目標学習を伴う強化学習エージェント集団の共同プランニング

4G－11　　前沢 力　　渥美 雅保
創価大学工学研究科情報システム学専攻　　創価大学工学部情報システム学科

1 はじめに

複数エージェントによる協調問題解決を，強化学習の枠組みで扱おうという研究も，近年，活発化しつつある．先駆的な研究として，[Weiss 93]はクラシファイアーシステム[Holland 86]を内部に持つエージェント集団が，局所環境情報のみに基づき共同プランを学習するアーキテクチャを提案している．しかし，局所環境情報のみに即応的な強化学習では，複雑な問題に対する協調の学習は困難で，その解決のためにモジュール化された複数のクラシファイアーシステムの上位にそれらの調整をするクラシファイアーシステムを設けた階層型クラシファイアーシステム[Dorigo 93]や，クラシファイアーシステムにエージェント間の通信構造を学習する仕組みを組み込んだシステム[前沢 96]等が提案されている．

ところで，協調行動が高度であればあるほど，反応的行動の学習に加えて熟考に基づく行動の学習が必要となる．本研究では，熟考的要素を自ら達成すべき目標の設定としてとらえ，その目標，目標の達成に必要な行為，協調のために必要とされる通信構造を学習するための３つのクラシファイアーシステムを構造的に持つエージェントアーキテクチャを提案し，視界を設定した３次元ブロックワールド問題を実験タスクとして取り上げ，目標学習の有効性を調べる．

2 ３次元ブロックワールドタスク

2.1 タスクの概要

図１のように，高さを含んだ３次元グリッド環境に，エージェントとブロックがランダムに配置されている環境を考える．ここで，この環境の東西，南北の両端はそれぞれ接続されているものとする．エージェント集団に課せられるタスクは，ブロックをある目標状態に配置することである．

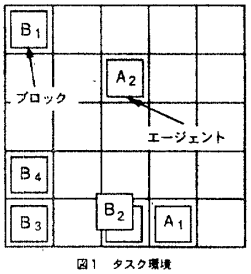


図1 タスク環境

2.2 エージェント

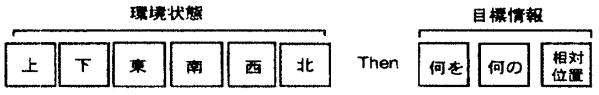
エージェントは視界が限定されたセンサーを持ち，視界内にブロックやエージェントがいれば，その方向および距離を特定できるものとする．エージェントがとりうる基本行動は，東西南北の方向に移動する，それら方向のブロックを持ち上げる，それら方向にブロックを降ろす，及び動かないの１３種類の行動で，さらにブロックや他のエージェントの上に１段だけ登ることもできる．エージェントが，自分の位置から２段上の高さにブロックを置くためには，土台となるブロックを用意したり他のエージェント上に登ることが必要で，本タスクには高度な協調行動が要求される．

3 目標学習を伴うエージェントアーキテクチャ

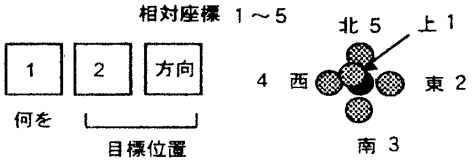
3.1 ３学習層の学習内容

本アーキテクチャは，目標層，行為層，通信層の３つの学習層からなる．以下ではそれぞれの層について説明する．

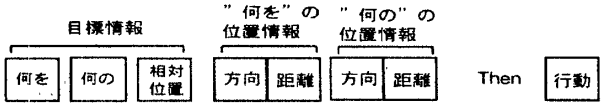
目標層：目標層では，環境状態の観測に基づき目標を選択することを学習する．これら目標は全体目標に対する中間目標としての働きをする．目標層のルールは以下のような形式を持つ．



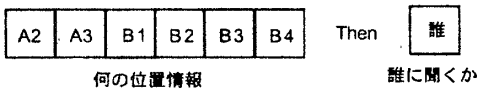
ここで目標情報は「何をある物の相対位置に移動する」ことを目標とするかを表す．



行為層：行為層は目標層が提出した目標を達成するための行動系列を提出することを学習する．行為層のルールは以下のような形式を持つ．



通信層：通信層は行為層がルールを適用するために必要とする環境情報のうち自分の視野外にある情報を，どのエージェントに尋ねればよいかという通信相手の選択を学習する層である．通信層のルールは以下のような形式を持つ．



Collaborative Planning of Agents with Structured Classsifier Systems for Goal and Action Learning , Chikara Maezawa, Masayasu Atsumi , Soka University ,1-236 Tangi-cho, Hachioji-shi, Tokyo 192,Japan

3.2 各層間報酬割り当て

エージェントの学習内容を階層化したことによって，エージェントに環境から与えられる報酬を各学習層に分配する必要が生じる．以下では各学習層の評価の方法について述べる．

目標層の評価：全体ゴールが達成したときに与えられる外部報酬を目標系列に対してプロフィットシェアリング法により分配する．

行為層の評価：目標層の提出した目標が達成された場合に，目標を提出した目標層のルールの評価値に比例した報酬を，行動系列に対してプロフィットシェアリング法により分配する．

通信層の評価：通信層は行為層が必要とする情報の通信を行うので，通信が成功した場合に通信を依頼したルールの評価値に比例した報酬を通信ルールに分配する．

4 実験結果

4.1 実験の内容

本実験では，エージェント数2，ブロック数5，ワールドサイズ5*5，全体ゴールとしてブロックを3つ重ねるタスクを考える．このタスクは簡単ではあるが，ブロックまたはエージェントを土台として設定することが必要となるタスクである．実験パラメータとしては，ルールの評価値の初期値は1000，タスク成功の際に受け取る強化信号は100である．また，交配確率は0.6，突然変異確率は0.001，交配は一点交配を使う．世代ギャップはルールが1000回起動するごとに行った．

4.2 実験方法

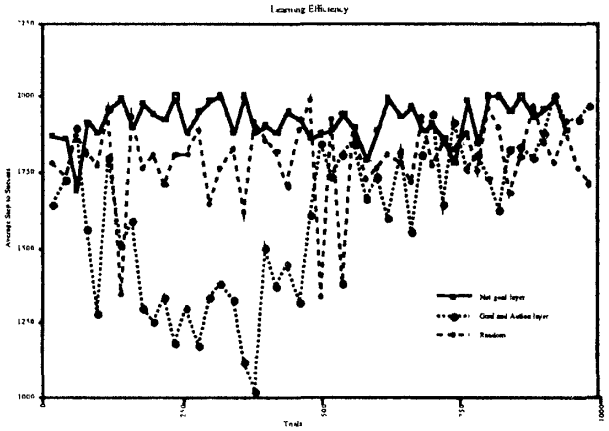
次の3種類のエージェントを用いて実験を行い学習パフォーマンスの比較評価をする．

(1) 各エージェントが行動をランダムに決定するケース
(2) エージェントを，行為層クラシファイアーと通信層クラシファイアーの2層で実現するケース(ルールの個体数は行為層100，通信層60)
(3) 目標学習を伴う3層のクラシファイアーから成るエージェントアーキテクチャのケース(ルールの個体数は目標層400，行為層100，通信層60)

4.3 学習曲線

グラフは，3種類のエージェントの学習性能を示している．横軸はトライアル数，縦軸は成功するまでの平均ステップ数で，50トライアル毎の平均ステップ数をプロットしている．ただし悪く学習した結果としてデッドロックに陥る場合があるので，今回の実験では2000ステップまででトライアルを打ち切っている．

ランダムなエージェントの平均成功ステップ数は約1800で，部分タスクの干渉により，本例のように単純なタスクでさえ多くのステップを要する．



グラフ・学習曲線の比較

(2)のケースでは，あるエージェントが共同プランとなりえない局所解を強く学習してしまう結果，デッドロックに陥る状況がおこり，共同プランの学習に失敗する場合があるが，(3)のケースではそのような状況は回避されている．これは，他のエージェントと通信することによって，状況を確認しながら行動を行うルールが，進化のアルゴリズムにより発生するためである．

4.4 学習されたプランの評価

(3)のケースで行為層が学習したルールは，視界内に目標ブロックが存在する場合はある程度妥当な制御を行うが視界内に目標物が存在しない場合はランダムな行動を取る．実験では行為層の学習の後に目標層の学習をしており，2段階の学習を行っていることが解った．しかし，学習が進むと全体の成功率が悪くなる現象が見られた．これはルールの設計の問題があり，視界内に目標物がないときの学習ルールが他の環境状態に対して障害となっているという問題である．

5 むすび

本論文では，目標学習を伴う強化学習エージェントアーキテクチャを提案した．そして，実験により本アーキテクチャの有効性を調べた．

◇ 参考文献 ◇

[Weiβ 93] G.Weiβ : Learning to C oodinate Actions i n Multi-Agent Systems, P roc. 13th I nt.Joint Conf. o n Artificial Intelligence, pp. 331-316 (1993).

[Holland 86] Holland, J.H., Holyoak, K.J., Nisbett, R .E. and Thagard, P.R. : Induction, MIT Press (1986)

[Dorigo 93] M.Dorigo,"Genetic-Based Machine Learning and Behabior-Based Robotics: A New S ynthesis",IEEE Transactions on Systems, Man and C ybanetics, Vol.23,No.1

[前沢96]前沢，渥美:強化学習エージェント集団の共同プラン生成における通信構造の相互進化,情処第52回全国大会講演論文集(2),pp.7-8(1996).