BackTrack探索を用いた強化学習の効率化

日大生産工（院）　○上野　智章
日大生産工　　松田　聖

1 はじめに

近年の自律エージェントの研究の活発化により、強化学習（Reinforcement Learning）が注目されている。強化学習は、学習するエージェント自体に知識が何も無い状態から始め、エージェントが経験したことのみを自分の知識として獲得していき、その知識を用いて学習を進めていく手法である。しかしその反面、エージェントが十分な知識を獲得するまでに莫大な時間がかかり学習スピードが他の学習手法より遅いという欠点を持っている。強化学習を現実の問題で使用できるようにするには学習スピードを上げる、つまり効率化を図ることが必要である。そこで、本研究ではBackTrack探索を用いて強化学習の効率化を行う。

次節で強化学習、3でBackTrack探索について述べることとする。

2 強化学習

強化学習は人工知能の機械学習の1つで、他に帰納的学習、演繹的学習、類推的学習、発見的学習がある。これら５つの学習手法は知識を持った外部の教師が存在する教師あり学習とそのような教師が存在しない教師なし学習に分けられる。強化学習は後者の教師なし学習である。また、強化学習は行動の評価を直後にするのではなく目標とする状態に到達した時にすることが出来る。この性質はゲームなどの今の行動が良かったかどうかはゲームが終了してみないと判らないといった環境下に適している。

次に、強化学習の基本構成を図1を用いて説明する。

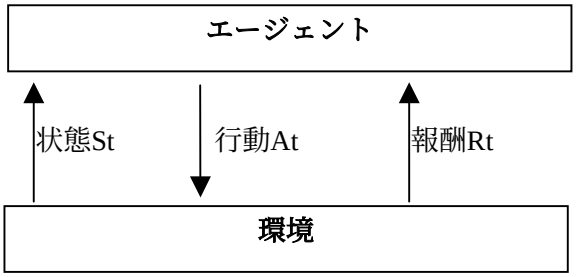

図1：強化学習の基本構成

時間tにおいてエージェントは環境の状態を認識し、これによって行動Atを環境に対しなす。そしてエージェントはこの行動により報酬Rtを環境から受け取る。エージェントは状態Stに対する行動Atの評価値を持っており、評価値によって行動が決定される。また、この評価値を更新することにより、最適な行動が選択されるようになる。

強化学習の代表的な学習手法にQ学習（Q-learning）やTD学習（TD-learning）がある。本研究ではQ学習を用いて検証を行うため、以下でQ学習について説明する。

Q学習は状態sで行動aをとった時の将来に獲得することが出来る報酬の期待値$Q(s,a)$を次式のように学習する。

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right] \quad (1)$$

但し、s,a,rはそれぞれ状態、行動、報酬を表し、s'は状態sで行動aをとった後の状態、

Efficient Reinforcement Learning with BackTrack

Tomoaki UENO and Satoshi MATSUDA

a'は状態s'における行動を表す。αは学習率で$0 < \alpha \leq 1$の値をとる。また、γは割引率で$0 \leq \gamma < 1$の値をとる。

式(1)を使うことにより、良い行動の期待値は上がるため、状態Sでの最良の行動Aをエージェントが学習することが出来る。しかし、過去の経験から期待値を更新することだけを優先していると学習結果が正確なものではなくなってしまう。最良の結果を得るには過去に選択されていない行動も試みる必要がある。このトレードオフを解決することも重要である。

有効な状態のみを選択し、学習を進めていくと学習スピードは増加する。そこで本研究では不必要な状態を削除するためBackTrack探索を使用する。

3　BackTrack探索

探索とは、ある問題において想定される多くの状態の中からその問題に最適な状態を見つけ出す方法である。しかし、探索方法自体に問題は無くとも状態の評価が間違っており結果として選択した状態が最適ではない場合が存在する。また、状態の評価値の最大値に同値ものが2つ以上存在するとき縦型探索などの効率の悪い全数探索をしない限り、評価値が最大のすべての状態を選択する保証はない。また、強化学習の環境下では選択した手が最適であったかどうかは終端ノードまで探索をしないと判断できない。この点を考慮して探索を行う必要がある。そこでBackTrack探索を使用する。以下にBackTrack探索の進め方を記した。

1. 図2の状態aにおいて子ノードの評価値の最大値のノードが複数存在するとき状態aにフラグを立てる

2. 評価値が最大の子ノードの一方bに進み、終端ノードまで１～２を繰り返す(図3参照)

3. フラグが立っている状態aに戻り２で選択されなかった評価値が最大の残りのノードdに進み１～２を繰り返す。(図4参照)

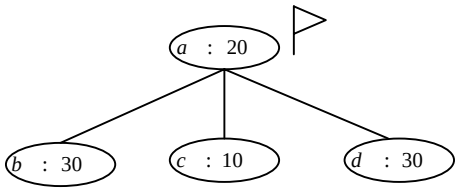

図2：BackTrack探索step1

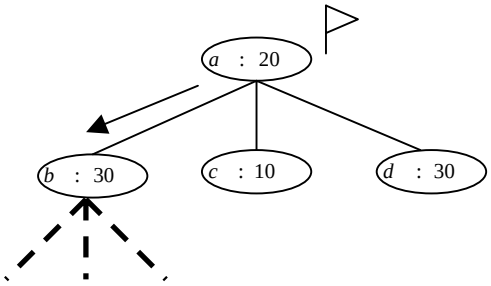

図3：BackTrack探索step2

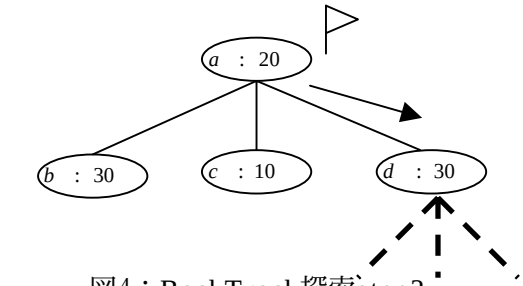

図4：BackTrack探索step3

BackTrack探索をすることにより、ゲーム等を始めからやり直すのでなく、途中の有力と思われるルートだけを選択して吟味するので無駄と思われる試行をしないですみ効率的と言える。また、選択されなかったノードに進むことにより探索の幅が増加するため、2章で述べたトレードオフの問題を解決できるといえる。

フラグの立て方を上記の説明では最大値が複数の時としたが、最大値に近い値にもフラグを立てることによってさらに幅広い探索をすることが出来る。しかし、フラグを数多く立てることにより全数探索に近くなり探索時間は無論増加し効率化とは程遠くなるためユーザの注意が必要である。

４　おわりに

今回、強化学習にBackTrack探索を適用することを提案した。現在はBackTrack探索の強化学習の有効性を検証するため、オセロの最適手を学習対象としシミュレーション中である。

参考文献

1) Richard S. Sutton, Andrew G. Barto
「強化学習」 三上　貞芳・皆川　雅章
共訳　森北出版株式会社,（2000）
2) 前田　隆・青木　文夫　共著　「新しい人工知能」基本・発展編　オーム社,(1999)