社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－GI－11 (7)
2004／3／8

サッカー協調プレイの強化学習のための状態縮約

The contraction state for reinforcement learning in soccer cooperation play

鹿田 和之進[1]　西野 順二
Kazunoshin Shikada　Junji Nishino

電気通信大学 システム工学専攻
Dept. of Systems engineering
The University of Electro-Communications

Abstract: In this paper, we show an empirical result on effects of state contraction in cooperative multi agent behavior rule using reinforcement learning method for 'centering tactics' in RoboCup soccer. We examine a difference between finely divided state space and roughly one. Learning result of contracted state with 30 games perform as well as fine-divided state with 60 games. The result of fine-contract combined method also perform well.

1 はじめに

サッカーのような複数のエージェントが、協調、競合して点を取り合うマルチエージェント環境における協調行動の実現は、非常に興味のある話題であり、これまで様々な研究がされてきた。マルチエージェント環境では、学習者が1人であったとしても、他のエージェントの行動政策が未知の場合、容易に状態空間を構成できない。内部 [1] らは、学習者の観測と行動を通して、学習者と他者の行動の関係を局所予測モデルとして推定し、その結果をもとに強化学習を行った。また、複数エージェントが同時にゼロから学習する困難な問題に対して、浅田 [2] は、進化的手法を導入し協調行動の実現に成功している。

本研究で扱う RoboCup Soccer シミュレーションリーグでは、ドリブルやシュートなどといった個人レベルの基礎スキルや、フォーメーションなどが主となって研究されている。現実のサッカーではさらに上のレベル、戦術や戦略といった協調プレイが必要とされる。しかし、協調プレイを Q-learning や強化学習によって実現するうえで、状態、行動の増加による計算量が問題となり、学習が成り立たないことが多い。

本研究では、代表的な強化学習手法である Q-learning において、その学習効率を上げる為の状態縮約について検討する。そして、戦術の中で特に重要な協調的プレイを獲得することを目的とする。

2 強化学習における状態の縮約

強化学習とは、環境から与えられる報酬（reward）をもとに、期待報酬を最大化するような行動を徐々に選択するように、行動を修正していく教師なしの学習法である。教師付き学習と異なり、状態入力に対する正しい行動出力を明示的に示す教師が存在しない。また、報酬にはノイズや遅れがあり、行動を実行した直後の報酬をみるだけでは、エージェント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。

2.1 強化学習の枠組み

強化学習の枠組を図1に示す。エージェントは環境の状態を表す状態集合 S を観測し、行動集合 A の中の一つの行動を選択する。このと

[1]鹿田 和之進　電気通信大学
〒 182-0021 東京都調布市調布ヶ丘 1-5-1
Tel:0424-43-5800,Fax:0424-43-8020
E-mail:shika@fs.se.uec.ac.jp
本研究の一部は科学研究費補助金、課題番号 14750179 によって行われた。

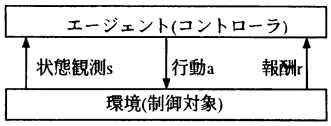


図 1: 強化学習の枠組

き、環境はマルコフ過程としてモデル化され、現在のエージェントがとった行動により確率的な状態遷移を行う。

現在の状態を s、エージェントがとった行動を a、次状態を s' とすると、$T(s,a,s')$ は、そのときの状態遷移確率を表す。ここで、状態から次に選択すべき行動への写像が決まる。この写像を方策（policy）という。また、それぞれの状態・行動のペア (s,a) に対し、報酬 $r(s,a)$ が定義される。

エージェントの目的は、環境中を行動しながらタスクを完遂するまでに、できるだけ多くの報酬を獲得することである。ここで、報酬の合計である積算報酬（retrun）R_t は、

$$R_t = \sum_{n=0}^{\infty} \gamma^n r_{t+n} \qquad (1)$$

で定義される。ここで、γ は割引報酬率を表し、将来の報酬がどの程度行動価値に影響を与えるかを決める。

2.2 Q-learning

実際の環境では、状態遷移確率と報酬の分布が完全に既知である場合は少なく、試行錯誤的に探索しながら、最適な行動を学習することが考えられる。以下では、このような探索的な学習法である Q-learning について説明する。

状態 s で行動 a をとり、それ以降最適政策をとったときの積算報酬の期待値、もしくは最適行動価値関数を $Q^*(s,a)$ をとると、

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \underset{a' \in A}{max} Q^*(s',a') \qquad (2)$$

と定義される。最初は、遷移確率や報酬が未知なので、オンラインで逐次的に行動価値 Q（以下Q値）を更新する。Q値は初期値0から始め、行動がとられるごとに、Q値を更新する。

本研究では、以下の Q-learning アルゴリズム [4] を用いた。

1. エージェントは環境の状態 s_t を観測する。
2. エージェントは任意の行動選択方法 (探査戦略) に従って行動 a_t を実行する。
3. 環境から報酬 r_t を受け取る。
4. 状態遷移後の状態 s_{t+1} を観測する。
5. 以下の更新式により Q 値を更新：

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha[r_t + \gamma \underset{a}{max} Q(s_{s+1},a_t)] \qquad (3)$$

6. 時間ステップ t を $t+1$ へ進めて手順1へ戻る。

2.3 サッカーでの状態とその縮約

サッカーでは、$108[m] \times 50[m]$ の連続二次元空間のフィールドで試合が行われる。フィールドには、2チーム22人のプレイヤーとボールが存在し、計23の状態変数がある。試合の状態数は、1[m] 単位に離散化したとしても、$(108 \times 50)^{23}$ 状態あることになる。この状態では、学習を行う環境に適していない。本研究では、大雑把に空間を分割することにより状態縮約し、このような環境で Q-learning を用い学習を行う。

3 縮約の効果の検証

縮約の効果について実験的に検証する。

3.1 RoboCup サッカーサーバシステム

RoboCup サッカーサーバは仮想的なフィールドを提供し、ボールとプレイヤーの全ての動

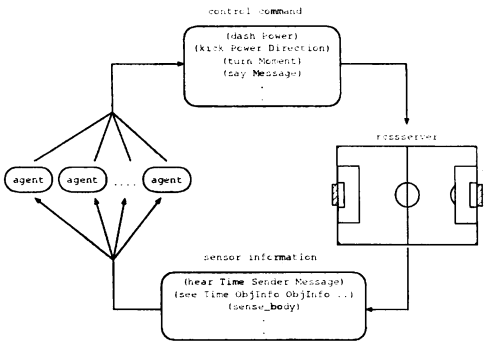


図 2: サッカーサーバシステム

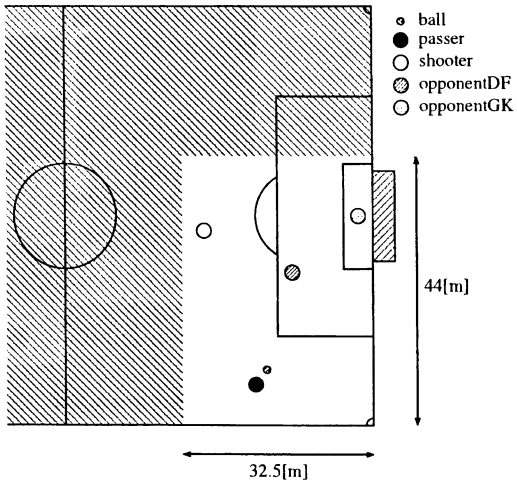


図 3: 実験環境

きを 2 次元でシミュレートする。また、サッカーサーバはプレイヤーの視野である視覚情報や、プレイヤーの位置や体力などが含まれる身体情報などを各プレイヤーに送り、各プレイヤーはこれらの情報から周囲の状況や自らの状態を分析し、走る（dash）、ボールを蹴る（kick）といった行動をサッカーサーバに送る。この一回の送受信を 1 ステップ (0.1 秒) として、試合は 6000 ステップ (10 分) 繰り返し行なわれる。システム構成を図 2 に示す。また、プレイヤー間では情報共有ができないが、かけ声（say）を用いることによりコミュニケーションをとれるが、本研究では用いないこととする。

3.2 サッカーエージェント

プレイヤーは RoboCup2003（世界大会）で優勝した UvA_trilearn2003[3] を味方 shooter、敵ディフェンダー（以下ＤＦ）、敵ゴールキーパー（以下ＧＫ）で使用した。また、学習者である passer プレイヤーは、UvA_base2003 を使用し作成する。

この UvA_trilearn2003 はアムステルダム大学の Jelle Kok が作成したプレイヤーで、coordination graph を用いてマルチエージェント環境での高度な意志決定を行っている。また、得点力が強くＧＫと 1 対 1 では高い確率で得点する事ができる。

UvA_base2003 は UvA_trilearn2003 の基礎ライブラリ（ドリブルやパスなど）を公開している。本研究では、このライブラリを使用し高度な意志決定の学習を試みる。

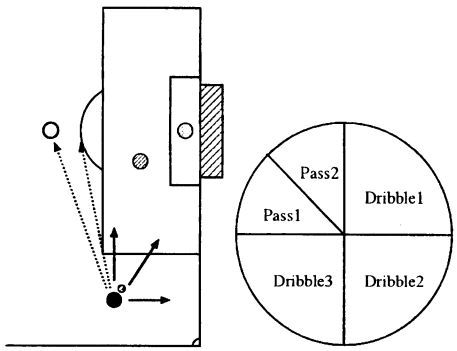
3.3 実験設定

本実験では、図 3 のような 44[m]×32.5[m] のフィールドで行う。この実験では、オフェンスの学習者 passer、味方 shooter、ディフェンスの敵ＤＦ、敵ＧＫの計 4 エージェントのみとする。passer は Q-learning により意志決定方策を学習していく。ここでは、passer がサイドから shooter にセンタリングをあげ、ゴールを狙う戦術を行う。

passer はボールを保持しているとき、ドリブルをしてゴールに近づくか、あるいはセンタリングをするかの意志決定を学習する。ボールを保持していないときは、shooter よりボールに近いときはボールを取りに行き、それ以外のときは、決められたポジションに移動する。

passer の行動として図 4(a) に示した以下の行動を用意した。

1. 上へドリブルする。(Dribble1)
2. 右上へドリブルする。(Dribble2)
3. 右へドリブルする。(Dribble3)



(a) passer の行動　　(b) ルーレット方策

図 4: passer の行動

4. shooter の方向へセンタリングをあげる。(Pass1)

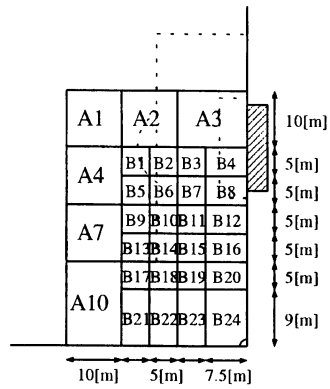
5. shooter の 5m 前方へセンタリングをあげる。(Pass2)

また、学習の過程で用いる方策は図 4(b) のようなルーレット方策を用いた。これは、パスをするとその時点でエピソードが終了し、学習の効率が著しく低くなることを避ける為である。ルーレット方策は、パスをする確率を低くすることで早い終了を回避している。

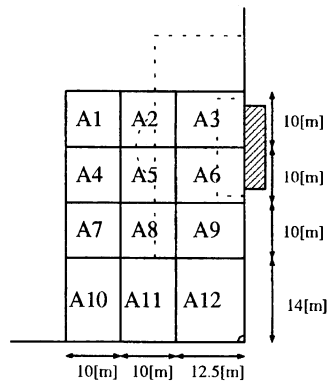
実験 1 では、空間を細かくして精度は良いが、収束速度の遅い学習を行う。実験 2 では状態空間を縮約し、空間を粗く分割し収束速度が早い学習を行う。実験 3 では、まず実験 2 で得られた Q 値を利用して、実験 1 で使用できるように Q 値を拡張し、さらに、実験 1 と同様の状態空間で学習を行う。これより、実験 3 のような状態を改めて変更することで、収束速度が速く精度の良い学習できたかを検証する。

3.4 実験 1（細分）

実験 1 では、図 5(a) のように空間を細かく分割し学習を行った。passer の位置を X_p、shooter の位置を X_s、敵 DF の位置を X_d とした。ここで、X_p, X_s, X_d は以下のように定義する。



(a) 細かい空間分割



(b) 縮約した空間分割

図 5: 空間分割

$$X_p \in \{A7, A10, B9, B10, .., B24, o\} \quad (4)$$

$$X_s \in \{A1, A2, A3, A4, A5, A6, o\} \quad (5)$$

$$X_d \in \{A2, A3, A11, A12, B1, B2, .., B16, o\} \quad (6)$$

ここで、o はそれ以外の場所である。

敵 GK はゴール前にいるものとして、状態変数としていない。よって、状態空間 (X_p, X_s, X_d) は、$19 \times 7 \times 21 = 2793$ 通りある。また、Q 値の状態数は $2793 \times 5 = 13965$ 通りとなる。

状態の初期値 $(X_{p_0}, X_{s_0}, X_{d_0})$ は、「その他」のエリアを除き、

$$X_{p_0} \in \{A7, A8, A9, A10, A11, A12\} \quad (7)$$

$$X_{s_0} \in \{A1, A2, A4, A5\} \quad (8)$$

$$X_{d_0} \in \{A2, A3, A5, A6, A8, A9, A11, A12\} \quad (9)$$

で定義され、この $6 \times 4 \times 8 = 192$ 通りからランダムで決められる。シュートが入るか、敵GKにボールを取られるか、フィールド外にでたら1エピソード終了とする。報酬はシュートが入った時に100を与える。

サッカーサーバは1試合6000[step]（10分）であり、本実験では、試合回数を60回に設定し学習を行った。1試合におけるQ値の更新回数は約600回、エピソード数は約130回であり、60試合のQ値の更新回数は約36000回、エピソード数は約7800回であった。また、学習率 α を0.5とし、割引報酬率 γ を0.7に設定した。

学習が終了した後、学習の効果測定して得られたQ値を使用としてさらに30回試合を行った。passerの方策にQ値の最大値を選ぶグリーディ方策を用いた。最大のQ値が0の場合、ランダムで行動を決定した。

3.5 実験2（縮約）

実験1では図5(b)のような A_1〜A_{12} の空間分割を行い、passerの位置を X'_p、shooterの位置を X'_s、敵DFの位置を X'_d とした。ここで、X'_p, X'_s, X'_d は以下のように定義する。

$$X'_p \in \{A7, A8, A9, A10, A11, A12, o\} \quad (10)$$

$$X'_s \in \{A1, A2, A3, A4, A5, A6, o\} \quad (11)$$

$$X'_d \in \{A2, A3, A5, A6, A8, A9, A11, A12, o\} \quad (12)$$

ここで、o はそれ以外の場所である。

状態空間 (X'_p, X'_s, X'_d) は $7 \times 7 \times 9 = 441$ 通りある。また、Q値の状態数は $441 \times 5 = 2205$ 通りとなる。状態空間の初期値、学習率、割引報酬、学習者の方策は実験1と同様である。但し、試合回数は実験1の半分の30回である。

表1: 各30試合の結果

	得点の平均	得点の標準偏差
学習前	26.7	3.28
実験1（細分）	35.4	4.75
実験2（縮約）	33.4	4.61
実験3（段階）	34.7	3.72

学習が終了した後、実験1と同様に学習の効果測定として得られたQ値を利用してさらに30回試合を行った。

3.6 実験3（段階的縮約）

実験2のQ値を用いて実験1の状態空間に適応できるように拡張した。例えば、$B1, B2, B5, B6$ に $A5$ のデータを使う。そして、実験1と同様な条件で学習させる。但し、学習回数は半分のは30回とする。

学習が終了した後、実験1、2と同様に学習の効果測定として得られたQ値を利用してさらに30回試合を行った。

4 縮約した状態の効果

表1は実験1，2，3の各30試合の結果である。平均得点、得点の標準偏差を記す。

学習前と実験1の得点を比較すると、学習者のセンタリングにおける、行動の学習に成功していることがわかる。試合を詳細にみると、図6のような、協調行動がみられる。passerは敵DFを引きつける為に、ドリブルで敵DFに接近し、近づいて味方にセンタリングをあげるという、高度な意志決定を行っており、戦術な協調行動を獲得することができた。しかし、成功率が低い理由として、空間の分割が粗く、大雑把な意志決定しかできないこと、行動を5個に制限したため、パス精度が悪いことも多かったためと考えらる。

また、 shooterであるUvA_trilearnが強すぎ、DFとGKの2者がいてさえも、シュートを入ることがあり、最適な解にQ値が収束して

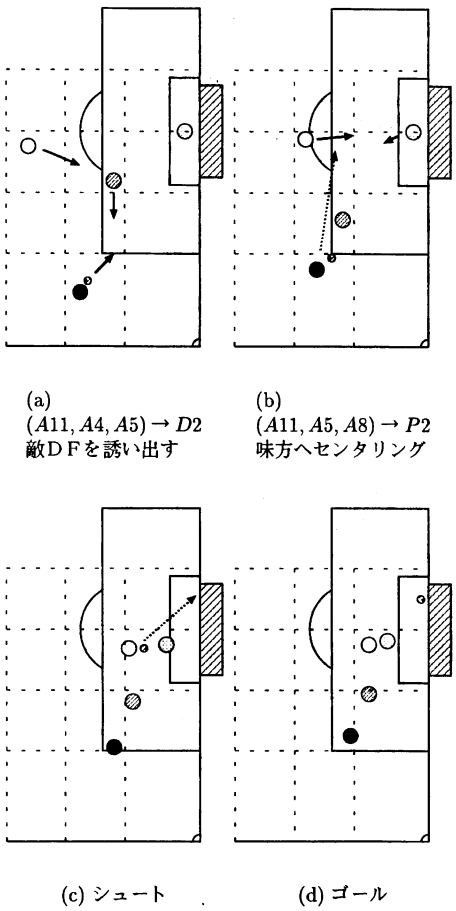


図 6: 獲得した協調行動

いない可能性もあると考えられる。

また、実験1，2の結果を比較すると、改善はみられたが小さいものであった。しかし、縮約した状態によって30回の試行でもある程度学習することが示され、問題に適応して分割を変更することより、より良い学習ができると考えられる。

5 おわりに

本研究により、Q-learningによりマルチエージェント間での、センタリングにおける戦術的な協調行動を獲得した。また、状態縮約を用いることにより、短期間である程度学習できることを示した。一方、段階的に細分化することで、学習の効果を上げることは、今回の実験では明らかではない。今後は、初期における状態縮約の方法、段階的な分割を行うための指針について検討が必要である。

参考文献

[1] 内部英治、浅田稔、細田耕，
複数の学習するロボットの存在する環境における協調行動獲得のための状態空間の構成，日本ロボット学会誌 Vol.20 No.3,pp.281-289，2002

[2] 浅田稔，ロボットの行動学習・発達・進化，共立出版，2002

[3] Jelle Kok，UvA_trilearn2003
http://carol.wins.uva.nl/ jellekok/
robocup/index_en.html

[4] Richard S.Sutton、Andrew G.Barto
強化学習，森北出版，2000