

# 「芝浦将棋」のチーム紹介

2010年3月24日

芝浦工業大学工学部情報工学科

五十嵐 治一, 黛 恵輔

## 1. はじめに

本稿は、第20回世界コンピュータ将棋選手権（2010年5月）に出場予定の「芝浦将棋チーム」の紹介文です。本チームは黛恵輔が4年次の卒業研究として行った研究の成果[1]を取り入れたコンピュータ将棋プログラムです。本チームのベースとなっているのは、保木邦仁さんが開発し、インターネット上でソースコードが公開されている“Bonanza” ([http://www.geocities.jp/bonanza\\_shogi/](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。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、コンピュータが最終的に人間（プロ棋士）を超えた棋力を身につけられるという保証はな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、これらのパラメータの値を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 TD( $\lambda$ )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。

## 2. TD( $\lambda$ )法の組み込み

これまでに TD( $\lambda$ )法による「駒の価値」や「王の安全度」を学習する実験は行われていましたが[2]、強いプログラムを作ることまでには成功していなかったようです。この原因の一つには、評価関数中の評価項目が少なかったことがあるように思われます。一方、Bonanza の評価関数では、駒の位置関係（2項関係）に対するパラメータが大量に導入されています。かつ、強さには実績がありますので、今回はこの Bonanza の評価関数の形式をそのまま使用することにしました。

TD( $\lambda$ )法によるパラメータの更新式は(1)のように与えられます [2][4].

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

ここで、 $W_t$  は時刻  $t$  における評価関数の重みベクトルであり、 $W = (w_1, w_2, \dots, w_n)$  と表します。Bonanza ではパラメータ数  $n$  は1万を超え、各  $w$  の値は約-30000~18000の値を取っていま

す。また、 $\alpha$ は学習係数、 $r$ は報酬です。強化学習は局面の優劣からシステム設計者が報酬を与え、その結果を評価関数中のパラメータへフィードバックするという方式を取ります。ここで重要なのが、今着手したその手の善し悪しをその場で報酬として与えるのではなく、局面が進んで優劣がはっきりしてから与えること（遅延報酬）が強化学習では可能です。前者は即時報酬と言い、Bonanzaの学習方式はこの即時報酬を与える方式でした。さらにBonanzaのような教師付学習では、報酬の代わりに正解手を教師データとして与える必要がありました。それに対して強化学習の場合は、その場で正解手を与える必要はありません。局面の優劣がはっきりしたところでその局面の優劣の度合いを数量化して与えてやるだけで良いのです。

さて、(1)の $P_t$ は時刻 $t$ の局面から推測した勝敗の予測確率です。ある時点での勝利の予測確率を正確にするために、それ以降の時間に出現した局面の勝利予測確率を参考にしますが、遠い将来の局面ほど参考にする程度を割り引くことにします。その割引率が $\lambda$ です。

また、(1)の $\nabla_w P_i$ は各パラメータ $w_i (i=1,2,\dots,n)$ による勾配ベクトルを表し、

$$\nabla_w P_i \equiv \left( \frac{\partial}{\partial w_1} P_i, \frac{\partial}{\partial w_2} P_i, \dots, \frac{\partial}{\partial w_n} P_i \right) \quad (2)$$

と定義されています。

本実験では[2]と同様に、 $P$ を局面 $K$ の評価値 $E(K)$ によって次の式(3)で表します。

$$P(K) = 1 / \left( 1 + e^{-\frac{E(K)}{\tau}} \right) \quad (3)$$

ここで $E(K)$ として、Bonanzaの局面評価関数[3]、

$$E(K) = \sum_{j=1}^N w_j x_j(K) \quad (4)$$

を利用します。右辺の $x_j$ は局面 $K$ において特徴 $j$ の要素が含まれている場合は1を、含まれていない場合は0をとる2値関数です。 $w_j$ は重み係数で、これを学習により定めます。

### 3. 学習実験

本研究ではBonanza同士を50試合対局させる学習実験を行いました。一方のBonanzaの評価値は初期値に固定し、学習側のBonanzaは、初期値を2. で述べたTD( $\lambda$ )法を用いて学習させました。思考時間は両者ともに3秒に固定しました。報酬 $r$ については、対局中は $r=0$ 、終局時に勝ったら $r=1$ 、負けたら $r=-1$ を与えると言ったかなりシンプルなものを設定しました。学習係数 $\alpha$ については、いくつかの値を設定してそれぞれ学習実験を行いました。

さて、学習の効果についてです。学習後のBonanzaを未学習のBonanzaと、学習時と同様に思考時間3秒という条件で対局させました（対局数50）。その結果、学習係数 $\alpha$ を適切

に調整すると、勝率を約 6 割にまでアップできることがわかりました。まだ、評価の際の対局数や、思考時間も限られた場合しか実験していません。これが本大会の試合条件下での学習実験や、Bonanza 以外のチームとの対局による学習実験がもっと必要になると考えています。

#### 参考文献

- [1] 黛 恵輔, “コンピュータ将棋 Bonanza の局面評価関数の調整 : TD( $\lambda$ )法の適用”, 芝浦工業大学工学部情報工学科 2009 年度卒業論文概要集, pp.177-178(2009).
- [2] 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).
- [3] 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).
- [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.