1C2-4

インタラクションにおけるニューラルネットを利用した強化学習の効果

Reinforcement Learning with Neural Networks for Human-Agent Interactions

竹内誉羽　　辻野広司
Johane Takeuchi　Hiroshi Tsujino

株式会社ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd

We have constructed a modular neural network model based on reinforcement learning and demonstrated that the model can learn multiple kinds of state transitions with the same architectures and parameter values, and without pre-designed models of environments. In this paper, our previously proposed learning model is applied for a task of human-agent interactions, where the agent should be taught from a person about a sequence of pushing buttons. We compare the performance of our learning model in the task with an ordinal reinforcement learning and demonstrate the efficiency of our learning model for a human-agent interaction task.

1. はじめに

人型ロボットなど、将来、人と自由な対話等のインタラクションを通じて、人の要望に答えるような機械の実現が待たれている。本報告では、そういった人型ロボットやコンピュータ上のキャラクターエージェントが人との音声対話を通じて、なんらかの行動手順・系列を学習するような場面を考える。そこでは人からの指示は必ずしも一定しておらず、人の指示以外の情報を参考に行動系列を実行しなければならなくなる場合が考えられる。そのためキャラクターエージェントが自分自身で試行錯誤しながら正解の手順を学習しなければならなくなる。試行錯誤を通じた機械学習として強化学習 [Sutton 98] がよく知られており、これにより報酬信号をたよりに試行錯誤によって行動系列を学習できる。

一方で機械学習器をキャラクターエージェントに組み込んで人とのインタラクションに応用する場合、人とのインタラクションは非常に多様であり、従ってあらかじめ想定できない状況に対して適応できるようにすることが学習器本来の役割であると思われる。そのためには第一段階として、機械学習器から出来る限り事前設定の要素を減らすことが必要であると考える。特に強化学習では入力情報の作り込みなどの事前設定によって、その適用範囲が限定されてしまう。そのため、われわれはニューラルネットを使用することを考えた。ニューラルネットによる学習器は、入力情報に対して柔軟である。複数種類の情報をニューラルネットに並列に入力して学習させると、特定の情報が得られなくなったときには別の入力情報が適当に補ってくれると期待できる。また、分類するパターンが少ない場合には、低次元の入力よりも高次元の入力に対して、より学習速度も速くなる傾向がある [Buonomano 09]。パターン認識においては、これらは過学習の要因であり、利点とはみなされないことが多い。しかし、行動学習の場合、パターン認識のように過学習になっているかどうか調べるための一般性のある基準がない。行動学習の場合は、その基礎となる入力パターンの分類は学習器任せでよく、結果として正しい行動出力が得られればそれでよい。従ってニューラルネットのこうした特性を強化学習に活用することができると考えた。

しかしながら強化学習でよく使われる表（テーブル）を使った学習器よりニューラルネットによる学習器は、学習速度が非常に遅くなる。この問題は単一のニューラルネットではなく、複数のニューラルネットを使って分散して学習させることにより解決できる。本報告では複数のニューラルネットに分散して学習させるためのモジュール構造をもったニューラルネットによる強化学習器を使用した。これにより、入力情報を特別に作り込まなくても入力情報の変化に対応でき、従って人からの情報の欠損に対して有効であることを示す。また学習速度については表を使った一般的な強化学習器とほぼ同等になった。

連絡先: 竹内誉羽，（株）ホンダ・リサーチ・インスティチュート・ジャパン，〒 351-0188 埼玉県和光市本町８－１，TEL 048-462-5219（代表），FAX 048-462-5221，E-mail johane.takeuchi@jp.honda-ri.com

2. 関連研究

2.1 人・ロボット間インタラクション

ロボットが人とのインタラクションを通じて学習を行う研究の多くでは、人からの教示に関し対話などの自然言語情報を用いてはいない（[Quinton 07, Yamashita 08]。また強化学習との関連では、[Conn 07] が人からの教示と強化学習との組み合わせを試みている。以上の研究では人からの教示は出力に対して直接的に影響を及ぼす。たとえばロボットの腕の動作であれば、教示は直接腕を人が動かすなどして与えられ、そのまま学習装置の出力の規範となる。本報告では人・機械間の音声対話を陽に扱い、人からの音声による教示は直接出力に関係せずに入力情報となる。特に人との対話の不確定性を扱った関連研究として [Inamura 05] があり、人が物の色を指定する時のあいまいさを確率的に取り扱っている。対して本研究では、入力情報である音声認識器や人の発話の不安定性から生じる情報の欠落を他の情報によって補完することを目指している。

2.2 モジュール型強化学習

我々は，MOSAIC モデル [Haruno 01]、モジュール型強化学習 (MMML)[Doya 02]、mnSOM[Nishida 06] 等を参考に、ニューラルネットを用いたモジュール型強化学習 SOMRL(Self-organizing Modular Reinforcement Learning) を提案した [Takeuchi 08, Takeuchi 07]。図 1 にこのモデルの概要を示す。それぞれのモジュールは、状態予測のための 3 層バックプロパゲーションネットワークと 2 層の強化学習ネットワークからなる。強化学習部分には、最急降下 SARSA(λ) アルゴリズム

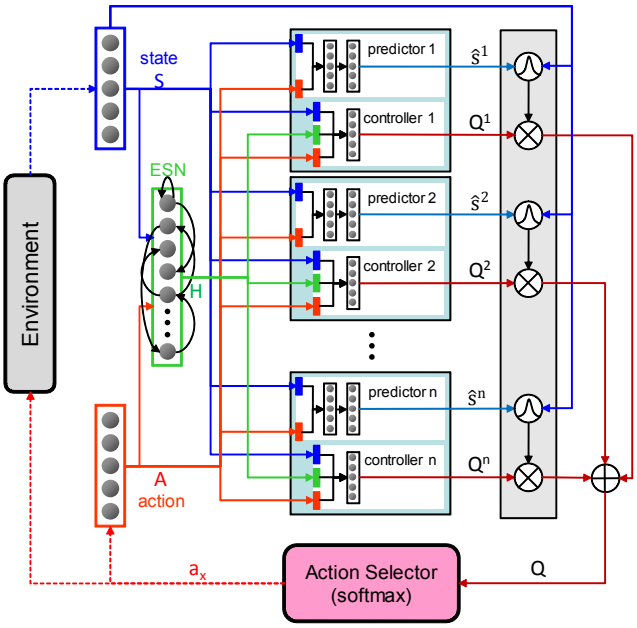

図 1: SOMRL の模式図。

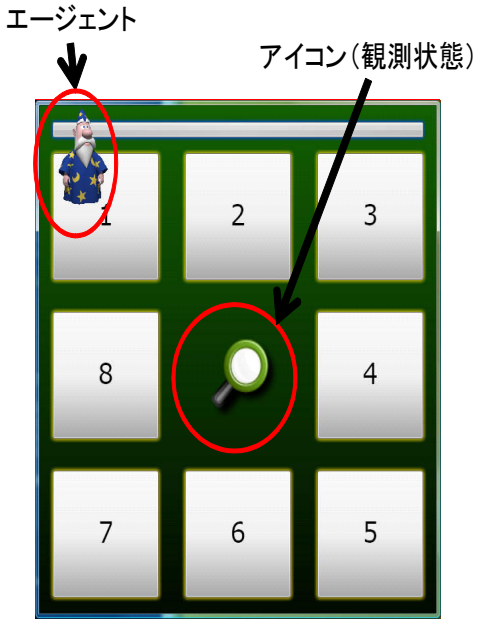

図 2: 作成した「ボタン押し」タスクのアプリケーション。

[Sutton 98] が実装されている。我々はモジュール内のそれら強化学習ネットワークの入力部分に観測状態（図中の S）と、ひとつ前の自分の行動選択の情報 (A)、さらにリカレントニューラルネットによる履歴情報 (H) を入力とした。履歴情報は、いわゆる部分観測 Markov 決定過程 (partially observable Markov decision process, POMDP) で有効であるが [Lin 93]、普通の Markov 決定過程 (MDP) では余分な情報である。しかし、この履歴情報がある場合のほうが、POMDP に限らず MDP でも学習速度の改善につながった。つまり問題ごとに学習器の構造もパラメータも変えずに対応できる可能性が示されたのである。本研究はこの SOMRL を入力情報に不確実性があり、途中からその一部分が情報として無意味になってしまう場合に応用したものである。

ここで、モジュールの選択は MMML に倣って、モジュール内の状態予測ネットワークの現在の観測状態に対する予測がより合致しているモジュールの強化学習ネットワークの結果が用いられる。強化学習ネットワークのほうは MOSAIC 的なルールでネットワークが更新され、状態予測の方は SOM 的なルールで更新される。詳しくは文献 [Takeuchi 08] を参考にされたい。強化学習ネットワークの出力は、それぞれの行動選択肢に対する行動価値関数値（いわゆる Q 値）である。これをもとにソフトマックス手法によって最終的な行動が決定される。

3. タスク設定とシステム構成

強化学習の学習課題は MDP で規定され、それは n 種類の異なる行動選択肢と m 種類の観測状態からなり、ある観測状態である行動選択肢を選ぶと次の観測状態に遷移し、その時の状態遷移は定常分布で記述される。本研究では最初の試みとして、この MDP の状態遷移をほぼそのままの形でエージェントが学習するタスクにした。図 2 に作成したアプリケーションを示す。ボタンが行動選択肢であり全部で 8 個ある。また、観測状態は中央のアイコンで示され、ボタンを押すと定められた状態遷移に従ってアイコンが変わる。ユーザはこの図では左上にいるキャラクターエージェントと対話する。このキャラクターエージェントがユーザの指示によってボタンを実際に押しに行く。従ってユーザはこのキャラクターエージェントに指示して、特定の順番でボタンを押すように仕向けなければならない。

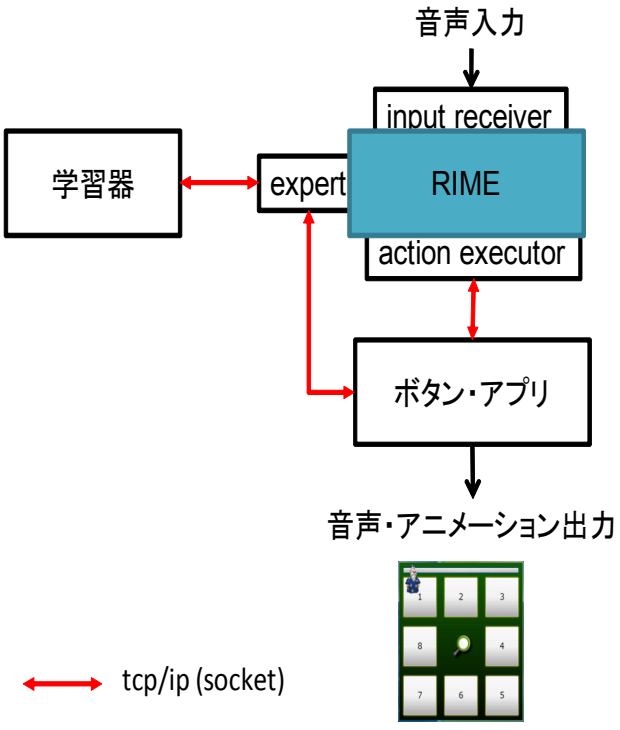

図 3: 全体のシステム構成。

システム構成（図 3）としては対話制御に RIME[Nakano 08] と呼ばれるシステムを使い、音声認識には julian[Kawahara 04] を使用している。音声認識文法としては、「１を押して」「１」「ボタン１」等を認識するが、学習器にはこれらをシンボル化

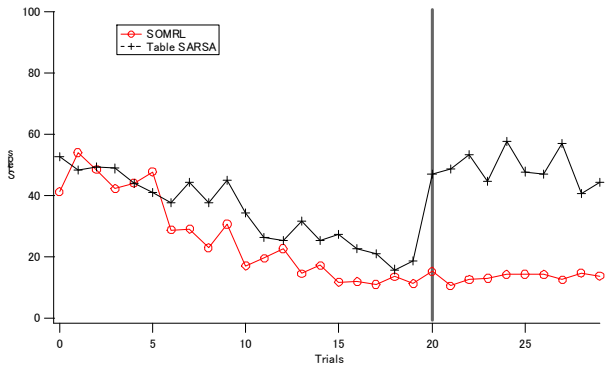

図 4: シミュレーション結果。それぞれ２０回分の試行の平均である。縦軸はそのトライアルに要したステップ数を表し、横軸はトライアルを初期状態から順番にならべたものである。

した情報、例えば “(push-button)(button-number 1)” といった形で送られる。学習器は、初期状態では送られて来るシンボルの意味をわかっていない。つまり、これらのシンボルの情報も適当にベクトル数値化されニューラルネットの入力情報の一部になるが、この入力情報を「ボタン１を押す」という行動選択肢に学習によってマップしなければならない。従って、最初は「１を押して」とユーザが発話しても、実際にボタン１を押しに行くとは限らない。文法にない発話、例えば「じゃ、次」「はい次」などは全て “(unknown)” というシンボルになる。学習器にとってこうした文法外の発話は、全て同じ入力情報となり、区別ができないので実質的に無意味になる。ただし、ユーザが「違う」「そうじゃない」といった発話をした場合だけは “(cancel)” というシンボルが送られ、これは直ちにキャラクターエージェントの行動停止を引き起こし、学習器には比較的小さな罰報酬が与えられる。こうしたキャンセル動作は強化学習上はエピソードの終端として扱われる。このキャンセル動作は学習器側からは行動として選択することはできない。また、ボタンを押す順番について一定の条件を満たした時には、成功したことがユーザにも通知され学習器にも報酬信号が送られ、強化学習上はこの時もエピソードの終端になる。成功時の報酬値は＋１ であり、キャンセル動作時の罰報酬値は −0.1 である。

その他、観測状態を表すアイコンについてもシンボル化された情報（例えば “(icon 1)”）として学習器に送られ、それぞれベクトルに変換されてニューラルネットに入力される。従って、観測状態としては、上記のユーザ発話の情報と表示されているアイコンの情報が学習器に与えられる。

4. シミュレーション結果

作成したシステムを使って音声対話により人がキャラクターエージェントにボタンを押す順番を教えられることを確認したが、本論文では音声対話システムを使わない簡単なユーザ・シミュレーションによる学習器の性能評価のみを示す。このシミュレーションによる評価では SOMRL がアイコンの情報とユーザ発話の情報を単純に並列に入力し特別に作り込んだりしなくても、ある状況の変化に対応できることを示す。このシミュレーションは、前半と後半部分とに分かれる。前半では正確に「ボタン n を押して」などといった指示が学習器に与えられ、学習器が間違った行動選択肢を選んだ場合にはただちにキャンセル動作が発行される。後半ではユーザからの指示はすべて “(unknown)” となる。これは実際の場合には、ユーザがキャラクターエージェントが十分に学習したと思って、正確な指示から「じゃ、次」のような発話に変化したことに対応する。いわゆる “Q テーブル” と呼ばれる表を使った強化学習器の場合、ユーザ発話とアイコンの情報の２つの入力情報について、あらかじめ Q テーブルを恣意的に作り込まない限り、この状況変化に適切に対応できない。こうした状況変化をあらかじめ、全て想定して作り込むようなことは非常に困難である。

図 4 に表の場合 (Table SARSA) と SOMRL の場合の比較を示す。ここで Table SARSA では、入力情報は２つの情報の集合の単純な積として表が作られている。このシミュレーションではボタン押しのシーケンスに成功し報酬 +1 を学習器がもらえるまでを１トライアルとしている。２０トライアルまでは想定ユーザが正確な指示を発話しているとし、２０トライアル目からは発話情報が “(unknow)” となるようにシミュレートした。この結果の示す通り、前半の学習スピードなどは両者にほとんど差が見られないが、２０トライアル目以降は Table SARSA は学習のしなおしになる。この時ユーザ発話からキャンセル動作がないため学習器自身の試行錯誤で学習しないといけなくなり、さらに学習スピードが遅くなる。対して SOMRL の場合は２０トライアル目以降も変わりなく前半での学習結果が保持されていることがわかる。

5. 考察

本稿では、キャラクターエージェントが人から手順を音声対話を通じて学習するような場面に、我々が以前に提案した SOMRL が応用できる可能性を示した。こうした課題は将来の人型ロボットや会話エージェントに必要な機能であり、その際、出来得る限り作り込みの要素を排除することがその学習システムにとって重要であると考えている。もし人とのインタラクションで起こる様々な事象をあらかじめ想定し作り込めるならば、そもそも学習器がいらなくなる。実際にはそうしたことは困難であり、あらかじめ作り込めないことに対応できる学習システムが必要であると考える。今後、この SOMRL を元に、様々な状況変化に対応するように拡張していく予定である。

参考文献

[Buonomano 09] Buonomano, D. V. and Maass, W.: State-dependent computations: spatiotemporal processing in cortical networks, *Nature Reviews Neuroscience*, Vol. 10, pp. 113–125 (2009)

[Conn 07] Conn, K. and Peters, R. A.: Reinforcement learning with a supervisor for mobile robot in a real world environment, in *Proceedings of the 7th IEEE International Symposium on Computational Intelligence and Robotics and Automation (CIRA 2007)*, pp. 73–78, Jacksonville, FL (2007)

[Doya 02] Doya, K., Samejima, K., Katagiri, K.-i., and Kawato, M.: Multiple Model-Based Reinforcement Learning, *Neural Computation*, Vol. 14, pp. 1347–1369 (2002)

[Haruno 01] Haruno, M., Wolpert, D. M., and Kawato, M.: MOSAIC Model for Sensorimotor Learning and Control, *Neural Computation*, Vol. 13, pp. 2201–2220 (2001)

[Inamura 05] Inamura, T., Inaba, M., and Inoue, H.: A Dialogue Control Model Based on Ambiguity Evaluation of Users 'Instructions and Stochastic Representation of Experiences, *Journal of Rbotics and Mechatronics*, Vol. 17, No. 6, pp. 697–704 (2005)

[Kawahara 04] Kawahara, T., Lee, A., Takeda, K., Itou, K., and Shikano, K.: RECENT PROGRESS OF OPEN-SOURCE LVCSR ENGINE JULIUS AND JAPANESE MODEL REPOSITORY, in *Proceedings of Interspeech-2004 (ICSLP)*, pp. 3069–3072 (2004)

[Lin 93] Lin, L.-J.: *Reinforcement Learning for Robots using Neural Networks*, PhD thesis, Carnegie Mellon University (1993)

[Nakano 08] Nakano, M., Funakoshi, K., Hasegawa, Y., and Tsujino, H.: A framework for building conversational agents based on a multi-expert model, in *SIGdial '08: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 88–91, Morristown, NJ, USA (2008), Association for Computational Linguistics

[Nishida 06] Nishida, S., Ishii, K., and Furukawa, T.: An Online Adaptation Control System Using mnSOM, in *Proceedings of 13th International Conference of Neural Information Processing (ICONIP)*, pp. 935–942 (2006)

[Quinton 07] Quinton, J. C. and Inamura, T.: Human-Robot Interaction Based Learning for Task-Independent Dynamics Prediction, in *Proceedings of the Seventh International Conference on Epigenetic Robotic*, No. 133–140 (2007)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA (1998)

[Takeuchi 07] Takeuchi, J., Shouno, O., and Tsujino, H.: Modular Neural Networks for Reinforcement Learning with Temporal Intrinsic Rewards, in *Proceedings of 2007 International Joint Conference on Neural Networks (IJCNN)* (2007), CD-ROM

[Takeuchi 08] Takeuchi, J., Shouno, O., and Tsujino, H.: Modular Neural Networks for Model-Free Behavioral Learning, in *Proceedings of the 18th International Conference on Artificial Neural Networks (ICANN)*, Vol. I, pp. 730–739 (2008)

[Yamashita 08] Yamashita, Y. and Tani, J.: Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment, *PLoS Computational Biology*, Vol. 4, p. e1000220 (2008)