

GA 将!!!!!!! 第 2 回電王トーナメント PR 文書

2014 年 9 月 25 日 森岡 祐一

1. はじめに

- ◆ GA 将!!!!!!!は私（森岡）がフルスクラッチで作成したコンピュータ将棋ソフトです。
- ◆ 開発の方針としては、評価関数の精度向上を最優先目標としています。
- ◆ 評価関数のパラメータ学習には、強化学習という手法を用います（詳しくは後述）。
- ◆ 最初のバージョンでは GA（遺伝的アルゴリズム）を用いてパラメータ学習を行っていた為、この様な名前にしました。
- ◆ 名前の最後の「!」はバージョン番号です。

2. 目標

- ◆ 棋譜や定跡といったエキスパート（≒プロ棋士）の知識データを用いずに、ほとんど事前知識が無い状態から学習を行い、その結果どのような棋風になるかを知りたいと思って、GA 将!!!!!!!の開発を始めました。
- ◆ 例えば「軽く囲ってから急戦を仕掛けるのがベスト」という結果になり新しい戦法を作り出せるのか、あるいは「矢倉や穴熊に囲うのが良い」という人間の経験に追従する結果になるのか、それが知りたいです。
- ◆ ただし、GA 将!!!!!!!のプログラム（主に評価関数と探索ルーチン）には、私が棋書を読んで知った、将棋固有の知識が含まれています。ですので、将棋の知識が完全に 0 の状態から学習を行う訳ではありません。
- ◆ 余談ですが、多層ニューラルネットワークや SVM (SVR)、ディープラーニングといったアルゴリズムを用いれば、将棋の知識を全く使用せずにコンピュータ将棋を作成出来る可能性があります。ただ、私は評価関数・探索ルーチンに関してはコンピュータ将棋専用のものを用いるのがベストだと考えています。

3. 探索ルーチン

- ◆ 現在、Softmax 実現確率探索という新しい探索ルーチンの実装中です。
- ◆ このルーチンの特徴として、評価関数のパラメータ学習と同様の枠組みで局面遷移確率の学習が可能である事が挙げられます。
- ◆ トーナメントまでにパラメータ学習が間に合うかは微妙ですが、上手く行けば大幅に棋力向上を果たせると見込んでいます。

4. 評価関数

- ◆ 以下の特徴量の線形和で評価値を決める、コンピュータ将棋用としては比較的普通の作りになっています。
 - 駒割の評価
 - 駒の絶対位置の評価
 - 二つの駒の相対位置関係の評価
 - 二つの駒の絶対位置関係の評価
 - 王将の移動可能範囲の評価
 - 駒の自由度の評価
 - 入玉関連の評価
- ◆ パラメータ数は上記全てで 120 万個程です。最近のソフトとしては少ない方かと思えます。

5. 強化学習

- ◆ 評価関数のパラメータ調整には、Bonanza Method 等の既存の手法ではなく、新しく考案した SR-PGLeaf (Split Reward-Policy Gradient with Leaf) というアルゴリズムを用いています。
- ◆ これは、強化学習の既存手法である方策勾配法 (Policy Gradient Method) をベースとして、指し手選択・パラメータ修正部分に α β 探索を組み合わせたものです。
- ◆ SR-PGLeaf のベースとなっている PGLeaf の詳細は、GPW での発表資料に記載してあります。詳しく知りたい方は下記 PDF を参照して下さい。

http://gasyou.is-mine.net/archive/GPW2012_P-3.pdf

- ◆ Bonanza Method 等の「教師あり学習」では、将棋の局面に対して正解となる手が与えられ、それを元に学習します。
- ◆ それに対して、「強化学習」では「正解」は与えられません。その代わり、例えば終局時に勝敗に応じた「報酬」を与えられ、その報酬をより多く得る為にはどうすれば良いか (= コンピュータ将棋の場合なら、勝率を上げるにはどうすれば良いか) を試行錯誤しながら学習を行います。
- ◆ つまり、「教師あり学習」は HOW (どの様に手を選べばいいか) を指示し、「強化学習」は WHAT (何を最大化すればいいか) を指示する学習手法です。「強化学習」においては、HOW は強化学習エージェント (学習を行うプログラム) が、自分で考えて決定します。

6. SR-PGLeaf のアルゴリズム

- ◆ PGLeaf では終局時に勝敗に応じた報酬を与え、これを手掛かりに学習を行います。この為、例えば「中盤まで優勢に進めていたが、終盤で逆転負けした」という場合に、「序盤～中盤では良い手を指していた」とエージェントが判断出来ません。
- ◆ そこで、自己対戦での対局中の評価値の上下を元に指し手の良し悪しを判断させる事にしました。
- ◆ ここで、以下の通り用語を定義します。
 - 確定報酬：終局時に与えられる報酬。勝ちなら+1、負けなら-1、引き分けなら 0。
 - 予測報酬：ある局面から十分な数の対局を行った際に得られる、確定報酬の期待値。ただし、終局時の予測報酬は確定報酬と同じ値とする。
 - 分割報酬：ある局面 P より先の局面の予測報酬と、局面 P での予測報酬の差。
- ◆ 予測報酬ですが、局面の評価値とシグモイド関数を用いて「 $2 * \text{sigmoid}(\text{評価値}, g) - 1$ 」で近似します。シグモイド関数のゲイン g は、対局中の評価値とその対局の確定報酬を全て保存しておき、一定間隔で最適な値に設定します。
- ◆ PGLeaf と SR-PGLeaf の違いは下記の 2 点です。
 1. SR-PGLeaf では、1 回の指し手選択を 1 エピソードとして扱う（1 手ごとに指し手の良し悪しを判断する）。
 2. 1 エピソード終了ごとに分割報酬を強化学習エージェントに与える。
- ◆ 上記の修正により、対局結果にかかわらず評価値の上昇した（＝分割報酬がプラスになった）手は良い手だと判断します。同様に、評価値の下降した手は悪手と判断します。

7. GA 将!!!!!!における強化学習の適用

- ◆ SR-PGLeaf では自己対戦や他のソフトとの対局の結果から学習が可能なので、既存手法と異なり棋譜・定跡等の人間の知識データを必要とせずに学習が可能です。
- ◆ GA 将!!!!!!における SR-PGLeaf では、自己対戦した結果から評価値が上昇した指し手は正例（良い手）、評価値が下がった指し手は負例（悪い手）と判断してパラメータを修正します。これは、人間が将棋を勉強する際の考え方と近いものだと考えています。
- ◆ GA 将!!!!!!では自己対戦を何十～何百万局も行い、その結果をベースに学習（評価関数のパラメータ修正）を行います。
- ◆ 自己対戦・パラメータ修正時には、先手と後手で同じ評価関数を使用します。
- ◆ つまり、人間が自分一人での対局を行い、その結果から学習する様な感じです。
- ◆ 大体 50 万～100 万局程度で学習が収束します。全幅 2 手+静止探索 6 手で自己対戦すると、24 時間あたり 4 万局前後の速度なので、学習が完了するまで 2～4 週間程度かかります。（Core i7 4930K 6 コアマシン使用時。）

8. その他

- ◆ 定跡は一切使用していません。なので、初手から延々数十秒考えます。
- ◆ 数十秒消費した挙句に、単に飛車先の歩を突くだけとか角道を開けるだけとか、そういう風にもなりますので、序盤に関しては自己対戦の棋譜から定跡を構築する等の対策をしたいところです。
- ◆ 詰将棋ルーチンは ABC 探索をベースにした簡単なものを実装しています。10 秒前後の思考時間があれば、20〜30 手程度の詰みを見付ける事は可能です。なので、実戦用の詰将棋ルーチンとしては必要最小限の機能・速度はあります。

9. 棋力

- ◆ 9 月 25 日現在、floodgate の 2 週間レーティングが 1836 です (Athlon 5350 搭載 PC での結果)。
- ◆ 他の思考エンジンとの対局結果ですが、対 ssp (プチ将棋付属の思考エンジン) では勝率 83% です。
- ◆ PGLeaf では対 ssp で勝率 65% 前後が限界だったので、SR-PGLeaf への拡張により大きく棋力が向上しました。

10. おわりに

- ◆ 以前の世界コンピュータ将棋選手権では 5 八玉戦法 (勝手に命名) で二次予選に進出したりと、色々と見ていて面白い将棋を指していました。
- ◆ 今回も観戦者の方に楽しんで頂ける様、頑張って開発してきます。
- ◆ ブログ・Twitter 等もやってます。GA 将!!!!!!に興味を持って頂ければ、こちらの方もどうぞよろしくお願いします。

<http://d.hatena.ne.jp/Gasyou>

<https://twitter.com/MoriokaYuichi>

<http://gasyou.is-mine.net/>