知能と情報（日本知能情報ファジィ学会誌）Vol.24, No.5, pp.1002－1011（2012）

原著論文

異質のマルチエージェント間のインタラクションを考慮した学習モデル†

張 坤 *1・前田 陽一郎 *2・高橋 泰岳 *2

強化学習はシングルエージェントを対象に開発された手法であり，マルチエージェント環境になると，個々のエージェントがどのように相互に影響を与えれば，全体の適切な協調行動を獲得できるかという問題における決定的な解決方法は示されていない．そこで本研究では，エージェント間の強化値インタラクションを通じて，優れた協調能力をもつインタラクティブ学習システムの構築手法を提案する．本手法では，個々のエージェントは環境との試行錯誤を繰り返しながら，エージェント間でも目標の達成度と協調度に応じて，相互信頼度を自律的に生成し，更新する．その信頼度の高さにより，各エージェントは他エージェントの強化値を利用するレベルを決める．強化値を相互に利用することで，エージェント間でインタラクティブ学習が可能なシステムを構築する．そのため，各エージェントは周囲のエージェントからも自身に有効な経験を学習することができる．環境や他エージェントとのインタラクションを通じて，マルチエージェントの協調行動と集団戦略を効率的に学習させる．

キーワード：マルチエージェント強化学習，インタラクティブ学習，強化値，信頼度

1．はじめに

近年，ロボットは多方面で実用化されているのに伴い，社会で担う役割が多くなり，多機能で汎用的なロボットが期待されている．特に，未知の複雑な環境で１台のロボットでは達成が困難であるが，複数の自律ロボットが協力することにより目標を達成することができるマルチエージェントシステムが盛んに研究されている[1, 2]．分散協調システムの一つとして，マルチエージェントシステムでは複数のエージェントが存在し，各エージェントは自分以外のエージェントを環境の一部として観測し，自身に与えられたタスクを自律的に達成することで全体の協調作業を行なう[3]．マルチエージェントシステムは問題解決能力，適応能力，ロバスト性，並列性，モジュール性などの点で利点があるため，シングルエージェントにできない作業でもマルチエージェントには達成できる可能性がある[4-6]．

また，強化学習は未知環境下で，エージェントが環境との相互作用により，報酬または罰を受け取ることで，動物の条件反射のようにある環境に対応する適切な行動を起こしやすくなる学習手法である．教師信号などの環境に対する事前知識を必要とせず，報酬を頼りに試行錯誤を繰り返すことで適切な行動を自律的に学習できるため，多くの研究者に注目されている[7]．他の機械学習と比べ，動的な環境変化に柔軟に対応しやすく，予想以上の行動を得たりする場合が多いため，複数のエージェントに協調動作を学習させるマルチエージェント強化学習の研究にも期待が寄せられている[8-10]．

しかしながら，強化学習はシングルエージェントを対象に開発された手法であり，マルチエージェントで強化学習をそのまま適用すると，不完全知覚問題，同時学習問題，報酬分配問題などいくつかの問題が生じることが知られている[11]．協調行動を獲得するため，個々の自律エージェントは，どのように相互に影響を与えれば，全体の適切な協調行動を獲得できるかという問題における決定的な解決方法は示されていない．

協力作業では，センシング情報，エピソード，学習方策やアドバイスのやり取りなどを共有することで，協調行動を学ぶことが有効である[12]．しかし，エージェントの自律性により，相手に有効なことを自身に適応できない場合が多い．マルチエージェント強化学習では，各エージェントの報酬はエージェントの行動

† Learning Model Considering the Interaction among Heterogeneous Multi－Agents

Kun ZHANG, Yoichiro MAEDA and Yasutake TAKAHASHI
*1 福井大学 大学院工学研究科 システム設計工学専攻
Dept. of System Design Engineering, Graduate School of Engineering, University of Fukui
*2 福井大学 大学院工学研究科 知能システム工学専攻
Dept. of Human and Artificial Intelligent Systems, Graduate School of Engineering, University of Fukui

の組み合わせで決まるため，Q値は他のエージェントの方策に依存し，確率ゲームの枠組みでとらえることが有用であり，ナッシュ均衡点を学習する手法がある[13-14]．また，エージェント同士がお互いの行動を観測し，その観測情報を基にして相手の政策を推測することが必要となる．この場合状況により，マルコフ性が成立せず，学習の収束は保障されない．マルチエージェント強化学習ではエージェントの経験の共有がより速く類似したタスクを学習することに役に立つ[15]．例えば，熟練したエージェントは学習者のための教師としてそれらの行動選択を支援する，または学習者が熟練したエージェントの行動を見て模倣する[16]．しかし，経験の共有は類似した環境の学習に限るため，学習環境の状況が複雑になると，自身に有効なことのみを自律的に学習することが必要となる．

マルチエージェントシステム環境では，学習プロセスの間にインタラクションによりどのような良い行動を獲得するかが重要となる[17]．そのため，エージェント間のソーシャルインタラクションを通じて，人間に似た優れた協調能力をもつマルチエージェントの構築を行なうことを本研究の主な研究目標とする．

本研究では人間社会のように，自律エージェント間の強化値インタラクションを通じて，インタラクティブ学習が可能なシステムの構築を提案する．ここでは従来の「環境-エージェント個体」のみの学習モデルと異なり，「エージェント個体-他エージェント個体」のインタラクティブ学習モデルも含めている．本手法では，各エージェントが目標達成行動に携わりながら，それぞれのエージェントと一緒に目標を達成した時，環境から得た報酬により，相手との信頼度を構築する．報酬は高いほど信頼度が高まる．学習しなかった環境または学習経験がほとんどなかった環境になると，信頼度をもったエージェントの強化値を利用することができる．複数のエージェントは信頼度によって，相互的強化値を利用することも可能になる．このように，エージェント間のインタラクションを利用して，相互協力を行うグループ作りを促進し，グループ行動の協調戦略を向上させる．

提案手法の有効性を検証するため，獲物追跡問題を例題にシミュレーション実験を行った．比較実験により，エージェント間には強化値に基づくインタラクティブ学習が有効であることがわかった．

2．マルチエージェント強化学習

マルチエージェントシステム(Multi-Agent System: MAS)とは，複数の自律的に行動するエージェントから構成されるシステムであり，個々のエージェントが協力して，複雑なタスクをシステム全体として達成する．一般に集中制御ではなく，複数のエージェントが自律的にシステム全体のタスクを達成するマルチエージェントシステムは問題解決能力とロバスト性などの利点があるため，多くの研究が行なわれている[18]．MASにおける各エージェントは環境を観測し，自分に与えられたタスクを達成する行動を通じて，システム全体のタスクを達成する．MASでは，全てのエージェントが同じ行動ルールで行動する均質(Homogeneous)型と各エージェントが異なる行動ルールで行動する異質(Heterogeneous)型に分類される[19]．

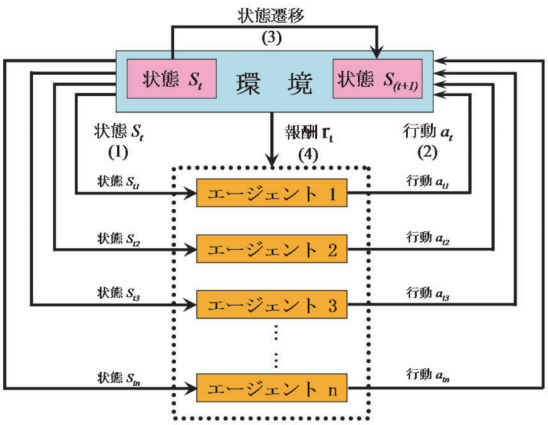


図1　マルチエージェント強化学習モデル

強化学習は未知の動的な環境でも柔軟な適応能力をもち，MASに適応するマルチエージェント強化学習は図1のようにモデル化できる．マルチエージェント強化学習においては，環境は複数のエージェントが相互に行動することによる影響で新しい環境に遷移する．自律的に考え行動するシングルエージェントと比べると，MASの特徴として，複数のエージェントが相互に影響を及ぼすため，自身以外のエージェントの行動選択基準にも影響を与える．そのため，相互的な協調は非常に重要となる．特に，役割分担など複雑な協調作業では他のエージェントの協力がないと問題を解決できないため，エージェント同士の協調を自律的に学習させることが必要になる．

3．強化値インタラクションに基づくインタラクティブ学習

マルチエージェント強化学習では，エピソードや学習方策を共有することで，協調行動を効率的に学ぶことができる．しかし，異質エージェントは相手に有効な方策を自身にそのまま適応できない場合が多い．そのため，相手の学習方策がどのような状況で，どの程

度利用されるかなどを自律的に学習することが必要となる．そこで，他エージェントの有効な学習経験を間接的に活かせれば，マルチエージェント全体の協調性が高まるものと考えられる．

本研究では，マルチエージェント間の強化値インタラクションを通じて，相互に学習できるインタラクティブ学習手法を提案する．ここでは能力の異なるエージェントに対し同じ能力のエージェント同士を同じグループにする．グループ間には目標の達成度により，グループ信頼度を構築する．一方，各エージェントはそれぞれのエージェント間にも個体信頼度を構築できる．グループ信頼度と個体信頼度により，知覚範囲内のエージェントの強化値をどの程度利用するかを判断する．毎回の学習結果に基づいて，グループ信頼度と個体信頼度が更新され，このようなエージェント間のインタラクションを繰り返すことで，協調行動を自律的に学習する．

3.1　強化値インタラクションモデル

通常のマルチエージェント強化学習では，エージェントは報酬の最大化を目的として，ある環境との試行錯誤を繰り返すことで，すべての強化値は状態観測から行動出力への適切なマッピングを獲得する．各エージェントは自身の強化値により行動を選択する．これに対し，本研究では図2のように，各エージェントは自身の強化値のみではなく，知覚範囲内のエージェントの強化値を考慮し，目標となる適切な行動を選択する．ここで言う強化値とは，例えば Q-learningの場合のQ値に相当するものを指す．

つまり，すべての強化値は自身の試行錯誤で獲得するため，他エージェントの強化値を考慮し，取り入れることで学習効率を高められると考える．他エージェントの強化値をそのまま共有するのとは異なり，自身に適用できる強化値のみを取り入れるため，他エージェントとの信頼度を構築することで，いろいろな強化値から自身に適用できる強化値のみを選択することを自律的に行なう．

3.2　信頼度に基づく行動選択戦略

異質エージェントは相手の戦略を自身に適用するため，自身に有効な戦略のみを利用することが必要である．そのため，試行錯誤でエージェント間には各グループの各個体への信頼度を生成し，更新する．その信頼度に基づいた行動選択戦略獲得の流れを図3に示す．ここでは，エージェント間の強化値の相互利用に基づいて，自己の利益のみで行動するだけではなく，他のエージェントとの共同利益も考慮して行動し，インタラクション機能によるマルチエージェントの集団戦略形成を目標とする．

3.3　Q-learningを用いたインタラクティブ学習

強化学習における実現方法は様々な手法が提案され，大きく分けて「環境同定型」と「経験強化型」の二つに分けることができる．以下ではQ-learningとProfit Sharingの二つの代表的な手法によるアプローチについて説明する．

信頼度に基づく強化値のインタラクティブ学習システムにおけるQ-learningの処理手順を以下に示す．

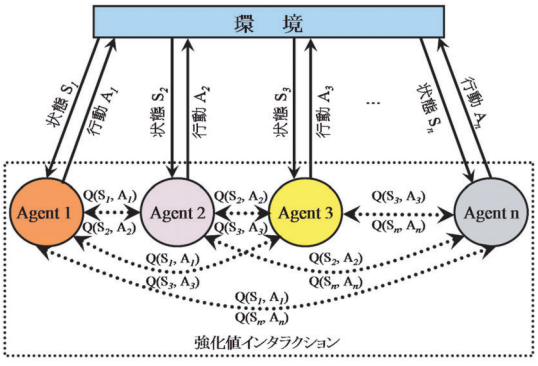


図2　強化値インタラクションモデル

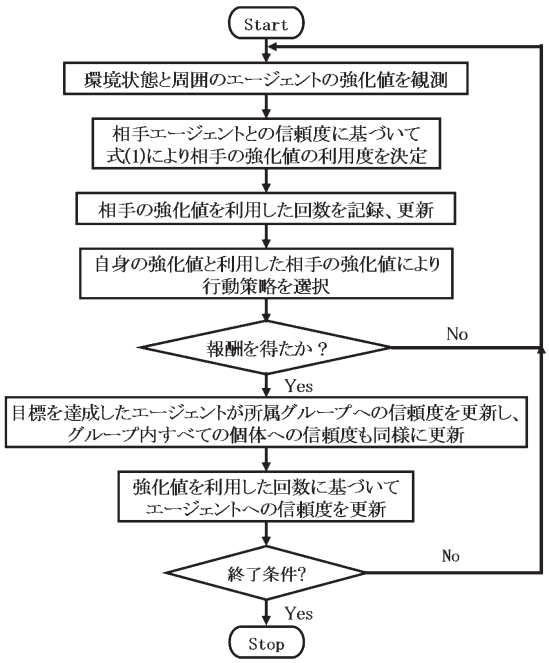


図3　信頼度に基づく行動選択戦略獲得の流れ

- $Q_t(s_t, a_t)$を初期化
- 各エピソードに対して以下を繰り返す
  - 環境状態 s_t を初期化
  - 各エピソードの各ステップに対して以下を繰り返す
    - s_t を観測
    - 式(1)により利用する強化値 $Q_t^*(s_t, a_t)$を計算
    - 強化値 $Q_t^*(s_t, a_t)$の利用回数を更新
    - 式(2)により方策$\pi(s_t, a_t)$を用いて選択行動 a_tを実行
    - 報酬 r_t と次状態 s_{t+1} を観測し，$Q_t(s_t, a_t)$を以下の式で更新

$$Q_t(s_t,a_t) \leftarrow Q_t(s_t,a_t)+\alpha[r_t + \gamma\, max_{a\in A}\, Q_t(s_{t+1},a) - Q_t(s_t,a_t)]$$

    - 式(3)～(5)により，信頼度を更新
    - $s_t \leftarrow s_{t+1}$
- 終了条件を満たせば終了

$$Q_t^*(s_t,a_t) = \sum_{o=1}^{n}(Q_t^o(s_t,a_t)\cdot\frac{C_t^o}{\sum_{i=1}^{n} C_t^i}) \tag{1}$$

$$\pi(s_t,a_t) = \frac{\exp[(Q(s_t,a_t)+Q^*(s_t,a_t))/T]}{\sum_{b_t\in possibleactions}\exp[(Q(s_t,b_t)+Q^*(s_t,b_t))/T]} \tag{2}$$

$$C_t^o = C_{グループ}^o + C_{個体}^o \tag{3}$$

$$C_{グループ}^o = \sum_{i=1}^{t}\frac{r_i^o}{R^*} \tag{4}$$

$$C_{個体}^o = \sum_{i=1}^{t}\frac{e_i^o}{E^*} \tag{5}$$

ここで，

s_t ：時刻 t における状態
a_t ：時刻 t における行動
$Q_t^*(s_t, a_t)$：利用する強化値
n ：知覚範囲内のエージェント数
T ：温度定数
$possibleactions$：すべての選択可能な行動の集合
b_t ：時刻 t で選択した行動
C_t^o：時刻 t におけるエージェント oの総合信頼度
$C_{グループ}^o$：エージェント oの所属グループの信頼度
$C_{個体}^o$ ：エージェント oの個体信頼度
$Q_t^o(s_t, a_t)$：時刻 t におけるエージェントoの強化値
r_i^o：エージェント oの所属グループが得た報酬
R^*：すべてのエージェントグループが得た平均報酬
e_i^o：時刻iにおけるエージェントoの強化値利用回数
E^*：すべてのエージェントの強化値利用回数
α ：学習率
γ ：割引率

学習前にはすべてのエージェントの間には信頼度が存在しない．目標を達成した報酬を得たとき，携わったエージェントの所属グループの間には式(4)のようなグループ信頼度が生成される．そのグループ信頼度が同様にグループ内の各エージェント間の信頼度に転用される．さらに，途中で他のエージェントの強化値を利用した回数により，式(5)のような個体と個体の信頼度を生成する．グループ信頼度と個体信頼度を構築することで，自身に適応できる戦略のみを利用することが可能になる．

3.4 Profit Sharingを用いたインタラクティブ学習

Profit Sharing法は学習解に最適性が保障されないが，連続した状態と行動の系列を短時間で学習できるという特徴がある．信頼度に基づく強化値のインタラクティブ学習システムでは，信頼度C_t^oの計算は式(3)と同様である．他のエージェントの行動評価値 $w_t^*(s_t, a_t)$は式(6)のようになる．行動を選択する際には式(7)のようなルーレット選択で行動を決定させる．その他の処理はQ-learningの場合と同様である．

$$w_t^*(s_t,a_t) = \sum_{o=1}^{n}(w_t^o(s_t,a_t)\cdot\frac{C_t^o}{\sum_{i=1}^{n} C_t^i}) \tag{6}$$

$$P(s_t,a_t) = \frac{w_t(s_t,a_t)+w_t^*(s_t,a_t)}{\sum_{b_t\in possibleactions}(w_t(s_t,b_t)+w_t^*(s_t,b_t))} \tag{7}$$

ここで，

$w_t(s_t, a_t)$ ：自身の行動評価値
$w^o{}_t(s_t, a_t)$ ：エージェント oの行動評価値
$w_t{}^*(s_t, a_t)$：利用する行動評価値
$P(s_t, a_t)$ ：状態 s_t において行動 a_t が選択される確率

4．シミュレーション実験

前節で提案したインタラクティブ学習手法を検証するために，ダイナミックに変化する現象をリアルタイムで分析できるマルチエージェント・シミュレータ(artisoc)を用いて獲物追跡問題のシミュレーション実験を行なった[20]．シミュレータの概観を図4に示す．このシミュレータは試行スペースでのエージェントの動きを表示しながら，各獲物エージェントが捕獲された時間，各グループ(A～F)が得た報酬量と各グ

図 4　シミュレーション実験で用いたシミュレータ

ループ間の信頼度関係などの実験結果を同時に出力することができる．

4.1　シミュレーション条件

このシミュレーションでの学習環境は 2 次元格子状で，100×100のグリッド空間を考える．6 種類(A〜F)のハンター(シミュレータ上では四角形と丸)と 3 種類(1〜3)の獲物エージェント(シミュレータ上では三角形)グループが存在する．各グループのエージェントは10体おり，すべての設定は同じである．ハンターエージェントの目標行動はできるだけ早くすべての獲物エージェントを捕獲することである．今回の実験では，いろいろな特性をもつ異質エージェントを設定し，自身の特性に似たエージェントが多くの経験を学習できるかどうかを検証するため，各ハンターエージェントの獲物を捕獲する能力を表 1 のように異なったものに設定した．

単体エージェントは自らの能力で捕獲行動を達成できず，獲物エージェントの位置に隣接する 8 つのグリッドに存在するすべてのハンターエージェントの合計能力が1Nに達すると，単体獲物の目標捕獲達成となる．そのため，各ハンターエージェントは捕獲しやすい獲物を優先選択すれば，効率的な捕獲目標の達成が期待される．すべての獲物エージェントが捕獲されたとき，1 回の学習試行とする．そのため，エージェント間で協調作業を学習することが必要である．

表 1　エージェントグループの捕獲能力（単位：N）

グループ	獲物 1	獲物 2	獲物 3
A	0.6	0.3	0.1
B	0.6	0.1	0.3
C	0.3	0.6	0.1
D	0.1	0.6	0.3
E	0.3	0.1	0.6
F	0.1	0.3	0.6

すべてのエージェントの視野は10×10のフィールド範囲とした．視野範囲内のエージェント間では強化値を利用することができるものとする．ハンターエージェント，獲物エージェントは共に上，下，左，右，左上，左下，右上，右下の 8 方向のいずれか 1 マスずつ動くことができる．また獲物エージェントはいつも一番近いハンターエージェントから最も遠ざかる方向に逃げるものとする．

基本報酬 R を10とし，目標捕獲を達成すると，各ハンターエージェントが得られる報酬 r_t は表 1 の自らの能力に基本報酬を乗じた値とする．本実験では強化学習のパラメータをα=0.5，γ=0.8，T=0.6とした．提案手法の有効性を検証するため，通常の強化学習(Q-learningとProfit Sharing)との比較実験を行なった．

4.2　シミュレーション結果および考察

すべての手法は 5 回ずつシミュレーションを行い，その平均結果の学習曲線を図 5 と図 6 に示す．横軸は試行回数で，縦軸はすべての獲物を捕獲したステップ数を表す．これらの実験結果により，提案手法はQ-learningとProfit Sharingの両方共に通常の強化学習手

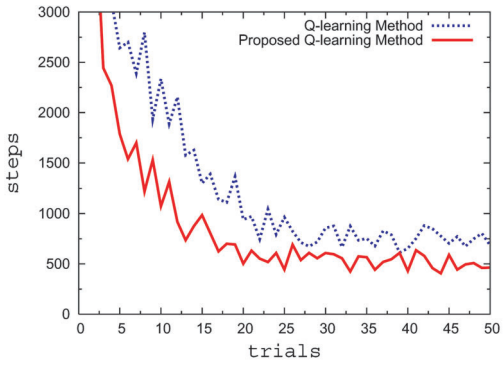


図 5　Q-learning を用いた学習結果

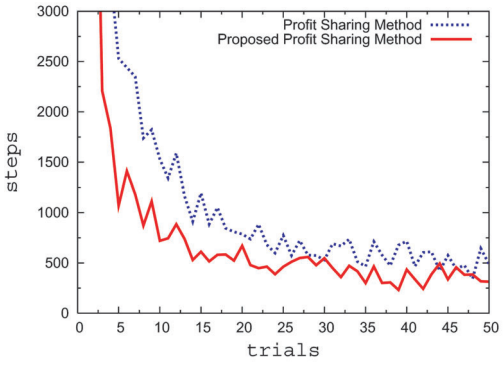


図 6　Profit Sharing を用いた学習結果

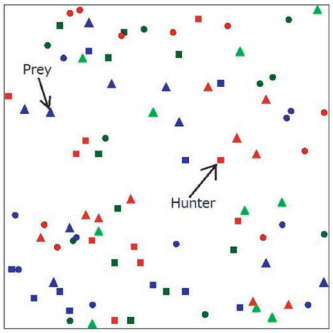
法と比べて圧倒的に短い時間で学習を終了したことが分かった．これにより効率的に獲物を捕獲する協調行動を取得できたことが検証された．本手法により，エージェント間に信頼度を構築することで，周囲のエージェントの強化値を利用することを通じて，協調学習の効率を高めることができる．

Q-learningでは報酬が徐々にその周辺の状態に伝搬していく形で進行するため，一般に学習速度が遅い．これに対し，Profit Sharingでは目標状態に到達し報酬を獲得した際に，その報酬を今までの行動系列に分配し，学習を進める．そのため学習で獲得した解の最適性は保証されないが，学習の立ち上がりに優れている特徴を持つため，Q-learningより早く収束することがわかった．

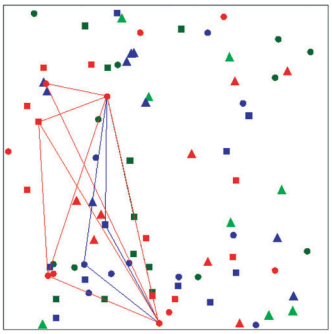
図7は各ハンターエージェント(四角形と丸)間に試行錯誤で獲物(三角形)を捕獲した場合に，線で繋ぐことで信頼度が構築されたという信頼関係を時系列順に示している．各ハンターエージェントは他のエージェントと信頼度を構築すると，自身の色の線で繋ぐことができる．線で繋がっているハンターエージェントは初めて一緒に目標を達成したという意味を持つが，次回の学習で相手の強化値を取り入れることができることを表す．これによりそれぞれのハンターエージェント間の信頼関係の生成過程が見えるようになった．捕獲数の増加に伴い，多くのハンターエージェント間には協調関係が生まれ，信頼関係のネットワークが徐々に拡散していくことがわかる．

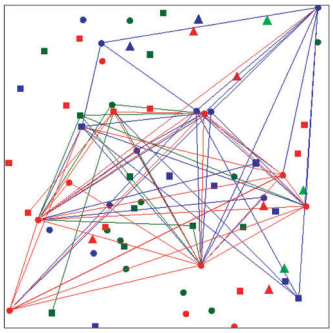
また，本実験では表1のような6種類のハンターエージェントが設定され，各グループのハンターエージェントは獲物エージェントグループに対する能力が異なるため，それぞれのグループ間の信頼度は学習前はゼロであったが，学習した後の信頼度関係は図8と図9のようになった．中心におる各グループはそれぞれのグループとの信頼度の大きさを繋ぐ線の幅で表している．幅が広いほど信頼関係が強いことを示す．

結果より，自身が所属するグループへの信頼度が一番高く，AとBグループ，CとDグループ，EとFグループはペアのような信頼度が確立された．表1によると，AとBグループは獲物1グループに対する捕獲能力は0.6Nで，獲物2と獲物3グループに対する捕獲能力より，非常に高いことが起因していると考えられる．そのため，両方のハンターエージェント個体は相手の強化値を利用すれば，獲物1グループの目標捕獲は2体で達成できることから，有効な経験になる可能性が高い．逆に，AとBグループのハンターエージェントは他のグループの個体と協調すれば，最低3体必要になる．AとBグループはペアとしての信頼関



(a) 学習前の信頼関係


(b) 5体の獲物が捕獲された時の信頼関係


(c) 20体の獲物が捕獲された時の信頼関係

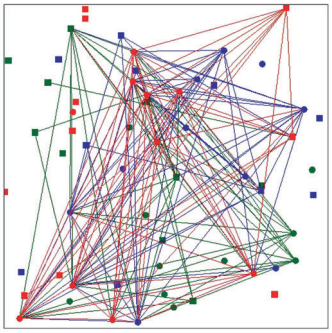

d) 30体の獲物が捕獲された時の信頼関係

図7 信頼関係の変化を示す可視化グラフ

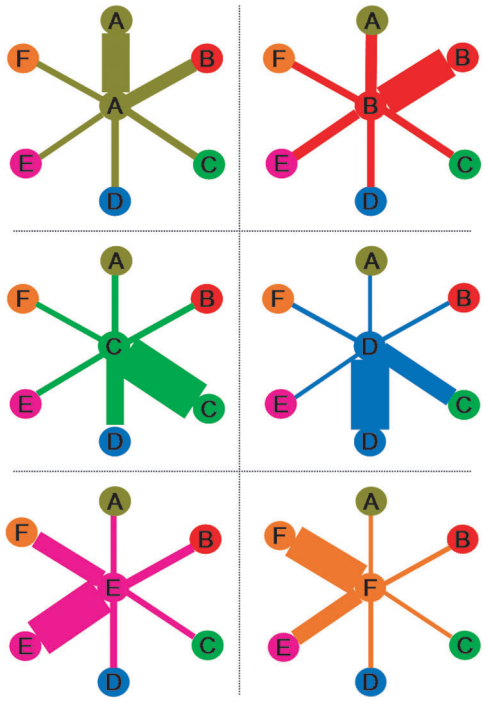


図 8　各グループ間の信頼度関係（Q－learning）

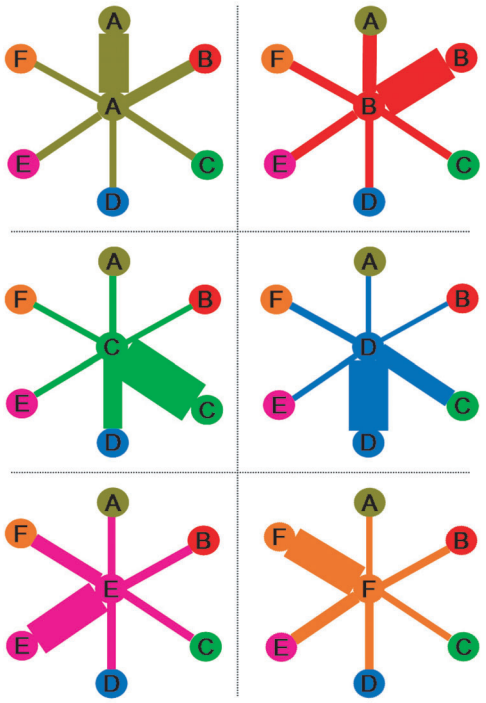


図 9　各グループ間の信頼度関係（Profit Sharing）

係になったことから，提案手法は自律的に適応できるグループ行動の協調戦略を学習できることが検証された．

エージェント間の協調行動の様子を見るため，獲物を捕獲するまでの30ステップの軌跡を記録した．最初の獲物を捕獲した様子を図10に示す．横軸と縦軸は座標で，三角，丸，四角で表されるエージェント位置は最後に捕獲された状態を示し，引かれている線はそれぞれのエージェントの軌跡を示す．最初の獲物を捕獲した時，ハンターエージェントは経験がなく，単純にランダムな行動で捕獲タスクを達成したため，エージェント間の協調の様子はほとんど見られなかった．

さらに学習後期には図11のような軌跡となり，ハンターB6，C9，E0の３エージェントは遠いところからPrey1という獲物エージェントに向かって効率よく協調して追跡していたことを示している．図10と図11を比較すると，ソーシャルエージェントの協調行動が獲得されたことが確認できる．

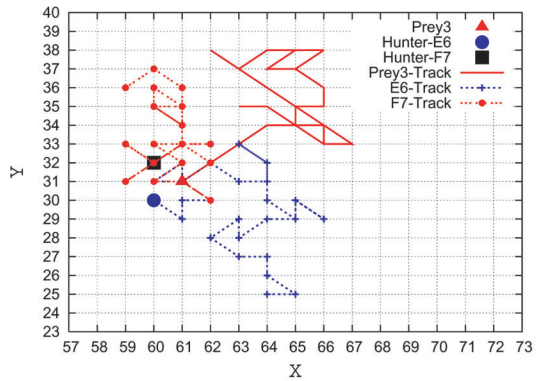


図 10　最初の獲物を捕獲した軌跡の例

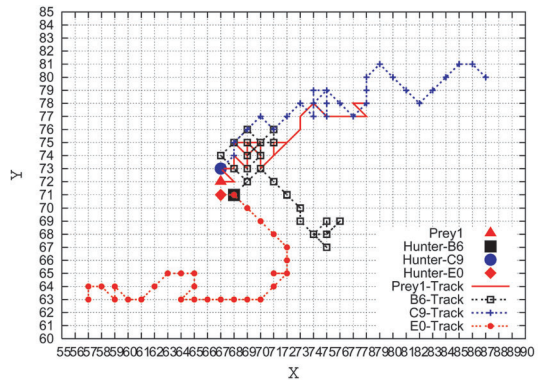


図11　学習後期に獲物を捕獲した軌跡の例

5．おわりに

本研究ではエージェント間の強化値インタラクションを通じて，インタラクティブ学習が可能な手法を提案した．各エージェントは学習プロセス中に他のエージェントとのインタラクションを利用し，相手との個体信頼度のみではなく，それぞれのグループ全体の信頼度も構築することができる．信頼度に基づいて，相手の強化値の利用戦略も学習できる．協調作業では，自身に適応できる強化値を利用しながら，他のエージェントとの協調行動を自律的に学習することが可能になる．

提案手法の有効性を検証するために獲物追跡問題を例題にシミュレーション実験を行った．Q-learningとProfit Sharing法を用いて，通常手法より効率的に協調関係が学習できることがわかった．またそれぞれのエージェント間の信頼度生成の過程を可視化して，信頼関係のネットワークが徐々に拡散していくことも確認できた．さらに，能力の異なるエージェントとして，自身に無効な戦略を放棄し，自身に有効な戦略のみを利用することが必要であるが，本研究では，6種類のエージェントグループを設定し，信頼度関係の構築により，各グループの特徴にあわせた学習ができることも検証された．信頼関係を学習することで，良い集団戦略をもったインタラクティブ学習システムが協調戦略を向上させた．

今後の課題として，エージェント間のインタラクションを深め，更に優れたインタラクティブ学習システムを構築する予定である．強化値の学習のみではなく，より良いコミュニケーションモデルの構造を構築することで，相互に学習機能をもったソーシャルインタラクティブ学習システムへの発展が考えられる．

謝辞

本研究のシミュレーションモデルは構造計画研究所のartisocを用いて作成された．本研究にご協力頂いた全ての方々に謝意を表す．

参 考 文 献

[1] 前田陽一郎，“マルチエージェントロボットにおける協調行動学習のための進化シミュレーション，”日本ファジィ学会誌，Vol.13，No.3，pp.281-291，2001.
[2] M. J. Mataric, “Reinforcement Learning in the Multi-Robot Domain,”Autonomous Robots, 4, pp.73-83, 1997.
[3] 藤田和幸，松尾啓志，“状態空間の部分的高次元化法によるマルチエージェント強化学習，”電子情報通信学会論文誌，D-I，Vol.J88-D-I，No.4，pp.864-872，2005.
[4] 高玉圭樹 著,マルチエージェント学習−相互作用の謎に迫る−，コロナ社，2003.
[5] T. Matsuura and Y. Maeda, “Deadlock Avoidance of a Multi-Agent Robot Based on a Network of Chaotic Elements,” Advanced Robotics, Vol.13, No.3, pp.249-251, 1999.
[6] S. Sen and M. Sekaran,“Multiagent Coordination with Learning Classifier Systems,” in Weiss, G. and Sen, S. (eds.), Adaption and Learning in Multi-agent systems, Berlin, Sprsinger-Verlag, Heidelberg, pp.218-233, 1995.
[7] A. Gosavi, “Reinforcement Learning: A Tutorial Survey and Recent Advances,” Avionics Circle Wright Laboratory Wright State University, 2000. INFORMS Journal on Computing, Vol.21, No.2, pp.178-192, Spring, 2009.
[8] 宮崎和光，木村元，小林重信，“特集「計算学習理論の進展と応用可能性」Profit Sharing に基づく強化学習の理論と応用，”人工知能学会誌，Vol.14，No.5，pp.40-47，1999.
[9] U.Hu and M. P. Wellman, “Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm,” Proc. of International Conf. on Machine Learning (ICML-98), pp.242-250, 1998.
[10] 福澤桂，バフマン ケ ルマンシャヒ，“マルチエージェント強化学習，”電子情報通信学会，NC，97(623)，pp.147-151，1998.
[11] 荒井幸代，“マルチエージェント強化学習−実用化に向けての課題・理論・諸技術との融合−，”人工知能学会誌，Vol.16，No.4，pp.476-481，2001.
[12] M. Tan, “Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents,” Proc. of the 10th International Conf. on Machine Learning, pp.330-337, 1993.
[13] 北原頌士，谷川裕一，鶴岡久，“ナッシュQ学習における協調行動の生成，”福岡工業大学研究論集，Vol.40，No.1，pp.15-20，2007.
[14] J.Hu, and M. P. Wellman, “Nash Q-Learning for General-Sum Stochastic Games,” Journal of Machine Learning Research 4, pp.1039-1069, 2003.
[15] L. Busoniu, R. Babuska, and B. D. Schutter, “A Comprehensive Survey of Multi-Agent Reinforcement Learning,” IEEE Transactions on Systems, Man and Cybernetics, Vol.38, No.2, pp.156-172, 2008.
[16] B.Price and C.Boutilier, “Accelerating reinforcement learning through implicit imitation,” Journal of Artificial Intelligence Research, Vol.19, No.1, pp.569-629, 2003.
[17] L. Nunes, E. Oliveira,“Cooperative learning using advice exchange,”Adaptive Agents and Multiagent Systems, Lecture Notes in Computer Science, pp.33-48, 2003.
[18] M. J. Wooldridge, An Introduction to MultiAgent Systems, John Wiley and Sons, Ltd. England, 2002.
[19] C. Claus, C. Boutilier,“The dynamics of reinforcement learning in cooperative multiagent systems,” AAAI/IAAI, pp.746-752, 1998.
[20] MASコミュニティ，http://mas.kke.co.jp/ modules/tinyd0/index.php?id＝9

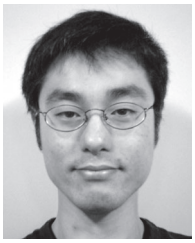
（2012年４月１日　受付）
（2012年７月25日　採録）

［問い合わせ先］
〒910-8507　福井県福井市文京３-９-１
福井大学 大学院工学研究科 知能システム工学専攻
前田　陽一郎
TEL：0776-27-8050
FAX：0776-27-8050
E-mail：maeda@ir.his.u-fukui.ac.jp

著 者 紹 介

張　坤［学生会員］

2007年南昌航空大学航空機械エンジニアリング学科卒業．2010年福井大学大学院工学研究科知能システム工学専攻博士前期課程修了．同年福井大学大学院工学研究科システム設計工学専攻博士後期課程入学，現在に至る．



高橋　泰岳［正会員］

1994年大阪大学大学院工学研究科博士前期課程修了．2000年同大学博士後期課程中退，同年同大助手となり助教を経て，2009年から福井大学大学院工学研究科講師，2012年同大准教授となり現在に至る．博士(工学)．人工知能学会，日本ロボット学会など各会員．知能ロボットの行動獲得に関する研究に従事．

前田　陽一郎［正会員］

1981年大阪大学基礎工学部機械工学科卒業．1983年同大学院修士課程修了．同年三菱電機(株)中央研究所入社，その後，産業システム研究所主事．1989年から３年間，国際ファジィ工学研究所へ出向．1995年より大阪電気通信大学総合情報学部情報工学科助教授．博士(工学)．1999年から１年間，カナダ・ブリティッシュコロンビア大学客員研究員．2002年福井大学工学部知能システム工学科助教授，2007年同大学院教授，現在に至る．主として，人とロボットの双方向インタラクションに関する人間共生システム研究に従事．計測自動制御学会，日本ロボット学会，人工知能学会，電子情報通信学会などの会員．

Learning Model Considering the Interaction among Heterogeneous Multi-Agents

by

Kun ZHANG, Yoichiro MAEDA and Yasutake TAKAHASHI

Abstract：

Reinforcement learning is a technique developed for a single agent. If it's used for the cooperative behavior in multi-agent environment, one of the main problems is how to benefit from mutual interaction during the learning process. In this research, we propose an interactive learning system with cooperative ability through the interaction of reinforcement value among agents. In this method, when each agent repeats trial and error, the confidence degree between other agents could be generated and updated based on the degree of goal achievement and cooperation. The adoption strategy of reinforcement value is determined through the confidence degree. Each agent is able to adopt reinforcement value of others, and an interactive learning system can be built among agents. Therefore, each agent could learn the available experience from others. The cooperative behavior and group strategy of multi-agent is also learned through the interaction with environment and other agents.

Keywords：Multi-Agent Reinforcement Learning, Interactive Learning, Reinforcement Value, Confidence Degree

Contact Address： **Yoichiro MAEDA**

Dept. of Human and Artificial Intelligent Systems, Graduate School of Engineering, University of Fukui
3-9-1, Bunkyo, Fukui-shi, Fukui, 910-8507, JAPAN

TEL ：0776-27-8050
FAX ：0776-27-8050
E-mail ：maeda@ir.his.u-fukui.ac.jp