対戦相手に合わせた評価関数の学習システム

加藤 俊明，鈴木 豪，小谷 善行，堤 正義
早稲田大学大学院理工学研究科，東京農工大学大学院工学研究科
東京農工大学工学部，早稲田大学理工学部
{kato,tutumi}@tutumi.phys.waseda.ac.jp,{go,kotani}@fairy.ei.tuat.ac.jp

要約

近年、ＴＤ(λ)法を用いた駒の価値や駒の位置による価値を学習する研究がいくつか行われ、成果を上げている。今までは、いかに人の感覚に近い価値を学習できるかということのみが問題にされて来た。しかし、そろそろ評価関数が自動的に学習できることで何が可能になるかを考えても良いのではないか。本論文では、相手に合わせた評価関数の微調整への可能性を探る。ただし、相手との直接対局で学習を行うのは現実的ではないので、相手の情報を別のシステムにまとめる必要があった。そこで、今回はそのシステムを評価関数に反映させたときに適切な学習が行われるかを確認した。

The learning system of evaluation function against the opponent

Kato Toshiaki, Suzuki Go, Kotani Yoshiyuki, Tutumi Masayoshi
Department of Science and Engineering, Waseda University
Faculty of Engineering, Tokyo University of Agriculture and Technology
Faculty of Engineering, Tokyo University of Agriculture and Technology
Department of Science and Engineering, Waseda University
{kato,tutumi}@tutumi.phys.waseda.ac.jp,{go,kotani}@fairy.ei.tuat.ac.jp

Abstract

Recently, some studies of learning the elements of evaluation function with TD(λ) were made and got to enough results. Till now, we only considered whether the outcome fitted with the thinking of human. But by now, we may be in the place where we can study what we do with the learning of evaluation function. In this paper, we search the way of adjusting evaluation function against the opponent as a new possibility by the learning with TD(λ).

1. 始めに

コンピュータ将棋の強さを決める要素として、ゲーム木の探索と同様に重要な役割を担うのが、局面の有利不利を数値化する評価関数である。ところが、殆どの場合、評価関数は人の経験により作られ、微調整も実際に対戦を繰り返すことで行われる。探索の研究の発展に対して遅れていると言わざるを得ない。しかし、近年 TD 法を応用することで、評価関数の学習に対する研究が進みつつある。そこで、今回は評価関数の学習を応用し、対戦相手ごとに最適な評価関数を学習するシステムの構築を目指す。

2. 評価関数を学習するためのシステム

2.1.評価関数の要素

評価関数を、各局面での評価要素の特徴量と評価要素に与えられた重みとの積の線形和により実現することにする。評価要素には、局面の有利不利を判断するのに適当と思われるものを選ぶ必要があるが、今回は駒の価値とその盤面での位置の価値を要素とする。

$$v_t = \sum w_i c_{t,i}$$

v_t : 局面の評価値

w_i : 評価要素の重み

$c_{t,i}$: 評価要素の特徴量（駒 , 位置）

ここで特徴量は評価要素と局面が与えられれば自ずと定まるので、重要なのはその重みであり、この重みの選択こそが評価関数を特徴付ける。今まで重みは経験により決められてきた。今回はこれをTD(λ)法により学習させる。

2.2.評価関数の学習

重みをTD(λ)法により学習すること、それ自体は、チェスや将棋、その他のゲームにより何度も実験が行われ、既に十分な成果を収めてきた。今回はそこから一歩進めて、評価要素の重みを対戦相手に応じて変化させることを考えたい。もちろん、今までの人力での調整ではほぼ不可能であったことだが、プログラムが自動的に重みを学習できるとすれば、対戦相手ごとの微妙な調整も可能になるのではないか。もちろん、全ての場合において理想的な重みは存在するのかもしれないが、そのような重みを見つけ出すのは不可能であろう。それならば、特定の相手に対してより有効な重みを求める方が現実的と言える。ただし、TD(λ)法による評価関数の学習には数千局の対局が必要となる。相手もコンピュータ将棋プログラムであれば良いが、相手が人間である場合は時間がかりすぎて実質的に不可能である。従って、なんらかの妥協が必要である。

2.3.相手の違いの評価関数への反映

私は以前の研究で、局面の進行状況を判断するためのシステムを考案した。このシステムは32の入力要素を持つニューラルネットワークとして構成され、十分な数の棋譜をもとにして、適切な判断を行うための重みを学習する。そして、ある局面が与えられると、その局面が終局まであとどれくらいかを0.1から0.9までの数値で表現する。

$$y = f(x)$$

x : 局面

y : 進行度（開始局面 :0.1　投了局面 :0.9）

しかし、人によって将棋の指し方は違うので、特定の相手との対局は特定の傾向を持つ。その傾向は進行

状況の判断のシステムに重みのずれとして現れるはずである。この重みの違いを相手の違いとして評価関数に反映させたい。そこで、以下のようにしてシステムを評価関数に組み込むことにする。

$$v_t = \sum (w_{i1} y_t c_{t,i} + w_{i2} c_{t,i})$$

v_t：局面の評価値

w_{i1}, w_{i2}：評価要素の重み

y_t：進行度

c_t：評価要素の特徴量

この評価関数の式の重みをＴＤ(λ)法による自己対戦で決定する。重みの違いによって相手をモデル化しているわけではなく、相手との将棋そのものの反映であるので、全く同じ評価関数の自己対戦による学習は妥当であろう。

3. TD(λ)法

Samuelによって導入されたＴＤ法をSuttonが拡張して定式化されたＴＤ(λ)法は、以下の式によって重みを更新していく。

$$\Delta w_t = \alpha (P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \nabla_w P_k$$

Δw_t：t手目におけるwの更新量

α：学習の効率

λ：過去の予言の影響力

P_t：t手目における先手の勝つ予想確率

$$P_t = \frac{1}{1+e^{v_t}}$$

$$v_t = \sum w c_t$$

v_t：t手目の評価値

$$\nabla_w P_k = \left(\frac{\partial}{\partial w_1} P_k, \frac{\partial}{\partial w_2} P_k, \ldots, \frac{\partial}{\partial w_n} P_k \right)$$

評価値vがシグモイド関数により勝ちの予想確率Pへと変換され、それを元にして更新量が決まる。ここで、学習の効率を決定するαを重みごとに分けて考えると式は以下のように変形される。

$$\Delta w_{i,t} = c \alpha_i r_{i,t}$$

$$c = const$$

$$r_{i,t} = (P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \nabla_{w_i} P_k$$

4. 実験方法

今回は局面の進行状況を判断するシステムの評価関数への組み込みの有効性を確認するために以下の実験を行った。

まず、プロの棋譜をもとに局面の進行状況の判断を判断するシステムの重みを学習する。次にこのシステムを用いた評価関数の駒の重みとその位置の違いによる重みをＴＤ(λ)法により自己対戦による学習で決める。駒の重みの初期値は100、位置の違いによる重みの初期値は10とした。さらに、λはλ＝0.97で固定し、αはD.F.Beal,とM.C.Smithにより提案された以下の式により一局ごとに決定する。

$$\alpha_i = \frac{N_i}{A_i}$$

$$N_i = N_i + \sum_{t=1}^{end-1} r_{i,t}$$

$$A_i = A_i + \sum_{t=1}^{end-1} |r_{i,t}|$$

$$r_{i,t} = (P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \nabla_{w_i} P_k$$

重みの更新もまた一局ごとに行い、上限を5000局として更新量が十分小さくなるまで繰り返す。

5. 実験結果（別紙参照）

6. 考察（別紙参照）

参考文献

[1]R.S.Sutton. Learning to Predict by the Methods of Temporal Differences. Machine Learning,3:9-44,1988

[2]D.F.Beal,M.C.Smith. Learning piece values using temporal differences. ICCA Journal,147-151,September 1997

[3]D.F.Beal,M.C.Smith. Temporal Coherence and Prediction Decay in TD Learning.