シミュレータを用いた SCCF の性能調査　実験設定

1　実験概要

2019 年度実施した SCCF の特性調査実験では、カテゴリ数が状態行動数の削減のために 6 個であることや、2 回前までの発話内容を格納するなど、状態行動数が少ない状態での学習ができることを確認した。

本実験では、状態行動数を増やし、深層強化学習アルゴリズムを SCCF に適用することで学習が可能であるかについて、性能を調査する。

ただし、実験を簡単にすることで、サイクルを早くするため、IDA は実際の利用者ではなく、エージェントを評価するシミュレータを対象として実験を実施する。

〇対象アルゴリズム

- REINFORCE
- Deep-Q-Network (DQN)
- Double DQN
- Dueling Network
- PPO(Proximal Policy Optimization Algorithms) など
- その他、時間があれば以下の環境で色々試したいです。

2　エージェントのパラメータ設定

エージェントのもつパラメータは以下の通り。

- 状態 $s \in \mathbf{S}$: 過去 T 回までの発話したカテゴリ
- 行動 $a \in \mathbf{A}$: 次に発話するカテゴリ
- エージェントが発話するカテゴリ: $C = |\mathbf{C}|, \mathbf{C}$: エージェントの発話するカテゴリの集合
- エージェントが持つ、過去の発話履歴の長さ: T (T ステップ前までの発話を保存)

以上より、エージェントのもつ状態行動の総数は,

$$|\mathbf{S}| \cdot |\mathbf{A}| = C^T \cdot C = C^{T+1}$$

3 実験設定

実験では、エージェントのパラメータを次のように変化させ、シミュレータを対象として学習の性能を確認する。

- $C = 6,\ T = 2,\ |\mathbf{S}| \cdot |\mathbf{A}| = 6^3 = 216$
- $C = 10,\ T = 2,\ |\mathbf{S}| \cdot |\mathbf{A}| = 10^3 = 1,000$
- $C = 15,\ T = 3,\ |\mathbf{S}| \cdot |\mathbf{A}| = 15^4 = 50,625$
- $C = 20,\ T = 3,\ |\mathbf{S}| \cdot |\mathbf{A}| = 20^4 = 160,000$
- $C = 50,\ T = 3,\ |\mathbf{S}| \cdot |\mathbf{A}| = 50^4 = 6,250,000$
- $C = 75,\ T = 3,\ |\mathbf{S}| \cdot |\mathbf{A}| = 75^4 = 31,640,625$
- $C = 50,\ T = 4,\ |\mathbf{S}| \cdot |\mathbf{A}| = 50^5 = 312,500,000$
- $C = 100,\ T = 3,\ |\mathbf{S}| \cdot |\mathbf{A}| = 100^4 = 100,000,000$

参考：エージェントの学習アルゴリズムに DQN などを使用する場合の NN の入力と出力

- 入力: 状態 + 行動　（入力ノード数： $T + 1$）
- 出力: 行動価値

よって、T の大きさは入力ノードの数と直結する一方、C の大きさは、入力ノード数とは関係ない。

4 シミュレータの設定

シミュレータは、エージェントの振舞を評価する。

4.1 シミュレータの与える報酬

2019 年度の REINFORCE の実験で性能がよかった実験設定（報酬設定 B）を採用

$$r \in \{1, 0, -1\}$$

4.2 シミュレータがエージェントに与える報酬について

シミュレータがカテゴリ番号 c の内容を、t 回連続して発話した条件の元、エージェントに与える報酬が r である確率を $P(r|c,t)$ と定義する。

ただし、

$$R_1 = P(r = 1|c = c^*, t = t^*)$$
$$R_0 = P(r = 0|c = c^*, t = t^*)$$
$$R_{-1} = P(r = -1|c = c^*, t = t^*)$$

とおいたうえで、$R_1 + R_0 + R_{-1} \neq 1$ の時、ルーレット選択を行う。

すなわち

$$R_r = \frac{R_r}{R_1 + R_0 + R_{-1}} (r \in \{1, 0, -1\})$$

4.3 報酬付与確率の具体例

報酬付与確率の具体例を下記に示す。

```
P_ex(r|c,t) =
[ (r = 1)
    // 配列に示された値は固定値である。
    // 確定値の配列番号は、同じ発話を繰り返した回数を示す。
    ([t=1の時r=1を与える確率, t=2の時, t=3, t=4])
    [0.52, 0.91, 0.92, 0.9] (α = 0.98) (カテゴリ番号c:1)
    [0.4 , 0.44] (α = 0.3) (カテゴリ番号c:2)
    [0.48, 0.13] (α = 0.27) (カテゴリ番号c:3)
    [0.35, 0.5] (α = 0.3) (カテゴリ番号c:4)
    [0.45] (α = 0.3) (カテゴリ番号c:5)
    [0.38] (α = 0.3) (カテゴリ番号c:6)
],
```

ただし、ここで α は減衰係数であり、固定値で定められない範囲の報酬確率を決める定数である。

ex.:

$$P_{ex}(r=1|c=3,n) = [0.48, 0.13, 0.13 \cdot 0.27, 0.13 \cdot 0.27^2, 0.13 \cdot 0.27^3 ...]$$

減衰係数の決め方:

- 減衰係数がかかるまでに 2 つデータがある箇所: 直前の勾配を利用

ex.

$$P_x(r=1|c=1,t) = [0.4.0.45, 0.1, ...] \quad \text{この時 } \alpha = 0.1/0.45 = 0.22$$

- ただし、直前の勾配が 1 より大きい時: $\alpha = 0.3$
- 直前の勾配が 1 の時: $\alpha = 0.7$
- 直前のデータがない箇所: $\alpha = 0.3$

5 実験に利用するシミュレータのパラメータ

以上の設定の元、2019 年度の SCCF 特性調査実験において、特徴的だった実験協力者 4 名のデータを用いてパラメータ（報酬付与確率）を用意した。下記に示す報酬付与確率の設定は、$C = 6$ の時のみ適用できる。

$C = 20$ など、カテゴリ数を増加させる場合のカテゴリごとの報酬付与確率は、6 つパラメータと全く同じものをランダムにコピーして、合計 20 個になるように指定する。

具体例: user1

- 1 位から 6 位まで順に発話確率が高かった。（2019 年度の実験結果）
- 報酬の与え方の特徴: 報酬 1or0 を与える確率が高かった。

```
P(r|c,t) =
[ (r = 1)
    // [確定値]（その後の値: 減衰係数をかけていく）
    // 確定値の配列番号は、同じ発話を繰り返した回数を示す。
    [0.52, 0.91, 0.92, 0.9]（α = 0.98）（カテゴリ番号:1）
    [0.4 , 0.44]（α = 0.3）（カテゴリ番号:2）
    [0.48, 0.13]（α = 0.27）（カテゴリ番号:3）
    [0.35, 0.5]（α = 0.3）（カテゴリ番号:4）
    [0.45]（α = 0.3）（カテゴリ番号:5）
    [0.38]（α = 0.3）（カテゴリ番号:6）
],
[ (r = 0)
    [0.45, 0.09, 0.08]（α = 0.89）
    [0.55, 0.33]（α = 0.3）
    [0.5, 0.88]（α = 0.3）
    [0.55, 0.5]（α = 0.91）
    [0.4, 1, 1]（α = 0.7）
    [0.5, 1, 1]（α = 0.7）
],
[ (r = -1)
    [0.01]（α = 0.3）
    [0.05, 0.2]（α = 0.7）
    [0.02]（α = 0.3）
    [0.1]（α = 0.3）
    [0.15]（α = 0.3）
    [0.1]（α = 0.3）
]
```