ISSN 2186-5647

2-31

ベイジアン逆強化学習における複数環境の軌跡を用いた複数報酬の推定

日大生産工（学部）　○五十嵐　充　　日大生産工　山内　ゆかり

1. まえがき

強化学習1)とは，ある環境と目的が与えられた時にその環境における価値を最大化し報酬を得ることを目的とする機械学習の一種である．しかし実世界では，例えば自動車運転においていい運転とは何か，という問題を定義した時，何をもっていい運転とするかという部分で報酬の設計が困難な場合が多い．そこで，2000年頃にAndrew Ng，Stuart J. Russellらによって逆強化学習2)という手法が提案された．この手法は，エキスパートの行動履歴からどういった目的(報酬)を持っているかを推定することで複雑な報酬の設定が容易となり，行動の意図を知ることが出来るというものである．これを発展させたものとしてベイズの定理を加えたベイジアン逆強化学習(BIRL:Bayesian Inverse Reinforcement Learning) 3)という手法が提案されている．ここで中田らは複数の環境で生成した軌跡からエキスパートの報酬分布を推定する問題をBIRLの枠組みを用いてBIRL-MD(MD: Multiple Dynamics) 4)という手法として提案した．これにより従来の報酬推定よりもエキスパートの報酬に近い推定が行われたことが報告されている．しかし従来のBIRLで扱われていた複数種類の報酬の推定に拡張されていないという問題がある．

本研究では，この問題を解決するためにDPM-BIRL(DPM:Dirichlet Process Mixture Models) 5)という既存のアルゴリズムを用いた改善手法を提案し，Windy Grid World環境における計算機実験により従来のDPM-BIRLにて扱われている複数種類の報酬の推定よりもエキスパートに近い推定が行われていることを比較し，その結果について報告する．

2. 従来研究

2-1. BIRL-MD

BIRL-MDは中田らにより提案された逆強化学習の一種である．BIRLではエキスパートがある環境において生成したデータ群から報酬の分布を推定しているが，BIRL-MDでは状態遷移確率が異なる環境とエキスパートが各環境で生成したデータ群から報酬の事後分布を推定する手法である．このイメージを図1に示す．

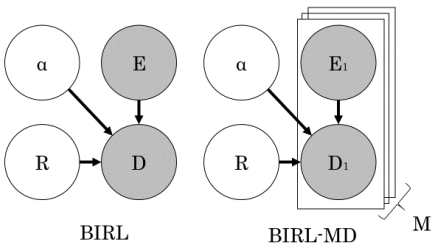

図1.BIRLとBIRL-MD

ここでRは報酬，Eは環境，Dはエキスパートのデータ群である．これらを報酬の事後分布として定式化したものを式(1)に示す．

$$P(R|\{(D_m,E_m)\}_{m=1}^{M}) = \frac{\prod_{m=1}^{M}P(D_m|R,E_m)}{\prod_{m=1}^{M}P(D_m|E_m)}P(R) \quad (1)$$

ここでBIRLのエキスパートモデルを式(2)に示す．

$$P(D|R) = \frac{1}{Z}exp\left(\frac{1}{\kappa}\sum_{(s,a)\in D}Q^*(s,a,R)\right)P(R) \quad (2)$$

ここで式(1)を式(2)で拡張したものとして式(3)に示す．

$$P(R|\{(D_m,E_m)\}_{m=1}^{M}) = \frac{1}{Z}\exp\left(\frac{1}{k}\sum_{m=1}^{M}\sum_{(s,a)\in D_m}Q^*(s,a,R,E_m)\right)P(R) \quad (3)$$

この式により報酬の事後分布を表すことができる．

2-2. DPM-BIRL

DPM-BIRLはChoiらにより提案された逆強化学習の一種である．逆強化学習では多くの場合行動データが単一のエージェントから生成されたものを用いるという点でデータの希薄性や

Estimating Multiple Rewards Using Multiple Environmental Trajectories in Bayesian Inverse Reinforcement Learning

Mitsuru IGARASHI and Yukari YAMAUCHI

データが十分にある場合でも報酬が変化するなどのデメリットがある．そのため，異なる環境下での複数のエージェントから行動データを収集する必要がある．そこでDPM(ディリクレ過程混合)モデルをBIRLに統合したDPM-BIRLを使いノンパラメトリックベイズを扱う手法が提案された．これを図2に示す．

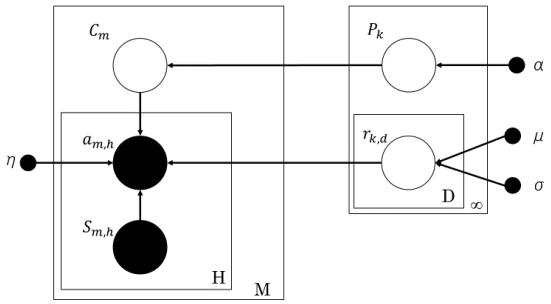

図2.DPM-BIRL

3. 提案手法

本研究では，BIRL-MD で課題とされている複数報酬の推定を行うという問題に対し，DPM-BIRL で扱われた DPM モデルを用いたノンパラメトリックベイズアルゴリズムを組み合わせた手法を提案し，複数環境において複数の報酬がある場合に従来の DPM-BIRL と比較し，エキスパートにより近い報酬の推定を行わせる．

4. 実験および検討

本実験ではWindy Grid World環境を用いる．この環境は格子状に設計された迷路空間であるGrid World空間において，ある確率で上下左右の風が吹く方向にそれぞれ30%の確率で遷移する環境である．また，無風状態も存在する．これにより複数の環境を表現している．

エージェントは上下左右のいずれかの行動をとりスタートからゴールを目指していく．この環境のイメージを図3に示す．

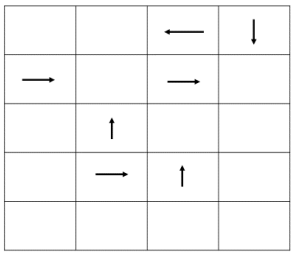

図3.Windy Grid World環境

この環境で，従来の手法と提案手法で，より複数の報酬の推定が行われていることを比較していく．

評価には逆強化学習で一般に用いられるEVDという指標を使う．これを式(4)に示す．

$$EVD = E_{\pi_{exp}}\left[\sum_{t=0}^{\infty}\gamma^t R_{exp}(s_t)\right] - E_{\hat{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t R_{exp}(s_t)\right] \quad (4)$$

5. まとめ

今回, BIRL-MDにおける複数報酬の推定への拡張問題に対し，DPM-BIRLで使われているDPMモデルを統合することでノンパラメトリックベイズを扱う手法を提案し，複数環境において複数報酬ある場合でもエキスパートに近い推定を行うという改善を試みた．

【参考文献】

1)Sutton,R.S. and Barto, A. G.: Reinforcement Learning:An Introduction, MIT press (2018)

2) Ng, A. Y., Russell, S. J., et al.: Algorithms for Inverse Reinforcement Learning., in Proceedings of The Seventeenth International Conference on Machine Learning,(2000)

3) Ramachandran, D. and Amir, E.: Bayesian Inverse Reinforcement Learning,IJCAI International Joint Conference on Artificial Intelligence, (2007)

4) 中田勇介,荒井幸代,複数環境におけるエキスパート軌跡を用いたベイジアン逆強化学習，人工知能学会論文誌, G-J73_1-10, (2020)

5) Choi, J. and Kim, K.-E.: Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions, in Advances in Neural Information Processing Systems, pp. 305–313 (2012)