プロの棋譜を用いたTD法による将棋の評価関数の学習

薄井克俊，鈴木豪，小谷善行
（東京農工大学）

概要

本稿では、将棋の駒の価値を TD 法を用いて調整する方法について述べる。既に自己対戦による将棋の駒価値の学習は行われている（D.F.Beal and M.C.Smith, 1998）が、ここではプロの対局の棋譜と TD 法を組み合わせた駒価値の学習について述べる。

実験では、東京農工大学小谷研究室で開発した将棋システムに TD 法による学習ルーチンを組み込んだものを学習プログラムとして用いた。コンピュータにプロの棋譜をなぞらせながら局面の評価をさせ、学習を行った。評価関数には線型一次関数を用い、探索は $\alpha\beta$ 法による全幅探索を行なった。実験に使用する棋譜は主に将棋年鑑から無作為に選び、一回の試行につき 3500 対局分の棋譜の学習を行った。

Learning Values of Shogi Pieces from Expert-Level Games using Temporal Differences

Katsutoshi USUI, Tsuyoshi SUZUKI, Yoshiyuki KOTANI
[ukkun, go]@fairy.ei.tuat.ac.jp　kotani@cc.tuat.ac.jp
Tokyo Univ. of Agri. and Tech., 2-24-16 Nakamachi, Koganei, Tokyo, JAPAN

Abstract

This paper describes a technique for learning values of shogi pieces using temporal differences. In previous research, temporal learning has been used to learn values of shogi pieces from self-play (D.F.Beal and M.C.Smith, 1998). In the approach described here, we use games played by expert players to learn the values of pieces.

For the experiment, we developed a learning system that uses iterative-deepening with alpha-beta pruning for searching the game tree. The moves for each game were based on actual games between expert-level players. In the experiment, data from 3500 expert-level games was used for learning the values of shogi pieces. The result shows that the learning is effective.

1．はじめに

将棋の王を除いた 13 種類の駒の価値を、Temporal Difference Learning （TD 法）を用いて学習することについて述べる。TD 法は 1959 年に Samuel によって導入され、1988 年に Sutton が拡張・形式化を行なった。以後、バックギャモン、チェス、将棋などで TD 法による学習が行なわれている。すでに自己対戦で将棋の駒価値を求める実験が行なわれている[1]が、我々は自己対戦ではなくプロの棋譜を用いて学習を行うことを試みた。

局面を評価する静的評価関数はコンピュータ将棋を強くするために重要な要素の一つであるが、評価関数のパラメータの調整は人手による部分が大きい。評価関数の重みを自動的に最適化する研究は古くから行なわれており、TD 法もその一つである。

実験では、東京農工大学小谷研究室で開発した将棋システムに TD 法による学習ルーチンを組み込んだものを学習プログラムとして用いた。コンピュータにプロの棋譜をなぞらせながら局面の評価をさせ、学習を行った。評価関数には線型一次関数を用い、探索は α β 法による全幅探索を行なった。実験に使用する棋譜は主に将棋年鑑から無作為に選び、一回の試行につき 1500 対局分の棋譜の学習を行った。

2．プロの棋譜を使った TD 法による学習の概要

2．1 TD 法の概要

TD 法は 1959 年に Samuel によって導入され、1988 年に Sutton が拡張・定式化を行なった。TD 法は近い未来の予言を利用して学習を行う手法である。TD 法では一つの対局中にインクリメンタルに学習を行うため、計算コストは小さい。パラメータを更新するためには従来の手法のように対局の終了を待つ必要がなく、また過去の経験を効果的に利用できるため学習時間を短縮できるという利点がある。さらに、任意の段階でパラメータを調整できるので、将棋やチェスにおける TD 法の学習は、一手一手を細かく学習できるという点で、一回の学習に一対局必要な学習アルゴリズムと比べて学習に向いていると考えられている。

2．2 TD 法の学習式

実験では次の式によって駒価値を調整した。W は評価関数の重みのベクトル、P は予想確率、α は学習率、λ は予想確率に対する重み、である。

$$\Delta W_t = \alpha(P_{t+1} - P_t)\sum_{i=1}^{t} \lambda^{t-i} \nabla_W P_i \qquad (1)$$

ここで、$\nabla_W P_i$ は P を W で偏微分した勾配ベクトルであり、式(2)のように表される。

$$\nabla_W P_i = \left(\frac{\partial}{\partial w_1} P_i, \frac{\partial}{\partial w_2} P_i, \ldots, \frac{\partial}{\partial w_n} P_i \right) \qquad (2)$$

実験では、局面の評価値を最終的な結果を予測する値であるとみなすことにする。探索によって返された評価値を標準的なシグモイド関数によって勝つ確率の予想確率に変換する。式(2)において、局面によって与えられる予想確率 P を式(3)によって与える。

$$P(E(K)) = \frac{1}{1+e^{-E(K)}} \quad (3)$$

2．3 プロの棋譜を使った学習

今回述べる実験では、学習するときの指し手はプロの棋譜によって決定し、探索はあくまでも評価関数を調整するために行う。TD 法による評価関数の重みの調整は式(1)によって行うが、探索はプロの棋譜によって得られた局面の予想確率 P を求めるために用いる。

ここで問題になるのが、プロとコンピュータとの技量の差である。プロの指し手は深い先読みのもとに決定されることが多いが、現在のコンピュータにとってはプロのように深く読むのは難しい。よって、プロの指し手の意味をコンピュータが理解するのは難しく、評価関数をおかしな方向に修正する（つまり「誤解」する）こともある。これが、単純にプロの指し手とコンピュータの指し手を比較して評価関数を調整する場合の問題点であった。しかし、TD 法では近い未来の予測を用いて学習するため、「誤解」の程度を軽減できるのではないかと考えた（「誤解」しないわけではない）。

局面の多様性については、局面を網羅的に学習することになる自己学習に対して、プロの指し手だけを選択して学習させる以上、ある程度学習結果が偏ることは考えられる。しかし、ある程度無駄な局面が出る自己学習よりも、プロの棋譜を用いた方が少ない学習回数で収束するのではないかと考えた。

3． プロの棋譜を用いた駒価値の学習実験

プロの棋譜と TD 法を組み合わせて、将棋の王を除く 13 種類の駒価値を学習する実験を行った。

実験で使用したシステムは α β 法による探索をおこなう。評価関数は駒価値だけの重み付き線型和である。駒の損得以外の知識（定跡など）は用いなかった。指し手の決定はプロの棋譜によって行われ、探索と評価は式(1)の P を決定するために行った。探索の深さは 3 とした。

実験は、将棋年鑑から 3500 対局分のプロの棋譜を用いて行った。重みの初期値は 1000 とした。重みの更新は式(1)によって行い、λ=0.95、α は Temporal Coherence[5]によって駒ごとに自動的に変動するようにした。また、重みの更新は一つの対局が終了するごとに行った。

4．実験結果

4．1学習の様子

駒価値の学習の様子を図 1 と図 2 に示す。α は 500 から 10 の間で変動させた。図 1 は成っていない駒の価値の学習の様子である。各駒ともに、学習の最後まで値が大きく振動しているが、1500 対局を超えたあたりからは一定の範囲内で振動していることがわかる。

図 2 は成駒の価値の学習の様子である。成駒は成っていない駒に比べると変動が小さく、学習が収束していないことがわかる。この現象は自己対戦による学習でも起きており、対局中に成駒がでてくる頻度が低いことが原因であると考えられる。

4．2得られた駒価値

図 3 に得られた駒価値を示す。図 3 の縦軸は学習の最後の 1000 対局分の駒価値の平均をとり、飛車の価値を 5 とした場合の値である。成っていない駒の価値は一般的な駒の価値の順番（歩、香、桂、銀、

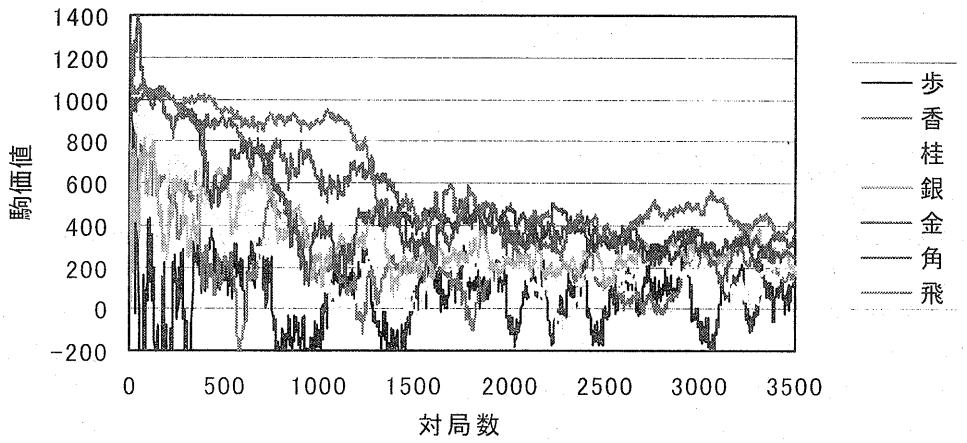

図1：駒価値の学習の様子

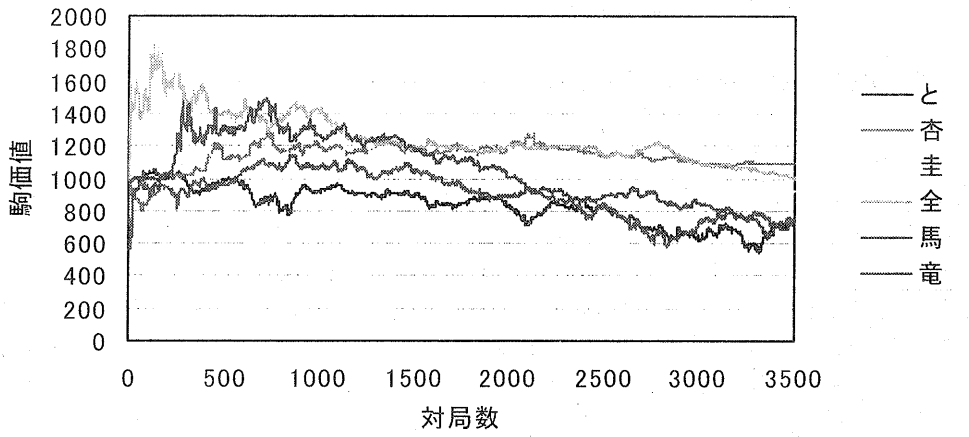

図2：駒価値の学習の様子（成り駒）

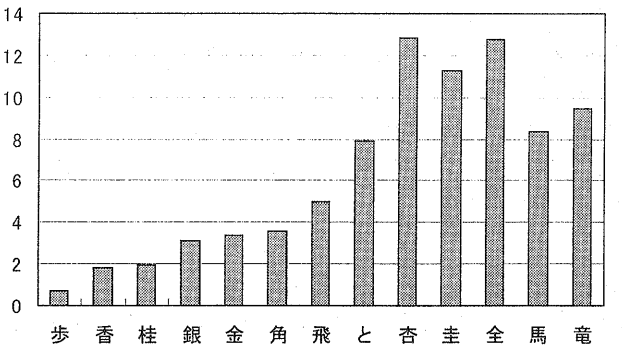

図3：得られた駒価値

金、角、飛車）になっているが、成駒についてはと金、成香、成桂、成銀が非常に大きな値になってしまった。

5．おわりに

本稿では、プロの棋譜と TD 法を組み合わせて将棋の駒価値を学習することについて述べた。成っていない駒の価値については一般的な値に近い駒価値を得られたが、成駒についてはあまり良い結果は得られなかった。成駒の学習が遅かった理由の一つには成駒が盤面にあらわれる頻度が低い（自己対戦の場合もプロの棋譜の場合も）ことが考えられるが、プロの棋譜の方が着実な読みに基づいて指しているはずなので、それを上手く読み取れれば自己対戦よりも良い結果が得られると思う。

また、駒価値だけの評価関数ではなくもっと複雑な（駒の働きや王の安全度などを考慮する）評価関数を学習することを考えた場合も、着実な発想に基づいているぶんプロの棋譜を用いた学習の方が効率がよいと思う。今後の課題として、駒価値以外の評価要素の学習にも取り組んでいきたい。

参考文献

[1] D.F.Beal and M.C.Smith : First Results from Using Temporal Difference Learning in Shogi, Computers and Games, pp.113-125, 1998

[2] J.Baxter, A.Tridgell, L.Weaver : Experiments in Parameter Learning using Temporal Differences, ICCA Journal Vol.21 No.2 pp.84-99, 1998

[3] D.F.Beal and M.C.Smith : Learning Piece Values Using Temporal Differences, ICCA Journal Vol.20 No.3, pp.147-151, 1997

[4] D.F.Beal and M.C.Smith : Learning Piece-Square Values using Temporal Differences, ICCA Journal 1999 December, pp.223-235, 1999

[5] D.F.Beal and M.C.Smith : Temporal Coherence and Prediction Decay in TD Learning, IJCAI99, pp.564-569, 1999

[6] G.Tesauro : TD-Gammon, a Self-Teaching Backgammon Program, achieves Master-Level Play, Neural Computation Vol.6 No.2, pp.215-219, 1994

[7] R. S.Sutton : Learning to Predict by the Methods of Temporal Differences, Machine Learning 3, pp.9-44, 1988

[8] 薄井克俊,鈴木豪,小谷善行：TD 法を用いた将棋の評価関数の学習，ゲームプログラミング・ワークショップ'99, pp.31-38, 1999

[9] 薄井克俊，鈴木豪，野瀬隆，乾伸雄，小谷善行：将棋における cost function を用いた評価関数の調整，第58回全国大会公演論文集 Vol.2, pp.237-238, 1999