動的環境を考慮した効率的な強化学習手法

高田　沙都子 †　　宮内　新 ‡　　荒井　秀一 §
武蔵工業大学 †　　武蔵工業大学 ‡　　武蔵工業大学 §

1. はじめに

強化学習とは，試行錯誤を通じて未知の環境に適応する学習制御の枠組である．環境との相互作用の繰り返しを通じて, 最適または合理的な政策を学習することが強化学習の目的である．強化学習の問題点の一つとして, 学習時間があげられる. 強化学習ではエージェントが試行錯誤をしながら学習していくことから, 学習が収束するまでにはかなりの学習時間を要する.

さらに、エージェントが学習する環境は、学習中に環境が大きく変化することも考えられる。環境が変化するとエージェントは新たな環境で学習し直す必要があるため、学習の効率が著しく落ちてしまう傾向がある。そのため、環境が変化しても素早く環境に適応し、効率的な学習ことのできる学習手法が望まれる。

学習の効率を改善した手法として Dynamic Profit Sharing[1][2] がある. 本研究では、この Dynamic Profit Sharing(DPS) を用いて研究を行う。

2. 研究背景

2.1 Profit Sharing

強化学習手法として代表的なものに Profit Sharing がある. 学習を行なうエージェントは環境から状態を認識し, 行動を選択する。その状態と行動を対 (以下ルール) として記憶し, 目標を達成した時に得る報酬をルールに分配することで, 各状態でとるべき行動を学習していく. このとき与える報酬を, 目標状態に近い方のルールから一つ状態を遠ざかる毎に報酬割引率 S で割り小さくしていく. また、適応できる環境のクラスが広く, 報酬割引率をエージェントが行動できる数にすることで、学習が必ず収束することを保証する合理性の定理が示されている [3].

2.2 Dynamic Profit Sharing

Profit Sharing は合理性を完全に保証している反面、報酬を目標状態から遠いルールまで伝搬することができない。

そこで、Dynamic Profit Sharing では Profit Sharing で学習中固定だった報酬割引率を状態ごとに適正な値を求めることで合理性を十分保証しながら、学習効率を向上させている。合理性を保つためには無効ルールが競合する有効ルールを差し置いて一番に強化されないようにすればよい. つまり学習において無効ルールに与えられる報酬が有効ルールに与えられる報酬よりも少なければよい. 無効ルールを抑制することが最も困難な環境は, ある状態が持つルールのうち一本のみが無効ルールで, その無効ルールが二番目の大きさの報酬値を持つルールであった場合である.

ある状態の中で最大の報酬値を持つルールが選択される確率を P_{max} とし, 同様に、二番目に大きい報酬値を持つ無効ルールが選択される確率を P_{ine} とおく. 合理性を満たすためには,$R_{max} > R_{ine}$ を満たせばよいので, 無効ルールが連続で選択される最悪の値を W_{ine} とおき, 報酬割引率を S とおくと

$$\frac{P_{max}}{1-P_{ine}} > \sum_{i=1}^{W_{ine}} \frac{1}{S^i} \tag{1}$$

が成り立っていればよい. この式を用いて、状態ごとに動的に報酬割引率を求めることで約 99%保証しながら効率の良い学習が可能となっている。

3. 研究目的

本稿では、Dynamic Profit Sharing で向上した学習の効率を維持しながら、環境の変化にも適応できる手法を提案することを目的とする。

4. 変化のある環境, ない環境での学習状態

環境が学習途中に大きく変化した場合と変化しなかった場合の学習効率の変化について Profit Sharing と Dynamic Profit Sharing を用いて検証を行った。

4.1 変化しない環境での学習

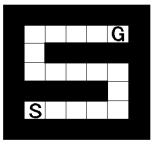

図 1: 変化しない場合の実験環境

図 1 のような 16 ステップで目標状態に到達することのできる迷路環境で実験を行った。10000 エピソードの学習を Profit Sharing、Dynamic Profit Sharing で行い結果を比較した結果、以下のようになった。

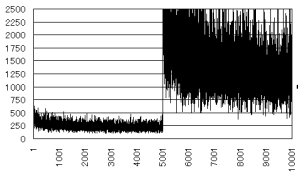

図 2: PS 学習曲線

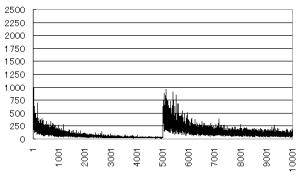

図 3: DPS 学習曲線

The Efficient Learning Method in Dynamic Environment
†Satoko Takada, Musashi Institute of Technology
‡Arata Miyauchi, Musashi Institute of Technology
§Syuichi Arai, Musashi Institute of Technology

4.2 変化する環境での学習

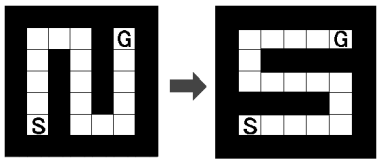

図 4: 変化する場合の実験環境

図 4 のように学習中、5000 エピソード後に環境が変化するような環境で 10000 エピソードの学習を Profit Sharing、Dynamic Profit Sharing で行い結果を比較した結果、以下のようになった。

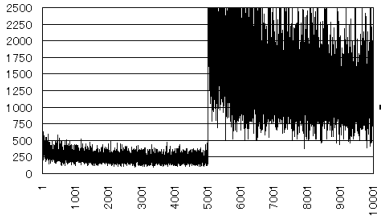

図 5: PS 学習曲線

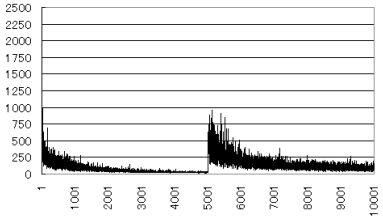

図 6: DPS 学習曲線

これらの結果から、学習中に環境が変化した場合は、学習の効率が著しく落ちることがわかる。これは、エージェントは環境が変化したあとも変化する前の学習結果を用いて学習を行わなくてはならないことに原因があると考えられる。しかし、エージェントは環境が変化したことを認識することはできないため、環境に変化に適応できる手法が必要となってくる。

5. 提案手法

学習中、変化の起こらない環境では Dynamic Profit Sharing で十分効率的に学習することができる。

環境が変化する前には有効ルールだったものが無効ルールとなり、無効ルールだったものが有効ルールとなる場合がある。この現象が学習の効率を悪化させる大きな要因となっている。

学習中、有効ルールはそのエピソードで最大の報酬値を獲得する。直前のエピソードで得た報酬の割合とそれまでの学習で得てきた報酬の割合に大きな差が生じた場合、それまで有効ルールだったものが無効ルールになるということが起こっていると考えられる。

このことから、直前のエピソードの報酬の割合とそれまでの学習から得られた報酬の割合に大きな差が生じた場合、次のエピソードでその状態の行動選択確率を調整すればよい。逆に、あまり報酬の割合に差が生じなかった場合は環境に変化がほとんど見られなかったと考えることができるため、調整を行わなくても Dynamic Profit Sharing で十分学習可能であると考えられる。

この手法を実現するアルゴリズムは以下のようになると考えられる。

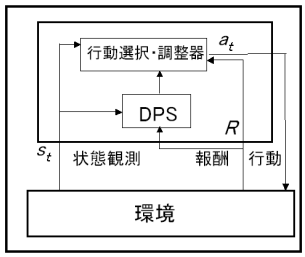

図 7: 提案手法の流れ

1. 前エピソードで得た報酬の割合とそれまでの学習で得た報酬の割合を比較して調整値を導出
2. 調整値に基づいて各ルールの行動選択確率を調整
3. 調整した行動選択確率を用いて行動を選択
4. 目標状態に到達するまで 3. を実行
5. 目標状態に到達したら DPS で学習を行うとともにエピソード分の報酬を調整器に分配

この手法を用いれば変化しない環境においては DPS の学習効率を保ちながら、環境が変化した場合においても報酬値を調整することで DPS よりも効率よく変化に適応できると考えられる。

6. 今後の方針

環境が変化する場合において、この手法と DPS について比較を行い、報酬の割合に差が生じた場合、次のエピソードでどの程度行動選択確率を調整するのが最も効率の良い学習が行えるのか、その割合を検討する必要がある。また、この調整によって合理性を保証する割合にどのような影響があるのか検証する必要がある。

参考文献

[1] 長谷川 雄吾, 高田 沙都子, 宮内 新, 荒井 秀一 : Profit Sharing を改良したより効率的な強化学習手法 (1)-選択確率による報酬割引率決定手法- , 情報科学技術フォーラム Vol.FIT 2003, 情報技術レターズ Vol.2, Page125-126

[2] 高田 沙都子, 長谷川 雄吾, 宮内 新, 荒井 秀一 : Profit Sharing を改良したより効率的な強化学習手法-Dynamic Profit Sharing での合理性の検討- , 情報科学技術フォーラム Vol.FIT 2003, 情報技術レターズ Vol.2, Page127-128

[3] 宮崎 和光, 山村 雅幸, 小林 重信 : 強化学習における報酬割当ての理論的考察, 人工知能学会誌, Vol.9, No.4, pp580-587(1994)