強化学習における
危険回避行動獲得のための負の報酬伝搬法

寺田 賢二

北陸先端科学技術大学院大学 情報科学研究科

1998 年 2 月 13 日

キーワード： 強化学習、分類子システム、バケツリレーアルゴリズム、罰回避.

概要

自律ロボット等の主体に複雑な環境または動的な環境下で仕事を行わせる場合、予め作業環境を踏まえてロボットを設計する事が困難である。そのため主体には学習能力などの何らかの適応能力が求められる。また自律ロボット等による学習を考えた場合ロボット作成に高いコストがかかっているため、効率の良い仕事の達成よりも自己の保全が重要になることがある。よって本研究では強化学習により危険回避行動の獲得を目的とする。

強化学習はその名が示すように、正の報酬をもたらす行動を「強化」することにより学習を進める。そのため強化学習による従来の研究のほとんどが危険回避 (罰回避) は重要視ぜず、危険回避行動の学習は直接罰を受ける行動の評価値を下げることによる相対的な強化しか行わない。これは、学習主体が環境の探索をつづけていけば相対的に強度が高くとも、いずれはその回避行動は無効ルールとして処理されてしまうことを意味している。よって危険回避行動への直接的な強化を行うことで危険回避行動を行うルールを有効なルールと判断させることが本研究の基本方針である。

提案する手法は経験強化型の強化学習である分類子システムにおいて用いられている報酬伝搬処理のバケツリレーアルゴリズムをもとにして、それに危険回避行動に対して明示的な報酬を与える処理を追加したものである。

しかし目的とするタスクとは別に報酬を新たに設定する事は、学習主体が新たに設定した報酬を得ることのみで満足をして、目的とするタスクを得るための探索を行わなくなる問題を抱えている。この問題を解決するために危険回避の報酬と同じ量の負の報酬の伝搬をさせる。しかし正値の報酬が実際に選択した行動の評価値を伝搬させるのに対して負値の報酬を伝搬は、どの負値の評価値を伝搬させるかの選択が難しい。そこで、実際に行っ

た行動の評価のみを伝搬させるのではなく、その状態で活性化した全ての行動の評価を伝搬させることにより、負値の報酬の伝搬を選択すること自体を本研究では行わない。

ルール生成処理における変更点は、提案手法では負値の報酬の伝搬を行うため、負値の評価値を持つルールを保持する必要がある。そのためルール生成処理に付随する廃棄ルールの決定はルールの評価値の低いルールではなく評価値の絶対値の低いルールを破棄することにした。また遺伝的アルゴリズムのみによるルール生成では強化学習時に新たなルール生成ができない。これは学習主体が危険な状態 (学習主体がとることのできる行動のうち罰をうける行動が含まれている状態) にいるとき、その状態に適合する分類子がない場合遺伝的アルゴリズム起動時まで罰行動をとる可能性を減らすことが不可能であることを示している。よって本研究では適合する分類子が無い場合に随時その状態に適合するルールを生成する。

本研究ではルール生成処理に上記の変更を加え、3 種類のシミュレーション実験により 6 種類の報酬伝搬処理の比較を行った。実験に用いた 6 種類の報酬伝搬処理と 3 種類のシミュレーションを次に示す。

- 報酬伝搬処理の種類
  - 普通のバケツリレーアルゴリズム
  - 負の報酬伝搬のみを追加した手法
  - 負の報酬伝搬と危険回避に対する報酬を与える手法
  - 普通のバケツリレーアルゴリズムに活性化した全ての行動の評価を伝搬させる手法
  - 負の報酬伝搬を追加し活性化した全ての行動の評価を伝搬させる手法
  - 提案手法
- シミュレーションの種類と特徴

  燃料 (ゴミ) 拾い問題 正値の報酬がランダムに存在する。また、正値の報酬獲得のための行動系列のループは存在しない。

  迷路問題 危険な状態と隣合わせで成功報酬がある。また、正値の報酬獲得のための行動系列のループが存在する。

  宣教師と人喰い人問題 危険な状態を潜り抜けないと成功報酬にたどり着かない。また、正値の報酬獲得のための行動系列のループが存在する。

これらの実験により次のことがわかった。

- 提案手法は全ての実験において通常のバケツリレーアルゴリズムより失敗数が少なかった。この結果から提案手法に失敗数を減少させる能力があることを実験により検証した。

- 成功報酬ではゴミ拾い問題で探索能力の劣化がみられた。迷路問題, 宣教師と人喰い人問題では従来手法と同程度の成功報酬を獲得した。しかし危険回避行動に対する報酬を調整することにより、実験 1 でも従来手法と同程度の成功報酬の獲得が可能であった。この結果から今回の実験では提案手法が成功報酬を妨げないことを示した。
- 宣教師と人喰い人問題で、活性化した全ての行動の評価を伝搬する方式が選択された行動の評価のみを伝搬する方式と比較して成功数で大きな差を示した。この結果から付け値の全てを伝搬させることにより状態全体の評価をする方式が探索空間を広げる可能性があることがわかった。