7D-02

# SOMに基づく多腕バンディットアルゴリズム

馬目　信人†‡　　篠原　修二‡　　鈴木　康大†‡　　朝長　康介†‡　　光吉　俊二‡

ソフトバンクロボティクス株式会社†　　東京大学大学院工学系研究科‡

## 1. はじめに

コミュニケーションロボットにおいて対面するユーザを満足させるには，ロボットの持つ多くの行動選択肢からより早くそのユーザに適した行動を出力する必要がある．このような問題は，多腕バンディット問題[1]として定式化される．多腕バンディット問題とは，レバーを引くとある確率で報酬が得られる腕が複数ある環境下においてどのように腕を選択すれば得られる報酬を最大化できるか考える問題である．コミュニケーションロボットにおいては，ロボットの行動選択肢を腕，ユーザの満足度を報酬と考えると，より腕の多い環境下においてより早く報酬確率の高い腕を選択することが求められる．

本稿では，人工ニューラルネットワークの一つであるSelf-Organizing Map (SOM) [2]を用いた多腕バンディット問題のための新しいアルゴリズムを提案する．また，数値実験により提案アルゴリズムが既存アルゴリズム UCB1[3]，UCB1-Tuned[3]，Thompson Sampling[4]に比べ，より腕の多い環境下においてより早く報酬確率の高い腕を選択できることを示す．

## 2. 提案アルゴリズム

提案アルゴリズムは，腕$i$を選択したときの報酬を腕$i$に対応する$\mathrm{SOM}^i$の入力とすることで腕の報酬確率を推定し，選択する腕を逐次決定していくものである．アルゴリズムを次に示す．

**Step 1.** 腕$i$に対応する$\mathrm{SOM}^i$を腕の数分用意する．このとき，$\mathrm{SOM}^i$の各ノードの重みの初期値$\mathbf{y}_k^i(0)$は全て 1 次元ベクトルで$\mathbf{y}_k^i(0) = (1)$とする．そして，時刻$t = t + 1$としてStep 2 に進む．

**Step 2.** 次式により全ての腕の評価値を計算し，その評価値が最も高い腕$j$を選択する．

$$j = \mathrm{argmax}_i \frac{1}{N^i}\sum_{k=1}^{N^i}\left\|\mathbf{y}_k^i(t-1)\right\|$$

このとき，$N^i$は$\mathrm{SOM}^i$のノード数である．

**Step 3.** 腕$j$を選択したときの報酬を確認する．このとき，得られた報酬が1の場合$\mathbf{x} = (1)$，報酬が 0 の場合$\mathbf{x} = (0)$とする．

**Step 4.** 選択した腕$j$に対応する$\mathrm{SOM}^j$について，$\mathbf{x}$に対する勝者ノード$c$を次式により決定する．

$$c = \mathrm{argmin}_k\left\|\mathbf{x} - \mathbf{y}_k^j(t-1)\right\|^2$$

**Step 5.** $\mathrm{SOM}^j$について，各ノードの重みを次式により更新する．

$$\mathbf{y}_k^j(t) = \mathbf{y}_k^j(t-1) + \alpha(t^j)h_{ck}(t^j)\{\mathbf{x} - \mathbf{y}_k^j(t-1)\}$$

このとき，$t^j$は腕$j$の選択回数である．また，$\alpha$は学習率，$h_{ck}$は近傍関数と呼ばれ，次式で与える．

$$\alpha(t) = \alpha_0\left(1 - \frac{t}{T}\right)$$

$$h_{ck}(t) = \exp\left(-\frac{{d_{ck}}^2}{2\sigma(t)^2}\right)$$

このとき，$\alpha_0$は学習率の初期値，$T$は学習率の縮小スピードを決める時定数，$d_{ck}$は勝者ノード$c$と近傍ノード$k$のユークリッド距離である．また，$\sigma$は近傍半径と呼ばれ，次式で与える．

$$\sigma(t) = \max\left\{\sigma_0\exp\left(-\frac{t}{\tau}\right), \sigma_{min}\right\}$$

このとき，$\sigma_0$は近傍半径の初期値，$\sigma_{min}$は近傍半径の最小値，$\tau$は近傍半径の縮小スピードを決める時定数である．そして，時刻$t = t + 1$としてStep 2 に戻り処理を繰り返す．

## 3. 数値実験

### 3.1. 設定

本稿では，多腕バンディット問題の中でも報酬が腕ごとに関連付けられた確率分布に従い与えられる確率的バンディット問題を対象とし提案アルゴリズムと既存アルゴリズム UCB1，UCB1-Tuned，Thompson Sampling との比較を行った．

確率的バンディット問題における報酬は，腕$i$ごとに設定された報酬確率$P_i$に基づいて決定される．プレイヤーは，腕$i$を選択すると確率$P_i$で報酬 1，確率$1 - P_i$で報酬 0 を得る．報酬確率$P_i$は，試行ごとに[0,1]区間の一様乱数で決定する．

---

Multi-armed bandit algorithm using self-organizing maps
†SoftBank Robotics Corp.
‡Graduate School of Engineering, The University of Tokyo

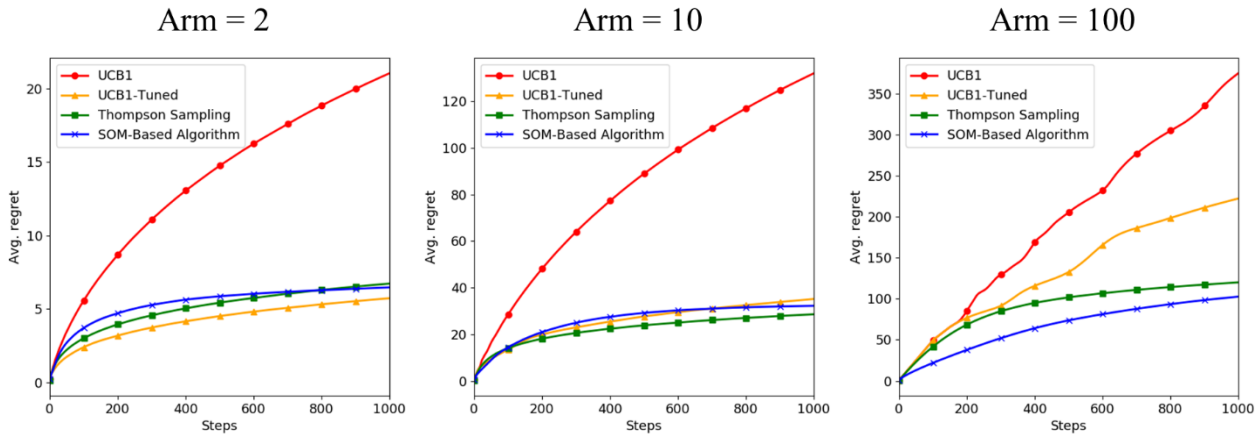


図 1. 腕の本数 2，10，100 の場合における Regret の 10,000 回平均値

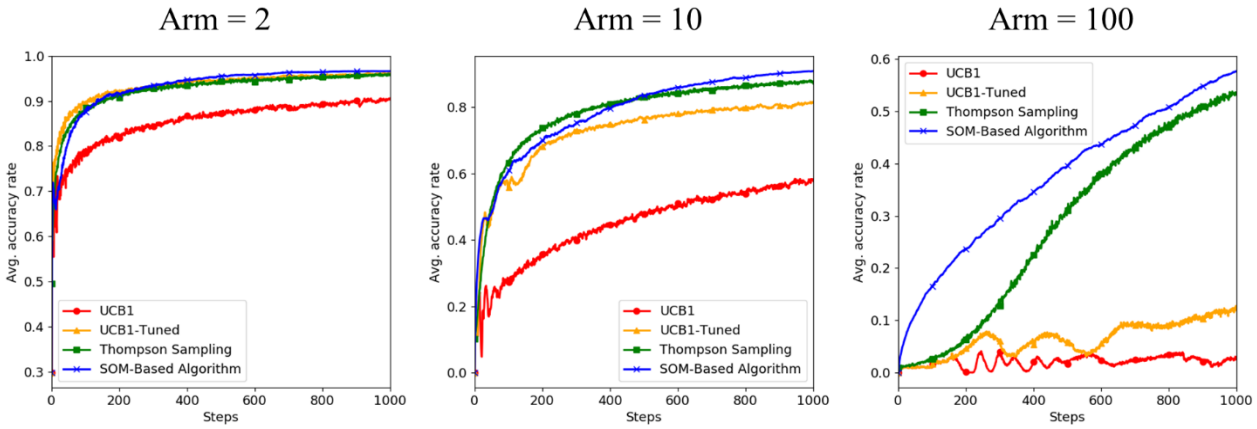


図 2. 腕の本数 2，10，100 の場合における Accuracy Rate の 10,000 回平均値

本実験では，性能評価のために腕の本数 2，10，100 の場合について，1,000 Step のシミュレーションを 10,000 回試行し，Regret と Accuracy Rate（AR）の平均値を算出した．Regret とは，全ての Step において最良の腕（$\mathrm{argmax}_i P_i$となる腕）を選択した場合と実際に選択した腕の報酬期待値の差であり，この値が小さいほど良いアルゴリズムと考える．AR とは，各 Step において最良の腕を選択した割合を表し，この値が大きいほど良いアルゴリズムと考える．提案アルゴリズムにおける SOM は，競合層の大きさ$10 \times 10$，非トーラス型の正方格子とし，$\alpha_0 = 0.1$，$\mathrm{T} = 1{,}000$，$\sigma_0 = 10$，$\sigma_{min} = 5$，$\tau = 50$とした．

### 3.2. 結果

腕の本数 2，10，100 の場合における Regret の結果を図 1 に，AR の結果を図 2 に示す．

図 1 より腕の本数 100 の場合について，全ての Step において提案アルゴリズムの Regret が一番小さい値となった．また，図 2 より腕の本数 100 の場合について，全ての Step において提案アルゴリズムの AR が一番大きい値となった．

## 4. 考察とまとめ

提案アルゴリズムは，腕の本数 100 の場合について Regret，AR ともに最良の結果となっている．これは，提案アルゴリズムが UCB1，UCB1-Tuned，Thompson Sampling と比べ，より腕の多い環境下においてより早く報酬確率の高い腕を見つけることができることを示している．

本稿では，SOM を用いた多腕バンディット問題のための新しいアルゴリズムを提案した．今後は，SOM のノード数を 1 とした場合の単純なモデルや学習率を変えたモデルについて分析を行う．

## 参考文献

[1] H. Robbins, “Some aspects of the sequential design of experiments,” *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[2] T. Kohonen, “Self-organizing Maps,” *Springer*, 1995.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[4] W. R. Thompson, “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples,” *Biometrika*, vol. 25, pp. 285-294, 1933.