TD(λ)と Bonanza の学習法との性能比較

築地毅　　柴原一友　　但馬康宏　　小谷善行
東京農工大学

概要

2006 年の世界コンピュータ将棋選手権において、将棋プログラム Bonanza が優勝を収めた。それは、新しい学習手法によったものであった。そこで、本稿では将棋を例に挙げ、この Bonanza による新しい学習手法に注目し、今まで一定の成果を挙げてきた TD(λ)と比較することにより、2 つの学習手法の性能を比較する。

Performance comparison with learning method of TD(λ) and Bonanza

TSUKIJI Tsuyoshi, SHIBAHARA Kazutomo, TAJIMA Yasuhiro, KOTANI Yoshiyuki
Tokyo University of agriculture and technology

Abstract

The Shogi program Bonanza won the first prize in the World Computer Shogi Championship in 2006. It has a new learning method. We focus a new learning method by this Bonanza and compare it with the learning method TD(λ) which have had significant results.

１．はじめに

本稿では、TD(λ)と Bonanza による全合法手による学習（以下 Bonanza の学習と呼ぶ）との性能比較を行う。思考ゲームの学習において、予言間の差を利用した TD(λ) [2][3][4]が一定の成果を挙げてきた。その後、実際に指された手と他の合法手を差別化するという新しい手法[1]が提案された。2006 年の第１６回世界コンピュータ将棋選手権において、後者の学習手法を用いた将棋プログラム Bonanza が優勝し、新しく有効な学習手法として注目されている。そこで本稿では、この二つの手法を、将棋を用いて同じ状況下において学習させ、その性能を比較するものとする。

１．１．目的

2006 年の第１６回世界コンピュータ将棋選手権において将棋プログラム Bonanza は、実際に指された手と他の合法手を差別化するという、新しい学習によって評価関数を構築した。そこで本稿において、学習手法に焦点を当て、ある同様の条件下において既存の手法であった TD(λ)と対照実験を行うことにより、Bonanza の学習の性能を調べるものとする。

２．関連研究

２．１．TD(λ)

プロの棋譜を用いた教師あり学習手法の一つが TD 法である。各局面における勝率（以下「予言」と呼ぶ）を評価関数から計算し、その予言間の差を利用して学習を行う手法である。更新方針は各予言間の差を最小にするように重みを変化させるものになっている。これは、勝ち局面の近くの（例えば１手前の）局面は、恐らく勝ちに近い局面であるからであり、ある一局において評価値が大きく変動することは考えにくいからである。さらに、現在の観測状態は、過去の観測状態に関係して起こると考えられる。そこで現在の観測状態と過去の観測状態とを関連させ、予言の勾配に指数型の重み(適格度トレース)λを導入した手法を TD(λ)と呼ぶ。[2][3][4]

２．２．Bonanza の学習

プロの棋譜を用いた、教師あり学習である。プロが理想的な手を指していると仮定するならば、プロの指した手による局面は、他に考えられる合法手による局面よりも評価値が高いと考えられる。従って、棋譜からプロが指した手による局面の評価値を高く、その他の手を低くするように重みを変化させることにより、プロが持つと思われる評価関数を学習することが出来る。[1]

３．TD(λ)と Bonanza の学習の理論

３．１．記号の定義

ここで、今回の実験で用いる記号について表１で定義する。

表１：記号の定義

$\vec{x}_{t,m}$	ある合法手 m によって指された t 手目の手による局面の特徴ベクトル。棋譜によって実際に指された手は m=0 とする。始局面は $\vec{x}_{0,0}$ と定義する。
$\vec{w}_t$	t 手目における特徴ベクトルの重みベクトル。
$E(a)$	評価関数
$P(a)$	シグモイド関数
α	学習率。
λ	適格度トレース。過去への依存度を表す。0 より大きく 1 以下の値を取る。
$l_{g,t1}(\vec{w}_{t2})$	局面 t2 における重みによって計算される、対局 g の局面 t1 に棋譜で示された局面とその他の合法手による評価値の違いの度合い。 $l_{g,t}(\vec{w})=\sum_{m=1}^{M}\left(T\left(E\left(\vec{x}_{t2,0}\right)\right)-T\left(E\left(\vec{x}_{t2,m}\right)\right)\right)$ と定義する。

３．２．TD(λ)の理論

TD(λ)は、実際に棋譜で指された局面ごとの評価値の差を小さくするように重みベクトルを変化させていく。したがって、目的関数は式１の様になり、この値を出来るだけ小さくすれば良い。ただし、G は学習する対局数とし、$T(g)$は対局 g における終局までの総手数とする。ただし、$P_i \equiv P\left(E\left(\vec{x}_{i,0}\right)\right)$とする。

$$E_T=\sum_{g=1}^{G}\left(\sum_{t=1}^{T(g)}\left(P_{t+1}-P_t\right)\right) \qquad (1)$$

また、重みベクトルの更新には式２を用いる。重みベクトルの更新は、1 対局をすべて学習した後に行う。

$$\vec{w}_{t+1} = \vec{w}_t + \alpha(P_{t+1} - P_t)\sum_{i=1}^{t} \lambda^{t-i} \nabla_w P_i \qquad (2)$$

3．3．Bonanza の学習の理論

Bonanza の学習は、棋譜によって実際に指された手と、その他の合法手によって行う。棋譜によって指された手が他の合法手より評価値が高くなれば良いので、目的関数は式３の様になり、この値を出来るだけ大きな値にすればよいことが分かる。

$$E_B = \sum_{g=1}^{G} \left(\sum_{t=1}^{T(g)} l_{g,t}(\vec{w}_t) \right) \qquad (3)$$

また、重みベクトルの更新には式４を用いる。重みベクトルの更新は、１対局をすべて学習した後に行う。

$$\vec{w}_{t+1} = \vec{w}_t + \alpha \nabla_w l_{g,t}(\vec{w}_t). \qquad (4)$$

式４は最急降下法に基づいている。傾きが正ならばプラス方向に、負ならばマイナス方向に微少量 w を変化させることによって、適切な値を得ることが出来る。

3．4．TD(λ)と Bonanza の学習との比較

TD(λ)と Bonanza の学習の特徴について述べる。TD(λ)は、棋譜だけの局面を評価すれば良いのに対し、Bonanza の学習では、全ての合法手を求め、さらにそれを全て評価しなければならないために時間がかかる。しかし逆に、TD(λ)では、人間が指した、良いとされる手しか学習できないのに対し、Bonanza の学習では、良くないとされる局面までも評価できるという強みがあると考えられる。このように二つの学習手法には、「学習する局面数」という点において、互いに相反する長所、短所があると考えられる。

4．実験

4．1．実験方法

３章で述べたように、２つの学習手法の相反する特徴の要素となっているのは「評価する局面数」であった。そこで、本稿では２つの学習手法を実装し、「評価する局面数」が２つの学習手法について等しくなるようにループ回数を調整し、学習の様子を得ることにする。学習対象となる棋譜の数は 3595 対局用意した。全ての対局を学習する時の局面数は、TD(λ)が、415708 個、Bonanza の学習が 30425150 個であった。従って、局面数を等しくするため、3595 対局の棋譜に対して、TD(λ)を 29275 回、Bonanza の学習を 400 回学習させ、その時の目的関数の収束の様子を得る。本実験では、13 個の駒価値を重みベクトルに取り、初期値を 1000 とする。また、適格度トレースλを 0.95 とし、学習率αを 0.2 とした。今回の実験においては、収束の速さを求めることが目的のため、学習率は変化させないことにする。また、重みベクトルを更新した後、重みが定数倍の値をとることを防ぐため、重みベクトルのノルムを変化させないように正規化を行った。

4．2．実験結果

図１と図２に、二つの学習によって得られた目的関数の収束の様子を示す。

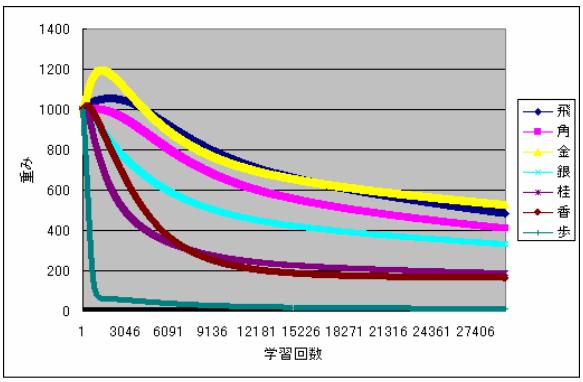

図１：TD(λ)目的関数の収束の様子

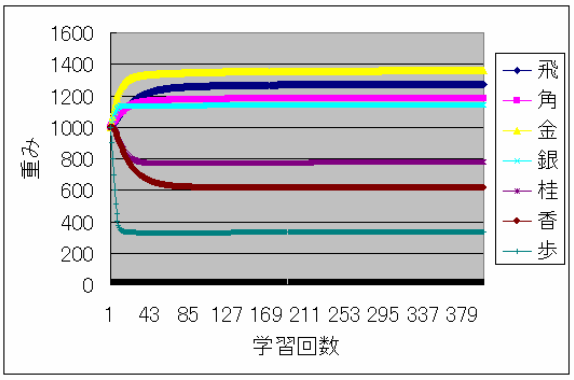

図２：Bonanza の学習目的関数の収束の様子

５．考察

図１、２のグラフで見ると、Bonanza の学習のほうが早く収束していることが分かり、今回の実験においては、収束の速さの点においては Bonanza の学習のほうが優れているという結果になった。これは、人間が指さなかった手まで学習することにより、良い局面と悪い局面の両方を学習できたからだと考えられる。但し、図はスペースの関係上載せることができなかったが最終的な駒価値の重みは妥当なものといえる結果ではなかった。これは重みベクトルが定数倍の値を取ることを防ぐために、ノルムが一定になるように行った正規化が今回の実験に対しては不適切だった可能性がある。

TD(λ)で学習すると、金の価値が大きくなることが多い[3]。これは終盤の詰めの段階で金を重視することが、適格度トレースによって、序盤にまでも重視するという結果になったと考えられる。従って、序盤、中盤、終盤に分けて実験してみる必要があると考えられ、Bonanza の学習においても実験する必要があると考えられる。

今回の実験では特徴ベクトルは駒価値のみで行ったが、例えば「駒の絶対的な位置」「駒の相対的な位置」「駒の効き」などの特徴ベクトルを増やした時、同じような結果になるか実験する必要がありそうだ。また、得られた評価関数を用いて、対局を行い学習結果の重みベクトルが適切な値かを確かめる実験も必要であると考えている。他に、今回は Bonanza の学習の比較対象として TD(λ)を取り上げたが、他の学習法と比較したり、他のゲームを例に挙げて実験をしたりする必要があると考えられる。

６．おわりに

本稿では、今までに一定の成果を挙げてきた TD(λ)と、新しく提案された Bonanza の学習との性能比較を行った。

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. GPW'06,pp78-83,2006

[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. GPW'99,pp.31-38,1999

[3] 薄井克俊,鈴木豪,小谷善行. TD 法を用いたプロの棋譜からの評価関数の効率的な学習. 東京農工大学修士論文,2001

[4]J.Baxter,A.Tridgell,L.Weaver. EXPERIMENTS IN PARAMETER LERNING USING TEMPORAL DIFFERNCES. ICGA Journal, 21(2), p84-99, 1998