「芝浦将棋」のチーム紹介

2012 年 3 月 20 日
芝浦工業大学情報工学科
五十嵐 治一，山本一将，川内博世，濱村綾

1. はじめに

本稿は，第 22 回世界コンピュータ将棋選手権（2012 年 5 月）に出場予定の「芝浦将棋チーム」の紹介文です．本チームは芝浦工業大学工学部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．本大会でも CSA 公認の使用可能ライブラリとして登録されている Bonanza Ver. 6.0.0 を使用しています．

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着手選択や強さが左右されてしまいます．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，教師付学習では評価関数の学習に対して何らかの限界があるのではないかと我々は考えています．

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これまで試みてきました．具体的には，Bonanza 同士の対局において，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．この学習結果を持って，第 20 回大会(2010)と第 21 回大会(2011)に参加して来ました．今回は，これまで適用してきた TD(λ)法ではなく，その発展版である TDleaf(λ)法による学習を行うなど，いくつかの点で改良を試みました．以下の章でそれらの特徴を簡単にまとめておきます．

2. 本年度の「芝浦将棋」の特徴

本年度の芝浦将棋の特徴は，以下の１）～６）のようにまとめることができます．

１）最急降下 TDleaf(λ)法を用いた評価関数パラメータの強化学習：

従来用いてきた TD(λ)法では学習時の出現局面を利用しましたが，TDleaf(λ)法では最善応手手順（principal variation）の末端局面（leaf または principal position）を利用するので，予測勝利確率のより正確な近似関数を得る可能性があると期待されます．これまでに，Beal と Smith が TDleaf(λ)法を将棋へ適用した例があります．そこでは学習時の読みの深さはあまり必要ないと結論付けています[2]．これは TDleaf(λ)法に対しては否定的な結論です．し

かし，彼らの実験では評価関数に用いられたのは駒の価値だけで，読みの深さも 4 までででした. 一方, チェスにおいては Baxter らにより TDleaf(λ)法の有効性が報告されています[3]. 我々は学習時にはなるべく深い探索が有効であろうと考えています.

２）序盤定跡の不使用：

学習時ならびに大会対局時には定跡データベースを使用しません．これは，序盤の定跡局面以外の局面も積極的に学習に貢献させたいと考えたからです．もし，従来定跡になかった新手らしきものが得られるようであれば，コンピュータ将棋の一つの成果と言えるのではないかと期待しています.

３）学習時の対局相手について：

学習時の対戦相手は，Bonanza Ver. 6.0.0 ですが，評価関数の値に正規分布に従うランダムノイズを加えています．これは対戦相手の着手傾向の偏りを少なくすることにより，特定の相手だけに強くなること（過学習の弊害）を防止するための対策です.

４）学習条件について：

学習時に必要な各種パラメータは予備実験により適切と思われる値を決定しています. また，学習時の対局時間は大会ルールと同じく２５分切れ負けとする予定です．しかし，かなり学習時間がかってしまいますので，これよりも短縮して，その分学習対局数の方を増やすことも検討しています.

５）探索エンジンについて：

現在，Bonanza Ver. 6.0.0 を参考にして，ゼロから探索エンジンのプログラムを書き上げているところです．もし大会までに間に合わない場合には，Bonanza Ver. 6.0.0 の探索エンジンを使用します.

６）評価関数について：

評価関数の形式は Bonanza Ver. 6.0.0 のものを採用しています. パラメータの値は Bonanza Ver. 6.0.0 の値を学習の初期値として使用していますが，１）で記したように TDleaf(λ)法により学習しています．ただし，駒の価値は初期値のままです．大会までには，駒の価値についても学習することを検討しています.

３. 最急降下 TDleaf(λ)法を用いた予測勝利確率の学習

本章では，強化学習の一種である最急降下 TDleaf(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. TDleaf(λ)法と TD(λ)法との相違点は，学習則において利用する局面の違いだけですので，最初に最急降下 TD(λ)法

の説明をします．詳細は文献[4][5]等をご覧頂くことにして，以下では文献[4]の要約だけを記します．

将棋において自己の t 回目(t≧0)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 $P_t(s_t; \omega_t)$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[6]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[2][7].

$$P_t(s_t;\omega_t) = 1\Big/\left(1+e^{-E(s_t;\omega_t)/\tau}\right) \tag{1}$$

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[2]ではτは用いられていません．芝浦将棋では，予備実験で適当な値を設定しています．また，TDleaf(λ)法では，(1)式の s_tが対局中に出現した局面ではなく，その局面をルートとする探索木の最善応手手順の末端局面(leaf) s_t^*である点が異なります．

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かが問題です．以下の章で説明します．

4．Bonanza の評価関数を利用

Bonanza 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．

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

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.6.0.0 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．

5. 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．

$$\omega_{t+1} = \omega_t + \alpha\delta_t e_t \tag{3}$$

ただし，δ_t と e_t は次の式により，学習プログラム側の手番ごとに計算することができます．

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

$$e_t = \sum_{k=0}^{t} \lambda^{t-k} \nabla_\omega P_k(s_k;\omega_k) \tag{5}$$

$$= \lambda e_{t-1} + \nabla_\omega P_t(s_t;\omega_t) \tag{6}$$

$$= \lambda e_{t-1} + (1/\tau)\left[1 - P_t(s_t;\omega_t)\right] P_t(s_t;\omega_t)\partial E/\partial\omega_t \tag{7}$$

ただし， $e_0 = 0$ と定義しておきます．また，(1)の時と同様に，(4)～(7)において出現局面 s_t を PV の leaf 局面 s_t^* に置き換えると最急降下 TDleaf(λ)法の学習則を得ることができます．

6. 今後の開発方針

「芝浦将棋」はまだまだ開発途上にあります．今後，開発したい点や課題などをいくつかあげておきます．

- TDleaf(λ)法の学習／評価実験を行う．学習局数を増やすとともに，読みの深さを深くした学習実験と，その学習効果を検証するための評価実験を行う必要がある．
- 学習によるパラメータ値の変化を分析する．
- 予測勝利確率の近似関数の精度向上を棋力向上へ結び付けるための検討が必要である．学習時の対戦相手の選択なども重要と思われます．
- 予測勝利確率の近似関数の精度向上以外の学習目的を考える．そのために，TDleaf(λ)法ではなく，「方策勾配法」という強化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．

7. おわりに

芝浦将棋の生い立ちや，一昨年度（2010 年）の選手権大会への初参加の様子，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頂きました[5]．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かと思います．

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．共同研究やチーム開発に参加，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] 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).

[3] J. Baxter, A. Tridgell, and L. Weaver, “KnightCap: A chess program that learns by combining TD(λ) with game-tree search,” Proceedings of the Fifteenth International Conference (ICML '98), pp.28-36 (1998)

[4] 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）．

[5] 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, pp.35-47（2011 年 4 月発行）

[6] 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.

[7] 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).