3D2-OS-37b-2

多層マルチモーダルLDAを用いた報酬のモデル化

Modeling reward using multilayer multimodal LDA

宮澤 和貴 Kazuki Miyazawa　青木 達哉 Tatsuya Aoki　日永田 智絵 Chie Hieida　岩田 健輔 Kensuke Iwata　中村 友昭 Tomoaki Nakamura　長井 隆行 Takayuki Nagai

電気通信大学情報理工学研究科
Faculty of Infomatics and Engineering, The University of Electro-Communications

In order to perform better in the real world for robot, it is important to understand the real world correctly. The real understanding of the real world is that you get more rewards when you take actions based on your own understanding. Therefore, in this research we propose a framework to learn better concepts based on rewards. We aim to realize this by reinforcement learning using a model incorporating rewards for multilayer multimodal LDA.

1. はじめに

実世界で我々と共存し生活を支援するような知的なロボットが期待されている．しかしながら，そのようなロボットの普及には至っていない．これは，ロボットが実世界を理解しておらず，未知の環境や物事に対して対処できないことが原因であると考える．この問題を解決するために，文献 [1] ではロボットによる実世界の理解について検討している．これらの取り組みでは，ロボットが経験によって取得する聴覚・視覚・触覚などのマルチモーダル情報をカテゴリ分類し，概念をボトムアップに形成することを基盤としている [2]．ボトムアップに獲得した概念を用いることで，ロボットは目前の観測データから，モダリティーを超えた未観測情報を予測することが可能となる．筆者は，この予測こそが理解の本質であると考えている．ただしこのモデルは，知識の確率的表現であり，行動を計画したり決定する仕組みは含まれていない．そもそもロボットはどのように経験し，どのようにデータを収集するのか，逆に獲得した知識をどのように行動決定に利用するのか．これらを明らかにすることで，試行錯誤による運動学習から，概念・言語獲得，実世界の理解，行動計画などが一つの枠組みで結びつくと考えている．

多層マルチモーダルLDA (mMLDA) はマルチモーダルLDAの拡張であり，複数の概念とその概念間の関係を確率的に表現する [3]．文献 [4] は，mMLDA と強化学習を併用した枠組みを提案し，概念学習，知識獲得と様々なレベルでの行動決定について検討している．

関連研究としては，多くの強化学習に関する研究を挙げることができる [5]．特に近年の深層学習の成功によって，DQN が非常に注目を集めている [6]．Levine らは，ロボットの様々なタスクにおいて，深層学習をベースとした強化学習の有用性を示した [7]．しかし，深層学習を用いたアプローチの問題は，学習した知識を抽象化して他のタスクに生かすことができない点にある．また，行動と言語の学習を同時に行う枠組みは提案されていない．一方我々は，既に述べたように，ロボットが経験することで得るマルチモーダル情報を構造化することで知識を獲得する確率モデルの有用性を示してきた．こうした知識は抽象化されており，言語とも結びついているため，様々なタスクに再利用することができる．しかしこうした知識は，ロボットが経験することで受動的に構築される．しかし，構築された知識に基づいて行動することで能動的に知識を獲得することも重要である．これまでに，強化学習による行動学習と概念獲得，さらには言語学習までを統一的に扱った例は存在しない．行動と知識獲得のループを実現するための統一的な枠組みを mMLDA を中心に組み立てることを目指す．本研究ではその初期段階として，mMLDA を用いた報酬のモデル化を行い mMLDA と強化学習の統合について検討する．

連絡先: 宮澤 和貴，電気通信大学，東京都調布市調布ヶ丘 1-5-1，miyazawa@apple.ee.uec.ac.jp

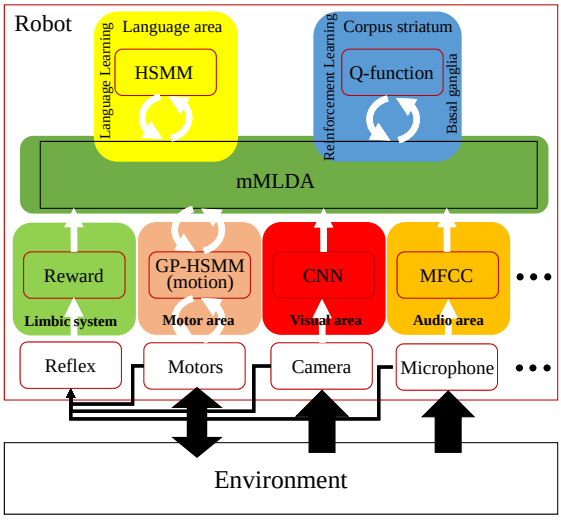


図 1: 統合モデルの概要.

2. Framework

図 1 に統合モデルの概要を示す．このモデルは， mMLDA を中心として，いくつかのモジュールが結合することで成り立っている．モジュールとしては，行動を決定するための強化学習部がある．また，強化学習で用いられる行動はガウス過程隠れセミマルコフモデル (GP-HSMM)[8] を用いた動作学習によって実現する． mMLDA の役割は，センサ—モータ情報を分類することで概念を形成し，それらの関係から強化学習のための状態空間や行動を作りだすことである．また，単語情報は mMLDA を通して実世界情報と結びついている．この単語情報に， HSMM によって表現される統語情報を適用することで，文を構成することが可能である [9]．また，逆に文章を分解して実世界情報を予測することで，文の意味を理解できるこ

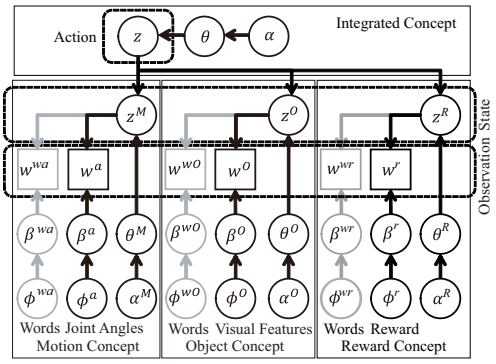


図 2: 本研究で用いた mMLDA のグラフィカルモデル．

とになる．この枠組みにより動作学習から言語操作までを統一的に学習できる．

本論文では，図 1 に示す統合モデルの中で強化学習と mMLDA と報酬の関係について考え，報酬に基づいたより良い概念を形成することを目指す．次章からこの枠組みの詳細についてそれぞれ述べる． 3. 章では動作の分節化について述べる．連続的なデータである動きデータから動き概念を形成するためには，データを分節化する必要があるためである．4. 章では，mMLDA と強化学習について述べる．

3. 基本動作の学習

本論文では，ロボットが初めに動作に関する知識を何も持っていない状態から学習することを想定する．そのためにロボットは，何かしらの動作から基本動作を学習し，それを組み合わせた動作を生成できる必要がある．これは，実際に体を動かした際に得られる関節角時系列の分節化・範疇化の問題である．体を動かすためには何かしらの反射行動や見まね動作，養育者の補助による動作などが必要であるが，後に示す実験では養育者役のユーザがロボットの手をつかんで動作の補助をすることによってロボットが動作を行い，その際の関節角時系列を分節化・範疇化することで基本動作を学習する．動作の分節化・範疇化は，ガウス過程隠れセミマルコフモデルに基づく手法によって行う ．

4. mMLDA と強化学習

ここではまず基本となる mMLDA について述べる．そして，mMLDA と強化学習の統合について説明する．

4.1 多層マルチモーダル LDA

多層マルチモーダル LDA (mMLDA) は，下位層に物体，動作，場所などの下位概念を表現するマルチモーダル LDA (MLDA) を，上位層にそれらを統合する MLDA を配置した階層的な構造をもつ確率モデルである．これにより，動作，場所，物体など各々のカテゴリ分類を行うと同時に，それらの概念間の関係を教師なしで学習することができる [3]．図 2 に，本研究の実験で用いた mMLDA のグラフィカルモデルを示す．図 2 において，z は統合概念を表すカテゴリであり，z^O，z^M，z^R はそれぞれ下位概念に相当する，物体，動作，報酬カテゴリである．上位カテゴリ z は，下位カテゴリ間の関係性を捉えており，ロボットの行動を表現する．w^o, w^a, w^r, w^w は観測データであり，それぞれ，物体情報，ロボットの動作情報，報酬情報，言語情報である．β^*，θ^* は多項分布のパラメータであり，ϕ^*，α^* はハイパーパラメータである．

4.1.1 下位概念

物体情報としては，ロボットの前の机の上の物体を奥行情報に基づくクラスタリングによって抽出し，抽出した物体画像を畳み込みニューラルネットワーク (CNN) によって特徴量に変換したものを利用する．ただし LDA では頻度情報を用いる必要があるため，特徴量の負の部分を 0 とし，全体を正規化することとした．動作情報としては，前述の動作分節化を行い，動作の中に現れる基本動作の頻度を利用する．報酬情報としては，ロボットが行動を行った際に得られた報酬値を利用する．言語情報としては，ロボットが動作している際に人から得られる発話を音声認識器で認識し，形態素解析を行いその中に含まれる単語の頻度を利用する．

4.1.2 パラメータ推定と予測

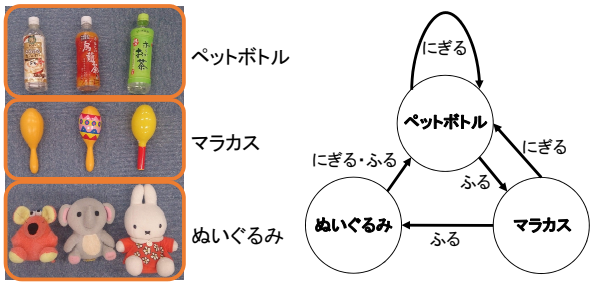
mMLDA では，各概念を表す隠れ変数 z，$z^C \in \{z^O, z^M, z^R\}$ を同時に学習する．学習にはギブスサンプリングを用い，各概念を表すカテゴリ z，z^C を，観測データ $w^m \in \{w^o, w^{wO}, w^a, w^{wM}, w^r, w^{wr}\}$ を用いてサンプリングする．サンプリングには，$\theta, \theta^C, \beta^m$ を周辺化した事後分布を用いる．さらに，学習モデルを用いることで，物体や動作の認識だけでなく，概念間の予測も可能となる．

4.2 強化学習

本研究では、強化学習として Q 学習を用いる．Q 学習における行動価値関数はある状態に対する行動価値を表現しており、図 1 に示すように，mMLDA と連携する形で存在している．mMLDA の報酬概念が即時的な報酬の予測であるのに対し，Q 関数では報酬の伝播が考慮されており，長期的に見た行動の価値が表現されている．学習手順は基本的な Q 学習をそのまま用いることができる．ただし、状態空間は mMLDA によって生成された物体概念 z^O によって規定する．また行動セットは、動作概念 z^M によって規定される．したがって、mMLDA を更新することで状態空間が変化することになる．これに対して、学習データを保持しておき、更新された状態空間を用いて再度オフラインで強化学習することで対応する．

5. 実験

実験には，図 4 に示すロボット (Baxter) を用いた．ロボットは，図 4 に示す動作を学習する．実験に用いた物体を図 3(a) に示す．物体は，図 3(b) に示すルールにしたがってロボットに与えられる．ロボットが，ある物体に対して行った動作により次に与えられる物体が変わる．物体と動作と報酬の関係を表 1 に示す．ロボットは，このように設定された中で，物体や動作などの知識を持たない状態から行動を行い，累積報酬を最大とするように行動選択することが目標である．



(a) 物体のカテゴリ　(b) 物体の出現ルール

図 3: 実験で用いた物体

表 1: 物体と動作と報酬の関係

	にぎる	ふる
ペットボトル	-1	-1
マラカス	-1	1
ぬいぐるみ	1	-1

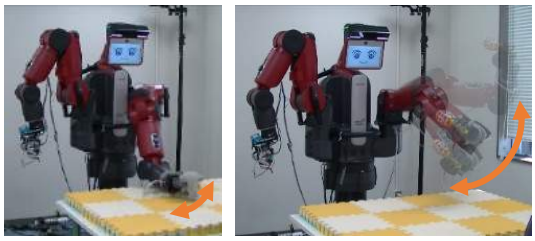


図 4: ロボットの行動: (a) にぎる (b) ふる.

5.1 動作の学習

まず始めに動作を獲得させた．実験者がロボットの腕をつかむことで動かし，動作データを取得した．この動作データを基に動作の分節化・範疇化を行い基本動作を学習した．分類した基本動作を人手によってつなげることで図 4 に示す握る・振るという動作を生成した．

5.2 mMLDA と強化学習

ロボットによる学習は，以下のような流れで行った．

1. ロボットが机の上の物体を見る
2. 物体に対して行動を選択する
3. 行動をしている間に人の発話情報を取得する
4. 行動に応じて報酬を得る
5. 取得したデータを保存して 1 に戻る

行動選択は，mMLDA に対して視覚情報のみを与え予測された物体概念と Q 関数に基づいて行った．今回は簡単のために，報酬はロボットの手首にあるボタンを押すことで与えた．報酬は表 1 に従ってロボットに与えた．

5.3 結果

報酬の予測と Q 関数についての結果を図 5 に示す．

報酬の予測は，mMLDA を用いて物体情報と動作情報から報酬概念を予測することである．(a) 報酬の予測で示している値は，正の報酬概念の予測確率から負の報酬概念の予測確率を引いた値である．表 1 と比較すると，ほぼすべての状態で設定した報酬と同じ報酬を予測できている．マラカスをにぎるときの報酬が誤って高く想起されているのは，学習時の経験としてマラカスを握るという行動を行った頻度が他の動作と比べて極端に少なかったため，上手く予測できなかったと考える．

(b)Q 値は，ロボットが mMLDA を用いて形成した状態空間上で Q 学習を行った結果である．この結果を見ると，(a) 報酬の予測で低く予測されている動作のペットボトルをふるという動作の Q 値が高くなっていることがわかる．この結果は，今回の実験設定から考えると報酬が多くもらえるように行動選択をしているといえる．他の状態に対しても同じように報酬を多く取得するように学習を行えている．つまり，順序関係のある行動学習を行えたといえる．

	にぎる	ふる
ペットボトル	−0.95	−0.96
マラカス	0.74	0.96
ぬいぐるみ	0.94	−0.97

(a) 報酬の予測

	にぎる	ふる
ペットボトル	−0.61	0.28
マラカス	0.00	2.03
ぬいぐるみ	1.20	−0.55

(b) Q 値

図 5: 報酬予測と Q 値

6. まとめ

本論文では，mMLDA を用いて報酬をモデル化することにより，報酬の予測が可能なことを示した．また，mMLDA と Q 学習を組み合わせることにより，ロボット自身が状態空間を形成しながら順序関係のある行動を学習できることを示した．さらに，mMLDA と HSMM を組み合わせることで，語彙や文法を獲得することができ自身の行動を文章で表現することや，言われた文章を解釈することができる．今後の課題としては，報酬に基づいてより良い概念構造の獲得を可能にするようにモデルを拡張することが挙げられる．

謝辞

本研究は，JST，CREST の支援を受け実施したものである．ここに感謝の意を表する．

参考文献

[1] T.Taniguchi, T.Nagai, T.Nakamura, N.Iwahashi, T.Ogata, and H.Asohe:Symbol Emergence in Robotics: a Survey, Advanced Robotics, Vol.30, pp.706–728, 2016

[2] 長井隆行, 中村友昭:マルチモーダルカテゴリゼーション 経験を通して概念を形成し言葉の意味を理解するロボットの実現に向けて, 人工知能 Vol.27, No.6, pp.555–562, 2012

[3] M. Attamimi, M. Fadlil, K. Abe, T. Nakamura, K. Funakoshi, and T. Nagai:Integration of Various Concepts and Grounding of Word Meanings Using Multi-layered Multimodal LDA for Sentence Generation, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.2194–2201, 2014

[4] 長井隆行, 中村友昭, アッタミミムハンマド, 持橋大地, 小林一郎, 麻生英樹:多層マルチモーダル LDA と強化学習による意味理解に基づく行動決定, 人工知能学会全国大会, 2F4-OS-01a-7, 2015

[5] R.S.Sutton, A.G.Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.

[6] Mnih *et al*, "Human-level control through deep reinforcement learning" Nature,Vol.518 pp.529–533, 2015.

[7] Sergey Levine and Chelsea Finn and Trevor Darrell and Pieter Abbeel, ''End-to-End Training of Deep Visuomotor Policies", Journal of Machine Learning Research, Vol.17, pp.1-40, 2016.

[8] 中村友昭, アッタミミムハンマド, 長井隆行, 持橋大地, 小林一郎, 麻生英樹, 金子正秀:ガウス過程の隠れセミマルコフモデルに基づく身体動作の分節化,人工知能学会全国大会，1O3-5，2016.

[9] M.Attamimi, Y.Ando, T.Nakamura, T.Nagai, D.Mochihashi, I.Kobayashi, H.Asoh, "Learning Word Meanings and Grammar for Verbalization of Daily Life Activities Using Multilayered Multimodal Latent Dirichlet Allocation and Bayesian Hidden Markov Models," Advanced Robotics 30(11-12), pp.806–824, 2016.