合議アルゴリズムによる AlphaDDA の動的強さ調整の改良

1265099　久保田　留奈　【 高度プログラミング研究室 】

Improving Dynamic Strength Adjustment of AlphaDDA with Ensemble Algorithms

1265099　KUBOTA, Runa　【 High-Level Programming Lab. 】

1　はじめに

近年，人工知能の発展は目覚ましく，ゲーム分野では人間のトッププロを超える強さの AI が開発されている．またそのような人間を超える強さを持つ AI を，人間の練習相手のために用いることを目標とした研究 [1] も行われている．人間相手にプレイする場合，AI プレイヤーが極端に強い手・弱い手を打つと人間は不快を感じるため，対戦相手の強さに合わせて AI プレイヤーの強さを適切に調整する動的強さ調整が重要である．

先行研究 [1] では，プレイヤーはおよそ 50 %の勝率を達成できていたが，対局の勝率のみによる評価では不十分だと考え，各局面における局面評価値の変動による再評価を行った [2]．その結果，実際は局面評価値は大きく振れる場面があり，ゲーム全体に渡って適切に強さを制御できているわけではなかった．

そこで本研究では，強い AI プレイヤーを作成する手法の一つとして成功を収めている [3] 合議と呼ばれる，ある局面において複数の思考プログラムの候補手から一つの手を選択するような手法を用い，より互角に近い手を選択させることで強さが調整可能か明らかにする．

2　AlphaDDA

動的強さ調整手法に，Fujita による AlphaDDA[1] がある．AlphaDDA は，ニューラルネットワーク・強化学習・モンテカルロ木探索（MCTS）を組み合わせた AlphaZero により得られる強いプレイヤーに対し，3 種類の動的な強さの調整を導入した AI プレイヤーである．それぞれ，MCTS のシミュレーション回数（DDA-M）・ドロップアウト確率（DDA-D）・UCT スコア（DDA-U）の調整を行うプレイヤーである．さらに本研究では，softmax 方策を用いた強さ調整手法（DDA-S）[4] を加えた 4 種類の DDA プレイヤーを使用する．

先行研究 [1] では，DDA プレイヤーと他の AI プレイヤーを先手後手入れ替えて計 100 試合させ，勝率をもとに動的な強さ調整が実現できたかの評価を行っていた．しかし，最終的な勝率が 50 %だとしても対局途中では DDA プレイヤーが優勢（または劣勢）な可能性があり，評価が不十分であると考えた．

3　AlphaDDA の再評価

リバーシ（オセロ）を対象のゲームとし，先行研究で得られた学習結果のパラメータをそのまま用いて，対局中の各手における局面評価値の変動の大きさの再評価を行う．対戦相手となるプレイヤーは，ミニマックス探索ベースの Minimax，モンテカルロ木探索ベースの MCTS1 と MCTS2（シミュレーション回数は，1 手あたり 300 回と 100 回），動的強さ調整を導入する前の AlphaZero プレイヤーである．各相手 AI プレイヤーに対し，先手後手を入れ替えて計 100 試合を行い，その棋譜を取得した．そうして得られた棋譜の各局面について，非常に強力なオープンソース AI プレイヤーである Edax（`https://github.com/abulmo/edax-reversi`）を用いて評価値を求めた．なお，Edax の評価値は，石数を表すようスケーリングされている．

MCTS1 を相手 AI プレイヤーとした場合の，対局途中の局面評価値の変動を図 1 に示す．また，勝敗の内訳，対局を通した評価値の平均絶対誤差（ただし，序盤の 8 手分は計算から除いた），30 手目の評価値の結果を表 1 に示す．なお，図 1，表 1 の結果は先行研究 [1] において勝率の調整におよそ成功していた DDA-M に対する評価結果である．

勝率に関しては，先行研究と同等の結果が得られた．しかしながら，図 1 から分かるように，ゲームの中盤までに評価値が大きく振れてしまい，その評価値の幅はゲーム終盤でもほとんど小さくなっていない．

よって，AlphaDDA は勝率の点では目標を達成できているものの，ゲーム全体に渡って適切に強さを制御できているとは言えない．

4　DDA プレイヤーの合議

合議により，複数の候補手から互角に近づく手を選択する合議プレイヤーを作成し，動的な強さの調整の改良を行う．本研究では動的強さ調整を実現するための合議の手法として，多数決合議と楽観合議を使用する．多数決合議は，最も多くのプレイヤーが支持した手を選択するものとし，楽観合議は，AlphaZero により求められた評価値が最も 0 に近づく手を選択するものとして，

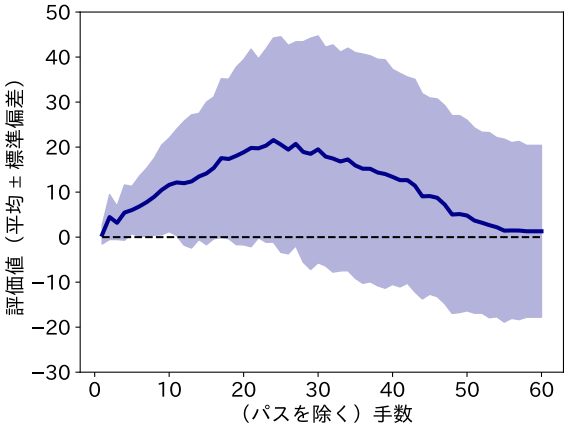

図 1　DDA-M vs MCTS1 の局面評価

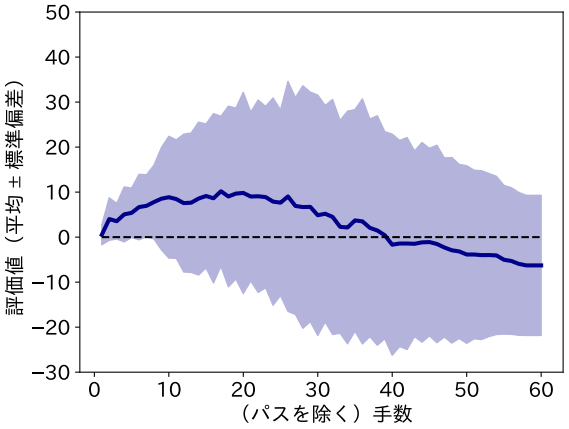

図 2　多数決合議による局面評価

表 1　DDA-M の対局結果

相手	勝敗			平均絶対誤差（平均 ±SD）	30 手目（平均 ±SD）
	勝	負	分		
Minimax	62	35	3	16.3 ± 8.3	11.5 ± 19.4
MCTS1	42	50	8	19.6 ± 10.1	19.5 ± 25.2
MCTS2	59	36	5	22.9 ± 11.7	22.2 ± 28.5
AlphaZero	58	41	1	14.4 ± 5.8	−4.8 ± 17.7

それぞれの手法における勝敗及び局面評価を行う．

また，各 DDA プレイヤーが同じ手しか選択しない場合，合議による複数の手から選ぶ部分が機能しない．そこで，DDA プレイヤーが手を選択する際の評価値に乱数を与えて各 DDA プレイヤーについて複数プレイヤーの生成を行なった．

伊藤らによる Bonanza の多数決合議実験では，プレイヤー数が増えると合議プレイヤーはより強くなる傾向があった [3]．本研究では総プレイヤー数を 16 とした場合の結果を示す．これは，[3] の実験で使用された最大のプレイヤー数である．

5　実験

MCTS1 と 4 種類の DDA プレイヤーを多数決合議・楽観合議のそれぞれの手法において，先手後手入れ替えて計 100 試合行う．

図 2 に DDA-M，DDA-D，DDA-U，DDA-S に乱数を加えて生成した計 16 プレイヤーによる多数決合議の局面評価の結果を示す．また表 2 に各合議プレイヤーと MCTS1 の対局結果を示す．

多数決合議では，勝率は下がったものの DDA-M のみの合議なしの場合と比べてピーク時の評価値の揺れは小さくすることができた．

楽観合議では，勝敗・ゲーム全体の局面評価はともに互角にはならなかった．動的強さ調整を考えない合法手全てで評価値 0 を目指す場合も互角に近づかないという結果であったため，評価値 0 に近づけるという楽観の考え方による動的強さ調整は実現困難であると考える．

表 2　各合議プレイヤーと MCTS1 の対局結果

合議	勝敗			平均絶対誤差（平均 ±SD）	30 手目（平均 ±SD）
	勝	負	分		
多数決合議	35	57	8	16.9 ± 7.2	4.9 ± 26.7
楽観合議	5	84	11	13.7 ± 5.4	−4.7 ± 18.0

6　まとめ

本研究では，既存の DDA プレイヤーの更なる強さの調節を目指し，複数の DDA プレイヤーを用いた多数決合議・楽観合議による合議プレイヤーの作成を行なった．多数決合議においては，単一の DDA プレイヤーに比べて局面評価値の変動を抑えることに成功した．一方，楽観合議では局面評価値の変動を抑えることができなかった．

参考文献

[1] K. Fujita, “AlphaDDA: strategies for adjusting the playing strength of a fully trained AlphaZero system to a suitable human training partner.” PeerJ Computer Science, 2022.

[2] 久保田 留奈, 松崎 公紀, “AlphaDDA の局面評価値を用いた再評価”, 電気・電子・情報関係学会四国支部連合大会論文集, 2023.

[3] 伊藤 毅志, 小幡 拓弥, 杉山 卓弥, 保木 邦仁, “将棋における合議アルゴリズムー多数決による手の選択”, 情報処理学会論文誌, Vol.52, No.11, pp.3030-3037, 2011.

[4] N. Sephton, P. I. Cowling, N. H. Slaven, “An Experimental Study of Action Selection Mechanisms to Create an Entertaining Opponent.”IEEE Conference on Computational Intelligence and Games (CIG), pp.122–129, 2015.