3M-06

強化学習を用いたロボットの行動に関する概念獲得への取り組み

恒川 英里† 小林 一郎† 麻生 英樹‡ 持橋 大地§ 中村 友昭¶ 長井 隆行¶
†お茶の水女子大学 ‡産業技術総合研究所 §統計数理研究所 ¶電気通信大学

1 はじめに

ロボットが人と共に暮らすことを考えた時，新しい状況にも適用可能である必要がある．強化学習は，経験によって最適行動を獲得出来るため，この問題の解決に有用だと考えられる．ロボットが未知な課題に対する行動を考慮した際，多くの観測情報を持つことにより，適切な行動が獲得出来ると考えられる．しかし，経験が蓄積されると，それまで観測した情報全てが必要でない場合があり，観測した情報を全て用いると，学習に時間がかかるというデメリットが発生する．そこで，効率良く学習するため，多層マルチモーダル LDA(mMLDA)[3] を用いる．本研究では，未知な状況である「片付け」という課題に対し，Q 学習と mMLDA を組み合わせることにより，適切な状態設定の獲得と，効率的な学習を行うことを目的とする．

2 Q 学習と mMLDA を用いた状態設定

2.1 概要

最初に，Q 学習を用いて報酬を得られる行動を探索し，報酬の得られた状態と行動のデータを集める．次に，そのデータセットを mMLDA を用いて分類する．分類結果を元のデータセットに適用し，報酬行動を推測できているかを確認する．この作業を繰り返し，一番結果が良かった mMLDA を用いて，元のデータセットで収束するまで Q 学習を行う．そして，データを再び集め，再学習を繰り返す．この 3 段階の学習を繰り返すことによって，最適な状態設定と，効率的な学習を目指す．学習の流れを図 1 に示す．

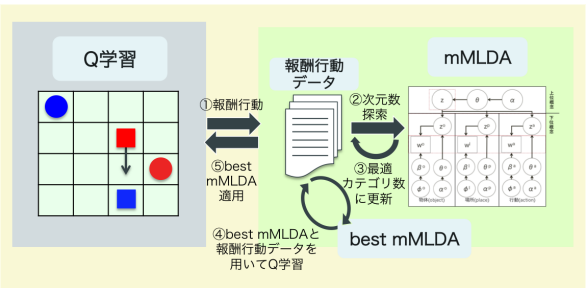

図 1: 最適状態獲得までの流れ

2.2 mMLDA

マルチモーダル LDA(MLDA)[2] を 2 層に拡張したモデルである．観測したマルチモーダル情報からその関係性を学習し，抽象化した概念を獲得することができる [3]．また，mMLDA は学習の過程で，多次元な情報から，指定されたカテゴリ数に情報を分類するという処理を行っている．本研究では，この特徴を利用し，観測した多次元のマルチモーダルな情報から他カテゴリも考慮した Q 学習の状態数削減を行う．

2.3 状態空間への mMLDA の結果の適用

観測した情報の分類を mMLDA を用いて行う．分類結果を元データと照合し，一番報酬を予測している確率の高い分類を新たな状態空間として Q 学習に適用する．これにより，獲得した概念に基づき，次元圧縮された状態空間での効率の良い Q 学習を実現する．

3 実験

3.1 作業課題

片付け課題の様子を図 2 に示す．4 × 4 のグリッド状に仕切られた机の上に置いてある物体を赤い物体に関しては左上，青い物体に関しては右下に置くことをゴールと設定し，学習を行う．

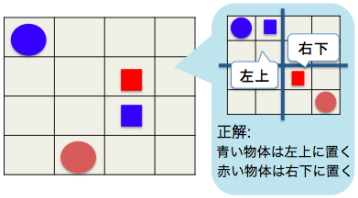

図 2: 片付け課題

3.2 設定

状態，行動，報酬を以下のように設定する．後ろに次元数を記す．

状態： 円形度 (10)，rgb 値 (10^3)，物体の面積 (10)，場所（机の上の座標）(16)，動き (3)

行動： (座標に) 掴んで置く，押して置く，そのまま

報酬： 正しい場所に置かれたら正の報酬（10）

今回用いるデータは上記設定の Q 学習プログラムを 100 エピソード× 10 回行って報酬の得られた行動 715 回分である．また，実験での，行動カテゴリのカテゴリ数は 3 に固定している．

An Approach to Concept Acquisition for the Act by a Robot Using Reinforcement Learning

†Eri TSUNEKAWA(tsunekawa.eri@is.ocha.ac.jp)
†Ichiro KOBAYASHI ‡Hideki ASOH §Daichi MOCHIHASHI
¶Tomoaki NAKAMURA ¶Takayuki NAGAI

3.3 実験結果

- カテゴリ分類

  分類の様子を見るため，人手で正しいと感じる2種類の設定を行い，学習させた．実験1では物体カテゴリ4つ(赤大丸，青大丸，赤小四角，青小四角)，右下と左上の領域のマス目から場所カテゴリ8つと設定し，実験2では正解に合わせて，物体カテゴリ2つ(赤，青)，場所カテゴリ2つ(左上，右下)に設定した．結果を表1に示す．値は識別精度の確率である．

表 1: 識別結果

カテゴリ	実験 1	実験 2
物体	0.51	0.95
行動	0.44	0.40
場所	0.28	0.52
行動時の報酬獲得確率	0.62	0.51

  表1の結果を受けて，物体カテゴリは，物体の観測情報に明確な差があるため，カテゴリ数が2つの方が良い精度であったと考えられる．行動カテゴリは，設定に差異が少ないため，値の変化は誤差範囲と考えられる．場所カテゴリについては，座標が連続しているため，区別がしにくいと考えられる．また，カテゴリごとの識別精度は実験2の方が良かったのに対して，報酬行動の識別は実験1の方が良い．カテゴリ毎のカテゴリ数の組み合わせによる適切な報酬行動の識別率の差異が生じる可能性があると考え，次にカテゴリ数の組み合わせを考える．

- 最適カテゴリ数組み合わせ探索

  物体と場所カテゴリ数をそれぞれ2，4，8と組み合わせて報酬を獲得する確率を比較した(表2)．物体のカテゴリ数が8と場所のカテゴリ数が8の時，カテゴリが余るなど，カテゴリ数が多すぎると判断し，表を空欄としている．

表 2: 最適カテゴリ数探索

報酬獲得確率		物体カテゴリ数		
		2	4	8
場所カテゴリ数	2	0.51	0.50	0.47
	4	0.53	0.51	0.50
	8	0.40	0.62	—

- カテゴリ数変化による下位概念の識別確率の変化

  物体と場所について，カテゴリ数の変化によるカテゴリ毎の識別の正解率を比較する(表3)．

  どちらもカテゴリ数が小さい方が精度が良かった．

- Q学習への適用

表 3: カテゴリ数変化による下位概念の識別確率

カテゴリ数	2	4	8
物体 (場所カテゴリ数=2)	0.95	0.81	0.33
場所 (物体カテゴリ数=2)	0.52	0.42	0.21

表2より，一番報酬獲得確率が高かった，物体カテゴリ数4，場所カテゴリ数8を採用し，Q学習を再度行う．状態の設定を以下のように更新する．状態の後ろには次元数を示す．

状態： 赤または青，丸または四角の物体(4)，場所(机の上の座標)(16)，動き(3)

各行動の報酬の遷移について，次元を削減する前と比較した(図3)．

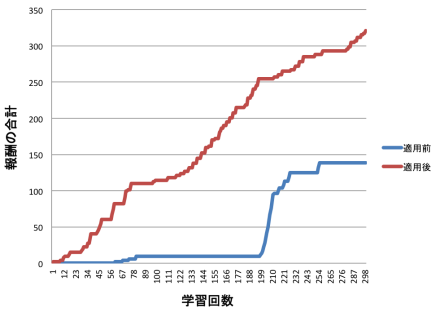

図 3: 報酬値の遷移の様子

3.4 考察

mMLDAを用いることにより，状態空間の次元圧縮に成功し，学習効率が良くなったことが確認された．一方で場所カテゴリ数が少なくなり，状態空間の次元が圧縮されているにもかかわらず，報酬獲得確率が減少することを確認した．これはmMLDAに与えたデータ数が少なかったために，獲得された概念が正確でなかったと思われる．

4 おわりに

Q学習とmMLDAを用いて，Q学習の持つ状態の適切な設定と効率の良い学習に取り組んだ．下位のカテゴリについて，カテゴリ毎の一番確率の高い識別結果に比例して，正しい報酬を獲得出来る確率も共に高くならないことを確認した．この点について，さらなる考察を行い，最適カテゴリ数のmMLDAと元データを用いたQ学習に取り組む．

参考文献

[1] Watkins, C.J.C.H.,Learning from Delayed Rewards. PhD thesis, Cambridge University, Cambridge, England. 1989.

[2] T. Nakamura et al., Grounding of Word Mean- ings in Multimodal Concepts Using LDA, in Proc. of IROS 2009, pp.3943-3948, 2009.

[3] アッタミミ，ムハンマド，阿部，中村，船越，長井，多層マルチモーダルLDAを用いた人の動きと物体の統合概念の形成，日本ロボット学会誌，Vol.32，no.8，pp89-100，2014.