2H1-3in

メタヒューリティクスによる強化学習のパラメータの最適化

安政 駿[*1] 手塚 太郎[*2]
Yasumasa Shun Tezuka Taro

[*1] 筑波大学 図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

[*2] 筑波大学 図書館情報メディア系

Faculty of Library, Information and Media Science, University of Tsukuba

1. 概要

近年，人間の学習方法を機械に行わせる手法として，機械学習と呼ばれる手法が広まっている．機械学習ではラベル付きデータに対してラベルなしのデータをフィッティングすることで学習を行う教師あり学習やラベルなしのデータからモデルを構築する教師なし学習が一般的である．しかし，これらの学習方法はデータがない新たな環境に適応するために試行錯誤を行い，新たなデータを集めて最適な行動を見つけるといった問題には向いていない．このような問題に適した学習方法として強化学習が注目を浴びている．強化学習では，ある環境に対して，方策や報酬関数，価値関数を定義することで，学習エージェントの行動を決定する[三上 2000]．方策とはある時点での学習エージェントのふるまい方を示し，確率的に遷移する．報酬関数は強化学習問題の目標を定義したものであり，エージェントは最終的に総報酬を最大化するように行動する．また，価値関数とは報酬関数にどの程度の価値を置くかを定義することでその状態の報酬を高める行動をとるか，環境における行動全体の報酬を高めるために行動を行うかを定めるものである．つまり，報酬関数に重きを置いた場合，エージェントは環境の全体を把握せずにその状態で最適な行動を行うことになる．逆に価値関数に重きを置いた場合はある状態での最適な行動ではなく，環境における目的を達する際に一番良いと思われる行動を行うこととなる．人間においては，環境や状態に応じて，報酬や価値における比重を分析し最適な行動をとることができる．しかし，機械においてはどちらに重点を置くかを機械的に判断する必要がある．このために，強化学習ではいくつかのパラメータを設定することで報酬重視の行動をとるか，価値重視の行動をとるかを判断することとなる．現在，このパラメータの設定は実験者の経験により決定されるため，パラメータを決定する際に環境に応じて適切なパラメータを調べる必要がある．これは，環境が変化する状況に対応させる強化学習の中では大きなコストとなり得るため，自動的に決定させる手法を導き出すことにより，強化学習の発展に貢献できるものと考えられる．最適パラメータを推定しようとした研究の一つに亀井らの研究が挙げられる[亀井 2007]．亀井らは探索環境に環境複雑性という尺度を定義することで，強化学習パラメータを推定することを試みている．しかし，この手法であると，環境ごとに複雑性の定義が必要である．そこで，より単純な指標によりパラメータを推定できる手法が必要である．

2. 手法

本研究ではメタヒューリスティクスの枠組みを用いて強化学習パラメータの推定を行う．メタヒューリティクスとは，最適化問題を解くための経験的手法のことである[久保 2009]．理論的に最適な解の値を得られるわけではないが，短時間で大域的な近似解を得ることができる．メタヒューリスティクスの手法としては，局所探索法や遺伝的アルゴリズム，タブー探索法など様々なものが存在するが，本研究では焼きなまし法を用いる．焼きなまし法は，ランダム性を用いることにより局所的最適解に陥る事を防ぐためのアルゴリズムである．探索の序盤はランダム性により大域解を探し，終盤に近付くにつれ局所解周辺を探索することができる．これにより，局所解に陥ることでパラメータが最適な値に近づかないという問題に対処する．また，終盤に局所解周辺を探索することにより，メタヒューリティクスの手法であっても，最適な解の値に近づくことができる．また，本研究では強化学習の課題から実験者がどのような指標に対して，最適なパラメータの値を探索するのか選択する必要がある．例えば，強化学習の探索時間を短くしたいのならば，強化学習課題から一回の探索にかかった時間をフィードバックする必要がある．この指標が一番高いものとなるようにパラメータの値を調節することとなる．

図 1 は本手法のアルゴリズムである．N は試行回数，θ はパラメータ，e は強化学習課題，W は対象としたい評価指標である．U[0,1]は区間[0,1]からの一様分布によるサンプリングを表す．評価指標の最大値を更新するようなパラメータが出現した場合には，$\hat{\theta}$，$\widehat{W}$ を更新することで新たな探索の起点とする．試行を進めていくにつれて，$\hat{\theta}$ に近い値を探索していくこととなる．

For i = 1 : N

$r_i = U[0,1]$;

$\theta_i = \frac{i}{N}\hat{\theta} + \left(1 - \frac{i}{N}\right) \cdot r_i$;

$W_i = \mathrm{e}(\theta_i)$;

If $\widehat{W} < W_i$

$\hat{\theta} = \theta_i$; $\widehat{W} = W_i$;

End

図 1 パラメータ推定アルゴリズム

[1] s1321656@u.tsukuba.ac.jp
[2] tezuka@slis.tsukuba.ac.jp

3. 実験

実験では最初に単一のパラメータを焼きなまし法により推定する. 実験の課題としては, モンテカルロ法を用いた三目並べの解法を学習するプログラムを用いた[八谷 2008]. このプログラムでは二つのパラメータ ε とγを変化させることで, ε -greedy 法による学習の挙動を変化させることができる. この実験ではγを 0.9 に固定した際の ε パラメータの推定を行う. 図 2 はγを 0.9 固定したときの ε をしらみつぶしに探索した際の ε の変化を表した図である. また, ε = 0.01 から ε = 1 までの 100 回の探索が必要となる.

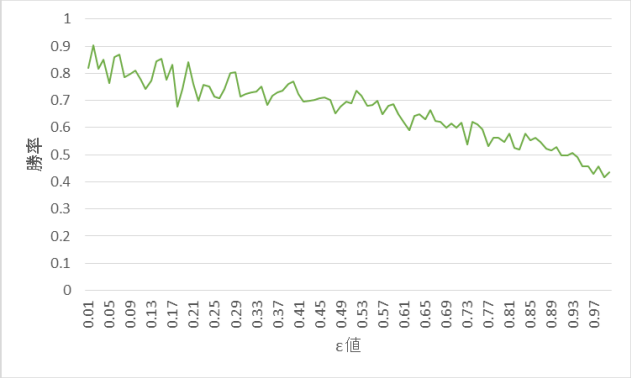

図 2 三目並べにおけるγを固定したときの ε の変化

これに対し, 本研究におけるパラメータ推定を行った結果を表に示す. 表 1 は初期 ε パラメータを 0.3, 0.5, 0.7 と置いた時の最大勝率とその時の ε の値である. 図 2 の結果, ε は小さい値を取る時, 勝率が高くなるという結果が見て取れる. 推定結果から, 初期値としてどのようなパラメータの値を取っていても, 一定以上の勝率を取るパラメータを得ることができると言える. しかし, 最適なパラメータを得るためには初期値と共に探索序盤の乱数に依存するという傾向が見られた.

表 1 γ=0.9 の時のパラメータ推定結果

初期 ε	ε	勝率
ε = 0.3	0.018	0.946
ε = 0.5	0.234	0.833
ε = 0.7	0.117	0.889

また, 実際の強化学習においてはパラメータが複数あることが一般的である. よって, 本実験では二つのパラメータに対し焼きなまし法を用いた推定を行う. 具体的には, ε とγに対して交互に焼きなまし法を適用することで, パラメータを変化させ, より高い勝率を持つパラメータへと変化させることができる. 二つのパラメータに対し, 0 から 1 までのパラメータを粒度 0.01 でしらみつぶしに探索を行った. この時 10000 回の試行を行うことになる. この結果, γ = 0.76, ε = 0.02 の時最大の勝率 0.942 を出すことが示された. これに対して, パラメータ推定では, 50 回の試行を行った際の勝率を示す. パラメータの初期値はγ = 0.3, 0.5, 0.7 と ε = 0.3, 0.5, 0.7 の値を取る. 表 2 は各初期値を指定した際に 3 回ずつ推定を行った時の最終的な勝率の平均値である. 各勝率の平均値は 0.883 であり, しらみつぶしに探索を行った時と比べて 93.8%の勝率の値を示している. つまり, このパラメータ推定法を用いることにより, 50 回の試行で最適値の 93%の勝率を持つパラメータを推定できることを示している. また, 表 3 は各初期値を用いてパラメータ推定を行った際の最終的なパラメータの値と勝率である.

表 2 初期値ごとの最終的な勝率の平均値

	ε = 0.3	**ε = 0.5**	**ε = 0.7**
γ = 0.3	0.856	0.905	0.891
γ = 0.5	0.889	0.914	0.894
γ = 0.7	0.849	0.887	0.868

表 3 初期値ごとの最高の勝率を示した時のパラメータ

初期値	ε	γ	勝率
ε = 0.3, γ = 0.3	0.107	0.873	0.877
ε = 0.3, γ = 0.5	0.022	0.503	0.919
ε = 0.3, γ = 0.7	0.100	0.206	0.9
ε = 0.5, γ = 0.3	0.059	0.232	0.9
ε = 0.5, γ = 0.5	0.039	0.230	0.931
ε = 0.5, γ = 0.7	0.035	0.070	0.913
ε = 0.7, γ = 0.3	0.205	0.419	0.849
ε = 0.7, γ = 0.5	0.090	0.651	0.891
ε = 0.7, γ = 0.7	0.121	0.085	0.886

この手法を用いることにより, 一定以上の勝率を持つパラメータを推定できたと言える.

4. まとめ

本手法の利点としては, パラメータの依存関係にとらわれることなくパラメータを推定できる点にある. また, どのような初期パラメータを取ったとしても一定の勝率に定まることから, 強化学習課題のパラメータを初期設定するだけで, 適した値に自動的に定めることができると言える.

今後の課題として, 推定の終盤においても一様分布を用いた乱数の生成を行っているため, 焼きなまし法の利点である, 局所探索がうまく行えていない可能性がある点が挙げられる. このことを解決するために, 一様分布ではない分布を用いることにより, より適したパラメータを推定できると考えられる. また, 今回の課題では適したパラメータが推定できるという結果が示されたが, 他の課題・環境においても同様な効果が得られるかを調査する必要がある.

参考文献

[三上 2000] 三上貞芭, 皆川雅章:強化学習, 森北出版, 2000.
[亀井 2007] 亀井圭史, 石川眞澄: パラメータの相互依存性を考慮した強化学習の最適パラメータ推定, 信学技報, 電子情報通信学会 , 2007.
[久保 2009] 久保幹雄, J.P.ペドロン: メタヒューリティクスの数理, 共立出版, 2009.
[八谷 2008] 八谷大岳, 杉山将: 強くなるロボティック・プレイヤーのつくり方, 毎日コミュニケーションズ, 2008.