ＴＤ学習とニューラルネットワークを融合した
オセロの盤面評価関数の獲得

日大生産工（学部）　○廣田　稔　　松田　聖

１．はじめに

近年、コンピュータボードゲームには様々な人工知能が組み込まれている。しかし多くの場合、人間が定石や好手などの情報を入力しておくことが必要となっている。本研究では、ＴＤ学習とニューラルネットワークを用いることで、コンピュータ自らがその様な情報を学習することを目指している。

よって、コンピュータが盤面の適切な評価をすることが出来るようになり、ある程度の強さに達することが出来るか確認することが、本研究の目的である。

２．ＴＤ学習

ＴＤ学習とは、現在の状態の評価値と行動後の状態の評価値を比べ、現在の評価値を更新していく学習方法である。行動後に現在より良い状態になるならば、現在の状態の評価値を上方修正する。逆の場合では評価値を下方修正する。またＴＤ学習では、学習の１サイクルが終了したとき報酬が支払われる。本研究では、オセロゲームが終了した時を１サイクルの終了としている。勝利した場合にはプラスの報酬が、敗北した場合にはマイナスの報酬が支払われることになる。ＴＤ学習は以下の(1)式で表される。

$$y^t \leftarrow y^t + \alpha(r + \gamma \cdot y^{t+1} - y^t) \qquad (1)$$

ここでy^tは現在の状態の評価値、y^{t+1}は行動後の状態の評価値、rは報酬を表している。αは学習率、γは割引率といわれる0から1の間の定数である。本研究では、オセロの盤面の善し悪しを評価値とする。

ＴＤ学習は、教師なし学習といわれる学習方法の一つである。教師なし学習とは、こちらから正解などの情報を与えて学習をするのではなく、状況が変化した際などに自立して学習を行う方法である。その為、本研究の学習に適しているといえる。ＴＤ学習をオセロゲームに用いる場合、盤面の一つ一つを評価し、その評価結果を記憶しておく必要がある。だが、全ての盤面を学習によって評価することは難しく、またコンピュータが記憶する状態の数も多くなり過ぎてしまう。この問題点を解決する為、ニューラルネットワークを用いる。

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、脳の構造を模した数学モデルの一つである。本研究では、オセロゲームの盤面の状態を入力とし、その状態の評価値を出力とする。図１は、本研究で用いるニューラルネットワークを表している。

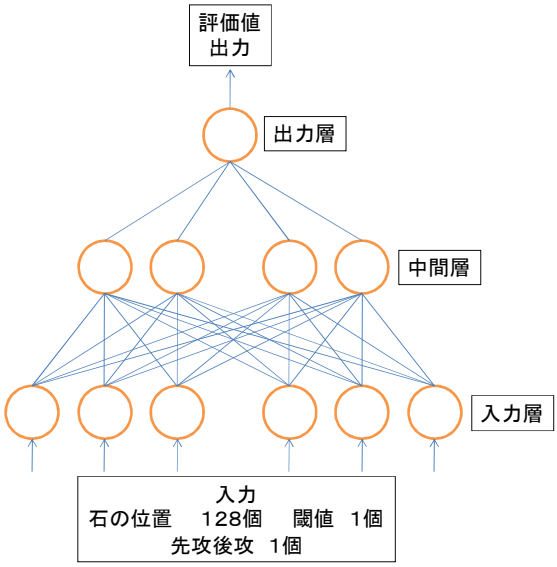

図１．ニューラルネットワーク

Acquisition of an Evaluation Function of Reversi with TD-Learning and Neural-Network

Minoru HIROTA & Satoshi MATSUDA

通常、ニューラルネットワークは教師信号といわれる出力ニューロンの正解の値と、実際の出力値を比較することで状態を評価する。そしてその評価結果を元に、各層のニューロン間のシナプス結合荷重を変更し学習をしていく。しかし、教師信号を与えることは本研究の目的に反する。そのため、ニューラルネットワークとＴＤ学習を融合されることで新しい評価方法を作成する。ニューラルネットワークの評価関数は以下の(2)式で、シナプス結合荷重の変更式は(3)式で表される。

$$E = \frac{1}{2}(t-o)^2 \tag{2}$$

$$\Delta w = -\alpha \frac{\partial E}{\partial w} \tag{3}$$

ここでtは教師信号、oは出力値、wはシナプス結合荷重、αは学習率といわれる0から1の定数を表している。

４．ＴＤ学習とニューラルネットワーク

双方の問題点を解決させる為に、ＴＤ学習とニューラルネットワークを融合させる。

ＴＤ学習においての、評価する状態の数が多すぎるという問題点は、一つ一つの盤面を評価し記憶するのではなく、ニューラルネットワークを用いて盤面の評価値を近似的に算出するシステムを構築する。このシステムによって、どのような盤面でも評価値を得ることができるようになる。また、盤面が変わるごとに盤面の状態から評価値を計算するため、一つ一つの盤面の評価値を記憶しておく必要がなくなる。

ニューラルネットワークにおける評価関数の問題点は、教師信号と出力値を使って行う評価を、ＴＤ学習における現在の評価値と行動後の評価値によって評価することで解消される。(1)式と(2)式を融合し作成した、新しい評価関数は以下の(4)式のように表される。

$$E = \frac{1}{2}(r + y^{t+1} - y^t)^2 \tag{4}$$

rは報酬、y^{t+1}は行動後の状態の評価値、y^tは現在の状態の評価値を表している。

この評価関数を用いることによって、教師信号を用いずに、各ニューロン間のシナプス結合荷重を更新することが出来る。本研究では、出力層と中間層、中間層と入力層の間の各ニューロンのシナプス結合荷重を変更する。その変更式は(4)式を用いて、以下の(5)、(6)式のように表される。

$$\Delta w = -\alpha \frac{\partial E}{\partial w} \tag{5}$$

$$\Delta v = -\beta \frac{\partial E}{\partial v} \tag{6}$$

wは出力層と中間層の間、vは中間層と入力層の間のニューロン間のシナプス結合荷重を表している。α、βはそれぞれ学習率といわれる0から1の間の定数である

５．実験方法

現在プログラムを作成中の為、細部を変更する可能性があるが、大まかに実験方法について説明する。

本研究における実験は、対戦を重ねシナプス結合荷重の値を変更していく学習の実験と、どの程度の強さになったのかを判断するための勝率の実験の二つに分けられる。

学習の実験では、全く同じ性能のコンピュータ同士を対戦させ、一方のコンピュータのシナプス結合荷重を変更していく。次の実験では、学習した方のコンピュータを使い同じことを繰り返す。学習がうまくいけばコンピュータは、少しずつだが着実に強くなっていくと考えられる。

勝率の実験では、評価値が固定で且つある程度の強さのコンピュータと戦わせ、その勝率の推移をみることで、コンピュータの学習が、どのように進んでいるかを判断することが目的である。

参考文献

1) 吉冨　康成,“ニューラルネットワーク”, 朝倉書店,(2002),p27-30
2) 三上　貞芳,皆川　雅章,“強化学習”,森北出版,(2000),p209-217
3) 高玉　圭樹,“マルチエージェント学習”, コロナ社,(2003),p21-44
4) 大内　東,山本　雅人,川村　秀憲,“マルチエージェントシステムの基礎と応用”, コロナ社,(2002),p70-90
5) Gerald Tesaur, “Temporal Difference Learning and TD-Gammon”,(1995)