TD 法を用いた将棋の評価関数の学習

薄井克俊, 鈴木 豪, 小谷善行
[ukkun, go]@fairy.ei.tuat.ac.jp　kotani@cc.tuat.ac.jp
東京農工大学

概要

本稿ではコンピュータ将棋の静的評価関数を、TD 法を用いて学習することについて述べる。すでに TD 法で将棋の駒の価値を求める実験が行なわれている (D.F.Beal and M.C.Smith, 1998) が、我々は駒の価値に持ち駒の価値・王の安全度を加えて学習実験を行った。

本実験では、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側プログラムとして用いた。評価関数は線型一次関数で、探索は $\alpha\beta$ 法による全幅探索を行ない、対戦相手の評価関数のパラメータは一般的な値に設定した。実験は一回につき 6000 局以上の対局を行なった。

Parameter Learning Using Temporal Differences in Shogi

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 learning evaluation function using Temporal Difference learning in shogi. We examined the learning the value of shogi pieces and some evaluation function features : pieces in hand and King safety.

We give our shogi program including liner evaluation function and min-max search algorithm the learning examination. The learning is obtained from randomised self-play. Opponent's evaluation function uses general parameter values. The program played over 6000 games for a trial.

1 はじめに

コンピュータ将棋の静的評価関数を、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 TD 法で将棋の駒の価値を求める実験が行なわれている[1]が、我々は駒の価値に加えてさらに多くのパラメータを学習することを試みた。

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。

学習実験では、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側プログラムとして用いた。評価関数は線型一次関数、探索は $\alpha\ \beta$ 法による全幅探索を行なった。学習の初期段階では各パラメータ間で値の差がないために評価値が同じ局面が多くでてくるので、いろいろな手を学習させるためにパラメータの初期値と比べてごく小さい乱数を評価値に加えた。

2 TD 法による評価関数の学習

2.1 TD 法の概要

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、計算時間が短縮できる。また、パラメータを更新するために対局の終了を待つ必要がなく、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

今回の実験では次の式によって評価値を調整する。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。

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

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。

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

今回の実験では、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。

$$P(E(K)) = \frac{1}{1 + e^{\frac{-E(K)}{1000}}} \tag{3}$$

$E(K)$は局面Kの評価値である。$E(K)$は式(4)のように表される。

$$E(K) = \sum_{j=1}^{N} w_j x_j(K) \tag{4}$$

j は学習する評価関数の評価要素の数、x は評価要素の特徴量である。特徴量の計算は、例えば駒価値の場合、(味方側の駒 j の枚数) − (敵側の駒 j の枚数) というように行なう。

シグモイド関数は、導関数が簡単になるという利点がある。

$$\frac{dP}{dE} = \frac{1}{1000} P(1-P) \tag{5}$$

(4)を使うと式(5)は次のように表される。

$$\frac{\partial}{\partial w_j} P_i = \frac{1}{1000} x_j P_i (1 - P_i) \tag{6}$$

図 1 は局面の評価値と予想値の関係をあらわしている。例えば、歩 一枚の価値を 100 としたとき、歩 2 枚分有利な場合の予想確率は 0.55 となる。

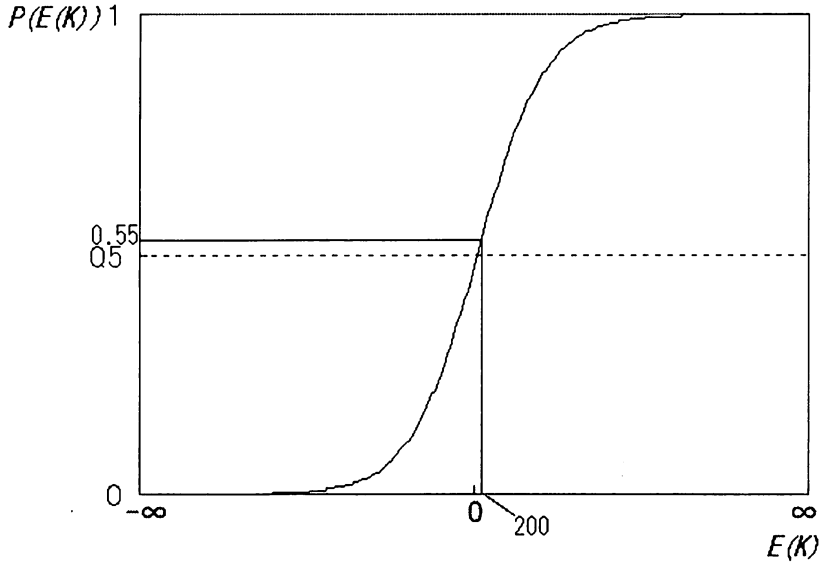

図 1 評価値から予想確率への変換

3 TD法による評価関数の学習実験

3.1 実験に使用したシステム

実験に用いたコンピュータ将棋システムについて述べる。実験には東京農工大学小谷研究室で開発した将棋システムを用いた。ゲーム木の探索では、$\alpha\beta$法と反復深化を用いた全幅探索を行なっている。評価関数は式(4)に示したように、特徴量に重みを掛けた線型和である。

評価関数の重みと特徴量を表1に示す。盤上の駒の価値と持ち駒の価値については、駒種ごとに、(敵の枚数－味方の枚数) に重みを掛けたものを加点する。王の周りの利きについては、王の周囲8近傍について、王以外に駒の利きが1つでもあるマスの数に、重みを掛けたものを加点することとする。重みは味方王と敵王の周りについて別々に2種類用意する。王と駒の距離については、王と駒との距離が近い場合 (3マス以内) に、その駒の価値に重みの0.1%を掛けたものに、さらに (4－王との距離) を掛けたものを加点することとする。重みは味方王と敵王との距離について別々に2種類用意する。

表1 評価関数の重みと特徴量

評価要素	学習する重み	特徴量
盤上の駒の価値	各駒の価値	駒の枚数
持ち駒の価値	各駒の価値	駒の枚数
王の周りの利き	王の周りのマスに利きがある場合の価値	利きがあるマスの数
王と駒の距離	王と駒の距離が近い場合の価値 (駒の価値に対する割合)	王と駒の距離 (マス)

3.2 実験方法

3.1節で述べた将棋システムに、2節で述べたようなTD法による学習システムを組み込み実験を行った。探索の深さは3とし、定跡などの特別な知識は用いていない。また、評価値にごく小さい乱数を加えて、同じ手順の対局を繰り返さないようにした。

学習する評価関数の重みの初期値はすべて1000とした。重み調整のための学習式は式(2)を用い、λ=0.95とした。さらに、αは対局数をこなすにつれて少しずつ減少するようにした。また、重みの調整は自分の手番がくるごとに行なっている。

実験は評価値を学習するプログラムを先手、対戦相手を後手とした。対戦相手も3.1節で述べた方法で指し手を選択することとし、対戦相手が用いる評価要素の重みは表2のようにした。

表2 対戦相手の重み

王	飛	角	金	銀	桂	香	歩	持飛	持角	持金	持銀	持桂	持香	持歩
100000	1000	900	700	600	400	300	100	1100	990	770	660	440	330	110
	竜	馬		全	圭	杏	と							
	1200	1100		700	700	700	700							

味方王の周りの利き	味方王との距離
25	100
敵王の周りの利き	敵王との距離
50	100

4 実験結果と考察

4.1 盤上の駒の価値

盤上の駒の価値の学習の様子を図 2 と図 3 に示す。実験は 6000 対局の学習を 1 回、10000 対局の学習を 2 回行った。予備実験の結果から、学習率 α は 250 から 1 まで、対局が進むにつれて徐々に減少するようにした。

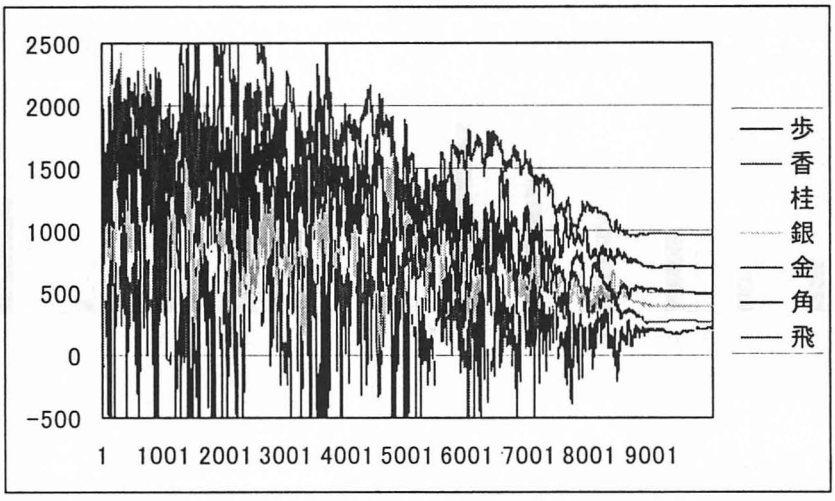

図 2　盤上の駒価値の学習の様子

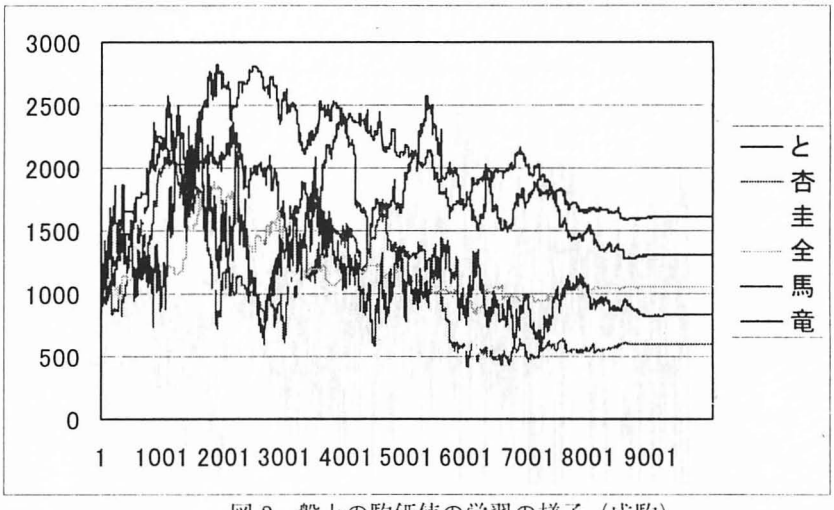

図 3　盤上の駒価値の学習の様子（成駒）

図 2 と図 3 のグラフは 10000 回対局させた実験の一つである。グラフでは一見値が収束しているようにも見えるが、3 回の実験で結果が大きく異なる駒もあった。

図 4 と図 5 のグラフに 3 回の実験の結果、飛車の価値を 5 とした場合の、他の駒との価値の比を示す。

どの実験でも、成っていない駒の価値についてはおおむね学習できている。いずれの場合も「桂・歩香・銀・金・角・飛」の順に学習されている。歩の価値が飛車の5分の1と、高めの値になった。おそらく、と金の価値が非常に高くなっていることが影響していると考えられる。成り駒については、図3を見てもわかるように学習の頻度が低く、現実的ではない結果になった。図5ではいずれの学習でもと金の価値が非常に高くなっていることがわかる。3手読みではと金を作るリスクまでは学習できなかったと考えている。

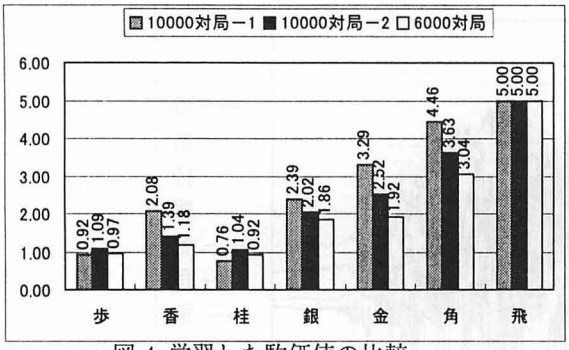

図4 学習した駒価値の比較

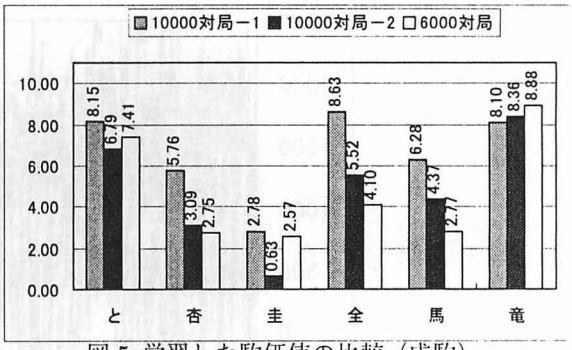

図5 学習した駒価値の比較（成駒）

4.2 持ち駒の価値

図6に持ち駒の価値の学習の様子を示す。実験では盤上の駒の価値を表2の値に固定し、持ち駒の価値の学習だけを行った。対局は6000回行い、学習率αは250から1まで、対局をこなすにつれて徐々に減少するようにした。

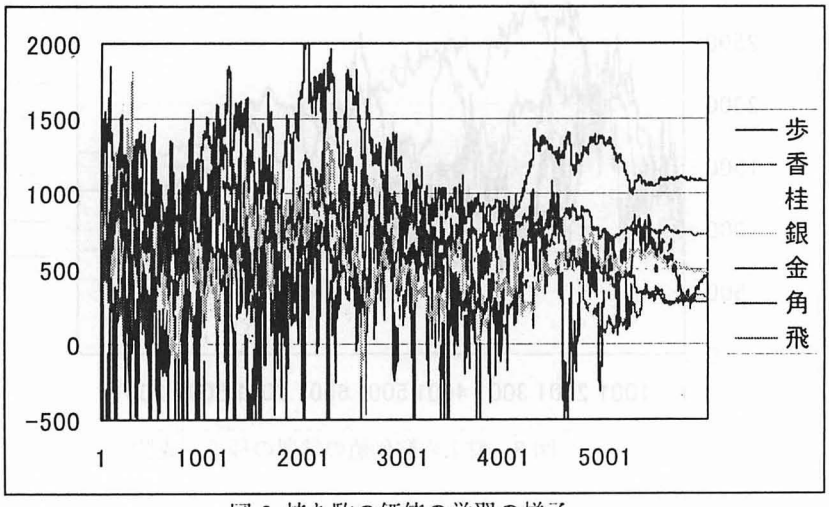

図6 持ち駒の価値の学習の様子

図6では対局が6000回に近づいても駒価値の順位が入れ替わるほど値が変動しており、さらに学習することが必要だと考えられる。

図7はもとの駒の価値を1とした場合に持ち駒の価値がどのくらいになるか、その割合を表している。多くの場合、持ち駒の価値は盤上にある駒の価値の1割増し程度の価値であると考えられているが、今

回の実験ではそのような結果にはならなかった。まず、他の駒の増加率が1.0前後であるのに対して、歩の価値の増加率は4.5と高くなっている。これは駒価値の学習でも述べたように、と金を作ることによるリスクが学習できなかったことによると思われる。また、金・銀・飛の価値は盤上の価値よりも低くなってしまっている。これは図6で値の変動が大きいことから、学習不足による誤差の影響と考えられる。

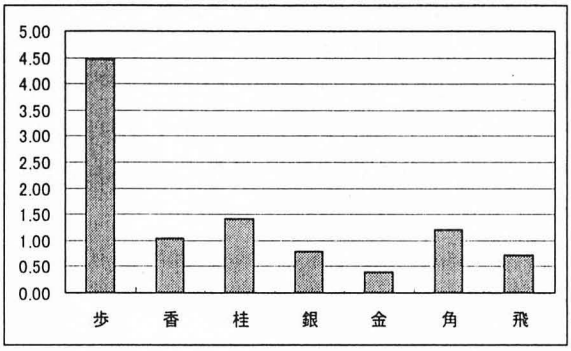

図7 元の駒の価値に対する割合

4.3 王の周りの利き・王と駒の距離

王の周りの利きと王と駒の距離は、どちらも王の安全度に関わってくる評価要素である。図8と図9に学習の様子を示す。実験では盤上の駒の価値を表2の値に固定し、王の周りの利きと王と駒の距離、それぞれの価値の学習だけを行った。対局は6000回行い、学習率αは50から0.2まで、対局をこなすにつれて徐々に減少するようにした。

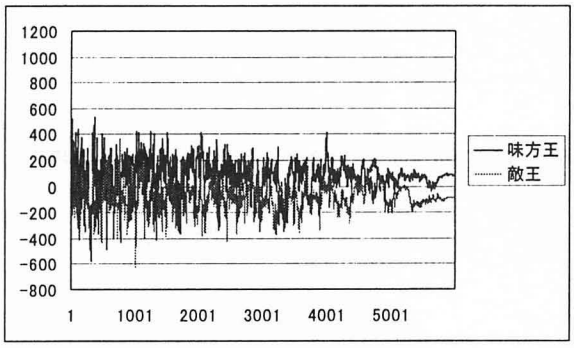

図8 王の周りの利きに関する価値

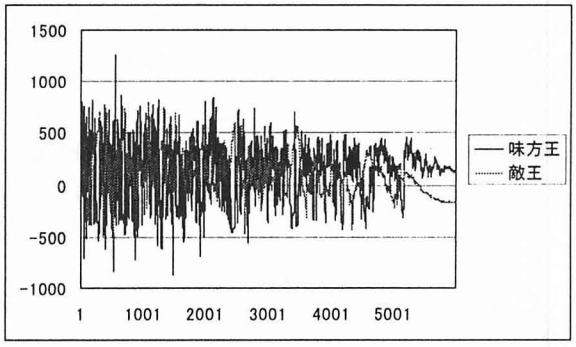

図9 駒と王との距離に関する価値

これらの評価要素は駒価値と比べて値が小さい（0に近い）と考えられており、また学習する評価要素の数が少ないことから、学習率を駒価値の学習時の5分の1としたが、それでも十分学習できている。重みは利き・距離の両方が、味方王については正の値、敵王については負の値になった。このことは敵王よりは味方王の近くに駒があったほうがよいということを表している。しかし、将棋は王を詰ますゲームであるから、敵王の近くに駒がないと勝つことができないはずである。今回の実験では読みの浅さなどから、敵王の近くに駒を移動したり打ったりしてもそのメリットが学習できず（敵王に近い≒敵の駒が多い＝すぐ取られることが多い）、結果として負の値になったと考えられる。

5 おわりに

コンピュータ将棋の静的評価関数について、評価要素の重みを TD 法を用いて学習する実験について述べた。実験の結果、盤上の駒の価値については D.F.Beal と M.C.Smith の実験[1]とおおむね同じ結果が得られた。しかし、持ち駒の価値・王の周りの利き・駒と王との距離についてはあまり良い結果は得られなかった。

学習実験で問題になるのは、まず学習させる評価要素の数である。これは収束の早さに関わってくるが、将棋の場合、駒価値だけでも 13 種類ある。駒価値に加えて他の評価要素も学習させようとするとそれだけ時間がかかるが、今回は駒価値と他の評価要素は別々に学習させた。次に読みの深さであるが、当然深く読めばより正確な学習ができる。しかし、これも学習にかかる時間の問題から深さ 3 で実験を行った。さらに、学習率 α や予想確率に対する重み λ といった定数をどう設定するかという問題もある。

今後の課題としては、

・評価要素を同時に学習させる。

・探索の深さを深くする

・学習に適した α や λ を探す

・いろいろな対戦相手と対戦させて学習させる

といったことがある。

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

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