7-41

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

日大生産工（学部）　○平林　佑一
日大生産工　松田　聖

1 はじめに

近年になり、コンピュータ囲碁のプログラムを作成する際に人間の考えを模倣した手法からモンテカルロ木探索に移り変わってきた。それにより困難であった評価関数を作成する必要がなくなり、コンピュータが人間に勝利する日は格段に近くなったといわれている。

現在、そのモンテカルロ木探索を用いたコンピュータ囲碁をより効率化させるために探索部分を強化しようとする動きが見られる。しかし､それは人間の編み出した囲碁知識(定石やパターン比較）を用いており、コンピュータは学習をしていない。

本研究ではモンテカルロ木探索の元となっているモンテカルロ法に学習機能を持たせることを目的とし、その手法としてニューラルネットワークを用いる。これによりコンピュータ自らが囲碁を学習するシステムを獲得する。なお、学習対象は９路盤の囲碁の終盤からとし、中盤をあらかじめ入力してから始めるとする。

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（ただし目は除く）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iの勝率$\overline{X_i}$は

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

によって表わされる。この値は合法手i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。

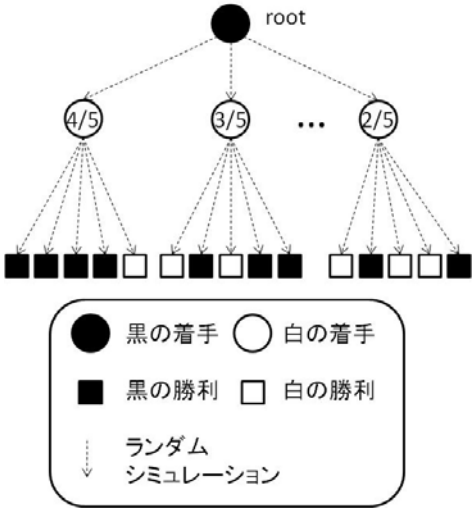

Fig1.モンテカルロの構成

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。

Computer Go Learning with Monte Carlo Method and Neural Network

Yuichi HIRABAYASHI and Satoshi MATSUDA

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　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。

本研究のニューラルネットワークの構造をFig2に示す。

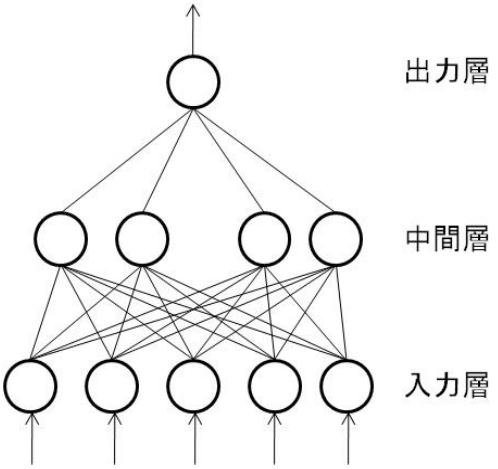

Fig2.ニューラルネットワークの構成

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。

4 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し、囲碁を学習することができる。すると、ルートノードから次の手を選ぶ際に学習した評価値の中から高いものを選択しやすくなり、良いと思われる候補手に重点的にモンテカルロシミュレーションをすることができる。こうすることで打つ可能性の低い手を省き、モンテカルロ法の欠点を解消することができる。

Fig3にモンテカルロ法とニューラルネットを組み合わせたモンテカルロ木を示す。

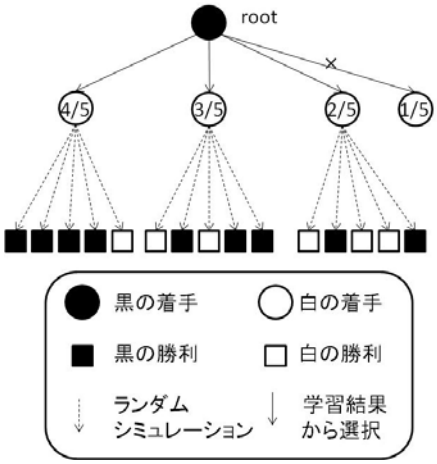

Fig3.モンテカルロ木

本研究のモンテカルロ木はニューラルネットワークが良い手と思われる手を選択する際に次に打つ全ての盤面を入力として渡し、それぞれの盤面を評価する。その中から
ε-greedy法を用いて3手を選ぶ。

その3手についてモンテカルロシミュレーションを行い、末端局面からそれぞれの元の手の勝率を計算し、それを新しい評価値とする。新しい評価値の中から最も良い評価値を選択し、それを次の手とする。そして、その盤面をニューラルネットワークに入力として渡す。そのときの教師信号を先ほど更新した評価値とし、結合荷重を修正してくことで囲碁を学習させていく。

実戦時には自分の手番の際にこのモンテカルロ木の学習したニューラルネットワークによって選ばれた3手についてモンテカルロシミュレーションを行い、その中から最善手を選ぶ。すると、モンテカルロ法単体よりも効率的に次の一手を選択することができる。

5　実験方法

実験は結合荷重を修正して学習させるものと強さを測る実験の２つを行う。前者は学習するコンピュータと学習をしないコンピュータを対局させ、結合荷重を修正させていく。後者は学習したコンピュータと学習する前のコンピュータを対局させ、勝率で強さを判断する。

「参考文献」

1) R. Coukom. Computing elo ratings of move patterns in the game of Go. In Computer Games Workshop, 2007.
2) 三上　貞芳・皆川　雅章,　強化学習, 森北出版, (2000)