

「芝浦将棋」のチーム紹介

2011年3月18日

芝浦工業大学工学部情報工学科

五十嵐 治一, 山本一将

1. はじめに

本稿は、第21回世界コンピュータ将棋選手権(2011年5月)に出場予定の「芝浦将棋チーム」の紹介文です。本チームは本学情報工学科の学生と教員により構成されており、教育と研究の一環として活動しています。本チームのベースとなっているのは、保木邦仁さん(現在、電気通信大学教員)が開発し、インターネット上でソースコードを公開している“Bonanza”(http://www.geocities.jp/bonanza_shogi/)です。このBonanzaでは、局面の評価関数中に含まれる大量のパラメータの値を、独自の教師付学習の手法によりプロ棋士の公式対局の棋譜から自動学習しています[1]。しかし、教師付学習では教師データとして採用した棋譜データにより強さが左右されてしまいます。また、プロ同士の棋譜と言っても指し手すべてが必ずしも最善のものであるとは限りません。そのため教師付学習では限界があり、コンピュータプログラムの棋力が最終的に教師である人間(プロ棋士)を超えるのは難しいのではないかと我々は考えました。

そこで、本チームでは、これらのBonanzaのパラメータ値を初期値として、強化学習の手法により、さらに追加的に学習することを試みました。この追加学習では、Bonanza同士を対局させて、一方のBonanzaに対しては強化学習の代表的手法の一つである最急降下TD(λ)法によりパラメータ値を学習させました。以下では、その追加学習の内容について簡単に紹介します。なお、昨年度の芝浦将棋のバージョンでは、アルゴリズムやプログラムにおいて、いくつもの誤りやバグがありました。今年度のバージョンでは、それらはかなり修正されています。しかし、本格的な学習実験や、学習で得られた評価関数の評価実験についてはまだ十分に行うまでに至っておりません。まだまだ開発途中あり、5月の2011年度選手権大会へ向けて、できる限りいろいろなことを試してみるつもりでいます。

2. 最急降下TD(λ)法を用いた予測勝利確率の学習

本章以降では、強化学習の一種である最急降下TD(λ)法を用いて、将棋の手番局面において予測勝利確率を与える近似関数の学習法について説明します。詳細は、文献[2][3]等をご覧ください。本紹介文における2~4章の説明は文献[2]の要約です。

将棋において自己の t 回目の手番の局面を状態 s_t とし、終局時($t=L$)における勝敗を z (勝てば $z=1$, 負ければ $z=0$)で表すことにします。時刻 t は手番ごとに1ステップずつ経過するものとします。ここで、各時刻 t に与える報酬 r_t を、 $t=L$ においては $r_t=z$, それ以外の時

刻では $r_t=0$ とします。このとき、局面 s において学習プログラムが勝利する確率の予測値 $P^\pi(s) \equiv E_\pi[z|s_t=s]$ (以下、予測勝利確率) を定義します。 π は指手の選択方法で方策と呼ばれ、 $E_\pi[x]$ は方策 π に従ったときの確率変数 x の期待値を表しています。また、評価関数中の重みパラメータ ω の更新 (=学習) は学習プログラムの手番ごとに行うものとします。

さらに、時刻 t における予測勝利確率 $P^\pi(s)$ の近似関数を $P_t(s; \omega_t)$ と表すことにします。バックギャモンで大成功を収めた TD-Gammon[4] では階層型のニューラルネットワークモデル (ω はニューロン間の結合重み) が用いられましたが、将棋では以下のシグモイド関数が用いられています[5][6]。

$$P_t(s; \omega_t) = 1 / (1 + e^{-E(s; \omega_t)/\tau}) \quad (1)$$

ここで、 $E(s; \omega)$ は局面 s の静的評価関数、 ω は評価関数中の重みパラメータです。ただし、将棋において最初に(1)を用いた文献[5]では τ は用いられていません。芝浦将棋では、予備実験として Bonanza 同士の対戦を行った結果、 Bonanza ver.4.1.3 同士の予測勝利確率を(1)のシグモイド関数で近似する場合は、 $\tau \sim 1250$ 程度であれば良いことが分かりましたので、今のところこの値を採用しています。

さて、(1)の評価関数をどう設定し、その中の重みパラメータをどのような学習則で学習していくのが問題です。以下の章で説明します。

3. Bonanza の評価関数を利用

Bonanza ver.4.1.3 の評価関数を $E_B(s)$ とすると、次の式で表すことができます。

$$E_B(s; \omega) = \sum_{j=1}^N \omega_j [x_j(s^1) - x_j(s^2)] \quad (2)$$

ただし、関数 x_j は特徴量 j が局面に現れているときに 1、それ以外は 0 をとります。 Bonanza ver.4.1.3 では、各駒の価値と、2種類の3駒の位置関係 (①自分の王1駒と、相手の王を除く2駒の計3駒、②自分と相手の王の2駒と、自分の1駒の計3駒) を局面 s の特徴量と考え、評価関数は各特徴量の線形和で表されています。なお、2駒の位置関係は①の中に含まれています。

また、(2)の右辺において、 s^1/s^2 は局面 s における先手/後手側から見た駒配置です。(2)の定義から、先手側が優勢であるときには $E_B > 0$ となります。したがって、4章の学習則を用いる際には、学習プログラムが先手であるときには、 $E(s) = E_B(s)$ とし、後手であるときにはマイナス符号を付けて $E(s) = -E_B(s)$ として用います。

4. 重みパラメータ ω の学習則

(1)の形の近似関数を強化学習における状態価値関数と見なすと、最急降下 TD(λ)法によ

り， ω の学習則として次の更新式を得ることができます。

$$\omega_{t+1} = \omega_t + \alpha \delta_t e_t \quad (3)$$

ただし， δ_t と e_t は次の式で，学習プログラムの手番ごとに計算することができます。

$$\delta_t = r_{t+1} + P_t(s_{t+1}; \omega_t) - P_t(s_t; \omega_t) \quad (4)$$

$$e_t = \lambda e_{t-1} + \nabla_w P_t(s_t; \omega_t) \quad (5)$$

$$= \lambda e_{t-1} + [1 - P_t(s_t; \omega_t)] P_t(s_t; \omega_t) \delta E / \partial \omega_t \quad (6)$$

5. 問題点

実際に，2～4章で述べた最急降下TD(λ)法の学習則をBonanza ver.4.1.3へ組み込んで，学習させてみると，以下の問題点や疑問点が現れました。

- ・学習のための対局時に学習プログラム側の探索量が大幅に低下する
- ・公式試合と同じ持ち時間（25分切れ負け）という環境下で学習しようとする，かなりの計算時間がかかってしまう
- ・評価値と勝率の曲線を(1)のシグモイド関数で近似するのが最適であるという保証はないのでは？
- ・予測勝利確率の近似関数の精度向上が本当に棋力向上につながるのか？
- ・予測勝利確率の近似関数の精度向上以外の学習目的には無力ではないか？

上記の問題点に対して，現在の芝浦将棋では，プログラム上の改良で学習計算の速度向上を図ることや，学習時の対局相手を自分よりも強い相手に徐々に変えていくことで棋力向上を試みています。また，学習目的を自由に選択できる他の強化学習法として，「方策勾配法」という学習法の適用も長期的な開発方針として検討しています。

6. おわりに

芝浦将棋の生い立ちや，昨年度（2010年）の選手権大会への参加の様子，学習理論の詳細，今後の展開について，コンピュータ将棋協会の会誌の中でまとめさせていただきました[3]。ご興味がある方はそちらも併せてご覧頂ければ参考になると思います。

最後になりましたが，芝浦将棋は強化学習を中心として棋力向上を目指すことを基本方針としています。芝浦将棋の理念，方向性にご賛同頂ける方であれば，学生，社会人の如何を問わず，どなたでも歓迎いたします。参加，協力をご希望の方は，arashi50@sic.shibaura-it.ac.jpまでご連絡下さい。よろしくお願いいたします。

参考文献

- [1] 保木 邦仁, “局面評価の学習を目指した探索結果の最適制御”, 第 11 回ゲーム・プログラミングワークショップ, pp.78-83(2006).
- [2] 五十嵐治一, 山本一将, “コンピュータ将棋への TD(λ)法の適用:Bonanza の評価関数パラメータ値”, 情報処理学会第 73 回全国大会講演論文集, 講演番号 3C-3, 第 2 分冊, pp.5-6 (2011 年 3 月 2-4 日, 東京).
- [3] 五十嵐治一, ”教育・研究プロジェクト「芝浦将棋」の展望“, コンピュータ将棋協会誌, Vol.22 (2011 年 4 月発行予定)
- [4] Richard S.Sutton, Andrew G.Barto(著), 三上 貞芳, 皆川 雅章(訳), 「強化学習」, 森北出版, 第 8 章, 2000.
- [5] D. F. Beal and M. C. Smith, “Temporal difference learning applied to game playing and the results of application to shogi,” Theoretical Computer Science, Vol.252, pp.105-119 (2001).
- [6] 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”, ゲームプログラミングワークショップ'99, pp.31-38(1999).