F-015

模倣学習における多様な典型的行動列抽出による説明性向上
Discovering Diverse Prototypical Sequences of State-Action Pairs for Interpretable Imitation Learning

入江 理政‡　　吉永 直生‡
Michimasa Irie　Naoki Yoshinaga

1. はじめに

模倣学習では、エキスパートの状態と行動の履歴データを用いて、エキスパートの行動選択を模倣するように学習する。この模倣学習を医療の治療診断推薦などの重要なケースに適用する際に、模倣精度だけでなく判断根拠が求められる場合がある。そこで、判断根拠の説明材料として、エキスパートの状態行動履歴中から典型例となる列(プロトタイプ)を抽出し、行動の予測時に根拠としてそれらの中から１つ選択するように学習することで説明性を向上させる手法がある。しかしながら、推論の根拠として特定のプロトタイプが出力されてしまうことがあり、適切な根拠の出力が難しい場合がある。そこで本研究では、上記課題である偏りを解消し、より有用で多様なプロトタイプを推論根拠として提示するための改良手法を提案し、検証を行った。

2. 従来手法

まず、本改良手法のベースとなる手法である prototypical option discovery for interpretable imitation learning(IPOD)について述べる。

2.1 オプション

エキスパートの状態s_tと行動a_tの列$\tau = \{s_1, a_1, \dots, s_T, a_T\}$を学習し、得られたモデルに対して状態$s_t$を入力したときのエキスパートがとりそうな行動$a_t$を推論する。ここで、行動のサブゴールとなるオプション$o_k \in \mathbb{R}^n, k = \{1, \dots, K\}$を導入する($K$は自然数)。図１に示すように、行動$a_t$を推論するにあたり、まず状態$s_t$の入力に対し、方策$\pi_h(o_k|s_t)$に従いオプション$o_t$が選択される。次に、状態$s_t$と選択されたオプション$o_t$を入力して方策$\pi_\theta(a_t|s_t, o_k)$に従い行動$a_t$が推論される。この時方策$\pi_\theta(a_t|s_t, o_k)$は式(1) に示す真の行動方策$\pi_E$との損失関数$L_{imt}$によって学習され、オプション方策$\pi_h(o_k|s_t)$は行動の予測精度に応じた報酬が貰える強化学習によって学習される。

$$L_{imt} = -\sum_{n=1}^{N} \pi_E(a_n|s_n) \log \pi_o(a_n|s_n, o_k) \quad (1)$$

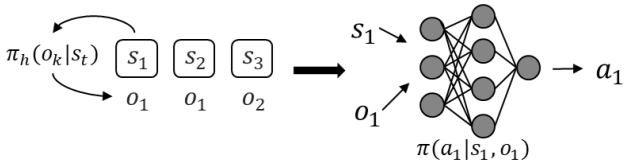

図１　行動推論

2.2 プロトタイプ

本手法におけるプロトタイプの概要と更新方法について述べる。

‡ NEC　デジタルテクノロジー開発研究所

2.2.1 プロトタイプの概要

プロトタイプはエキスパートの状態行動履歴中に含まれる典型的な部分列を表し、オプションごとに１つのプロトタイプを持つように学習される。すなわち、サブゴールごとに典型的な状態行動列を提示することで、予測の説明性を向上することができる。

2.2.2 プロトタイプの更新

図２に示すようにプロトタイプはエキスパートの状態列をエンコードした特徴ベクトルから決定される。はじめに、状態列のうち、選択されたオプションが連続する列ごとに分割してセグメントにする。次に LSTM を用いてセグメントをエンコードして特徴ベクトル$f_\phi(s_{v_{m'}:v_m})$を得る($v_{m'}, v_m$はそれぞれセグメントの開始と終了のステップ)。こうして得られた状態列セグメントの特徴ベクトル達から、プロトタイプを特徴ベクトル空間上にて更新していく。

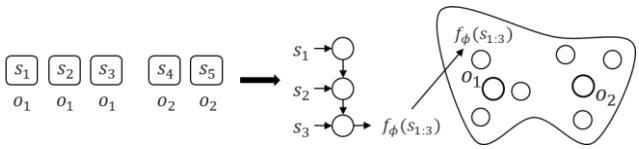

図２　プロトタイプの更新

プロトタイプ特徴ベクトルおよびエンコーダのパラメータは以下の式(2),(3)から求められる損失を最小とするように学習される。L_{emb}によってセグメントの特徴ベクトルがプロトタイプに近くなるような引力作用が働き、オプションごとにクラスタ構造が形成されやすくなる。また、L_{opt}の第１項はプロトタイプが最も近い状態列セグメントにより近づくことで根拠提示時の説明性を向上させ、第２項はプロトタイプ間の距離が近い場合に遠ざかるように学習することでプロトタイプの多様性を向上させる。ここで、$\lambda_1, \lambda_2 \in [0,1]$は重みのバランスを決めるハイパーパラメータ、$d_{min}$は２つのプロトタイプの距離の近さの応じた損失を決める閾値を表す。

$$L_{emb} = \sum_{m=1}^{M} \min_{k=1}^{K} \left\| f_\phi\left(s_{v_{m'}:v_m}\right) - o_k \right\|_2^2 \quad (2)$$

$$L_{option} = \lambda_1 * \sum_{l=1}^{K} \min_{m=1}^{M} \left\| f_\phi\left(s_{v_{m'}:v_m}\right) - o_k \right\|_2^2 + \lambda_2 * \sum_{l=1}^{K} \sum_{j=l+1}^{K} \max\left(0, d_{min} - \left\| o_i - o_j \right\|_2^2\right) \quad (3)$$

2.3 課題

しかしながら、上記手法では状態列の特徴ベクトルが特徴空間上に分散されないために、行動の推論時に特定のオプションに偏って選択されやすく説明性が損なわれることがある。

3. 提案手法

本手法では前述の課題の解決のために、状態列の特徴をより掴み、特徴空間上により分散させるように特徴ベクトルの抽出を工夫することでプロトタイプの特徴ベクトルの偏りを抑制し、オプションの選択を 1 つに偏りにくくする。

3.1 VAE

前述の従来手法のようにエンコーダのみである場合、状態列の特徴表現能力が不足している。そこで、特徴ベクトルから元の状態列を再構成するデコーダを導入し、オートエンコーダの構造とすることで状態列の特徴表現をより掴み、特徴空間上に特徴ベクトルを分散させることができる。本手法では特に、図 3 に示すようなオートエンコーダの構造をした生成モデルである Variational Autoencoder(VAE)を用いることを考える。VAE では、事前分布として標準正規分布$\mathcal{N}(0, I)$が仮定された潜在変数を導入し、学習対象となるデータが与えられた時の潜在変数の事後分布$\mathcal{N}\left(\mu_{v_{m'}:v_m}, \sigma^2_{v_{m'}:v_m} I\right)$の平均と分散のパラメータ$\mu_{v_{m'}:v_m}$、$\sigma^2_{v_{m'}:v_m}$を学習する。このように表現に制約がかかった潜在変数を特徴ベクトルとして扱うことで、特徴ベクトルの分布の偏りを防ぎ、偏った値を持ったプロトタイプの特徴ベクトルが生じることをより抑制することができる。VAE による損失関数を式(4)に示す。ここで$D_{KL}(P \parallel Q)$はQに対するPの KL-ダイバージェンスを表す。

$$L_{VAE} = \sum_{t=1}^{T} \|\hat{s}_t - s_t\|_2^2 + D_{KL}\left(\mathcal{N}\left(\hat{\mu}_{v_{m'}:v_m}, \hat{\sigma}^2{}_{v_{m'}:v_m} I\right) \parallel \mathcal{N}(0, I)\right) \quad (4)$$

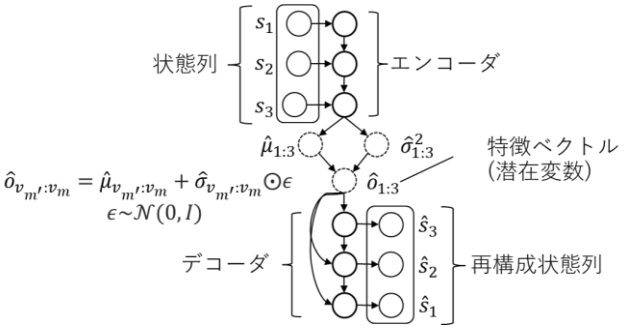

図 3　VAE による特徴抽出

3.2 プロトタイプの更新

従来手法のように状態列をエンコードして直接特徴ベクトルを学習する代わりに、入力に対する潜在変数の事後分布を学習することに伴い、プロトタイプの更新式を変更する。すなわち、従来手法のように特徴ベクトル間のユークリッド距離からプロトタイプを求めるのではなく、分布間のダイバージェンスを損失関数として利用する。ただし、式が持つ効果に関しては従来手法と同様となる。オプションkのプロトタイプの潜在変数の分布を$\mathcal{N}\left(o_{\mu k}, o^2_{\sigma k} I\right)$とすると、損失関数は以下のように書き直される。ここで、$D_{JS}(P \parallel Q)$はPとQの JS-ダイバージェンスを表す。

$$L_{emb} = \sum_{m=1}^{M} \min_{k=1}^{K} D_{KL}\left(\mathcal{N}\left(\mu_{v_{m'}:v_m}, \sigma^2_{v_{m'}:v_m:} I\right) \parallel \mathcal{N}\left(o_{\mu k}, o^2_{\sigma k} I\right)\right) \quad (5)$$

$$L_{option} = \lambda_1 * \sum_{i=1}^{K} \min_{m=1}^{M} D_{KL}\left(\mathcal{N}(o_{\mu k}, o^2_{\sigma k} I) \parallel \mathcal{N}\left(\mu_{v_{m'}:v_m}, \sigma^2_{v_{m'}:v_m} I\right)\right) + \lambda_2 * \sum_{i=1}^{K} \sum_{j=i+1}^{K} \max\left(0, d_{min} - D_{JS}\left(\mathcal{N}(o_{\mu i}, o^2_{\sigma i} I) \parallel \mathcal{N}(o_{\mu j}, o^2_{\sigma j} I)\right)\right) \quad (6)$$

3.3 実験

ファッション通販サイトの商品ページクリック履歴データを用いて手法の比較検証を行った。過去にクリックした商品情報(商品 ID、カテゴリ、一覧表示時の位置など)とアクセス元の国から、ユーザーが次にアクセスしそうな商品の予測を行った。訓練データから抽出した状態列セグメントの特徴ベクトルの従来手法と提案手法の比較を図 4 に示す。特徴ベクトルは 2 次元に設定し、対応するオプションごとにラベル付けしている。また、提案手法は分布の平均パラメータをプロットした。図 4 において、従来手法では選択されるオプションが偏っているのに対し、提案手法では多様なオプションが選択されていることが確認できる。また、従来手法では分布が集中している箇所があるのに対し、提案手法では分散されていることが確認できる。

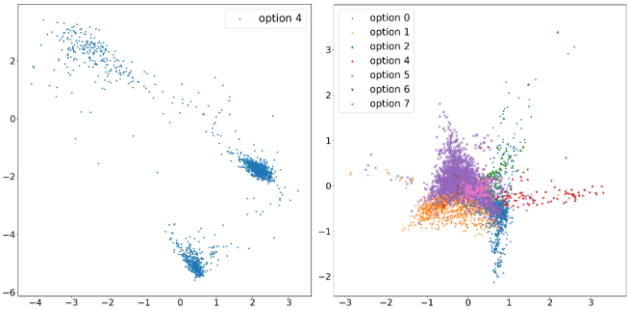

図 4　特徴ベクトルの比較
(左：既存手法、右：提案手法)

4. おわりに

判断根拠の説明材料としてプロトタイプを提示する模倣学習とプロトタイプを決定する特徴ベクトルの抽出方法について述べた。本手法は VAE を利用することで特徴ベクトルを分散し、オプション選択の偏りを抑制して行動の予測の判断根拠の説明性を向上することができた。

参考文献

[1] Kingma, Diederik P, Welling, Max. “Auto-encoding variational Bayes.” In Proceedings of the International Conference on Learning Representations (ICLR), (2014).
[2] Łapczyński M., Białowąs S, “Discovering patterns of Users' behaviour in an e-shop-comparison of consumer buying behaviours in Poland and Other European Countries”, Studia Ekonomiczne 151, p. 144-153, (2013)
[3] Wenchao Yu, Haifeng Chen, Wei Cheng, Interpretable Imitation Learning via Prototypical Option Discovery, US-20210374612-A1, (2021).