反実仮想機械学習におけるオフポリシー評価の一考察

杉村 真理子　(指導教員：小林 一郎)

1　はじめに

反実仮想機械学習 (CFML) とは，観測され得たが実際には観測されなかった（反実仮想）データを捉えるための機械学習技術である．オフポリシー評価 (Off-Policy Evaluation:OPE) は CFML の研究領域の 1 つであり，過去に別のポリシーによって集められたログデータを用いて仮想的なポリシーを評価することを目的とする．これによって，リスクやコストを伴うオンライン実験を行うことなく，新たなポリシーの評価やより良いポリシーの獲得が可能となる．これまで様々な OPE 手法が提案されてきたが，適用するドメインの環境設定（以下，単に「環境」）によって各 OPE 手法がポリシーを評価する精度は異なるため，1 つの実験環境だけでその手法の性能を図ることはできない．したがって，複数の実験環境を用いて OPE 手法を評価する必要がある．

本研究では，OPE のベンチマークスイートである Caltech OPE Benchmarking Suite (COBS) [1] を用いて，3 種類の OPE 手法について異なる実験環境で性能評価を行い，その結果から手法と環境の相性について考察を行う．

2　OPE 手法

OPE は, 評価対象のポリシー π_e, オンライン実験済みのポリシー π_b, π_b のログデータ D を入力とし, π_e の報酬総和の推定値を出力する関数 $\hat{V}(\pi_e, \pi_b, D)$ を定義する．OPE は主に Direct Method，Inverse Propensity Scoring，Hybrid Method の 3 つに分類される．

2.1　Direct Method (DM)

Direct Method (DM) は，ログデータを元に未観測データを直接推定するモデルを作成し，その推定値を基に π_e の報酬期待値を推定する手法である．以下は，DM の中で最も一般的な手法である Approximate Model (AM) である．

$$\hat{V}_{AM} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\hat{r}(a, x_i)\pi_e(a|x_i) \tag{1}$$

ここで，A は行動の集合，n は D のサイズ，$\hat{r}(a, x_i)$ は D を元に学習された報酬推定モデルである．$\hat{r}(a, x_i)$ は D に存在するデータの偏り（選択バイアス）の影響を受ける可能性があるが，$\hat{r}$ を適切にモデル化できれば未観測領域を含めた網羅的な評価が可能となる．

2.2　Inverse Propensity Scoring (IPS)

Inverse Propensity Scoring (IPS) は，選択バイアスによる推定精度の低下を解決するための OPE 手法である．π_b と π_e 間のサンプリング比を用いて重みづけすることで，π_b において選択されやすい行動による影響を軽減する．以下は IPS の中で最も基本的な手法である Importance Sampling (IS) である．

$$\hat{V}_{IS} = \frac{1}{n}\sum_{i=1}^{n} w(a_i, x_i) r_i \tag{2}$$

$$w(a_i, x_i) = \frac{\pi_e(a_i|x_i)}{\pi_b(a_i|x_i)} \tag{3}$$

式 (3) は重みの計算を表しており, $\pi_e(a_i|x_i)$ と $\pi_b(a_i|x_i)$ はそれぞれ π_e と π_b が状況 x_i で行動 a_i をサンプリングする確率である．

2.3　Hybrid Method (HM)

DM の推定モデルはログデータに依存するため，ログデータ中の選択バイアスの影響を受けてしまう．一方で IPS は重みづけによって選択バイアスを解消するが，ログデータに含まれる情報に対してのみ評価を行うため過学習を引き起こしやすい．Hybrid Method (HM) [2] は，DM と IPS を組み合わせることでこれらの問題に対応する手法である．最も基本的な HM 手法である Doubly Robust (DR) を以下に示す．

$$\hat{V}_{DR} = \hat{V}_{DM} + \frac{1}{n}\sum_{i=1}^{n} w(a_i, x_i)(r_i - \hat{r}(a_i, x_i)) \tag{4}$$

ここで右辺の第 2 項は第 1 項の DM において $\hat{r}$ が誤った推定をしたときに IPS と同様の重みづけを行っており，DM におけるバイアスに弱いという弱点を補っている．

2.4　評価方法

OPE 手法の性能を評価するには，OPE で推定した π_e の報酬総和の推定値 $\hat{V}$ と実際に π_e を運用した際の報酬総和の平均二乗誤差 (MSE) を計算する．

$$MSE = (\hat{V}(\pi_e) - V(\pi_e))^2 \tag{5}$$

MSE の値が小さいほど，真値に近いオフポリシー評価が可能な OPE 手法であると言える．今回は MSE に平方根を取った二乗平均平方根誤差 (RMSE) を用いて OPE の性能を測る．

3　実験

3.1　ベンチマーク

本実験では COBS を用いてベンチマークテストを行った．COBS では 8 つの実験環境が用意されており，環境設定を変えた時に OPE の推定精度がどれだけ変化するかを観察することができる．また，OPE 手法として DM が 8 種類，IPS が 4 種類，HM が 28 種類用意されている．

3.2　実験設定

本実験では，OPE 手法として 2 節で説明した AM, IS，DR を使用する．

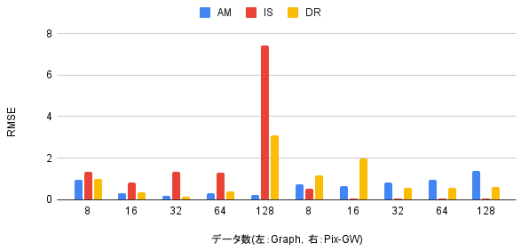

図 1: Graph 環境と Pix-GW 環境の比較

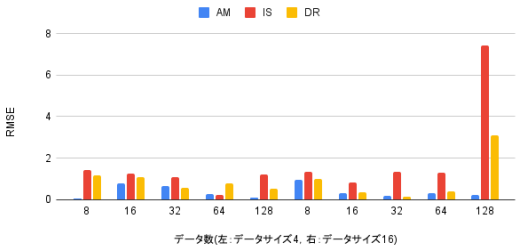

図 2: Graph 環境における T=4 と T=16 の比較

3.2.1 実験 1

実験環境として Graph と Pixel-Gridworld(Pix-GW) を用いて，それぞれデータ数を 8, 16, 32, 64, 128 と変えた時の各 OPE 手法の精度の変化を観察する．各実験設定についてそれぞれ 5 回ずつ実験を行い平均を取った．Graph 環境と Pix-GW 環境の環境要因の違いについては表 1 に示す．また今回 Graph 環境においてデータサイズ T は 16 に設定した．ここでデータは（状況，行動，報酬）の 3 つ組の集合で構成されるエピソードを 1 単位としている．

表 1: Graph 環境と Pix-GW 環境の比較

環境	Graph	Pix-GW
状態記述	位置	ピクセル
データサイズ T	4 or 16	25
報酬	確率的	決定的
報酬リターン	様々	逐次
ポリシーの表現形式	テーブル	NN
初期状態	0	様々
収束状態	2T	ゼロ画像

3.2.2 実験 2

実験環境として Graph を用いて，データサイズ T が 4, 16 それぞれの場合について，データ数を 8, 16, 32, 64, 128 と変えた時の各 OPE 手法の精度の変化を観察する．各実験設定についてそれぞれ 5 回ずつ実験を行い平均を取った．

実験環境として Graph を用いて，データサイズ T が 4, 16 それぞれの場合について，データ数を 8, 16, 32, 64, 128 と変えて，5 回ずつ実験を行い RMSE の平均を取った．

3.3 実験結果

実験 1 の結果を図 1 に，実験 2 の結果を図 2 に示す．横軸はデータ数であり，図 1 の左半分は Graph 環境，右半分は Pix-GW 環境，図 2 の左半分は T=4，右半分は T=16 での結果になっている．縦軸は 2.4 節で示した RMSE であり，この値が小さいほど OPE 手法の性能は高いと言える．

3.4 考察

図 1 では，Graph 環境では AM が良い評価なのに対して Pix-GW 環境では IS が良い結果となっている．これは各環境のポリシーの表現形式の違いによるものと考えられる．Graph 環境では離散的で明確なポリシー表現が可能なテーブル形式であるため，AM が効率的に学習を行えた可能性がある．一方で Pix-GW 環境は連続的で複雑なポリシー表現を可能とするニューラルネットワーク (NN) であり，AM での報酬推定モデルが学習不足になってしまった可能性がある．

また図 2 を見ると，データサイズ T が 16，データ数が 128 とそれぞれ最大の時に IS の精度が飛び抜けて悪くなっている．これは，式 (3) で示した通り IS が $\pi_b(a_i|x_i)$ を分母とする重みを取ることで，$\pi_b(a_i|x_i)$ が非常に小さな値だった場合に重みが極端に大きくなってしまい精度に悪影響を与えてしまったと考えられる．データサイズやデータ数との相関については，データサイズ・データ数を大きくするほどそのような外れ値を引く回数が増えると考えられる．このような過度な重みづけを防止するために，重みの上限を設定する Clipped IPS (CIPS) [3] や，重みを正規化する Self-Normalized IPS (SNIPS) [4] といった IPS 手法も提案されている．

4 まとめと今後の課題

本研究では COBS を用いて，異なる実験環境において基本的な OPE 手法である AM，IS，DR について性能評価実験を行い，各手法の精度がどのように変化するかを観察した．その結果，ポリシー表現がテーブルの場合は AM，NN の場合は IS が良い傾向にあること，また IS はデータの量が増えると精度が悪化する傾向にあることが分かった．

今後は他の実験環境についても各環境要因がもたらす影響をさらに詳しく確認するほか，より発展的な OPE 手法における各手法の特性や環境との相性についても分析を進めていきたい．

参考文献

[1] Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *ArXiv*, Vol. abs/1911.06854, , 2019.

[2] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning, 2016.

[3] Alex Strehl, John Langford, Sham Kakade, and Lihong Li. Learning from logged implicit exploration data, 2010.

[4] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc., 2015.