様式 C-19

科学研究費補助金研究成果報告書

平成　24　年　6　月　11　日現在

機関番号：１２６０１
研究種目：若手研究（B）
研究期間：2008～2010
課題番号：20700126
研究課題名（和文）帰納的強化学習の計算理論～環境の探索と帰納的再構成のベイズ推定

研究課題名（英文）　Computational Theory of Inductive Reinforcement Learning - Bayesian Inference on Environment Search and Inductive Reconstruction

研究代表者
牧野　貴樹（MAKINO　TAKAKI）
東京大学・生産技術研究所・特任准教授
研究者番号：20418651

研究成果の概要（和文）：
強化学習における環境の探索と帰納的再構成を、ベイズ推論手法に基づいて再構築する研究を行った。強化学習においては、エージェントは試行錯誤しながら環境モデルを学習するが、ベイズ理論に基づいた適切な環境モデルがあれば、不確実性を表現することで最適な探索が実現できるはずである。この目的のために、本研究では、TD-Network と呼ばれる予測的状態表現に基づく環境記述手法について、学習能力を高める提案を行った。また、隠れマルコフモデルのノンパラメトリックベイズモデルを拡張し、隠れ状態の階層的クラスタリングを実現する方法を提案した。さらに、徒弟学習の枠組みを応用し、他者の行動から環境についてのモデルをベイズ推定に基づいて構築する手法を提案した。これらは環境を探索しながら再構成してゆくプロセスのベイズ的再構成に必要となる要素技術である。

研究成果の概要（英文）：
This study focuses on environmental model reconstruction in reinforcement learning based on Bayesian inference techniques.　In reinforcement learning, an agent learns environment model by trial-and-error; if we have a suitable Bayesian environment model that represents uncertainty in the environment, an optimal exploration can be achieved.　For this purpose, we proposed new approaches that improve TD-network, an environment description framework based on predictive state representation. In addition, we extended a nonparametric Bayesian model for hidden Markov model to represent hierarchical clustering of hidden states.　Moreover, we applied the framework of apprenticeship learning and proposed a method that constructs environment model from other’s actions based on Bayesian inference. These are elements that are required for Bayesian reconstruction of the process of environmental search and reconstruction.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	800,000	240,000	1,040,000
2009 年度	500,000	150,000	650,000
2010 年度	500,000	150,000	650,000
年度			
年度			
総　計	1,800,000	540,000	2,340,000

研究分野：工学
科研費の分科・細目：情報学・知能情報学
キーワード：強化学習，Restricted Collapsed Draws，ベイズ推論,徒弟学習，無限隠れマルコフモデル，クラスタリング，中華料理店過程，TD-Network

１．研究開始当初の背景

環境の不確実性を探索する強化学習と、不確実性を確率分布の形で表現するベイズ推論の手法が結びつけられておらず、強化学習の探索コストに関する理論的解明が遅れていた。この２つを結びつけることで、これまで存在する種々の環境探索手法を統一的に説明することができ、また、最適に環境モデルを再構築するための指針が得られると考えられた。

２．研究の目的

本研究においては、環境を適切に表現するベイズモデルを構築し、エージェントの経験から環境の事後分布を得る方法について研究することを目的とした。

３．研究の方法

表現力の高い適切なモデルをコンピュータ上で構築し、シミュレーション環境において学習を行わせることでモデルおよび手法の評価を行った。

４．研究成果

主に以下の 3 点に関する成果が得られた。

1) TD-Network に関する提案

環境モデルの記述力を高める手段として、TD-Network と呼ばれる予測的状態表現に基づく環境記述手法について、学習能力を高める提案を行った。具体的には、Question network の自動生成のための新たな手法を提案したほか、Question network が不完全な場合においても学習が可能となる Simple Recurrent TD Network (SR-TDN) を提案した。これらの成果は、国際会議 International Conference of Machine Learning にて発表し、論文集に収録されたほか、国内の学会においても発表している。

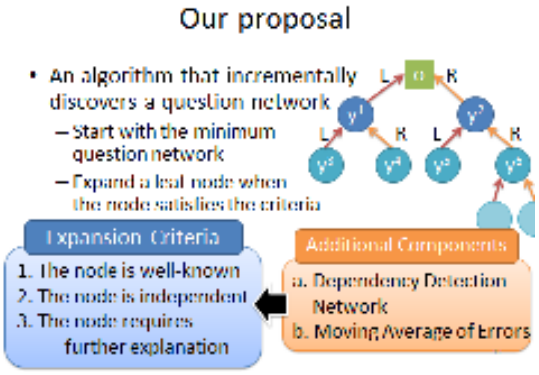

図 1: Question Network 自動生成手法

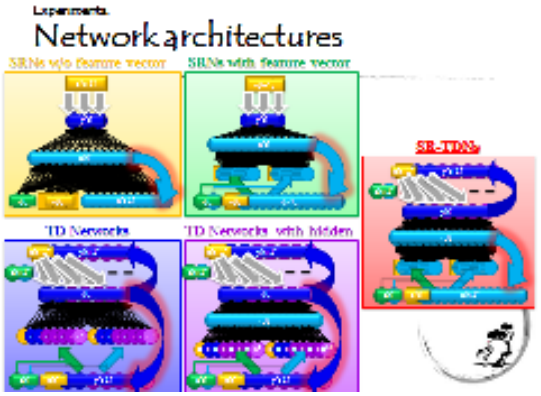

図 2: 不完全な Question Network から環境モデルを獲得する SR-TDN

2) 環境モデルのノンパラメトリックベイズモデル

隠れマルコフモデルのノンパラメトリックベイズモデルを拡張し、隠れ状態の階層的クラスタリングを実現する方法を提案した。これは、隠れ状態を階層化し、各階層が表すクラスタが前後の遷移確率に関して相関を持つような事前分布の構成法からなっているものである。また、それに付随して、新たなサンプリング推論手法 (Restricted Collapsed Draws: 制約付き周辺化分布サンプリング) の提案を行った。この手法を使うことで、提案するモデルにおいて、階層化した中華料理店過程を利用して表現した周辺化分布から直接サンプリングすることができ、高速な推論が可能となった。この成果は、IBISML 研究会にて招待講演, BayesComp2012 ワークショップなどで発表している。

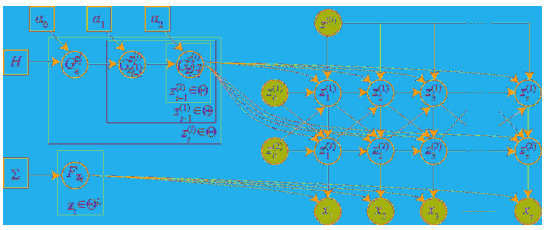

図 3. 隠れマルコフモデルの状態の階層的クラスタリングモデル (2 階層の場合)。

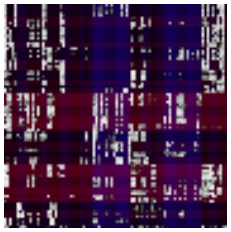

図 4. 3 階層の階層クラスタリング隠れマルコフモデルにおける遷移行列の例。

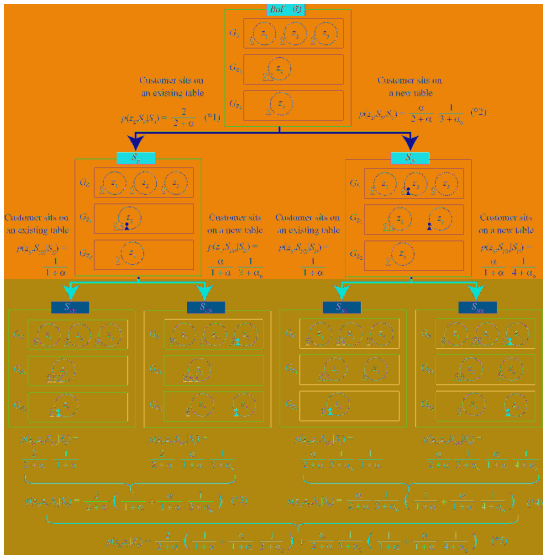

図 5: Restricted Collapsed Draws 推論手法における動作

3) 他者の行動からの環境モデルの学習
さらに、徒弟学習の枠組みを応用し、他者の行動から環境についてのモデルをベイズ推定に基づいて構築する手法を提案した。これは、自らの行動の結果だけではなく、他者が行動するに至る推論過程について考慮することで、環境に関するより正確なモデルを獲得するものである。これらにより、ベイズ推論を利用した環境モデルの構築に関しての研究を進めることが可能となった。
この成果については、国際会議 International Conference of Machine Learning にて発表し、論文集に収録されたほか、国内の学会においても発表している。

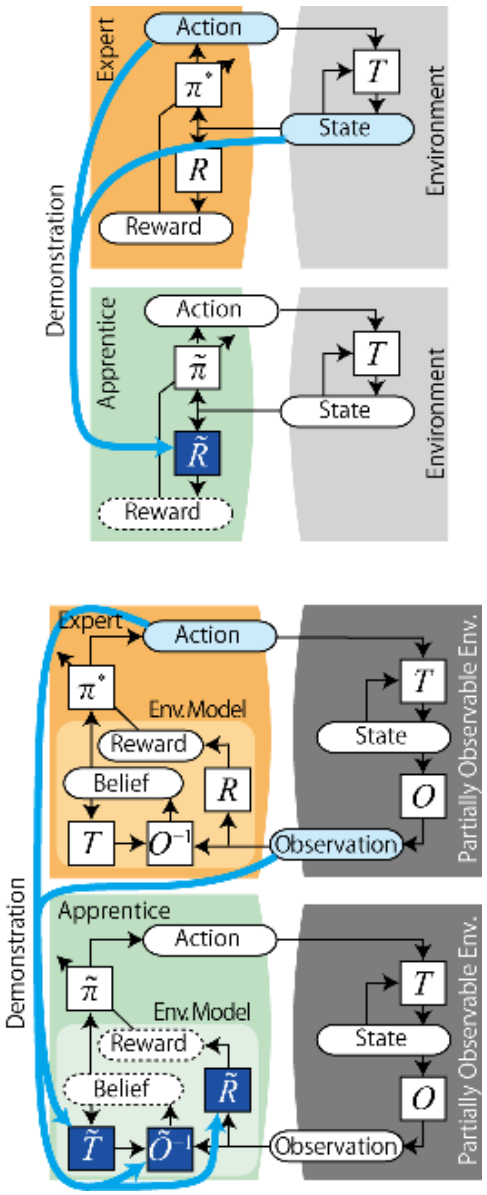

図 6. 従来の徒弟学習手法である逆強化学習（上）と、提案する徒弟学習による環境モデル獲得（下）

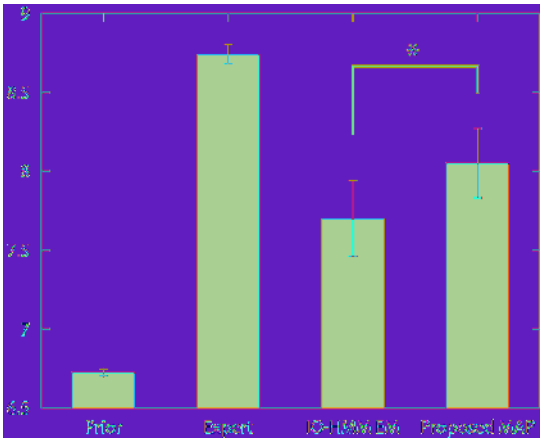

図 7. 学習なし、他者（エキスパート）、 従来手法、提案手法習での報酬値の比較

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 7 件）

[1] Takaki Makino and Johane Takeuchi. **Apprenticeship learning for model parameters of partially observable environments**. To be appeared in *ICML '12: Proceedings of the 29th Annual international conference on machine learning*, Edinburgh, June 2012. ［査読あり］

[2] 牧野貴樹, 竹内誉羽. **部分観測環境のモデルパラメータに対する徒弟学習**. 信学技報, Vol. 111, No. 480, pp. 49-54, March 2012. IBISML2011-94. ［査読なし］

[3] 牧野貴樹. **強化学習(私のブックマーク)**. 人工知能学会誌, Vol. 26, No. 3, pp. 301-303, 2011. ［査読なし］

[4] 牧野貴樹, 滝久雄, 合原一幸. **利他的行動と再帰的他者推定**. 生産研究, Vol. 62, No. 3, pp. 259-265, 2010. ［査読あり］

[5] Taiki Takahashi, Tarik Hadzibeganovic, Sergio A. Cannas, Takaki Makino, Hiroki Fukui, and Shinobu Kitayama. **Cultural neuroeconomics of intertemporal choice.** *Neuroendocrinology Letters*, Vol. 30, No. 2, pp. 185-191, 2009. ［査読あり］

[6]Takaki Makino. **Proto-predictive representation of states with simple recurrent temporal-difference networks**. In Léon Bottou and Michael Littman, editors, *ICML '09: Proceedings of the 26th Annual international conference on machine learning*, vol. 26, pp. 697-704, Montreal, June 2009. Omnipress. ［査読あり］

[7] Takaki Makino and Toshihisa Takagi. **On-line discovery of temporal-difference networks**. In Andrew McCallum and Sam Roweis, editors, *ICML '08: Proceedings of the 25th Annual International Conference on Machine Learning*, vol. 25, pp. 632-639, Helsinki, 2008. Omnipress. ［査読あり］

〔学会発表〕（計 9 件）

[1] Takaki Makino. **Hierarchical Nested Infinite Hidden Markov Models**. Bayesian Inference and Stochastic Computation 2012 workshop, 立川市, 2012/6/22.

[2] Takaki Makino and Johane Takeuchi. **Learning model parameters of partially observable markov decision process from demonstration**. In *Proc. of the 2nd International Symposium on Innovative Mathematical Modeling*. 東京, 2012/5/13.

[3] Takaki Makino. **Slice sampling for chinese restaurant process**. In *Proc. of the 2nd Asian Conference on Machine Learning (ACML 2010)*. Tokyo, 2010/11/8.

[4] 牧野貴樹. **ノンパラメトリックベイズに基づく統計的機械学習**. 電子情報通信学会技術研究報告 IBISML2010-14, 電子情報通信学会, 東京, 2010/6/15. ［IBISML 第 1 回研究会 招待講演］

[5] 牧野貴樹. **階層状態無限隠れマルコフモデル**. 情報論的学習理論 (IBIS2009) ポスター発表 (福岡市), 2009/10/20.

[6] Takaki Makino, Taiki Takahashi, Hirofumi Nishinaka, and Hiroki Fukui. **Probabilistic discounting for modeling behaviors in Iowa gambling task**. In *Proceedings of Multi-disciplinary Symposium on Reinforcement Learning (MSRL 2009)*. Montreal, Canada, 2009/6/18.

[7] Takaki Makino. **Simple recurrent temporal-difference networks**. 情報論的学習理論ワークショップ (IBIS2008), 仙台市, 2008/10/29.

[8] 牧野貴樹, 合原一幸. **自己観測原理：他者認知の数理的枠組**. 第22回 人工知能学会全国大会, 旭川市, 2008/6/13.

[9] 牧野貴樹.**POMDP 環境中での TD-network の自動獲得：単純再帰構造による拡張**. 第 22 回 人工知能学会全国大会 予稿集, 旭川市,June 2008/6/13.

〔図書〕（計 2 件）

[1] 牧野貴樹. **コミュニケーションの自己組織化**. 国武 豊喜 (監修), 自己組織化ハンドブック, NTS 出版, pp. 438-443, 2009.

[2] Taiki Takahashi, Takaki Makino, Yu Ohmura, and Hiroki Fukui. **Employing delay and probability discounting frameworks for a neuroeconomic understanding of gambling behavior**. In M. J. Esposito, editor, *Psychology of Gambling*, pp. 67-82. Nova Science, 2008.

〔産業財産権〕
○出願状況（計０　件）
○取得状況（計 0 件）

〔その他〕
ホームページ等
http://www.sat.t.u-tokyo.ac.jp/~mak/

６．研究組織
(1)研究代表者
牧野　貴樹（MAKINO TAKAKI）
東京大学・生産技術研究所・特任准教授
研究者番号：20418651