UCTを用いた訓練初期局面の多様化によるTD学習法の改善*

三木 理斗† 三輪 誠‡ 田浦 健次朗¶ 近山 隆‡

† 東京大学工学部 ‡ 東京大学大学院新領域創成科学研究科 ¶ 東京大学大学院情報理工学系研究科

1 はじめに

強化学習によるコンピュータゲームプレイヤーの評価関数の学習法として、TD(λ) 法 [1] はさまざまなゲームで注目すべき成果を挙げている。対戦の経験から学習を行う場合には、なるべく多様な局面を展開することが望ましい。本研究では TD(λ) 法においてエージェントの手の選択を多様化する方法として、序盤数手を UCT 法によって決定する手法を提案する。実験としてブロックスデュオの評価関数の重みを学習し、従来の ϵ-greedy 方策や完全にランダムに序盤の手を決める手法と比較して評価し、提案手法が少ない対局数でよりよい重みを学習することを確認した。

2 TD(λ) 学習法

2.1 ゲームにおける TD(λ) 法の概要

ゲームで多く用いられる評価関数は局面 s のさまざまな特徴（駒の配置、持ち駒の種類など）を評価要素 $\vec{\phi}(s)$ とし、その各要素の重みつき線形和で表した次のような形をしている。

$$V(s) = \vec{\theta} \cdot \vec{\phi}(s)$$

TD(λ) 法では各局面 s で任意の方策にしたがって手を決めて次の局面 s' を決定し、誤差 δ を用いて重みを更新する。

$$\delta = r + \gamma V(s') - V(s)$$

$$\vec{e} = \gamma\lambda\vec{e} + \nabla_{\vec{\theta}} V(s)$$

$$\vec{\theta} = \vec{\theta} + \alpha\delta\vec{e}$$

r は対局終了時のみ与えられる報酬で、勝ちなら 1、負けなら 0、引き分けなら 0.5 とする。γ は割引率と呼ばれるパラメータである。$\vec{e}$ は適格度トレースと呼ばれるもので各対局の最初に零ベクトルに初期化する。α は学習率と呼ばれるパラメータである。

2.2 学習局面の多様化

チェスのような確定ゲームで、かつ棋譜やネットワーク上の対戦サーバを用いずに学習を行う場合は、意図的に選択する手をばらつかせ、ある程度多様な局面を学習するような方策を用いる必要がある。

そのために従来用いられてきた手法に ϵ-greedy 方策がある。これは各局面において通常はもっとも評価値の高い手を選択するが、ある確率 ϵ でランダムな手を選択する方法である。この方法では選択する手が完全にランダムであるため明らかな悪手を打つことが多いことと、ランダムな手を打った前後で TD(λ) におけるバックアップのつながりが途絶えてしまうという問題点がある。

3 提案手法

本研究では序盤 UCT 方策という手法を提案する。この手法では序盤の数手を UCT 法 [2] によって決定する。UCT 法はゲーム木を探索する手法のひとつで、ランダムシミュレーションで手を決定するモンテカルロ法に最良優先探索を組み合わせてより有望な部分木を優先的に展開するように改良した手法である。これを用いた囲碁のプログラムは世界トップクラスの成績を収めている [3]。UCT 法を用いることでランダムよりも良い手が選ばれることが期待できる。また、序盤にのみ指し手の変化をつけるため、TD(λ) のバックアップが切れ切れになりすぎないのでより高い精度の学習ができると考えられる。

4 評価

評価のための実験として、ブロックスデュオ (http://www.blokus.com/) の評価関数の学習を行った。ブロックスデュオは陣取り形式でランダム要素のない二人対戦ゲームである。

ϵ-greedy 方策、序盤 UCT 方策と、比較のため序盤数手に完全ランダムに手を決定する方策 (以下、序盤

*Improving Temporal-Difference Learning Using a Variety of Opening Moves Obtained by UCT
Ayato Miki†, Makoto Miwa‡, Kenjiro Taura¶ and Takashi Chikayama‡
†Faculty of Engineering, The University of Tokyo
‡Graduate School of Frontier Sciences, The University of Tokyo
¶ Graduate School of Information Science and Technology, The University of Tokyo

ランダム方策) の 3 方策についてそれぞれ自己対戦による学習を行った。評価要素は 2007 年度の GPCC のブロックスデュオ大会 2 位のプログラムと同じものを用い、それぞれの重みの初期値は 0 に設定した。λ は 0.8、割引率 γ は 1 とし、学習率 α は次式にしたがって減衰させた [4]。

$$\alpha = \alpha_0 \frac{N_0 + 1}{N_0 + \text{Episode\#}^{1.1}}$$

今回は α_0 は 0.01、N_0 は 100 とした。Episode# は対局回数である。ϵ-greedy 方策は 1000000 対局、序盤 UCT 方策および序盤ランダム方策は 100 対局の学習を行った。序盤に UCT やランダムを用いる深さは 4 手目までとし、UCT のシミュレーションは 20000 回行った。

図 1 に代表的な評価要素の最初の 100 対局での重みの変化を示す。序盤 UCT 方策と序盤ランダム方策は重みの学習の様子が見られるが、ϵ-greedy 方策では重みにほとんど変化がない。200 対局あたりまで重みが変化していない状態が続いたが、その後変化を始め、2000 対局ほどで他の方策と同じ重みに到達した。これは終局の形が大きく変わるほどの手の変化が発生しなかったためと考えられる。序盤ランダム方策も 100 対局で序盤 UCT 方策とほぼ同じ重みを学習しているが、最初の 20 局ほどまではやや不安定な変動が見られる。

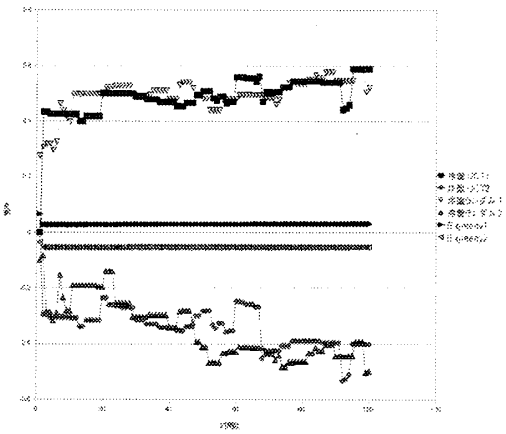

図 1: 重みの変化

5 おわりに

序盤 UCT 方策を用いることで、完全にランダムな従来の手法よりも安定した学習を行うことができる。今回実装した UCT 方策による学習では、学習中に UCT 法を用いてシミュレーションを行うため手の決定に非常に時間がかかっており、学習時間は ϵ-greedy 方策とそれほど変わらない。しかしあらかじめ UCT 法のシミュレーションを行って序盤の手をデータベース化しておけば、以後はその中からランダムに選択することで序盤ランダム方策と同じ速度で学習を行うことができる。本手法で得られた評価関数を大会 2 位のプログラムと対戦させたところ、勝率は 1 割程度であった。自己対戦のみで強いプレイヤーを作るにはまだ改善の必要があるだろう。

参考文献

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[2] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *ECML*, Vol. 4212 of *Lecture Notes in Computer Science*, pp. 282–293. Springer, 2006.

[3] Sylvain Gelly, Yizao Wang, Rémi Munos, and Olivier Teytaud. Modification of UCT with patterns in monte-carlo go. Rapport, HAL - CCSd - CNRS, 2006.

[4] Alborz Geramifard, Michael Bowling, Martin Zinkevich, and Richard S. Sutton. iLSTD: Eligibility traces and convergence analysis. In Bernhard Schölkopf, John Platt, and Thomas Hoffman, editors, *NIPS*, pp. 441–448. MIT Press, 2006.