様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月　 ７ 日現在

機関番号：３４５０４
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１２８０９

研究課題名（和文）認知的満足化に基づく探索技法の開発

研究課題名（英文）Development of Search Technique based on Cognitive Satisficing

研究代表者
大用　庫智（Oyo, Kuratomo）

関西学院大学・総合政策学部・講師

研究者番号：６０７５５６８５

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：ゲームAI やロボティクスを初めとした巨大な探索空間において、強化学習で最も重要な課題である「速さと正確さのトレードオフ」の既存の限界を超える手法の開発が行われている。そこで本研究では、その既存の限界を突破するため、既存研究とは別のアイディアとして人間の満足化の概念を探索技法の中心的な役割を果たす木探索（モンテカルロ木探索）へ実装し、新しい探索技法の開発を進めた。また、モンテカルロ木探索において満足化の優れた性能を示した。

研究成果の学術的意義や社会的意義
本研究は、人間の適応的な意思決定方法を探索能力として活用する。強化学習技術の枠組を用いて、その探索能力を探索技法の中心的な役割を果たす木探索への実装に着目した。本研究は一般性が高いものであると考えており、各既存問題に応用する際に、複雑なアルゴリズム化等が不要であることから、バンディット問題の応用例である様々なゲームAI や、スケジューリング、最適化問題等の幅広い領域での結果の一般性が期待でき、幅広い探索問題への波及効果があると考えている。

研究成果の概要（英文）：In huge search spaces such as game AI and robotics, the purpose of reinforcement learning is the development of methods that exceed the existing limits of the trade-off between speed and accuracy, which is the most important task. In this study, in order to exceed the existing performance limits, we applied the concept of human satisficing which is another idea different from existing research to tree search (Monte Carlo tree search) , which plays a central role in search methods. In Monte Carlo tree search, we showed the efficiency realized by the satisficing model.

研究分野：人工知能，知能情報学

キーワード：強化学習　機械学習　バンディット問題

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

国内外問わず、囲碁や将棋のコンピュータが人間のプロに勝利した、という報道が頻繁に行われている。それらのような巨大な探索空間を持つ問題に対しての成果は、近年の強化学習技術の発展によるものである。技術的には囲碁研究を劇的に進展させたモンテカルロ木探索 [Kocsis & Szepesvari, 2006] が中核にある。これはランダムな行動による情報収集(プレイアウト)とその配分の工夫により「最も勝ちやすい次の選択肢」を決定する方法である。モンテカルロ木探索の汎用性の高さはスケジューリングや最適化問題等の広い研究領域で示されている[Browne et al., 2012]。しかし、モンテカルロ木探索にはランダムなサンプリングを行うため膨大なサンプリングが必要であり、そのため計算コストの高さが知られていた。また、効率的な探索を阻害する要因として、次の着手からゲーム終了までの間が長い、深い探索が困難であると指摘され続けていた。

研究代表者はこれまで、強化学習を含む人工知能研究において、認知心理学や行動経済学の分野の中での確率規則や論理規則などの合理的エージェントにとっての規範からの逸脱（人間認知の偏り）の適応的な意味を明らかにしてきた。特に強化学習の最も基礎的な課題（n 本腕バンディット問題）において、人間の満足化、相対的な評価とリスクに対する態度を経験ベイズ法の形式を持つ行動価値関数（緩い対称性モデル[篠原, 2007]）として整備し、それをアルゴリズムとして実装してきた。本研究計画と関連が深い満足化は「受容可能な基準を満たす選択肢を見つけると探索をやめる」という概念である。研究代表者は強化学習の中で従来の解法よりも少ない試行回数で高い性能を実現する、という効率的な満足化を研究した。

研究代表者は研究対象であった緩い対称性モデルが複雑であったため、そのモデルから効率的な満足化の機能を抽出し、より単純化した満足化価値関数 RS（reference satisficing）が、木探索において効率的な探索を実現できると考えた。そこで認知心理学や行動経済学の知見を活用しながら、強化学習の枠組みで、単純な計算論的モデル（RS）をモンテカルロ木探索に応用し、巨大な探索空間でも効率的な探索が行える探索技法の開発を目指した。

２．研究の目的

近年、ゲーム AI やロボティクスを初めとして、巨大な探索空間において、強化学習で最も重要な課題である「速さと正確さのトレードオフ」の既存の限界を超える手法の開発が行われ続けている。そこで本研究では、その既存の限界を突破するため、既存研究とは別のアイディアとして人間の満足化の概念を探索技法の中心的な役割を果たす木探索（モンテカルロ木探索）へ実装し、新しい探索技法を開発する。そして、満足化の探索能力の性能を最も単純な課題と一般性を持つ探索問題において示しつつ、モンテカルロ木探索の性能向上を試みる。本研究は一般性の高い課題において検証し、研究方法にて説明する満足化価値関数は期待値や条件付き確率という簡単な数式の枠組にそのまま利用することができるため、波及効果として幅広い探索問題とその応用例の性能向上が期待できる。

３．研究の方法

まず、モンテカルロ木探索で広く用いられている既存の行動価値関数 UCB を説明しながら本研究で扱う満足化価値関数 RS を説明する。

モンテカルロ木探索を活用して効率的に探索を行うためには、プレイアウトの配分の工夫として速さと正確さのトレードオフに上手く対処することが最も重要な問題になる。従来のモンテカルロ木探索は、プレイアウトによって得られた情報（例えば囲碁などでは勝ち負け）から、行動価値関数 UCB に従い探索空間を探索する。それにはバンディット問題の解法として「試行錯誤が不十分な行動には本当は良いかもしれないと下駄をはかせて評価する」という楽観主義（Upper Confidence Bound：UCB）的アルゴリズムの様々なバージョンがある[Garivier & Cappe 2011]。そのため、ここでは簡略化のために、二つの選択肢があるとして UCB を定義する。モンテカルロ木探索では勝ち負けの情報(バンディット問題では報酬の有無)のみで十分であるため、選択肢 X の客観的な価値は条件付き確率 P(勝ち|選択= X) = P(X,勝ち)/P(X)と一致する。この条件付き確率に、サンプル数が少ない選択肢への探索を促すための項（$\sqrt{2\ln n/n_x}$）を加えた UCB(X) = P(勝ち|選択= X)+$\sqrt{2\ln n/n_x}$ が UCB である。ここで n_x は選択肢 X の選択回数，n は（X に限らず）全体の選択回数を意味する。この UCB の値が最も大きい選択肢を選びながら行動することにより、プレイアウトの配分の工夫が可能になる。しかし、UCB は探索を促すための項（$\sqrt{2\ln n/n_x}$）により正確性を重視する代わりに速さが犠牲になるため、扱う問題の探索空間が広大であればあるほど、膨大な計算コストが必要となる。

そこで、本研究では UCB の代わりに満足化価値関数を用いる。満足化価値関数は選択肢 X を RS(X)=n_x(P(勝ち|選択=X)-R)と価値付けする。ここで R は各選択肢の価値に対する共通の基準である。この基準はモンテカルロ木探索において、次のような満足化行動を関数レベルで実現する。もし基準 R 以上の P(勝ち|選択=X)を持つ選択肢があればそれを選択する（知識利用）。また、そのような行動が一つもなければ、選択肢の中からランダムに選択肢を選択する（探査）。

本研究課題では次のように研究を進めた。
(1) モンテカルロ木探索において効率的に探索するためのプレイアウトの配分には、N 本腕バンディット問題において、これまで得た知識の活用（知識利用）による速さの優先か、知識を増やすための情報収集（探査）による正確さの優先かのトレードオフに対処する必要があると考えられている。そこで広大な探索空間において満足化の性能を検証する前に、モンテカルロ木探索の基礎となるバンディット問題のシミュレーション実験にて、各アルゴリズムのトレードオフへの対処方法と経験的な基本性能を明らかにする。その際に最も初期の UCB1 から最も高性能な KL-UCB+ [Garivier & Cappe 2011] を満足化の比較対象とした。

(2)満足化と探索技法の中心的な役割を果たす木探索を連携させる。これにより巨大な探索空間でのトレードオフに対処可能で、モンテカルロ木探索の問題点を克服可能な探索技法を開発する。満足化価値関数を活用したモンテカルロ木探索により効率的な探索を実現する探索技法を提案した。そして、一般化された探索問題を通して、その提案手法の基本的な性能を示し、モンテカルロ木探索の性能向上を目指す。代表的な UCB を比較対象としてコンピュータシミュレーションを行った。

以下に(1)と(2)の具体的な方法を述べる。
(1)N 本腕バンディット問題での研究方法
本研究課題で扱う N 本腕バンディット問題は、不確実性下の逐次的な意思決定を分析する心理学の実験課題としても知られている。現在では、この問題は人工知能の機械学習の一種である強化学習という分野の中で最も基礎的な問題とみなされている。この N 本腕バンディット問題において RS を行動価値関数として用いて、RS の特性を検証した。N 本腕バンディット問題では n 種類の選択肢があり、それぞれに報酬確率とそれに従い得られる報酬が設定されている。この設定から後悔（最も良い選択肢の期待値と選択した選択肢の期待値の差）という指標を用いて RS の性能を調査することができる。また N 本腕バンディット問題は、正解率と切り替え率、満足度などの指標が知られている。ここでは最も高い報酬確率を持つ選択肢を選択した割合を正解率、n 回選択までに選択肢を切り替えた割合を切り替え率、設定した基準以上の報酬確率を持つ選択肢を選択した割合を満足度とした。比較対象としては n 本バンディット問題の最も知られている UCB シリーズのアルゴリズム、Thompson Sampling などを網羅的に実装し、その中でも最も性能が高い(後悔が低い)KL-UCB+アルゴリズムを RS の比較対象とした。このような設定において、モンテカルロ木探索に RS を実装することを想定し、選択肢の数、報酬確率の設定を網羅的に変更し、RS の特性を検証した。

(2) モンテカルロ木探索での研究方法
本研究課題で扱うモンテカルロ木探索は、N 本腕バンディット問題を応用した探索技法であり、ゲーム木上で N 本腕バンディット問題同様な逐次的な意思決定を行う。モンテカルロ木探索はランダムな行動による情報収集とその配分の工夫により、最善となる選択肢を決定することができる。このモンテカルロ木探索において RS をサンプリングの配分の工夫するための行動価値関数として用いて、RS の探索技法としての特性と性能を検証した。本研究課題では RS の探索技法としての一般的な有効性を示すため、探索問題固有の知識（例えばルール等）を用いない一般化された探索問題においてシミュレーションによる検証を行った。この探索問題は、次はどこに打てるのかのルールなどを加えることで囲碁などの具体的な探索問題を構築する事が可能であり、この問題において高成績を示すことは具体的な問題での性能向上に直接繋がるので一般性がある。そのため、比較対象は UCB として、一般化された探索問題において探索空間を広げながらコンピュータシミュレーションを行った。

４．研究成果

研究の方法で述べた(1)と(2)の研究結果と今後の展望をまとめる。
(1) N 本腕バンディット問題での RS の特性と性能
N 本腕バンディット問題の行動価値関数として RS を利用し、RS の特性を検証した。まず、最も単純な 2 本腕バンディット問題において RS の性能と特性を調べた。RS の基準を元に、低確率設定(全ての選択肢の報酬確率が基準 R 未満)、高確率設定(全ての選択肢の報酬確率が基準 R 以上)、単確率設定(一つの選択肢のみ報酬確率が基準 R 以上)の 3 つの設定でシミュレーションを行った。その結果、正解率と切り替え率の二つの指標から RS には、低確率設定では満足する選択肢がないため探索を優先する傾向、高確率設定では満足する選択肢があるため知識利用を優先する傾向があることを確認した。同様に単確率設定では RS は従来手法よりも性能が高い（後悔が低い）ことが分かった。これから 2 本腕バンディット問題において RS が従来手法よりも性能を向上させるためには、単確率設定のように基準を設定すれば良いことを確認した。次に N 本腕バンディット問題においても RS の性能と特性を調査した。その結果、RS はこれまでの N 本腕バンディット問題での価値関数の定式化よりも単純な形式であるが、それは適切な基準の設定がなされれば、従来の手法よりも高い性能を示すことを確認した。また N 本腕バンディット問題の選択肢の数や報酬確率を変化せても、基準を適切に設定すれば、RS には 2 本腕バンディット

問題同様の振る舞いがあることを確認した。以上のようにモンテカルロ木探索に RS を実装する目処がついた。

(2) モンテカルロ木探索での RS の特性と性能
研究方法の(2)でも述べたように、探索問題固有の知識を用いない一般化された探索問題のおけるモンテカルロ木探索に RS を行動価値関数として組み込んだ探索技法の性能を検証した。まず、探索空間のオーダーが 2^{20}-1 の設定において 2 本腕バンディット問題同様な結果が得られるかを検証した。その結果、一般化された探索問題においても 2 本腕バンディット問題同様な RS の探索と知識利用の振る舞いを確認した。また、一般化された探索問題は探索の難しさの調整として探索空間の広さを容易に調節できるため、探索空間を 2^{25}-1 から 2^{28}-1 まで広げた場合の RS の性能を検証した(具体的には単確率設定のように問題と基準を設定し、プレイアウト回数が 1000 の場合の最適解を選択した割合を求めて検証した)。その結果、探索空間の広さが 2^{26}-1 までは RS と UCB の最適解を選択した割合が 100% に近づくことが分かった。しかし、探索空間の広さが 2^{27}-1 では RS の最適解を選択した割合は 100%、UCB の最適解を選択した割合は 71%であり、さらに探索空間の広さが 2^{28}-1 では RS の最適解を選択した割合は 98%、UCB は 40%以下となった。これらのことから、探索空間が広大になっても基準以上の選択肢が一つある場合では RS の性能がほぼ低下せず、UCB は探索空間が広大になればなるほど性能が悪くなった。つまり、RS は UCB と比較して探索空間に対するスケーラビリティにも優れており、RS が巨大な探査空間においてもトレードオフの既存の限界を突破できることを示した。さらに一般化された探索問題においても N 本腕バンディット問題同様な RS の振る舞いがみられるかを検証した。その結果、一般化された探索問題においても N 本腕バンディット問題同様な RS の探索と知識利用の振る舞いの変化を確認した。

(3) 今後の展望
これまでの特性の検証の結果から、基準未満の選択肢しかない設定では、RS には選択肢を高い確率で探索する傾向がみられた。その傾向は N 本腕バンディット問題のシミュレーションでも確認したが、モンテカルロ木探索のシミュレーションではより顕著に現れた。このような場合は RS が探索を行い過ぎてしまい、その探索は一見無駄のようにも考えられる。その傾向はさらに複雑な問題において、より顕著に現れる可能性がある。そのため、この部分については基準の変更等で対処する必要がある。

本研究課題は、人間の適応的な意思決定方法や能力を探索能力として活用した。強化学習技術の枠組を用いて、その探索能力を探索技法の中心的な役割を果たす木探索への実装に着目した。本研究課題の巨大な探索手空間でも対処可能な探索技法の開発は進展した。また、本研究課題の RS と関連があり、人間の因果関係を推定する人間認知のモデルでも成果を得ることができた。例えば、その認知的なモデルが、不要な情報を効率的に削ぎ落とせる性能があり、一種の前処理として有効に機能することが分かった。これは本研究課題のテーマが探索であったため付随して明らかにすることができた。本研究課題は一般性が高いものであると考えているため、人間の認知を統合的に扱う枠組や、広く有効な探索技法の開発が可能になると考えられ、より広い分野での研究の波及効果があると考えられる。

今後、本研究課題で得られた成果は学術論文でまとめ、精力的に学会発表するなど研究結果の情報公開に努める。また、これまで得られた結果の簡単な解説などホームページなどを通して発信するように努める。

<引用文献>
(1).Kocsis, L. and Szepesvári, C, Bandit based Monte Carlo Planning, Machine Learning, Proceedings of the 17th European Conference on Machine Learning, 2006.
(2).Browne, C., Powley, E. Whitehouse, D., Lucas, S.,Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., Colton, S.: A survey of Monte Carlo tree search methods, IEEE Transactions on Computational Intelligence and AI in Games, 4(1), 1-43, 2012.
(3). 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
(4). Garivier, A. and Cappé, O. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond, In COLT 2011 - The 24th Annual Conference on Learning Theory, pp. 359-376, 2011.

５．主な発表論文等

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 3 件）

1. Oyo, K., Yamada, T., Sandoh, H. Segmentation of Media Users According to Life Value, 29th European Conference on Operational Research conference, 2018
2. Yokokawa, J., Oyo, K., Takahashi, T. Causal induction under rarity and small data, In Proceedings of The Twenty-Third International Symposium on Artificial Life and Robotics 2018, 991-994, 2018.
3. 高橋　達二，大用　庫智，玉造　晃弘，横川　純貴，稀少性仮定の下での非独立性の判断としての人間の観察的因果推論，2017 年度人工知能学会全国大会(第 31 回)予稿集，4M1-1in1，ウインクあいち（愛知県産業労働センター）愛知県名古屋市，2017.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
なし

６．研究組織

(1)研究分担者
なし

(2)研究協力者
なし