一見良さそうな領域に探索を重点化すると失敗する” ことをだまし構造があると言う[6]．前章で紹介した例も，着手 a は正解ではないが一見良く見え，着手 b は正解なのに一見悪く見えることが問題であった．そこで，このような探索木も**だまし構造がある**と言うことにする．

また，クラス分類問題等の呼び方になぞらえて，一見良く見えることを false positive（偽陽性），一見悪く見えることを false negative（偽陰性）と呼ぶことにする．例えば，医療ではしばしば簡易検査を行い陽性のものを精密検査するが，このような場合，偽陽性は精密検査のコストがかかる，偽陰性は病気を見逃すリスクがあるということになる．探索においては，偽陽性は本来探索すべきでない箇所を探索するコストがかかる，偽陰性は正解を見逃すリスクがある，と読み替えることができるだろう．

3.1 ベンチマーク

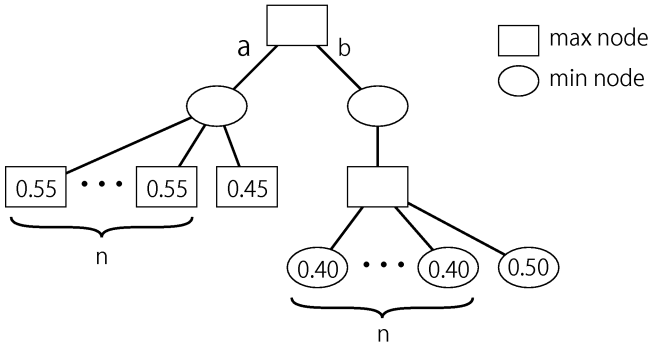

図 3 だまし構造のあるベンチマーク問題の例

図 3 は，だまし構造のある二人ゲーム探索木の例である．末端ノードでは，書かれた数字の確率に従って報酬 1 を返す（自分が勝つ確率であると考える）．木の構造があらかじめ分かっていれば，着手 a の勝率は相手が min を選択するので 0.45，着手 b の勝率は自分が max を選択するので 0.50 であり，b が正解である．しかし，着手 a 側には多くの勝率 0.55 の手があり，相手が正しく選択しない序盤では全体として有望に見える（false positive）．一方着手 b 側は多くの勝率 0.40 の手があり，自分が正しく選択しない序盤では全体として悪く見える（false negative）．

図 4 は，UCT（tuned ではない）で本ベンチマーク ($n = 5$) を解かせた場合の，着手 a,b の勝率の期待値と訪問回数である．前章の実験と同様，序盤は着手 a 側の勝率期待値が高く（左）探索も重点化されている（右）．中盤，勝率が逆転し，終盤は着手 b 側にだけ探索が集中している．そして，45000〜65000 回訪問のあたりでは，訪問回数最大の手 (a) が正解でない，という点も囲碁のケースと同じである．

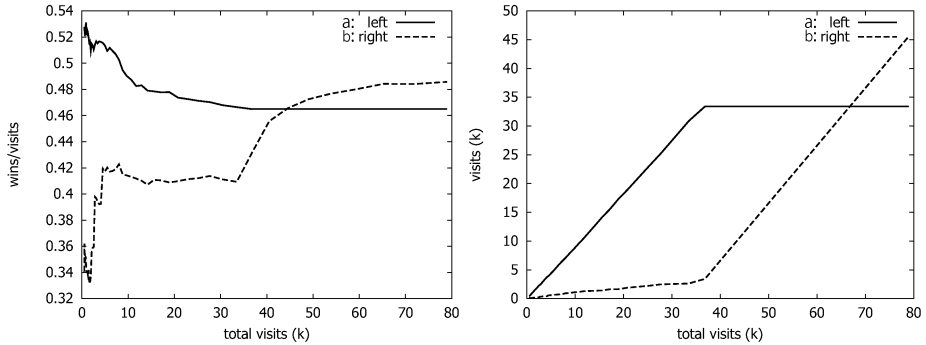

図 4 UCT による探索，勝率の期待値（左）と訪問回数（右）

図 5 は，100 回同様の探索を行い，40000 回訪問時点での状況をプロットしたものである．横軸は，着手 a の勝率期待値－着手 b の勝率期待値を表し，これが正の場合は，勝率の推定・比較に失敗しているということになる．縦軸は，着手 a の訪問回数－着手 b の訪問回数を表し，これが正の場合は正解を打てないということになる．図を見ると，第一象限（勝率推定も打つ手も不正解），第三象限（勝率推定も打つ手も正解）に加え，第二象限（勝率推定には成功しているが，打つ手は不正解）にも相当数の点があることがわかる．これもまた，訪問回数最大を選択することの弱点を証明している．

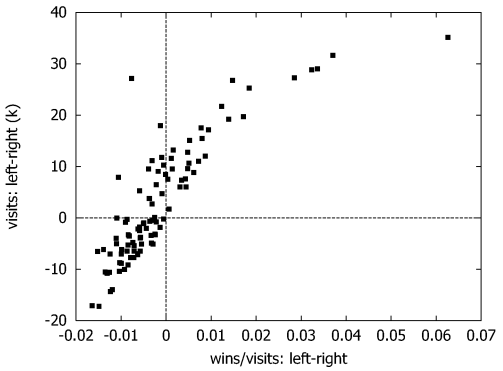

図 5 40000 回訪問時点での，勝率差（横軸）と訪問回数差（縦軸），100 試行

3.2 バリエーション

本稿で提示した木（図 3）は，だまし構造のあるベンチマークの一例に過ぎない．他にも重要なものとしてだまし構造が positive negatve の片側しかなかったり，0.3 や 0.7 などの不純物的なノードが付いていたり，だまし構造が 2 手目と 4 手目に二重になっていたりなど，多くのものが考えられる．一方で，現実に存在しないようなケースをあれこれ想像して極端なベンチマークを作ることもあまり生産的とは言えない．今後は，現実の問題を分析しつつ，系統的なベンチマークの整備を行う予定である．

4. おわりに

本稿では，だまし構造が現実の問題に存在することを示し，その本質を抜き出してベンチマーク化したが，それに対処する方法については述べてこなかった．ここでは，幾つかの方向性を提示することで結びとし，具体的な提案と実験は別稿に譲ることにする．

まず，UCB 値の補正項パラメータ C を大きくすることは，有望そうでないノードにもある程度の探索資源を割くことで，だまされる程度を小さくすることができると考えられる．しかし一方，本当に有望でないノードにも資源を割くため，だましがない場合の性能の低下が深刻であると予測する．

次に，勝率の推定には成功しているのに着手を誤る点については，訪問回数最大の手を選択するのではなく，LCB 値（UCB 値の補正項の符号をマイナスにしたもの）最大の手を選択するなどの方法で回避できる．ただし，これはあくまで探索終了後の話であり，探索中にだまされることの改善にはならない．

一見良い自分の手に対して，重点探索ののち相手の好手が発見された場合，相手はその手を選ぶ可能性が高い訳であり，今まで探索した他の手の勝率を尊重し平均化して親ノードに渡すことは適切でない．このような場合，過去の記憶を忘れる，あるいは min(max) 値を親ノードに渡すなどの手法も有望であろう．ただしこの手法も，だましがない場合の性能低下とのトレードオフは避けられない．

いずれにせよ，各着手の勝率が静的なものであるとして開発された UCB をそのまま二人ゲームに適用することには限界があると考える．さまざまなベンチマーク問題を作り，それらに対応できる手法を新規に考える必要があるのではないか．

参 考 文 献

1) Auer, P., Cesa-Bianchi, N. and Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem, *Machine Learning*, Vol.47, pp.235–256 (2002).
2) Coulom, R.: Efficient Selectivity and Backup Operators in Monte-Carlo Tree Searcg, *CG06* (2006).
3) Gelly, S., Wang, Y., Munos, R. and Teytaud, O.: Modification of UCT with Patterns in Monte-Carlo Go, Technical Report RR-6062, Inria (2006).
4) Kocsis, L. and Szepevári, C.: Bandit based Monte-Carlo Planning, *European Conference on Machine Learning*, pp.282–293 (2006).
5) Tom, D. and Müller, M.: A Study of UCT and Its Enhancements in an Artificial Game, *ACG 2009*, pp.55–64 (2009).
6) 池田　心，小林重信：GA の探索における UV 現象と UV 構造仮説，人工知能学会論文誌，Vol.17, No.3G, pp.239–246 (2002).