線形時系列予測と強化学習を用いた動的な学習適応システム
Dynamic Learning System Using Liner Time-Series Prediction and Reinforcement Learning

今井　智也[*1*2]
Tomoya Imai

三上　貞芳[*2]
Sadayoshi Mikami

[*1] 公立はこだて未来大学大学院システム情報科学研究科
Future University Hakodate, Graduate School of System Information Science

[*2] 公立はこだて未来大学
Future University Hakodate

Abstract: The purpose is to propose a dynamic system corresponding to flexibility to real world by combining the Linear Time-Series Predicting and Reinforcement Learning, detecting the change of environment, and forecasting other agent's behavior. The research showed a result that more excellent than past technique though it had error margin of forecast. It is possible for a higher performance by dynamically correcting the parameter of the forecast algorithm because decrease in the prediction error stops by a constant value.

1. はじめに

現在学習システムを用いたアプリケーションはロボットの行動制御やコンピュータゲームの戦略，音声・文字・画像認識など多分野に用いられている．時間とともに複雑に変化する環境や複数のエージェントが存在する場合，エージェント間の相互作用による競合が発生し学習効率が低下するという問題が生じる．そこで強化学習エージェントが時系列予測を用いて他エージェント群の挙動を予測し，競合を回避することで学習効率の向上を目指すことが本研究の目的となる．マルチエージェント[三上 99]における競合回避と目標最大化の実例として大型建造物の空調システムがあげられる．様々な外的要因で変化する環境を，エージェント同士が協調行動をとることで必要最小限の制御で温度を一定に保つことが可能である．また交通やネットワークのトラフィック管理のように混雑を予測し，学習により経路や交通指示器を自律的に制御するシステムがあげられる．本研究では単純な時系列予測・強化学習[Sutton 98]アルゴリズムを組み合わせたエージェントによるシミュレーション実験を用いてエージェントの競合回避および目標最大化への収束精度を検証する．

2. 提案手法および実験概要

本節ではマルチエージェントの競合回避および協調行動の獲得モデルとして扱う椅子取りゲーム問題[Chishima 07]について述べる．

2.1 椅子取りゲーム問題

マルチエージェントによる協調行動の獲得問題として椅子取りゲーム問題(図 1)を扱う．ここで扱う椅子取りゲーム問題とはエージェントと椅子が複数存在している状態で開始され，それぞれのエージェントは同時に椅子を 1 つ選択する．選択した椅子が競合した場合にはエージェントに－1，競合しなかった場合には＋1 の報酬を与える．これを 1 試行とし，エージェントは 1 試行毎に得た報酬の値によって学習を行う．試行が進むことでエージェント群は互いに競合を回避するような行動選択を行う．この状態への移行によってマルチエージェントが協調行動を獲得したと言える．

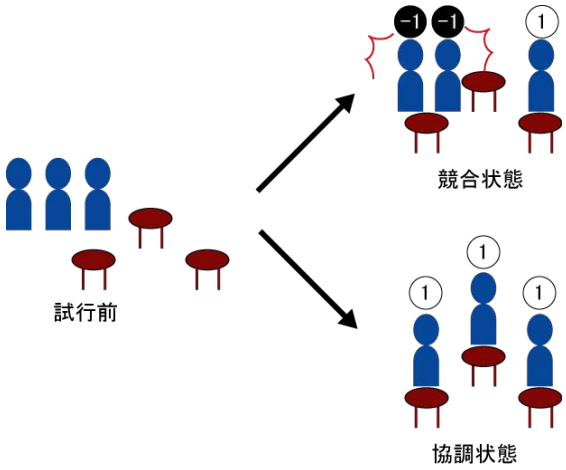

図 1: 椅子取りゲーム問題

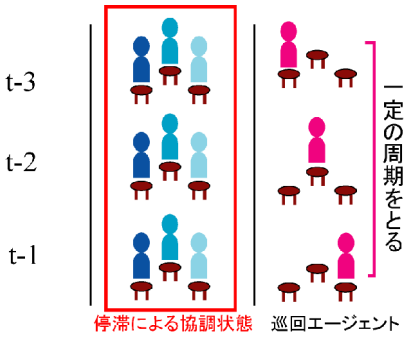

図 2: 巡回エージェント

学習エージェントのみで実験を行った場合，試行初期にエージェントがそれぞれ競合しない椅子を連続で選択し続ける(停滞する) ことによって簡易的に協調状態を獲得できてしまう問題が

連絡先：今井　智也，公立はこだて未来大学大学院　システム情報科学研究科，　〒041-0806 北海道函館市亀田中野町 116，℡0138-34-6419，g2109004@fun.ac.jp

生じる．そこで実験を行うにあたり学習エージェントの他に周期的行動選択を行う巡回エージェントを配置した(図 2)．巡回エージェントは試行毎に各椅子を巡回するような行動選択を行う．巡回エージェントが環境を常に変化させ続けるので，学習エージェントは停滞によって報酬を取得し続けることが出来ないため，先に述べた問題を回避することが出来る．

2.2 提案手法

本研究の提案手法(図 3)は線形時系列予測と強化学習を組み合わせることでマルチエージェント環境において，目標最大化とそれに伴い生じる競合によるパフォーマンスの低下を回避する手法である。

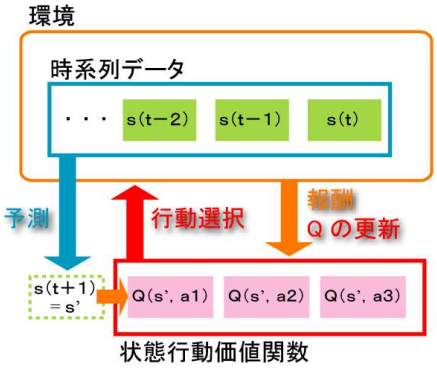

図 3: 提案手法

本手法では選択・行動・報酬を 1 試行(ステップ)として学習を行うことを前提とし，エージェントは試行毎に他エージェント群の行動結果から構成される状態 s を取得する．他エージェント群の行動結果を本手法では時系列データとして扱う．取得した時系列データに対して線形時系列予測を用いて次に取り得る行動予測を行う．行動予測を他の全てのエージェントに対して行いそれらを結合することで，次ステップで予測される全体の状態 s'を構成する．予測される状態 s'に対して最も価値の高い行動を状態行動価値関数 Q(s',a)より選択し行動を行う．行動結果に対して得られた報酬によって状態行動価値関数を更新し次のステップへ移行する．

椅子の数とエージェントの数が多いほど状態が広大になり，試行初期の目標最大化精度は低い．試行初期は予測される状態に対してどのような行動を取れば高い報酬が得られるか不明であるため，試行錯誤を行う必要がある．しかし試行が進むことで予測される状態に対して最も高い報酬が期待できる行動を，経験によって選択することが出来る．

3. 提案手法の評価実験

本手法の性能を評価するため，計算機実験を行う.実験は提案手法を用いたエージェント群に加え，比較として強化学習のみを用いて行動選択を行うエージェント群による実験を行った.

3.1 実験設定

実験は 3 椅子 3 エージェントで行う． エージェントの 1 つは 2.1 節で記述した巡回エージェントとし，残りのエージェントに各手法を用いる．予測手法として AR 回帰予測[Brockwell 04]，強化学習の学習方式に Q− Learning を用いた．行動選択規則は ε-greedy 法に従う．行動選択の ε 値はアニーリングスケジュールを用いた．これは試行初期にはランダムな行動を多く選択することで広く探索を行い，学習が進むにつれ探索を緩め報酬を優先する．本実験では椅子の番号をそれぞれ 0 番，1 番，2 番として扱っており，巡回エージェントは各椅子を 012012...のように一定のパターンで巡回を行う．

3.2 実験結果と考察

各手法における実験結果を図 4 に示す．図の縦軸はエージェント全体の競合率に対して 16 区間の移動平均を取ったものである．両手法とも試行が進むことで学習により競合率の低下が見られる．強化学習のみによって行動選択を行うエージェントは，ある一定の試行で学習による競合率の低下が打ち止めとなる．しかし提案手法によって行動選択を行うエージェントは試行が進むにつれ競合率が 0 に収束し協調状態への移行が確認できる．

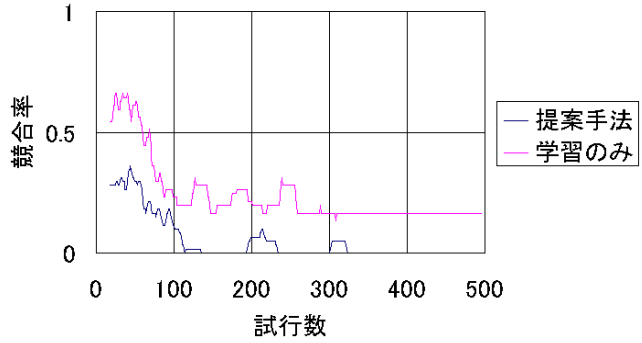

図 4: 離散場における実験での比較

提案手法は他エージェントの行動履歴を参照し，AR モデルを構築することで次の行動を予測する．本実験においてエージェントは 0 番，1 番，2 番という計算機上の番号によって行動選択を行っているが，椅子が離散的に存在し椅子番号自体に数学的な関連性が無いため，AR モデルを用いて正確に挙動を予測することは困難である．そのため実際には提案手法を用いたエージェントは，他エージェントの行動履歴を辿りその情報を用いて行動選択を用いることで，直前の行動のみを参照している強化学習エージェントに比べて行動選択の幅が広く，優れた競合回避精度を示したといえる．

4. 連続場での提案手法の評価実験

ここでは前節の実験条件を連続場へ拡張することで再度提案手法の有意性を確認する．

4.1 実験設定

3 節で扱った椅子取りゲーム問題は椅子に番号が振られ，エージェントが番号を選択することで行動が行われる．ここでは椅子番号を離散値から連続値へと変更し実験を行う．椅子を連続場に拡張することで，エージェントの行動選択と状態が分離され，AR モデルの構築がより正確に行われる．

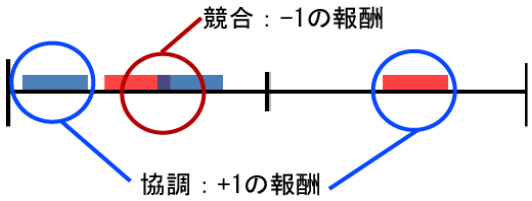

図 5: 連続場における椅子取りゲーム問題

実験は幅 10 の 1 次元フィールド(図 5) で行いエージェントは 1エピソード毎に正の方向に 2 まで，負の方向に 2 まで移動することが可能である．つまりエージェントは+2，+1，±0(停滞)，-1，-2 の 5 つの行動から 1 つを選択する．エージェント同士の場所が競合した場合は-1，競合しなければ+1 の報酬を与える．実験は 4 つのエージェントによって行い 2 つは学習エージェント，残りの 2 つは巡回エージェントを用いる．ここでの巡回エージェントは 1 エピソード毎に 1，2 の速度でフィールドを往復する．

4.2 実験結果

各手法における実験結果を図 6 に示す．図の縦軸は各手法を用いたエージェントの報酬を表す．

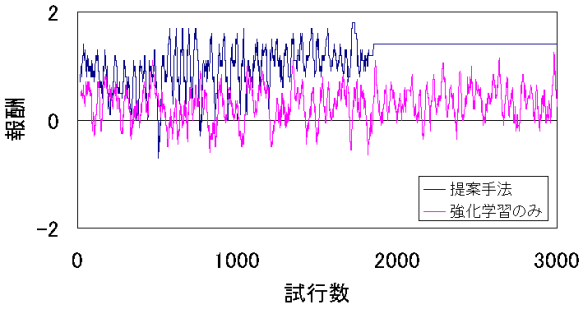

図 6: 各手法の比較

強化学習のみを用いて行動選択を行うエージェントは試行を重ねても協調状態への収束が見られない．強化学習エージェントは他エージェントの直前の座標を参照し行動選択を行うが，他エージェントの座標情報だけではエージェントの進行方向や速度に着目し競合を回避することができず，学習の進行は行われない．提案手法を用いて行動選択を行うエージェントは試行が一定数経過すると協調状態へ収束した．提案手法は他エージェントの状態履歴を参照することでエージェントが次にどのような挙動を示すかを予測し，予測される状態において最も報酬が期待できる行動を選択する．

図 7 は実験における予測誤差を表す．予測誤差の算出は次エピソードにおける予測結果と，実際の結果の差である．試行数 2000 手前で学習が収束し，エージェントが同じ行動を繰り返す状態となる．この状態へ移行することで予測の誤差に変動が生じなくなり，これ以上のパフォーマンスの向上は打ち止めとなる．

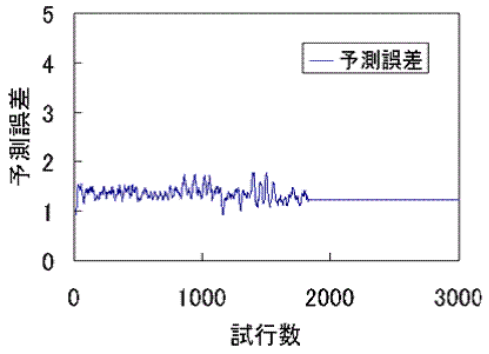

図 7: 予測誤差

5. 考察

本実験に用いた椅子取りゲーム問題はマルチエージェント環境におけるシンプルな競合回避問題であるが，シンプル故に挙動の解析が困難であり予測の精度が低い．また連続場での実験においても予測精度の頭打ちと同時に学習による経験の獲得が停止する現象が生じる．しかし両実験とも予測精度の低さに関わらず強化学習のみを用いて行動選択を行うエージェントに比べ高いパフォーマンスを示した．AR モデルは線形予測の中でも最も簡単なモデルである．連続場での実験において得られた結果から強化学習単体によって行動選択を行うエージェントでは収束できない問題を，単純な時系列予測を組み込むだけで収束させることが可能であるという点に本手法の優位性を評価することができた．

今後の課題として多次元フィールドやエージェント数を大幅に増加させた場合において，提案手法がどのような性能を示すかを確認する必要がある．

参考文献

[三上 99] 三上貞芳，嘉数侑昇:マルチエージェント系における機能創発，計測と制御, 38-10,pp.630-635, 1999.

[Sutton 98] Sutton, R. S., and Bart, A. G.: “Reinforcement Learning：An Introduction”, MIT Press， 1998. (三上 貞芳, 皆川 雅章 共訳：強化学習，森北出版，2000.)

[Chishima 07] 千島洋徳: 時系列予測と強化学習を用いたマルチエージェント型学習コントローラ, IPSJ2007, 2007.

[Brockwell 04] Introduction to Time Series and Forecasting : シーエーピー出版, 2000.