TD 法を用いた Amazon の静的評価関数の学習

西條 良輔
jo@fairy.ei.tuat.ac.jp

鈴木 豪
go@fairy.ei.tuat.ac.jp

小谷 善行
kotani@cc.tuat.ac.jp

東京農工大学

概要

今回我々は，Amazon ゲームにおける評価関数と学習について提案する．Amazon とは 2 人対戦型ゲームで，Amazon の勝敗は，どちらかの可能手の消失で決まる．そのため，序盤は可能手の多さの有効性が考えられる．自分が相手より先に到達できる領域も重要である．両者の支配空間の大きさと，同じく両者の可能手の多さの 4 つの評価に，どのような重みを与えるとうまく動作するかを，TD 法を用いて検証する．

Evaluation Function of Amazon with the Temporal Difference Learning

Ryosuke SAIJO
jo@fairy.ei.tuat.ac.jp

Tsuyoshi SUZUKI
go@fairy.ei.tuat.ac.jp

Yoshiyuki KOTANI
kotani@cc.tuat.ac.jp

Tokyo Univ. of Agri. and Tech.
2-24-16 Nakamachi, Koganei, Tokyo, JAPAN

Abstract

We propose an evaluation function for amazon game with Temporal Difference Learning. Outcome of amazon is decided by disappearing of possible move. We think number of possible move is effective for opening game. The place that is reached smaller move than enemy is important. We inspect successfully move used by Temporal Difference Learning for how many weight given for size of that place and number of possible move.

1 はじめに

Amazonゲームとは2人型対戦ゲームである．駒はチェスのクイーンと同じ動きをとり，移動先からクイーンが到達可能なマスに，石を置く．この動作を交互に繰り返し，先に動けなくなったほうが負けである．駒は，駒や石を飛び越えて移動することはできない．

今回，Amazon ゲームにおける可能手の数と自分の支配領域に着目した評価関数について考えた．

2 可能手について

Amazon ゲームでは可能手がなくなることが負けとなる．したがって他に評価すべき要素がない場合，可能手の多さを評価することは有効である．特に序盤は盤上の石の数も少なく，お互いが張り合っている状態が多いので，この評価が有効であると思われる．

2.1 可能手の評価関数

1 手先読みをおこない，各局面で自分および相手の可能手を評価する．ここでは，自分の駒が何通りの動きを実現できるかをカウントする．石の設置を含めた可能手は考えない．また，同じ局面での相手の可能手も同じように評価する．相手の可能手の評価は次の相手の手番ですぐに反映されるが，自分の可能手の評価は，相手の手番での手によっては評価値どおりの可能手がなくなることがある．

3 支配領域について

自分だけが 1 手で到達できるようなマスを，自分が支配している領域とみなし，その多さを評価するのが，支配領域の評価である．双方が 1 手で到達できるマスや，到達するのに 2 手以上必要な領域は競合領域とする．

3.1 支配領域の評価関数

1 手先読みをし，各局面で自分および相手の支配領域の広さを評価する．1 種類の駒しか存在しない閉じた領域もその駒の支配領域と考えられるが，今回は 1 手で到達できるマスしかカウントしていない．

図:1 は，Amazon 初期配置における支配領域の例である．

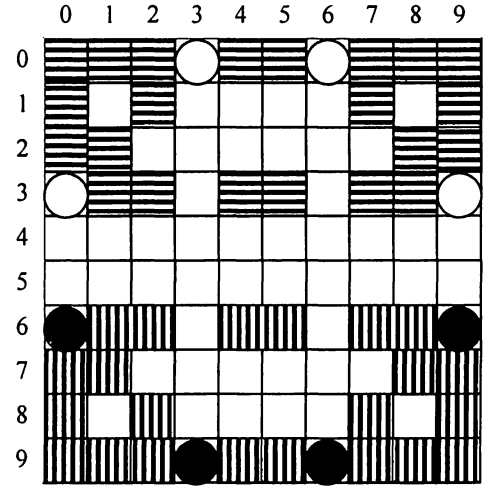

図:1 初期配置における支配領域
横縞が○の支配領域
縦縞が●の支配領域

4 評価式

自分の可能手の数を評価する評価関数を F1，相手の可能手の数を評価する評価関数を F2，自分の支配領域を評価する評価関数を F3，相手の支配領域を評価する評価関数を F4 とする．また，各関数にかかる重みを w1，w2，w3，w4 とし，次のように評価式 F を設定する．

$$F=(w1\times F1)-(w2\times F2)+(w3\times F3)-(w4\times F4)$$

自分が負けたときの局面では，F1，F3 が 0 になり，F2，F4 が少なくとも 0 以外なので，評価値は負の値になる．自分が勝ったときの局面での評価値は正の値になる．

5 Temporal Difference Learning (TD 法)

設定した評価式を，次のような更新式を用いて更新する．

$$w \leftarrow w + \sum_{t=1}^{m} \Delta w_t$$

$$\Delta w_t = \alpha\left(w^T f_{t+1} - \frac{1}{1+e^{-(W_t \bullet F_t)}} \right) \sum_{k=1}^{t} \lambda^{t-k} \nabla f_k$$

f は評価値をあらわし，t は手を単位とする時間である．W および F はそれぞれ w, f を成分とするベクトルであり，$W \cdot F$ はその内積である．

α は学習パラメータである．λ は，0 に近づくほど過去の評価を考慮する．λ=1 のときは過去の評価を考えない．

6 実験

TD 法において，λ=1 とすると，LMS 法の更新式となる．今回，予備実験として LMS 法での w1～w4 の収束を調べた．また，学習パラメータ α は，0.01，0.10，0.20 を使用した．結果を図:3，図:4，図:5 に示す．

いずれの図も，重みは発散している．

原因として，評価値が考えられる．勝負に勝ったとき，負けたとき，その 2 つの評価値にはっきりした差がなければ学習はうまく行われない．評価関数の調整が必要である．

7 今後の展望

評価関数と重みの関係をもう少し工夫し，学習がうまく行われるようなものを検討中である．調整が終われば，TD 法と LMS 法の比較実験を行う予定である．

また，次のような作戦も思いついているので，順次実現していきたい．

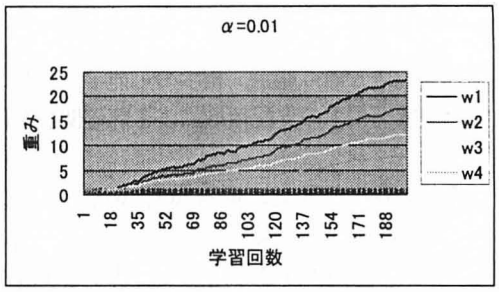

図 2: α=0.01 のとき

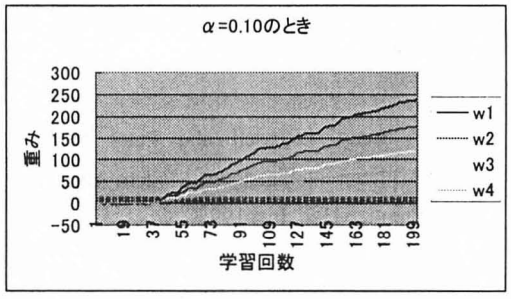

図 3: α=0.10 のとき

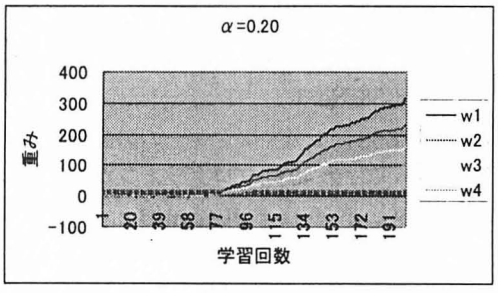

図 4: α=0.20 のとき

・閉じた領域の判定

閉じた領域について評価を行う．閉じた領域内に自分の駒だけが存在し，相手の駒がないときは，他の駒の可能手が無くなるまで動かさない．また，閉じた領域内で場所取りが展開されているときは，どの領域での場所取りを優先させるかという評価関数を用いて優先度を設ける．

・先読みの深さ

Amazon ではゲームが終盤に近づくにしたがって可能手の数が少なくなるので，ゲームの進行状況に応じて先読みの深さを変化させる．特に，終盤では広さ，深さともにくまなく探索することが可能だと思われる．

8 まとめ

今回，自分の可能手，相手の可能手，自分の支配領域，相手の支配領域による評価関数を用意した．

また，各評価要素に重み w1，w2，w3，w4 を掛けた評価式を提案した．設定した評価式の重みについて，LMS 法を用いて更新する実験をおこなったが良い結果は得られなかった．

※ 参考文献

[1] Stuart J.Russell and Peter Norvig : Artificial Intelligence A Modern Approach, Prentice-Hall, Inc. 1995

[2] 安西祐一郎：認識と学習，岩波講座ソフトウェア科学 15，1996

[3] Gerald Tesauro : Temporal Difference Learning and TD-Gammon , Communications of the ACM, March 1995/ Vol. 38, No. 3.

[4] 小谷善行：アマゾンゲームの評価関数について，第 40 回プログラミングシンポジウム，1999.1，pp104-pp105