修 士 論 文

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

指導教官　國藤進 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

寺田賢二

1998 年 2 月 13 日

要 旨

別紙

目 次

1 序論 **1**

1.1 はじめに 1

1.2 研究の背景 2

1.2.1 強化学習 2

1.2.2 強化学習による危険回避行動 3

1.3 研究の特色 3

1.4 研究の目的 4

1.5 本論文の構成 4

2 危険回避行動の強化を行う学習法 **5**

2.1 強化学習の枠組 5

2.2 分類子システム 5

2.2.1 報酬伝搬処理：バケツリレーアルゴリズム 7

2.2.2 ルールの生成：遺伝的アルゴリズム 8

2.3 実験に用いる学習器 9

2.3.1 ルール生成の変更点 11

3 実験 **13**

3.1 実験手順 13

3.2 評価項目 13

3.3 実験 1:燃料 (ゴミ) 拾い問題 13

3.3.1 問題の特徴 13

3.3.2 タスク 14

3.3.3 設定 14
3.3.4 実験結果 17
3.4 実験 2:迷路問題 18
3.4.1 問題の特徴 18
3.4.2 タスク 19
3.4.3 設定 19
3.4.4 実験結果 20
3.5 実験 3:宣教師と人喰い人の問題 23
3.5.1 問題の特徴 23
3.5.2 タスク 23
3.5.3 設定 23
3.5.4 実験結果 24

4 考察と検討 **27**

5 結論 **33**

謝辞 **35**

A ルール生成に対する考察 **38**

図一覧

2.1 強化学習の枠組 6
2.2 付け値の計算法 (例) 7
2.3 分類子のランダム生成 12

3.1 作業空間 (学習開始直後) 15
3.2 センサ入力と文字列との対応 16
3.3 衝突回数の累積値 (平均値):実験 1 17
3.4 収拾した燃料の累積値 (平均値):実験 1 18
3.5 作業空間:迷路 (学習開始直後 19
3.6 迂回経路と行動系列のループ 21
3.7 燃料の収集数 (平均値):迷路問題 21
3.8 障害物への衝突回数 (平均値):迷路問題 22
3.9 障害物への衝突回数の累積値 (平均値):迷路問題 22
3.10 作業空間 (学習開始直後) 25
3.11 失敗回数 (平均値):宣教師問題 25
3.12 成功回数 (平均値):宣教師問題 26

4.1 罰報酬と隣接しているときの状態 29
4.2 成功数の累計値：ゴミ拾い問題 29
4.3 衝突数の累計値：ゴミ拾い問題 30
4.4 成功数：迷路問題 30
4.5 衝突数の累計値：迷路問題 31
4.6 衝突数：迷路問題 31
4.7 成功数：宣教師と人喰い人問題 32

図一覧

4.8　失敗数：宣教師と人喰い人問題 32

表一覧

第 1 章

序論

1.1 はじめに

近年、未知の環境や人間が到達不可能な環境下で作業を行う自律ロボットが注目されている。火星探査機「Mars Pathfinder」[1] に搭載された移動探査ローバー「Sojouner」は部分的に自律であるロボットではあるが、その代表例であろう。

ローバーは地球にいるオペレーターの指示により制御される。しかしながらオペレーターとローバーとの間の通信には 6 分から 41 分の遅延があるために、ローバーは自律制御を行うことによって通信による遅延を補っている。今回の火星探査では、つぎのような理由からローバーの自動制御はシンプルなアルゴリズムで目的を達成する事ができた。

- ローバーのは基本的にオペレータの指示に従って行動する。
- 目的達成のためにローバーに複雑な行動が求めらない。
- ローバー以外の環境に作用できる物が少ない。

しかし将来的にオペレーターを介さない行動や複雑な行動、そして複数のローバーによる協調作業を考えた場合、学習等の何らかの適応能力が必要となるであろう。

また、このローバーのように非常にコストをかけている物には、効率の良い仕事の達成よりも自己の保全が重要になる場合がある。

よって本論文では強化学習を題材とした危険回避行動の学習をあつかう。

1.2 研究の背景

今回研究対象とした強化学習は、数年前から自律ロボット等の能動的な主体に対する学習法の一つとして注目されている [2][3][4]。これは強化学習が下記のような利点を持つからである。

- サブサンプションアーキテクチャ[5] のような環境との相互作用を重視する考えと相性が良い。
- 不確実性と報酬遅れを伴う弱い情報源から学習可能である。

1.2.1 強化学習

強化学習 (reinforcement learning)[6] とは、学習させたいタスクの達成状態として設定された報酬を手掛かりに、環境に適応する学習方式である。強化学習はその学習方針により 2 つに大別できる。一つは環境を同定することにより最適な行動系列の獲得を目的とする環境同定型 (explorantion oriented) の強化学習で、代表的なものに Q 学習が挙げられる。もう一つは正の報酬を受け続けることを目的とする経験強化型 (exploitation oriented) の強化学習である。両者とも自律ロボット等の学習に用いられ、各々下記のような利点・欠点を持つ。

- 環境同定型

  利点 マルコフの環境では学習が終了した時点での最適な行動系列を知ることが可能である。

  欠点 試行錯誤により環境を同定するために、学習時にあらゆる行動を行う必要がある。よって学習主体は罰報酬を与えられる危険な行動も行わなければならない。また環境の複雑さの従った膨大な量の試行錯誤も必要とする。

- 経験強化型

  利点 環境同定型の強化学習と比較すると学習が収束するのに要する試行錯誤が少ない。

  欠点 学習の目的が継続的に報酬の得られる行動パターンの獲得であるため、その行動系列が最適な系列で無くても満足してしまう。

本研究では危険回避を重要視して考えるため、学習システムは罰報酬に直接繋がる行動も行う必要のある環境同定型の強化学習を避け、経験強化型の強化学習の一手法である分類子システム (classifier system)[11] をもとにする。

1.2.2 強化学習による危険回避行動

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) は重要視せず、危険回避行動の学習は直接罰を受ける行動の強化値を下げることによる相対的な強化しか行わない。危険回避を取り上げた研究としては、罰ルールを判定し罰回避政策形成アルゴリズム [7] や罰状態脱出政策形成アルゴリズム [8] により罰回避の政策を獲得する手法や倒立振子などを扱った研究 [9] がある。

前者の罰ルールを判定する手法では、罰ルール判定の十分条件の 1 つを「直接罰を得た経験のあるルール」としているため、ルールの条件部における抽象化が許されない。例えば学習ロボットを想定した時、「全ての状態に対して前進する」というルールが存在したなら。そのルールは壁に衝突することによって罰ルールとなる。この罰ルールは全ての状態で活性化するため、他のルールは全て罰状態に隣接していることになる。この結果、罰ルール判定の十分条件の「選択されたルールが罰ルールのみである」により他のルールが罰ルールとして判定される可能性が高くなる。これらのことから罰ルールを判定し処理を行う方式は状態空間が大きくなった時に対応が困難であることが予想される。後者は罰回避状態が目的となっているため。罰回避をしつつ正値の報酬を獲得することについては取り扱われていない。

学習主体のシステムに差があるため、比較しづらいが本研究の基本方針と同じ危険回避行動に報酬を与える処理をおこなっている研究 [10] もある。この研究での方式を普通の分類子システムに組み入れると、罰報酬が大きい時に危険回避に 1 度に与えられる報酬が大きいためその回避行動に固執してしまう問題が予想される。

1.3 研究の特色

このような背景から大きな状態空間で対応可能であり、更に罰報酬と成功報酬が混在する環境において危険回避行動が獲得できる学習システムを提案する。その手法は経験強化

型の強化学習であるバケツリレーアルゴリズム [11] を報酬配分に用いた分類子システムをもとにして危険回避行動に対して明示的な報酬を与えることにより学習主体の負の報酬の回避を行わせる。

1.4 研究の目的

危険回避行動に対して報酬を与えることにより罰報酬の減少が可能であることを実験的に検証する。また、考えられる問題点として、提案する手法は環境から得られる正値の報酬とは別に学習主体の内部で報酬を作り出す。そのため主体内部で作り出される報酬を受けることで満足をして、環境からの正値の報酬を得るための探索を行わない可能性がある。すなわち正値の報酬の獲得に対する性能劣化が考えられる。よって性能劣化の程度をシミュレーション実験を行うことによって確かめる。

1.5 本論文の構成

本論文は本章も含め 5 章から構成される. 2 章では提案した学習システムについて説明する。3 章では 2 章で述べた学習システムによるシミュレーション実験について述べる。4 章では実験結果をふまえて、提案手法に対する検討を行なう。最後に本論文の結論と今後の課題について 5 章で述べる.

第 2 章

危険回避行動の強化を行う学習法

2.1 強化学習の枠組

未知なる環境に置かれたロボットのような学習主体を考えた場合、強化学習の概念的枠組は図 2.1 のようになる [6]。この枠組によると強化学習を行う学習主体は 3 つの構成要素からなる。各構成要素の役割を次に示す。

状態認識器 感覚入力から行動の候補となる競合集合を作る。

行動選択器 競合集合から行動を選択する。

学習器 報酬にしたがって状態認識器がより適切な行動の候補を生成するように更新する。

2.2 分類子システム

本研究で学習システムのもととした分類子システムによる強化学習も図 2.1 の枠組で表現することができる。次に分類子システムの構成要素を強化学習の枠組に沿って述べる。

- 状態認識器

  入力 感覚入力から {0,1} の 2 種類の文字による固定した長さ k の文字列が得られる。

  構成要素 分類子と呼ばれるルールを多数持ち、それらをルールベースにより蓄える。分類子の条件部は {0,1,#} の 3 種類の文字の固定された長さ k の文字列

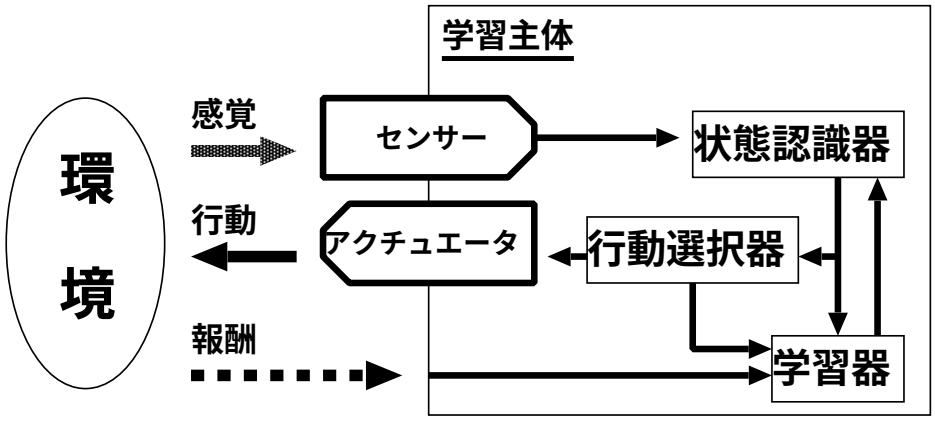

図 2.1: 強化学習の枠組

で構成される。ここで#は「0 でも、1 でもどちらでもよい (don't care)」ということを示している。動作部は同様に {0,1,#} の 3 種類の文字で構成されることもあるが、本研究では#は使わず {0,1} の 2 値で構成した。また動作部での#は「通り抜け (pass through)」要素となり、対応した入力メッセージの内容をそのまま出力メッセージに出力するのに用いられる。

処理 メッセージに適合するルールベース内の分類子をすべてマークする。

出力 マークされた分類子を競合集合として行動選択器に渡す。

- 行動選択器

  入力 状態認識器から競合集合を受け取る。

  処理 入札 (2.2.1 参照) により競合集合から分類子を 1 つ選択する。

  出力 選択された分類子の動作部内容をアクチュエータに渡す。

- 学習器 学習器はルールの信頼値配分とルール生成の 2 つの機能により分類子の信頼度の調節と生成を行う。前者を 2.2.1 の報酬伝搬処理で述べ、後者を 2.2.2 のルール生成で述べる。

```
分類子Cの付け値Bid(C,t)は分類子Cの
特殊度R(C)・強度S(C,t)・定数k
により決定され、それぞれ以下の式で表される。
Bid(C,t) = kR(C)S(C,t)
R(C) = (条件部の文字数ー条件部の#の数) / 条件部の文字数
S(C,t) = その分類子の強度
k = 0〜1の定数。本稿では0.1を使用
(例 )      LHS                RHS   強度
          10#100###0      01      600
R(C): ( 10 - 4 ) / 10 = 0.6              S(C,t): 600
Bid(C,t): 0.1 x 0.6 x 600 = 36
```

図 2.2: 付け値の計算法 (例)

2.2.1 報酬伝搬処理：バケツリレーアルゴリズム

分類子システム [11][12] における信頼値配分はバケツリレーアルゴリズムによって行われる。それは各々の分類子に強度 (strength) を割り当て、その強度を修正することによって行われる。

まず分類子の強度は、その分類子による行動が環境によって予め設定された「報酬」を得る動作に直接関与する時、報酬の値に従って修正される。しかし、この処理だけでは報酬を生む状態を作り出した分類子に報酬を割り当てることができない。この問題をバケツリレーアルゴリズムによって解消している。その手続きは行動選択器で行われる入札処理に付随して行われる。入札による競合解消は、まず競合集合として集められた分類子に付け値 (bid) が付ける。付け値は次の 2 つのパラメータに比例した大きさとなる。1 つは過去の経験による分類子の有用性で、分類子の持つ強度によって決定される。もう 1 つは分類子の特殊性で分類子の条件部の非#の記号の数を条件部の長さで割り計算される。その計算式と例を図 2.2 に示す。そして最大の付け値を提出したものを入札の勝利者とする戦略 [12] や付け値の大きさに比例した確率で勝利者を決定する戦略 [11] により勝利者となった分類子を選択することで競合解消を行う。

強度の修正は付け値を用いた以下の 2 つの式で行われる。まず入札により勝者となった

分類子 C が強度 S(C,t) を付け値の量 B(C,t) だけ減らされる。すなわち

$$S(C, t+1) = S(C, t) - B(C, t)$$

となる。一方勝者となった分類子 C に適合する状態を作り出した前回の入札処理の勝利者 C^w は、C の付け値の分だけ、強度が増やされる。すなわち

$$S(C^w, t+1) = S(C^w, t) + B(C, t)$$

この報酬の受渡しの繰り返しにより強度の修正をして行動系列の獲得を可能にしている。

2.2.2 ルールの生成：遺伝的アルゴリズム

分類子システムでは全ての状態と動作の対についての分類子を保持しているわけではない。そのため利用できる規則が不十分であるとき新しいもっともらしい規則を生成する必要がある。そのため遺伝的アルゴリズム [13][14] により規則を生成する。その処理を次に示す。

選択 分類子の強度によりルール生成のもととなる親の対を選択する。

交叉 選択した親の対に交叉処理を行い新しいルールを生成する。

突然変移 交叉により生成されたルールに対して突然変移処理を行う。

淘汰 最も重要度 (強度) の低い分類子を子孫によって置き換える。

本研究では次のような設定で遺伝的アルゴリズムによるルール生成を行う。

- 本研究では条件部と動作部を連結したものを一つの遺伝子とする。
- 交叉処理は 2 点交叉を用いて 1 度の交叉で 2 つの遺伝子を作成する。ただし全ての分類子中の最大の強度が初期値 (500) より少し大きい値 (700) 未満の場合は、新たに 2 つの分類子をランダム生成する。またルーレット選択の相手となる分類子の強度が低い場合 (正の強度を持つ分類子の強度の総和が 100 未満の場合) は交叉の相手はランダムに生成した分類子とする。

2.3 実験に用いる学習器

2.2 節で述べた分類子システムの枠組では、危険回避行動の強化は罰を与えられる行動に対しての相対的な強化である。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による回避行動をとる分類子の無効ルール化を防ぐことが本研究の基本方針である。本研究では次に示す 6 種類の報酬伝搬処理を用いてシミュレーション実験を行い。危険回避行動への直接的な強化による報酬伝搬と他の報酬伝搬法の性能を比較する。

BB Bucket Brigade

通常のバケツリレーアルゴリズムであり入札時における報酬の伝搬は次の 2 つの式による伝搬のみである。

$$S(C, t+1) = S(C, t) - B(C, t)$$

$$S(C^w, t+1) = S(C^w, t) + B(C, t)$$

NP Negative Propagation

上記の BB の処理に次に示す負値の付け値の伝搬を追加した。

- 入札時に負値の付け値を提出する分類子があった場合、その最小値を提出した分類子 C^- に対して BB の勝者 C に対する処理と同様の処理を行う。つまり NP では BB の処理に次の処理を加える。

$$S(C^-, t+1) = S(C^-, t) - B(C^-, t)$$

$$S(C^w, t+1) = S(C^w, t) + B(C^-, t)$$

注 $B(C^-, t)$ は負値であるため $S(C^-, t+1)$ は増加し $S(C^w, t+1)$ は減少する。つまり共に強度は 0 に近付く。

この処理により、まるで学習主体が危険な状態に対して「恐怖心」を憶えたかのように危険な状態に近付くのを避けさせることを目的とする。

RA Reward for Avoidance

上記の NP の処理に次に示す回避行動への直接的な強化を行う。

- NP 処理の時に選択された C^-からの回避に対する報酬として $B(C^-,t)$ を勝者 C に与える。つまり RA では NP の処理にさらに次の処理を加えることになる。

$$S(C,t+1) = S(C,t) - B(C^-,t)$$

注 $B(C^-,t)$ は負値であるため $S(C,t+1)$ は増加する。

この処理により回避行動を価値のあるものとして強化することで回避行動を獲得させる。また、危険回避への報酬を BB の処理に追加せずに NP の処理に追加するのは、危険回避行動による成功報酬の獲得の性能劣化を防ぐためである。例えば「足踏み」を行うことの可能な学習ロボットが壁際で「足踏み」をした場合、ロボットは壁に衝突しなかったため危険回避の報酬を獲得する。しかし状態が変化をしていないためロボットは再び「足踏み」をして報酬を獲得することが可能である。このループを形成させないために負の報酬の伝搬が作用するからである。

NP,RA は正値の報酬の伝搬は実際に行った行動の期待値を伝搬するのに対して負値の報酬伝搬はその行動とは関係なく、その状態の最小値の負の強度を伝搬する。そのため行動の評価のバランスを保って伝搬させることを大きく妨げると考えられる。よって行動に対する評価値を伝搬するのを諦め、視点を変えて状態全体に対する評価値の伝搬を行うことによるバランスの悪化を防ぐ。状態全体に対する評価値の伝搬法の、上記の 3 種類の報酬伝搬処理との処理の違いを次に示す。

- 伝搬に用いる付け値は最初の 3 種類の伝搬法では勝者の付け値のみである。すなわち $B(C,t)$ と $B(C^-,t)$ のみである。この 2 つの値を正値の付け値の総和とに負の付け値の総和に変更する。付け値の総和が状態全体の信頼度を示していると見なし、これにより分類子の強度を次の状態の信頼度に対する評価値としている。この変更は入札時に正の付け値を提出した分類子の集合を $\{C^{+*}\}$ 負の付け値を提出した分類子集合を $\{C^{-*}\}$ とするならば次のようになる。

$$B(C,t) \rightarrow \sum_{c^+ \in \{C^{+*}\}} B(c^+,t)$$

$$B(C^-,t) \rightarrow \sum_{c^- \in \{C^{-*}\}} B(c^-,t)$$

すなわち残りの 3 種類の伝搬処理は次のようになる。

BB_AB Bucket Brigade (propagation of All Bids)

$$S(C^w, t+1) = S(C^w, t) + \sum_{c^+ \in \{C^{+*}\}} B(c^+, t)$$

全ての $c^+(c^+ \in \{C^{+*}\})$ に対して

$$S(c^+, t+1) = S(c^+, t) - B(c^+, t)$$

NP_AB Negative Propagation (propagation of All Bids)

$$S(C^w, t+1) = S(C^w, t) + \sum_{c \in \{\{C^{+*}\}, \{C^{-*}\}\}} B(c, t)$$

付け値を提出した全ての分類子 $c(c \in \{\{C^{+*}\}, \{C^{-*}\}\})$ に対して

$$S(c, t+1) = S(c, t) - B(c, t)$$

RA_AB Reward for Avoidance (propagation of All Bids)

$$S(C^w, t+1) = S(C^w, t) + \sum_{c \in \{\{C^{+*}\}, \{C^{-*}\}\}} B(c, t)$$

$$S(C, t+1) = S(C, t) - \sum_{c^- \in \{C^{-*}\}} B(c^-, t)$$

付け値を提出した全ての分類子 $c(c \in \{\{C^{+*}\}, \{C^{-*}\}\})$ に対して

$$S(c, t+1) = S(c, t) - B(c, t)$$

2.3.1 ルール生成の変更点

まず従来のバケツリレーアルゴリズムでは強度の小さい分類子が重要度の低い分類子として考えられルール生成処理で廃棄されていた。しかし報酬伝搬法NP,RA,NP_AB,RA_ABでは負値の強度も使用するため、強度の値がそのまま分類子の重要度として結び付かない。よってNP,RA,NP_AB,RA_AB では分類子のもつ強度の絶対値を分類子の重要度とし

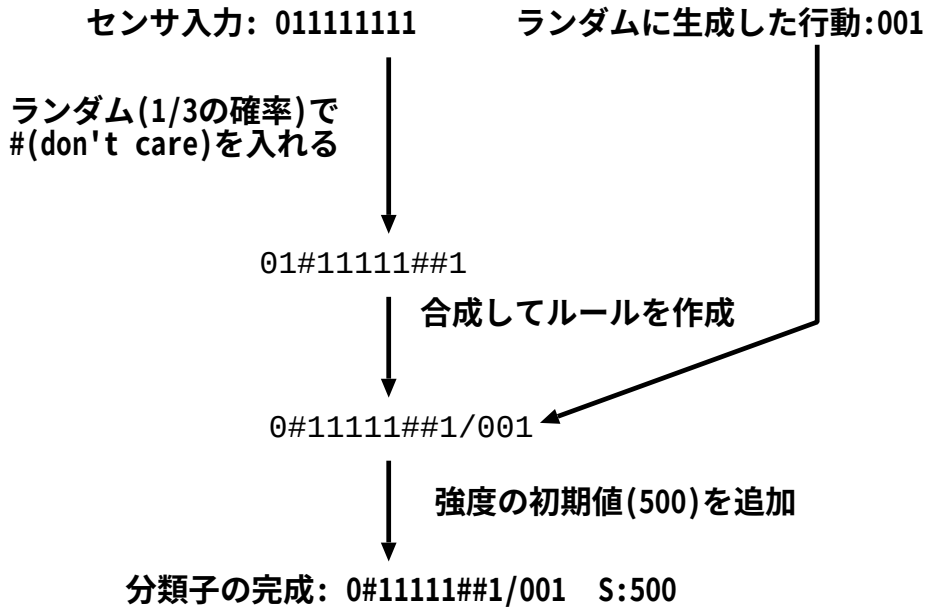

図 2.3: 分類子のランダム生成

て廃棄する分類子の選択を行う。ただし負値の強度を持つ分類子でルールベースが溢れないようにするため、負値の強度をもつ分類子の数は全体の分類子の半分までに制限する。

次に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。(付録 A 参照) よって本研究では適合する分類子が無い場合に図 2.3 に示すルール生成処理を行う。

第 3 章

実験

3.1 実験手順

2.3 節で述べた 6 種類の報酬伝搬法 (BB,NP,RA,BB_AB,NP_AB,RA_AB) を用いて 3 種類のシミュレーション実験を行う。実験 1 では成功報酬と危険回避の関係が薄い問題、実験 2 では危険な状態と隣合わせで成功報酬がある問題、実験 3 では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問題を取り上げ実験を行う。3 種類のシミュレーションとも、分類子の初期集団は全てランダムに生成する。

3.2 評価項目

成功報酬を受ける回数と罰報酬を受ける回数によって報酬伝搬法の比較を行う。比較に用いたグラフは 10〜20 回の試行の平均値である。

3.3 実験 1:燃料 (ゴミ) 拾い問題

3.3.1 問題の特徴

成功回数は、次に示す成功報酬の特徴のために作業空間上の探索能力を示すものと考えられる。

- 成功報酬は作業空間にランダムに広がっている。

- 1 度得た報酬はその場所からは無くなる。そのため成功報酬への行動系列は短く、3 STEP 程度である。

また成功報酬への行動系列が短く、成功報酬によるループを形成することは無い。そのため成功報酬への行動系列獲得が失敗数の削減に結び付かない。また行動系列が短いことは危険回避行動により成功報酬への行動系列を破壊する可能性をも少なくしている。

以上のことから提案する危険回避行動へ報酬を与える手法 (RA,RA_AB) に有利であると考えられる。

3.3.2 タスク

燃料を拾い集める学習主体によるシミュレーション実験を行った。ここでのタスクは図 3.1 で示す作業空間に散らばっている燃料を障害物を避けながら収拾することである。

3.3.3 設定

主体への入力 環境の状態を表現する長さ 21 の文字列がセンサを通じて獲得される。(図 3.2)

- 16 文字を使用した 8 方向の障害物を (隣接/近い/遠い/無し)4 段階で識別するセンサ
- 3 文字を使用した現在の位置、左、右の餌の有無を識別するセンサ
- 2 文字を使用した前方の物体を識別するセンサ

主体の行動 (燃料を拾う行動/左旋回/右旋回/前進) の 4 種類の行動を長さ 2 の文字列で持つ。

主体の持つ分類子の数 1000 個

報酬 報酬伝搬処理以外の報酬は以下の 3 種類がある。

- 正の報酬は燃料を拾ったときに 500 が与えられる。
- 負値の報酬は障害物に衝突した際に-2000 が与えられる。
- 行動のコストとして行動を行う毎に-10 が与えられる。

作業空間 30× 30 の大きさを持ち、濃い灰色で示される燃料はランダムに配置される。また黒色で示される障害物は常に同じ場所に存在する (図 3.1)。

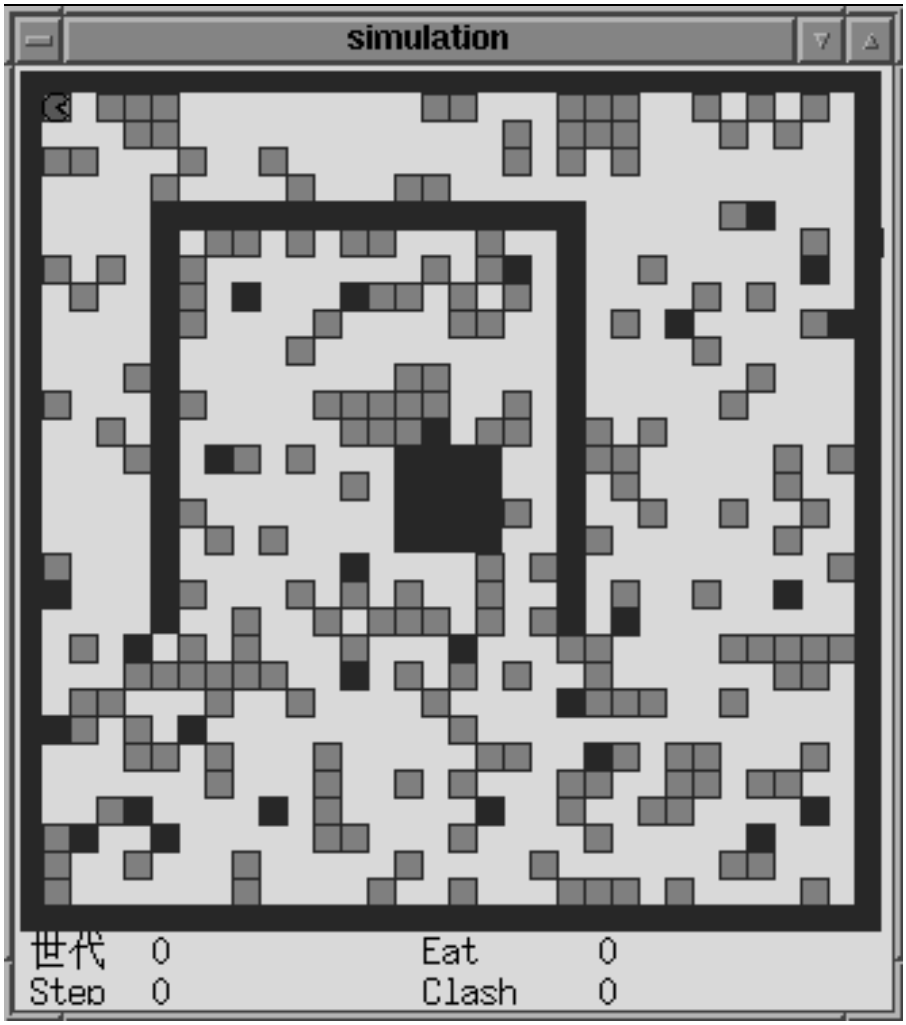

図 3.1: 作業空間 (学習開始直後)

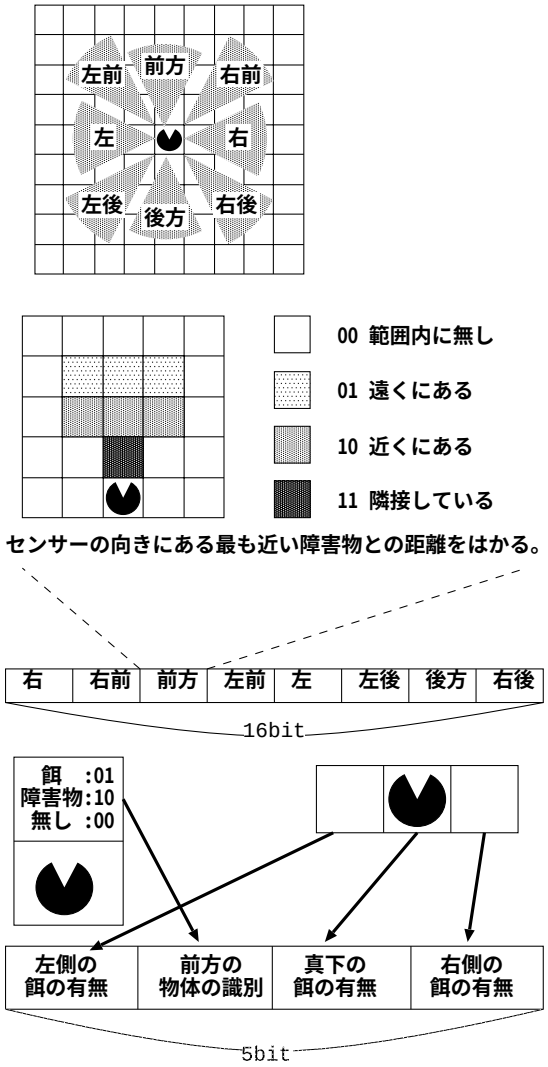

図 3.2: センサ入力と文字列との対応

遺伝操作の設定

交叉処理の親の選択 ルーレット選択法におより強度に比例した確率で選択

交叉方法 2 点交叉

突然変移率 0.05

起動間隔 主体が 50 回動作を行う毎に起動

分類子の生成と入れ替え 1 度の起動で全体の分類子数の 1 割を新たに生成し、重要度の低い分類子と入れ替える。

3.3.4 実験結果

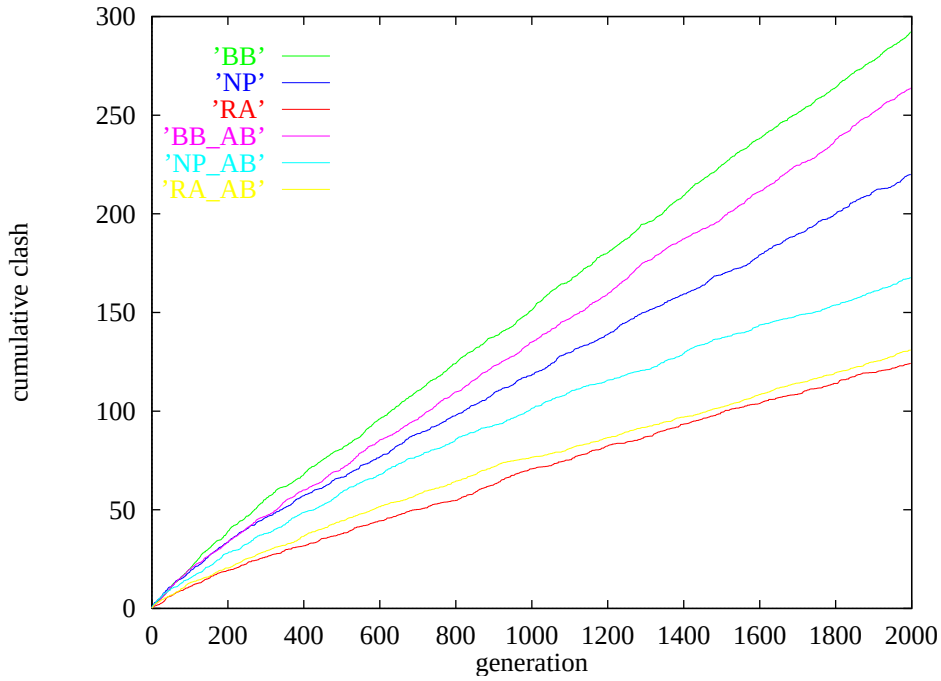

図 3.3: 衝突回数の累積値 (平均値):実験 1

図 3.3 の衝突回数では、負値の強度について操作をしない BB,BB_AB が最も衝突回数が多く、ついで負値の報酬伝搬を行う NP,NP_AB、そして最も衝突回数の削減に成功したのが危険回避行動に対して報酬を与える RA,RA_AB となった。RA,RA_AB では衝突回数を BB,BB_AB の半分以下に押さえることに成功している。

図 3.4 成功数では、成功数の多い順に NP_AB,NP,BB_AB,BB,RA_AB,RA となった。各方法の 100 回の成功報酬を受けるのに要した世代数とその世代数における失敗数の累積値を次に示す。

方法	NP_AB	NP	BB_AB	BB	RA_AB	RA
世代数	720	763	812	913	1008	1018
失敗数	79	93.6	110.3	138.6	76.95	71.6

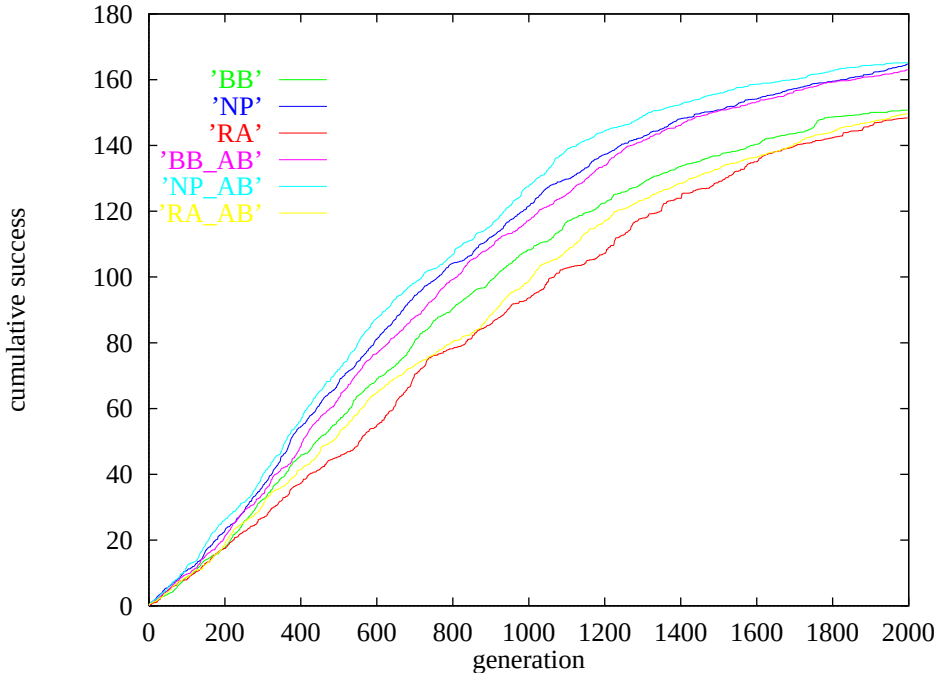

図 3.4: 収拾した燃料の累積値 (平均値):実験 1

危険回避行動に対して報酬を与える RA_AB,RA は BB と比較して 1 割程度多く世代数が必要となっている。また、最も成績の良い NP_AB は、BB と比べて 2 割程度少ない世代数で 100 回の成功数に達している。

3.4　実験 2:迷路問題

3.4.1　問題の特徴

この問題は成功報酬への行動系列による完全なループを形成することが可能である。そのため成功報酬への行動系列の獲得により失敗をしなくなる。成功報酬の最短のループは 35STEP である。険な状態と隣合わせで成功報酬がある。

3.4.2 タスク

学習主体のタスクは図 3.5 で示す作業空間上の障害物を避けながら燃料を拾うことである。

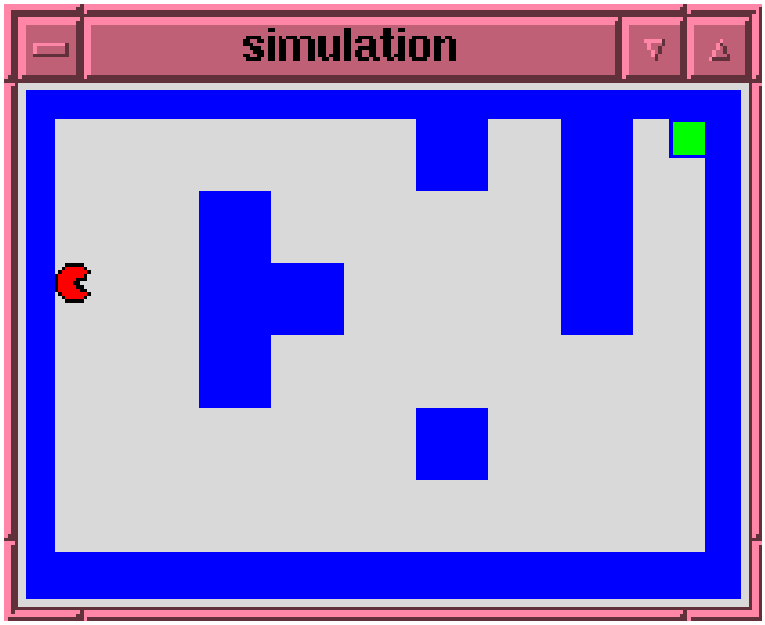

図 3.5: 作業空間:迷路 (学習開始直後

3.4.3 設定

設定のほとんどが実験 1 と同じである。

主体への入力 環境の状態を表現する長さ 21 の文字列がセンサを通じて獲得される。

- 16 文字を使用した 8 方向の障害物を (隣接/近い/遠い/無し)4 段階で識別するセンサ
- 3 文字を使用した現在の位置、左、右の餌の有無を識別するセンサ
- 2 文字を使用した前方の物体を識別するセンサ

主体の行動 (燃料を拾う行動/左旋回/右旋回/前進) の 4 種類の行動を長さ 2 の文字列で持つ。

主体の持つ分類子の数 1000 個

報酬 報酬伝搬処理以外の報酬は以下の 3 種類がある。

- 正の報酬は燃料を拾ったときに 500 が与えられる。
- 負値の報酬は障害物に衝突した際に-2000 が与えられる。
- 行動のコストとして行動を行う毎に-10 が与えられる。

作業空間 図 3.5 で示すように作業空間は灰色で示された燃料と円弧で示された学習主体、そして黒色の障害物 (壁) からなる。学習主体のスタート位置及び燃料と障害物の位置は図 3.5 に示した位置に固定されている。学習主体が燃料を拾った時点で初期状態へ再配置される。また、迷路の形は文献 [15] を参考に作成した。

遺伝操作の設定

交叉処理の親の選択 ルーレット選択法におより強度に比例した確率で選択

交叉方法 2 点交叉

突然変移率 0.05

起動間隔 主体が 10000 回動作を行う毎に起動

分類子の生成と入れ替え 1 度の起動で全体の分類子数の 1 割を新たに生成し、重要度の低い分類子と入れ替える。

3.4.4 実験結果

ここでも衝突回数は負値の強度に対して操作を行わない BB,BB_AB が最も多く、NP_AB,AB を間に挟んで RA,RA_AB が最も衝突回数が少なかった。成功数では全ての方式で行動系列の獲得を成功させている。しかし、NP,RA は迂回経路 (図 3.6) を多く使用しているため成功数に差がでた。BB,BB_AB,NP_AB,RA_AB では成功数の差は僅かで、殆どの学習方式が 39STEP の報酬への行動系列のループ (図 3.6) を作成していた。

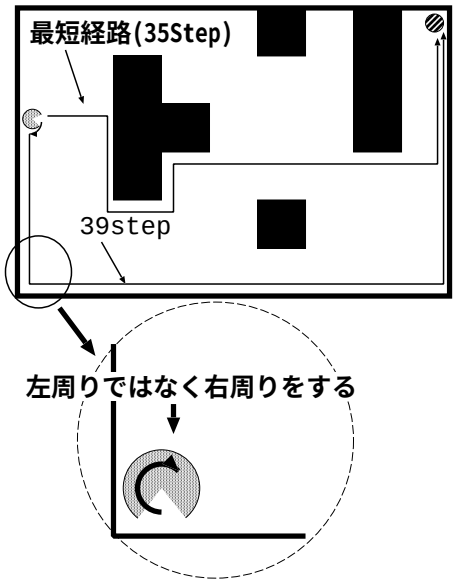

図 3.6: 迂回経路と行動系列のループ

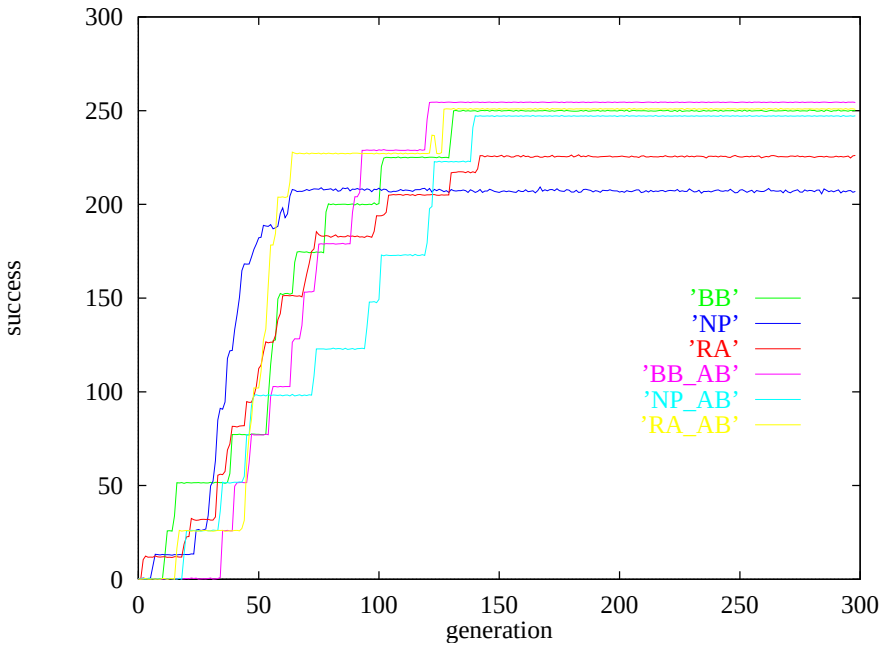

図 3.7: 燃料の収集数 (平均値):迷路問題

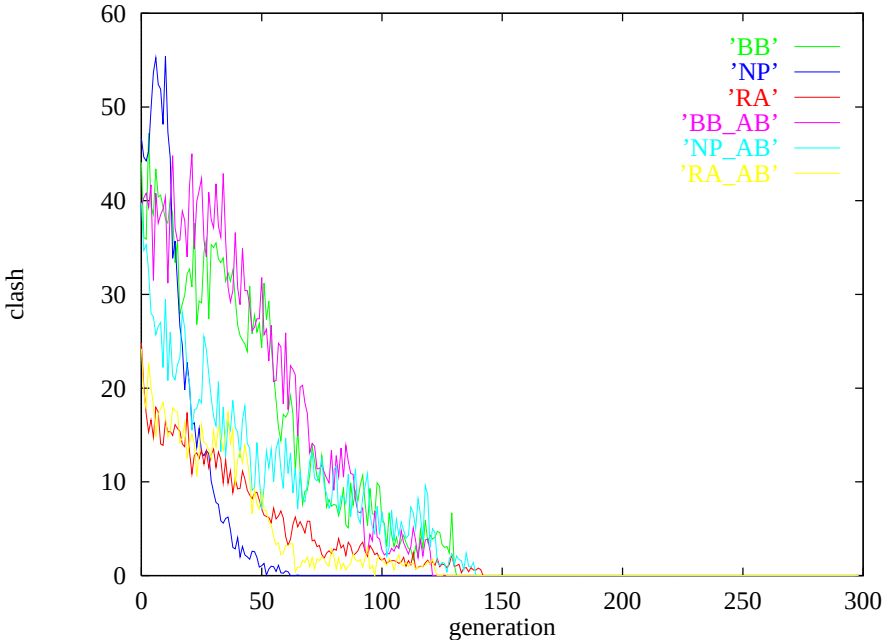

図 3.8: 障害物への衝突回数 (平均値):迷路問題

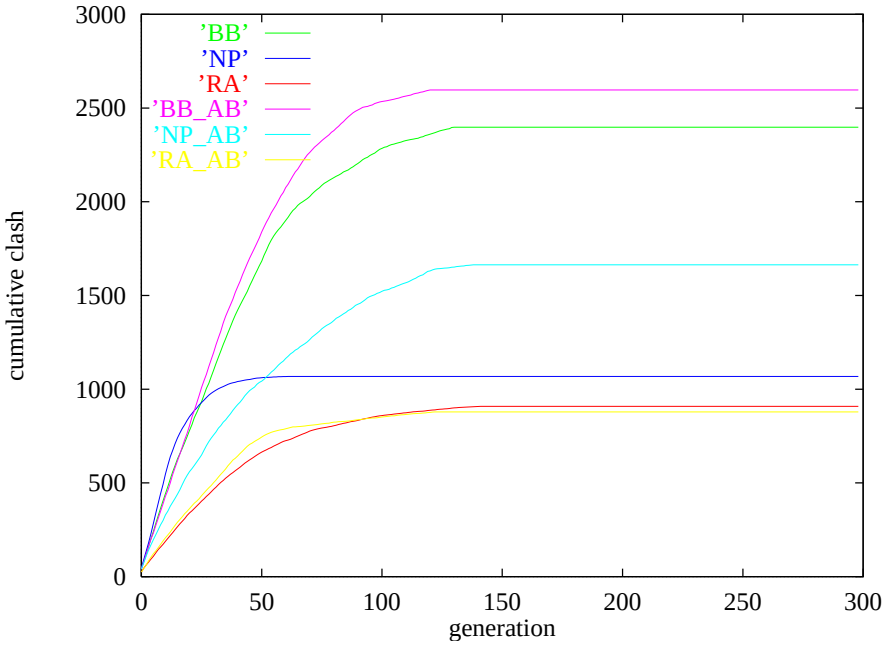

図 3.9: 障害物への衝突回数の累積値 (平均値):迷路問題

3.5 実験 3:宣教師と人喰い人の問題

3.5.1 問題の特徴

成功報酬への行動系列はループを形成することができ、ループ獲得により環境から罰報酬を受けることは無くなる。成功報酬の最短のループは 11STEP である。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。

3.5.2 タスク

宣教師と人喰い人の問題 (Missionaries-and-CannibalsProblem)[16] は次に示すような探索問題である。

川の左岸に 3 人の宣教師と 3 人の人喰い人がいる。これを 1 そうの舟を使って全員を右岸に移動させる問題である。ただし移動はつぎに示す制約に縛られる。

- 舟は小さく定員は 2 人である。
- 川の左岸・右岸・舟の上のいずれにおいても宣教師は人喰い人より人数が少なかった場合食べられてしまう。

3.5.3 設定

システムへの入力 右岸の宣教師の数 (2 進数:2bit)
右岸の人喰い人の数 (2 進数:2bit)
舟の位置 (左 (0) 右 (1):1bit)

システムの行動 宣教師 (2 進数:2bit) と人喰い人 (2 進数:2bit) の移動人数

システムの持つ分類子の数 1000 個

報酬 タスク達成時に 10000 がを与えられる。
宣教師が食べられた時に-1000 が与えられる。
行動のコストとして行動を行う毎に-10 が与えられる。

作業空間 宣教師と人喰い人問題は右岸の宣教師の数 (横軸) と人喰い人の数 (縦軸) を 2 次元表現した図 3.10 に示す迷路問題に置き換えることができる。この迷路上を移動する学習主体は次のような制約に従う。

- 迷路を進む主体は図 3.10 の左上に示す白丸から出発し右下の丸に到達した時点で成功報酬を得る。つまり左上の白丸が全員が左岸にある状態であり、右下の丸が全員が右岸にいる状態である。
- 中央の青色の地点は、その地点に主体が来た時罰報酬をうける状態であり宣教師が食べられてしまう状態を示している。
- 主体は移動方向を右下の方向と左下の方向の交互に変化させる。移動量は 1 マスか 2 マスに限定される。
- 周囲を囲む茶色は移動不可能状態を示しており、主体はその状態へ移る行動をとることができない。

また成功報酬を受けた場合は再びスタート位置 (左上) に戻り、シミュレーションを続ける。

遺伝操作の設定

交叉処理の親の選択 ルーレット選択法におより強度に比例した確率で選択

交叉方法 2 点交叉

突然変移率 0.05

起動間隔 主体が 10000 回動作を行う毎に起動

分類子の生成と入れ替え 1 度の起動で全体の分類子数の 1 割を新たに生成し、重要度の低い分類子と入れ替える。

3.5.4 実験結果

成功数、失敗数共に全ての付け値を渡す方式が優れていた。100 世代以内で唯一成功数の平均値が最適値の 909 回に届いた。

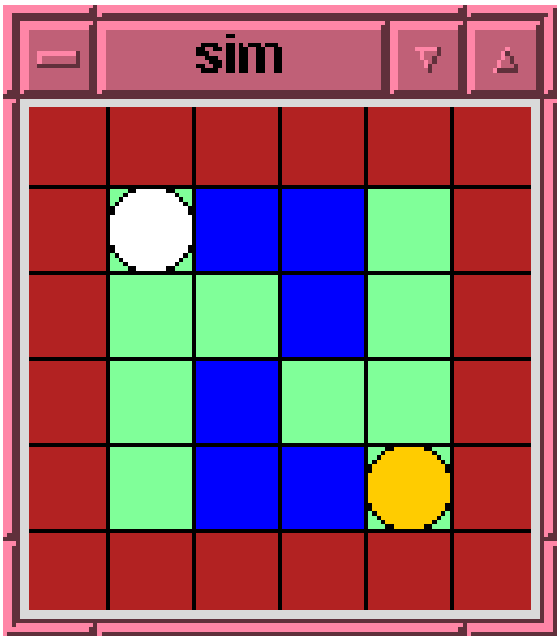

図 3.10: 作業空間 (学習開始直後)

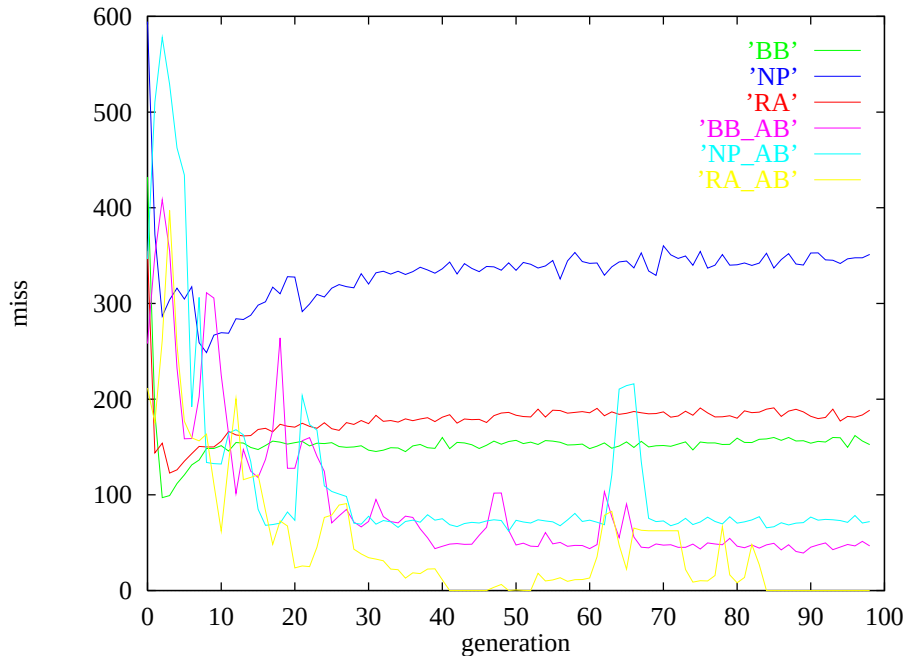

図 3.11: 失敗回数 (平均値):宣教師問題

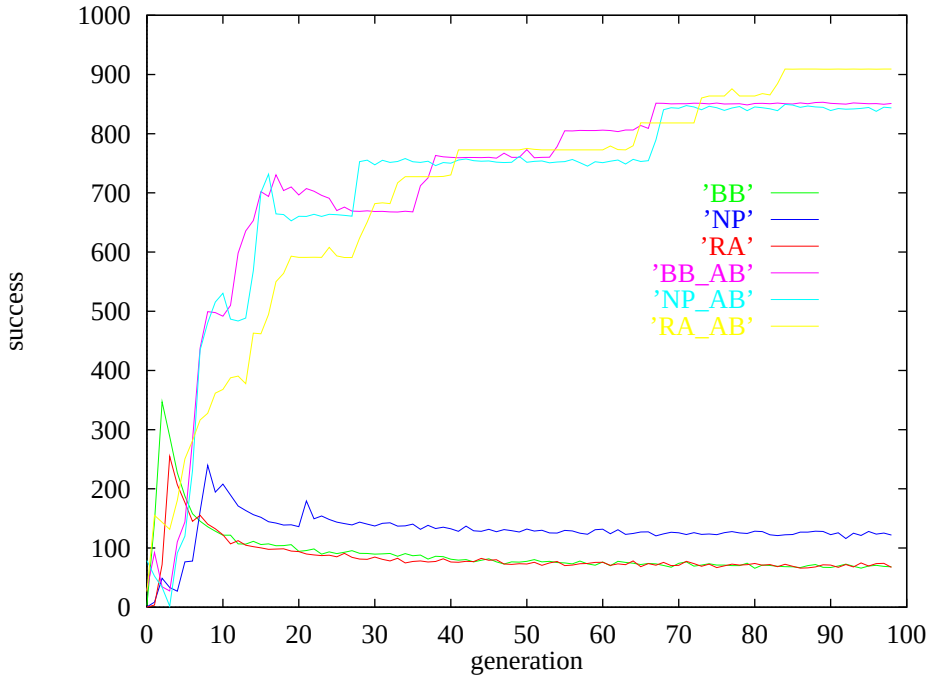

図 3.12: 成功回数 (平均値):宣教師問題

第 4 章

考察と検討

失敗数において実験 1 の燃料拾い問題・実験 2 の迷路問題では期待通りの結果となり、危険回避行動に対して報酬を与える方式 (RA,RA_AB) が最も失敗数が少く、次に負値の報酬伝搬を行う方式 (NP,NP_AB) が続き、最も悪いのが負値の強度について処理を行わない (BB,BB_AB) となった。成功数では実験 1 において危険回避行動に報酬を与える方式が最も成功数が少なかった。この原因は、学習主体が危険回避行動をとることのみで満足してしまう状態を防ぐことを失敗しているからである。つまり危険回避行動による正のタスク達成の妨害を妨げる役割とした負値の報酬伝搬が効いていないことがわかる。例えば図 4.1 のような環境とルール保持を想定し、その時状態 B で eat 行動を取り続けることで n 時間後の eat 行動を支持する分類子 C_Eの強度は clash の行動を支持する分類子 C_C から以下のような影響をうける。

$$S(C_E, t+n) = 2000kR(C_C) - 2000(1 - kR(C_C))^n$$

この式は分類子 Cc により本来強化されるべきでないと考えられる行動 eat の動作が強化されることをしめしている。このことから危険回避行動に対する報酬を一定の割り引き率で減らすことにした。割り引き率 0.9 にした RA を RA_r 同じく RA_AB を RA_AB_r とし、次の表と図 4.2〜図 4.8 を得た。

方法	NP_AB	NP	BB_AB	BB	RA_AB	RA	RA_r	RA_AB_r
世代数	720	763	812	913	1008	1018	863	845
失敗数	79	93.6	110.3	138.6	76.95	71.6	53.6	62.1

表:100 回成功するまでにかかった世代数とその世代までの失敗数

このように変更することにより、実験 1 においても BB と比較して性能劣化はみられなかった。また負値の報酬伝搬を与える方式が最も成功数が多かったのは、正値の報酬に辿りつくにはセンサーがエネルギーを発見している状態以外では、ランダムに行動するほうが良い。そのため、エネルギーがセンサーで発見されていない状態の分類子の強度を比較的素早く減らせる負値の報酬伝搬を行う方式が最も良い成功数を得た。

実験 2 の成功数では RA,NP では、正値の報酬へ至る行動系列のループに迂回経路を多く入れてしまっている。全ての付け値を伝搬させる方式 (BB_AB,NP_AB,RA_AB) や負の報酬伝搬を行わない方式 (BB,BB_AB) が迂回経路が少ないことを考えると、RA,NP の報酬伝搬処理では、正値の報酬がルーレット選択法によって実際に選ばれた勝者の行動の付け値が伝搬されるのに対し負値の報酬は最小の付け値が伝搬されることや勝者の行動と負値の行動が異なる場合でも負値の報酬が伝搬されるといったことが、行動の評価を伝搬するバランスを崩しているからであると考えられる。しかし実験 2 においても割引を導入することにより RA は他の手法と同程度の性能になった。

実験 3 は実験 1,2 とは異なる特徴を示し、負値の報酬の取り扱いの差よりも付け値を全部伝搬させることにより状態全体の評価値を伝搬させる方式 (BB_AB,NP_AB,RA_AB) と、勝者となる 1 つのみを伝搬させることにより行動の評価値を伝搬させる方式 (BB,NP,RA) の差が顕著に出た。これは状態全体の評価をする方式では主体が試行すべき行動が多数存在する場合、それだけ多くの報酬が伝搬され、その状態にたどり着きやすくなるからであると考えられる。つまり罰状態に囲まれ罰状態を通り抜けるパスが獲得されにくい時、もし、そのパスを通り抜けることによって新たな広い探索空間存在するならそのパスに対して報酬が伝搬されやすい。これを宣教師問題に当てはめるなら宣教師を 3 人右岸に渡らせる処理は罰状態に囲まれ獲得され難い。そのため宣教師が 3 人の状態に対するルールを持つ分類子はルールベース上に多く残っている。そこに宣教師が 3 人いる状態にたどり着いたなら、その行動を取った分類子はより多くの報酬を得ることができる。すなわち宣教師を 3 人右岸に移す報酬が効率的に獲得されるのである。

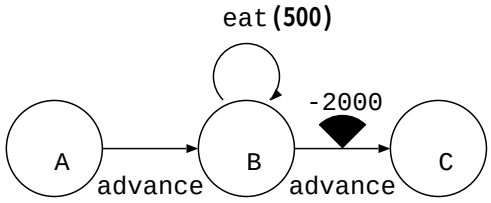

図 4.1: 罰報酬と隣接しているときの状態

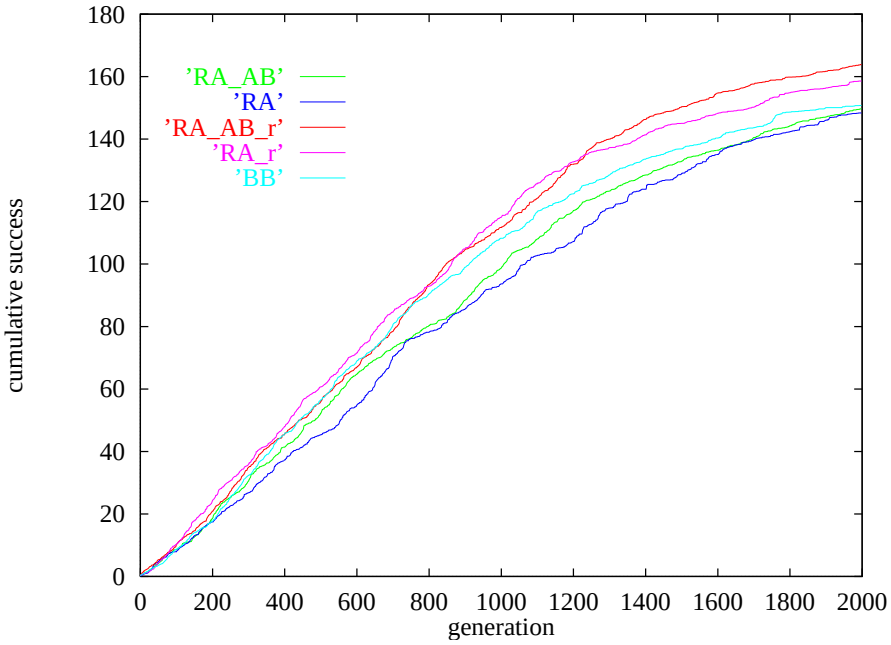

図 4.2: 成功数の累計値：ゴミ拾い問題

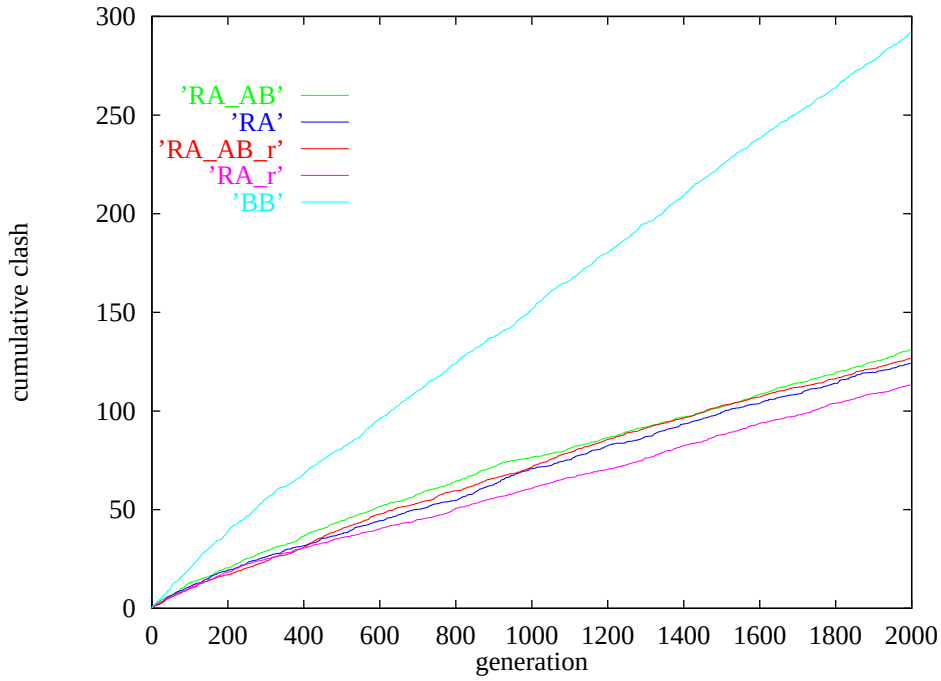

図 4.3: 衝突数の累計値：ゴミ拾い問題

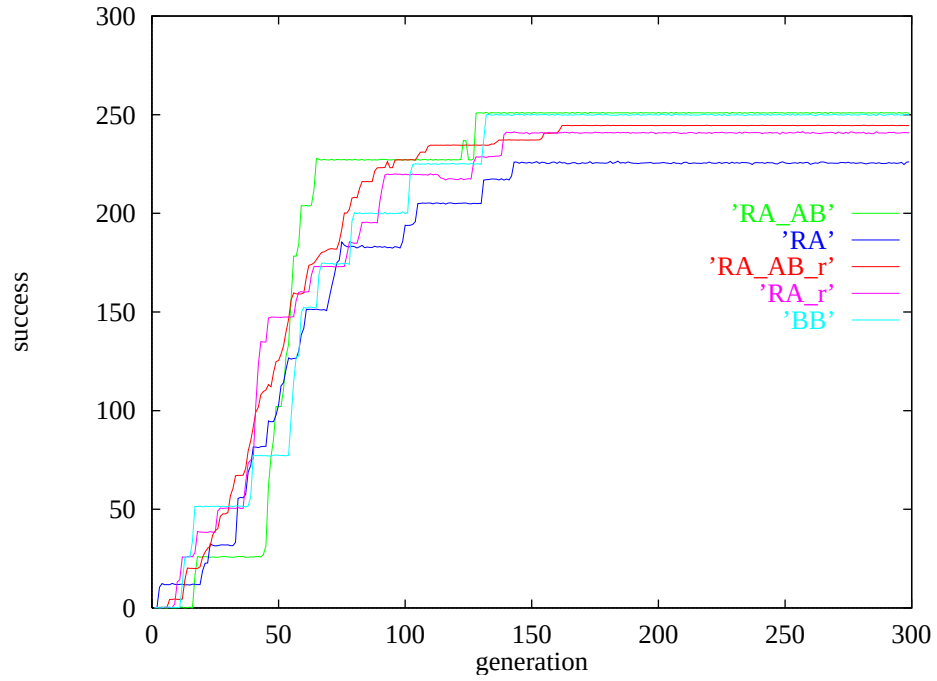

図 4.4: 成功数：迷路問題

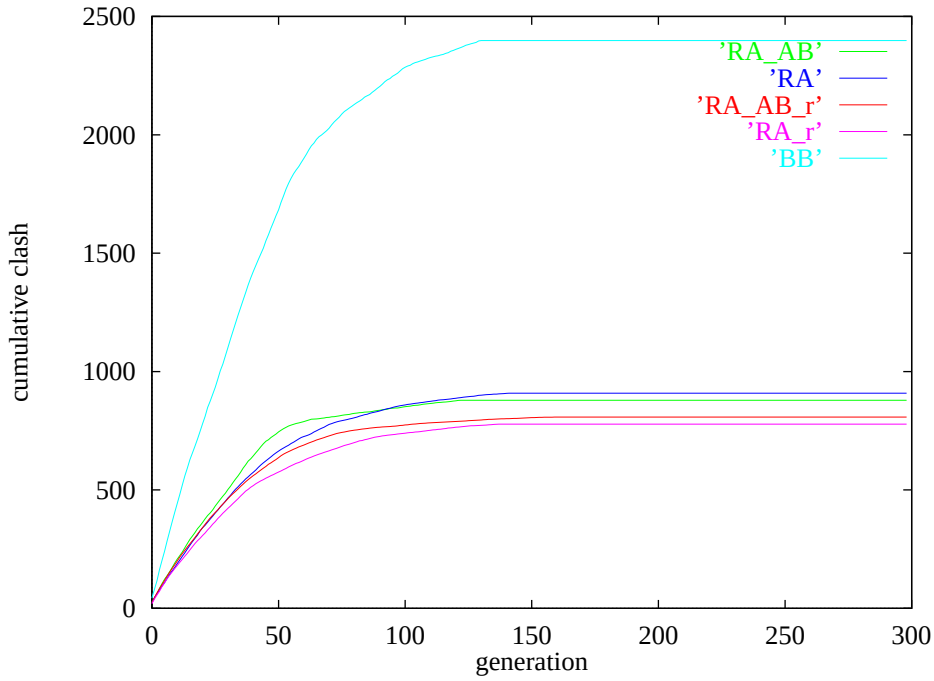

図 4.5: 衝突数の累計値：迷路問題

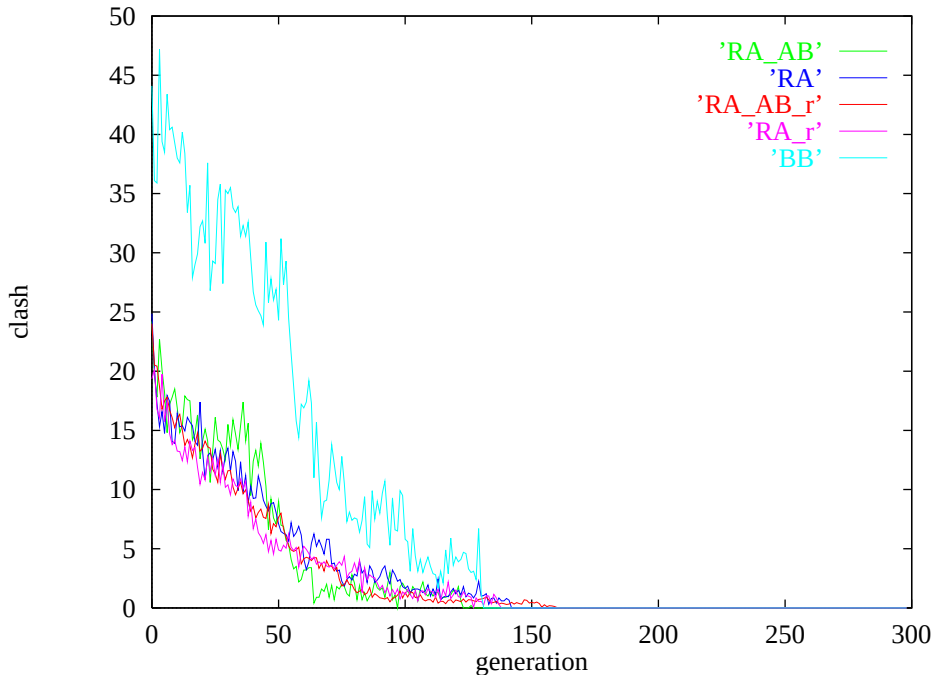

図 4.6: 衝突数：迷路問題

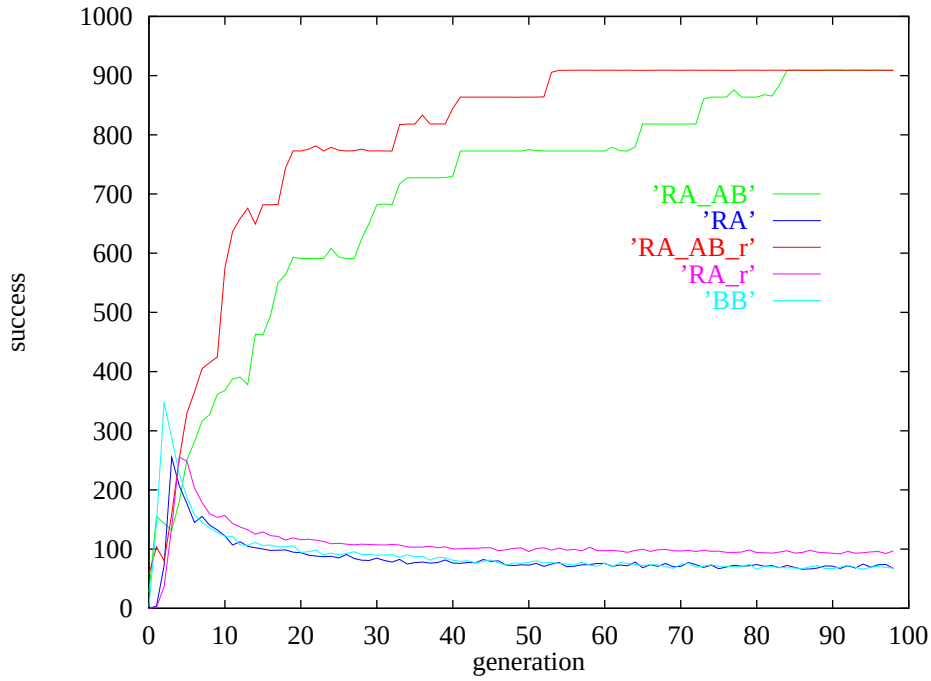

図 4.7: 成功数：宣教師と人喰い人問題

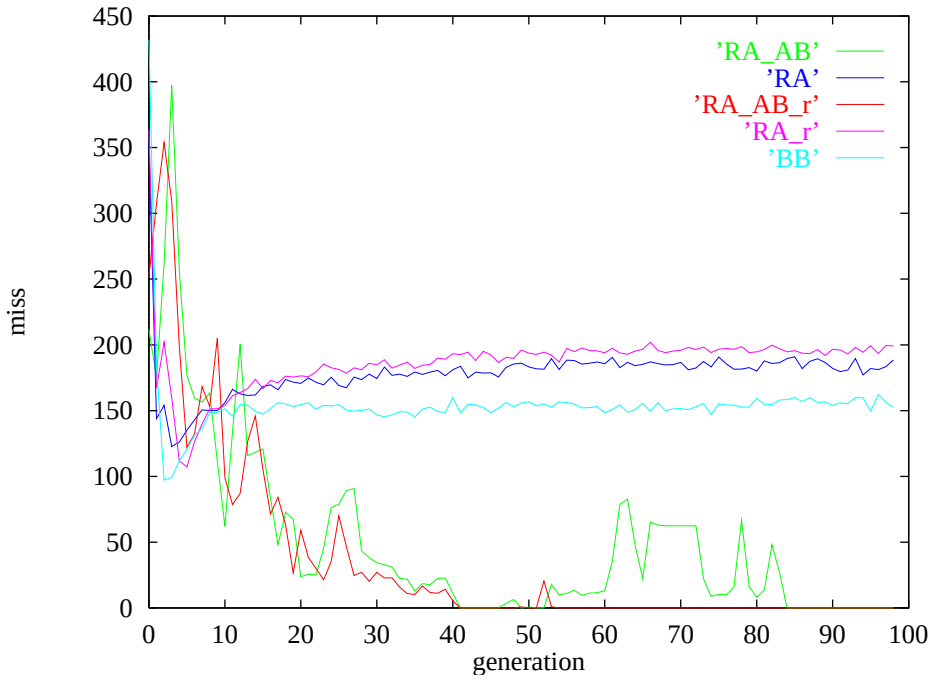

図 4.8: 失敗数：宣教師と人喰い人問題

第 5 章

結論

提案手法は、基本となるバケツリレーアルゴリズムに簡単な変更を加えることにより以下の結果を得ることができた。

- 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。

今後の課題として次に示すものがある。

- 提案手法の動的な環境での検証
- 提案手法の大きな状態空間においての性能の検証

- 本研究で用いた状態全体を評価する方式に推察どおり探索空間を広げる能力が存在するかの検証

謝辞

本研究を進めるにあたって, 多くの方にご支援をいただきました。この場を借りて感謝の気持ちを表したいと思います。

指導教官の國藤進教授には自由な研究環境をはじめとする日頃の研究生活全般においても常に配慮していただき大変感謝しています。

また國藤研究室の方々には日頃から研究に関して議論を重ねていただき、研究活動以外にも私生活の面でも非常にお世話になりました。心から感謝しております。

1998 年 2 月 13 日

寺田 賢二

参考文献

[1] 久保田孝：“火星探査機「Mars Pathfinder」”, 日本ロボット学会誌,Vol.15,No.7,pp.986-992,1997.

[2] 開一夫、松原仁 : “機械学習からみたロボットの学習 — 能動的学習機構に向けて”, 日本ロボット学会誌,Vol.13,No.1,pp.5-10, 1995.

[3] 浅田稔、野田彰一、俵積田健、細田耕：“視覚に基づく強化学習によるロボットの行動獲得”, 日本ロボット学会誌,Vol.13,No.1,pp.68-74, 1995.

[4] 山口智浩、増淵元臣、田中康祐、谷内田正彦 : “経験型強化学習における仮想個体から実ロボットへの学習行動の伝搬”, 人工知能学会誌,Vol.12,No.4,pp.570-581, 1997.

[5] Brooks,R.A : “A Robust Layered Control System For A Mobile Robot”, IEEE journal of robotics and automation, Vol.RA-2,No.1,pp.14-23, 1986.

[6] 山村雅幸、宮崎和光、小林重信 : “エージェントの学習”, 人工知能学会誌, Vol.10, No.5, pp.683-689, 1995.

[7] 宮崎和光、小林重信 : “マルコフ決定過程下での統合的強化学習システム”, 第 9 回自律分散システムシンポジウム予稿集,pp.85-90, 1997.

[8] 坪井創吾、宮崎和光、小林重信 : “強化学習に基づくオセロゲームの政策形成”, 第 24 回知能システムシンポジウム, 1997.

[9] 畝見達夫 : “実例に基づく強化学習法による失敗しない制御方法の学習”, 人工知能学会誌,Vol.7,No.6,pp.1001-1008, 1992.

[10] 幡野七子 : “状況に応じた推論モジュールの切り替えに関する研究”, 北陸先端科学技術大学院大学修士論文, 1996.

参考文献

[11] J.H.Holland,K.J.Holyoak,R.E.Nisbett and P.R.Thangard: “induction”, MIT Press(1986).

市川伸一ほか (訳), “インダクション — 推論・学習・発見の統合理論へ向けて”, 新曜社, 1991.

[12] J.H.Holland: “The Possibilities of General-Purpose Learning Algorithms Applied to Parallel Rule-Based System”, in R.S.Michalske, et al.(eds.),Machine Learning, An Artificial Intelligence Approach, Vol.2,pp.593-623, Morgan Kaufmann (1986).

田中卓史 (訳), “脆弱性の回避：並列型の規則に基づくシステムへ適用した汎用学習アルゴリズムの可能性”, pp.91-128. 知識獲得と学習シリーズ 7, 類推学習, 共立出版 (1988).

[13] 北野宏明 (編) : “遺伝的アルゴリズム”, 産業図書,1993.

[14] 和田健之介 : “デジタル生命の進化”, 岩波科学ライブラリー 11, 岩波書店,1994.

[15] 上原忠弘、山村雅幸、小林重信 : “不完全知覚下での強化学習のための環境モデル構築”, 第 21 回知能システムシンポジウム予稿集,pp.25-30, 1995.

[16] 長尾真 : “知識と推論”, 岩波講座ソフトウェア科学 14, 岩波書店 (1988).

付録 A

ルール生成に対する考察

分類子システムを GA(遺伝的アルゴリズム) の方向から見ると、バケツリレーアルゴリズム等による強化学習は次に示す GA の処理の手順 [13] における適応度の評価にあたる。

1. 初期集団の生成
2. 終了条件がみたされるまでループ
   (a) 適応度の評価
   (b) 選択
   (c) 交叉
   (d) 突然変移

よって、強化学習により行動する学習主体は与えられた分類子を評価するための道具ということになる。つまり実際に起こり得る状態に対する「状態-動作」対を生成できるかは、GA によるルール生成で確率に高くはなるが、結局は運次第である。特に学習初期には、その確率が低いと考えられる。

しかし自律ロボットにおける強化学習を考えた場合、評価すべき状態は少なくともセンサー入力により確実に判断できる。よってルール生成は、その状態に対する行動を作成することに行い、その分類子を評価したほうが自然であると考えられる。

また何も行うべき行動が無い状態の場合、その時の行動の学習は報酬の受け皿となる分類子の強度がないため行われない。これは試行を無駄しており効率の面からも悪く、さら

に学習されない状態が罰報酬を受ける行動を含んでいる場合は、その行動は確率的に選択されるため、罰報酬の原因となる。

よって本研究では、選択可能な分類子が無い状態の場合は、新たにその状態に適合するルールを生成する。(図 2.3)