Good Arm Identification via Bandit Feedback

2017 年度 修了
東京大学大学院 新領域創成科学研究科
複雑理工学専攻 杉山・佐藤・本多研究室
47-166095 鹿野 英明
指導教員：杉山 将 教授

1 はじめに

多腕バンディット問題とは，1 人のエージェントが K 台のスロットマシンの中から，最適な台を 1 つ選びプレイすることを繰り返し収益を最大にする台の選び方を見つける問題である．この問題は知識の探索と活用のトレードオフを扱う基本的な問題であり，試行錯誤によって有用な行動の獲得を目指す強化学習の一分野として位置づけられている．多腕バンディット問題は，1930 年代から医療や農業の基礎として研究されていたが，近年，情報通信技術の発達によって，インターネット広告配信やゲーム木探索などの実問題に応用されるようになり，現在，大きな注目を集めている．

本稿では，初めに，累積報酬最大化と最適腕識別と呼ばれる多腕バンディット問題の古典的な定式化としきい値バンディット問題と呼ばれる最近の定式化を紹介する．次に，我々が提案する優良腕識別の定式化とその研究成果について述べる．

2 多腕バンディット問題の定式化

多腕バンディット問題は，エージェントが何の最適化を目指すかによって様々な定式化が考えられる．本節では，本研究に関連する三つの定式化を紹介する．

2.1 累積報酬最大化

累積報酬最大化とは，時刻 T までに得られる報酬の最大化を目指す定式化であり，報酬に関する探索と活用のジレンマが存在することが知られている．

代表的な方策である Upper Confidence Bound (UCB) 方策 [1] は，標本平均と補正項の和からなる UCB スコアが最大の台を選ぶ方策である．これによって，エージェントは各時刻 t で期待報酬が最大でない台を $\mathrm{O}(1/t)$ 程度の確率で選択することができる．UCB 方策によって得られる累積報酬のオーダーは理論限界のオーダーと一致することが示されている．

2.2 最適腕識別

最適腕識別とは，期待報酬が最大の台を誤り確率 δ 以下で見つけるために必要な探索回数の最小化を目指す定式化である．この定式化は純粋探索問題であるため，報酬のジレンマは発生しない．

代表的な方策である Lower and Upper Confidence Bounds (LUCB) 方策 [2] では，台の選択基準に UCB スコアを用い，最適腕と識別するための基準に LCB スコアを用いる．ただし，UCB スコアを求めるときに用いる補正項の値は累積報酬最大化のそれとは異なる．LUCB 方策に基づいて台を選択することで，最適腕を識別するのに必要な探索回数のオーダーは理論限界のオーダー $\mathrm{O}(\log(1/\delta))$ と一致する．

2.3 しきい値バンディット問題

しきい値バンディット問題とは，時刻 T で全ての台を優良腕と劣悪腕に分類したときの誤分類率の最小化を目指す定式化である，ここで，優良腕は期待報酬がしきい値以上の台を，劣悪腕は期待報酬がしきい値未満の台を表す．この定式化も純粋探索問題であるので，報酬のジレンマは発生しない．

代表的な方策である Anytime Parameter-free Thresholding (APT) 方策 [3] では，しきい値と標本平均が近い台を多く選ぶ．この方策の時刻 T での誤分類率は理論限界のオーダー $\mathrm{e}^{\mathrm{O}(-T)}$ と一致する．

3 優良腕識別

我々が提案する優良腕識別とは，優良腕を誤り確率 δ 以下で見つけるために必要な探索回数の最小化を目指す定式化である．ここで，優良腕とは期待報酬がしきい値以上の台を表す．この定式化は純粋探索問題であるため報酬のジレンマは発生しないが，信頼に関する探索と活用のジレンマという新たな困難がある．本節では，問題設定と優良腕識別に必要な探索回数の下界について述べる．

3.1 問題設定

1 人のエージェントが K 台のスロットマシンの中から 1 台を選びプレイすることを繰り返して，誤り確率 δ 以下で優良腕をできるだけ速やかに見つけることを考える，ここで，優良腕とは期待報酬が ξ 以上の台を表す．各台の期待報酬を $\{\mu_i\}_{i=1}^K$ で表し，一般性を失わず以下を仮定する．

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m \geq \xi \geq \mu_{m+1} \geq \cdots \geq \mu_K$$

ここで，m は優良腕の数を表す．エージェントは誤り確率 δ 以下で期待報酬がしきい値 ξ 以上の台を見つけたと判定したら，この台を優良腕として出力し，探索対象からこの台を除外する．そして，優良腕として出力可能な台がなくなったと判定したら $\perp$

(NULL) を出力し停止する．

エージェントが λ 台のスロットマシンを優良腕として出力するまでに必要な探索回数を τ_λ，エージェントが $\perp$ (NULL) を出力するまでに必要な探索回数を τ_{stop} と表し，ある方策が (λ,δ)-PAC (Probabily Approximately Correct) および δ-PAC であることを次のように定義する．

定義 1 ((λ,δ)-PAC)**.** 優良腕が λ 個以上あるとき，出力した λ 個の優良腕のいずれかが誤りである確率が δ 以下である．優良腕が λ 個未満のとき，$\perp$ (NULL) を出力するまでに λ 個以上の優良腕を出力する確率が δ 以下である．

定義 2 (δ-PAC)**.** 全ての $\lambda \in \{1,\dots,K\}$ に対して (λ,δ)-PAC である．

この定義より，優良腕識別は方策が δ-PAC であるという制約のもとで $\{\tau_1,\tau_2,\dots,\tau_{\mathrm{stop}}\}$ を同時に最小化することを目指す問題となる．

3.2 探索回数の下界

優良腕識別に必要な探索回数の下界は以下の通りである．

定理 1. 任意の (λ,δ)-PAC 方策は以下の条件を満たす．

$$\mathbb{E}[\tau_\lambda] \geq \left(\sum_{i=1}^{\lambda} \frac{1}{d(\mu_i,\xi)} \log\frac{1}{2\delta}\right) \frac{m}{d(\mu_\lambda,\xi)}$$

ここで，$d(\cdot,\cdot)$ はベルヌーイ分布のカルバック・ライブラーダイバージェンスを表す．

定義 1,2 より，(λ,δ)-PAC は δ-PAC より弱い制約であるので，λ が固定されている場合は，δ-PAC 方策よりも (λ,δ)-PAC 方策の方が小さい τ_λ を達成できることが期待される．しかしながら，(λ,δ)-PAC 方策と同程度の τ_λ を δ-PAC 方策が達成可能なことを 4.2 節で示す．

4 優良腕識別の方策

本節では，ナイーブな方策と提案方策の紹介と，これらを用いたときの探索回数の上界について述べる．

4.1 ナイーブな方策とその探索回数

ナイーブな方策として，LUCB 方策としきい値バンディット問題の APT 方策と同じ台の選択基準を用いることが考えられる．これらをそれぞれ LUCB-G 方策と APT-G 方策と呼ぶ．これらの方策を用いたときの探索回数は $\mathrm{O}(K\log(1/\delta))$ であり，これは優良腕と劣悪腕をほぼ一様に探索することを意味している．

4.2 提案方策とその探索回数

我々は優良腕識別の方策として，Hybrid algorithm for the Dilemma of Confidence (HDoC) 方策を提案する．HDoC 方策では，台の選択戦略に累積報酬最大化の UCB スコア $\tilde{\mu}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{\log t}{2N_i(t)}}$ を，優良腕の識別基準に最適腕識別の LCB スコア $\underline{\mu}_i(t) = \hat{\mu}_i(t) - \sqrt{\frac{\log(4KN_i^2(t)/\delta)}{2N_i(t)}}$ を用いる．ここで，$\hat{\mu}_i(t)$ は台 i の時刻 t での標本平均，$N_i(t)$ は台 i を時刻 t までに探索した回数を表す．

これを用いたときの探索回数は $\mathrm{O}(\lambda\log(1/\delta) + (K-\lambda)\log\log(1/\delta))$ であり，HDoC 方策がナイーブな方策よりも期待報酬が高い台を優先的に引くことができ，少ない探索回数での識別を実現していることを意味している．

5 数値実験

ナイーブな方策と提案方策を用いた数値実験の結果を示す．数値実験は，しきい値バンディット問題に基づく設定: Synthetic 1–3 と医療データに基づく設定: Medical 1,2 で，許容誤り確率を $\delta = 0.05$ として 1000 回行った．表 1 に各方策の探索回数の平均値を示す．この結果より，Medical 2 を除くほぼ全ての設定で HDoC 方策が効率的に探索ができることが確認できる．

表 1　各方策の平均探索回数.

Synth. 1	τ_1	τ_2	τ_3	τ_4	τ_5	τ_{stop}
HDoC	**114.0 ± 21.8**	**146.7 ± 22.6**	**186.8 ± 34.4**	**778.7 ± 741.1**	**5629.2 ± 1759.6**	10264.1 ± 2121.1
LUCB-G	134.4 ± 26.6	167.3 ± 27.5	197.0 ± 31.2	798.0 ± 246.5	5702.9 ± 1589.9	10258.2 ± 2054.9
APT-G	6067.5 ± 1789.3	6282.5 ± 1810.9	6473.4 ± 1820.9	8254.0 ± 1909.5	10161.3 ± 2062.0	10243.0 ± 2062.0

Synth. 2	τ_1	τ_2	τ_3	τ_{stop}
HDoC	**202.2 ± 106.7**	825.8 ± 1048.7	**5237.2 ± 1614.8**	10001.2 ± 2051.5
LUCB-G	259.5 ± 120.4	**763.7 ± 260.4**	5566.6 ± 1575.5	9961.8 ± 1957.7
APT-G	6891.3 ± 1776.4	7990.7 ± 1839.1	9971.4 ± 1976.5	10048.9 ± 1976.5

Synth. 3	τ_1	τ_2	τ_3	τ_{stop}
HDoC	**7081.3 ± 2808.4**	**10955.9 ± 2954.4**	18063.0 ± 9252.0	46136.6 ± 4699.4
LUCB-G	10333.0 ± 3330.0	14183.6 ± 3186.7	**17162.8 ± 2997.7**	46059.1 ± 4740.4
APT-G	44326.0 ± 4708.1	45212.9 ± 4727.2	45676.7 ± 4727.8	45852.7 ± 4727.8

Medical 1	τ_1	τ_{stop}
HDoC	**10170.1 ± 4276.6**	31897.1 ± 5791.0
LUCB-G	10524.9 ± 3189.7	31827.9 ± 5786.0
APT-G	30847.1 ± 5724.9	31571.7 ± 5829.4

Medical 2	τ_1	τ_2	τ_3	τ_4	τ_{stop}
HDoC	17748.1 ± 10045.8	36695.0 ± 14006.3	69362.3 ± 18170.8	829915.2 ± 9976.9	875900.0 ± 284915.5
LUCB-G	**17741.0 ± 10073.5**	**36169.2 ± 13504.2**	**67343.4 ± 18105.0**	**821671.8 ± 6713.5**	865808.6 ± 298620.9
APT-G	623465.0 ± 238518.3	674725.3 ± 232943.8	710539.1 ± 243000.9	855729.7 ± 286830.9	861223.9 ± 286280.9

6 まとめ

本論文では多腕バンディット問題の新たな枠組みである優良腕識別を提案した．そして，優良腕識別を効率的に解く HDoC 方策を開発しその性能が理論限界のオーダーと一致することを理論的に示した．さらに，実験的にも HDoC 方策が有効であることも確認した．

参考文献

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[2] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, pages 655–662, 2012.

[3] A. Locatelli, M. Gutzeit, and A. Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1690–1698, 2016.