岡山理科大学紀要 第41号 A pp. 105-112(2005)

RoboCupサッカーシミュレーションリーグにおける強化学習の有効性

金重 徹・片山 謙吾*・南原 英生*・成久 洋之*
岡山理科大学大学院工学研究科情報工学専攻
*岡山理科大学工学部情報工学科
(2005年9月30日受付、2005年11月7日受理)

1. まえがき

近年，単純な能力を持つエージェントを複数組み合わせ，エージェントの協調行動によって複雑な問題を解くマルチエージェントシステム (Multi Agent System : MAS) が注目されている．MASの標準問題として注目されているのがRoboCupサッカーである．

RoboCupサッカーは，“西暦2050年までに，サッカーの世界チャンピオンに勝てる，自律型ロボットチームを作る”という目標を掲げたランドマークプロジェクトである．サッカーはエージェント個人の能力だけでは勝つことができず，味方と協調することが重要となる．また，敵エージェントとの競合も問題となっておりマルチエージェント環境である．しかし，RoboCupサッカーのようなMASを設計する際，環境に存在する他のエージェントとの相互作用を考慮しながら設計しなければならず，予め設計者が全ての環境を想定し，プログラム化することは非常に困難である．そこで，エージェントが自ら環境を認識し，環境と相互作用しながら行動を獲得する強化学習 (Reinforcement Learning)[1] などが注目されている．

本論文では，RoboCupサッカーのシミュレーションリーグを対象とし，強化学習を有するRoboCupサッカーエージェントの有効性を検討する．

本論文の構成は，2. 章をRoboCupサッカーについての概要，3. 章を強化学習の概要，4. 章を強化学習の適用の詳細と問題設定，5. 章を実験結果と考察とし，最後にむすびとする．

2. RoboCupサッカー

RoboCupサッカーとは，1990年代前半に日本の研究者によって提唱されたランドマークプロジェクトである[2]．RoboCupサッカーの目的は，“西暦2050年までに，サッカーの世界チャンピオンに勝てる，自律型ロボットチームを作る”という目標を掲げ，その達成過程で作り出される様々な技術を，社会的・産業的に重要な分野に応用することである．

RoboCupサッカーには，以下に示す5種類のリーグが存在する．

- 小型ロボットリーグ
  卓球台ほどの大きさのフィールドで，直径18cm以内のロボット5台1チームで規定されたオレンジ色のゴルフボールを使って対戦するリーグである．
- 中型ロボットリーグ
  直径50cm以内のロボットが，卓球台9枚ほどの大きさのフィールドで，オレンジ色のボールを追う競技である．
- 四足ロボットリーグ
  ソニーのAIBOをプラットフォームとして使った4台1チームのサッカーリーグである．共通のプラットフォームを採用しているため，各チームのロボットプログラミングによって試合の勝敗が左右される．
- ヒューマノイドロボットリーグ
  このリーグは2002年大会から正式種目となった自律型2足歩行ロボットによるリーグである．「歩く」，「ボールを蹴る」といった基本動作を試す競技や，独自の機能を披露する「フリースタイル」競技や，2対2での対戦が行われる．

- シミュレーションリーグ
  ロボットの実機を使うことなく，コンピュータ上の仮想フィールドで，プログラミングされた 11 対 11 のバーチャルロボットが 5 分ハーフのサッカーを行うリーグである．各リーグの中で一番最初に提唱され，最も洗練された動きをする．

本論文では，5 種類のリーグの中でサッカーの技術面のみに集中できるシミュレーションリーグを対象とする．RoboCup サッカーシミュレーションリーグは，マルチエージェントシステムにおける協調行動のほか，

- 変化する環境において決められた時間単位に次の動作計画を行う実時間処理
- 一部分，誤差を含んだ入力情報で処理を行う不完全情報処理
- これらを実現するプログラムアーキテクチャ

などの研究分野のテーマを含んでいる[3][4]．

3. 強化学習

強化学習は最適な行動を人間が学習主体 (エージェント) に教えるのではなく，エージェント自身が環境との試行錯誤を通して得た行動の結果から，自律的に意思決定の方策をより良いものへと構築していく学習手法である．目的を達成した際にスカラー値の報酬を与える事によってのみ学習を行う．しかし報酬にはノイズや遅れがある．そのため，行動を実行した直後の報酬をみるだけでは，学習主体はその行動が正しかったかどうかを判断できないという困難を伴う．強化学習が注目を集めている理由は以下の 2 つである．

不確実性のある環境
多くの実世界の制御問題では，不確実性の扱いは厄介である．しかし，強化学習はエージェント自身が環境との試行錯誤を通して学習するので，不確実性を含む環境でも有効に制御することが可能となる．

離散的な状態遷移も含んだ段取り的な制御
設計者が目標状態で報酬を与えるという形で，させたいタスクをエージェントに指示しておけば，ゴールへの到達方法はエージェントの試行錯誤学習によって自動的に獲得される．つまり，設計者が「何をすべきか」をエージェントに報酬という形で 指示しておけば「どのように実現するか」をエージェントが学習によって自動的に獲得する枠組である．

3.1 強化学習の可能性

強化学習を適用することで，考えられる利点は以下の 3 つが挙げられる．

制御プログラミングの自動化
環境に不確実性や計測不可能な未知のパラメータが存在すると，タスクの達成方法及び，目標状態への到達方法は設計者にとって自明でない．よって設計者が予め設計することは非常に困難である．しかし達成すべき目標を報酬によって明示することは前述に比べ簡単である．そのため，目標達成までのプログラミングを強化学習で自動化することで，設計者の負担軽減が期待できる．

ハンドコーディングよりも優れた解の検出
強化学習は，試行錯誤を通じて学習するため，人間のエキスパートが得た解よりも優れた解を発見する可能性がある．特に人間の常識では対処しきれない未知なパラメータが多い場合，強化学習の効果が期待できる．

自立性と想定外の環境変化への対応
機械故障などの急激な変化など，予め事態を想定しプログラミングしておく事が困難な環境の変化に対しても自動的に行動を獲得することが期待できる．

3.2 強化学習エージェントの構成

強化学習エージェントの構成を図 1 に示す．強化学習エージェントは，状態認識器，学習器，行動選択器の 3 つのモジュールから成っている．状態認識器は，現在の状態を認識する．学習器は，各状態での各行動の重みを蓄えており，状態認識器によって認識した状態での，各行動の重みを行動選択器に渡す．行動選択器では，その重みにもとづき行動を選択する．行動した結果，環境が目標状態となった時に報酬を与え，各行動の重みを強化する．

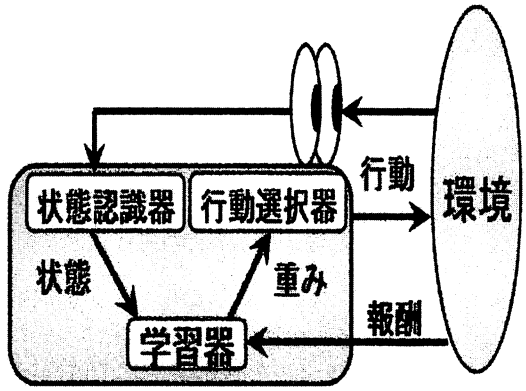


図 1　強化学習エージェントのモデル図

3.3 マルコフ決定過程

多くの場合，エージェントと環境との相互作用は，マルコフ決定過程 (Marukov Decision Processes : MDPs)[5] によってモデル化できる．マルコフ決定過程では，次の状態 s'，報酬 r，行動 a，履歴 $a_t, s_t, r_t, ..., a_0, s_0, r_1$ に対し，

$$P_r(s_{t+1} = s', r_{t+1} = r | a_t, s_t, r_t, ..., a_0, s_0, r_1) = P_r(s_{t+1} = s', r_{t+1} = r | a_t, s_t)$$

が成り立つ．つまりマルコフ決定過程では，「次の状態 s' が観測され，かつ報酬 r が得られる確率」は直前の状態と行動のみに依存する．

任意の状態 s と行動 a が与えられたときの，次に遷移可能な状態 s' の確率を $P^a_{ss'}$ と定義する．

$$P^a_{ss'} = P_r(s_{t+1} = s' | s_t = s, a_t = a)$$

同様に，任意の状態 s と行動 a および任意の次状態 s' が与えられた時の，次に得られる報酬の期待値を $R^a_{ss'}$ と表す．

$$R^a_{ss'} = E(r_{t+1} | s_{t+1} = s', s_t = s, a_t = a)$$

マルコフ決定過程は，環境が取りうる状態集合 S，エージェントの取りうる行動の集合 A，状態遷移確率 $P^a_{ss'}$，報酬の期待値 $R^a_{ss'}$ によって定義される．

3.4 強化学習の主な手法

強化学習は主に手法の違いによって「環境同定型学習」,「経験強化型学習」の 2 種類に分類される[6]．

- 環境同定型学習
  試行錯誤的に行動を繰り返し，環境を把握することで最適解を導く．しかしなるべく多くの環境状態を把握する必要があり，環境状態数が増加すると最適解を得るまでに非常に時間がかかり，解を得られないこともある．代表例として Q-Learning がある．

- 経験強化型学習
  環境を把握するよりも，いかにして多くの報酬を得るかということを目的とする．毎回の行動に対し報酬を与えるのではなく，目的達成時に報酬を与え，行動開始から目的達成までの状態行動対に振り分ける．環境の一部しか把握しないので最適解を得る可能性は低いが，収束が速く穏かな環境変化にも対応できる．代表例として Profit Sharing がある．

3.5 Profit Sharing 強化学習

Profit Sharing 強化学習は，非 MDP となるようなマルチエージェント環境において，有用であると期待されている．また，Profit Sharing は他の強化学習に比べ，学習の立ち上がりが素早く，不完全知覚に対しても有効であることが示されている[7][8]．このことから本論文では，Profit Sharing を学習手法として用いる．

Profit Sharing の合理性定理

Profit Sharing は，報酬を得たときにそれまでに使用された状態行動対（s_t, a_t）をエピソード単位で強化する手法である．エピソードとは，初期状態あるいは報酬を得た直後から次の報酬までのルールの選択系列のことである．

次式を用いて重み $w(s_t, a_t)$ を更新する．ここで，$w(s_t, a_t)$ はエピソード上の t 番目の重み，r は報酬値，f は強化関数である．

$$w(s_t, a_t) \leftarrow w(s_t, a_t) + f(r, t)$$

あるエピソードで，同一の感覚入力（状態）に対して異なるルールが選択されているとき，その間のルールを迂回系列という．常に迂回系列上にあるルールを無効ルールと呼び，それ以外を有効ルールと呼ぶ．無効ルールと有効ルールが競合するならば，無効ルールを強化すべきではないと考えられる．そこで本論文では，政策の局所的合理性を保証する必要十分条件が証明されている合理性定理にしたがい，無効ルールを抑制する次の強化関数を使用する[9]．

$$f(r, t) = \frac{1}{S} f(r, t-1), t = 1, 2, \cdots N-1.$$

ここで，N はエピソードの最大長，S は報酬割引率である．なお，報酬割引率は $S \geq L+1$ とする（L は同一感覚入力下に存在する有効ルールの最大個数である）．

3.6 ルーレット選択法

Profit Sharing の学習過程における行動選択法としては，ルーレット選択法が良い性能を示すことが知られている．ルーレット選択法は，ある状態 s において，各行動の重み $w(s, a_t)$ を全ての行動の重みの合計 $\sum w(s, a_t)$ で割り，確率 $P(a_t|s)$ を求め，その確率により行動を決定する方法である．

$$P(a_t|s) = w(s, a_t) / \sum w(s, a_t)$$

また，ルーレット選択法は，非 MDP 環境における行動選択法として有効である．このような理由から本論文では行動選択法としてルーレット選択法を使用する．

4. 強化学習の適用と実験設定

本章では，本研究での強化学習の構成と実験設定を述べる．

ベースエージェントプログラムは電気通信大学の YowAI2002 のベーシックプログラムとし，本研究で設計した強化学習を有する 2 人の味方エージェントに対し，ベースプログラムの敵エージェント（DF）2 人による 2 対 2 の対戦で実験を行う（図 2 参照）．なお，各エージェントの初期配置は固定とし，ボールの初期位置は味方エージェントの初期位置とする．

以下に，本研究で設計した強化学習エージェントの状態，行動，報酬について述べる．

強化学習の状態

サッカーのフィールドを離散化するために，敵陣ハーフコートを 8×8 の格子状に分割する．そこで認識する状態は，自分の位置と他のエージェントを観測した格子の位置とする (図 3 参照)．

強化学習の行動

本研究では，強化学習を階層的に設計している．図 4 は，本研究で用いた強化学習の階層構造を示したものである．

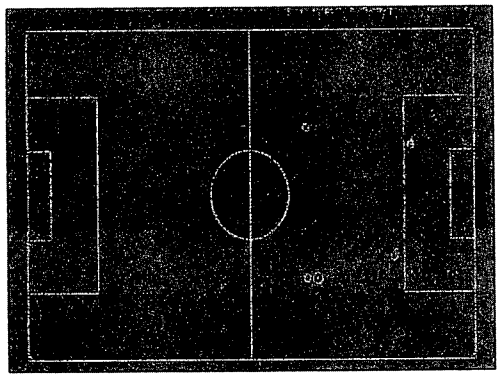


図 2　実験環境

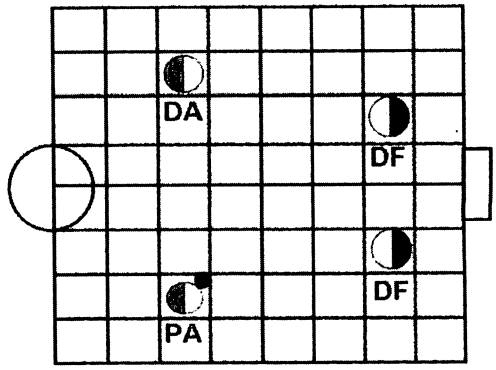
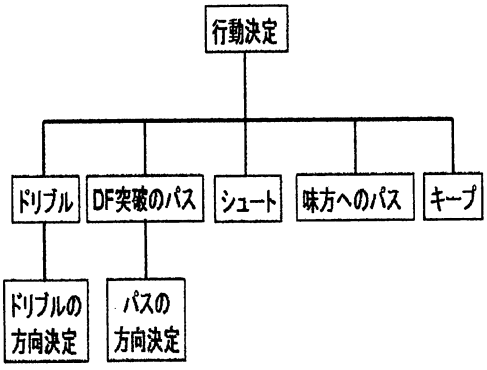
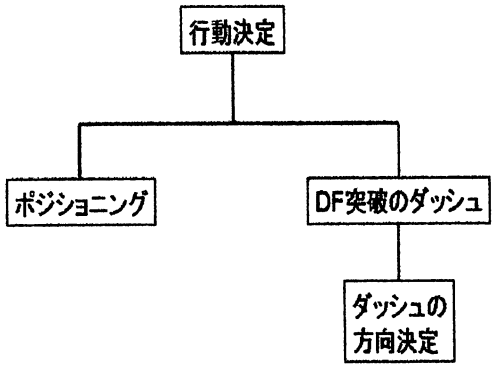


図 3　強化学習の状態



(a) PA の強化学習に関する階層構造



(b) DA の強化学習に関する階層構造

図 4　強化学習の階層構造図

まず，ボールを所有しているエージェント（PA）について説明する．図 4(a) は，PA の階層構造を示したものである．PA は，ドリブル，DF 突破のパス，シュート，味方へのパス，キープの 5 種類の行動決定に関して強化学習を用いる．さらに，行動決定においてドリブルまたは，DF 突破のパスが選択された場合は，方向決定に関して強化学習を用いる．

次に，ボールを所有していないエージェント（DA）について説明する．図 4(b) は，DA の階層構造を示したものである．DA はポジショニング，DF 突破のダッシュの 2 種類の行動決定に関して，強化学習を用いる．さらに，PA と同様に行動決定において DF 突破のダッシュが選択された場合は，方向決定に関して強化学習を用いる．

強化学習の報酬

行動決定及び，ドリブルの方向決定における報酬は，得点を上げるまでの行動群に与える．DF 突破のパスの方向決定における報酬は，パスを味方に渡すことができれば報酬を与え，DF 突破のダッシュの方向決定における報酬は，ダッシュを行っている間にボールを受け取る事ができれば報酬を与える．

ボールを DF に奪われる，もしくはセットプレーになった場合は学習を終了とする．なお，報酬の有無に関わらずここまでを 1 エピソードとする．

5. 実験結果と考察

RoboCup サッカーシミュレーションリーグを対象に，強化学習を有するエージェントの有効性を検討するために実験を行う．図 5 は学習回数を 1 万エピソードとし，強化学習を適用したエージェント及び，ベースプログラムの FW の得点率を 500 回ごとに平均し，プロットしたものである．なお，横軸は学習回数，縦軸は得点率である．図 5 の結果から，強化学習を適用したエージェントの学習初期の得点率は約 9 %であ

るのに対し，学習後期では約 25 %で明らかに得点率が向上していることが観測できる．さらに学習回数を重ねる事で得点率の向上が望める．一方ベースプログラムは，学習を行わないため得点率が約 9 %のまま一定である．この結果より，設計者が全ての環境を想定することが困難である場合，エージェント自らが適応していくことが非常に重要であると考えられる．よって RoboCup サッカーシミュレーションリーグに強化学習を適用することは有効である．

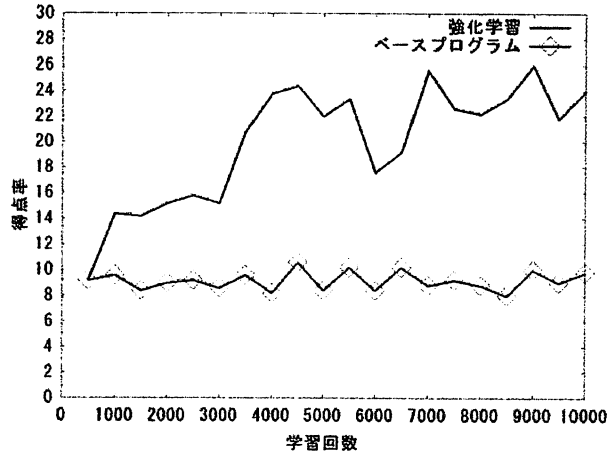


図 5 実験結果

6. むすび

本研究では，RoboCup サッカーシミュレーションリーグを対象とし，強化学習の有効性を検討した．その結果，本研究の実験設定において強化学習は，プログラム化されたエージェントの得点率をはるかに上回り，有効であることを示した．この結果から，RoboCup サッカーの様に我々設計者が全ての環境を想定することが困難である場合，エージェント自身が学習し，環境に適応することが非常に重要であると考えられる．

しかし，RoboCup サッカーの目的でもある現実問題の応用に，強化学習を適用することを考えた場合，学習速度の遅さが問題である．なぜならば，現実問題では迅速な対応・安全性・確実性が求められるからである．迅速な対応で言えば，実機（ロボット）に強化学習を適用した際，学習する前に壊れてしまうという可能性が考えられる．そこで，今後の課題は，文献[10] で提案されている手法を RoboCup サッカーシミュレーションリーグに適用し，その有効性を検討することである．

参考文献

[1] Richard S. Sutton, Andrew G. Barto [著] 三上 貞芳, 皆川 雅章 共訳, “強化学習,” 森北出版, 2000.

[2] 土屋 俊, 中島 秀之, 中川 祐志, 橋田 浩一, 松原 仁, 大澤 幸生, 高間 康史, “AI 辞典 第 2 版,” 共立出版社, 2003.

[3] 野田 五十樹, “RoboCup におけるマルチエージェントシミュレーション,” セルオートマトン・シンポジウム講演論文集, pp.63–68, 2001.

[4] 高橋 友一, 伊藤 暢浩 著, “RoboCup ではじめるエージェントプログラミング,” 共立出版社, 2001.

[5] Bellman, R. E., “A Markov decision process,” Journal of Mathematical Mechanics, Vol.6, 679–684, 1957.

[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.23-29, 1995.

[7] 宮崎和光, 木村元, 小林重信, “ProfitSharing に基づく強化学習の理論と応用,” 人工知能学会誌, Vol.14, No5, pp800-807, 1999.

[8] 宮崎和光, 小林重信, “離散マルコフ決定過程下での強化学習,” 人工知能学会誌, Vol.12, No6, pp.3-13, 1997.

[9] 宮崎和光, 山村雅幸, 小林重信, “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, Vol9, No4, pp580-587, 1993.

[10] 片山謙吾, 奥石尚宏, 成久洋之, “強化学習エージェントへの階層化意志決定法の導入 — 追跡問題を例に —,” 人工知能学会論文誌, Vol.19, No.4, pp.279-291, 2004.

Effectiveness of Reinforcement Learning for RoboCup Soccer Simulation League

Toru KANESHIGE, Kengo KATAYAMA*, Hideo MINAMIHARA* and Hiroyuki NARIHISA*

Graduate School of Engineering, Okayama University of Science.
**Department of Information and Computer Engineering, Faculty of Engineering,*
Okayama University of Science.
1 - 1 Ridai-cho, Okayama, 700-0005, Japan.

(Received September 30, 2005; accepted November 7, 2005)

Multi-agent systems in real world require agents to act effectively and autonomously. Particularly RoboCup Soccer is used as a standard multi-agent environment and is known to be one of the most difficult tasks and challenging problems in complex, real-time domains. It is difficult to create intelligent agents for RoboCup Soccer because the environment of soccer is very complex. The Reinforcement Learning is one of the most promising approaches to create agents for such complex environments. In this paper, we apply Reinforcement Learning to RoboCup Simulation League. As a result, we show the effectiveness of multi agents created by Reinforcement Learning.