社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－GI－15（6）
2006／3／7

機械学習と自動プレイを用いた将棋類の類似度比較について

佐々木宣介
県立広島大学経営情報学部
sasaki@pu-hiroshima.ac.jp

概 要

本研究の大きな目標は，世界の全将棋種を対象に，ルールの変遷が各将棋種にどのような影響を与えたかを探ることである．これまで計算機による自動プレイによって将棋とその変種のデータを調べ，質的類似度についての評価を行ってきた．また，自動プレイ実験で使われる思考アルゴリズムでは駒価値の評価を利用しているが，既に廃れてしまった将棋種の駒価値の学習を強化学習の一種である TD 学習法を適用することを提案し，実験を行ってきた．本論文では，TD 学習法による駒価値の学習に際して，一部の変種において学習の進行が不安定であった点について，これを改善した結果を報告する．

The Similarity Estimation Analysis of Shogi Variants Using Machine Learning and Self Play Experiments.

Nobusuke Sasaki

Faculty of Management and Information Systems,
Prefectural University of Hiroshima
sasaki@pu-hiroshima.ac.jp

Abstract

This study explores how the evolutionary changes of the rules affect the characteristics of the games in the Shogi species. The author proposed the self-play experiment to obtain the statistical game data and analyze the similarity of modern and ancient Shogi variants. The author made computer programs that performed lookahead search using an evaluation function based simply on piece-material balance, then, applied Temporal Difference Learning method to obtain the appropriate piece values even for obsolete variants. However, the learning procedure of some variants was not so stable in former experiment. In this paper, the author reports some improvement of this problem.

1　はじめに

本研究の目的は，世界の将棋類において，ゲームのルールの変遷がゲームの質にどのような影響を与えたかを探ることである．ゲームの進化の研究は，文献，フィールド調査等を使って行われるが，本研究はコンピュータ自動プレイによる解析により，文献等の調査とは異なる視点から知見を得ることを目指している．

先行研究において，それぞれ異なる進化を経て，異なるルールが定着し，生き残った世界三大将棋 (将棋，チェス，中国象棋) で，平均終了手数 D，平均合法手数 B から計算される，$\frac{\sqrt{B}}{D}$ の値がプロ棋士レベルのゲームでほぼ一定の値となっていることに着目してきた．$\frac{\sqrt{B}}{D}$ の値が将棋種のルールの進化論的変遷を評価する上で，重要な指標になると考え，D，B の他，この $\frac{\sqrt{B}}{D}$ の値も利用して将棋種間の質的類似度の評価を行ってきた [1],[2].

これらの先行研究では，既に廃れてしまい，プレイヤがいないような歴史的変種においても，ゲームのデータを簡便に採取する手法として，コンピュータプログラムによる自動プレイを提案した．また，自動プレイにおいては，完全にランダムにプレイするプログラムよりも，より人間のプレイヤに近いデータを採取するために，機械学習の手法を利用することにより，ある程度の強さを持つ思考アルゴリズムを自動的に作製することを提案し，計算機実験を行った．これは，駒の損得のみを評価関数とする思考アルゴリズムの自動プレイプログラムを作製し，強化学習の一種である TD 学習法 (Temporal Difference Learning) を利用して駒価値の学習を行うものである．

TD 学習法は強化学習の一手法であり，Samuel により導入され [3]，Sutton によって拡張された [4]．ゲームプログラミングにおいても多くの応用例があり，中でも Tesauro によるバックギャモンへの適用が有名である [5]．TD 学習法による現代将棋の駒価値の学習は Don Beal 等によって最初に試みられた [6].

先行研究において，古代の将棋である平安将棋と現代将棋，その中間形のルールを持つ変種について，各変種に対して，駒の損得のみを評価関数とするアルゴリズムに最適化して駒価値の学習を行い，それらの値は，そのアルゴリズムで動作させる限りにおいては，トップレベルのプログラムで使われている駒価値をこのアルゴリズムで動作させるよりも優れた結果を示した [2].

同じ思考アルゴリズムを用いて自動プレイ実験を行ってゲームのデータを収集，解析した結果，日本将棋における大きな 2 つのルールの変化，大駒ルールおよび持駒ルールにおいては，持駒ルールの付加の方がより大きな変化であること及び大駒の付加は，それ単独ではなく，持駒ルールと組み合わされることにより，より大きな影響を与えたと推測した．

この手法を利用すれば，既に廃れてしまって強いプレイヤが存在せず，適切な評価値のバランスを決定することができないゲームに対しても，ある程度適切な値を提供することが可能になると期待される．特に，平安将棋から現代将棋につながる小さい盤の将棋だけでなく，中将棋，大将棋などの大きな盤の将棋についての解析を進める上では，機械学習を用いた自動的な学習の重要性は高くなると考えられる．なぜなら，これらの大きな盤の将棋には現代将棋には存在しないさまざまな駒があるからである．これらの駒についても適切な駒価値を決定しようとしても，現在では廃れてプレイヤが存在しないものもあり，現代将棋につながるグループの変種と比較して，駒価値を決定する手がかりがさらに少なくなっている* ．

しかし，先行実験における学習結果は，一部の変種，特に平安将棋に持駒ルールを加えた変種においては学習が安定しないという問題があった．この問題を解消するため,「盤上の駒と持駒を別々の評価値に分けて学習を行う」,「学習パラメータの調整を行う」などの方法で，学習の安定化を目指したが，本論文では，学習パラメータの調整によってある程度安定した学習を行うことが可能となったので，その結果について報告する．

2 TD 学習法

本章では TD 学習法についての概要を示す．

ある局面 A における評価値 ν は，以下の式であらわされる．

$$\nu(A) = \sum_j w_j x_j(A) \tag{1}$$

j は評価要素，x_j は評価要素の特徴量を表す．

また，ある局面において，その局面からそのゲームが勝利に終わる予測確率を P として，以下のシグモイド関数で表わされる．

$$P = \frac{1}{1+e^{-\nu}} \tag{2}$$

TD 学習法においては，t 手目の局面における評価要素の重みの更新値 Δw_t は，以下の式で表現される．

$$\Delta w_t = \alpha(P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \nabla_w P_k \tag{3}$$

ここで，α は学習レートを制御するパラメータである．λ は $0 \leq \lambda \leq 1$ の範囲をとり，過去の状態を学習に反映させる程度を制御するパラメータである．λ が 0 であれば現在の状態のみを利用して更新を行なう．

本論文における実験では評価要素は駒の損得のみを利用しているので，式 1 は以下のように表わすことができる．

$$\nu(A) = \sum_j w_j (N_a(A) - N_b(A)) \qquad j = \{\, \text{歩}, \text{香}, \text{桂}, \cdots \} \tag{4}$$

ただし，$N_a(A)$ は味方の駒 j の枚数，$N_b(A)$ は敵方の駒 j の枚数をあらわす．式 4 における評価要素の重み w_j の値がすなわち駒価値を表わす．持駒ルールのない平安将棋及び平安将棋+大駒においては，駒を取った際の評価関数の値の変化は，取った駒の値だけ有利になる．一方，持駒ルールが存在する将棋及び平安将棋+持駒においては，駒を取った際の評価関数の値の変化は持駒ルールの存在しない変種と比べて 2 倍となる．

また，式 2 のシグモイド関数の導関数は以下のように単純な形で表わされる．

$$\frac{dP}{d\nu} = P(1-P) \tag{5}$$

なお，本実験では駒の損得のみを学習の対象としているが，式 1 にその他の評価要素を含んでいる場合に，TD 学習法はその評価要素の重みも同様に学習可能である．

3 実験

3.1 TD 学習法による駒価値の学習

TD 学習の実験は以下の手順で行った．

* 中将棋は現在もわずかであるがプレイヤがいる

- 学習に用いたプログラムは，$\alpha\beta$ 法で全幅探索を行なう．
- 先手後手双方とも同一のアルゴリズムで動作し，学習している駒価値の値をそのまま評価関数で利用する．
- 評価関数は駒の損得のみを計算している．また，深さ 3 の詰み探索も行なう．
- 先読みの深さは 3 とし，探索木の末端でさらに駒の取り合いが発生する際には，駒の取り合いのみの探索延長 (静けさ探索) を行なう．探索延長は，静かな局面になるか，深さ 6 に達した場合に先読みを打ち切る．
- 同じ評価値の最善手が複数ある場合には，その中からランダムに次の一手を選択する．
- 1 手進める度に式 3 によって駒価値 w_j の更新を行なう．
- 持駒と盤上の駒は区別していない．したがって，今回の計算機実験では，持駒の価値は学習していない．
- 持駒ルールを有しない平安将棋及び平安将棋+大駒では，一方が玉一枚になった時にはそこでゲームを打ち切り，次のゲームに移る．
- 1000 手以上経過しても勝負がつかなかった場合には，そこでゲームを止めて引き分けとして処理する．

学習は 10000 局行なった．駒価値 w_j の初期値は 1 とした．先行実験においては，次のような学習パラーメータを採用していた．学習レートを制御する α の初期値は 0.05 とし，学習が 1 局終了するごとに少しずつ減少させて，0.002 まで変化させる．約 3000 局の時点で 0.002 まで達した後は 0.002 で固定する．また，過去の局面の状態を現在の学習に反映させる割合を制御する λ の値は 0.95 であった．

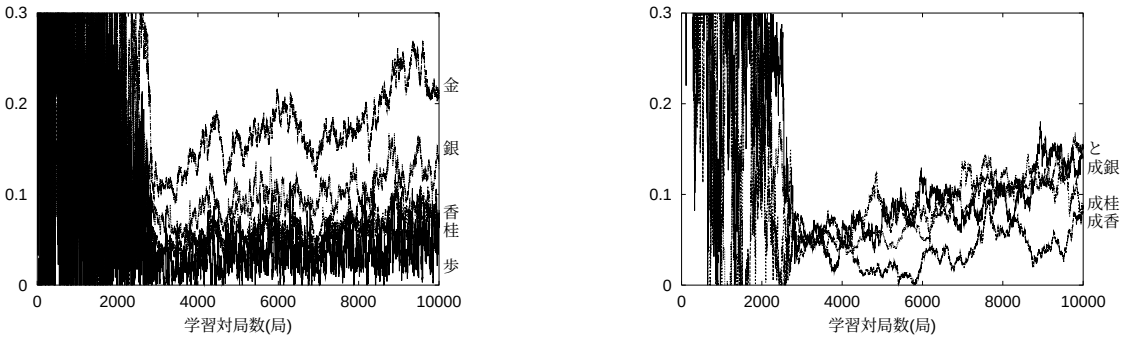
今回は，α の値を減少させる割合，最終的に固定する値，λ の値などを変化させて学習の安定性を調査した．

3.2 学習パラメータの調整結果

前節で述べたようにいくつかの学習パラメータの調整を行ったが，その結果，特に影響が大きかったのが過去の局面の状態をどの程度学習に反映させるかを制御する λ の値であった．

図 1 から図 4 まで，平安将棋+持駒ルールの変種において，λ の値を 0.95 から 0.7 まで変更して学習を行ったそれぞれの学習曲線を示す．なお，ここに示した図では，α については，先行研究と同様の設定を用いたものである．

λ の値が 0.95 の場合と比較して明らかに 0.8，0.7 の場合は学習の安定性が改善していることがわかる．



(a) 生駒　　(b) 成駒

図 1: λ の値が 0.95 の時の学習曲線 (平安将棋+持駒)

学習の安定性を評価するために，5501 局終了時から 6000 局終了時までの 500 点の学習値を取り出してその平均値を計算し，その平均値に対して，95%から 105% の間に位置するのが何点存在するのか計算した．その結果を表 1 に示す．先行実験で用いた値 λ=0.95 に対して，0.8, 0.7 のデータは，この 1 割の範囲に収まっている割合が明らかに高くなっており，値のばらつきが小さく，学習の安定性が増していることがわかる．

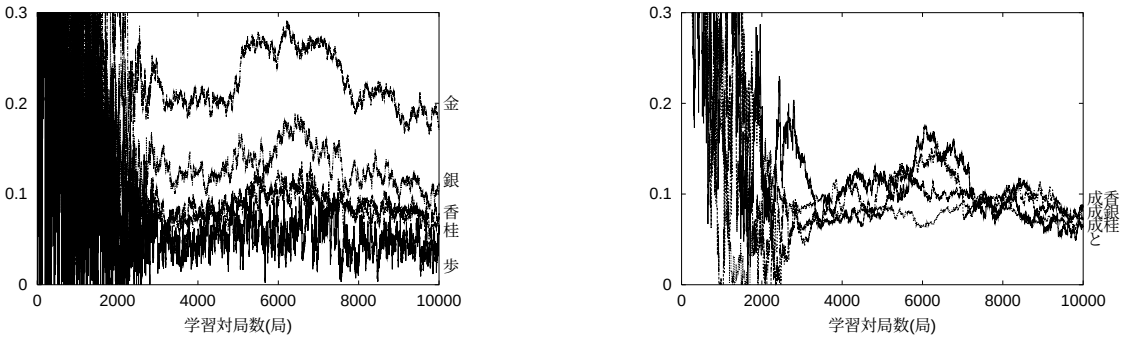
このように先行実験では安定した学習が行うことができなかった平安将棋+持駒ルールにおいても，学習時のパラメータの調整により，安定した学習が可能となるであることがわかった．

参考までに，図 5-7 に，これ以外の変種についても同じパラメータ設定で学習を行った場合の学習曲線を示す．将棋や他の変種についても同じく学習の経過が安定していることがわかる．

また，表 2 には，λ の値が 0.8 の時の，6000 局の学習を行なった時点で，学習した駒価値の値を直前の 500 局分について，平均値を求め，それを歩の価値で正規化した結果の一覧を示す．

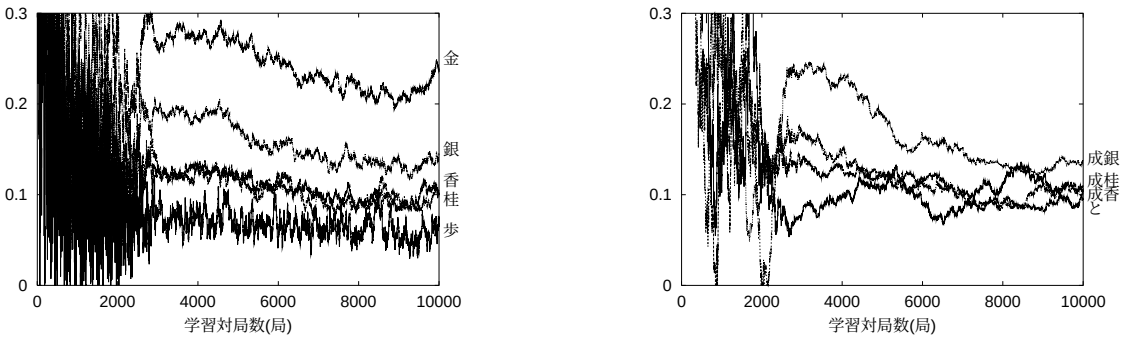
3.3 学習値の評価および自動プレイによる実験

駒価値を TD 学習で学習する時の設定を調整することにより，学習の安定性を向上させることができたが，得られた数値の妥当性を評価するために，先行実験と同様の計算機実験を行い，先行実験によって得られた結果と同様の傾向を



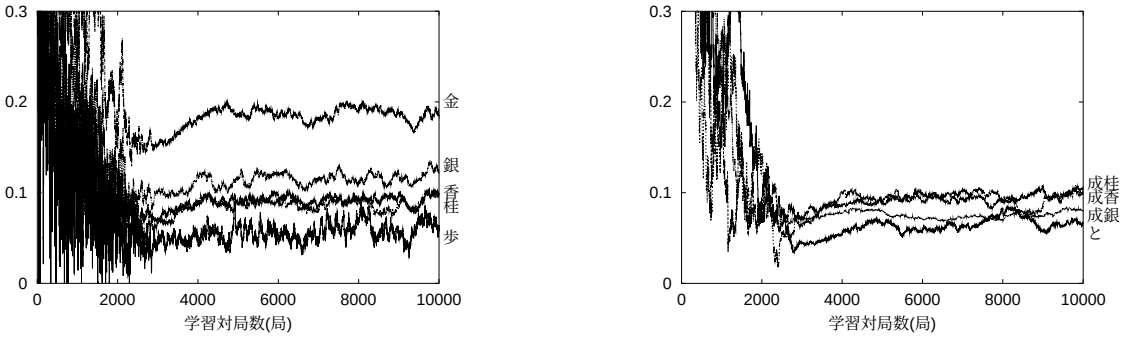
(a) 生駒　(b) 成駒

図 2: λ の値が 0.9 の時の学習曲線 (平安将棋+持駒)



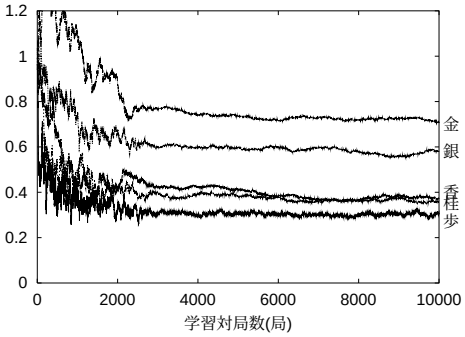
(a) 生駒　(b) 成駒

図 3: λ の値が 0.7 の時の学習曲線 (平安将棋+持駒)

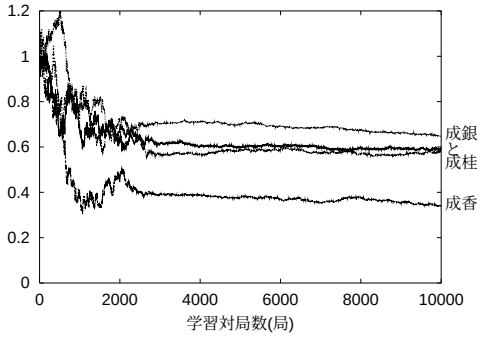


(a) 生駒　(b) 成駒

図 4: λ の値が 0.7 の時の学習曲線 (平安将棋+持駒)

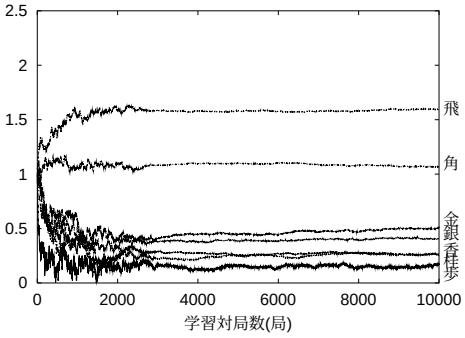


(a) 生駒

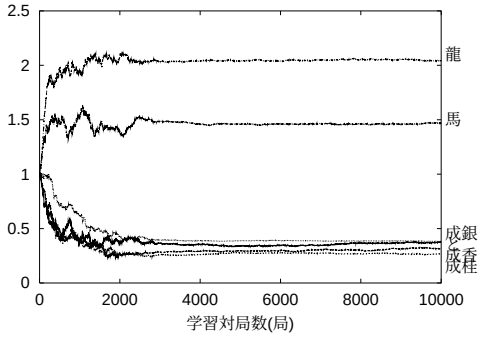


(b) 成駒

図 5: 平安将棋の学習曲線

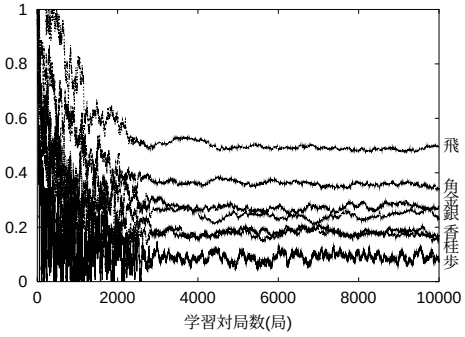


(a) 生駒

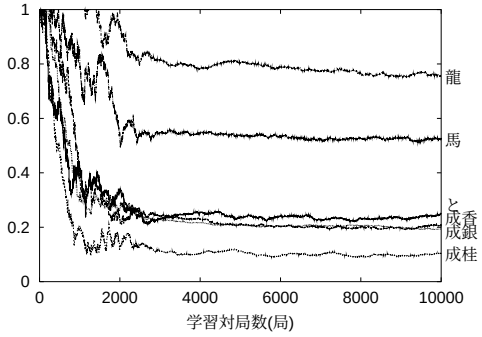


(b) 成駒

図 6: 平安将棋+大駒の学習曲線



(a) 生駒



(b) 成駒

図 7: 将棋の学習曲線

λ	歩	香	桂	銀	金	と	成香	成桂	成銀
0.95	7.4	18.8	41.0	38.8	48.2	17.4	4.8	37.2	45.0
0.9	9.0	45.0	44.6	37.8	86.2	17.4	52.6	46.4	23.2
0.8	28.8	86.8	62.0	96.0	100.0	71.0	86.0	87.6	92.8
0.7	28.4	85.6	98.8	100.0	100.0	91.0	93.2	93.8	100.0
0.6	26.2	95.4	71.6	86.6	98.4	34.0	96.0	70.6	100.0
0.5	32.2	96.4	98.6	48.2	97.4	99.8	96.8	100.0	100.0

表 1: 学習 6000 局の直前 500 局の学習値のばらつき (500 点の平均値の上下 1 割以内に入っている割合)

	平安将棋	平安将棋 +大駒	平安将棋 +持駒	現代将棋
歩	1.00	1.00	1.00	1.00
と	2.00	2.40	1.48	3.06
香	1.27	1.83	1.59	2.38
成香	1.24	2.07	1.58	2.69
桂	1.24	1.79	1.53	2.11
成桂	1.93	1.93	1.71	1.24
銀	1.96	2.70	2.19	3.19
成銀	2.28	2.73	2.20	2.67
金	2.37	3.12	3.60	3.49
角		7.67		4.81
馬		10.21		7.02
飛		11.07		6.48
龍		14.30		10.35

表 2: 学習した相対的駒価値．5501 局から 6000 局終了時の平均値 (λ=0.8 のデータ)

示すことを確認した．

まず，得られた駒価値がどの程度妥当であるか評価するために，トップグループの将棋プログラムである YSS の駒価値 [7] の値を用いた相手と対戦させた．

TD 学習で用いたプログラムと同一のアルゴリズムの，駒価値のみを評価関数とするプログラムで，一方のプレイヤを YSS が用いている駒価値を使い，もう一方のプレイヤが学習によって得られた値を用いる．なお学習した駒価値としては，表 2 の値を用いた．詰み探索の深さは 3 手，先読みの深さは 3 手で最大 6 手まで静けさ探索をおこなう．以上のプログラムを使用して，学習値が先手の場合と後手の場合それぞれで 1000 局の対戦を行なった．

表 3 にその結果を示す．

	学習値が先手			学習値が後手		
	勝ち	負け	引分	勝ち	負け	引分
平安将棋	171	132	697	173	108	719
平安+大駒	479	206	315	483	214	303
平安+持駒	417	394	189	424	395	181
将棋	521	476	3	527	465	8

表 3: TD 法で学習した駒価値と YSS の駒価値の対戦結果

どの変種においても，学習した値は YSS の駒価値を使った相手に勝ち越した．勝敗の数がかなり接近しているものもあるが，少なくとも YSS の駒価値に対して同等以上の対戦結果をあげていると言える．したがって，今回使用した「駒の損得のみを評価関数とした静けさ探索付き先読み」に対し，ある程度最適化された値が学習されていることが確認された．

次に，これも先行研究と同様に TD 学習により学習した駒価値の値を用いて，自動プレイの実験を行なった．

実験は以下の条件で行なった．

- 同一のアルゴリズムで動作するコンピュータプログラムを用いて，多数の対戦を行う．(100-10000 局)
- プログラムは以下の 3 種類のアルゴリズムで動作する．
  1. 全合法手の中から完全にランダムに指し手を選択する．(アルゴリズム 1)

2. 詰み探索の能力のみを有する．相手玉の連続王手の詰みのみを探索し，詰みを発見できなかった場合にはランダムに指し手を選択する．詰み探索の深さは 1，3，5，7 の 4 種類．(アルゴリズム 2)
3. 詰み探索に加え，駒の損得を評価関数として用いた先読みの能力を持つ．最初に相手玉の詰み探索を行って，詰みを発見できなかった場合には，評価関数を用いて先読みを行い，最善手を探す．先読みの深さは 1，3，5 の 3 種類，詰み探索の深さは 5 に固定する．また，先読みの末端局面で，駒の取り合いが生じている場合には，最大で末端局面+3 手まで静けさ探索をおこなう．駒の価値として，TD 学習で獲得した値を使用する．(アルゴリズム 3)

- 1000 手以上経過しても勝負がつかなかった場合には，そこでゲームを止めて引き分けとして処理する．
- 引き分けに終わったゲームのデータは D および B の算出には使用しない．

アルゴリズム 1，アルゴリズム 2 については 10000 局の自動対戦を行い，アルゴリズム 3 については先読み深さが 1 手，3 手の場合は 1000 局，5 手の場合は 100 局の自動対戦を行った．

その結果は，先行研究で得られたデータと同様で，平安将棋と平安将棋+大駒の変種がほぼ同じ特徴を示し，平安将棋+持駒や将棋とはある程度離れていることが確認された．この自動対戦実験の結果の一部として，$\frac{\sqrt{B}}{D}$ の値の変化をアルゴリズム 2 における結果を図 8，アルゴリズム 3 における結果を図 9 に示す．

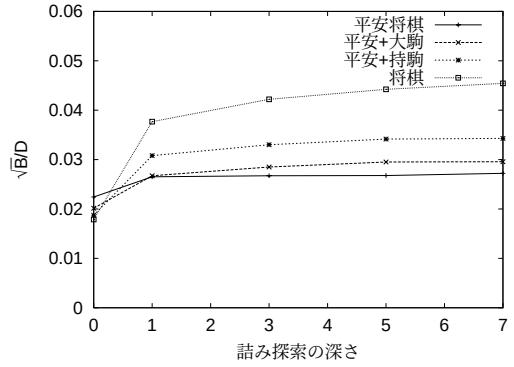


図 8: 将棋の歴史的変種において詰み探索の深さを変えた時の $\frac{\sqrt{B}}{D}$ の変化

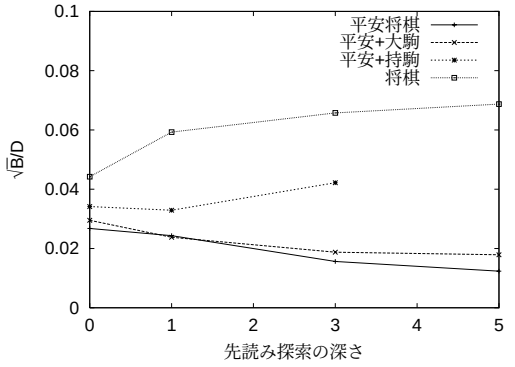


図 9: 将棋の歴史的変種のおいて先読み深さを変えた時の $\frac{\sqrt{B}}{D}$ の変化

4 まとめと今後の課題

本論文では，将棋及びその歴史的変種において TD 学習法を利用して相対的駒価値の学習を行なう手法について，これまでの実験においては，平安将棋+持駒ルールの駒価値の学習において，学習が十分に安定しないという問題があったが，学習の際のパラメータを調整することで，安定した学習が可能となることを示した．

また，今回の改良の結果得られた駒価値について，先行研究と同様に自動プレイの実験で使用してきた将棋プログラム YSS の駒価値とを対戦させて比較する実験および自動プレイ実験を行った．その結果は先行研究の計算機実験と同等の結果を示した．

今後は現代将棋につながるいわゆる「小将棋」の変種の他に中将棋，大将棋等の大きな盤の将棋類についても，解析を行っていく予定である．これらの大きな盤の変種においては，現代将棋には存在しない駒が多数あり，また，既にプレイヤが存在しない変種もあり，最適な駒価値のバランスはよくわかっていない部分もある．これらの駒価値の決定にも，機械学習を用いた自動的な学習がある程度有効であると期待できる．

なお，大きな盤の変種は，そのルールにおいても，一部特別なルールが存在する．例えば中将棋においては,「獅子」という駒について，獅子同士の駒の取り合いには一定の制約が設けられている．また,「酔象」という駒が成ると「太子」となり，太子が盤上にある場合には玉が取られてもゲームが続行できるといったルールも現代将棋には存在しない特別なルールである．

これらの特徴的なルールの有無がゲームの質にどのような影響を与えるのか，といった点も含めて解析を行っていく予定である．

謝辞

本研究は文部科学省科学研究費補助金（若手研究 (B)：16700240）による助成を受けた．

参考文献

[1] 佐々木宣介, 橋本剛, 飯田弘之 (1999). “自動プレイによるチェスライクゲームの歴史的進化の研究” **ゲームプログラミング・ワークショップ** *‘99 (IPSJ Symposium Series, vol. 99, No. 14)*, pp39-45.

[2] 佐々木宣介, 飯田弘之 (2002). “将棋種の歴史的変遷の解析” **情報処理学会論文誌**, *vol. 43 No. 10*, pp.2990-2997.

[3] A. L. Samuel, (1959). “Some Studies in Machine Learning Using the Game of Checkers” IBM Journal of Research and Development, 3, pp.210-229.

[4] R. Sutton (1988). “Learning to Predict by the Methods of Temporal Differences” Machine Learning, 3, pp.9-44.

[5] G. Tesauro, (1994). “TD-Gammon, a self-teaching backgammon program, achieves master-level play.” Neural Computation, 6, pp.215-219.

[6] Donald F. Beal and Martin C. Smith (1998). “First Results from Using Temporal Difference Learning in Shogi” in *Lecture Notes in Computer Science, LNCS 1558* (eds. Jaap van den Herik and Hiroyuki Iida), Springer-Verlag, pp.113–125.

[7] 山下宏, (1998). “YSS –そのデータ構造，およびアルゴリズムについて” **コンピュータ将棋の進歩２**, **松原仁** *(***編著***)*, pp.112-142, 共立出版, ISBN4-320-02892-9.

[8] 梅林勲, 岡野伸, (2000). “世界の将棋 改訂版”, 将棋天国社.