RoboCup 2Dサッカーシミュレーションにおける機械学習を用いた自律エージェントの行動決定

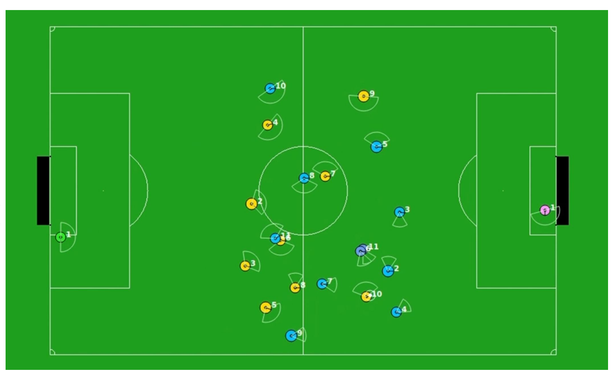
レネ研究室 | 2220014 | 小原　優陽

背景と目的

RoboCup 2Dサッカーシミュレーションは，マルチエージェントAI研究の標準プラットフォームとして活用されてきた．しかし，現在の強豪チームが採用するルールベースシステムは，複雑な戦略の実現に伴うルール設計・管理の困難さや，人間が設計するがゆえの行動の最適性の限界という課題を抱えている．

この課題に対し，近年成功を収めている深層学習を用いたデータ駆動型アプローチが注目される．これは，トップチームの試合ログデータからエージェントが自律的に最適な行動を学習する手法である．本研究は，学習した深層学習モデルをリアルタイムシミュレーション環境へ統合し，お手本を単に再現するだけでなく，さらに学習を通じて強化されるエージェントの開発を目指す．

具体的には，強豪チームのログデータを用いた模倣学習で基本行動モデルを構築し，それを自己対戦学習で強化することで，ルールベースシステムに依存しない高性能な自律エージェントを開発し，その有効性を実証することを目的とする．



研究方法

本研究は，RoboCup 2Dサッカー環境を対象とし，以下の手順で開発と評価を進める．まず，開発プロセスの第一段階として，強豪チームの試合ログを教師データに用い，模倣学習によってお手本の行動を忠実に「再現」するベースラインモデルを構築する．次に第二段階として，この再現モデルを初期エージェントとし，自己対戦（Self-Play）を繰り返す強化学習を適用することで，ベースラインの性能を超え，より最適な戦略を獲得した「強化」モデルへと発展させる．

模倣学習（パス予測）の数式表現

本アプローチは、専門家（プロ選手）の行動データを教師あり学習（**行動クローニング**）として扱い、選手の行動を予測する関数（ポリシー）を獲得するものです。

1. 専門家データセット D

N 個の「状態」と「行動」のペアからなるデータセット D を定義します。 S_i: **状態 (State)**、 $A_{E,i}$: **専門家の行動 (Expert Action)**。

$$D = \{(S_i, A_{E,i})\}_{i=1}^{N}$$

2. 状態 S の定義

「状態 S」は、パサー (x_p, y_p) と最も近い味方 (x_{tm1}, y_{tm1}) の座標ベクトルです。

$$S = (x_p, y_p, x_{tm1}, y_{tm1})$$

3. 学習するポリシー π_θ

学習の目的は、状態 S から予測行動 A_P を出力する**ポリシー (Policy)** π_θ（＝モデル）を獲得することです。(θ: パラメータ)

$$A_P = \pi_\theta(S)$$

4. 学習の目的（損失最小化）

学習は、ポリシーの予測 A_P が専門家の行動 A_E と一致するよう、**損失関数** L を最小化します。(θ^*: 最適化されたパラメータ)

$$\theta^* = \arg\min_\theta \sum_{(S_i, A_{E,i}) \in D} L(\pi_\theta(S_i), A_{E,i})$$

5. 性能評価（正解率）

学習済みポリシー π_{θ^*} の性能は、テストデータ D_{test} に対する**正解率 (Accuracy)** で評価します。($\mathbb{I}(\cdot)$: 指示関数)

$$\text{Accuracy} = \frac{1}{|D_{test}|} \sum_{(S, A_E) \in D_{test}} \mathbb{I}(\pi_{\theta^*}(S) = A_E)$$

図1「再現モデル」のフロー

オフライン強化学習 (CQL) の数式表現

本アプローチは、事前に収集したログデータ（**オフラインデータセット** D）のみを用いて、エージェントの行動価値関数 Q と最適方針 π を学習するオフライン強化学習です。

1. オフラインデータセット D

ログから抽出した「状態 S, 行動 A, 報酬 R, 次の状態 S'」の遷移の集合 D を定義します。

$$D = \{(S_i, A_i, R_i, S'_i)\}_{i=1}^{N}$$

2. 状態 S_t と 報酬 R_t の定義

状態 S_t: ボール、自分、味方11人、敵11人の座標（計48次元）ベクトル。 **報酬** R_t:「ボールの前進距離」と「ゴールボーナス」で計算されます。

$$R_t = 5.0 \times (x_{ball}^{(t+1)} - x_{ball}^{(t)}) + R_{goal} \quad (\text{where } R_{goal} = 1000 \text{ if } x_{ball}^{(t+1)} > 52.5)$$

3. 学習目的 (Discrete CQL)

CQL (Conservative Q-Learning) は、通常のQ学習損失 $L_Q(\theta)$ に加え、データセット D に存在しない行動の価値を低く見積もる**保守項** $L_{CQL}(\theta)$ を導入します。

$$\theta^* = \arg\min_\theta \left(\mathbb{E}_{(S,A,R,S') \sim D}[L_Q(\theta)] + \alpha \cdot L_{CQL}(\theta) \right)$$

4. 保守項 $L_{CQL}(\theta)$ (CQL誤差)

保守項は、全行動のQ値の期待値（第1項）を最小化しつつ、データセット内の行動のQ値（第2項）は保持するように働きます。

$$L_{CQL}(\theta) = \mathbb{E}_{S \sim D}\left[\log \sum_{A'} \exp(Q_\theta(S, A')) \right] - \mathbb{E}_{(S,A) \sim D}[Q_\theta(S, A)]$$

図2「強化モデル」のフロー

結果と考察

再現モデル

```
Accuracy: 0.12435476302205538

Classification Report:
                                precision    recall  f1-score   support

                  Aarón Martín Caricol       0.00      0.00      0.00         1
                  Adrián López Álvarez       0.00      0.00      0.00         1
                   Aitor Ruibal García       0.00      0.00      0.00         2
                  Alberto Soro Álvarez       0.00      0.00      0.00         1
             Alejandro Moreno Lopera         0.00      0.00      0.00         1
                   Allan Romeo Nyom          0.00      0.00      0.00         2
                   Anssumane Fati            0.00      0.00      0.00        31
                 Antoine Griezmann           0.14      0.01      0.02       113
               Antonio García Aranda         0.00      0.00      0.00         3
          Aridane Hernández Umpiérrez        0.00      0.00      0.00         1
               Borja Iglesias Quintas        0.00      0.00      0.00         2
                Brais Méndez Portela         0.00      0.00      0.00         6
               Carlos Clerc Martínez         0.00      0.00      0.00         1
            Carlos Domínguez Cáceres         0.00      0.00      0.00         3
              Carlos Fernández Luna          0.00      0.00      0.00         1
            Carlos Henrique Casimiro         0.00      0.00      0.00        13
              Carlos Soler Barragán          0.00      0.00      0.00         4
                  Clément Lenglet            0.12      0.22      0.15        86
              Cristian Tello Herrera         0.00      0.00      0.00         1
          Damián Nicolás Suárez Suárez       0.00      0.00      0.00         1
...
                          accuracy                               0.12      2131
                         macro avg           0.01      0.01      0.01      2131
                      weighted avg           0.07      0.12      0.07      2131
```

図3モデルの評価指標

1，研究結果：予測精度
- 決定木モデルによるパス予測精度：12.4%
- ランダム予測（バースライン）の精度：0.4%

2，課題と考察

モデルの予測が一部の選手に偏り、多様な選手を予測できていないという課題は、元データの量的な偏り（不均衡）を反映し、モデルが全体の精度を最大化しようと学習した結果であると考察されます。

強化モデル

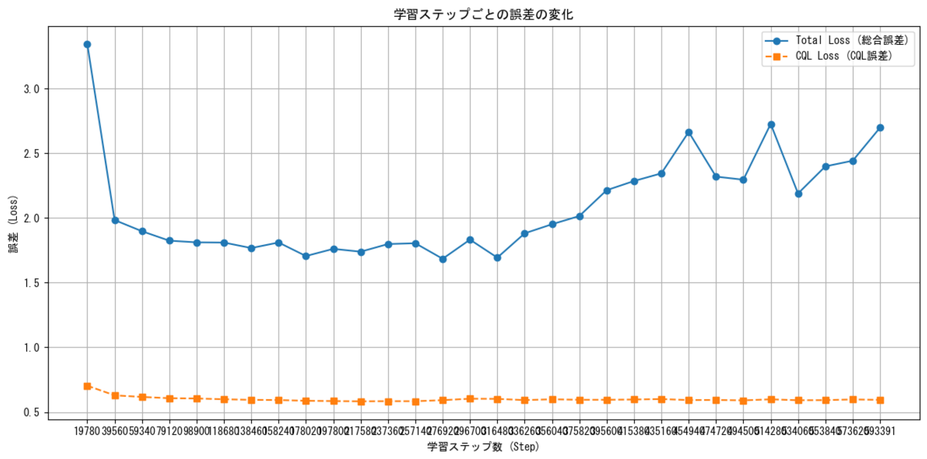


図4学習ステップごとの誤差の変化

1. 研究結果：最終学習誤差（59.3万ステップ時点）
   DiscreteCQLモデルの最終CQL誤差（保守性の指標）: 0.594
   DiscreteCQLモデルの最終総合誤差（模倣の指標）: 2.698
2. 課題と考察

モデルが学習したCQL誤差（オレンジ破線，保守性の指標）が、学習初期から0.6前後で安定して収束している という結果は、データセットの極端な行動の偏り（dash/turnが99%以上） を反映し、モデルが未知の行動を過大評価しないよう保守的な学習に成功した結果であると考察されます。

（一方で、総合誤差（青線）が後半に上昇傾向にあるのは、この保守的な制約の中で、モデルがログデータ（お手本）とは異なる新しい価値観（Q値）を学習しようと試みているためと考えられます。）

終わりに

本研究の中間成果として，ログデータの極端な行動の偏りに対して，DiscreteCQLを用いて保守的な模倣エージェントの構築に成功した．

卒論に向けて，パス予測モデルは特徴量追加や，試合数などのデータ量を増やし精度向上を，CQRエージェントは実際に試合を行ってみて有効的な結果になっているのか，また精度向上をそれぞれ目指し研究を進めます．