TD 法を用いた Amazons の優位距離の学習

太田　喜裕、Reijer Grimbergen

山形大学

概要

Amazons では、「Queen 距離」と「King 距離」を利用した評価値の算出方法が提案されている．本稿では、お互いの持つ Queen 距離と King 距離の差によって更に細かく評価を行う方法を提案し、各距離の差にどの程度の評価の違いがあるかを、TD 学習を用いて確認した．実験の結果、Queen・King 距離共に差が大きくなるに従って各評価値も大きくなった．また、学習値を用いて学習前のプログラムと対局を行わせたところ、結果を向上させる事に成功した．

Learning Domination-Distance of Amazons Using Temporal Differences

Yoshihiro Ota, Reijer Grimbergen

Yamagata University

Abstract

For an evaluation function of Amazons, the features「Queen-Distance」and「King-Distance」have been proposed. In this paper, we propose that evaluation can be done more accurately by using the difference between each player's Queen-Distance and King-Distance. Temporal Difference Learning was used to find the optimal evaluation function values for the difference between Queen-Distance and King-Distance. The experiments showed that a larger difference in Queen-Distance and King-Distance corresponds to a larger evaluation value. Also, self-play experiments showed that the learned values resulted in improved playing strength.

1. はじめに

Temporal Difference Learning（以下 TD 法）によって Amazons の評価関数を適正値に収束させる試みは以前、「TD 法を用いた Amazons の静的評価関数の学習」[1]で行われた．実験の結果、各重みとも発散し望ましい結果が得られなかった．この理由として、チェスや将棋とは異なる Amazons の特徴を考慮に入れていなかったためだと考えられる．今回の研究では学習要素を変えると供に、学習式に工夫を加え Amazons の評価値の重みの学習を行った．

2. Queen・King 距離の説明と問題点

現在、Amazons プログラムの評価関数には、「Queen 距離」と「King 距離」という 2 つの概念を利用した評価方法が広く用いられている．これは、「An Evaluation Function for the Game of Amazons」[2]内で提案された評価方法である．Queen 距離とは Amazons の駒がチェスの Queen と同じ動きをした場合、そのマスまで何手で行けるかを表した値である．一方、King 距離とは Amazons の駒がチェスの King と同じ動きをした場合、そのマスまで何手で行けるかを表した値である．図-1 に Queen 距離を、図-2 に King 距離の例を示す．各マスにおける左上の数字が白駒の、右下の数字が黒駒の Queen・King 距離を表しており、駒が到達できないマスは∞で表している．

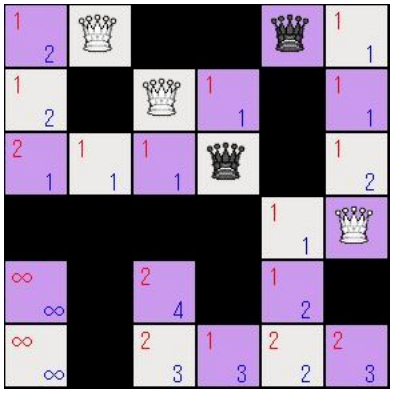


図-1　Queen 距離

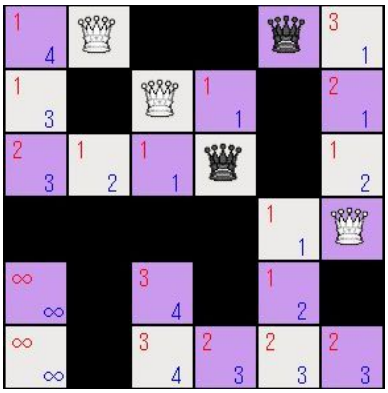


図-2　King 距離

[2]の論文中では、各マスにおける距離の差を利用し評価値の算出を行っていた．以下に現在が白の手番だと仮定した場合の評価値の計算方法を示す．

$$\Delta(\text{白の距離、黒の距離}) = \begin{cases} 0 & \text{if 白の距離} = \text{黒の距離} = \infty \\ k(\leqq 0.2) & \text{if 白の距離} = \text{黒の距離} < \infty \\ 1 & \text{if 白の距離} < \text{黒の距離} \\ -1 & \text{if 白の距離} > \text{黒の距離} \end{cases}$$

（点数　／　各空マスでのお互いの距離の差）

この評価関数の問題点として、距離の差がどの様な場合であっても、「白の距離 ＜ 黒の距離」ならば＋1、「白の距離 ＞ 黒の距離」ならば－1 が必ず評価値に加算される点である．そのため、本研究では「優位距離」を用いて更に正確に評価が行えるようにした．優位距離とは相手の Queen・King 距離と比較してあるマスまで何手の差で到達できるかを表す数字である．優位距離が＋1 ならば、相手より 1 手早くそのマスに到達でき、－1 なら相手より到達が一手遅くなる事を表している．この優位距離の差に従って更に細かく評価を行うと共に、各優位距離がどの程度の評価値を持っているかを、TD 法を用いて自動学習させ、その差を確かめた．

3. TD 法による評価関数の学習

3.1 TD 法の学習式

今回の実験では以下の式によって評価値の学習を行った．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．

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

予想確率Ｐは以下の式によって与えられる．

$$P(V) = \frac{1}{1 + e^{\frac{-V}{\omega}}} \qquad (2)$$

Ｖは評価関数の評価値で、ωはシグモイド関数の傾きを調整する値である．Ｖは以下の式の様によって与えられる．

$$V = \sum_{distancetypes} W_{type} C_{type} \qquad (3)$$

Wは各優位距離の重み、type は優位距離の種類、Ｃは評価要素の特徴量を表しており、以下の式によって与えられる．

$$C_{type} = \frac{\text{（味方の優位距離 type の合計）} - \text{（相手の優位距離 type の合計）}}{\text{（優位距離を持つマスの合計）}} \qquad (4)$$

優位距離 type の差を優位距離を持つマスの合計で除算を行うのは、Amazons の特徴を考慮に入れたためである．Amazons では一手打つ度に盤上に「矢」を配置し、各駒の移動可能位置が制限されていくため、優位距離が常に減少していく．そのため優位距離の差のみを使用すると、予想確率 P が正確に算出されず、重みが上手く学習できないという問題があった．式(4)では、優位距離を持つマスの合計との比をとる事によって時間 t と t+1 での予想確率を正確に算出できるようにした．(1)～(4)までの式を用いて各優位距離の重み、例えば相手との距離の差が＋1 の場合に与える点数、＋2 の場合に与える点数を学習させた．

4. TD 法による優位距離の学習実験

4.1 学習させる優位距離

研究を行うに当り、学習させる優位距離の種類を決めるため、100 戦を行い各優位距離がどの程度の割合で現れるか実験を行った．その結果を表-1、表-2 に示す．

表-1　Queen 距離の優位距離出現割合

優位距離の種類	出現割合（%）
領域	28.31
優位距離±0	29.30
優位距離±1	34.57
優位距離±2	5.21
優位距離±3	1.58
優位距離±4～±9	1.03

表-2　King 距離の優位距離出現割合

優位距離の種類	出現割合（%）
領域	28.31
優位距離±0	19.21
優位距離±1	32.32
優位距離±2	11.85
優位距離±3	4.54
優位距離±4～±14	3.77

Queen・King 距離共に、領域と優位距離の値が±0～±3 までが全体の 96%以上を占めており、これらが評価を行う上で重要な優位距離である事が分かる．そのため、領域と Queen・King 距離それぞれの優位距離の差 0、1、2、3 以上の 9 つの段階に分け、各重みの学習を行う事にした．

4.2 実験に使用したシステムの説明

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っている．評価関数には Queen・King 距離のみを使用し、4.1 に示した様に、自分と相手の距離の差に従い 9 つの段階に分け点数を与えた．

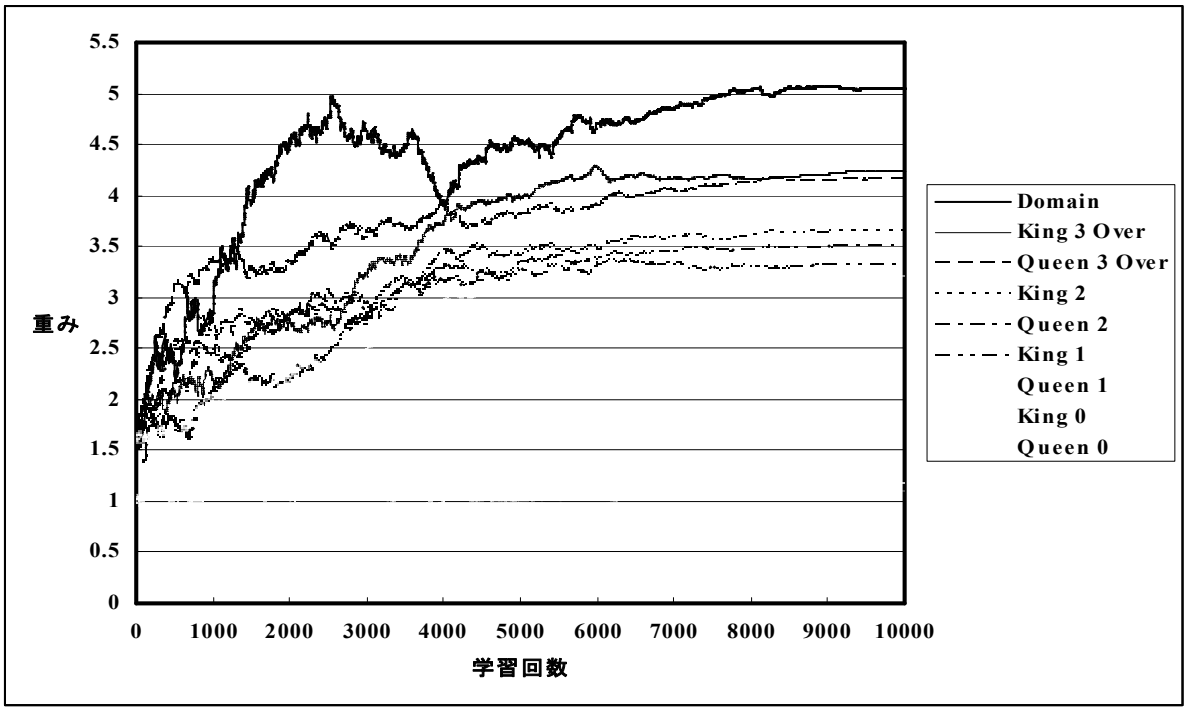
4.3 実験方法

4.2 で述べたシステムに 3 節で述べた TD 法による学習システムを組み込み、実験を行った．探索は白黒共に一手 1 秒で行い、Opening Book 等の特別な知識は使用していない．また、評価関数に小さな乱数を加え、同じ手順の対局を繰り返さないようにし、重みの調整は 1 つのゲームが終了する度に行った．

実験では 10000 回の対局を通し、各重みの学習を行った．各初期値は、学習する優位距離の重みは全て 1.0、ω の値は 1.0、λ の値は 0.95、学習値 α は初期値 0.1 とし、学習値は対局が進むに従い徐々に減少させた．

5. 実験結果

実験は学習 A、B、C で各 10000 回の対局を行った．図-3 に学習 A の結果を示す．



図−3　優位距離の重みの学習の様子

表-3 には各学習結果の収束値を示す．

表−3 各学習結果の収束値

	Queen 0	King 0	Queen 1	King 1	Queen 2	King 2	Queen 3Over	King 3Over	Domain
学習 A	1.091	1.169	3.205	3.312	3.507	3.656	4.155	4.244	5.342
学習 B	1.123	1.158	3.481	3.331	3.564	3.683	4.236	4.112	5.865
学習 C	1.135	1.211	3.455	3.674	3.212	3.555	4.711	4.789	6.105

表-4 には、Queen 0 を 1.000 にした場合の各優位距離の重みの比を示す．

表−4 各学習結果の収束値

	Queen 0	King 0	Queen 1	King 1	Queen 2	King 2	Queen 3Over	King 3Over	Domain
学習 A	1.000	1.071	2.938	3.036	3.214	3.351	3.808	3.890	4.896
学習 B	1.000	1.031	3.099	2.966	3.174	3.280	3.772	3.662	5.223
学習 C	1.000	1.067	3.044	3.237	2.830	3.132	4.151	4.219	5.379

6. 対局結果

表-5 には学習した優位距離の重みを使用したプログラムを、優位距離 0 ならば±0.2、他の優位距離の評価値の重みは±1.0 に固定した学習を行っていないプログラムと対局を行わせた．対局では、探索は白黒共に一手 1 秒で、α β法と反復深化を用いた全幅探索で行った．また、Opening Book 等の特別な知識は使用していない．その結果を表-5 に示す.

表−5　学習した優位距離の重みを用いた対局結果

	対局回数	勝ち	負け
学習 A	300	168	132
学習 B	300	164	136
学習 C	300	166	134

表-6 には学習した Queen の優位距離の評価値のみを、表-7 には学習した King の優位距離の評価値のみを使用し、学習を行っていないプログラムと対局させた結果を示す.

表−6　学習した Queen の優位距離の重みのみを用いた対局結果

	対局回数	勝ち	負け
学習 A	300	153	147
学習 B	300	143	157
学習 C	300	147	153

表−7　学習した King の優位距離の重みのみを用いた対局結果

	対局回数	勝ち	負け
学習 A	300	155	145
学習 B	300	152	148
学習 C	300	140	160

7. 結論

学習結果を見ると収束値は、Queen・King 0、Queen・King 1、Queen・King2、Queen・King 3Over、領域、という様に 5 つの段階に別れ、重みが次第に高くなっている．これは、「優位距離の大きいマスを多くする＝多くの領域を獲得する事に繋がる」という事を学習した結果だと考えられる．対局でも、各優位距離の重みが明確になっているため、各優位距離の境界が曖昧な学習前のプログラムと比較し、盤面の自分の優位・不利を正確に判断できるようになっていた．Queen と King の学習した優位距離の重みを別々に使用した対局では、Queen と King のお互いの評価値のバランスが上手くとれていない事が原因で、学習前のプログラムと比較してそれ程違いが出なかった．

8. おわりに

本稿では、コンピュータ Amazons における優位距離の評価値の重みを、TD 法を用いて学習する実験について述べた．今後の課題として、探索を深くして手を打った結果を学習に反映させる事である．今回は、時間の関係で、一手 1 秒で手を打ち学習を行っており、特に序盤で取得する特徴量の値が正確に取得できていない場合がある．そのため、探索深さ 3 で同じ様に学習を行い、より正確な学習結果を出したい．また、ゲームの進行段階によって各優位距離の重みが異なるので、進行度に従いゲームを何段階かに分け、それぞれの進行度の場合に学習値がどの様な値に収束するか確かめたい．

参考文献

[1] 西條 良輔, 鈴木 豪, 小谷 善行, “TD 法を用いた Amazons の静的評価関数の学習”, Game Programming Workshop’99, pp.105-108, 1999.

[2] Jens Lieberum, “An Evaluation Function for the Game of Amazons”, In H.J.Van den Herik, H.Iida, and E.A.Heinz, (Eds.), Advances in Computer Games vol.10, Kluwer Academic Publishers, Boston, USA, pp.299-308, 2003.

[3] 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”, Game Programming Workshop’99, pp.31-38, 1999.

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

[5] D.F.Beal and M.C.Smith, “First Results from Using Temporal Diffrence Learning in Shogi”, Computer and Games Conference, pp.113-125, 1998.