2N5-OS-03b-3

抽象的なゲーム木探索に対する因果的価値関数の有効性

Effectiveness of Causal Value Function for Game Tree Search

大用 庫智*1 Kuratomo Oyo
小川 翔太郎*2 Shoutarou Ogawa
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

Classical search methods on game trees are based on an evaluation function that enables quantitative treatment of the state and a strategy that utilizes the function such as Mini-Max method. Construction of the evaluation can be quite hard, especially for a game with huge search space like the game of Go. Recently, UCT, a Monte Carlo tree search method with bandit-like sampling allocation, has been shown to be very effective. We propose *LST* that utilizes an action value function implementing causal intuition of human. With tuning of an intuitive parameter, it enables faster search of the optimal actions with a kind of satisficing behavior.

1. はじめに

1950 年前後から、チェス・囲碁・将棋等の二人ゼロ和確定情報ゲームの研究は、人間よりも強いゲーム AI の作成を一つの目標として研究が進んできた[人工知能学会 08]。チェスの研究当初は有益な探索方法の研究であったが、現在はその効率的な探索方法が強いゲーム AI を作成する重要な要因になっている。1990 年代から 2014 年現在までに、チェッカー・リバーシ・チェス・将棋等の AI は人間のプロよりも遥かに強くなっている。ゲーム AI の研究は商業的に有益であり、また、チェッカーの最善手は引き分けになること[Schaeffer 07]が *Science* に掲載されていることからも学問的にも重要な研究分野であることが分かる。

これらのゲーム AI の着手の探索法は、ゲームの局面を点数化する評価関数とその評価を活かす戦略(Mini-Max 法)が基本になっている。しかしながら、人工知能研究の歴史の中で成功を収めてきたその探索法は、囲碁のゲーム AI には全く役に立たなかった。これは、囲碁が他のゲームと比較して、評価関数の作成が困難であり、かつ、探索空間が膨大であることが主な要因である。このために、より汎用的で効率的な探索方法の登場が望まれていた。

この問題点を解決するために、[Bruegmann 93]はランダムな着手を初期局面(ルートノード)の子ノードからゲーム終局まで繰り返してサンプリングを行う原始モンテカルロ法を提案した。この方法では評価関数を必要とせず、ゲーム終局の勝ち(1)負け(0)の情報から計算された価値を実際の着手の行動価値としている(当初はゲーム終局の差(得点差等)が用いられていたが、勝率を追求する方が遥かに強いゲーム AI になることが知られている[美添 2012])。

しかし、この方法ではサンプリングに無駄が多く、強い囲碁 AI の作成は困難であった。このため、この方法に木探索とランダムなサンプリングを工夫するバンディット問題のアルゴリズムを組み合わせたモンテカルロ木探索が登場し、特に着手の行動価値関数部分にバンディット問題の標準的なアルゴリズムである UCB を組み合わせた方法は UCT (UCB applied to trees) [Kocsis 06]と呼ばれている。40 年以上進展することがなかった囲碁の研究は UCT によって劇的に進展した[美添 2012]。この UCT では着手の行動価値関数となる UCB が最も重要である。

本研究では UCT の代替案となる *LST* (loosely symmetric model applied to trees)を提案する。本研究のアイディアは UCT の行動価値関数部分に、人間の因果的直感を正確に記述する価値関数(緩い対称性(*LS*)モデル)を用いることである。本研究では、囲碁や将棋、チェス等のゲーム固有の知識を利用しない抽象的なゲーム木を用いる。その木を利用して、[Kocsis 06]で行なわれた様なシミュレーションを通して、最適解収束の速さとその振る舞いの仕方の両方の観点から *LST* と UCT を比較する。

連絡先:大用 庫智、e-mail: kuratomo.oyo[at]gmail.com

2. 抽象的なゲーム木

Kocsis と Szepesv は抽象的なゲーム木[Smith 94]を利用して、UCT が最善手に収束する事を理論とシミュレーションの観点から示した[Kocsis 06]。本研究でも[Kocsis 06]と同様に図 1 のようなゲーム木を用いる。このゲーム木では、MAX と MIN の二人のプレイヤーが交互にゲームを進める。MAX は自身にとって最も有望な着手を選択する。逆に MIN は MAX にとって最も不利な選択をする。このような戦略が Mini-Max 法である 。ゲーム終局を意味する葉ノードには、ゲーム終局の評価値が一定の範囲から一定の確率で与えられる。この評価値が一定の基準よりも高ければ MAX の勝利(図 1 の赤い部分)となり、その基準以下であれば MAX の敗北となる(図 1 の青い部分)。

二人のプレイヤーの着手の行動価値は葉ノードに到達する事で得られる勝ち(1)負け(0)の情報から計算される。そして、現在の局面からその直接の子ノードの行動価値が高い方を選択し、葉ノードまで交互に着手を繰り返す(注意点として二人のプレイヤーは Mini-Max 値を行動価値とすることは出来ない)。この行動により勝ち負けの情報をサンプリングする。本来はランダムにゲーム終局まで着手を繰り返すことをプレイアウトと呼ぶが、本研究では上記のサンプリングをプレイアウトと呼ぶ。プレイアウトと実際の着手選択をするために、UCT ではバンディット問

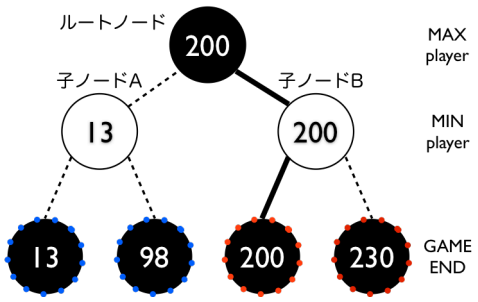

図 1:深さが 2、子ノード数が 2 のゲーム木. 赤と青の点線はそれぞれ MAX プレイヤーの勝ち負けを意味している。実線は Mini-Max 的に最適な選択経路である。

題のアルゴリズムが利用される。

2.1 バンディット問題（逐次的意思決定）

バンディット問題ではそれぞれの当たり確率が不明な複数の腕を持つ一台のスロットマシンから一度に一つの腕を選択することを続ける。目的は累積獲得報酬の最大化である。この問題は探索と知識利用のジレンマとそれが導く速さと正確さのトレードオフを表現する強化学習の最も基本的な問題と見なされている[Sutton 98]。ここで、探索はより良い結果を得ために即時の結果に結びつくとは限らない情報収集であり、知識利用はこれまでに得た情報から局所的な最適化として主観的にベストな選択肢を選ぶ報酬獲得である。バンディット問題は他の強化学習の問題と比較すると単純であるために、探索と知識利用のスイッチングまたはバランシングを行うアルゴリズムの研究が盛んに行われている[e.g. Auer 02]。バンディット問題の専門用語である腕はゲーム木において着手（子ノード）に対応し、コインの有無は勝ち負けに対応し、勝ち負けはそれぞれ 1 と 0 に対応付けられる。この探索と知識利用のバランスを取るアルゴリズムによって、囲碁のゲーム AI の研究は劇的に進展を遂げた。

(1) UCB (upper confidence bound) アルゴリズム

バンディット問題の現在最も標準的なアルゴリズムは UCB1 である[Auer 02]。このアルゴリズムはほぼ価値関数にすぎないが、十分な選択回数が許されれば高い成績を示し、期待損失の上界を保証（即ち最適解収束の保証）している。UCB1 アルゴリズムは最初に全ての子ノードを選択しなければならない。その後、価値関数$\mathrm{UCB1}(j)$の値が最も高い子ノードを選択する。

$$\mathrm{UCB1}(j) = \bar{X}_j + \sqrt{2\ln n / n_j}, \qquad (1).$$

ここで、$\bar{X}_j$は子ノード j の期待値（条件付き確率 $P(1|j)$ と一致）であり、n_jは子ノード j の選択回数、n は（jに限らず）各深さに到達した選択回数を意味する。UCB アルゴリズムは MoGo 等の囲碁 AI のモンテカルロ木探索に応用されている[Gelly 06]。

(2) 因果的価値関数（緩い対称性モデル）

認知心理学や行動経済学では、確率規則や論理法則などの規範からの逸脱（認知バイアス）の研究が行われ、我々人間はヒューリスティクスに由来するものと考えられる認知バイアスを強く持つことが知られてきた[e.g., Tversky 81]。本研究では着手の行動価値関数として、人間の因果的直感を再現する緩い対称性（loosely symmetric：LS）モデルと呼ばれる条件付き確率的な関数に注目した。LS は篠原によって認知バイアス（特に幼児の語彙獲得に必要不可欠とされている対称性と相互排他性バイアス）を定量的に扱うモデルとして経験的に発見された[篠原 07]。その後、LS は[Takahashi10, Takahashi 11a, 11b]で分析されている。LS は条件付き確率に含まれていない二つの特別な項（$\gamma_1 = bd/(b+d)$ and $\gamma_2 = ac/(a+c)$）を用いて認知バイアスを調整する。

$$LS(1|A) = \frac{a + \gamma_1}{a + b + \gamma_2 + \gamma_1} = \frac{P(A,1) + \gamma_1'}{P(A) + \gamma_2' + \gamma_1'}, \qquad (2).$$

$$LS(1|B) = \frac{c + \gamma_1}{c + d + \gamma_2 + \gamma_1} = \frac{P(B,1) + \gamma_1'}{P(B) + \gamma_2' + \gamma_1'}, \qquad (3).$$

ここで$\gamma_1' = \gamma_1/(a+b+c+d)$と $\gamma_2' = \gamma_2/(a+b+c+d)$である。$A$と$B$は図 1 の様な子ノード A と B に対応する。共変動情報aは1とAが共に発生した頻度を意味する$N(A,1)$である。同様にb, c, dはそれぞれ$N(A,0)$と$N(B,1)$, $N(B,0)$を意味する。

このモデルは原因と結果の共変動情報（a, b, c, d）から帰納的に因果関係を推論する因果帰納実験のメタ分析において、人間の因果的直感と高い相関関係（$r = 0.96$）を持つことが示されている[Oyo 13]。また、因果関係を学習した結果と意思決定が一貫しているかを調査した実験においても LS は高い相関関係（$r = 0.98$）を持つ事が示されており、因果帰納モデルを意思決定課題に使う正当性が示されている[大用 14]。そして、この因果的な価値を行動価値関数としてバンディット問題に用いると従来の方法（ϵ-greedy 法や Softmax 法、UCB）よりも高成績を示すことが分かっている[Oyo 13, 14, 大用 14]。また、このモデルは、我々人間が持つ一定の基準で満足に至る選択肢を探す満足化の原理[Simon 56] や振る舞いの心理的フレームを決定する参照点[Tversky 81]、複数の選択肢の順位付けをより容易にするための相対評価[Kahneman 79]などのヒューリスティクス（人間的特性）を実装している[大用 14]。満足化基準Rを持つ LS の計算は$a,\ c$に対して$2(1-R)$を乗算し、b, dに対して$2R$を乗算すれば良い。この基準はデフォルトで 0.5 である。LS は満足化基準に従い振る舞いが変化する[大用 14]。

モンテカルロ木探索に LS を実装する方法は容易である。UCT の行動価値$\mathrm{UCB1}(j)$を LS に置き換えて、UCB1 の初期方策を削るだけである（UCT の疑似コードは[Kocsis 06]を参照）。また、LS の計算は共通項が含まれるため見た目より簡単である。

3. シミュレーション

ここでは[Kocsis 06]と同様な設定を用いて、最適解収束の速さとその振る舞いの仕方の観点から LST と UCT を比較する。

3.1 設定と指標

抽象的なゲーム木の深さと一つの親ノードが持つ子ノードの数は、それぞれ 20 と 2 とした。ゲーム終局を意味する葉ノードには確率P_Xが設定されている。葉ノードは確率P_Xで[128, 254]、$1-P_X$で [0, 127]の評価値が割り振られる。このゲーム終局の評価値が 128 以上であれば MAX の勝敗は勝利となる。図 1 の様に子ノード A 以降の葉ノードにはP_A、子ノード B 以降の葉ノードにはP_Bの確率が設定されている。即ち、P_Xが高ければ MAX にとって有利な局面を意味する。この設定に従い 100 種類の木を生成した。そして、それぞれの木に対して 1000 プレイアウトを 100 回実行し、その平均を結果とした。

指標は正解率と切り替え率の二つを用いる。正解率は Minimax 的に最適なノードを選択した割合である（具体例は図 1 を参照）。切り替え率は前回の選択から選択肢を変えた割合である。これらの指標は各プレイアウト回数の後に計算される。

初期設定として UCB1 には先頭打着緊急度[Gelly 06]の考え方を用いて UCB1 の初期方策を実現する。LS の共変動情報には初期値として 1 が代入されている。この設定は LS が UCB1 などの他の方法と比較して相対的に最も不利な設定であるので、LS の性能を示す本論文では最もフェアな設定である[Oyo 13]。

3.2 結果 1

ここでは高確率環境と単高確率環境、低確率環境の三種類の確率設定毎に結果を示す。高確率環境のP_AとP_Bは 0.8 と 0.6 とし、単高確率環境のP_AとP_Bは 0.6 と 0.4、低確率環境のP_AとP_Bは 0.4 と 0.2 とした。つまり、高確率環境では MAX が常に勝利できるような有利な局面を想定した設定であり、低確率環境では常に敗北してしまうような不利な局面を想定した設定である。また単高確率環境では一つの選択肢が唯一勝利可能な局面を想定した設定である。ここで確率の設定は LS の基準である 0.5 を基準としている。

上記の様な環境毎に LS と UCB1 の結果を図 2 に示す。図 2 の結果から UCB1 は環境毎に殆ど切り替え率を変えず、正解率も殆ど同じである。一方、LS は環境毎に切り替え率を変化させており、高確率と単高確率環境では LS の切り替え率は低い

が、低確率環境では切り替え率が極端に高くなるという特徴がある。また、*LS* の正解率は高確率環境では UCB1 に劣るが、単高確率と低確率環境では UCB1 よりも高い事が分かる。

LS と UCB1 は探索方法に差異があるため、その差異に焦点をあてて、同一の Mini-Max 値を持つ木を用いて両者を視覚的に比較する。木は深さ 3 と子ノード数 2 である。訪問回数毎にノード内の色は白から赤への濃淡が設定する。プレイアウト 1000 回後の差異の例を図 3 に示す。図 2 同様に図 3 から UCB1 の振る舞いの変化はあまり見られない。図 2 の結果を踏まえると図 3 から *LS* は満足化基準に従い探索する空間の広さを調整している事が視覚的に分かる。*LS* は基準よりも良い局面（高確率と単高確率）では瞬時に勝利する着手に執着する。また、同様に *LS* は基準よりも低い局面では活路（基準よりも良い局面）を探し続けている。

3.3 結果 2

図 2 と図 3 から *LS* と UCB1 には振る舞いの仕方に大きな違いがあり、単高確率環境の様に満足化基準が適切であれば UCB1 よりも遥かに高い成績を *LS* は示す事ができる。そこで、ここでは図 1 と同じ設定の高確率と低確率環境において *LS* の満足化基準を$\min(\bar{P}_A, \bar{P}_B) + |\bar{P}_A - \bar{P}_B| \times 0.5$と設定した結果を図 4 に示す。ここで$\bar{P}_X$は子ノード X 以降のそれぞれの葉ノードで得られる勝ち負けの情報から計算した勝率である。

図 4 の結果から *LS* の正解率と切り替え率が図 2 の単高確率環境と同様になっていることが分かり、UCB1 よりも高成績であることが分かる。これは *LS* の行った満足化が木探索における最善手の発見という最適化に一致したためと考えられる。

4. 議論

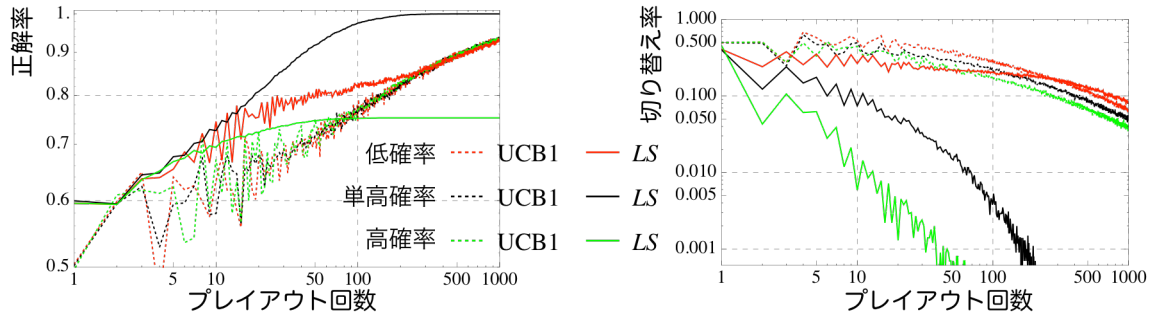

図 2:LS と UCB1 の正解率（左）と切り替え率（右）

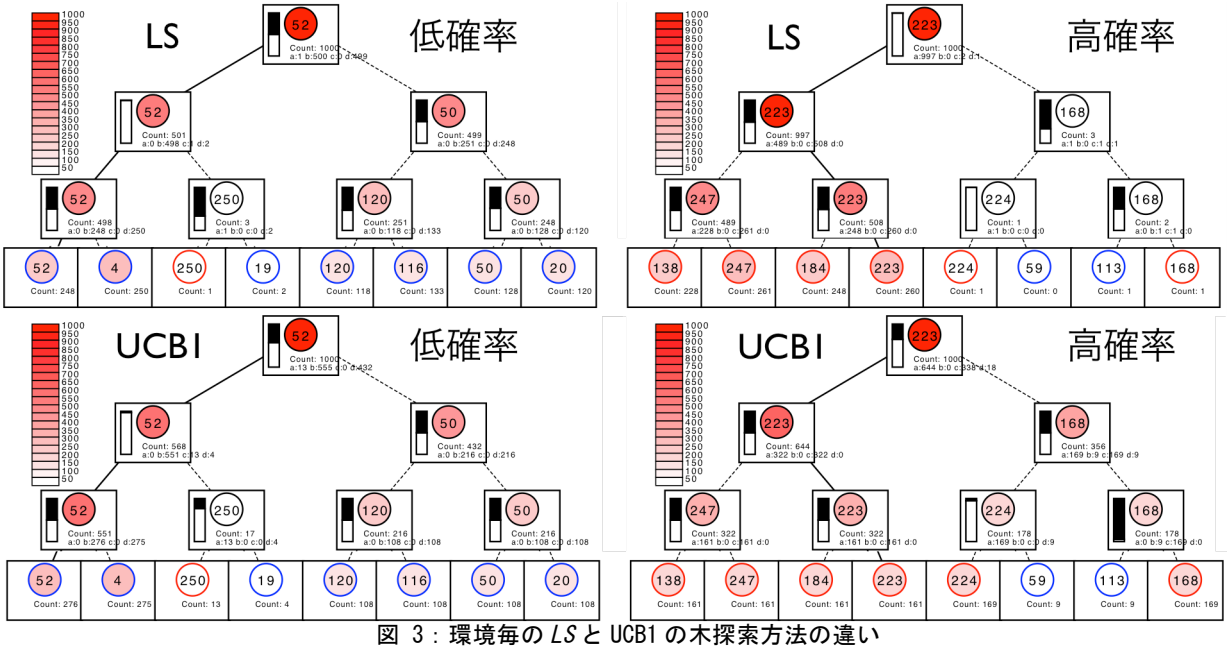

図 3：環境毎の *LS* と UCB1 の木探索方法の違い

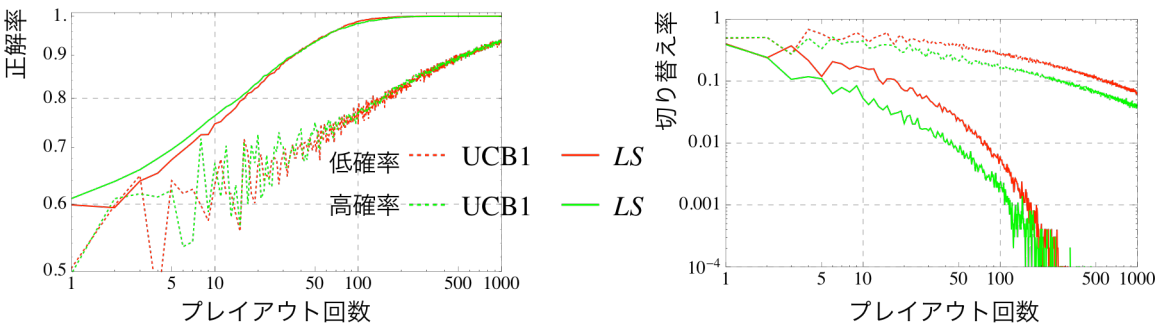

図 4：満足化基準を変化させた *LS* の正解率（左）と切り替え率（右）

本研究では、囲碁 AI 等で多用されているモンテカルロ木探索に人間認知から触発された *LS* を実装した *LST* を提案した。そして、UCT の提案論文[Kocsis 06]と同様なシミュレーションを通して *LST* と UCT を比較した。その結果、*LST* は UCT よりも速く最適解に収束することが分かった（図 2, 4 を参照）。また、*LS* は木探索においても人間の認知に見られる満足化基準に従い特徴的な振る舞いをすることが分かった（図 3 を参照）。本論文の残りでは、モンテカルロ木探索を用いた方法での勝率の向上方法とその振る舞いの仕方を紹介する。そして、それらの内容と本研究の提案手法がどのような場合に有効であるかを議論する。

モンテカルロ木探索登場以前では、ゲーム AI が効率的に探索し、その強さを向上させるためには優秀な評価関数が必要不可欠であった。そのために、その作り込みに膨大な時間を費やしていた。強いゲーム AI を作成するために、モンテカルロ木探索においても優秀な評価関数の作り込み同様な方向でプレイアウトの強化が進んでいる[e.g., Gelly 06]。ここで言うプレイアウトとは本来の意味であるランダムな着手をゲーム終局まで行うという意味である。現在の多くの囲碁 AI 等では、このランダムな部分にゲーム依存の知識を利用した良質なプレイアウトを用いることで、一回のプレイアウトに長い処理時間を費やしてもプレイアウトの回数自体を減らす方向に進んでいる。しかし、これは UCB1 が膨大なサンプリング数を前提に高性能を示す、という考えとは逆方向である。このため、今後は *LS* のように速くかつ正確な方法がより必要になると考えられる。

モンテカルロ木探索を用いたゲーム AI は勝率の最大化を目標に振る舞う。このような振る舞いは従来の虱潰し的アルゴリズムには難しく、初めてモンテカルロ木探索を実装した Crazy Stone [Coulom 06]の登場時にはその振る舞い仕方に注目が集まった[美添 2012]。本研究の結果から、UCB1 は局面の優劣毎に各指標の変化は現れていないため、モンテカルロ木探索自体がコンピュータの特徴的な振る舞を生む要因であると考えられる。一方、*LS* は満足化基準に従い自律的に振る舞いを変化させている。*LS* が満足化基準よりも現在の局面が良いと判断した場合には、勝利する着手を楽観的に素早く選択し執着する。この行動の中でも（満足化基準よりも高い）勝利する着手が一つのみの場合は最適な行動をする（図 2 の単高確率環境と図 4 を参照）。逆に *LS* が現在の局面状況では勝利の見込みは薄いと判断した場合には、一定の基準（満足化基準）を超える着手を探し続けるため良く探索する空間を広げて広大に探索する。このような *LS* の振る舞いは現在のゲーム AI においては稀少であると考えられ、単に一つの満足化基準 *R* を直感的に調整すれば良いので、強いまたは弱いゲーム AI の作成も容易である。

次に本研究で提案した *LST* が具体的にどのような場合に有効であるかを議論する。最善手と次善手の評価値が著しく離れるゲームの種類がある（例えばチェス等）。このような状況において図 4 で示したような満足化基準よりも高い選択肢を瞬時に見つける *LS* の性質が有効に働くと考える。上記とは逆に、最善手と次善手の評価値に差があまりないゲームの種類もある（例えば囲碁等）。また、モンテカルロ木探索にはほぼ敗北する状況の中で数少ない活路がある場合に収束が遅くなる欠点がある。*LS* は二つの選択肢の評価値の差が著しく小さいバンディット問題でも UCB よりも速く最適解に収束することが示されており、また準最適解から抜け出して最適解を得られることも示されている[Oyo 13, 14]。これらの *LS* の性質と本研究で示した探索を促す能力が囲碁等や収束遅延の問題に有効に働くと考えられる。

LS は選択肢の数が多くなればなるほど UCB よりも高成績を示すため[大用 14]、選択肢が多い状況（特に囲碁の様なゲームの種類）ではより高い成果を期待できる。

5. 結語

人間認知に触発された *LS* を用いて *LST* を提案した。この方法は UCT よりも速く最適解に収束し、人間に見られるような満足化を基準とした特徴的な振る舞い示した。*LST* はゲームの知識を利用して作り込まれた囲碁 AI により有効に働き、満足化基準を変化させることで多くの例で有効に働く可能性を議論した。本研究の結果から実際のゲーム AI に *LST* を実装することは有益であると考えられるため、強いゲーム AI またはエンターテイメント用のゲーム AI として *LST* を具体例に今後実装する。

参考文献

[Auer 02] Auer, P., Cesa-Bianchi, N., and Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem, *Machine learning*, 47, 23–256 (2002).

[人工知能学会 08] 人工知能学会：デジタル人工知能学事典, 共立出版(2008).

[Bruegmann 93] Bruegmann, B.: Monte carlo go (1993).

[Coulom 06] Coulom, R.: Efficient selectivity and backup operators in monte carlo tree search. In P. Ciancarini and H. J. van den Herik, editors, *Proceedings of the 5th International Conference on Computers and Games*, Turin, Italy (2006).

[Gelly 06] Gelly, S., Wang, Y., Munos, R., and Teytaud, O.: Modification of UCT with Patterns in Monte-Carlo Go, Technical Report 6062, INRIA (2006).

[Oyo 13] Oyo, K. and Takahashi, T.: A Cognitively Inspired Heuristic for Two-Armed Bandit Problems: The Loosely Symmetric (LS) Model, *Procedia Computer Science*, 24, 194–204 (2013).

[Oyo 14] Oyo, K. and Takahashi, T.: A Human Causal Value Function and Its Optimality under Greedy Method for Two-Armed Bandit Problems, (AROB 19th 2014), 113－118 (2014).

[大用 14] 大用庫智, 高橋達二：緩い対称性を持つ因果的価値関数の妥当性とバンディット問題に対するその有効性. (準備中)

[Simon 56] Simon, H. A.: Rational choice and the structure of the environment, Psychological Review, 63(2), 129–138 (1956).

[Schaeffer 07] Schaeffer, J., Burch, N., Björnsson, Y., Kishimoto, A., Müller, M., Lake, R., Lu, P. and Sutphen, S.: Checkers Is Solved, *Science*, 317(5844), 1518–1522 (2007).

[Sutton 98] Sutton, R. S., and Barto, A. G.: *Reinforcement Learning: An Introduction*, Cambridge, MIT Press (1998).

[篠原 07] 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(1), 58–68 (2007).

[Smith 94] Smith, S.J.J., and Nau, D.S.: An analysis of forward pruning, *AAAI*, 1386–1391, 1994.

[Kahneman 79] Kahneman, D. and Tversky, A.: Prospect theory: An analysis of decision under risk, Econometrica, 47, 263–291 (1979).

[Kocsis 06] Kocsis, L. and Szepesv, C.: Bandit based Monte-Carlo Planning, *Machine Learning: ECML 2006 In Proceedings of the 17th European conference on Machine Learning*, 4212, 282–293 (2006).

[Takahashi 10] Takahashi, T., Nakano, M., and Shinohara, S.: Cognitive symmetry: Illogical but rational biases, *Symmetry, Culture and Science*, 21(1-3), 275–294 (2010).

[Takahashi 11a] Takahashi, T., Shinohara, S., Oyo, K., and Nishikawa, A.: Cognitive symmetries as bases for anticipation: A model of Vygotskyan development applied to word learning, *International Journal of Computing Anticipatory Systems*, 24, 95–106 (2011).

[Takahashi 11b] Takahashi, T., Oyo, K., and Shinohara, S.: A loosely symmetric model of cognition, *Lecture Notes in Computer Science*, 5778, 234–241 (2011).

[Tversky 81] Tversky, A. and Kahneman, D.: The framing of decisions and the psychology of choice, Science, 211, 453–458 (1981).

[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–, 共立出版 (2012).