7R-02

自己組織化特徴マップに基づいた確率的連想メモリによるProfit Sharingの実現

片山貴裕　長名優子

東京工科大学 コンピュータサイエンス学部

1　はじめに

教師信号を用いずに環境との相互作用により適切な行動を行うための政策を獲得するための学習方法として，強化学習に関する様々な研究が行われている．強化学習では，報酬を適切に設定しておくことで，未知の環境においても試行錯誤を繰り返すことで学習を行うことができる．

ニューラルネットワークを用いた強化学習では，ニューラルネットワークのもつ耐ノイズ性により，ノイズに強い学習を実現することができると考えられる．しかしながら，多くのニューラルネットワークは追加学習が行えないため，強化学習による学習とニューラルネットワークの学習が同時に行える方法として，自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリ [1] を用いた強化学習 [2] が提案されている．

本研究では，自己組織化特徴マップに基づいた確率的連想メモリ [3] を用いて Proit Sharing[4] を実現する．この手法は，自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリを用いた強化学習 [2] と同様，強化学習による学習とニューラルネットワークの学習を同時に行うことができる．

2　自己組織化特徴マップに基づいた確率的連想メモリ

2.1　構造

自己組織化特徴マップに基づいた確率的連想メモリは入出力層とマップ層，ニューロンの間をつなぐ重みから構成される．入出力層は，入力部，出力部，信頼度を表す 3 つの部分から構成されている．

2.2　学習

自己組織化特徴マップに基づいた確率的連想メモリの学習は，基本的には自己組織化特徴マップの学習アルゴリズムに基づいた方法で行われる．ただし，このモデルでは，学習がある程度進んだ段階で，各学習ベクトルが 1 つのマップ層のニューロンに対応づけられ，特定のパターンと対応づけられたマップ層のニューロンに結合する重みベクトルの値が固定される．このようにすることで，既学習パターンの情報を破壊することなく，新しいパターンを学習することが可能となる．また，必要に応じてマップ層にニューロンを追加することもでき，初期のマップ層のニューロン数にしばられることなく追加学習が行えるようになっている．

2.3　想起

想起時には，入出力層にパターンを入力する．マップ層のニューロン i の内部状態の値 u_i^{MAP} は

$$u_i^{MAP} = r_i \cdot g\left(1 - \frac{||\boldsymbol{x}^{IN} - \boldsymbol{w}_i^{IN}||}{\sqrt{N^{IN}}}\right) \qquad (1)$$

のように与えられる．ここで，$\boldsymbol{x}^{IN}$ は入力パターンのうち，入力部に入力される部分のパターンのベクトル (以下，入力ベクトルと呼ぶ)，$\boldsymbol{w}_i^{IN}$ はニューロン i の入力部の重みベクトル，N^{IN} は入力部のニューロン数，r_i はマップ層のニューロン i と入出力層の信頼度を表すニューロンの間の重みを表す．また，$||\cdot||$ はベクトルのノルムを表す．関数 $g(u)$ は，

$$g(u) = \begin{cases} u, & (u > \theta^R \text{かつ } u > \theta_{min}^R) \\ 0, & (\text{それ以外}) \end{cases} \qquad (2)$$

で与えられる．ここで，θ^R は入力ベクトルと重みベクトルの距離のしきい値であり，

$$\theta^R = d_{max} - a(d_{max} - d_{min}) \qquad (3)$$

のように与えられる．ここで，a $(0 < a < 1)$ はしきい値を決める係数である．d_{max} と d_{min} は

$$d_{max} = \max_i \left(1 - \frac{||\boldsymbol{x}^{IN} - \boldsymbol{w}_i^{IN}||}{\sqrt{N^{IN}}}\right) \qquad (4)$$

Profit Sharing using Self-Organizing Map-based Probabilistic Associative Memory
Takahiro Katayama and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

$$d_{min} = \min_i \left(1 - \frac{||\boldsymbol{x}^{IN} - \boldsymbol{w}_i^{IN}||}{\sqrt{N^{IN}}}\right) \quad (5)$$

で与えられる．また，θ_{min}^R はしきい値の最小値である．式 (3) の a を小さな値に設定することで，θ^R は d_{max} よりも小さく，なおかつ d_{max} に非常に近い値になる．こうすることで，入力と重みの距離が d_{max} に非常に近くなる重みをもつニューロンだけが，信頼度に基づいた内部状態の値をもつことができる．

マップ層のニューロン i が 1 を出力する確率 $P(x_i^{MAP} = 1)$ は

$$P(x_i^{MAP} = 1) = \frac{u_i^{MAP}}{\sum_{i'=1}^{N^{MAP}} u_i^{MAP}} \quad (6)$$

で与えられる．ここで，N^{MAP} はマップ層のニューロン数である．信頼度に基づいて決定される内部状態に基づいて出力が決定されるため，信頼度に基づいた確率的な連想を行うことができる．

3 自己組織化特徴マップに基づいた確率的連想メモリによる Profit Sharing の実現

Profit Sharing[4] では，エージェントの観測と行動の組をルールとし，報酬を基にルールの価値を更新することで学習を行う．エージェントが報酬を獲得したときに，初期状態から報酬を得るまでの一連のルール (エピソード) に報酬を以下のように分配する．

$$q(o_x, a_x) \leftarrow q(o_x, a_x) + r \cdot F(x) \quad (7)$$

ここで，$q(o_x, a_x)$ は時刻 x における観測 o_x のときに行動 a_x を取るというルールの価値，r は報酬量を表し，以前のルールの価値に報酬分配関数 $F(x)$ に基づいて分配された報酬を加算することで価値を更新している．報酬分配関数 $F(x)$ は

$$F(x) = \frac{1}{(|C^A| + 1)^{W-x}} \quad (8)$$

で与えられる．ここで，$|C^A|$ はエージェントの取りうる行動の数，W はエピソードの長さ，x は時刻を表す．報酬獲得の直前のルールに最も多く報酬が分配され，報酬獲得時の時刻から離れるほど分配される報酬の量が減るようになっている．

自己組織化特徴マップに基づいた確率的連想メモリにより Profit Sharing を実現する際には，入力部にエージェントの観測，出力部にエージェントの行動，信頼度にルールの価値を割り当てる．また，ボルツマン選択により行動選択を行うために，想起時にマップ層のニューロン i が 1 を出力する確率 $P(x_i^{MAP} = 1)$ を以下のように変更する．

$$P(x_i^{MAP} = 1) = \frac{\exp(u_i^{MAP}/T(t))}{\sum_{i'=1}^{N^{MAP}} \exp(u_i^{MAP}/T(t))} \quad (9)$$

ここで，$T(t)$ は t 回目の試行における温度パラメータであり，学習の進行に伴い 0 に近づけていく．$T(t)$ の値は学習の開始直後では大きな値に設定されるため，行動はほぼランダムに選択される．学習が進むにつれて $T(t)$ の値は 0 に近づくため，価値の高いルールの行動が高確率で選択されるようになる．

4 計算機実験

格子状に区切られたフィールドを一体のエージェントが一体の獲物を追う獲物捕獲問題を例題として実験を行った．エージェントは獲物の相対的な位置，エージェントの上下左右のマスの壁の有無を観測として扱う．獲物はエージェントから逃げるように行動する．エージェントは獲物と隣接したマスに移動することで獲物を捕獲できるものとし，獲物を捕獲すると報酬を獲得することができる．提案手法を用いて学習を行い，獲物を捕獲できるような行動が学習できることを確認した．

参考文献

[1] J. Niitsuma and Y. Osana : “Self-organizing map based probabilistic associative memory for sequential patterns,” Proceedings of IEEE and INNS Interna- tional Joint Conference on Neural Networks, Killarney, 2015.

[2] 新妻純, 長名優子: “自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習の実現,” 情処理学会第 77 回全国大会, 2015.

[3] Y. Osana : “Self-organizing map-based probabilistic associative memory,” Proceedings of International Conference on Neural Information Processing, Kuching, 2014.

[4] J. J. Grefenstette : “Credit assignment in rule discovery systems based on genetic algorithms,” Machine Learning, Vol.3, pp.225–245, 1988.