着手決定の複雑さの指標とゲームの進化論的変遷

佐々木宣介 * 武下信夫 ** 橋本剛 ** 飯田弘之 **
* 広島県立大学 ** 静岡大学
sasaki@bus.hiroshima-pu.ac.jp
{cs6056, hasimoto, iida}@cs.inf.shizuoka.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，ルールの進化論的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数多くの将棋種のデータを調べてきたが，既に廃れてしまった将棋種の駒価値などの評価値はて適切な値が不明で，やむを得ず現代将棋のプログラムで用いている値をそのまま利用してきた．本論文では，TD 学習法を適用して自動的に適切な駒価値を学習させることを試みた．現代将棋及びその歴史的変種に対して TD 学習法を適用し，駒の相対的価値の自動学習を行なった結果，これまで自動プレイの実験で用いてきた YSS の駒価値よりも最適化された値を獲得した．これによって得られた駒価値を使用して駒の損得のみを評価関数とする探索アルゴリズムを用いた自動プレイにより将棋種のデータを調査・評価した．今後，本論文の手法を発展させることにより，将棋種のデータを調査する際に，自動的に評価関数のバランスを調整して実験を行なうための標準的な手法の確立につながることが期待される．

Decision-Complexity Estimate in Evolutionary Changes of Games

Abstract

This study explores how the evolutionary changes of the rules affect the strategic characteristics of the games in the chess species, whose variants have been distributed world wide, from the viewpoint of evolutionary selection. We have developed a method of computer self-play experiments to collect the game data and analyzed the strategic characteristics of the games using a proposed estimate of decision complexity for chess species. Since we do not have sufficient knowledge about obsolete variants, we have used piece values for the material balance in evaluation function after a SHOGI program for all shogi variants in the previous self-play experiments. However, these values may not suitable for old variants. In this paper, we apply Temporal Difference Learning to acquire the piece values of all SHOGI variants automatically. Then, computer self-play experiments were performed to compare each other based on the proposed estimate of decision complexity in chess species. We expect that the method proposed in this paper will be a standard procedure for self-play experiments to obtain game statiscis of obsolete variants.

1 はじめに

本研究の目的は，世界の将棋類において，進化論の見地から，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でゲームの着手決定の複雑さに関わる性質の類似度を評価することで，文献等の調査とは異なる視点から新たな知見を得るものである．

先行研究において，われわれは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，この $\frac{\sqrt{B}}{D}$ の値を利用して将棋種間の質的類似度の評価を行ってきた [1].

既に廃れてしまった歴史的変種の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，計算機実験を行った．世界三大将棋とされるチェス，象棋，将棋の他，将棋種の祖先とされている Chaturanga に対しても，自動プレイによって，数多くの対局のデータを採取し，その解析を試みた．また，日本将棋とその歴史的変種においては，単純なランダムプレイの他，駒の損得

のみを評価関数として持つアルゴリズムによる自動プレイによって D, B のデータも採取した [2]. これには以下の狙いがあった.

- より (強い) 人間プレイヤに近い対局データを求める
- 種々のアルゴリズムの違いによって D, B がどのように変化するかを比較し, 似たようなデータの変化が生じた変種は質的に近いという観点から, 変種間の類似度の評価を試みる

その結果, 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.

また, 自動プレイを行うゲームアナライザを作成し, ゲームの戦略的複雑さを表す指標 $\frac{\sqrt{B}}{D}$ を初めとする統計データを調べたところ, SHATRANJ からチェスへの進化において, 最初にフィルツェーン・フィールからクイーン・ビショップへの変更が起こり, その後ポーンの初手 2 マス移動が加わったことを示唆する結果を得た [3].

しかし, 自動プレイを用いた計算機実験において, より人間のプレイヤに近いデータを採取しようと考えた時, 既に廃れたゲームにおいては, 強いプレイヤが存在しないため, 人間が自分で評価要素のバランスを決定することが困難であった. これまで日本将棋とその歴史的変種に対して行った実験において, 駒の損得を評価関数とするアルゴリズムを利用したが, その計算機実験では, 日本将棋の既存のプログラムでトップレベルのプログラムのひとつである YSS で用いられている値を駒の評価値として使用した. しかし, 歴史的変種においては, ルール, 駒数の違いなどがあり, その値は最適な値ではないと考えられる. また, 現代将棋プログラムについても, この値は駒価値だけでなく, さまざまな評価要素を考慮したアルゴリズムにおいて最適化された値であるため, 駒価値のみを評価関数とするアルゴリズムには最適化されていない. より人間の対局に近いデータを採取するためには, 各変種それぞれについて, 最適な駒の評価値を決定する必要がある.

本論文では, 各変種それぞれにおいて, 駒の損得のみを評価関数とするアルゴリズムに対してチューニングした相対的駒価値を求めるために, TD 学習法 (Temporal Difference Learning) を利用して駒価値の学習を行った結果を報告する. TD 学習法は強化学習の一手法であり, Samuel により導入され [4], Sutton によって拡張された. [5] 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. [6] 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 [7]. 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よりも優れた結果を示した. この他に, 駒価値以外の評価要素の学習への適用も試みられている [8]. 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. 本手法の適用を通じ, 既に廃れたゲームに関して, 自動的に評価要素のバランスを決定し, より信頼性の高いデータを得ることが出来る, 標準的な手法を確立することを目指す.

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．

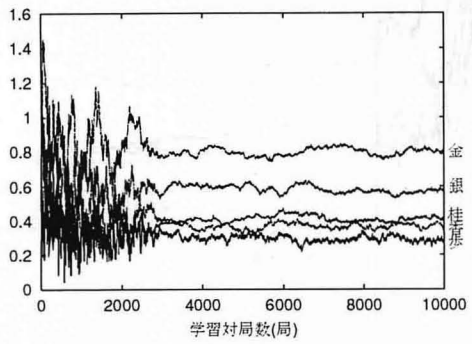
- 学習に用いたプログラムは，$\alpha\beta$ 法で全幅探索を行なう．
- 先手後手双方とも同一のアルゴリズムで動作し，学習している駒価値の値をそのまま評価関数で利用する．
- 評価関数は駒の損得のみを計算している．また，深さ 3 の詰み探索も行なう．
- 先読みの深さは 3 とし，探索木の末端でさらに駒の取り合いが発生する際には，駒の取り合いのみの探索延長 (静けさ探索) を行なう．探索延長は，静かな局面になるか，深さ 6 に達した場合に先読みを打ち切る．
- 同じ評価値の最善手が複数ある場合には，その中からランダムに次の一手を選択する．
- 1 手進める度に式 3 によって駒価値 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．また，α の初期値は 0.05 で，学習が 1 局終了するごとに少しずつ減少させ，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．λ の値は 0.95 とした．

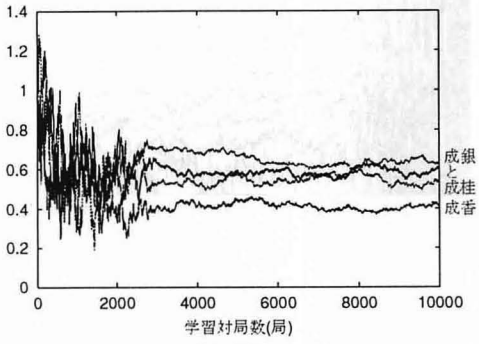
3.2 学習結果

将棋とその歴史的変種について，TD 学習法により，駒価値の学習を行なった結果を以下に示す．

図 1-4 にそれぞれの変種の学習の様子を示す．また，図 5-8 及び表 1 に各変種における，歩の価値を 1 とした時の各駒の相対的評価値を示す．

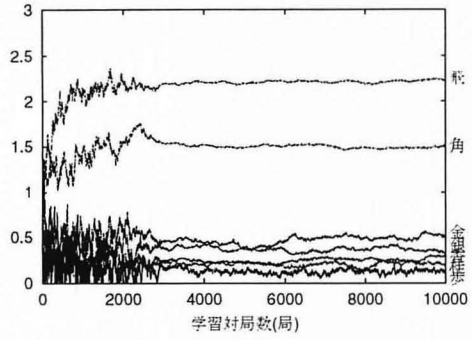


(a) 生駒

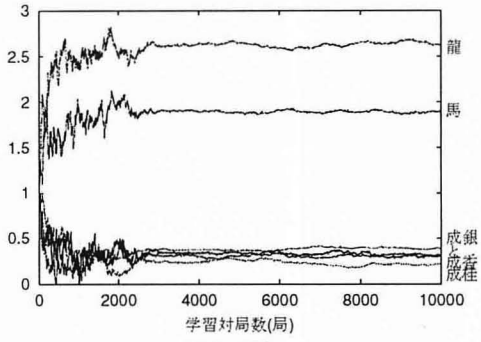


(b) 成駒

図 1: 平安将棋の学習曲線

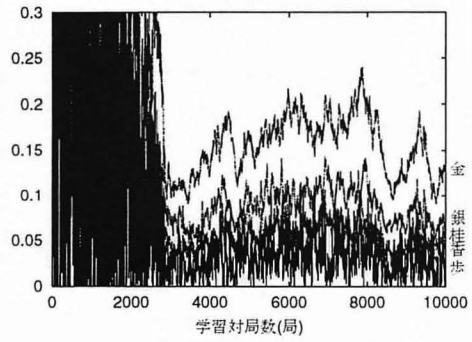


(a) 生駒

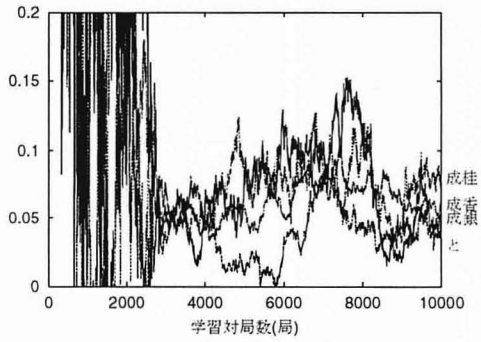


(b) 成駒

図 2: 平安将棋+大駒の学習曲線

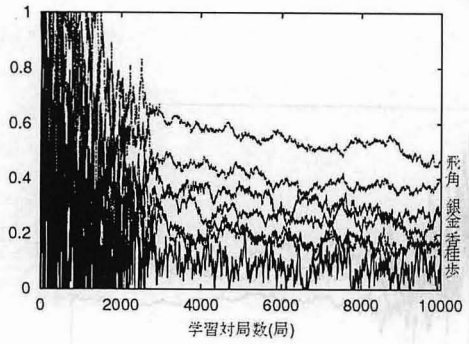


(a) 生駒

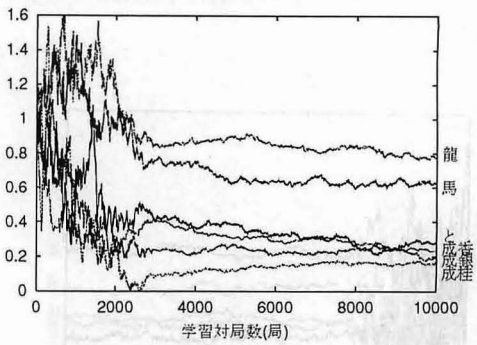


(b) 成駒

図 3: 平安将棋+持駒の学習曲線



(a) 生駒



(b) 成駒

図 4: 将棋の学習曲線

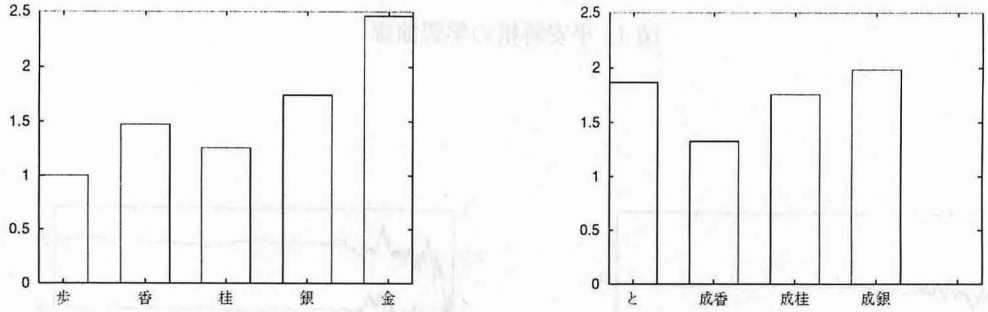


図 5: 学習によって得られた平安将棋の相対的駒価値 (歩を 1 とする)

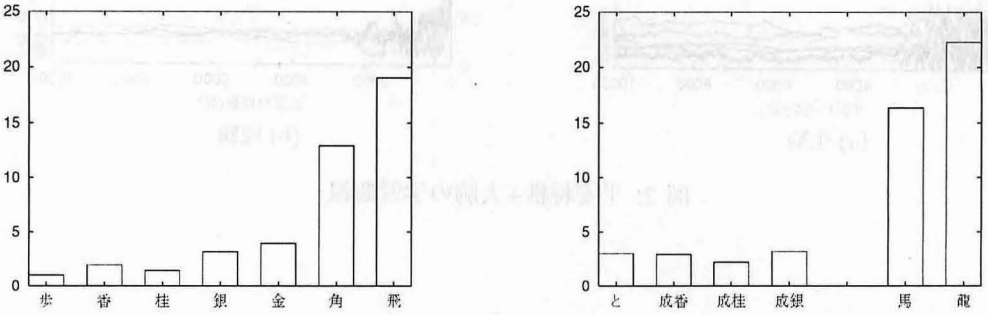


図 6: 学習によって得られた平安将棋+大駒の相対的駒価値 (歩を 1 とする)

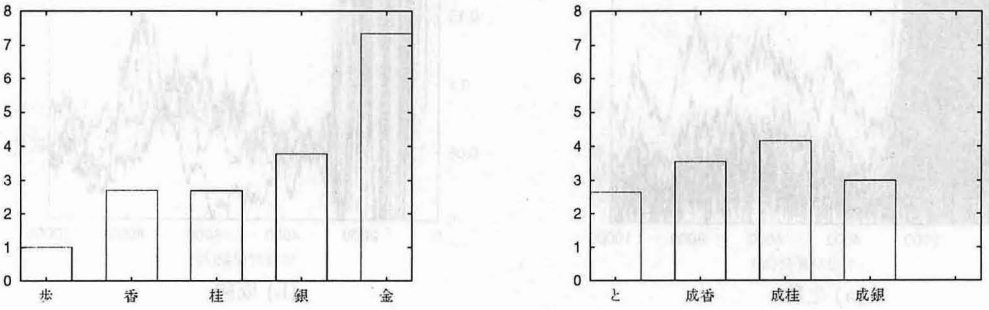


図 7: 学習によって得られた平安将棋+持駒の相対的駒価値 (歩を 1 とする)

	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	1.87	2.98	2.64	6.85
香	1.47	1.92	2.70	3.45
成香	1.33	2.92	3.56	4.48
桂	1.26	1.41	2.69	4.26
成桂	1.76	2.20	4.17	2.75
銀	1.74	3.13	3.78	5.81
成銀	1.98	3.17	3.00	5.98
金	2.46	3.90	7.34	6.22
角		12.89		8.18
馬		16.36		13.08
飛		19.02		11.51
龍		22.25		17.44

表 1: 学習した相対的駒価値 (歩の価値を 1 とする). 平安将棋+持駒は 10000 局終了時, その他は 6000 局終了時

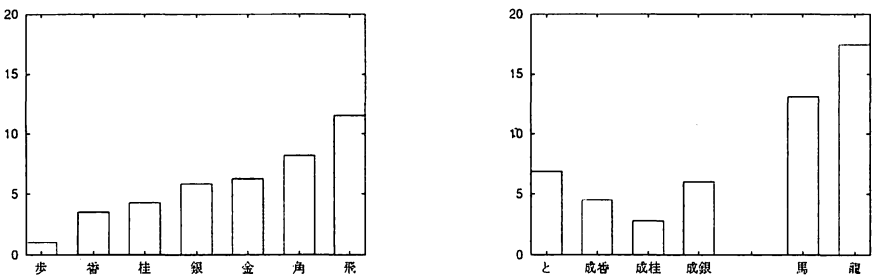


図 8: 学習によって得られた将棋の相対的駒価値 (歩を 1 とする)

図 5-8 及び表 1 の数値は, 平安将棋+持駒を除く 3 種の変種については, 6000 局の学習を行なった時点の学習結果である. 平安将棋+持駒ルールの変種では学習が十分に安定していなかったため, 10000 局終了時点での結果を示す.

以上の結果からいくつかの特徴が読み取れる.

1. 平安将棋+持駒の駒価値の学習の様子が十分に安定していない. (図 3)
2. ほとんどの変種で, と金の価値が他の小駒の成駒と比べて大きくなる傾向がある.
3. 将棋における大駒と小駒の価値の差よりも平安将棋+大駒の大駒と小駒の価値の差が大きい.

1 については, 実験に使用している先読み探索アルゴリズムがいまだ十分ではないことが考えられる. プログラムが適切でない指し手を選択するケースが多い場合, ある時点で有利であるプレイヤが後で形勢を逆転されてしまうことが多く生じてしまい, うまく学習が進まなくなる. また, 駒価値の評価で盤上の駒と持駒とを区別していないことによる影響も考慮する必要がある.

2, 3 については興味深い結果ではあるが, 値の再現性も含め, 今回の実験だけでなく, 今後さらに詳細に実験を行なって評価する必要があると考えられる.

3.3 学習の評価

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 [9] 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で動作する. 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. 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 2 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	150	133	717	151	126	723
平安+大駒	508	201	291	499	201	300
平安+持駒	466	334	200	443	348	209
将棋	519	477	4	520	476	4

表 2: TD 法で学習した駒価値と YSS の駒価値の対戦結果

どの変種のおいても，学習した値は YSS の駒価値を使った相手に勝ち越し，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．

3.4 自動プレイによる実験

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．
将棋の各変種における自動対戦の結果を図に示す．アルゴリズム 2 における結果を図 9，アルゴリズム 3 における結果を図 10 に示す．
先行研究で得られたデータと同様に，平安将棋と平安将棋+大駒の変種がほぼ同じ特徴を示すことが確認された．

4 結論と今後の課題

本論文では，TD 学習法を利用して将棋及びその歴史的変種に関してその相対的駒価値の学習を行なった．学習の結果得られた駒価値と，これまで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したところ，全ての変種において，学習した値が YSS の値に勝ち越すという結果が得られた．このことは本論文で使用しているアルゴリズムに対してある程度最適化された相対的駒価値を学習したと言える．以上のことから，TD 学習法によって評価関数の最適化を行い，その結果をもとに自動プレイ実験を行うという手法によってこれまでよりも信頼性の高いデータを採取可能となることが期待できる．
また，本論文では駒価値の学習のみ対象としているが，この TD 学習法を利用すれば，単に駒価値を学習するだけでなく，その他の評価要素についても調整が可能である．従って本研究で対象としたような，

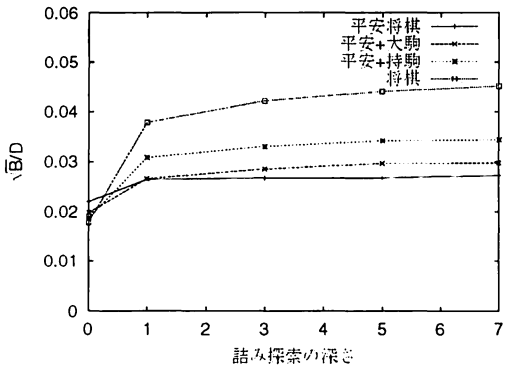


図 9: 将棋の歴史的変種において詰み探索の深さを変えた時の $\frac{\sqrt{B}}{D}$ の変化

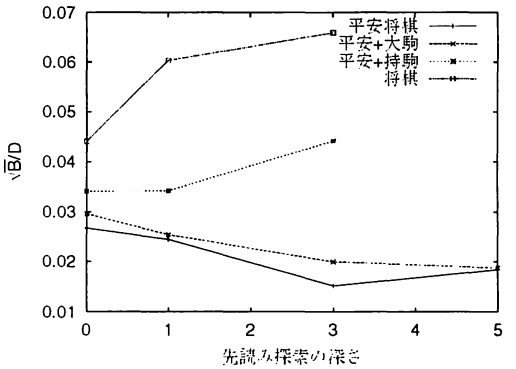


図 10: 将棋の歴史的変種のおいて先読み深さを変えた時の $\frac{\sqrt{B}}{D}$ の変化

既に廃れており，強いプレイヤが存在しないゲームに関するデータを採取する際の標準的な手法として利用可能と考えられ，非常に有用である．しかし，本論文で行なった TD 学習は，駒価値のみの学習で，持駒を区別していないこと，また，平安将棋+持駒に対する実験では，学習が十分に安定していないという問題もあり，まだ試験的な適用という意味合いが強い．

今後，将棋及び平安将棋+持駒については，盤上の駒と持駒の価値を区別して学習を行う必要がある．さらに，廃れたゲームも含めた将棋類一般でゲームのデータを採取する標準的な手法として適用可能となるように駒価値の学習に留まらず，多くの評価要素を含む評価関数の調整にも利用可能となるように本手法を洗練させることがこれからの重要な課題である．

参考文献

[1] H. Iida, N. Sasaki, and T. Hashimoto (1999). "Towards a Classification of Games using Computer Analyses" *Proceedings of Third International Colloquium of Board Games in Academia III*, Firenze, Italy.

[2] 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.

[3] H. Iida and N. Takeshita (2001). "Two-person Chaturanga and four-handed Chaturanga compared using computing analysis" *Proceedings of Fourth International Colloquium of Board Games in Academia IV*, Switzerland. (in press)

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

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

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

[7] 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.

[8] 薄井克俊, 鈴木豪, 小谷善行, (1999). "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 (IPSJ Symposium Series Vol. 99, No. 14)*, pp.31-38.

[9] 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.