連続状態入力のQ学習における入力次元削減の検討

紫尾太朗[†1]　水間大資[†2]　田中智紘[†3]　渕田孝康[†4]

概要：近年，ロボット技術の発展により自らが学習を行う機械学習への期待が高まってきており，マルチエージェントを用いた研究が行われている．本研究では，連続状態入力における状態入力空間の次元数の削減を行う．環境のモデルはハーフフィールドサッカーを用いる．攻撃者と守備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，守備者はパスカットを目的として行動する．学習にはQ学習を使用し，状態空間は他のエージェント，ボール，ゴールを対象として，それらへの距離と角度から高次元入力空間を構成し実験を行った．

キーワード：マルチエージェント，強化学習，機械学習，ニューラルネット

Consideration on input dimension reduction in Q learning of continuous state input

TARO SHIBI[†1]　TAISHI MIZUMA[†2]
TOMOHIRO TANAKA[†3]　TAKAYASU FUCHIDA[†4]

Abstract: In recent years, with the development of robot technology, the expectation for machine learning which learns by itself is increasing, and research using multi agents is carried out. In this research, we assume an environment where multiple autonomous agents cooperate and learn by using high dimensional input space, and use half field soccer for environment model. Learning is conducted using multiple agents of attackers and defenders, attackers act as goals and assistants, and defenders act as passcuts. Learning uses Q learning, and the state space is composed of high-dimensional input space from distances and angles to other agents, balls and goals, and experiments were performed.

Keywords: Multi agent, Reinforcement learning, Machine learning, Neural net

1. 研究背景・目的

近年の機械学習において，人の手によって制御された行動だけを行うロボットでは不測の事態に対応できないといった問題点がある．例えば，自然災害によって道をふさがれた場合には，人間の予測の範囲外のことであるため，人間によってあらかじめプログラム制御されたものでは，その後の行動が不可能になってしまうことも多い.そのため，自律ロボットがその場の環境の変化に対応し，自ら学習を行っていく機械学習への期待が高まってきている．学習を行う初期段階では，何もわからずに動作している状態である.そのため，次元数や入力が多数存在したり，環境が様々に変化したりする設計などでは，難しい学習となってしまう．

本研究では，動的な連続状態空間について，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．そこで，連続状態入力における状態入力空間の次元数の削減を行う．観測される次元数はAutoEncoderを用いて次元削減を行う．次元数を削減することができれば，エージェント数の増加や実行時間の短縮が行えると考えた．自律的な状態空間分割手法を用いて観測した次元数の削減を行い，高次元かつ動的環境の中でどこまで有効であるか検証する

2. 外部状況

高次元連続状態空間において関数近似を用いた移動ロボットの障害物回避[1]，実問題への適用，複数のエージェントによる追跡問題[2]など多岐にわたる．これらの 研究の多くは，環境依存，次元の呪い，現実環境 への適応，不完全知覚，報酬の分配，高次元空間 での学習時間の問題などがある．

3. Q学習（Q-Learning）

ある環境の状態sの下で，行動aを選択する行動価値$Q(s,a)$を学習する方法となる．この行動の価値を示すものQ値と呼ぶ． Q値を用いて，状態空間を軸ごとに細かく区切って表の形にしたQ-tableに保持させる．エージェントは観測された状態においてQ-tableを参照し，最も価値の高い行動a_tをとる．そして，行動a_tをとって状態S_tからS_{t+1}に遷移した場合，環境から返された報酬rと次の状態S_{t+1}をもとに下記の学習式でQ値の更新を行う．

$$Q(s_{t+1},a_{t+1}) \leftarrow Q(s_t,a_t)+\alpha(r_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t))$$

Q 学習についての学習の流れを以下に示す.

1. 状態を離散化
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
4. 状態 s から ε-greedy 法で行動 a を選択し，上記の式に基づき Q 値を更新し，状態は s' に移行
5. 1~4 を繰り返し

4. 実験環境モデルと提案手法

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をモデルとする．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い，連続状態入力における状態入力空間の次元数を AutoEncoder を用いて削減する手法を提案する．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 5 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．

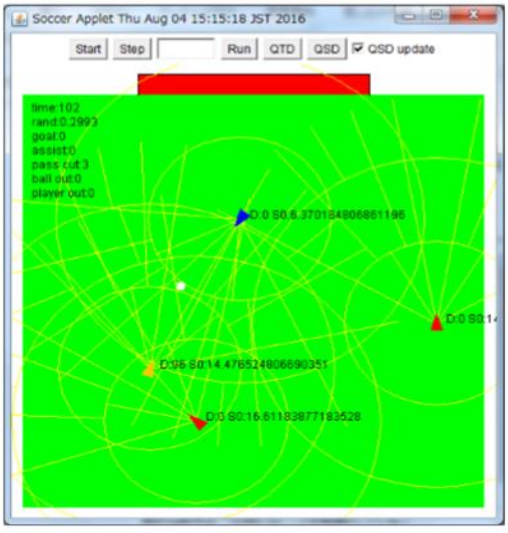

図 5　実験モデル（ハーフコートサッカー3vs1）

4.1 AutoEncoder（自己符号化器）

AutoEncoder は，DeepLearning の 1 つであり 3 層ニューラルネットにおいて，入力層と出力層に同じデータを用いて教師あり学習させたものである．入力ベクトルとして X，教師ベクトルにも X を与えて，誤差関数を最小化することで学習を行う．AutoEncoder では，教師ベクトル=入力ベクトルとして入力ベクトルを与えると出力として自分自身を復元している．3 層の AutoEncoder を図 1 に示す．しかし，目的としているのは入力を復元することではなく中間層の出力である．中間層で得られた Z の値を取り出すことで 4 次元のデータを 3 次元のデータで表すことができ，次元削減が行える．

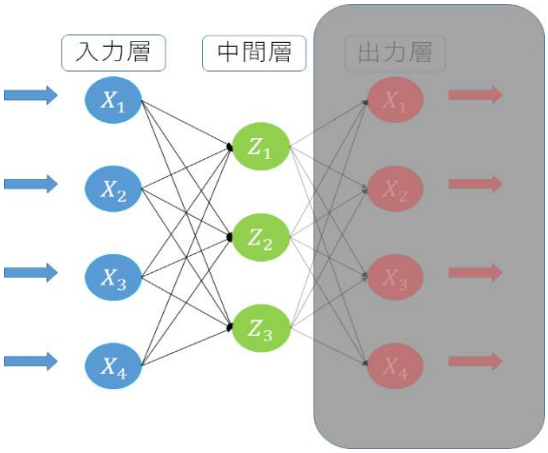

図 1　3 層の AutoEncoder

4.2 状態空間の分割

エージェントは各エージェント，ボール，ゴールについて観測を行っている．観測対象は，各エージェント，ボール，ゴールへの距離と角度である．人数によって各エージェントにおける状態ベクトルの次元数は異なる．図 2 に観測の例を示す．

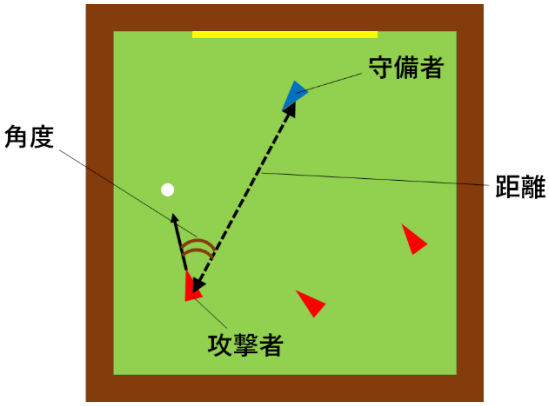

図 2　状態観測の例

4.3 状態空間の離散化

本研究では、距離と角度を入力とし，階層構造を用いた円によるボロノイ領域の生成を木構造により行う自律的

に離散化する分割手法を扱う．

4.4 PseudoVoronoi 分割

階層構造を導入したボロノイ図を利用した分割手法のことを PseudoVoronoi 分割と呼ぶ．この手法では、初めの状態観測の前に次元数から半径を決め，その半径の円をあらかじめ追加しておく．この円のことを Top と呼ぶ．その後，あらかじめ作られた円の内部に距離と角度から観測点が入力されると Top の半径を数倍に収縮した半径の Q 空間を追加していく．この場合 Q 空間は、Top かすでに生成されている Q 空間の内部に必ずできることとなる．つまり，Q 空間領域内に新たな Q 空間領域が追加されることとなることから木構造の関係性を作り出すこととなる．この木構造を作り出すことで状態の参照効率を上げることができる．図 3 に 2 次元での PseudoVoronoi 分割の例を示す．

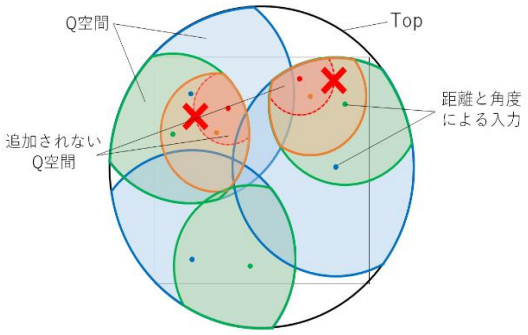

図 3　PseudoVoronoi 分割の例

5. 実験

5.1 実験 1

まず実験 1 として観測された次元数を AutoEncoder を用いて圧縮を行う．次元圧縮の際にニューロン同士の重みが必要となる．その重みを取り出すための実験を行う．AutoEncoder の学習データにはサッカーエージェントがそれぞれ距離と角度について観測した状態を用いる．この実験には 7 層構造の AutoEncoder を用いて学習を行う．120000 個の学習データをそれぞれ 30 回学習させる．使用する AutoEncoder は｛10-19-14-*X*-14-19-10｝で構成し観測された 10 次元のデータを 1 次元ずつ圧縮する．各数字はそれぞれの層のニューロンの数を表しており，*X* の部分には 9~5 のニューロンを与える．教師信号と学習した出力の結果の許容誤差は±5%以内とする．7 層構造 Autoencoder を図 4 に示す．

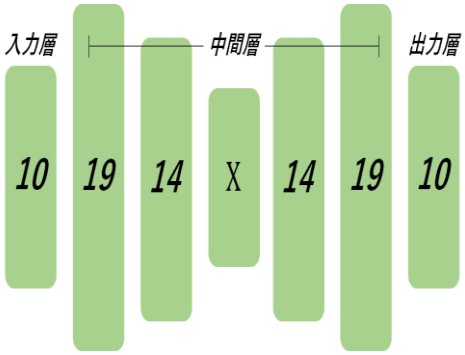

図 4　7 層構造 AutoEncoder

5.2 実験 1 結果

サッカーエージェントでそれぞれ距離と角度を観測した 10 次元データを 9 次元~5 次元の範囲で圧縮を行った．その結果を以下に示す．

9 次元　→　成功：97%　　失敗：3%
8 次元　→　成功：90%　　失敗：10%
7 次元　→　成功：85%　　失敗：15%
6 次元　→　成功：77%　　失敗：23%
5 次元　→　成功：67%　　失敗：33%

教師信号と学習した出力の結果の誤差が±5%のものを成功，誤差が±5%以上のものを失敗としている．

5.3 実験 2

実験 1 で得られた次元圧縮に必要な重みを用いて，実際に観測されたデータを削減して学習させる．各次元数の報酬獲得数と Q 空間の探索時間を測る実験を行う．学習を行う流れとして，まず状態の観測をする．その後行動を決定する前に観測された状態を削減し，その値を用いて状態空間の分割を行い，行動を決定し学習させていく．次元削減の際には，実験 1 で用いた AutoEncoder の左半分をサッカーエージェントに組み込み実験を行っている．実験 2 の学習の流れを図 6 に示す．

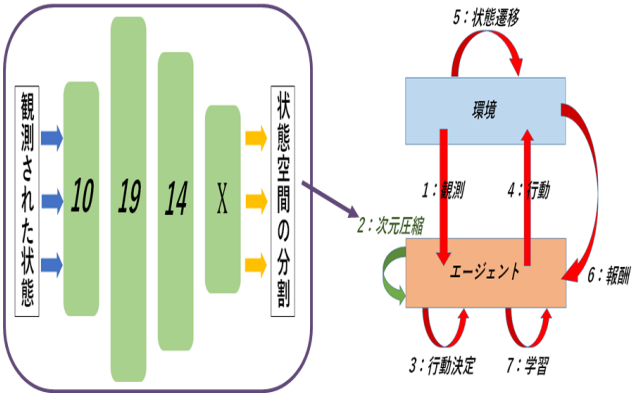

図 6　学習の流れ

エージェントは攻撃者 3 人と守備者 1 人で行い，学習には Q 学習を使用する．PseudoVoronoi 分割で状態空間の分割を行う．行動選択手法は ε -greedy 法を使用し，ランダム行動

率を 0.3，学習率は 0.1，割引率は 0.9 とする．ランダム行動率は行動回数が増えるごとにシクモイド関数で減衰させる．PseudoVoronoi 分割の半径収縮率は 0.2，追加される最小半径は 0.3 とする．600000 回行動で乱数種を 1~20 まで変化させ，1000 回ごとのデータの平均を取る．

5.4 実験 2 結果

各次元の実験結果を以下のグラフに示す．

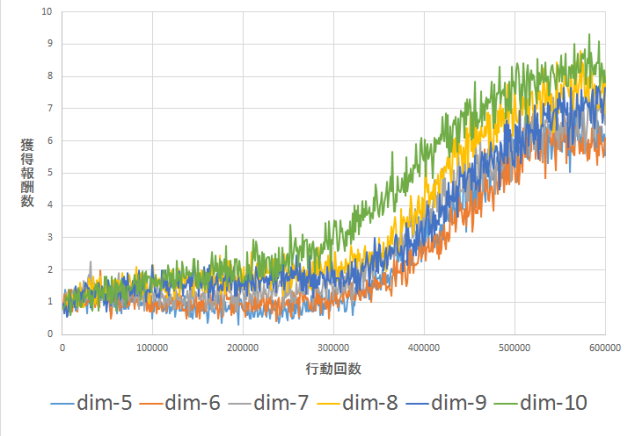

図 7　各次元の獲得報酬数の比較

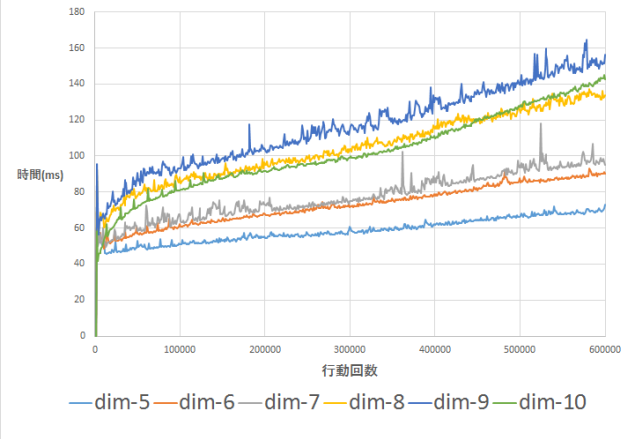

図 8　各次元の実行時間の比較

5.5 考察

次元削減後は次元削減前に比べ，獲得報酬数（ゴール数）はやや劣るが行動回数が増えるごとにゴール数も増えている．次元を削減しても学習が行えていると言える．やや劣る原因として 2 つ考えられる．1 つ目は，AutoEncoder で学習させた際に生じた誤差が影響していると考えられる． 2 つ目は，AutoEncoder で学習させるデータ数が少ないと考える．実行時間は次元削減前と比較して全体的に向上している．しかし，9 次元と 8 次元に圧縮したときには向上していない．これは次元圧縮する際に時間がかかってしまうためであると考えられる

6. まとめと今後の課題

今回，{10-19-14-*X*-14-19-10} の 7 層構造の AutoEncoder を用いて観測される次元数を削減する実験をし，獲得報酬数と実行時間を見る実験を行った．実行時間は次元数を削減したことによって向上させることができた．しかし獲得報酬数が次元削減前より劣ってしまったのは，AutoEncoder の学習による誤差と学習するデータ数の少なさが原因考えている．

今後の課題として状態空間の分割に関しては現在 Q 空間の追加のみを行っている．そのため，Q 空間が常に増加していく問題が生じる．学習をしていく中で使わなくなった無駄な Q 空間あると考え，必要に応じて削除していかなければならない．また， Q 空間を移動させることで Q 空間の最適化を行うことも必要であると考える．次元削減については，現状では AutoEncoder の学習成功率は次元を減らせば減らすほど下がってしまう．そのため AutoEncoder の改良が必要だと考える．AutoEncoder の精度が向上すればもっと多くのエージェントを用いた学習でも適応でき，エージェントの増加で学習精度が悪くなっていく問題の解決にもなると考える．また，攻撃者だけの学習ではなく，守備者も含めた学習にも対応させていくのも課題の 1 つである．

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] 松尾豊“Deep Learning” 情報・システムソサイエティ誌 Vol.19，No.1 pp.12-13(2014)