高次元入力空間におけるマルチエージェント強化学習
Multi-agent reinforcement learning in high-dimensional input space

紫尾太朗　水間大資　田中智紘　渕田孝康
（鹿児島大学大学院　理工学研究科）

1　研究背景・目的

あらかじめプログラムされた行動だけを行う自律ロボットでは，活動中の不測の事態に対応できないという問題がある．例えば，予期せぬ地震などの際に道がふさがれた場合には，人間の予測から逸脱しているためにその後の動作は不可能になってしまう．このため，自律ロボットが環境の変化を認識して自ら学習を行う機械学習に期待が高まっている．本研究では，動的な連続状態空間についてQ学習と円による状態空間の分割法を用い，単一エージェントではなく複数のエージェントが目的を達成しようと学習を行うことにより，多次元での複数エージェントの学習を可能にすることを目的としている．

2　Q学習について

Q学習についての学習の流れを以下に示す．

(1)　他のエージェントへの距離と角度を観測し，状態を離散化する．
(2)　全ての状態とその時に取り得る行動 s, a の組について，初期のQ値をランダムに決める．
(3)　離散化された状態を最初の状態 s にセットする．
(4)　状態 s から ε −greedy 法で行動 a を選択し，①の式に基づきQ値を更新し，状態は s' に移行する．

$$Q_{s,a} = \alpha(r + \gamma \mathrm{MAX}_i Q_{s',i} - Q_{s,a}) \cdots ①$$

3　状態空間の分割

状態の分割には距離と角度を用いて円による分割を行う．本研究では CorrectVoronoi 分割と PseudoVoronoi 分割の 2 種類の分割手法を用いて実験を行う．

3．1　CorrectVoronoi 分割

ボロノイ図を利用して状態空間を分割していく手法をCorrectVoronoi 分割と呼ぶ．この手法では，円によるボロノイ図を作り出している．この円を Q 空間と呼ぶ．図1に2次元での簡単な例を示す．

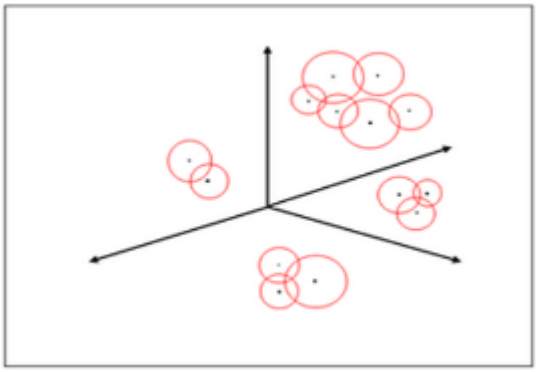

図1　CorrectVoronoi 分割による状態空間の分割

3．2　PseudoVoronoi 分割

CorrectVoronoi 分割の問題点として空間の分割数が増加するたびに実行時間が遅くなってしまう問題が挙げられる．そこで，階層構造を導入した空間分割を用いる．これを PseudoVoronoi 分割と呼ぶ．この分割は、正しいボロノイ分割を与えるわけではないが、高次の空間においては母点を高速に探索可能となることが期待できる。

初めの状態観測の前に次元数から半径を決め，次元を囲むような半径の円をあらかじめ追加しておく．その後，あらかじめ作られた円の内部に距離と角度から入力が入ると その円の半径を定数倍に収縮した半径の Q 空間を追加していく．Q 空間内で出来る状態の分布を円により示している．分割の例を図2に示す．

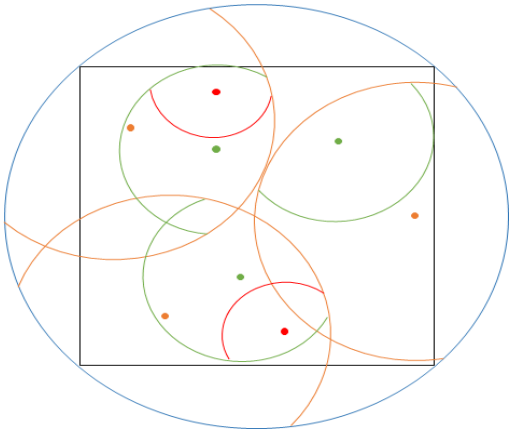

図2　PseudoVoronoi 分割による状態空間の分割

4 実験環境モデル

本研究では，複数のエージェントでの学習を行うためにハーフフィールドサッカーをモデルとしたミニゲームで，複数のエージェントが協調あるいは競合しながら同時に学習を行うモデルを提案する．この際，学習に対しては恣意的な条件は一切使わないことを制限としている．

このモデルは，攻撃者のエージェントを複数体用いて学習を行い，攻撃者はゴールとアシストを，守備者はボールに近づく．学習には Q 学習を使用．図3では，青色の守備者が1人，赤色の攻撃者が3人の例を示しており，白色がボール，赤色がゴールを示している．

現在学習には，一般的な学習 Q 学習を用いて行っている．攻撃者3人で行い，エージェントの行動は，シュート，パス，ボールに向かう 3 種類．報酬は，攻撃者はゴールすること，またはアシストすることができれば報酬1を与える．また，攻撃者についてはボールを奪われると報酬－1を与える．

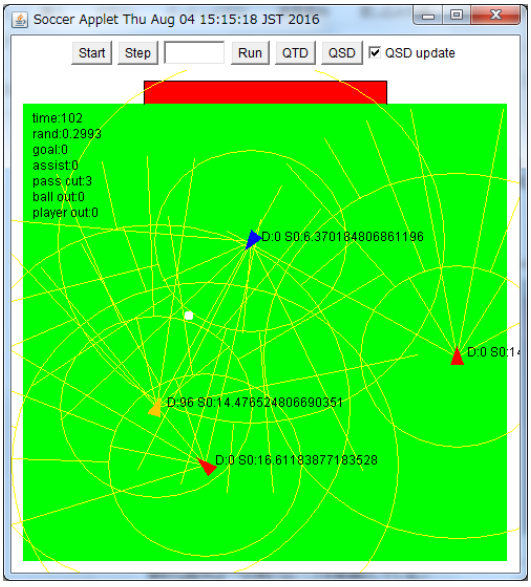

図3　ハーフコートサッカー環境(1vs3)

5 実験結果

今回，CorrectVoronoi 分割法の追加する円の半径と PseudoVoronoi 分割法の追加される円の収縮率を変化させ，各半径と各収縮率での実験とエージェント数を増やし，報酬数を見る実験を多数行った．その結果を基に，CorrectVoronoi 分割法で最適と思われる半径は 0.3 であり，PseudoVoronoi 分割法で最適と思われる収縮率は 0.2 であった．2種類の分割手法の比較を示したグラフを図4に示す．

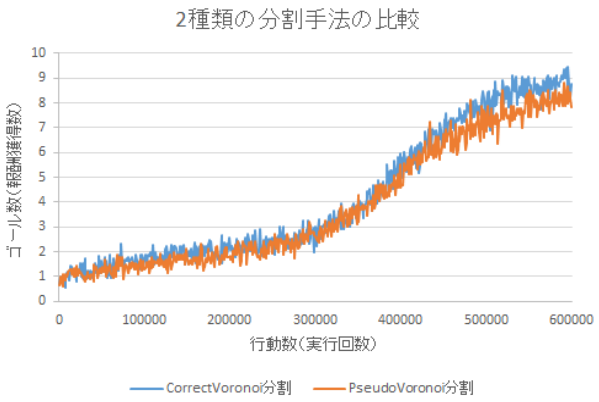

図4　2 種類の分割手法の比較

6 まとめ

これまで，距離と角度を用いて円によるボロノイ分割により多次元空間の分割を行っていたが，本研究では，円による擬似ボロノイ分割からマルチエージェント環境への適応を試みた．その結果として，報酬数はほとんど差異なく，実行時間は CorrectVoronoi 分割に比べ，PseudoVoronoi 分割のほうが速くなった。
CorrectVoronoi 分割の方は線形探索を行っているため実行時間が遅くなってしまうが，PseudoVoronoi 分割のほうは二分木探索を行っているため実行時間を速くすることができた．

現在，状態数の多さが原因のため少ないエージェント数での実験しか行うことができない．エージェントを増加させて実験するために，状態数をさらに削減しなければならない．また，攻撃者だけが学習を行うのではなく，守備者にも学習を適応させることが今後の課題であると考えている．

7 参考文献

[1] 田中昭雄，中田洋平，松本隆:“動的環境下の強化学習アルゴリズム:Sequential Monte Carlo とサンプル初期化”電子情報通信学会技術研究報告 . NC，ニューロコンピューティ ン グ，Vol. 104，No. 759，pp. 101-106(2005)

[2] 一井宏次，釜屋博行，阿部健一:“高次元連続状態空間における局所重み付き回帰手法を用いた強化学習:動的環境下での移動ロボットの障害物回避”自動制御連合講演会講演論文集，Vol. 52，pp. 218-218(2009)

[3] 伊賀上大輔，市村匠:“階層型モジュラー強化学習による動的環境に適した学習手法を用いる児童見守りアプリケーション”ファジィシステムシンポジウム講演論文，Vol. 28，pp. 69-74(2012)