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

4C－4

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においては，各エージェントは大域的制約を満たすために，直面する状況のもとで誰と協調して部分問題を解決すればよいかを学習することが必要となる．複数の分業化した強化学習エージェントによる協調問題解決の先駆的な研究として，Weiβ[Weiβ 93] は各エージェントが共有ゴールを達成する共同プランを，局所情報に基づき学習する方法を提案しているが，大域的な制約を満たすプランの学習には時間がかかるため，エージェント間で適切な通信をすることが重要となってくる．

本研究では，クラシファイアーシステム[Holland 86]を持つエージェント集団が協調のために必要とされるエージェント間の通信構造を進化的に学習するモデルを提案する．そして通信構造の学習が，共同プランの学習速度を向上させることをマルチエージェントブロックワールド問題を用いて実験的に明らかにする．

2 マルチエージェントクラシファイアーシステム

2.1 特徴

(1) エージェントは，自らが観測可能な局所情報に加えて特定のエージェントの観測情報を一対一通信により収集して，行動を選択する機構を持つ．
(2) エージェントは，共有ゴールを達成する共同プランを生成するために，通信すべき相手を進化的に学習する機構を持つ．

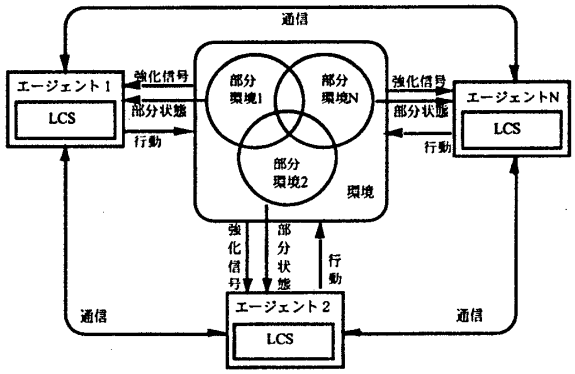
2.2 構成

マルチエージェントクラシファイアーシステムの構成を図１に示す．エージェントは局所環境状態の観測S_Lと他のエージェントとの通信により得た非局所環境状態情報S_Nにもとづき，自らの行動Acを決定するルールを持つ．すなわちルールは

IF $C_L \cup C_N$ THEN Ac

C_L：局所環境条件項　C_N：非局所環境条件項

という形式を持つ．各ルールには「評価値」が割り当てられており，条件部がマッチした実行可能ルール間の競合は，評価値に基づき解消される．ある程度実行が進み，ルールの評価値に十分な変更がなされた時点で評価値を適応度とみなして遺伝的アルゴリズムを適用する．これにより無駄なルールは淘汰され，有効なルールを組み合わせて新たなルールの生成が行われる．



LCS：Learning Classifier System

図１：マルチエージェントクラシファイアーシステムにおける情報の流れ

3 共同プラン生成における協調学習

3.1 マルチエージェントブロックワールド問題

協調問題解決の例として，図２のようなブロックワールド問題において，各エージェントが１つのブロックの移動を担当し，Start状態からGoal状態への共同プランを作成する問題を取り上げる．

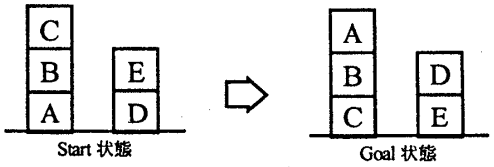


図２：ブロックワールドタスク

この例題では，エージェントの観測能力は，自分の上のブロックの有無と自分の下のブロック名のみに限られる．また，複数のエージェントが同時に同じ場所に移動しようとする場合，評価値が高い行動が優先される．

タスクが成功すると各エージェントに報酬がProfit Sharing[宮崎 94]により分配される．

3.2 通信構造の自己組織化

通信を伴う行動決定と通信構造の自己組織化のアルゴリズムを示す．

step1 $S_P = S_L$ （S_P：現在の環境についての知識）

step2 S_Pのすべての情報と照合するルール集合R を選択する．

step3 R から，各ルールの評価値にノイズを加えた結果の評価値が最大のルールr^*を選択する．

step4 r^*の条件部C_{r^*}に未照合の項があるとき，その項についての情報を持つエージェントと通信し，通

Co-Evolution of Communication Structure for Collaborative Planning by Reinforcement learning agents
Chikara Maezawa,Masayasu Atsumi　Soka University
1-236 Tangi-cho,Hachioji-shi,Tokyo 192,Japan

信コストを評価値から引くと共に，S_pを更新する．
step5 C_r．がすべて満たされたならば，行動Ac_r．を出力する．そうでないならばstep2へいく．

ゴールが達成された場合，実行したルールに報酬を分配することにより，ルールの評価値の強化を行う．通信するたびに通信コストを引くため報酬の与えられないルールは評価値が下がるのみとなり，進化の際に淘汰される可能性が高くなる．それによって有効な通信を含むルールが残り，その結果，通信構造の自己組織化がなされる．

4 実験

4.1 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) 各エージェントが，局所環境情報のみに基づく行動を強化学習するケース
(3) 進化的に通信構造を学習するケース

以上の3つのエージェントの実験プログラムをUNIX/C環境で構築し，シュミレーション実験を行った．

4.2 学習曲線

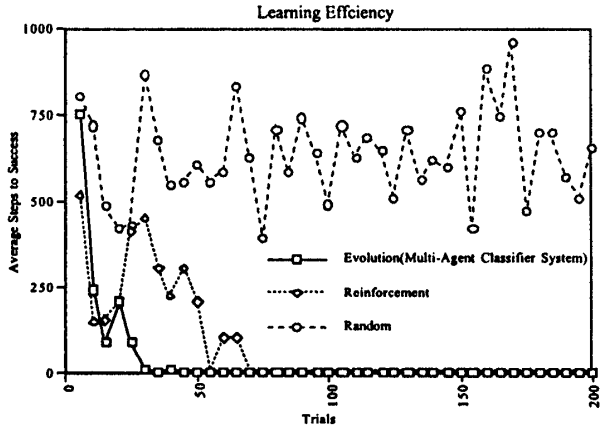


図3：学習曲線の比較

図3は，3種類のエージェントによる分業タスク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，5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1000ステップまででトライアルを打ち切っている．なお，ルールの評価値の初期値は1000，ノイズを表す乱数は100，通信コストは（通信相手の数*0.001*強化信号値），タスク成功の際に受け取る強化信号は5である．また，ルールの個体数Nは各エージェントごとに500，交配確率P_cは0.6，突然変異確率P_mは0.001，交配は一点交配を使う．世代ギャップGは1で行った．

ランダムなエージェントの平均成功ステップ数は約700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．

エージェントが局所環境情報に基づく行動を強化学習するケースでは，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，進化的に通信構造を学習する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．

4.3 通信構造の組織化

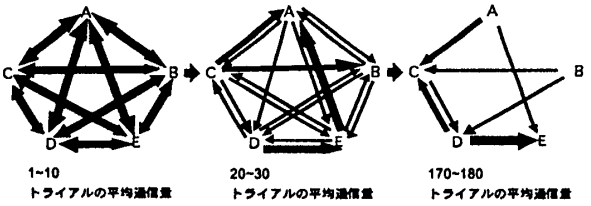


図4：学習過程での通信構造の変化

図4は，エージェント間の通信構造の進化を有効グラフを用いて表したものである．枝の太さは通信の頻度を示す．図4を見ると，学習が進むにつれて全体的な通信量は減少し，特定のエージェント間に通信が限定されることがわかる．進化的に学習されたルールは，「共同プランの成功のために，自分の行動の前に達成されていなければならない非局所的状態を確認して行動を実行する」というものが多くなっている．

通信構造の自己組織化においては，通信にはコストがかかるため，コストをかけても効果を上げられるルールのみが淘汰により選択されてくる．従って全く通信する必要のないタスクでは，通信を用いるルールは淘汰され，逆に，複雑なタスクでは，通信を多く用いるルールが残ることが予想される．

5 むすび

本研究では，クラシファイアーシステムを持つエージェント集団が協調のために必要とされるエージェント間の通信構造を進化的に学習するモデルを提案した．そして，通信構造の学習が，共同プランの学習速度を向上させることを示した．

◇ 参考文献 ◇

[Weiβ 93] Weiβ, G. : Learning to Coodinate Actions in Multi-Agent Systems, Proc. 13th Int.Joint Conf. on 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)
[宮崎 94] 宮崎和光,他：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 (1994)