平 成 18 年 度　　卒 業 論 文 概 要

入学年度 平成 11 年度	**学生番号** 11117697	**氏名** 田中 靖章
卒業研究題目 つながりがある複数の環境での強化学習の研究		和田 · 犬塚 **研究室**

1 はじめに

現在, 複雑な動作が可能なロボットの開発と共に, その動作を獲得するための手段として強化学習の重要性が高まっている. 強化学習とは, 学習主体である**エージェント** (コントローラ) が, 設計者により目標 (**報酬**) を設定された制御対象である**環境**に対し, 試行錯誤する事により, 目標を達成するための最適な動作を学習する方法である.
強化学習では環境が複雑になってくると, 学習の成果を得るまで膨大な時間が必要となる. 本研究では一定の規則でつながれている複数の単純な環境の合成と捉えられる場合, 複数の単純な環境で学習をおこない学習結果を合成する今回の提案手法によって, より効率的に学習ができないか実験をおこない, 検証した.

2 強化学習

通常の強化学習では環境と学習主体であるエージェントが以下のやりとりによって学習が進む.

1. エージェントが環境を観測し**状態** s を同定し, 探査戦略に従い**行動** a を選択実行する.
2. 状態と行動によりエージェントは報酬 r を得て, 状態 s' に遷移する.
3. 報酬を元にエージェントは学習をおこない,1. を繰り返す.

強化学習の目的はこのやりとりを繰り返す事により, 各状態でどんな行動を実行するのが最も多く報酬を得られる可能性があるか (**最適方策**) を学習することが目的である.
本研究では学習アルゴリズムとして状態と行動の組 (**ルール**) に Q 値というスカラー値をつけ評価する Q 学習を用いた. Q 学習の更新式は次のようになる

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

α ($0<\alpha\leq 1$) は学習率,γ ($0\leq\gamma\leq 1$) は割引率で, 将来得られるであろう報酬をどれだけ評価するかを決める.$\max_{a'} Q(s',a')$ はルール (s,a) により遷移した状態での最大Q値をあらわす.

3 提案手法

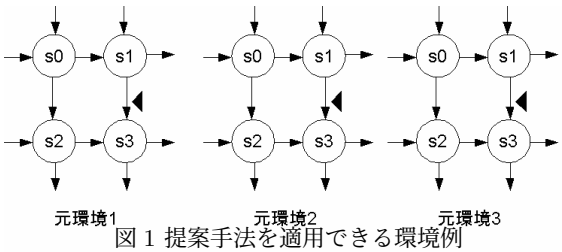

図 1 提案手法を適用できる環境例

例えば, 図 1 のような三つの環境が存在し, 一つの環境でのみ下に行動でき, 他の環境では右にしか行けないというつながりがあった場合, それぞれの環境に強化学習を適用することはできない. 強化学習を適用するには三つの環境で同時に行動を選択する事により形成される, 状態数が各環境の状態数の積, この場合 $4\times4\times4=64$, 行動が {(下, 右, 右),(右, 下, 右),(右, 右, 下)} の 3 通りの新たな環境を構築し, その環境で学習する必要がある. このように構築される環境をここでは**全体環境**とよび, 元となる環境を**元環境**とよび, 区別する. 全体環境では, 状態, 行動, 報酬に, 元環境の情報をパラメータとして持っていると推測できる. しかし通常の強化学習では状態, 行動区別ができて, 報酬も合成されたものさえ分かっていれば学習できるため, 元環境の情報は必要とされない. それに対して提案手法では状態, 行動, 報酬の中に含まれる元環境のパラメータで学習をおこなう. 元環境が 1 から n までの n 個存在するときの提案手法のアルゴリズムを図 2 に示す. 関数 com$Q$ のアルゴリズムとして本研究では二通り提案している. 全体環境でのルール (s,a) に含まれる元環境 i のパラメータのルールを (s^i,a^i) とするとき, ひとつめの方法では, $Q(s,a)\leftarrow\sum_{i=1}^{n} Q^i(s^i,a^a)$ により, 全体環境での Q 値を導出している. もう一つの方法では,(s,a) に含まれる元環境のパラメータの Q 値の内, 最大値 を $Q(s,a)$ としている.

1: 状態 $s=(s^1,\ldots,s^n)$ の Q 値を関数 com$Q$ により
　　各 Q テーブルの Q 値を元に導出する
2: 状態 s の Q 値から, 探査戦略に従い,
　　行動 $a=(a^1,\ldots,a^n)$ を選択する
3: 行動 a を実行する
4: 報酬 $r=r^1+\cdots+r^n$ を得る
5: 状態 $s'=(s^{1'},\ldots,s^{n'})$ を観測する
6: **for**　$i\leftarrow 1$　**to**　n　**do**
7:　　$Q^i(s^i,a^i)\leftarrow(1-\alpha)Q^i(s^i,a^i)$
　　$+\alpha\left\{r^i+\gamma\max_{a'} Q^i(s^{i'},a^{i'})\right\}$
8: $s\leftarrow s'$;
9: 1 へループ

com$Q$: 各 Q テーブルの結果から全体環境での Q 値を決定する関数
α : 学習率　γ : 割引率

図 2 提案手法による強化学習のアルゴリズム

4 実験とその結果

いくつか環境を用意し, 既存手法と提案手法を比較した. 既存手法より収束の効率が上がっているか. 提案手法の学習結果が最適方策となっているか. com$Q$ が元環境の Q 値の総和か最大値どちらがよいかに注目し実験をおこなった. 提案手法で例として述べた全体環境を基本形とし, その基本形に対して変化を加えて結果がどのように変わるか実験した. 結果として,com$Q$ は最大値よりも, 総和の方がよい結果が得られた. また全体環境に既存手法を適用した場合に比べ, 提案手法がより効率的に収束することが確認できた. 結果の例を図 3 に示す. さらに基本形の場合, 元環境の数を増加させた場合, および元環境の形が不均一の場合も最適方策を導出できた. しかし複数の元環境から合成された報酬として全体環境で得られる場合, 最適方策にならなかった. また元環境の報酬ではなく全体環境側から報酬を設定し, 提案手法の学習方法では報酬がマルコフ性を満たさないとき, 収束せずに, 最適方策を得ることができなかった.

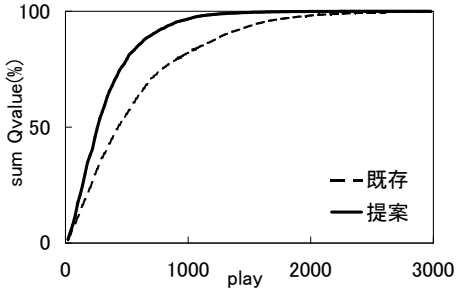

図 3 Q 値収束に関する実験結果の例

5 今後の課題とまとめ

本研究ではつながりがある複数の環境における学習において, 学習効率を改善するための強化学習のアルゴリズムを提案した. また, 実験をいくつかの環境に対しおこなうことにより, 提案手法がどのような条件を満たす環境に対して有効に働くか検証をおこなった.

参考文献

[1] Richard S Sutton & Andrew G.Barto, 三上貞芳 · 皆川雅章 共訳. 強化学習. 森北出版.