ＴＤニューラルネットワークを用いたオセロの性能評価

日大生産工（院）　○廣田　稔
日大生産工　　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、それら人工知能が考えるための知識、つまり定石や好手などの情報は製作者が事前にコンピュータへ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という二つの人工知能の学習方法を融合した、ＴＤニューラルネットワークを用いることで、コンピュータが自らその様な情報を学習することを確認し、コンピュータが自己学習するシステムの一端を示すことを目的としている。

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のならば、現在の状態の評価値を上方修正する。また、現在より悪い状態になる場合では評価値を下方修正する。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、製作者が正解などの情報を与えて学習をさせるのではなく、状況が変化した際などにコンピュータが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学習で各状態の評価値を得るには、１サイクルが終了した時に支払われる報酬を元に、そのサイクルで通過した状態を逆に辿って計算していくことになる。したがって、全ての状態の評価値を得るためには、何サイクルもシミュレーションをし、全ての状態を経験する必要がある。ＴＤ学習をオセロゲームに用いる場合、盤面の状態数は膨大になるため、全ての盤面を経験し、その評価値を記憶しておくことは簡単ではない。この欠点を補填する為、ニューラルネットワークという学習方法に注目する。

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。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。

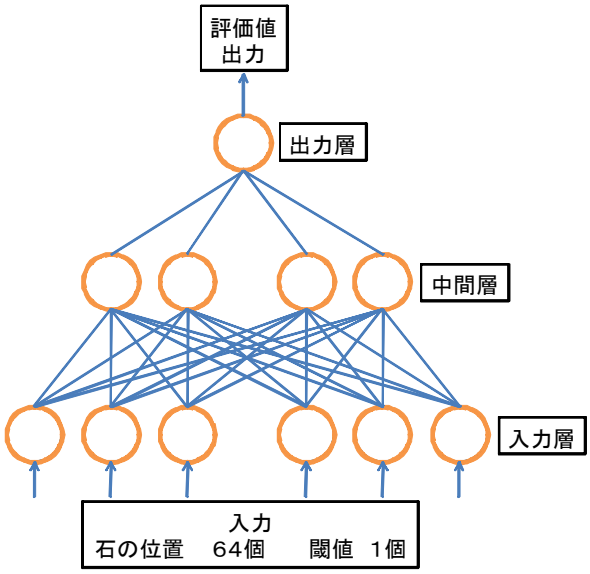

図１．ニューラルネットワーク

Performance Evaluation of Reversi with TD-Neural-Network

Minoru HIROTA & Satoshi MATSUDA

ニューラルネットワークは文字認識などに用いられており、入力の値が似ていると出力値が同じになるという特性がある。この特性を用い、本研究では入力をオセロゲームの盤面、出力をその盤面の評価値とし、似た盤面の評価値が同じになるようにしている。また、これにより経験したことのない盤面に対しても、近似的な評価値を与えることができるようになる。

ニューラルネットワークはＴＤ学習の欠点を補填しているが、これのみを用いて学習するには問題がある。ニューラルネットワークの学習は、各層のニューロン間のシナプス結合荷重を変更することで行われる。この変更には、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し評価することが必要となる。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。

(2)、(3)式を見てわかるように、ニューラルネットワークは教師信号を与えなければ学習することは出来ない。しかし教師信号を与えることは、コンピュータに製作者の知識を与えることになるので、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学習の双方の利点をうまく利用できるようこの二つの学習方法を融合する。

４．ＴＤニューラルネットワーク

ＴＤ学習における問題、評価する状態の数が多すぎるという点に対しては、根本的に一つ一つの盤面を評価し記憶することをやめてしまう。そのかわりに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とで解決する。このシステムによって、盤面の状態さえ分かれば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、盤面の評価値を記憶しておく必要がなくなるのである。

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決する。ＴＤ学習の(1)式とニューラルネットワークの(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で、教師信号をつかわずにニューラルネットワークの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

コンピュータ上でオセロプログラムを戦わせ、その結果に対し考察する。

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)