ISSN 2186-5647

7-55

モンテカルロ法とニューラルネットワークによる α β法の効率化

日大生産工（院）○平林　佑一
日大生産工　　松田　聖

1　まえがき

モンテカルロ法は乱数により近似解を求める手法である。評価関数を使用することなく、ゲームが終了するまでシミュレーションすることによって評価値が得られるため、コンピュータ囲碁のＡＩにしばしば用いられる。しかし、単体で適用した場合には探索空間が大きいほどシミュレーションを多く施行しなければ尤もらしい評価値を得ることができず、時間を多大に費やしてしまう。そこで筆者は嘗てニューラルネットワークと組み合わせることにより、ニューラルネットワークの出力（評価値）が良い手に対してモンテカルロ法を実行する手法を提案した。これにより探索効率上昇が見込めたが再考の余地が残る結果となった。

本研究ではα β法にニューラルネットワークを組み合わせることにより尤もらしい手から探索させることで枝刈り効率上昇を目的としている。また、３手先読みα β法の葉ノードの評価値をモンテカルロ法で求めるため、枝刈りによるモンテカルロ法の効率上昇も兼ねている。なお、ニューラルネットワークの学習や実際のシミュレーションをする際の例として9路盤の囲碁を用い、ゲーム中盤から始めることとする。

2　モンテカルロ法

モンテカルロ法とは乱数を用いてシミュレーションを繰り返し、近似解を求める計算手法である。本研究では囲碁の盤面評価をモンテカルロ法で行う。

囲碁に適用する場合、囲碁のルールに則った全ての着手可能な場所（ただし眼は除く）に乱数で石を黒白交互に打つ。そしてそれを終局まで何度もシミュレーションする。終局時に最終的な勝率を見て最適手を選ぶ。

合法手iのプレイアウトを行った回数をs_i、プレイアウトによって得られた報酬の和をX_iとする。報酬はプレイアウトの結果、勝利であれば１、敗北であれば0を与える。

このとき、合法手 の勝率 は

$$\overline{Xi} = \frac{X_i}{s_i} \qquad (1)$$

によって表わされる。この値は合法手 を打った直後の盤面の評価値とする。なお、この評価値は後述するニューラルネットワークの教師信号として利用する。

Fig1にモンテカルロ法の構造を示す。

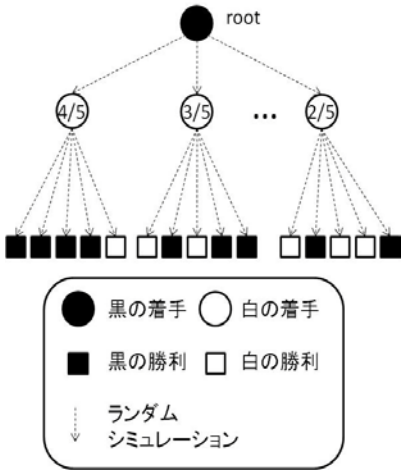

Fig1.モンテカルロの構成

このモンテカルロ法ではプレイアウトによって得られた結果のみを用いるため、プレイアウトの回数に依存して評価の精度が変わる。囲碁では一回の探索での時間が数秒～数分しかないため、評価値の精度を高くするにはプレイアウトによる合法手の選択を完全ランダムではなく、良さそうな手を重点的に探索する必要がある。そのため、本研究では現在の状態から次の手を選択する際にα β法とニューラルネットワークを用いて学習することで不要なプレイアウトを省くことを目的としている。

3　α β法

ミニマックス法は得られた評価値から相手の手番には最小値、自分の手番には最大値を取るように木を逆に辿ることで手を選ぶ手法である。このミニマックス法の探索時間を省略した方法をα β法という。

Efficient α β Method by Monte Carlo Method and Neural Network

Yuichi HIRABAYASHI and Satoshi MATSUDA

本研究ではこのαβ法を３手先読みで使用し、３手目の局面（葉ノード）の評価値をモンテカルロ法で算出する。
Fig2にαβ法のモデルを示す。

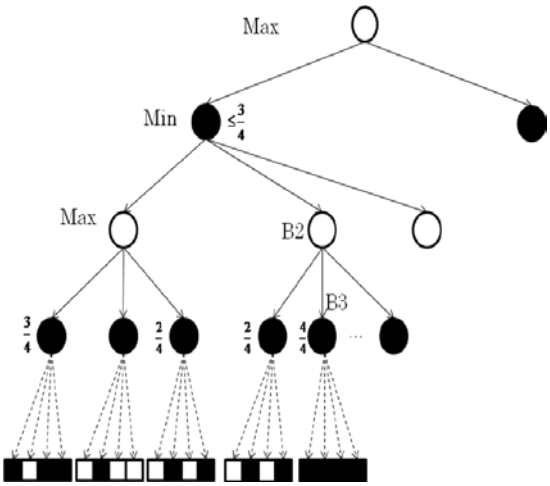

Fig2.αβ法

αβ法では探索効率化のためにβカットを用いる。Fig2におけるβ値は3/4となり、B2の評価値として早めに3/4以上のノードを探索できれば大きなβカットを行うことができる。そのため、B2の子ノードの評価値を予め予測し、評価値の高いものから探索をすることができれば大幅なカットが見込める。そこで、ニューラルネットワークを用いて学習した結果を元にノードの並び替えを行う。

４　ニューラルネットワーク

ニューラルネットワークとは人間の脳の構造を計算上のシミュレーションで表した数理モデルである。ニューロン間のシナプス結合荷重を修正することで出力値を正確な値になるように学習していく。
入力として石の位置、すなわち盤面の状態を入れる。入力層のニューロン数は9路盤のため、2×9×9の162個と閾値の1個を足し、163個とする。出力はその盤面の状態の評価値とし、出力層のニューロン数は1個とする。学習中における教師信号はモンテカルロ法によって得られた評価値を用いる。これにより、ニューラルネットワークの結合荷重を修正し、囲碁を学習していく。

５　モンテカルロ法とニューラルネット

モンテカルロ法とニューラルネットワークを組み合わせると、モンテカルロシミュレーションによって生成された末端局面から元の局面の勝率を評価し、その値を教師信号としてニューラルネットワークの結合荷重を修正することで囲碁を学習することができる。
αβ法にて子ノードを探索する際には子ノードの盤面１つ１つを学習したニューラルネットワークの入力層に入力する。その後出力として子ノードの評価値を得ることができる。通常のαβ法とは違い、子ノードの評価値が得られたことにより評価の良いものから順に探索をすることができる。そのため、早い段階でβ値以下の値、α値以上の値となる子ノードに対してモンテカルロ法を施行することができ、枝刈り数の増加及び木構造全体のモンテカルロ法試行回数減少が見込める。すなわち、αβ法の枝刈りすべきノードが遅い段階で得られる問題とモンテカルロ法の無駄な候補手に対するモンテカルロシミュレーションによる時間浪費の問題を解消することができる。また、ニューラルネットワークの精度の悪さによる誤判断はモンテカルロ法施行回数増加による探索時間減少の隔たりとなるが、木の解の質を低下させることはない。
Fig3にモンテカルロ法とニューラルネットワークにより効率化したαβモンテカルロ木を示す。

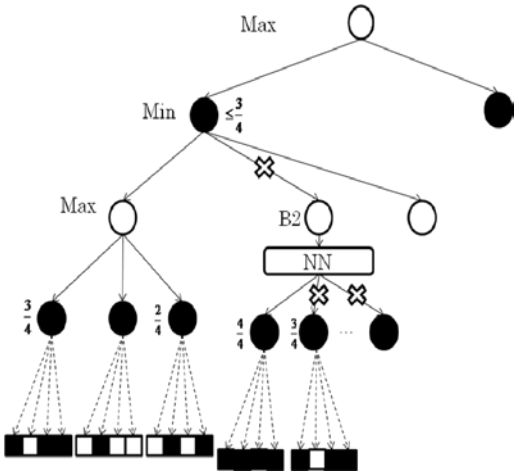

Fig3.**αβ**モンテカルロ木

本研究のモンテカルロ木は３手先読みのαβ法で行う。ルートノードから探索を行い、子ノードを探索する際には子の盤面をニューラルネットワークに入力として渡し、それぞれの盤面を評価する。その中からMinの際には評価値の低い順に、Maxの際には高い順にノードを探索する。
深さが３（葉）に到達した際にはモンテカルロ法を用いて評価値を算出する。
ニューラルネットワークの学習はルートノードにて子ノードの中から最も高い評価値を選択し、次の手として後に行う。選んだ盤面をニューラルネットワークの入力として渡し、そのときの教師信号を先ほど選んだ評価値とする。これにより結合荷重を修正していく。

参考文献

[1]R.Coulom. Computing elo ratings of move patterns in the game of Go. In Computer Games Workshop, 2007.
[2]平林　佑一、モンテカルロ法とニューラルネットワークによる学習するコンピュータ囲碁、日本大学生産工学部第42回学術講演会数理情報部会開講演概要(2009)、pp129-130